
Tests of Knowledge Recovery Can Introduce the Information They Seek to Recover

Neal Krishna

Independent researcher

With Apart Research

Abstract

When a model stops giving an answer, has it forgotten the information or only stopped expressing it? Tests of recovery intervene on the model to bring the answer back. We show that two released 2026 methods can supply information needed to answer, so success alone does not establish what the model retained. Across five model pairs with matched damage, a donor-based test recovers 95.3% of learned associations and 85.6% of associations the tested models never encountered. Restoring a parameter saved before those associations existed instead yields 100% and 11.9%. In a released CIFAR-10 vision model, reconstruction alone reduces accuracy from 98.54% to 35.39%; normalized full-SAE reconstruction yields 98.66%. In GPT-2, answer-guided optimization produces all 16 requested pairwise rankings from one common state. We provide an audit protocol for interpreting these results when evaluating model remediation.

1. Introduction

When a model stops producing an answer, it may have lost the relevant information or only the ability to express it. One way to investigate is to change the model's internal activations and ask whether the answer returns. But the intervention can supply the information needed to answer. A successful test therefore raises a further question: what did the model know before the test began?

We study this question in two released 2026 methods [1,2]. One brings in activations from another model; the other optimizes an activation change toward a specified answer. We call the model being tested the *recipient* and a model providing activations the *donor*. Both methods use a sparse autoencoder (SAE), which represents activations using learned features. Our concern is the information introduced through these operations.

The decisive experiment compares models with known learning histories. We create random associations after freezing one recipient, then teach them to a second model. After the same controlled damage, a released donor-based intervention recovers 95.3% of the learned associations and 85.6% of associations the never-exposed models did not learn. Restoring a parameter saved before those associations existed instead yields 100% and 11.9%. The donor-based test can make both histories look like successful recovery; the parameter repair distinguishes them. Swapping donors further shows that the returned answers follow the information supplied by the donor.

This distinction is critical when a recovered answer is used as evidence that remediation left the targeted information in the model. METR's investigation agenda asks whether evaluations accurately characterize

model propensities and whether remediation robustly mitigates them [17]. CoSAI calls for performance and safety validation before redeployment, followed by phased restoration with heightened monitoring [18]. If an investigator brings back an answer by editing a remediated model, the result must be interpreted in light of what that edit adds. Our protocol for incident investigators records those inputs and checks whether the same test also succeeds in a model known not to have learned the tested associations.

Earlier work already addresses the underlying inference. Deeb and Roger avoid reteaching evaluation facts when testing unlearning [3]. Makelov et al. show that successful activation manipulation need not identify a naturally used mechanism [4], and Kwon et al. treat injection into a null recipient as memory induction [5]. OpenUnlearning studies the faithfulness and robustness of unlearning metrics [6]; J-Access uses reference models and recovery controls [7]. Chatzoudis et al. show that SAE reconstruction itself can change predictions [8].

We build on these controls by measuring what two specific released implementations do. The experiments ask three questions: does a donor supply the recovered answer; does reconstruction change the behavior being measured; and can an optimizer introduce a requested answer into a common model state? Together, the results distinguish evidence about retained information from the ability to produce an answer under intervention.

2. What a test of recovery needs to distinguish

A recovery score measures the model and the intervention together. The intervention may provide donor activations, the correct answer as an optimization target, or labels used to select features. These inputs can help produce a correct answer even when the model being tested never learned the association.

We make that possibility observable by controlling the learning history. First, freeze a model and the components used to test it. Next, generate new random associations and train only a donor on them. A test that now makes the frozen model express those associations has introduced information from outside that model. Holding the recipient fixed while changing the donor's answers identifies which information it follows.

To compare the two learning histories, we also test a model that learned the associations. We apply the same damage to learned and never-exposed models, then compare donor-assisted recovery with a repair drawn from a checkpoint saved before the associations were generated. This assesses whether successful recovery identifies which model previously learned the associations.

Appendix C gives the information-theoretic formulation. Its central point is that recovering an answer constrains the combined information in the recipient and the test. Attributing that answer to the recipient requires accounting for information supplied by the test [9].

3. A donor can supply answers the recipient never learned

3.1. Freeze the recipient before creating the associations

We use the released intervention code, called a restoration hook, unchanged [10]. It changes selected SAE features at patch positions and moves the classification-token state (CLS) toward the donor's state. At strength one, it replaces CLS with the donor's CLS. We apply the primary intervention after the first of two trained Transformer blocks, so the recipient still processes the changed activations before producing an answer.

For each of five confirmation seeds, a width-32 Transformer first learns 128 entity-to-label associations spanning eight output classes. We call these the retain task and use them to check whether the intervention disrupts previously learned behavior. We then fix the recipient, SAE, selected features, and remaining

Transformer block and output head. Only afterward does an operating-system random seed assign balanced eight-way labels to 64 separate entities. Donor A subsequently learns this map by fitting entity embeddings. Donor B learns its cyclic shift, $Y_B = (Y_A + 1) \bmod 8$. B disagrees on every entity but carries the same information under a known permutation; the two maps are not independent.

Both donors share the recipient's frozen suffix and readout and reach 100% accuracy on their own associations. The primary hook receives no test labels or class gate. Recipient parameter hashes remain unchanged. Unpatched baseline logits and hidden states are checked for exact equality before and after the intervention runs. The pilot seed is excluded from confirmation.

This experiment tests the released intervention in a trained small Transformer with a known learning history. The SAE has 64 latents and 20 coordinates selected by variation on retain activations, rather than the original ViT and class-specific feature-selection pipeline. The recipient already knows all eight output labels. What it has never encountered is the subsequently sampled entity-to-label mapping.

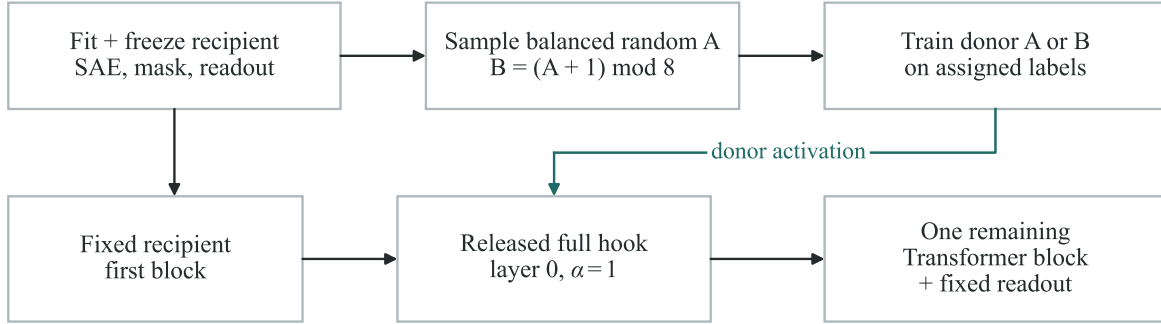
3.2. Changing the donor changes the recovered answer

The released hook strongly favors the supplied donor's mapping (Table 1). Changing only the donor changes the predicted class on 305/320 entities (95.3%). Donor-specific accuracy ranges from 82.8% to 93.8% across the five seed pairs. Accuracy on the 128 fitted retain entities per model remains 100% under either main intervention.

