# UNLEARNING CAN REMEMBER THE TEST AND FORGET THE REST

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Unlearning should remove specific facts from a model and leave everything else intact. Evaluations check the second part with a retain set: facts the model should still know afterwards. But many unlearning methods train on that same retain set. A method could therefore damage much of what the model knows, restore exactly the facts that will be tested, and look nearly as good as retraining from scratch. We show that this happens. We train language models of 0.6B to 1.24B parameters from two families on synthetic biographies, so we can also test people the model knows but no method targets or trains on. A method that resets parameters and then fine-tunes on the retain set can remove the targets while keeping retain accuracy high, yet in all 17 configurations of our main setting these untouched people lose 44 to 104% as much accuracy as the targets; where we retrained a matched model without the targets, the untouched people lose only 6 to 15% as much. The loss appears in both model families and equally for untouched facts whose questions were never trained. Prompting with their own biographies does not bring most of their facts back. The loss persists even when the untouched people are learned exactly like the targets, and only people in the retain set are spared. On the TOFU benchmark, a standard method, gradient difference with a strong retain weight, shows the same pattern: retain questions it trains on stay at the retrained level, while untouched questions of the same kind fall 20 points below it. A retain score therefore cannot show that unlearning was selective, and evaluations should also test known facts that no method trains on.

## 1 INTRODUCTION

Language models memorize private data, copyrighted text, and hazardous knowledge from their training corpora (Carlini et al., 2021; Eldan & Russinovich, 2023; Li et al., 2024), and their owners may be asked or required to remove it. Retraining without that data removes it exactly but is usually too expensive (Cao & Yang, 2015; Bourtoule et al., 2021), so approximate unlearning edits the trained model instead (Liu et al., 2025). Many methods pair a loss that suppresses the forget set with a loss that preserves a retain set; gradient difference, NPO with a retain term, and RMU all fine-tune on retained data (Maini et al., 2024; Zhang et al., 2024; Li et al., 2024). Benchmarks judge the result on the same two sets: the forget set should become unanswerable and the retain set should stay answerable, as close to retraining as possible (Maini et al., 2024; Shi et al., 2025b).

**These two scores cannot tell selective removal from broad removal followed by repair.** Low forget accuracy with high retain accuracy is read as evidence that the targets, and only the targets, were removed. Yet a method that erased everything the model knows and then fine-tuned it on the retain set would earn the same two scores. A retain score measured on the people a method trains on confirms only that the training worked. What the two scores leave out are the facts the model knows that are neither targeted nor repaired.

We measure what the two scores leave out, in a controlled setting. Following Allen-Zhu & Li (2024), we train models on synthetic biographies, so we know which people each model can answer questions about and how it learned them. Selectivity can be measured only where both are known, and benchmarks built on real text control neither (Section 6). We teach the facts in three settings, Inserted, Mixed, and Two-stage (Table 1), and in a crossover experiment. In all but Inserted the model



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.

can answer people who are in neither the forget set nor the retain set, and we call these people an untargeted holdout.

**Reset-and-repair removes the targets, and much of the holdout with them.** A single method that passes both scores while removing untargeted facts shows that the scores cannot establish selectivity. Reset-and-repair is built from two standard parts: a Fisher-ranked parameter mask, using the per-parameter weights of PerTA (Cai et al., 2026), and fine-tuning on the retain set, as gradient difference does. On a model with no answerable holdout it reaches 93% of retraining's forget gap. Where a holdout exists, reset-and-repair costs it 72.1 points on the model in Figure 1 and retraining without the targets 15.0. Retain accuracy recovers because repair trains on the retain questions.

Our contributions are:

1. A test for selectivity: an untargeted holdout, scored against matched retraining, separates selective removal from broad removal followed by repair on the retain set (Sections 2 and 4).
2. Controlled evidence (four designs, two families, 0.6B to 1.24B parameters): under reset-and-repair the holdout loses 44–104% as much as the targets in all 17 Mixed configurations, 12 passing both standard checks, against 6–15% under matched retraining where we ran it; on one model, held-out answers whose questions were and were not trained lose the same; the loss reaches biography completion and persists for a holdout learned like the targets (Sections 4 and 5).
3. What protects a group: membership in the retain set (17.2 points lost outside it on one model, at most 5.2 inside across eight Protocol U models), not a 487-fold concentration of the targets in parameters or a Fisher mask 13 times better at removing targets than random (Sections 4 and 5).
4. Gradient methods do not remove the targets selectively: in 247 evaluations of seven methods on Two-stage, retained and held-out people fall with the targets; tuned on Inserted, none comes near the success threshold; and on Llama-2-7B with the real MUSE-News corpus, none lifts retain knowledge above forget knowledge (Section 3, Appendix D.7).
5. A standard method on a standard benchmark: on TOFU, gradient difference keeps the retain half it trains on at the retrained level while the untouched half falls 19.7 points below it, in three seeds and with the halves swapped (Section 6).
6. A comparison of what current benchmarks score, and a four-step protocol for adding such a holdout (Section 6).

## 2 Setup

Each synthetic person has six facts, and people are split into forget, retain, and held-out sets. The attributes are birth date, birth city, university, major, employer, and work city (Allen-Zhu & Li, 2024). The model sees them in short biographies and answers one question per attribute. Splitting people rather than questions means a paraphrase can never move a forget person into the retain set.

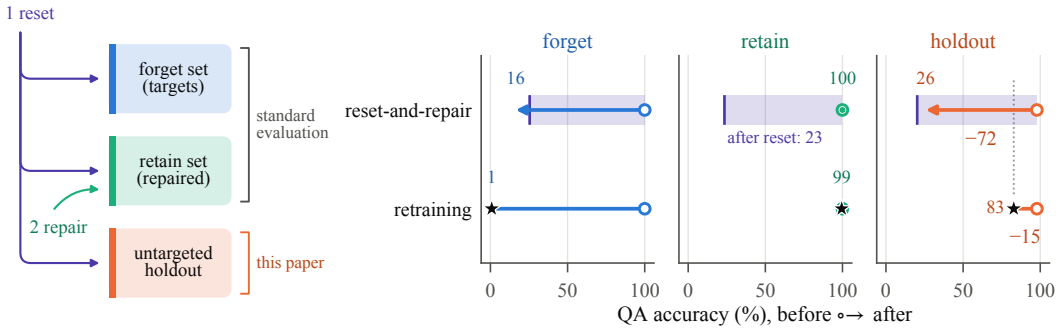



**Figure 1. High forget and retain scores hide damage that an untargeted holdout reveals.** *Left:* reset touches every group, repair fully restores only the retain set, and standard evaluation scores only forget and retain. *Right:* accuracy before (○) and after (arrowhead; ⋆: retrained) on one Mixed model (Qwen3, 500 people, 1,000 steps; reset-and-repair: four seeds; retraining: one run, Appendix E.1); purple band: accuracy right after the first reset (two seeds), when the retain set is as low as the targets.

**Table 1. Three ways the model learns the facts; only Mixed and Two-stage can answer people outside the forget and retain sets.** Every setting trains the forget and retain people's biographies and questions; *Mixed also trains all six questions of 83 held-out people and two of one more (of 100, or of 200 at 1,000 people). Mixed also runs with 80 / 720 / 200 people and on Llama-3.2-1B; Two-stage trains questions with LoRA after biographies.

| | Inserted | Mixed | Two-stage |
|---|---|---|---|
| Held-out biographies trained | ✗ | ✓ | ✓ |
| Held-out questions trained | ✗ | some* | ✗ |
| Accuracy before: forget / retain | 97.9 / 96.7 | ≥99 / ≥98.6 | 99.9 / 100 |
| held-out | 0.5 | 84–100 | 99.0 |
| People (forget / retain / held-out) | 40 / 360 / 100 | 40 / 360 / 100 | 1K / 49K / 50K |
| Initialization | Qwen3-0.6B-Base | Qwen3-0.6B-Base | random |

Three training settings teach these facts in different ways, and only two leave held-out people answerable (Table 1). In the *Inserted* setting, pretrained Qwen3-0.6B-Base is trained on the biographies and questions of forget and retain people only; held-out people never appear, and the model cannot answer them. The *Mixed* setting trains the biographies of all people together with the questions of forget and retain people and of some held-out people (Table 1). Training biographies and questions together lets the model answer questions about people whose questions it never saw (Allen-Zhu & Li, 2024), so it answers the other held-out people too, and all of them at 84–100% accuracy. The *Two-stage* setting trains a Qwen3-0.6B architecture from random initialization, first on biographies of about 100K people and then on questions for half of them, following the pretrain-then-fine-tune protocol of Allen-Zhu & Li (2024). Section 5 describes a fourth design, the crossover.

**An untargeted holdout is a group of people the model can answer who are in neither the forget set nor the retain set.** Mixed and Two-stage have one and Inserted does not. In Mixed, some held-out people are learned like the targets, from biographies and questions, and the rest from biographies alone; in the crossover's early arm, holdout and targets are learned the same way. Protocol U is Mixed with every held-out person's questions trained (eight models). On all eight we run reset-and-repair with the held-out people in the retain set; on one (500 people, 2,500 steps) we also restrict the retain set to the 360 original retain people.

**A method counts as selective only if it removes the targets, keeps the retain set, and costs the holdout no more than retraining does.** We report QA accuracy on each group. The forget gap is retain minus forget accuracy. The selectivity margin is the forget drop minus the holdout drop, with each drop measured from the model before unlearning. The margin measures how much a method favors the targets; whether it is selective is decided against retraining. Matched retraining is the same pipeline trained without the forget people's biographies and questions. Even retraining costs the holdout some accuracy (5.8 to 15.0 points in four Mixed retrainings), because the removed people no longer help train the question-answering skill. The retain set supplies both the Fisher estimate and the repair data. A method succeeds if the forget gap is at least 30 points, retain accuracy stays at or above 95, and, where an untargeted holdout exists, the holdout loses at most 1.5 points more than under retraining from the same data order.

We compare reset-and-repair, which resets the parameters the targets depend on most and then repairs on the retain set, with seven gradient-based methods: gradient ascent (GA), gradient difference (GD), NPO, SimNPO, DPO, RMU, and task arithmetic (Section 3). Reset-and-repair scores each parameter by the share of its diagonal Fisher information that comes from the forget set,

$$W_i = \frac{g_{F,i}^2 + \epsilon}{g_{F,i}^2 + g_{R,i}^2 + 2\epsilon},$$

where $g_{F,i}^2$ and $g_{R,i}^2$ are squared gradients on forget and retain questions. These are the weights of PerTA (Cai et al., 2026), which uses them to scale a forget-only task vector. We instead reset every parameter with $W_i$ above a threshold to its value before the facts were learned, then fine-tune on the retain set for 300 steps. We repeat this for four rounds (three on one Mixed model), raising the threshold from 0.55 or 0.60 to 0.95 or 0.975 (Appendix B). The repair step trains only on the retain set, which is why the holdout matters.