Intervention on the same recipient	Agreement with A	Agreement with B
No hook	38/320 (11.9%)	37/320 (11.6%)
Released hook with donor A	281/320 (87.8%)	6/320 (1.9%)
Released hook with donor B	2/320 (0.6%)	275/320 (85.9%)

Table 1. Five recipient training runs, each evaluated on the same 64 entities under alternative donors. The main intervention uses the first block, strength one, and no label gate. These are repeated predictions across five models, not independent model replications for every table entry.

(a) Target assignment follows recipient freeze



(b) The supplied donor changes the recipient's answers



Figure 1. Donor-controlled answers after the recipient is frozen. (a) The recipient, SAE, readout and label-free feature selection are fixed before assigning balanced eight-way labels A. $B = (A + 1) \bmod 8$ is a cyclic shift, not an independent map; only donors subsequently learn these associations. (b) Row-normalized confusion matrices pool five seeds with 64 canaries each (40 cases per A-label row). Within each seed, all conditions use the same recipient weights and cases. The unchanged released full hook at layer 0, $\alpha = 1$ leaves one Transformer block before the output: donor A produces diagonal agreement, whereas donor B produces shifted agreement. The unchanged released hook is tested in a trained small Transformer.

Across the same paired conditions, CLS-only intervention reaches 83.0% donor-map agreement, compared with 86.9% for the full hook and 15.0% for patch-only intervention. Component interventions are diagnostic wrappers that restore the recipient's non-tested tokens after calling the hook. Shuffling donor entity inputs moves agreement toward the shuffled assignment: 74.2%, versus 15.2% with the original assignment. A target-free recipient copy used as donor remains near chance.

The full CLS path accounts for much of the measured transfer. A final-block CLS replacement exactly reproduces donor predictions and serves as a readout-substitution control. The earlier-block result shows transfer through a remaining attention block, with unchanged recipient parameters and no label gate.

3.3. Recovery succeeds in both learned and never-exposed models

We next apply the same parameter damage to two histories: a model that learned A and the never-exposed recipient. The damage adds a predeclared, label-independent vector to their shared CLS parameter, impairing ordinary predictions. No new training occurs. A separate repair restores only that parameter from the checkpoint predating target-label generation, retaining each model's own entity embeddings.

Condition, scored against A	Learned A	Never exposed to A
Damaged, no intervention	40/320 (12.5%)	40/320 (12.5%)
Released hook with donor A	305/320 (95.3%)	274/320 (85.6%)
Target-free CLS parameter repair	320/320 (100%)	38/320 (11.9%)

Table 2. Matched parameter-damage controls across the same five checkpoints. The target-free repair supplies neither donor activations nor target associations and returns each model exactly to its own undamaged state.

The donor helps both models answer correctly, including the model that never learned the associations. Restoring the saved parameter recovers them only in the model that learned them. This comparison shows that the two histories can be distinguished in our setup, even though donor-assisted recovery succeeds in both.

The continuous hook scores retain potential diagnostic information: learned-recipient scores range from 93.8% to 98.4%, versus 81.3% to 90.6% in never-exposed recipients. Those ranges do not overlap in these five pairs. A cutoff between these observed ranges separates the two histories in these five pairs. To use recovery accuracy as a test of prior learning, an investigator would need to select the cutoff on calibration models and evaluate it on separate models with known learning histories.

3.4. Information can also pass through patch activations

To test transfer without whole-CLS replacement, a separate three-seed, mean-pooled Transformer preserves recipient CLS. Patch-only recovery reaches 76/192 A assignments and 66/192 B assignments. Shuffling donor IDs moves agreement toward the shuffled assignments (77/192 and 70/192), versus 19/192 and 18/192 for the original maps.

A prespecified follow-up fits the SAE dictionary and class-F1 experts after donor associations exist; these components therefore belong among the test's information-bearing inputs. For class 7, which has no positive recipient-training examples, unit-strength patch-only recovery is 1/48, and a TopK-encoded alternative yields 0/48. At strength ten, patch-only recovery reaches 27/48 in two of three seeds, with no class-7 false positives among 336 non-class-7 targets or 756 retain queries. Overall retain accuracy is 753/756. Holding dictionary and experts fixed while crossing donors produces 24 class-7 predictions, all following the activation donor. Appendix A retains the failed seed, reserved-membership controls, and reconstruction limits.

4. The recovery procedure can change the measurement

4.1. Reconstruction changes an unchanged vision model

We evaluate released original vision Transformers (ViTs), SAEs, statistics, and expert indices [11]. After initial 160-image calibrations, we freeze the compared paths and evaluate all 10,000 CIFAR-10 test images [12] and all 3,925 validation images from the 160-pixel ImageNet release [13]. The donor and recipient are identical copies of the original model; no training or unlearning occurs. We apply the intervention at layer 9 to every evaluated example, leaving two Transformer blocks.

Despite the donor and recipient being identical, the released hook reduces CIFAR accuracy from 98.54% to 35.39%. ImageNette remains above 99% (Table 3). Strengths zero and one and an independent raw dense reconstruction yield exactly identical logits on both datasets. Donor-prefix, manual-prefix, and unchanged-CLS checks also agree exactly. Because no information differs between donor and recipient, these controls isolate the effect of reconstructing the activations. The large CIFAR error and class bias arise within the recovery procedure itself.

Reconstruction path	CIFAR-10 correct / 10,000	ImageNette correct / 3,925
Original model	9,854	3,907
Released hook at zero or one; independent raw dense path	3,539	3,895
Raw full SAE: latent bias and TopK	9,696	3,887
Supplied-statistics normalization, full SAE, denormalization	9,866	3,902

Table 3. Same-model controls on complete evaluated populations. The full sets include the initial subsets; they are not independent model replications.

The raw CIFAR path predicts bird on 6,205 images: 997/1,000 birds and 5,208/9,000 nonbirds. Baseline and normalized reconstruction have 12 and 6 bird false positives. If the code applied its true-label gate, examples outside the chosen class would avoid the intervention and its errors. This follows from the implementation; we did not run a separate gated arm. The normalized full SAE changes 75 CIFAR and 16 ImageNette baseline predictions. It sharply reduces reconstruction error on both datasets, while utility changes differ: CIFAR’s large collapse does not recur on ImageNette. Appendix A reports geometry, subset history, and the unchanged full-population protocols.

4.2. Changing the decision rule also improves accuracy

A fixed application of Selective Synaptic Dampening (SSD) [14] lowers bird accuracy to 444/1,000. In an exploratory comparison, normalized same-image donor restoration at strength ten reaches 833/1,000 birds and 97.09% overall accuracy. Unpatched bird AUROC is already 0.9908. Subsequent donor-free fivefold scalar logit calibration, supervised by 8,000 labels per fold, yields 97.30% out-of-fold accuracy and 889 birds correct. Overall accuracy favors score calibration, while donor restoration correctly classifies 35 more nonbirds and produces one bird false positive versus 47. The apparent improvement therefore depends on which errors the evaluator counts. Appendix A reports all 20 interventions and this post-outcome baseline.

5. An optimizer can introduce the requested answer

For the second method, we hold the model state fixed and change which answer the optimizer is asked to produce. The released recovery code [15] adjusts internal activations to raise a target name above a paired subject name. The correct answer therefore enters the test through its objective. The adjustment is bounded in norm and constrained to preserve selected SAE features.

We start the input to GPT-2 block 4 from one identical 14-token activation tensor derived from a neutral sentence. All original indirect-object identification (IOI) token states and caches are absent; blocks 4–11 and the frozen output head remain. The reference and feature mask are shared across cases. For each requested target and paired subject, optimization starts from a zero perturbation and uses the same settings: 100 steps and norm budget 20. We enforce the feature constraint on the actual update after every step, correcting the implementation issue described below.

Using the prespecified zero-clamp condition, which sets the selected SAE feature values to zero, all 16 directed target-versus-subject rankings become positive, across eight unordered name pairs. Opposite directions begin from the same state, so the initial 8/16 positive margins are a structural balance property. They are not a chance estimate or 16 independent trials. The name pair is side context; optimization receives the target orientation. Success concerns pairwise ranking, not full-vocabulary top-one prediction. Mean selected-feature drift is 3.03×10^{-6} and mean encoder-rowspace fraction is 4.29×10^{-8} .

Separately fitted final-readout and released negative-five-clamp controls also yield all 16 pairwise rankings.

Replaying only the residual-change component succeeds on 5/16 directions under the zero clamp and 0/16 in the other two controls. Replaying the decoded change succeeds on all directions. These components are separate interventions and need not obey the full update's constraints or norm budget. The full intervention induces the requested rankings; residual-only replay does not reproduce that success.

5.1. Checking the constraint actually enforced

The implementation projects raw gradients before Adam. Coordinatewise rescaling can move the actual iterate outside the encoder-null subspace. We correct this by projecting the iterate after each update, followed by norm clipping. Fitting one shared perturbation across 32 real-prefix prompts, the original and corrected methods both restore exactly the same 14 of 19 clamp-induced failures. Mean selected-feature Euclidean drift falls from 4.64 to 9.52×10^{-6} . Both perturbations have norm 20. Correct enforcement preserves that recovery count, while overall pairwise accuracy falls from 27/32 to 25/32 because two previously unsuppressed cases are lost.

An established LayerNorm invariance [16] exposes a second boundary. Removing each update's mean preserves all tested pairwise signs and vocabulary logits within the prescribed tolerance, yet changes whether the update passes a descriptive feature-drift threshold. Centering changes membership in the coordinate-specific feasible set even when outputs agree to tolerance. Appendix B reports the numerical checks.

Newly fitted updates that also enforce zero mean preserve all 16 positive target rankings. The centered descriptor is re-referenced at the unchanged defended state; direct preactivation checks confirm the added constraint. Selected centered ReLU outputs remain inactive under both methods, so this establishes feasibility without mean cancellation, not an improved sparse-feature monitor.

6. What investigators should report

The three experiments concern different parts of the same inference. Donor activations can supply associations the recipient never learned. Reconstruction can change an unchanged model's predictions. An answer-guided optimizer can produce opposite rankings from the same starting state. Each result shows why the recovery procedure belongs in the explanation of a recovered answer.

An investigation should therefore distinguish the model's unaided behavior, the behavior obtainable under a specified intervention, and the evidence about what the model previously learned. Our protocol starts by saving unassisted outputs and recording every donor, label, optimization target, and external state supplied during recovery. It then applies the same intervention to a never-exposed recipient and checks reconstruction on an unchanged model. Corrections and new errors are reported together.

If the test also produces the answer in the never-exposed recipient, success establishes what that intervention can induce. A claim of retained information needs a comparison that separates the histories. One established approach learns a recovery map on independent facts, freezes it, and evaluates on held-out

facts [3]. J-Access instead evaluates model-level predictions against later relearning [7]. Our controls focus on information introduced by the recovery operation itself.

Appendix F provides a checklist for recording these comparisons and deciding whether a recovered answer supports a claim about retained information.

7. Limitations

Incident applicability. We test evaluation methods using controlled Transformers, released ViTs, and GPT-2. We do not test models involved in the public incidents or validate a decision to return an affected system to service.

Controlled learning histories. Shared frozen Transformer blocks and output heads deliberately favor donor compatibility. Targets are random associations learned by the donors, and retained-task utility is measured on fitted entities. The matched parameter damage is an experimental control, not an evaluated unlearning algorithm. Withholding class-7 positives still leaves negative softmax training and an output direction for that class. The reserved-membership primary fails; its codebook variant transfers rankings without establishing accurate concept restoration.

Vision-model experiments. The recovery extension uses one partially impaired SSD checkpoint and reuses its original SAE. Score calibration was added after observing outcomes on the same population and has a different information budget; out-of-fold predictions are not independent validation. We do not evaluate successfully unlearned recipients, the authors' class-specific unlearning pipeline, natural semantic deletion, or unseen-entity generalization. The unchanged-model calibrations cover complete test or validation sets for two released checkpoints, but do not identify which implementation produced the original paper's tables. The release does not specify the ImageNette source-resolution archive or establish the historical SAE normalization convention; we use the 160-pixel ImageNette archive.

GPT-2 experiments. We use a public-weight adapter and benign name contrasts, rather than the paper's full official execution. The common state removes instance context while preserving pretrained name knowledge. SAE reconstruction is poorly calibrated after optimization even under the milder zero clamp. Preserving selected features does not establish preservation of the full representation. The replayed components are separate, potentially much larger interventions. These experiments therefore do not refute SAE recovery or residual localization in general.

8. Conclusion

A released test of knowledge recovery produced 85.6% of target associations in models that had never encountered them, compared with 95.3% in models that had learned them. Successful recovery can therefore reflect information supplied by the test. For investigators evaluating remediation, a recovered answer alone does not establish that the model retained the targeted information.

The controls identify what additional evidence is needed. Restoring a parameter saved before target generation recovered learned associations while leaving never-exposed models near chance. Identical-model reconstruction checks exposed errors introduced by the test, and a common starting state revealed target information supplied by optimization. Evaluations should record these information sources, measure new errors alongside recovered answers, and validate retention criteria against known learning histories before using them to judge remediation.