**Table 2. Reset-and-repair keeps retain accuracy at 94 or more but not the holdout; retraining costs the holdout far less.** Accuracy after unlearning, forget gap (retain minus forget accuracy), and the drop on the untargeted holdout (–: no answerable holdout; n/m: not measured). ⋆ rows: retraining without the forget people. Reset-and-repair rows are means over runs ± population standard deviation; Mixed rows use 500 people (Appendix D); Protocol U is in Figure 4a.

| Setting | Method | Forget ↓ | Retain ↑ | Gap ↑ | Holdout drop ↓ |
|---|---|---|---|---|---|
| Two-stage | best of 247 gradient evals. | 29.0 | 30.7 | 1.7 | n/m |
| | ⋆ *retraining (smaller model)* | 0.6 | 100.0 | 99.4 | −1.5 |
| Inserted | RMU (best gradient) | 81.2 | 96.8 | $15.5_{\pm 0.9}$ | – |
| | NPO, SimNPO, GA, GD | 81.2–97.9 | 82.6–96.6 | −1.8 to 1.4 | – |
| | reset-and-repair | 5.3 | 96.5 | $91.1_{\pm 3.0}$ | – |
| | ⋆ *retraining* | 0.8 | 98.5 | 97.7 | – |
| Mixed, 2,500 steps | reset-and-repair | $46.2_{\pm 11.2}$ | 94.0 | 47.8 | $39.1_{\pm 9.9}$ |
| | ⋆ *retraining (matched)* | 1.2 | 100.0 | 98.8 | 5.8 |
| Mixed, 1,000 steps | reset-and-repair | $16.4_{\pm 3.2}$ | 100.0 | 83.6 | $72.1_{\pm 2.8}$ |
| | ⋆ *retraining (matched)* | 0.8 | 99.5 | 98.7 | 15.0 |

## 3 Gradient Methods Lower Targets and Retained Facts Together

**On the Two-stage model, no gradient configuration lowers the targets without lowering retained and held-out facts just as much.** We ran 47 configurations of GA, GD, NPO (Zhang et al., 2024), SimNPO (Fan et al., 2025), DPO (Rafailov et al., 2023), RMU (Li et al., 2024), and task arithmetic (Ilharco et al., 2023) (Appendix C), and evaluated most at several checkpoints, 247 evaluations in all. All retain drops lie within 3.1 points of the diagonal of Figure 2a, and held-out people fall with them: in all 59 evaluations that cost the targets 30 points or more, the holdout loses 94 to 103% as much. In 176 of all 247 the targets stay at least 90% answerable, and in all 51 that bring forget accuracy below 50%, retain accuracy falls below 50% too. The median forget gap is $-1.00$; the largest, $+1.67$ points, comes from an evaluation that has already lost most of both groups (Table 2). Retraining a 1,000-person two-stage model without the forget people reaches $+99.38$ (Appendix B).

**Tuned on the Inserted model, no gradient method comes near the success threshold.** We tuned five methods, each over three learning rates and three values of its main hyperparameter, and re-ran each method's best configuration with three seeds (one for GA). NPO, SimNPO, GA, and GD end with forget gaps within 2 points of zero. RMU, the best, reaches $+15.54$ with retain accuracy at 96.8, well short of the 30-point threshold (Table 2). On Llama-2-7B with the real MUSE-News corpus, none of the 31 GA, GD, and NPO configurations that did not diverge closes the 16.4-point gap by which retain knowledge starts below forget knowledge; the best reaches $-12.88$ (Appendix D.7).

On a Two-stage model, one shared skill answers forget and retain questions alike. On a second, smaller two-stage model (51K people; Appendix C.3), a LoRA adapter trained only on retained people's questions answers 99.3% of forget questions correctly. The answers come from each person's biography, read out by a question-answering skill that everyone shares, and three further diagnostics agree (Appendix C.3); the gradient updates we tried lower forget and retain accuracy together, consistent with damage to this skill.

## 4 Reset-and-Repair Removes What Repair Does Not Restore

**Unlike the gradient methods, reset-and-repair removes the targets, and on the Inserted model it nearly matches retraining.** Averaged over six runs covering four forget splits, its forget gap is $+91.1$ (range 87.7 to 95.2), 93% of the $+97.69$ that retraining reaches (96% on the split the retraining baseline uses), with retain accuracy at 96.5. Both the Fisher ranking and the repair contribute on Qwen3. In one round, resetting the parameters with the highest Fisher share opens a 6.1 times larger forget gap than resetting the same number of random parameters (40 random seeds); on Llama-3.2-1B a random mask does almost as well ($1.06\times$; Appendix D.1.1). Our reimplementation of PerTA's update, which has no repair step, reaches only $+9.3$, and retain accuracy falls to 9.7.

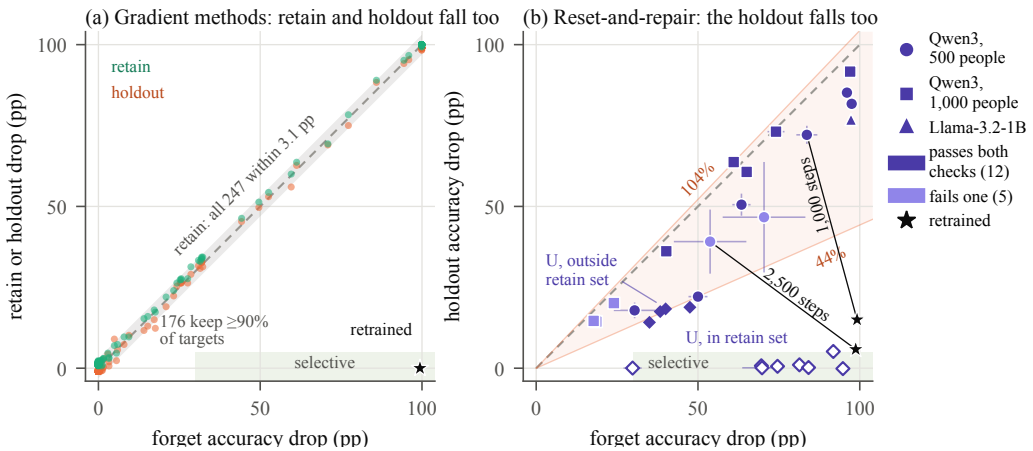



**Figure 2. Gradient methods lower retained people as much as the targets (a); reset-and-repair spares retained people but removes 44–104% as much holdout (b).** $x$: what the targets lose; $y$: what a kept group loses; diagonal: equal loss; shaded: forget drop $\geq$30, kept-group drop $<$5 (a guide, not the criterion). *(a)* Gradient methods on Two-stage, retain (green) and holdout (orange) drop; band: $\pm$3.1 points around the diagonal for retain; $\star$: retraining a smaller two-stage model. *(b)* Reset-and-repair on the 17 Mixed configurations (means $\pm$ sd over four seeds; dark: the 12 passing both standard checks) and on Protocol U (U; filled: holdout outside the retain set, one per seed; hollow: inside, one per model); wedge: 44–104%; $\star$: matched retraining, joined to its model (another data order: Appendix E.1).

Sampling, paraphrasing, a small relearning attack, and quantization recover no more forget answers than after retraining (Appendix D.1.3). Inserted cannot test selectivity, because no one outside the forget and retain sets is answerable (held-out accuracy 0.5 before unlearning) and repair trains on the questions that measure retain accuracy. Two checks still separate reset-and-repair from retraining. A loss-based membership test still tells forget people from unseen ones after reset-and-repair (AUC 0.848, against 0.435 after retraining). Relearning brings the targets back much faster: fine-tuning on all forget questions for 50 epochs lowers the forget gap to $+14.3$ after reset-and-repair but only to $+83.6$ after retraining (Appendix D.1.3).

**Where a holdout exists, reset-and-repair removes much of it along with the targets.** A Mixed model answers held-out people at 84–100% before unlearning, and these people are in neither the forget set nor the retain set. Figure 2b shows what the targets and this holdout lose in 17 Mixed configurations, each averaged over four seeds: Qwen3 with 500 or 1,000 people at several training lengths and forget splits, and Llama-3.2-1B. Twelve of the 17 configurations pass both standard checks, a forget gap of at least 30 and retain accuracy of at least 95, and each of them still loses 17.9 to 91.7 points of holdout. Across all 17, the holdout loses 44–104% as much as the targets, and at least 14.4 points. Three more Llama-3.2-1B training lengths, one seed each, also pass both checks and lose 84–95% as much (Appendix D.2). Whether a held-out question was trained makes no difference: on the model of Figure 1, the 500 held-out questions that were trained lose 70.2 points and the 100 that were not lose 70.1 (Appendix E.2). With 1,000 people the selectivity margin is between $-2.7$ and $+5.4$ points; with 500 people it favors the targets more ($+11$ to $+28$) but still removes 18 to 85 points of holdout (Appendix D.2).

**Matched retraining costs the holdout far less, and keeps every held-out fact the biography probe measures.** We retrained the 500-person Mixed model without the forget people's biographies and questions. It cannot answer the forget people (forget accuracy 1.2), and its holdout loses only 5.8 points of QA accuracy (8.7 with another data order), the small cost of training without them described in Section 2. It still completes 100% of held-out people's biographies (Appendix E.4). Reset-and-repair on the same model costs the holdout 39.1 points (Table 2), and 72.1 at 1,000 steps, where retraining costs 7 to 15 depending on the data order (Appendix E.1); paired by data order, it exceeds retraining's holdout drop by 33.3 and 57.1 points, 22 to 38 times the 1.5-point criterion. A second model trained with the same 2,500-step configuration, which has no matched retraining, loses 22.1 under reset-and-repair.




**Figure 3. Repair fully restores only what it trains on, and the holdout ends far below retraining.** QA accuracy through four rounds of reset (shaded) and repair on the Mixed model with 500 people and 1,000 training steps; thick lines: mean of two seeds, thin: each seed. Right: the matched retrained model ($\star$); the bracket spans the holdout's gap to it.

Each reset lowers all three groups and each repair fully restores only the retain set (Figure 3). Repair brings retained people back to 99.5 or more after every reset, while forget and held-out accuracy recover only partly and end at 15 and 27 to 31 in the two plotted seeds.

**Only membership in the retain set, which supplies both the Fisher estimate and the repair data, keeps a group's loss small (Figure 4a).** On the Protocol U model, which trains every held-out person's questions, reset-and-repair costs them 17.2 points when they are outside the retain set and nothing when they are inside it (at most 5.2 across all eight Protocol U models, including four whose forget drop is 81 to 95). The inside run removes less of the targets (29.8 against 40.2), but the eight models show protection at every level of removal.

The damage is not confined to question answering (Figure 4b). Prompted with the start of each held-out person's training biography, the model after reset-and-repair completes 25% of their attributes, down from 100% and close to its 28% QA accuracy in the same four reruns; it completes 86% for retained people, whose QA accuracy repair restored to 100. A retrained model completes 100% of held-out biographies, so retraining weakens only how the holdout's facts are read out.

**Among Qwen3 Mixed models, Fisher-tail mass, the fraction of parameters with $W_i > 0.55$ before unlearning, predicts how much reset-and-repair removes from targets and holdout alike (Figure 4c).** Across eight such models, its Spearman $\rho$ is 0.83 with the forget drop, 0.93 with the holdout drop, and 0.07 with the selectivity margin (Appendix D.3). A fit frozen before two further Mixed models were trained predicted both forget gaps, and both models lost most of their holdout. When each mask resets a fixed 20% of every weight matrix, the Fisher-ranked mask removes 13 times as much of the targets as a random mask and is more selective, yet the holdout still loses 65% as much as the targets (Appendix E.5).

## 5 A Holdout Learned Like the Targets

The crossover supplies a holdout learned exactly like the targets. In the Mixed models, how and when each held-out person is learned is not controlled; the crossover's early arm learns holdout and targets from biographies alone, in the same phase. Two models share a 50K-person population (Figure 5a); each learns one focal set of 400 people early, during continued pretraining, and another late, in a short final phase, and the second model swaps the sets. No focal person's questions are ever trained. A question-answering adapter trained on 2,000 other people reads out every fact. A holdout of 200 people is learned during continued pretraining, like the early focal set, and is in neither the forget set nor the retain set. Reset-and-repair resets each model to its own starting weights and repairs on 2,000 retained people. We fixed the decision rule, which compares the two arms, before tracing any trade-off curve; the holdout comparison is a secondary analysis. The test resets a fixed fraction of each weight matrix because the four-round schedule of Section 2 collapses these models, and there is no retraining reference (Appendix F).