References

- [1] Jang, Y., Lee, J., Kim, D., Jo, J., and Woo, S. S. Suppression or Deletion: A Restoration-Based Representation-Level Analysis of Machine Unlearning. The Web Conference, 2026. Version 1, February 18, 2026.
- [2] Cui, M., Shen, L., and Yang, X. SAE Interventions are Unreliable: Post-Intervention Recovery of Suppressed Behavior. arXiv:2606.18322v1, June 16, 2026.
- [3] Deeb, A., and Roger, F. Do Unlearning Methods Remove Information from Language Model Weights? arXiv:2410.08827, revised February 7, 2025.
- [4] Makelov, A., Lange, G., Geiger, A., and Nanda, N. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. ICLR, 2024. Conference paper.
- [5] Kwon, J., Kim, D.-K., Kim, J., Kim, Y., Kook, W., and Cha, M. AI Engram: In Search of Memory Traces in Artificial Intelligence. arXiv:2606.14997v1, June 12, 2026.
- [6] Dorna, V., Mekala, A., Zhao, W., McCallum, A., Lipton, Z. C., Kolter, J. Z., and Maini, P. OpenUnlearning: Accelerating LLM Unlearning via Unified Benchmarking of Methods and Metrics. arXiv:2506.12618v2, November 9, 2025. First submitted June 14, 2025.
- [7] Song, Z., Liu, H., Wang, X., Li, S., Li, X., Gao, L., Zhang, J., Lu, Z., Ji, F., Chang, X., and Chen, X. Measure, Don't Optimize: Forecasting Recovery in LLM Unlearning. arXiv:2608.11408v1, August 11, 2026.
- [8] Chatzoudis, G., Li, Z., Moran, G. E., Wang, H., and Metaxas, D. N. Beyond Interpretability: When, Why, and How Sparse Autoencoders Enable Label-Free Visual Steering. arXiv:2506.01247. Version 3, May 26, 2026. First submitted June 2, 2025.
- [9] Cover, T. M., and Thomas, J. A. Elements of Information Theory. Second edition. Wiley, 2006. Publisher record.
- [10] Jang et al. Released restoration hook, commit 4828ff93eb1d0031aa00319adf7726912e7126c9. Pinned source.
- [11] Jang et al. Released CIFAR-10 and ImageNet ViTs, SAEs, statistics, and expert assets, revision fd43ec57144116b2b77d4fb0e0fb695b021e353a. Pinned model repository.
- [12] Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. Report.
- [13] Howard, J., and fastai contributors. Imagenette dataset. Official repository. Accessed September 12, 2026.
- [14] Foster, J., Schoepf, S., and Brintrup, A. Fast Machine Unlearning without Retraining through Selective Synaptic Dampening. Proceedings of the AAAI Conference on Artificial Intelligence, 38(11):12043–12051, 2024. Published paper. Pinned implementation.
- [15] Cui et al. Released recovery core, commit 5a80dfca7bff0a06cfc62441e7ef62b12ef8a4a3. Pinned source.
- [16] TransformerLens contributors. Main Demo Notebook, section on folding LayerNorm and centering writing weights. Official documentation. Accessed September 12, 2026.
- [17] METR. How independent researchers could investigate AI propensities after misalignment incidents. July 28, 2026; updated September 5, 2026. Investigation agenda.
- [18] Coalition for Secure AI. AI Incident Response Framework V1.0. October 27, 2025. Sections 1.3.2 and 3.3.3. Framework.

Appendix A. Complete donor controls and replication units

The five confirmation seeds are 23, 41, 59, 83, and 97. All 57 planned rows per seed are retained; the seed-17 pilot is excluded. Multiple donors, strengths, gates, and component interventions reuse each recipient and target map. Every secondary true-label-gated row preserves nongated outputs exactly, as expected from the gate. Those rows are not evidence of intrinsic class selectivity. Strength zero still reconstructs patch tokens and is distinct from no hook.

Earlier-layer full hook	Correct A with donor A	Correct B with donor B
Strength zero	36/320	40/320
Strength one	281/320	275/320
Strength ten	246/320	241/320

The published extrapolation strength ten therefore also transfers associations in this fixture, but less accurately than the primary strength-one condition. At the final block, strength one reproduces both donors exactly; strength ten yields 320/320 and 319/320. Matched parameter-damage tests use strength one only. All 60 measured strength-one fitted-retain checks are 100%; retain utility for strengths zero and ten is not measured.

The matched damage uses a fixed zero-mean unit vector, scaled by 10,000, added to the shared CLS parameter. Hashes confirm that all other parameters are unchanged, hooks do not mutate the damaged models, and copying the pre-label CLS parameter returns each to its own original checkpoint. The descriptive rule of accuracy above one half is positive in every learned and never-exposed donor-restoration history. Choosing a separating threshold from their observed non-overlapping scores would require an independent calibration set.

The patch-coordinate extension uses seeds 131, 137, and 149, 126 retain IDs per model, mean pooling, and the public Autoencoder/TopK implementation with a return-shape adapter. Donors learn 190/192 A and 188/192 B associations. Their reserved-class competence is imperfect: 22/24 and 20/24. Earlier-layer class-7 recall remains 0/24 for both donors at every tested strength. The patch-only shuffled and target-free controls use frozen checkpoints without new training.

The separate reserved-membership study runs seeds 211, 223, and 227 for each of two prespecified readouts. Each cell has 32 positive and 32 negative new memberships. The learned-readout donors achieve sensitivities 23/32, 23/32, and 19/32, failing the 0.9 feasibility criterion. Fixed-codebook donors meet all prespecified criteria. Their main hooked recipient produces 5, 10, and 7 true-positive decisions, with no false positives, and AUROCs 0.9844, 0.9775, and 0.9951. With donor IDs permuted, mean AUROC is 0.5407 against the original assignment and 0.9831 against the donor's permuted assignment. No threshold is fitted on evaluation labels. Strength ten increases detections to 68/96 but reduces fitted-retain accuracy to 55.0%. The readout variants use different random memberships and are not a paired causal estimate of changing the classifier alone.

In the source-trained supplement, the public SAE uses 64 latents, TopK eight, and 400 fixed full-batch fitting steps on all donor patch states. Ten features follow a reconstructed class-F1 rule; feature selection is in-sample and uses donor labels. The runtime provides no complete training/selection pipeline, so activity threshold, tie handling, and fitting settings are declared reconstruction choices. The released hook bypasses the learned latent bias and TopK; the alternative uses the public forward latents before selected-coordinate interpolation. Recipient CLS is retained in the patch-only comparison. At strength ten, the TopK alternative recovers 17/48 class-7 cases with 754/756 retain predictions, versus 27/48 and 753/756 for the released encoding path. Both fail in seed 137. The dictionary alone is insufficient in this panel: target-free donors and empty expert lists yield no patch-only class-7 predictions. Normal donor SAE reconstruction itself reduces class-7 performance from 42/48 to 28/48; dense reconstruction reduces it to 11/48. Amplification-dependent recovery therefore mixes imported information with reconstruction quality.

A separate coordinate-compatibility control permutes actual recipient parameters and co-transforms its SAE, preserving standalone predictions. Across five learned, damaged recipients, full-hook target counts are 305, 50, and 305 out of 320 under original, unaligned-donor, and jointly aligned coordinates. Retain

counts are 639, 133, and 639 out of 640. This large utility loss allows a utility-aware test to reject the unaligned intervention. The prespecified float32 numerical certificate passes four of five seeds; one aligned-hook difference exceeds the 10^{-3} logit tolerance. A declared post-outcome float64 diagnostic preserves every class prediction and passes all five certificates, with maximum aligned-hook difference 1.75×10^{-12} , without revising the original failure. This is a compatibility and precision diagnostic, not a useful utility-preserving false-deletion example.

Released image-model calibration

The ViT-Base/16 checkpoint, SAE, statistics, and experts use the pinned revision. The initial CIFAR screen takes the first 16 images of each test class, then orders their original indices, retaining all 160. Baseline, released hook, raw full SAE, and normalized full SAE score 158, 51, 154, and 158 correct. Preprocessing uses bilinear resizing to 224 pixels and channel mean and standard deviation 0.5. The six primary arms preserve the original CLS token and use the released class-0 expert indices without a true-label gate. Identical donor and recipient patch states make expert interpolation zero.

The raw full-SAE comparison jointly restores latent bias and TopK, so it does not isolate their individual effects. In the initial CIFAR screen, the statistics-normalized path includes denormalization and changes logits by up to 2.4996 while preserving every baseline class prediction. Raw relative squared reconstruction error above one can coexist with high task accuracy: the raw full SAE scores 154/160. Geometric error and utility therefore remain separate measurements.

Four secondary arms zero the CLS token after layer 9. Zeroing alone leaves 157/160 correct and fails to create substantial suppression. Exact donor-CLS restoration returns baseline logits and 158/160; adding the released full hook gives 51/160, and raw patch reconstruction with zero CLS gives 17/160. These outcomes are retained as calibration controls, not promoted as an unlearning recovery demonstration.

Independent validation recomputes predictions, class counts, equality checks, and summaries from all ten saved full-logit arrays and per-image records. Reconstruction means are checked from the per-image metrics rather than an independent neural rerun. The raw path has bird precision 16/102 and false-positive rate 86/144; all ten class confusion matrices are preserved. These controls do not estimate a population frequency across models or unlearning methods.

The separate ImageNette calibration retains the first 16 validation JPEGs per class from the 160-pixel archive, using the released ImageFolder transform and layer-9, k32 SAE. Every primary arm remains correct on all 160 images. Raw dense, raw full-SAE, and normalized full-SAE mean relative squared patch errors are 1.64652, 0.914765, and 0.0202438; maximum logit changes are 4.58772, 3.63028, and 2.77861. Thus unchanged predicted classes do not imply numerical identity. Zeroing CLS alone retains 160/160 correct; zero CLS with raw patch reconstruction yields 153/160. No alternate archive, subset, or preprocessing was tried after observing the result.

The frozen CIFAR extension retains all 10,000 test examples, including the initial 160, with the same ten arms and batch size 16. On the 9,840 examples outside the screen, baseline, raw reconstruction, raw full SAE, and normalized full SAE score 9,696, 3,488, 9,542, and 9,708 correct. Raw reconstruction changes 6,346 predictions in this complement and 6,455 across the full set. Normalized reconstruction changes 75 predictions, all outside the screen. Full-set mean per-image relative squared patch errors are 49.4014, 1.01604, and 0.017749 for raw dense, raw full SAE, and normalized full SAE. Raw bird precision is 997/6,205 and false-positive rate is 5,208/9,000. The normalized path's 12 additional correct predictions are a descriptive aggregate change, not an established general improvement.

The final ImageNette extension retains all 3,925 validation JPEGs from the same 160-pixel archive with unchanged arms. Raw reconstruction changes 23 baseline predictions, and normalized reconstruction changes 16. Mean relative squared patch errors are 1.64668, 0.914645, and 0.0201772 for raw dense, raw full SAE, and normalized full SAE. Thus the full-set extension retains the negative result for a CIFAR-scale accuracy collapse, while revising the subset's exact class-prediction preservation. Neither the low geometric error nor the aggregate accuracy establishes per-example identity. On the 3,765 examples outside the initial ImageNette screen, the four primary path groups score 3,747, 3,735, 3,727, and 3,742 correct in Table 3 order. Across both datasets, the original 160 cases reproduce byte-identical logits in all ten arms of the full runs.

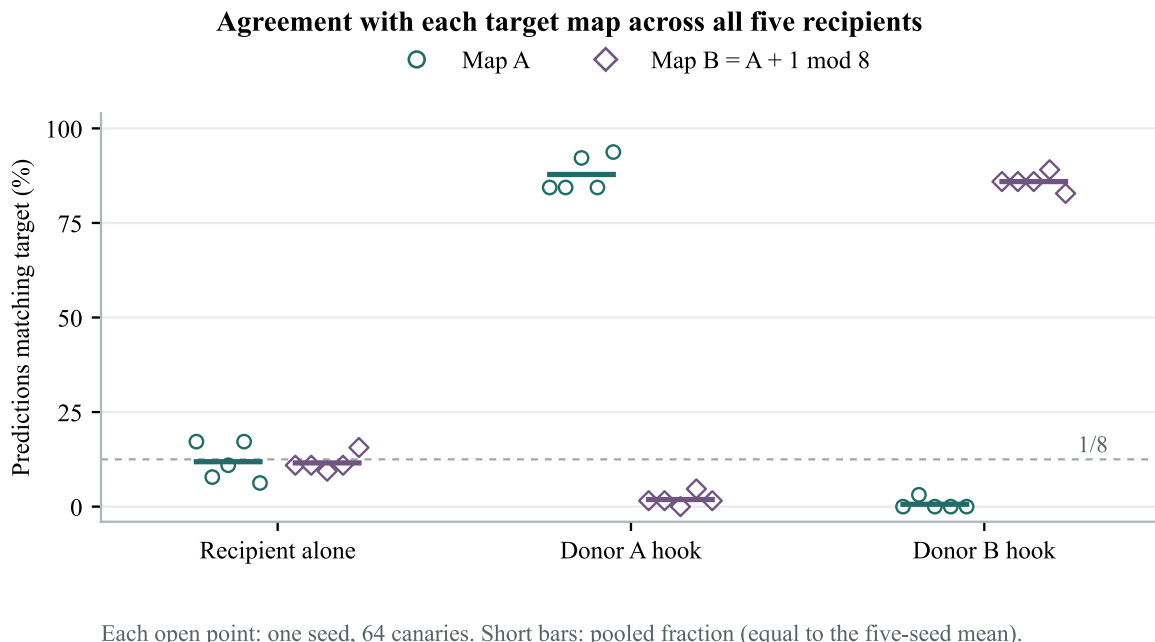


Figure S1. Every seed's agreement with the two target maps. Open points show five recipient seeds, each evaluated on the same 64 canaries across the three conditions. Horizontal marks show pooled fractions, equal to the seed mean because seed sizes are equal; they are not uncertainty intervals. Map B is the cyclic shift of A. The dashed 1/8 reference is the expected agreement of a target-independent prediction with a balanced eight-way assignment. No inferential independence across the 320 canaries is assumed.

Exploratory calibration on a partially impaired recipient

We apply the released SSD operator [14] once to the original CIFAR-10 ViT, using its full-class ViT defaults: selection weighting five, dampening one, batch size 64, and squared minibatch-mean cross-entropy gradients. Importance uses all 5,000 bird training images and all 50,000 training images, retaining source batch order and the checkpoint-compatible preprocessing. Bird accuracy falls from 988/1,000 to 444/1,000; nonbird accuracy changes from 8,866/9,000 to 8,872/9,000.

This run fails the original requirement of bird accuracy at most 20%, and its conditional restoration does not run. The following 20-arm study is designed after that outcome, with the saved partially impaired checkpoint fixed. It does not relax the failed criterion, tune SSD, or reproduce the original authors' unlearning table. Intervention settings are fixed before their outcomes are inspected. Both original-donor and recipient no-hook logits exactly reproduce their previous full-test references.

Every arm covers all 1,000 bird and 9,000 nonbird test images at layer 9. The 20 conditions comprise no hook; three strength-zero controls; raw, normalized-full-SAE, and residual-preserving paths at strengths

one and ten under same-image or shuffled original donors; and four matched CLS-only arms. All nonzero arms use the same whole-CLS interpolation. CLS-only preserves recipient patches. The original dictionary, statistics, and expert set remain fixed, without fitting an unlearned dictionary or alignment. Figure S2 reports every executed condition.