Early facts are favored by 487 times as many parameters as late ones. For early facts 51.17% of attention and MLP parameters have Fisher share $s > 0.5$ (the crossover's version of $W_i$; Appendix F), against 0.105% for late facts (Figure 5b), a 487-fold difference with no overlap across all 20 combinations of model, focal set, and seed. A threshold reset would therefore reset far more

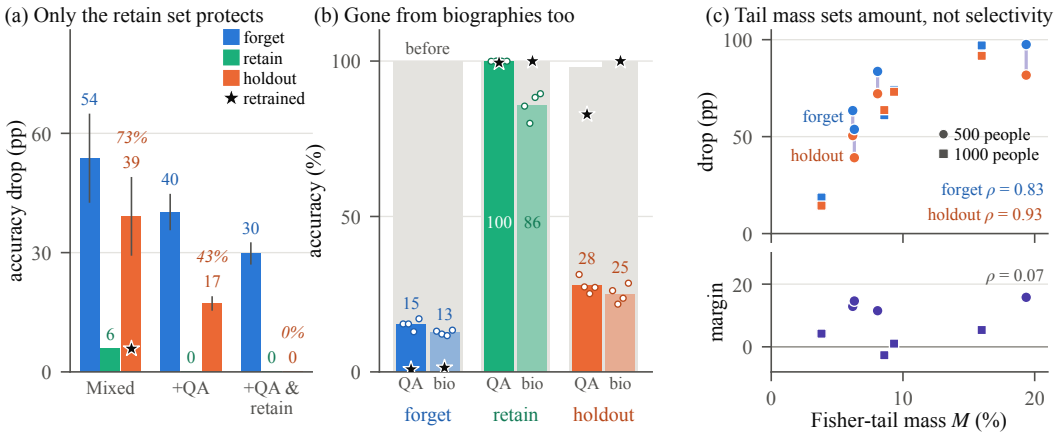



**Figure 4. Only retain-set membership keeps the holdout's loss small, the removal reaches the biographies, and Fisher-tail mass predicts how much is removed.** *(a)* Forget, retain and holdout drop on 2,500-step models, with every held-out question trained (+QA; Mixed trains 83 of 100 people's) and the holdout also in the retain set (+QA & retain); error bars: sd over four seeds; percentages: holdout over forget drop; ⋆: matched retraining. *(b)* QA (solid) and biography completion (light) after reset-and-repair (four reruns, dots; gray: before) and after retraining (⋆), Mixed, 1,000 steps. *(c)* Eight Mixed models against Fisher-tail mass $M$ (circles: 500 people; squares: 1,000): drops (top) and selectivity margin (bottom).

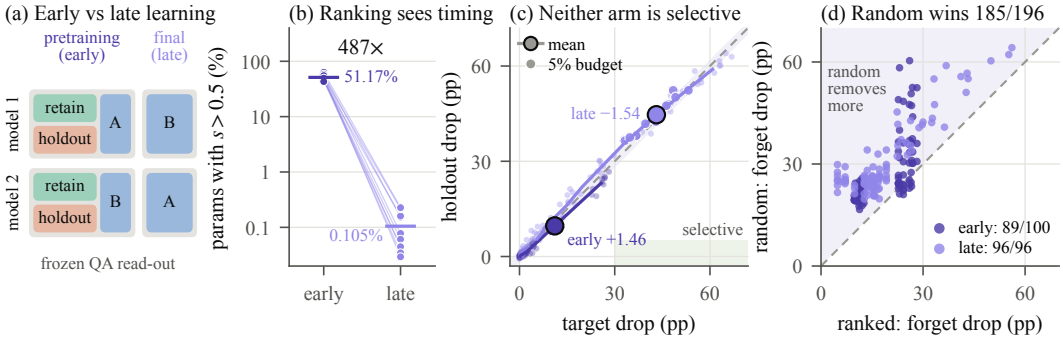



**Figure 5. When a fact is learned changes which parameters favor it in Fisher information, but not how selectively reset-and-repair removes it.** *(a)* Each focal set (A, B) is learned early in one model and late in the other; retain and holdout people are learned early. *(b)* Share of parameters with Fisher share $s > 0.5$; lines pair a focal set across models; bars: each arm's mean. *(c)* Target against holdout drop for every model and budget; faint dots: other budgets; dots: each model at the pre-registered 5% budget; large circles: each arm's mean there, labelled with the selectivity margin; curves: each arm's mean across budgets; faint band: the 2-point effect threshold. *(d)* Forget drop of random 5% masks against the Fisher-ranked mask at the same collateral damage; shaded: random removes more.

parameters in the early arm (4.63% against 0.098%), so the test resets the same fraction of every weight matrix in both arms.

**Neither arm is removed selectively (Figure 5c).** At the 5% budget of the pre-registered grid, the early arm loses 11.1 points of target accuracy and 9.7 of holdout accuracy, and the late arm loses 43.1 and 44.7: selectivity margins of $+1.46$ and $-1.54$, both inside the pre-registered 2-point effect threshold; without a retraining reference, the margin is the only measure here. The pre-registered rule asked whether the early arm is removed more selectively than the late arm. It fails: the arms' curves of target loss against collateral loss cross in 7 of the 10 comparisons over five training seeds, four of them added after the seed-0 freeze; the one criterion that passes does so only because the late arm loses more of everything (Appendix F.2). Under the one-shot threshold reset that defines the 487-fold premise, the early arm's holdout loses 88% as much as its target; resetting instead to a model that already knows the holdout spares it, but not the untargeted focal set (Appendix F.3).

**Table 3. No existing benchmark scores a set that is untouched, learned like the targets, and paired with retraining.** Scored sets beyond the forget set. Untouched: neither trained on during unlearning nor used to select methods or checkpoints. Learned like targets: same source and training stage as the forget set. Retrained: scores reported for a model retrained without the forget set. Every listed set is answerable before unlearning. ✓ holds, ○ partly, ✗ does not.

| Benchmark | Scored set | Untouched | Learned like targets | Retrained |
|---|---|---|---|---|
| TOFU (Maini et al., 2024) | retain questions | ✗ | ✓ | ✓ |
| | real authors, world facts | ✓ | ✗ | ✓ |
| DUSK (Jeung et al., 2026) | retain knowledge | ✗ | ✓ | ✓ |
| MUSE-News (Shi et al., 2025b) | held-out retain split | ○[a] | ✓ | ✓ |
| MUSE-Books (Shi et al., 2025b) | held-out retain split | ○[a] | ○[b] | ✓ |
| RWKU (Jin et al., 2024) | neighbor knowledge | ✓ | ○[c] | ✗ |
| WMDP (Li et al., 2024) | MMLU subsets | ○[e] | ✗ | ✗ |
| SUITE (Peleg et al., 2026) | held-out retain qs. | ✓ | ○[c] | ✗ |
| Ours (Mixed) | untargeted holdout | ✓ | ○[d] | ✓ |
| Ours (crossover, early) | untargeted holdout | ✓ | ✓ | ✗ |

[a] kept out of training but used as the early-stopping criterion. [b] same training stage, different source (fan wiki against the books). [c] pretraining knowledge whose acquisition is not controlled. [d] biographies trained like the targets', and questions for some held-out people. [e] used to select the unlearning method's hyperparameters.

Random masks remove more of the targets than the Fisher-ranked mask at equal collateral damage in 185 of the 196 comparable draws out of 200, in both arms and every seed (Figure 5d). On Mixed, where a trained question-answering skill reads the facts out, the Fisher-ranked mask wins instead (Section 4).

## 6 What This Means for Evaluating Unlearning

**Evaluations should report an untargeted-holdout drop, the accuracy lost on facts the model knows that are neither targeted nor used for repair.** A useful holdout is answerable before unlearning, unlike Inserted's held-out people; untouched by the method, since retain-set membership hides the loss (17.2 points outside, none inside); learned like the targets, so it depends on what they depend on; and scored against retraining, which itself costs the holdout 5.8 to 15.0 points. Adding one takes four steps: build it from the forget set's source and training stage, for example by splitting the retain data in two as MUSE and our TOFU experiment do; keep it out of training, repair, any reset reference, hyperparameter search, and checkpoint selection; report its drop next to the forget gap and retain accuracy, for each method and for retraining; and count a method as selective only if its holdout drop is within a fixed margin of retraining's from the same data order (we use 1.5 points).

No benchmark scores a set with all four properties; MUSE-News comes closest (Table 3), but uses its held-out retain split to decide when to stop unlearning (Shi et al., 2025b), so methods are still selected against it. TOFU's scored retain questions are all contained in the retain sets that gradient difference and KL regularization train on (Maini et al., 2024), and TOFU's Model Utility, also reported by OpenUnlearning (Dorna et al., 2025), combines them with real-author and world-fact questions. Other untouched sets, such as RWKU's neighbors (Jin et al., 2024), WMDP's MMLU subsets (Li et al., 2024) and SUITE's held-out questions (Peleg et al., 2026), are pretraining knowledge learned unlike the targets, mostly without a retraining reference.

**On TOFU, a standard method shows the same failure.** We split TOFU's retain authors in half, let gradient difference regularize on one half, and scored ROUGE-L recall on both with OpenUnlearning's Llama-3.2-1B-Instruct model (Figure 6, Appendix E.6). With a retain weight of 5, the trained-on half ends at 80.8 over three seeds, level with retraining's 81.4, while the untouched half ends 19.7 points below retraining; swapping the halves gives the same result. The untouched half loses 60% as much as the targets, inside the 44–104% range of our Mixed models. The gap needs a strong retain weight and appears for gradient difference: forget recall stays above retraining's 38.4, weaker retain weights lower both halves (half B still 15 to 20 points below half A), and at the same weight NPO lowers both halves together while SimNPO barely forgets. TOFU's retain score uses

only questions of the first kind; its real-author questions, which no method trains on, fall 12.0 points, and its world facts do not move.

## 7 Related Work

Prior work has found both incomplete forgetting and collateral damage; we identify how scored retention can hide collateral damage and measure it against matched retraining. Critiques of forgetting depth find that suppressed answers return or remain recoverable (Hu et al., 2025a; Zhang et al., 2025c; Lynch et al., 2024; Deeb & Roger, 2024; Xu et al., 2026b). Benchmarks and membership-inference evaluations can mislead (Thaker et al., 2025; Feng et al., 2025; Hayes et al., 2025), for example when forget and retain queries are combined (Thaker et al., 2025); a retain score misses broad removal even when they are independent. Held-out retain probes reveal over-forgetting that benchmarks miss (Peleg et al., 2026), and unlearned models fail up to 98.7% of generated benign questions the original model answered (Ko et al., 2025); we show that repair on the retain set can hide such loss. With matched retraining in a controlled testbed, Yang & Yeung (2026) find that trained probes miss under-forgetting of held-out facts; we find the complementary failure on the retain side.

Localization has repeatedly failed to predict where editing or unlearning works, and a better mask did not avoid collateral damage here. Causal-tracing locations do not predict where editing works (Hase et al., 2023), and for unlearning, random parameters do about as well as localized ones (Lee et al., 2025); localization helps when facts are written into predefined parameters (Boglioni et al., 2026). Saliency and Fisher-based methods choose weights to update, reset, or rescale (Fan et al., 2024; Jia et al., 2024; Liu et al., 2023; Foster et al., 2024; Shi et al., 2025a; Cai et al., 2026); in image classifiers a saliency mask forgets no better at the representation level than an equal-size random one (Habbati et al., 2026).

Unlearning damages retained knowledge that shares something with the targets, here the question-answering skill. Removing forget data damages retain data that shares its knowledge (Jeung et al., 2026; Hu et al., 2025b; Cheng et al., 2026b; Mavrothalassitis et al., 2025); damage concentrates on syntactically similar queries (Chang & Lee, 2025) or nearby knowledge (Rinberg et al., 2025; Su et al., 2026), grows with representational entanglement (Wybitul et al., 2026), and in image classifiers grows near the forget set even with retain-set fine-tuning (Dolgova & Stich, 2026); correlated knowledge can also let forgotten facts survive (Wei et al., 2025). Regularizing on similar retain queries preserves them (Chang & Lee, 2025), as repair preserves our retain set.

## 8 Limitations and Conclusion

One method whose retain score hides broad removal suffices; we show two (reset-and-repair, gradient difference), and the holdout test can check others. We measure behavior; weight-level erasure (Deeb & Roger, 2024; Łucki et al., 2025) and legal guarantees are outside our scope. The Mixed step grid and the crossover were built for other questions and are reanalyzed here. Every fixed reset budget we tried collapses Qwen3-1.7B, and Two-stage reset-and-repair used a mismatched reset target (Appendices D.6 and D.5). Our Qwen3 models use a non-standard sign for the rotary position encoding; the Llama models use the standard one and lose as much holdout, but the Fisher-tail ranking of Section 4 is untested under it (Appendix B.3).

**A retain score cannot show that unlearning was selective.** Forget and retain scores do not measure losses on facts a method neither targets nor trains on; scoring an untargeted holdout against matched retraining does, at the cost of one more evaluation set.

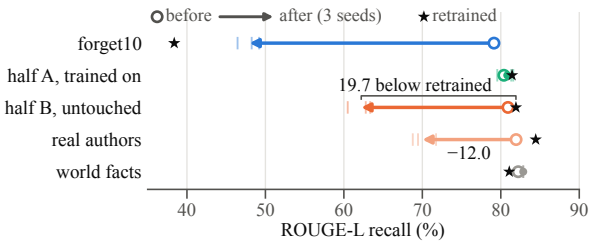



**Figure 6. On TOFU, the retain half that gradient difference trains on stays at the retrained level; the untouched half ends 19.7 points below it.** ROUGE-L recall before (○) and after (mean of three seeds; ticks: seeds), retain weight 5; ⋆: retrained without forget10. Bottom rows: TOFU's utility questions, on reruns.

### AI USE STATEMENT

In this work, we used generative AI tools (an AI coding assistant and chat models) to implement methods, to give feedback on the research methodology and experiment design, to support data analysis, to interpret results and refine hypotheses and the claims they support, and, in one appendix experiment, to rewrite the templated biographies as free prose (Appendix D.1.2). We have not used generative AI tools to generate any other data or to clean or reformat datasets: the synthetic people and question–answer pairs are generated deterministically by our own scripts following the attribute scheme of Allen-Zhu & Li (2024), and the benchmarks are public. Developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, qualitative and thematic data analysis, and translation are not applicable to this work. Additionally, we used generative AI tools to write and edit experiment and analysis code, run and monitor training and evaluation jobs, suggest experimental parameters, create figures, identify relevant literature and check citations, suggest the paper's structure and title, and draft and edit the manuscript. We have reviewed all AI-assisted work: every reported number was recomputed from the per-run evaluation results of its runs; for five artifacts that were not kept we report their recorded results (Appendix I), and the decision rules of the two pre-registered analyses were frozen before the tests they define were run, and later amendments are reported (Appendices F.2 and D.3). We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

Appendix B gives data generation, splits, training lengths, and every hyperparameter of reset-and-repair, and Appendix E describes the experiments run for this paper. Every reported number was recomputed from the per-run evaluation results of its runs, except that for five artifacts that were not kept we report their recorded results (Appendix I). For every reported configuration we report every run and every seed, and any run left out of a mean is named with the reason. Two analyses were pre-registered: the crossover's decision rule, frozen before any trade-off curve was traced (Appendix F.2), and the Fisher-tail fit, frozen before two further Mixed models were trained (Appendix D.3).

### ETHICS STATEMENT

All people in our controlled experiments are synthetic; small checks use public benchmarks of fictitious authors, real news articles, and Wikipedia biographies (Maini et al., 2024; Shi et al., 2025b; Ramakrishna et al., 2025), and no human subjects were involved. The main risk we see is the one the paper is about: a low forget score with a high retain score can be read as evidence that specific data was removed when it was not.

## REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1067–1077, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Ameya Anjarlekar and Sandeep Pombra. LLM unlearning using gradient ratio-based influence estimation and noise injection. *arXiv preprint arXiv:2508.06467*, 2025.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024.

Anna Bair, Yixuan Even Xu, Mingjie Sun, and J. Zico Kolter. Compressed sensing for capability localization in large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Matteo Boglioni, Thibault Rousset, Siva Reddy, Marius Mosbach, and Verna Dankers. LACUNA: A testbed for evaluating localization precision for LLM unlearning. *arXiv preprint arXiv:2607.02513*, 2026.

Anna Borisiuk, Andrey Savchenko, Alexander Panchenko, and Elena Tutubalina. Anatomy of unlearning: The dual impact of fact salience and model fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25846–25859. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1287.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019.

Chengyi Cai, Zesheng Ye, Jiangchao Yao, Jianzhong Qi, Bo Han, Xiaolu Zhang, Feng Liu, and Jun Zhou. Per-parameter task arithmetic for unlearning in large language models. *arXiv preprint arXiv:2601.22030*, 2026.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE, 2015. doi: 10.1109/SP.2015.35.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, 2021.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. Towards robust and parameter-efficient knowledge unlearning for LLMs. In *International Conference on Learning Representations (ICLR)*, 2025.

Hwan Chang and Hwanhee Lee. Which retain set matters for LLM unlearning? a case study on entity unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5966–5982, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.310. URL https://aclanthology.org/2025.findings-acl.310/.

Ting-Yun Chang, Jesse Thomason, and Robin Jia. Do localization methods actually localize memorized data in LLMs? a tale of two benchmarks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3190–3211. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.176. URL https://aclanthology.org/2024.naacl-long.176/.

Hang Chen, Jiaying Zhu, Hongyang Chen, Hongxu Liu, Xinyu Yang, and Wenya Wang. Navigating by old maps: The pitfalls of static mechanistic localization in LLM post-training. *arXiv preprint arXiv:2605.06076*, 2026a.

Hang Chen, Jiaying Zhu, Xinyu Yang, and Wenya Wang. CLUE: Conflict-guided localization for LLM unlearning framework. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Knowledge localization: Mission not accomplished? Enter query localization! In *The Thirteenth International Conference on Learning Representations*, 2025. Spotlight.

Jiali Cheng, Ziheng Chen, Chirag Agarwal, and Hadi Amiri. A mechanistic perspective and circuit-guided difficulty metric for unlearning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10950–10964. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.findings-acl.532.

Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. Machine unlearning under retain-forget entanglement. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026b. URL https://openreview.net/forum?id=4WMBSHHJEr.

Sungjun Cho, Dasol Hwang, Frederic Sala, Sangheum Hwang, Kyunghyun Cho, and Sungmin Cha. Reference-specific unlearning metrics can hide the truth: A reality check. *arXiv preprint arXiv:2510.12981*, 2025.

Alex Cloud, Jacob Goldman-Wetzler, Evžen Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient routing: Masking gradients to localize computation in neural networks. *arXiv preprint arXiv:2410.04332*, 2024.

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

Noam Diamant, Neta Glazer, and Ethan Fetaya. Stress testing unlearning algorithms. *arXiv preprint arXiv:2608.22527*, 2026.

Polina Dolgova and Sebastian U. Stich. Forgetting has neighbors: Localized collateral forgetting in machine unlearning. *arXiv preprint arXiv:2605.31317*, 2026.

Vineeth Dorna, Anmol Reddy Mekala, Wenlong Zhao, Andrew McCallum, J. Zico Kolter, Zachary C. Lipton, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Gy67Zh5X1i. Datasets and Benchmarks Track.

Ronen Eldan and Mark Russinovich. Who's Harry Potter? Approximate unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

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

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 *Advances in Neural Information Processing Systems*, 2025.

Eoin Farrell, Yeu-Tong Lau, and Arthur Conmy. Applying sparse autoencoders to unlearn knowledge in language models. *arXiv preprint arXiv:2410.19278*, 2024.

Lawrence Feng, Gaurav R. Ghosal, Jacob Mitchell Springer, Ziqian Zhong, and Aditi Raghunathan. Early data exposure improves robustness to subsequent fine-tuning. In *Conference on Language Modeling (COLM)*, 2026.

Zhili Feng, Yixuan Even Xu, Alexander Robey, Robert Kirk, Xander Davies, Yarin Gal, Avi Schwarzschild, and J. Zico Kolter. Existing large language model unlearning evaluations are inconclusive. *arXiv preprint arXiv:2506.00688*, 2025.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12043–12051, 2024. doi: 10.1609/aaai.v38i11.29092.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *The Seventh International Conference on Learning Representations (ICLR)*, 2019.

Gaurav R. Ghosal, Pratyush Maini, and Aditi Raghunathan. Natively unlearnable large language models. *arXiv preprint arXiv:2606.13873*, 2026.

Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20964–20992, 2025.

Billel Habbati, Alessio Merlo, Luca Verderame, and Meriem Guerar. Gradient concentration, not weight saliency, explains representation-level class unlearning. *arXiv preprint arXiv:2607.21353*, 2026. Preprint, not peer reviewed.

Xiaoqi Han, Víctor Gutiérrez-Basulto, Ru Li, Xiaoli Li, Jiye Liang, and Jeff Z. Pan. Consistency-aware editing for entity-level unlearning in language models. *arXiv preprint arXiv:2601.08840*, 2026.

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

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 497–519, 2025. doi: 10.1109/SaTML64287.2025.00034.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic test of unlearning using parametric knowledge traces. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19513–19535. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.985. URL https://aclanthology.org/2025.emnlp-main.985/.

Yihuai Hong, Yiran Zhao, Wei Tang, Yang Deng, Yu Rong, and Wenxuan Zhang. The rise of parameter specialization for knowledge storage in large language models. In *Advances in Neural Information Processing Systems*, 2025b.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *The Thirteenth International Conference on Learning Representations*, 2025a.

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*, 2025b.

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

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 (Volume 1: Long Papers)*, pp. 14389–14408. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Wonje Jeung, Sangyeon Yoon, Hyesoo Hong, Soeun Kim, Seungju Han, Youngjae Yu, and Albert No. DUSK: Do not unlearn shared knowledge. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42013–42031. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.2085.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient LLM unlearning framework from logit difference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hengrui Jia, Taoran Li, Jonas Guan, and Varun Chandrasekaran. The erasure illusion: Stress-testing the generalization of LLM forgetting evaluation. *arXiv preprint arXiv:2512.19025*, 2025.

Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. WAGLE: Strategic weight attribution for effective and modular unlearning in large language models. In *Advances in Neural Information Processing Systems*, 2024.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. RWKU: Benchmarking real-world knowledge unlearning for large language models. In *Advances in Neural Information Processing Systems*, 2024.

Andrej Karpathy. nanochat: The best chatgpt that $100 can buy, 2025. URL https://github.com/karpathy/nanochat.

Hyo Seo Kim, Dongyoon Han, and Junsuk Choe. NegMerge: Sign-consensual weight merging for machine unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30067–30085, 2025a.

Yejin Kim, Eunwon Kim, Buru Chang, and Junsuk Choe. Improving fisher information estimation and efficiency for LoRA-based LLM unlearning. In *Proceedings of the Second Conference on Language Modeling*, 2025b.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Myeongseob Ko, Hoang Anh Just, Charles Fleming, Ming Jin, and Ruoxi Jia. Probing hidden knowledge holes in unlearned LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Shingo Kodama, Niv Cohen, Micah Adler, and Nir Shavit. Understanding empirical unlearning with combinatorial interpretability. *arXiv preprint arXiv:2602.19215*, 2026.

Aravind Krishnan, Siva Reddy, and Marius Mosbach. Not all data are unlearned equally. In *Conference on Language Modeling (COLM)*, 2025.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4158–4169, 2019.

Jamie Lanyon, Axel Finke, Petros Andreou, and Georgina Cosma. On the importance of multiple training seeds for evaluating machine unlearning. *arXiv preprint arXiv:2510.26714*, 2025.

Hwiyeong Lee, Uiji Hwang, Hyelim Lim, and Taeuk Kim. Does localization inform unlearning? A rigorous examination of local parameter attribution for knowledge unlearning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21857–21869. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1109.

Jaeung Lee, Dohyun Kim, and Jaemin Jo. Measuring the depth of LLM unlearning via activation patching. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550, 2024.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. Large language model unlearning via embedding-corrupted prompts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7:181–194, 2025. doi: 10.1038/s42256-025-00985-0.