The normalized path combines supplied normalization, latent bias, TopK encoding, selected-coordinate mixing, decoding, and denormalization. The residual-preserving path adds only the decoded latent change to the original patches. For identical latent changes and CLS treatment, the difference between these paths is the unedited reconstruction error, a standard linear-decoder identity. Their comparison does not isolate normalization alone. Residual strength zero exactly reproduces no hook; normalized strength zero already changes 252 predictions, with 103 corrections and 47 new errors.

At strength ten with the same-image donor, normalized full-SAE restoration corrects 447 previous errors and introduces 54 new errors: 393 net additional correct answers. It reaches 833/1,000 birds and 8,876/9,000 nonbirds correct, with one bird false positive. Residual-preserving restoration reaches 817 and 8,865 correct, with 411 corrections, 45 new errors, and one bird false positive. CLS-only reaches 566 and 8,877, with 127 net additional correct answers and no bird false positives. Relative to its own zero-strength baseline, the normalized arm adds 337 correct answers; relative to matched CLS-only, it adds 266.

The released raw path behaves differently from the intact-model calibration. Strength zero produces no bird predictions and only 4,108/10,000 correct answers. Same-image strength ten yields 20/1,000 birds and 3,771/10,000 overall correct. The original-model broad bird bias does not recur; this recipient instead suffers severe default-prediction failure. Positive or negative responses to this operation therefore require their own calibration.

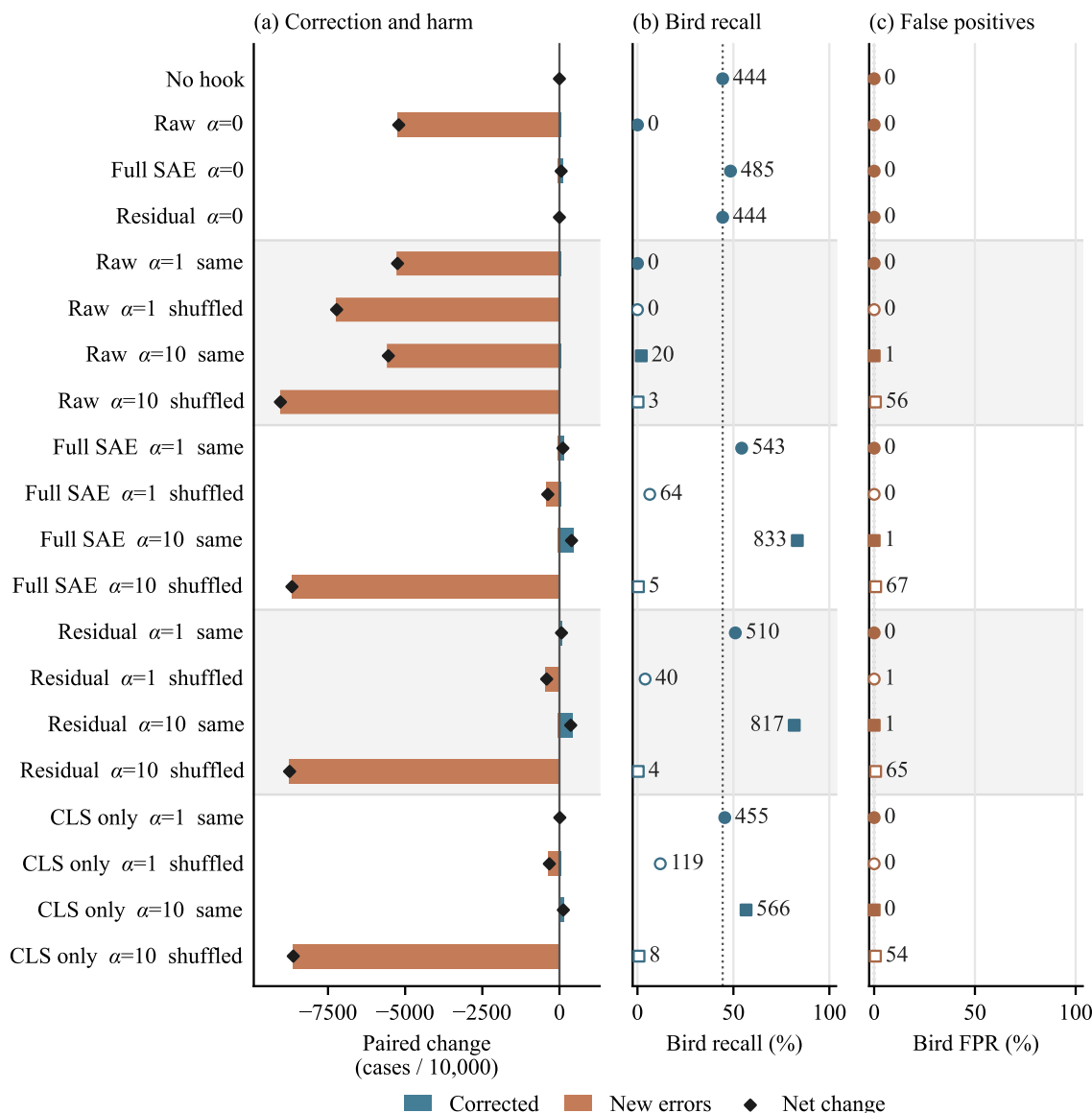


Figure S2. All 20 executed conditions on 10,000 CIFAR-10 images. The label "Full SAE" denotes normalized full-SAE reconstruction. (a) Corrected errors (blue), new errors (orange), and net correctness change (diamonds), relative to no hook (9,316/10,000 correct). (b) Recall among 1,000 birds. (c) False-positive rate among 9,000 nonbirds. Point labels give counts; dotted lines show no hook. Squares denote strength ten; hollow points use shuffled donors. Every nonzero SAE arm also interpolates CLS; matched CLS-only arms preserve patches. Derived oracle-label policies are excluded. This exploratory follow-up uses a partially impaired checkpoint; the earlier eligibility gate remains failed.

The first global shuffle from a fixed seed is retained without rejection: it has one fixed point, 1,035 same-class pairs, and 112 bird-to-bird pairs. On 8,965 true-label-discordant pairs, normalized strength-ten predictions agree with the donor class 7,021 times, versus 69 without a hook. Agreement remains above the class-conditional marginal expectation, a descriptive comparison rather than a retention test. Shuffled strength ten damages overall utility in every family: normalized, residual-preserving, and CLS-only arms leave only 646, 571, and 690 correct answers. Donor correspondence does not establish information originally present in the recipient.

Nineteen additional oracle policies are derived by taking each intervention's saved logits on true bird rows and no-hook logits elsewhere. They are logit-splicing calculations, not executed or deployable label-free arms. The raw same-image strength-ten policy shows 8,892 overall correct rather than 3,771 ungated, concealing 5,121 net lost nonbird answers. For the normalized arm, the derived policy removes four net nonbird gains, changing 9,709 to 9,705. Gating can hide harm or suppress benefit.

A timestamped secondary analysis was added after launch but before intervention-score inspection. It ranks the bird logit minus the log-sum-exp of the other nine logits, over all 9,000,000 bird/nonbird pairs, with half credit for ties. Table A1 gives the decision-versus-ranking counterpoint; all 21 readouts, including the original reference, remain in the artifact.