Yufang Liu, Changzhi Sun, Yuanbin Wu, and Aimin Zhou. Unlearning with Fisher masking. *arXiv preprint arXiv:2310.05331*, 2023.

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 (TMLR)*, 2025. NeurIPS 2024 SoLaR Workshop Best Technical Paper; arXiv:2409.18025.

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.

Weitao Ma, Xiaocheng Feng, Weihong Zhong, Lei Huang, Yangfan Ye, Xiachong Feng, and Bing Qin. Unveiling entity-level unlearning for large language models: A comprehensive analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5345–5363. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.358/.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *Proceedings of the First Conference on Language Modeling*, 2024. OpenReview: B41hNBoWLo.

Ioannis Mavrothalassitis, Pol Puigdemont, Noam Itzhak Levi, and Volkan Cevher. Ascent fails to forget. In *Advances in Neural Information Processing Systems*, 2025.

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

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

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, September 2024. Llama-3.2-1B-Base model card: https://huggingface.co/meta-llama/Llama-3.2-1B.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40034–40050, 2024.

Amit Peleg, Naman Deep Singh, Naama Pearl, Bibhabasu Mohapatra, and Matthias Hein. Forget narrowly, retain broadly: Unlearning as an asymmetric generalization problem. *arXiv preprint arXiv:2607.09236*, 2026.

Yuefeng Peng, Parnian Afshar, Megan Ganji, Thomas Butler, Amir Houmansadr, Mingxian Wang, and Dezhi Hong. Forget to know, remember to use: Context-aware unlearning for large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Nicholas Pochinkov and Nandi Schoots. Dissecting language models: Machine unlearning via selective pruning. *arXiv preprint arXiv:2403.01267*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Anil Ramakrishna, Yixin Wan, Xiaomeng Jin, Kai-Wei Chang, Zhiqi Bu, Bhanukiran Vinzamuri, Volkan Cevher, Mingyi Hong, and Rahul Gupta. LUME: LLM unlearning with multitask evaluations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6524–6535. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.347. URL https://aclanthology.org/2025.findings-emnlp.347/.

Hadi Reisizadeh, Jiajun Ruan, Yiwei Chen, Soumyadeep Pal, Sijia Liu, and Mingyi Hong. Leak@$k$: Unlearning does not make LLMs forget under probabilistic decoding. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Jie Ren, Zhenwei Dai, Xianfeng Tang, Hui Liu, Jingying Zeng, Zhen Li, Rahul Goutam, Suhang Wang, Yue Xing, Qi He, and Hui Liu. A general framework to enhance fine-tuning-based LLM unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18464–18476. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.949.

Roy Rinberg, Usha Bhalla, Igor Shilov, Flavio P. Calmon, and Rohit Gandikota. RippleBench: Capturing ripple effects using existing knowledge repositories. *arXiv preprint arXiv:2512.04144*, 2025.

Yan Scholten, Stephan Günnemann, and Leo Schwinn. A probabilistic perspective on unlearning and alignment for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. OpenReview: 51WraMid8K.

Aakriti Shah, Yifan Hu, and Thai Le. What the "spotless" mind remembers: How knowledge entanglement shapes what leaks after unlearning in LLMs. *arXiv preprint arXiv:2510.25732*, 2025a.

Raj Sanjay Shah, Jing Huang, Keerthiram Murugesan, Nathalie Baracaldo, and Diyi Yang. The unlearning mirage: A dynamic framework for evaluating LLM unlearning. In *Proceedings of the Second Conference on Language Modeling*, 2025b.

Jialei Shi, Kostis Gourgoulias, John F. Buford, Sean J. Moran, and Najah Ghalyan. DeepClean: Machine unlearning on the cheap by resetting privacy sensitive weights using the Fisher diagonal. In *Computer Vision – ECCV 2024 Workshops*, volume 15640 of *Lecture Notes in Computer Science*, pp. 1–16. Springer, 2025a. doi: 10.1007/978-3-031-91672-4_1.

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

Filip Sondej and Yushi Yang. Collapse of irrelevant representations (CIR) ensures robust and non-disruptive LLM unlearning. *arXiv preprint arXiv:2509.11816*, 2025.

Filip Sondej, Yushi Yang, and Adam Mahdi. RepSelect: Robust LLM unlearning via representation selectivity. *arXiv preprint arXiv:2606.17168*, 2026.

Bo Su, Ankit Shah, and Thai Le. PreUnlearn: Auditing collateral knowledge damage before large language model unlearning. *arXiv preprint arXiv:2606.18473*, 2026.

Guangzhi Sun, Potsawee Manakul, Xiao Zhan, and Mark Gales. Unlearning vs. obfuscation: Are we truly removing knowledge? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11457–11467, 2025.

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

Stefan Vasilev, Christian Herold, Baohao Liao, Seyyed Hadi Hashemi, Shahram Khadivi, and Christof Monz. Unilogit: Robust machine unlearning for LLMs using uniform-target self-distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22453–22472. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1154. URL https://aclanthology.org/2025.findings-acl.1154/.

Xu Wang, Zihao Li, Benyou Wang, Yan Hu, and Difan Zou. Model unlearning via sparse autoencoder subspace guided projections. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26530–26546. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.1348.

Yaxuan Wang, Jiaheng Wei, Chris Yuhao Liu, Jinlong Pang, Quan Liu, Ankit Parag Shah, Yujia Bao, Yang Liu, and Wei Wei. LLM unlearning via loss adjustment with only forget data. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Rongzhe Wei, Peizhi Niu, Hans Hao-Hsun Hsu, Ruihan Wu, Haoteng Yin, Mohsen Ghassemi, Yifan Li, Vamsi K. Potluru, Eli Chien, Kamalika Chaudhuri, Olgica Milenkovic, and Pan Li. Do LLMs really forget? evaluating unlearning with knowledge correlation and confidence awareness. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=BmEH70Wjcu.

Ruihan Wu, Chhavi Yadav, Russ Salakhutdinov, and Kamalika Chaudhuri. Evaluating deep unlearning in large language models. *arXiv preprint arXiv:2410.15153*, 2024.

Ruihan Wu, Konstantin Garov, and Kamalika Chaudhuri. Learning-time encoding shapes unlearning in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Abudukelimu Wuerkaixi, Qizhou Wang, Sen Cui, Wutong Xu, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Adaptive localization of knowledge negation for continual LLM unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68094–68117, 2025.

Evžen Wybitul, Tim G. J. Rudner, and Christian Schroeder de Witt. Entangled representations amplify collateral damage in unlearning. *arXiv preprint arXiv:2609.02285*, 2026.

Haoming Xu, Ningyuan Zhao, Liming Yang, Sendong Zhao, Shumin Deng, Mengru Wang, Bryan Hooi, Nay Oo, Huajun Chen, and Ningyu Zhang. ReLearn: Unlearning via learning for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5967–5987. Association for Computational Linguistics, 2025a.

Xiaoyu Xu, Minxin Du, Qingqing Ye, and Haibo Hu. OBLIVIATE: Robust and practical machine unlearning for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3696–3715. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.emnlp-main.183.

Xiaoyu Xu, Minxin Du, Kun Fang, Yaxin Xiao, Zhicong Huang, Cheng Hong, Qingqing Ye, and Haibo Hu. FIT to forget: Robust continual unlearning for large language models. *arXiv preprint arXiv:2601.21682*, 2026a.

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 *Proceedings of the 43rd International Conference on Machine Learning*, 2026b.

Tomoya Yamashita, Akira Ito, Yuuki Yamanaka, Masanori Yamada, Takayuki Miura, and Toshiki Shibahara. Sparse-autoencoder-guided internal representation unlearning for large language models. *arXiv preprint arXiv:2509.15631*, 2025.

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.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=8Dy42ThoNe.

Xiaotian Ye, Mengqi Zhang, and Shu Wu. LLM unlearning should be form-independent. *arXiv preprint arXiv:2506.07795*, 2025.

Sangyeon Yoon, Yeachan Jun, and Albert No. Position: The term "machine unlearning" is overused in LLMs. In *Proceedings of the 43rd International Conference on Machine Learning (Position Paper Track)*, 2026.

Jiatong Yu, Yinghui He, Anirudh Goyal, and Sanjeev Arora. On the impossibility of retrain equivalence in machine unlearning. In *Third Conference on Language Modeling (COLM)*, 2026.

Miao Yu, Liang Lin, Guibin Zhang, Xinfeng Li, Junfeng Fang, Xingrui Yu, Ivor Tsang, Ningyu Zhang, Kun Wang, and Yang Wang. UniErase: Towards balanced and precise unlearning in language models. *arXiv preprint arXiv:2505.15674*, 2025.

Qingjie Zhang, Haoting Qian, Zhicong Huang, Cheng Hong, Minlie Huang, Ke Xu, Chao Zhang, and Han Qiu. Understanding the dilemma of unlearning for large language models. *arXiv preprint arXiv:2509.24675*, 2025a.

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

Shengming Zhang, Le Zhang, Jingbo Zhou, Zhi Zheng, and Hui Xiong. LLM-Eraser: Optimizing large language model unlearning through selective pruning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1960–1971, 2025b. doi: 10.1145/3690624.3709312.

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

Nicolas Zucchet, Jörg Bornschein, Stephanie C. Y. Chan, Andrew Kyle Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? Dynamics, curricula and hallucinations. In *Second Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=vBcGnragkr.

# A Where the Evidence Is

Table 4 maps each main-text claim to the appendix section that gives its full numbers.

# B Setup Details

The data, training runs, retraining baselines, evaluation and reset-and-repair settings below lie behind Section 2 and Table 1. Every value comes from the run's own training configuration.

## B.1 Synthetic people and splits

Each synthetic person has six attributes: birth date, birth city, university, major, employer and work city. Each trained person has five biography variants and one question per attribute. The forget, retain and held-out sets are sets of people, so no paraphrase of a forget question can enter training through a retain person. Table 5 lists the sets in every setting.

Held-out accuracy means different things across settings. In Inserted the held-out people never appear in training, so their accuracy sits near zero by construction (0.5 on the 500-person model) and meets the holdout criterion trivially. In Two-stage the model saw their biographies but not their questions and answers them with its question-answering skill. In Mixed the training set contains the questions of some held-out people as well (83 of 100 at 500 people and 83 of 200 at 1,000 with all six, plus one with two) and only the biographies of the rest; reset-and-repair removes both kinds equally (Appendix E.2). In both settings held-out accuracy measures knowledge the model holds. In Protocol U their questions are trained as well. All eight Protocol U models are run with the held-out people inside the retain set, which is the 460 or 920 people listed in Table 5 (Appendix D.4). One of them (500 people, 2,500 steps) is also run with the retain set restricted to the 360 original retain people (Appendix E.3).

## B.2 Training runs

All Qwen3 models use the Qwen3-0.6B architecture (28 layers, width 1,024, 596M parameters with tied embeddings) and 512-token sequences. Training uses Muon for weight matrices and AdamW for the embedding and output layers, with base learning rates 0.02, 0.3 and 0.004 scaled by $\sqrt{\text{batch tokens}/524{,}288}$, no warmup, a linear warmdown over the last half of the configured steps, and weight decay 0.2. Table 6 lists the runs.

**Table 4. Every main-text claim has its full numbers in the appendix.**

| Main-text claim | Appendix |
|---|---|
| Three settings, their splits, epochs and pre-unlearning accuracy (Table 1) | B |
| Gradient methods: 247 evaluations on Two-stage, best $+1.67$, retraining $+99.38$ | C |
| Tuned gradient methods on Inserted, best $+15.54$ (RMU) | C.2 |
| Shared question-answering skill (D1 99.33) and the other diagnostics | C.3 |
| Reset-and-repair on Inserted, $+91.1$, 93% of retraining | D.1 |
| Fisher ranking matters on Inserted Qwen3 ($6.1\times$ random) | D.1.1 |
| Other Inserted models: Llama-3.2-1B, 1,000 and 2,000 people, continued pretraining, nanochat-d12, free-prose biographies | D.1.2 |
| Sampling, paraphrase, relearning, quantization and other question formats | D.1.3 |
| Mixed: holdout loses 44–104% of what the targets lose, 17 configurations | D.2, Fig. 17 |
| Matched retraining costs the holdout 5.8 and 8.7 points (2,500 steps, two data orders) | E.1, Fig. 17 |
| Only the retain set keeps a group's loss small | E.3, D.4 |
| The holdout's facts are gone from biography completion too | E.4 |
| Fisher-tail mass ranks forget and holdout loss together ($\rho \approx 0.83, 0.93$), stable across Fisher samples | D.3 |
| A fit frozen before training two further Mixed models predicts both (pre-registered) | D.3, Table 27 |
| Fisher-ranked mask on Mixed: $13\times$ random at 20%, holdout loses 65% as much | E.5, Fig. 23 |
| TOFU: gradient difference keeps its retain half at 80.8, untouched half falls to 62.2 (retrained 81.4 / 81.9; 3 seeds) | E.6, Fig. 24 |
| TOFU: real-author questions fall 12.0 under gradient difference, world facts do not; NPO lowers both halves; SimNPO barely forgets | E.6, Fig. 25 |
| Two-stage runs with a mismatched reset reference (appendix only) | D.5 |
| Qwen3-1.7B: the Fisher tail nearly vanishes and no tested reset budget preserves utility (appendix only) | D.6 |
| Real corpora: MUSE-News on Llama-2-7B (31 gradient configurations; Fisher, random and repair-only arms) and OLMo-1B on LUME | D.7 |
| Crossover: $487\times$ difference in Fisher share, margins $+1.46$ / $-1.54$, 7 of 10 crossings, 185 of 196 | F, Fig. 26 |
| Why the crossover uses fixed-budget resets: budget confound and collapse under the threshold schedule | F.3 |
| Held-out questions trained vs never trained lose the same (70.2 / 70.1 points) | E.2, Table 33 |
| Llama-3.2-1B at three more training lengths: holdout loses 84–95% as much as the targets | D.2, Table 24 |
| Rotary convention: disclosure and reruns with the standard one | B.3, Table 7 |
| Crossover: one-shot threshold reset (holdout loses 88%) and a reset reference that knows the holdout | F.3, Table 40 |
| Scope of the claims | G |

**Table 5. Every setting except Inserted trains the held-out people's biographies.** People in each set, and whose biographies and questions the model saw during training. "Held-out" is the untargeted holdout wherever the model can answer it. Two-stage evaluation samples 1,000 forget, 100 retain and 100 held-out people.

| Setting | Size | Forget / retain / held-out | Biographies trained for | QA trained for |
|---|---|---|---|---|
| Inserted | 500 | 40 / 360 / 100 | forget + retain | forget + retain |
| Inserted | 1,000 | 80 / 720 / 200 | forget + retain | forget + retain |
| Inserted | 2,000 | 160 / 1,440 / 400 | forget + retain | forget + retain |
| Mixed | 500 | 40 / 360 / 100 | everyone | forget + retain + 84 of 100 held-out |
| Mixed | 1,000 | 80 / 720 / 200 | everyone | forget + retain + 84 of 200 held-out |
| Protocol U | 500 | 40 / 460 (360 + 100) / 100 | everyone | everyone |
| Protocol U | 1,000 | 80 / 920 (720 + 200) / 200 | everyone | everyone |
| Two-stage | $\sim$100K | 1,000 / $\sim$49K / $\sim$50K | everyone (stage 1) | forget + retain (stage 2) |
| Crossover | 50K | focal sets of 400 | see Appendix F | scaffold people only |

The Mixed models were trained for several step counts to vary how saturated the model is. At 500 people, the model labelled "2,500 steps" in Table 23 and a later rerun at the same configuration are two different models; at 1,000 people, the grid model uses 4,000 steps and a separate model uses 2,500. Llama-3.2-1B-Base (Meta AI, 2024), with 1.24B parameters, replaces Qwen3 in one Mixed and one Inserted model. The nanochat-d12 model (Karpathy, 2025) of the cross-architecture check

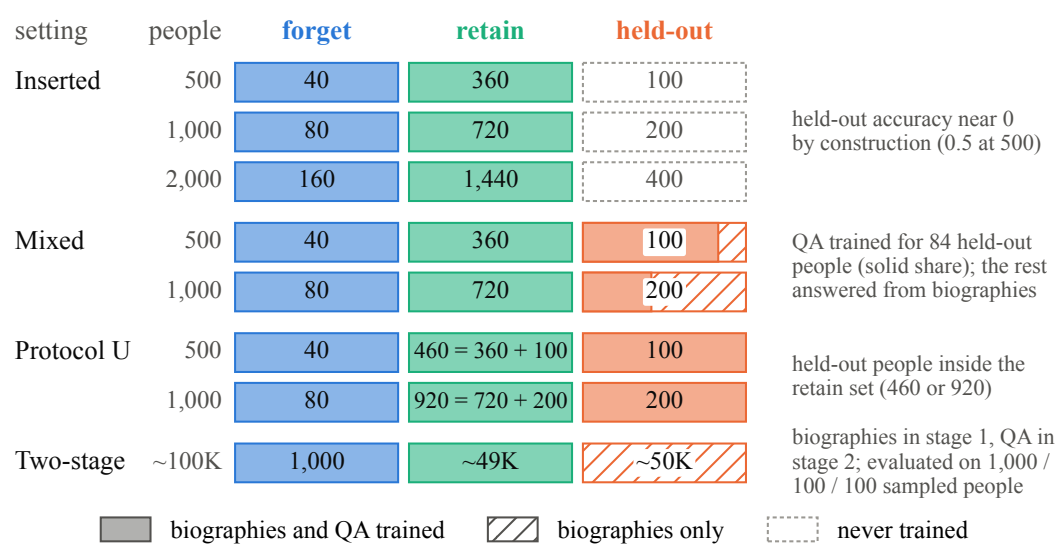



**Figure 7. Every setting except Inserted trains the held-out people's biographies, so only on Inserted is held-out accuracy near zero by construction.** Number of people in each set. Solid: biographies and questions trained; hatched: biographies only; dashed outline: never trained. In Mixed, the solid share of the held-out cell is the people whose questions are also trained (84 of 100; 84 of 200). In Protocol U the held-out people are also inside the retain set. The crossover is omitted (Table 5).

**Table 6. Two-stage is the only setting trained from random initialization; the others start from pretrained Qwen3-0.6B-Base.** "Epochs" is the number of passes over the training data at the checkpoint used. The Inserted checkpoint at step 2,500 comes from a run configured for 10,000 steps, so it was taken at full learning rate, before the warmdown began.

| Setting | Initialization and data | Steps | Tokens / step | Epochs |
|---|---|---|---|---|
| Inserted, 500 | Qwen3-0.6B-Base; biographies + QA for 400 people (2,000 biographies, 2,400 questions); loss on answer tokens only for QA rows | 2,500 | 6,144 | 70 |
| Inserted, 1,000 / 2,000 | same recipe | 5,000 | 6,144 | 70 / 35 |
| Mixed, 500 | Qwen3-0.6B-Base; biographies for all 500 people + QA for the 400 forget and retain people and for 83 of the 100 held-out people | 500–2,500 | 24,576 | 262 at 2,500 |
| Mixed, 1,000 | same recipe; QA for 84 of the 200 held-out people | 500–4,000 | 24,576 | 210 at 4,000 |
| Protocol U | as Mixed, plus QA for every held-out person | 1,250–12,000 | 24,576 | – |
| Two-stage, stage 1 | random init, Qwen3-0.6B architecture; ∼100K people's biographies (501K rows) | 190,000 | 49,152 | ∼199 |
| Two-stage, stage 2 | QA LoRA (attention rank 16, embedding rank 128, $\alpha$=4, dropout 0.2, AdamW $3\times10^{-4}$) on 50K people; merged at step 15,000 | 15,000 | 24,576 | – |

has 12 layers and width 768 and is trained from random initialization; of its 1.02B parameters, about 932M are embedding tables.

### B.3 Rotary convention and reruns with the standard one

The Qwen3 models in this paper apply rotary position embeddings with the opposite sign to the released Qwen3 weights. The Inserted, Mixed, Protocol U and crossover models are initialized from

Qwen3-0.6B-Base and trained to near-perfect accuracy under this convention; the Two-stage models start from random weights, for which the sign is only a choice of encoding. The Llama-3.2-1B models, trained with the same pipeline, use the standard convention. We found the mismatch after the experiments. Three kinds of evidence bear on it (Table 7).

**The Llama models use the standard convention and show the same holdout loss.** On Llama-3.2-1B Mixed (Table 23), reset-and-repair costs the holdout 76.8 points against the targets' 97.3, while retain accuracy stays at 97.6.

**Matched Inserted reruns.** Retraining the 500-person Inserted model from Qwen3-0.6B-Base with the standard convention, in the same code as the paper, reset-and-repair opens a forget gap of $+74.83$ (4 seeds) against $+91.1$ in the paper, and one Fisher-ranked reset of 21.4% of parameters opens $3.5\times$ the gap of a random reset of the same size (one Fisher run against four random seeds), against $6.1\times$. A fifth seed is excluded because its evaluation scored the whole population as the retain set. The removal is weaker but of the same kind; Inserted has no answerable holdout, so these runs do not test selectivity.

**Mixed and Protocol U models trained from scratch with the standard convention.** A later version of our code trained Mixed and Protocol U models from random weights with the standard convention and reset them to their random initialization. Its repair step ran in higher precision than the paper's, which its authors found makes repair remove more, so we use these runs only for the direction of each effect. On the Mixed model (500 people, 2,500 steps), reset-and-repair keeps retain accuracy at 98.9 while the holdout falls from 97.5 to 16.1 (6 seeds); with the embeddings excluded from the reset, as in the Protocol U runs, the holdout falls to 19.8. Held-out people inside the retain set (Protocol U) lose at most 1.7 points. No matched retraining exists for these models.

**The Fisher-tail ranking is untested under the standard convention.** The only standard-convention Fisher-tail grid is Inserted, with one run per cell and fixed reset budgets rather than the threshold schedule that ties $M$ to the first round's reset. Its forget gaps are near saturation ($+83$ to $+97$), and their Spearman correlation with $M$ is $+0.67$ over eight cells ($p = 0.07$) and $+0.14$ over 18. It does not test the Mixed relation of Appendix D.3, which we therefore claim only for the models studied there; the claims that the holdout loses much of what the targets lose, and that retain-set membership prevents it, do not depend on it.

**Table 7. Under the standard rotary convention, reset-and-repair still keeps retain accuracy and removes most of the holdout, except for held-out people in the retain set.** Accuracy in %, mean over seeds ($\pm$ population SD where given). Inserted: retrained from Qwen3-0.6B-Base with the paper's code. Mixed and Protocol U: trained from random weights with a later version of the code whose repair ran in higher precision; direction only.

| Model and method | Seeds | Forget | Retain | Held-out |
|---|---|---|---|---|
| Inserted 500, before unlearning | – | 100.00 | 99.77 | 0.83 |
| reset-and-repair, 4 rounds (gap $+74.83 \pm 1.22$) | 4 | 25.10 | 99.93 | 0.92 |
| Fisher-ranked reset of 21.4%, one round (gap $+45.00$) | 1 | 53.75 | 98.75 | 1.00 |
| random reset of 21.4%, one round (gap $+12.80 \pm 2.18$) | 4 | 87.08 | 99.88 | 0.75 |
| Mixed 500, 2,500 steps, from scratch, before unlearning | – | 98.75 | 99.44 | 97.50 |
| reset-and-repair, 4 rounds | 6 | $4.10_{\pm 1.66}$ | $98.94_{\pm 0.18}$ | $16.14_{\pm 1.65}$ |
| same, embeddings excluded from the reset | 4 | 20.42 | 99.31 | 19.75 |
| Protocol U 500, from scratch, reset-and-repair (before: 100 held-out) | 4 | $34.90_{\pm 3.80}$ | $99.40_{\pm 0.21}$ | $99.46_{\pm 0.18}$ |
| Protocol U 1,000, from scratch, reset-and-repair (before: 100 held-out) | 4 | $9.11_{\pm 1.94}$ | $98.59_{\pm 0.41}$ | $98.33_{\pm 0.54}$ |

### B.4 Retraining baselines

Each retraining baseline is a model trained without the forget people's data. It shows what selective removal looks like, and that our measurements detect it. Table 8 lists every baseline.