Readout	Bird correct / 1,000	Bird AUROC
Original reference	988	0.999848
Partially impaired, no hook	444	0.990754
Raw, same-image, strength ten	20	0.981163
Normalized full SAE, same-image, strength ten	833	0.999668
Residual preserving, same-image, strength ten	817	0.999732
CLS-only, same-image, strength ten	566	0.996664

Table A1. Default class decisions and continuous bird-versus-rest ranking on the same complete test set. The AUROC analysis fits no threshold.

The unpatched scores already distinguish birds well despite low recall. Even the raw strength-ten path retains high AUROC while almost never predicting bird. Its strength-zero AUROC is lower, 0.908264, so the failure is not simply a rank-preserving constant shift. The original donor carries task information; these measurements establish neither missing discrimination nor successful unlearning or semantic deletion.

Donor-free, label-supervised decision calibration

After the aggregate intervention outcomes were known, we declare one score-calibration follow-up. Its plan is recorded before individual saved-score inspection, followed by one fitted run. Five class-stratified folds are assigned by hashing the original test indices. In each fold, a single additive bird-logit bias maximizes ordinary ten-class accuracy on the other 8,000 labeled images and is applied once to 2,000 held-out images. All 10,000 predictions are out of fold. The search covers the complete finite float64 decision-state set, with exact class-index tie handling and a fixed closest-to-zero maximizing-bias rule.

No donor logits enter fitting, and model weights and representations remain unchanged. The procedure is donor-free but label-supervised: its calibration labels belong among the test's additional inputs. It differs from the pre-label parameter repair in Table 2, which supplies no target associations.

Condition	Correct / 10,000	Bird correct / 1,000	Nonbird correct / 9,000	Bird false positives / 9,000
Partially impaired, no hook	9,316	444	8,872	0
Cross-fitted scalar calibration	9,730	889	8,841	47
Normalized donor, same-image, strength ten	9,709	833	8,876	1
Original reference	9,854	988	8,866	12

Table A2. Label-supervised score calibration and donor intervention on the same test population. Counts describe different error tradeoffs, without an inferential superiority claim.

Against no hook, scalar calibration corrects 445 errors and introduces 31, versus 447 corrections and 54 new errors for donor restoration. All 445 scalar corrections are birds. Of its 47 bird false positives, 31 were previously correct nonbirds and 16 were already misclassified nonbirds. Comparing predictions directly, the scalar corrects 134 donor errors but loses 113 answers the donor gets right. Its 21-image overall advantage therefore does not dominate the donor's better nonbird accuracy and lower false-positive count.

The high unpatched AUROC and effective scalar correction show that much discrimination is already expressed in output scores. They do not establish that the donor operation acts only as a bias shift or is necessary for recovery. Cross-fitting excludes each row's own label from fitting its applied scalar, but study selection follows aggregate inspection of this same population. It is exploratory internal validation, with one partition and unequal information budgets, rather than an independent test of generalization. No new calibration algorithm, fitted pooled-AUROC claim, or latent-storage identification follows.

Appendix B. Optimizer, transfer, and calibration details

The primary GPT-2 comparison fits one shared perturbation on all 32 prompts; the task-independent controls fit a separate target-conditioned perturbation for each directed contrast. These protocols must not be pooled. In the primary comparison, 61 positive-attribution features remain after requesting the top 64, with encoder rank 61. The original perturbation has encoder-rowspace fraction 0.146, versus 7.33×10^{-8} after iterate projection. Both recover the same 14 valid failures, but corrected optimization loses two previously unsuppressed cases: total pairwise accuracy is 27/32 upstream and 25/32 corrected. This is not a Pareto improvement.

On 32 held-out prompts with 16 new names, frozen upstream and corrected shared perturbations recover 6/18 and 9/18 clamp-induced failures, respectively, without test-label fitting. The model still sees the full test prefixes. Partial transfer thus concerns recoverability through the full interface, not information uniquely stored in the locally intervened activation.

The released IOI script uses a negative-five decoder clamp; the paper's zero-clamp description motivates the separately prespecified sensitivity. A supplementary literal-zero test on the 32 real prefixes has 32/32 positive target margins both before and after clamping. Its valid-flip set is empty, so it stops before optimization without changing the feature mask. This negative concerns real-prefix suppression; the common-state zero-clamp induction experiment asks a different question. Each sensitivity run reoptimizes from zero under the same budget and projection settings. The final-readout control stores 32 rows, but

prompt removal leaves 16 distinct directed objectives; exact duplicate rows are retained. Layer-4 controls directly evaluate these 16 objectives. Opposite orientations remain dependent.

A direct source witness isolates the Adam problem: a selected encoder vector $(1, 2)$ and gradient $(2, -1)$ have zero inner product, but the first coordinatewise Adam step produces a nonzero encoder component. The source's ridge approximation is an additional reason not to claim exact projection. Projecting the actual parameter is an established correction. A final measured feasible witness remains useful even if the search procedure lacks a universal guarantee.

For SAE reconstruction $R(h) = D(E(h))$, the residual replay component is

$$\delta_{\text{res}} = \delta - [R(h + \delta) - R(h)] = [(h + \delta) - R(h + \delta)] - [h - R(h)]. \quad (1)$$

It is the optimized change in reconstruction error, not the original error. In the final-readout control, decoded components have norms 307–938 and residual components 292–931, although their sum has norm 20. They nearly cancel. Component replay therefore tests unconstrained sufficiency, and nonlinear downstream margins need not add.

Calibration further limits mechanism attribution. The zero-clamp whole-sequence state has relative SAE reconstruction error 0.593 before optimization and mean error 5.53 after it; active features increase from 395 to 2,674. The released negative-five primary setting produces much larger reconstruction error and feature activation. Direct preservation of selected features remains measurable in these states, but does not imply that the entire SAE decomposition is well calibrated.

The normalization replay keeps every saved defended tensor and perturbation fixed. It decomposes each update into its mean and centered components without renormalization, clipping, or refitting. Maximum full-versus-centered logit difference is 0.000561237; maximum baseline-versus-mean-only difference is 0.000328064. Both pass the prospective 0.001 tolerance with no pairwise sign disagreement. The feature threshold 0.0001 is descriptive, not a paper-defined or preregistered gauge-replay acceptance rule: the 16 original zero-clamp drifts are at most 3.34×10^{-6} , while centered drifts range from 0.0229 to 1.1183. For the corrected primary shared update, mean selected-feature Euclidean drift changes from 9.52×10^{-6} to 4.73 while the same 14/19 recoveries remain.

At the descriptive feature-drift threshold, all 16 original zero-clamp updates pass and no centered update does. In the stricter follow-up, maximum centered preactivation change falls from 2.13 to below 2.87×10^{-6} .

The stricter follow-up preserves the raw, whole-sequence defended state, original mask, objectives, and optimization settings. A separately centered SAE descriptor is re-referenced at that same state; the model input is not re-centered or re-clamped. With selected encoder columns A and mean direction $\mathbf{1}$, the actual iterate is projected into $\ker([A, \mathbf{1}]^T)$, increasing constraint rank from 64 to 65. Each new perturbation starts at zero. All 16 stored updates have absolute mean at most 1.13×10^{-8} ; maximum raw and centered preactivation changes are 3.22×10^{-6} and 2.87×10^{-6} . Pairwise margins remain positive but range from 1.21 to 83.25, compared with 53.63 to 77.04 for the original solutions. Only their success count is unchanged.