**Inserted.** The Inserted 500-person recipe is trained on the 360 retain people only (1,800 biographies and 2,160 questions); the 1,000- and 2,000-person versions leave out their forget people in the same way. Each reaches a large forget gap.

**Two-stage, 1,000 people.** Stage 1 pretrains from random initialization for 6,900 steps of 49,152 tokens on the biographies of the retain and held-out people (920 people, no forget people). Stage 2 trains a QA LoRA (attention rank 8, embedding rank 128) for 15,000 steps of 49,152 tokens on the 720 retain people's questions. The same recipe with the forget people included answers every forget and retain question, with Fisher-tail mass $M{=}21.3\%$. Without them the model cannot answer the forget people, keeps the retain people, and its held-out accuracy rises by 1.5. This smaller pipeline starts from random initialization and reads facts out through a QA LoRA, which is why its held-out accuracy starts at 67 while the Mixed models start near 100. Appendix E.1 gives a retraining baseline matched to the Mixed pipeline.

**Table 8. Every retraining baseline removes the forget people.** Accuracy on forget / retain / held-out people. Inserted held-out people are never trained, so their accuracy is near zero by construction. The Two-stage full-corpus row is the same recipe with the forget people included, not a baseline. Mixed rows are matched retraining (Appendix E.1), each against the full-corpus model of the same data order; their gap is not reported. TOFU's retraining reference (retain90) is in Table 37.

| Setting | Model | Forget | Retain | Held-out | Gap | Holdout drop |
|---|---|---|---|---|---|---|
| Inserted | Qwen3, 500 | 0.83 | 98.52 | 0.83 | +97.69 | – |
| | Qwen3, 1,000 | 1.04 | 98.33 | 0.92 | +97.29 | – |
| | Qwen3, 2,000 | 1.25 | 92.44 | 1.12 | +91.19 | – |
| | Qwen3, CPT | – | – | – | +94.17 | – |
| | nanochat-d12 | 2.08 | 99.81 | 1.67 | +97.73 | – |
| Two-stage | 1,000, full corpus | 100.00 | 100.00 | 67.33 | | |
| | 1,000, retrained | 0.62 | 100.00 | 68.83 | +99.38 | rises 1.5 |
| Mixed | 500, 1,000 steps, order 42 | 0.83 | 99.49 | 82.83 | – | 15.00 |
| | 500, 1,000 steps, order 1 | 1.67 | 99.81 | 87.83 | – | 6.83 |
| | 500, 2,500 steps, order 42 | 1.25 | 100.00 | 93.17 | – | 5.83 |
| | 500, 2,500 steps, order 1 | 0.83 | 100.00 | 89.33 | – | 8.67 |

### B.5 Evaluation

The prompt is the question followed by "`Answer:`". The model decodes greedily for up to 32 tokens, and an answer counts only if the stripped output equals the reference exactly. A set's accuracy is the unweighted mean of its six per-attribute accuracies. The forget gap is retain accuracy minus forget accuracy. The forget and holdout drops are measured against the same model before unlearning, and the selectivity margin is the forget drop minus the holdout drop.

Selective removal requires a forget gap of at least 30, retain accuracy of at least 95, and, wherever an untargeted holdout exists, a holdout drop at most 1.5 points larger than under retraining on the same pipeline.

### B.6 Reset-and-repair

Reset-and-repair adapts PerTA (Cai et al., 2026): we replace its continuous task-vector rescaling with a thresholded reset and add a repair step on the retain set after every round. Each round has three steps.

1. **Fisher estimate.** For each of 100 forget mini-batches of 8 sequences of 512 tokens, take the gradient of the mean token loss, square it elementwise, and sum over mini-batches to obtain $g^2_{F,i}$ for every parameter; $g^2_{R,i}$ is computed the same way from 100 retain mini-batches. The estimate is recomputed every round.
2. **Reset.** Compute $W_i = (g^2_{F,i} + \epsilon)/(g^2_{F,i} + g^2_{R,i} + 2\epsilon)$ with $\epsilon = 10^{-8}$, and copy every parameter with $W_i$ above the round's threshold back from the reference weights. All parameters are eligible, embeddings included. The reference is the pretrained base for Qwen3 and Llama models and the random initialization for nanochat-d12.
3. **Repair.** Train for 300 steps of 6,144 tokens on the retain people's biographies and questions, with the training optimizer, no warmup, and a linear decay over the last 150 steps.

We write "after $k$ rounds" for the checkpoint after the $k$-th repair; the Mixed 1,000-person 4,000-step model is reported after three rounds. Table 9 lists the threshold schedules. The Fisher-tail mass $M$ is the fraction of parameters with $W_i > 0.55$, which is also the fraction the first round resets under a 0.55 threshold.

**Table 9. Thresholds rise across rounds, so later rounds reset fewer parameters.** One threshold per round.

| Models | Thresholds |
|---|---|
| Mixed (all Qwen3 models, both forget splits), Protocol U, Inserted runs 5 and 6 | 0.55, 0.90, 0.95, 0.95 |
| Inserted runs 1–4 | 0.60, 0.90, 0.95, 0.95 |
| Llama-3.2-1B, Mixed and Inserted | 0.55, 0.85, 0.95, 0.975 |
| nanochat-d12, Inserted | 0.55, 0.85, 0.95 |
| Inserted 1,000 / 2,000, conservative | 0.55, 0.85, 0.95 |
| Inserted 1,000 / 2,000, strong | 0.45, 0.65, 0.85 (, 0.95) |
| Two-stage, aggressive / softened | 0.55 (or 0.60, 0.65, 0.75), 0.85 (0.90 with 0.75), 0.95, 0.95 |

# C Gradient Methods

No gradient configuration meets the success criterion (Section 3). In the 247-evaluation sweep on Two-stage, forget, retain and held-out accuracy fall together; on Inserted, the tuned methods leave the targets in place; and four diagnostics explain why.

## C.1 The Two-stage sweep

No evaluation in the sweep comes near the success threshold: the best forget gap is $+1.67$, against $+99.38$ for the retraining baseline. The sweep comprises 247 checkpoint evaluations from 47 configurations, with checkpoints saved every 250 steps; 237 of the evaluations use 50 people per group. Figure 8 shows every evaluation by method family, and Table 10 summarizes each family. The median gap is $-1.00$, and 8 of the 247 evaluations have any positive gap. The 222 LoRA evaluations never exceed $+0.67$; the overall maximum of $+1.67$ comes from the 25 full-parameter GA/GD evaluations on the merged 596M-parameter model.

The held-out people, whose questions no configuration trains on, fall with the targets as well. Measured against the Two-stage model before unlearning (99.0 held-out accuracy), the 59 evaluations that cost the targets 30 points or more cost the holdout 94.0% to 103.4% as much, and at least 30.8 points. No evaluation lowers forget accuracy by 10 points while losing less than 5 points of holdout, and the forget drop never exceeds the holdout drop by more than 5.25 points. Retain minus held-out accuracy after unlearning stays between $-4.0$ and $+3.0$. Scoring the evaluations against two other pre-unlearning checkpoints of the same model changes these ratios by at most 0.03.

**Table 10. Values behind Figure 8: no gradient method opens a forget gap on the Two-stage model.** Forget gap over the 247 checkpoint evaluations from 47 configurations, by family. LoRA runs fine-tune the Two-stage model's QA adapter at step 15,000 (attention rank 16, embedding rank 128); full-parameter runs update the same model with the adapter merged (99.91 / 100.00 / 99.00 before unlearning), or the stage-1 model. The success threshold is $+30$; the retraining baseline reaches $+99.38$.

| Family | Scope | Evaluations | Max | Median | Evaluations $> 0$ |
|---|---|---|---|---|---|
| GA | LoRA | 48 | $+0.67$ | $-1.33$ | 2 |
| GD | LoRA | 33 | 0.00 | $-0.67$ | 0 |
| NPO | LoRA | 38 | $-0.67$ | $-1.33$ | 0 |
| SimNPO | LoRA | 38 | $+0.67$ | $-1.67$ | 1 |
| DPO | LoRA | 30 | $-1.00$ | $-1.00$ | 0 |
| RMU | LoRA | 30 | $-1.00$ | $-1.33$ | 0 |
| Task arithmetic | LoRA | 5 | 0.00 | 0.00 | 0 |
| GA / GD | full parameter | 25 | $+1.67$ | 0.00 | 5 |
| All | | 247 | $+1.67$ | $-1.00$ | 8 |

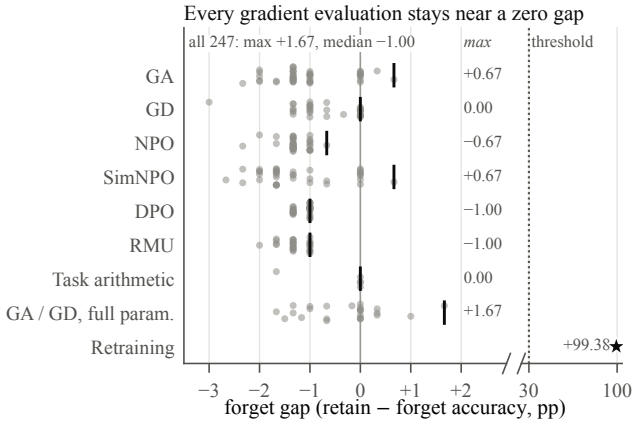



**Figure 8. No evaluation in the Two-stage gradient sweep comes near the success threshold.** Forget gap of each of the 247 checkpoint evaluations, by method family (points jittered vertically), on a broken axis: the left panel spans $-3.5$ to $+2.9$, where a tick marks each family's largest gap (printed on the right); the right panel holds the success threshold of $+30$ (dotted) and the retraining baseline's $+99.38$ (star). The largest gap is $+1.67$.

Unless listed otherwise, runs use learning rate $10^{-5}$, 5 epochs over the questions of the 1,000 forget people, 8 sequences per batch, 10% warmup and gradient clipping at 1.0. The sweep covers: GA at learning rates $10^{-5}$, $10^{-6}$, $5{\times}10^{-7}$ and $10^{-7}$; GD at $10^{-5}$, $10^{-4}$, $5{\times}10^{-5}$ and $2{\times}10^{-5}$; NPO (Zhang et al., 2024) at $10^{-5}$, $10^{-4}$, $10^{-3}$ and $5{\times}10^{-3}$, plus $\beta$=0.1; SimNPO (Fan et al., 2025) at $10^{-5}$, $10^{-4}$, $10^{-3}$ and $5{\times}10^{-4}$, plus $\beta$=0.1; DPO (Rafailov et al., 2023) at the default, $\beta$=0.1 and learning rate $10^{-4}$; RMU (Li et al., 2024) targeting layers 7, 14 or 27, plus two coefficient variants; task arithmetic (Ilharco et al., 2023) at $\alpha \in \{0.1, 0.3, 0.5, 1, 2\}$. The full-parameter runs, the only ones with every parameter unfrozen, are GA and GD on the merged model with QA or biography data, and GA directly on the stage-1 model before any QA training.

The sweep does not include model editors such as ROME (Meng et al., 2022) or MEMIT (Meng et al., 2023), which write a new object for a single subject–relation pair; our targets are six free-form attributes per person with no natural replacement answer. It also does not include recent variants of, and alternatives to, the gradient families (Wang et al., 2025b; Ji et al., 2024; Vasilev et al., 2025; Ren et al., 2025; Xu et al., 2025a; Cha et al., 2025).