Centered selected-feature activation drift is zero for both the old and new updates because those selected ReLUs remain inactive. Eleven of sixteen new states have entirely zero centered encoding. This descriptor change therefore does not demonstrate an improved feature monitor. Its reconstruction statistic compares Qh with the unprojected decoder output and is not a quotient-space reconstruction norm. Raw reconstruction error remains large, about 5.31. The supported constructive result is feasibility under one additional, directly verified preactivation constraint.

Appendix C. Information accounting

Let Y be a target with K possible values, Z declared side context, and H the recipient state before the test. Let S include every additional input: donor activations, optimization labels, selected indices, and any prompt or cache information excluded from H and Z . Let g be fixed, with any data-dependent fitting or selection included in S . For an answer $\hat{Y} = g(H, S, Z)$ with error probability e , conditional Fano inequality and the chain rule give

$$I(Y; H|Z) \geq H(Y|Z) - h_2(e) - e \log_2(K - 1) - I(Y; S|H, Z). \quad (2)$$

Here $h_2(e)$ is binary entropy, and all logarithms are base two. This standard bound [9] identifies the missing condition: recovery accuracy constrains the recipient plus test. Attribution to the recipient requires controlling the test's additional information. Small perturbation norm or preservation of selected features does not supply an information bound; the perturbation's direction can encode a target.

Equation 2 follows by bounding $H(Y|H, S, Z)$ with conditional Fano and decomposing $I(Y; H, S|Z)$ by the chain rule. It requires finite target support and a complete side-input inventory. We do not estimate information in bits from pooled canary accuracies: balanced maps couple entity labels, and the intervention arms reuse recipients and assignments. A label-aware selector, donor, or cache cannot be omitted because it is considered part of the test. An exact balanced finite construction in the artifact illustrates a target-writing perturbation with zero prior target information; it is an explanatory example, not a substitute for the model experiments.

For the canary control, the probability space is the experiment before a fresh association map is revealed. Here Z contains the entity ID and preassignment context, and the map M is sampled independently of the frozen recipient. The recorded map and generator seed are postassignment data, excluded from this Z . With $Y = M(\text{ID})$, $I(Y; H|Z) = 0$ refers to this randomization ensemble. It is not an empirical mutual-information estimate across IDs after fixing the realized map. An ID-preserving state can support a subsequently fitted lookup decoder; that decoder and its assignment-specific training information belong in S . We therefore establish lack of prior exposure to the randomized associations, not the inability of an arbitrary decoder to express them.

Our experiments isolate information introduced by donor activations and by answer-guided optimization. First, target associations are sampled after the recipient and its diagnostic components are fixed. Second, the complete task-dependent activation tensor is replaced by one common state, while the optimizer retains its explicitly target-informed objective. These controls preserve a usable downstream system and expose the source of its induced answer.

Appendix D. Code, data, and disclosures

Acknowledgements

Research conducted at the AI Incident Response Sprint, September 2026. All original experiments, analyses, and report preparation for this project were conducted during the sprint. The published methods, model assets, and prior findings on which we build are credited above.

Code and data

Appendix F provides the audit protocol. The public companion contains standard-library code, saved predictions and margins, and pinned references to the tested implementations. It reproduces the main tables from saved outputs; it does not rerun model training or inference. The full experiment archive retains the protocols, scripts, target maps, and source/result audits separately. Model assets are identified by revision

receipts. The image-model calibration and fixed SSD application use one A100 GPU. GPT-2 runs use a local RTX 4070; the donor and reserved-membership experiments run locally on CPU. No provider-model inference, private incident data, or malicious prompt set is required by these experiments.

LLM usage

Codex (OpenAI) assisted with code execution and literature search. The author defined the research question and design, interpreted the results, wrote the paper, and verified the reported claims.

Appendix E. Limitations and Dual-Use Considerations

Section 7 states the main scientific limitations, and Appendices A and B retain the failed controls and calibration limits. The experiments do not estimate how often recovery tests lead to mistaken retention conclusions in deployed systems.

The intended benefit is more reliable evaluation of model remediation. The controls help investigators distinguish information retained by a model from information supplied during testing, and detect new errors introduced by an intervention. This can reduce both false claims that remediation failed and false confidence that an intervention restored useful behavior without other costs.

The intervention methods also have potential misuse. Activation transfer or answer-guided optimization could be adapted to elicit outputs that a system was modified to suppress. Our benign experiments do not establish that these methods recover deleted sensitive information or bypass safeguards in deployed models. They show why recovering an output and identifying its prior source must be evaluated separately. Information supplied from outside the recipient can still create a harmful capability when that input is available during deployment; identifying its source does not establish that a system is safe to redeploy.

A second risk is manipulation of evaluation evidence: supplying answer information during a test, choosing thresholds after seeing outcomes, or reporting corrections while omitting new errors could make remediation appear more or less effective than it is. Preserving the information-input inventory, comparison models, full error counts, and failed controls makes those choices easier to detect.

The minimal companion is limited to benign saved-result reproduction and source references. It excludes model weights, private incident data, credentials, operational attack instructions, and target-specific bypass material. Applications involving restricted information or affected systems should use authorized access, controlled evaluation data, and disclosure procedures appropriate to the incident.

Appendix F. Audit checklist and one-month continuation

Audit checklist for incident investigators

Use this checklist when a recovered answer is offered as evidence that remediation left information in a model. Record the tested model, target associations, intervention, comparison models, and decision criterion before interpreting the result.

1. Record unassisted behavior. Save the recipient's outputs before the intervention, including continuous scores, target errors, and performance on the retain task.

2. Account for information supplied by the test. List donor activations, target labels, optimization objectives, feature-selection labels, prompts, caches, and fitted decoders. Record when each component acquired access to the target associations.

3. Compare known learning histories. Apply the same intervention to models that learned the associations and models frozen before those associations were generated. Where compatible donors are available,

change their associations while holding the recipient fixed. Report whether the answer follows the recipient's history or the donor.

4. Check what reconstruction changes. Run the intervention with identical donor and recipient states and at zero strength. Compare logits and predictions with the unassisted model; zero strength may still reconstruct activations.

5. Report the complete error tradeoff. Count corrected errors, new errors, false positives, and retain-task performance. Compare continuous ranking with the deployed decision rule. Record the information supplied to any fitted calibration baseline.

6. Validate the retention criterion. If recovery scores distinguish known histories, select the decision threshold on calibration models and evaluate it on separate models. If a never-exposed model also produces the answer, report induction under the tested intervention; that success alone does not identify prior retention. If a suitable learning-history control is unavailable, record that the retention claim remains unresolved by this test.

One-month continuation

This Open Track project targets investigators evaluating remediation after an AI incident. The next month would test whether the controlled-history findings support useful decisions on natural-language unlearning tasks.

During the first week, reconstruct a documented unlearning pipeline with fixed method settings, target facts, and held-out utility checks. During the second, compare recovery in models with matched learning histories and train recovery maps only on independent facts. During the third, select decision thresholds on calibration models and evaluate them on separate models and facts. During the fourth, package the resulting protocol with per-case outcomes and compare the retention conclusions against the known histories. Obtain any missing original training recipe before claiming exact replication of published unlearning results.