### C.2 Tuned gradient methods on Inserted

Tuned on Inserted, no gradient method comes near the success threshold, and four of the five leave the targets almost untouched (Figure 9, Table 11). Each method unlearns with full-parameter updates on the forget people's biographies (4 sequences per batch, 10% warmup, clipping 1.0, weight decay 0.01) and is tuned over a 3×3 grid: GA and GD over learning rates $\{3{\times}10^{-6}, 10^{-5}, 3{\times}10^{-5}\}$ and $\{200, 400, 800\}$ steps; NPO and SimNPO over the same rates and $\beta \in \{0.05, 0.1, 0.5\}$; RMU over rates $\{1.5{\times}10^{-5}, 5{\times}10^{-5}, 1.5{\times}10^{-4}\}$ and steering coefficients $\{30, 100, 300\}$. For each method we kept the configuration with the largest forget gap among those with retain accuracy of at least 80, a looser bar than the success criterion's 95, and reran it with seeds 234, 345 and 456 (GA with one seed).

**Table 11. Values behind Figure 9: tuned gradient methods on Inserted leave the targets in place.** Best configuration per method under the retain $\geq$ 80 rule; 3 seeds (GA 1 seed). Reset-and-repair on the same model reaches $+91.1 \pm 3.0$ (6 runs, Appendix D.1).

| Method | Selected configuration | Forget gap | Retain |
|---|---|---|---|
| NPO | lr $3{\times}10^{-6}$, $\beta$=0.5 | $-1.28 \pm 0.02$ | 96.6 |
| SimNPO | lr $3{\times}10^{-6}$, $\beta$=0.5 | $-1.82 \pm 0.02$ | 95.7 |
| GA | lr $3{\times}10^{-6}$, 400 steps | $+1.39$ | 82.6 |
| GD | lr $10^{-5}$, 200 steps | $+0.63 \pm 0.10$ | 89.8 |
| RMU | lr $1.5{\times}10^{-4}$, coefficient 30 | $+15.54 \pm 0.85$ | 96.8 |

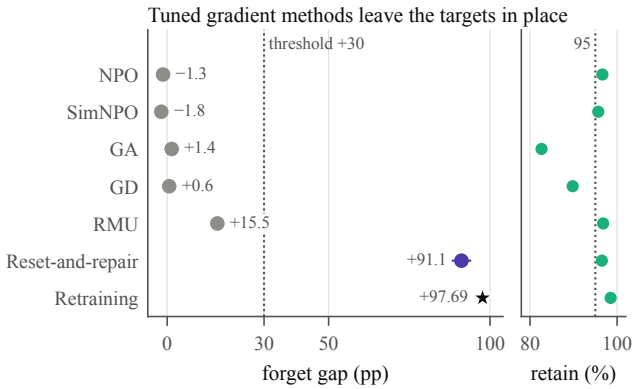



**Figure 9. Tuned on Inserted, the gradient methods leave the targets in place, far below reset-and-repair on the same model.** Left: forget gap of each method's best configuration under the retain $\geq 80$ rule, mean $\pm$ SD over seeds (3 seeds, GA 1; the bars are mostly narrower than the markers). Reset-and-repair is the mean of the six runs of Appendix D.1, and the star is the retraining baseline. The dotted line is the success threshold. Right: retain accuracy of each row, with the success criterion's 95 dotted.

### C.3 Diagnostics D1–D4

Four diagnostics point to one question-answering skill shared by forget and retain questions, which is why gradient methods lower both together (Table 12). All four use one two-stage model from an earlier pipeline than the sweep's: the Qwen3-0.6B architecture trained from random initialization for 15,000 steps of 49,152 tokens on the biographies of 51K people, followed by a QA LoRA (attention rank 16, embedding rank 128) for 15,000 steps. D1, D2 and D4 go through a LoRA adapter.

**Table 12. All four diagnostics point to one question-answering skill shared by forget and retain people.** Each row gives what the diagnostic tests and what we observe on the two-stage model of this subsection.

| Diagnostic | Procedure | What the diagnostic tests | Observed |
|---|---|---|---|
| D1 retain-only adapter | Train a QA LoRA on retain people only; ask forget questions | whether a retain-trained adapter transfers to forget people's biographies | 99.33 |
| D2 "Unknown" control | Train a LoRA to answer "Unknown" for forget people (one run) | whether this adapter intervention suppresses the scored answers | 94.20 |
| D3 probes | Logistic regression (5-fold) on the last name token at 8 layers, forget vs. retain | whether forget membership is linearly decodable at the probed layers | 0.505 random, 0.530 structural split |
| D4 adapter SVD | Effective rank (exp of singular-value entropy) of 56 attention adapters of rank 16 | whether the adapter updates are low-rank | 1.14; 96.7% of energy in the top direction |

D1 is the strongest of the four. An adapter that never saw a forget question answers 99.33% of them, so answering a forget question needs nothing beyond what the retain data trains. D2 shows that the model keeps the facts under an adapter trained to deny them. D3 and D4 are indirect: the probes find no separation between forget and retain people at the eight layers probed, and the SVD shows that the QA update is close to one direction.

## D Reset-and-Repair Results

Reset-and-repair removes the targets on Inserted and Mixed, and wherever an answerable untargeted holdout exists, it removes much of that holdout too (Section 4). We report every reset-and-repair model: Inserted with its controls and variants, all 17 Mixed models, the Fisher-tail grid, Protocol U, the Two-stage runs that the main text sets aside, Qwen3-1.7B, and two real corpora.

### D.1 INSERTED

On Inserted, reset-and-repair removes the targets and keeps the retain set, reaching 93% of the retraining baseline's forget gap. The main-text $+91.1 \pm 3.0$ is the mean of six runs covering four forget splits, ranging from $+87.7$ to $+95.2$; Figure 10(a) shows them and Table 13 lists them with their schedules. Three runs share the main forget split; the other three each use their own forget split on a separately trained 5,000-step model, and those splits have no held-out people.

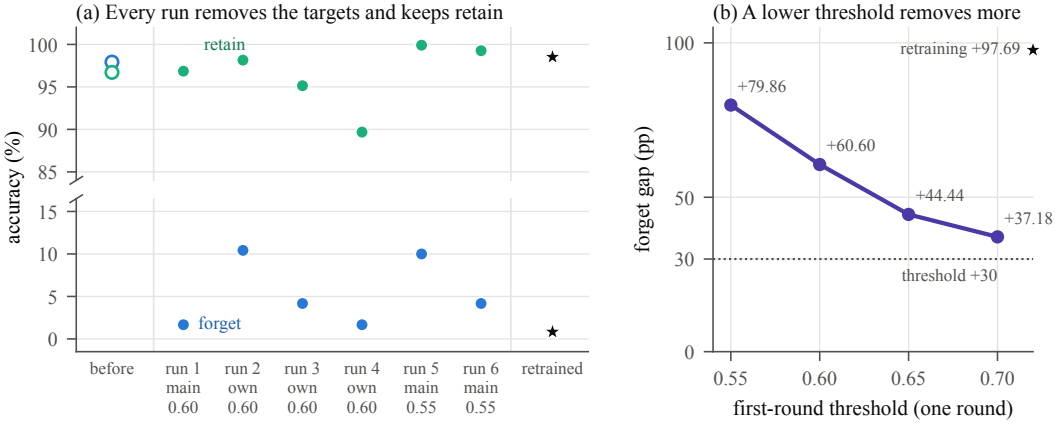



**Figure 10. Every reset-and-repair run on Inserted removes the targets and keeps retain accuracy at 89.7 or above, and a single round removes more the lower its first threshold.** (a) Forget and retain accuracy of the six runs of Table 13, on a y-axis broken between 15 and 85; under each run, its forget split (main or its own) and first threshold. Hollow circles mark the main model before unlearning and stars the retraining baseline. (b) Forget gap after a single round at each first threshold; the star marks the retraining baseline's $+97.69$ and the dotted line the success threshold.

**Table 13. Values behind Figure 10: every reset-and-repair run on Inserted removes the targets.** Four rounds; six runs over four forget splits, whose mean is the main-text number. The base model before unlearning scores 97.92 / 96.71 / 0.50. Held-out accuracy is "–" for runs whose split has no held-out people. The last two rows are our reimplementation of PerTA's update, gradient- and Fisher-weighted, applied in one step with continuous rescaling and no repair.

| Run | Forget split, model | First threshold | Forget | Retain | Held-out | Gap |
|---|---|---|---|---|---|---|
| run 1 | main, 2,500 steps | 0.60 | 1.67 | 96.85 | 1.00 | +95.19 |
| run 2 | own split, 5,000 steps | 0.60 | 10.42 | 98.15 | – | +87.73 |
| run 3 | own split, 5,000 steps | 0.60 | 4.17 | 95.14 | – | +90.97 |
| run 4 | own split, 5,000 steps | 0.60 | 1.67 | 89.68 | – | +88.01 |
| run 5 | main, 2,500 steps | 0.55 | 10.00 | 99.91 | 0.33 | +89.91 |
| run 6 | main, 2,500 steps | 0.55 | 4.17 | 99.26 | 0.17 | +95.09 |
| Mean | | | 5.35 | 96.50 | | $+91.1 \pm 3.0$ |
| Retraining baseline | main | | 0.83 | 98.52 | 0.83 | +97.69 |
| PerTA, gradient | | – | 0.42 | 9.72 | | +9.31 |
| PerTA, Fisher | | – | 0.42 | 8.89 | | +8.47 |

Runs 5 and 6 share the model, split and data order, so they differ only through GPU nondeterminism. The gap is already there before the fourth round: stopping after three rounds gives $+93.55 \pm 1.70$ on four runs of an earlier variant. Our reimplementation of PerTA's update, which applies the reset in one step with continuous rescaling and no repair, removes the targets and damages the whole model, in both its gradient- and Fisher-weighted variants (last rows of Table 13). PerTA had no released code at the time, so we did not run its pipeline or benchmarks.

A lower first threshold resets more and removes more, with no sharp optimum: a single round at 0.55, 0.60, 0.65 or 0.70 gives $+79.86$, $+60.60$, $+44.44$ and $+37.18$. Deleting the 40 forget people in two batches of 20, one round each, gives $+62.73$ after the first batch (all 40 scored as forget) and $+97.31$ after the second (one run).

### D.1.1 FISHER RANKING AGAINST RANDOM MASKS

On this Qwen3 model, resetting the parameters with the highest $W$ removes far more than resetting the same number of random parameters. Figure 11 and Table 14 compare masks after one round with repair, all resetting the 21.4% of parameters that a 0.55 threshold selects unless noted.

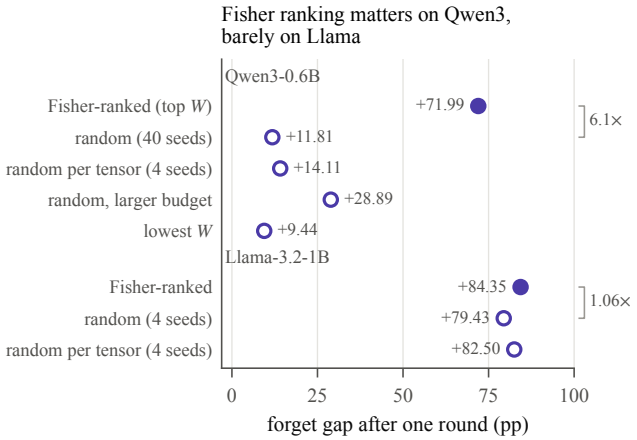



**Figure 11. On Inserted Qwen3 the Fisher-ranked mask opens $6.1\times$ the gap of a random mask of the same size; on Llama-3.2-1B random masks do almost as well.** Forget gap after one round with repair. Filled: Fisher-ranked; hollow: other masks, mean $\pm$ SD over mask seeds where several were run. "Per tensor" matches the Fisher mask's count in every tensor. Brackets give the ratio of the Fisher-ranked gap to the same-size random gap.

**Table 14. Values behind Figure 11: on Inserted Qwen3, the Fisher-ranked mask removes $6.1\times$ as much as a random mask of the same size.** One round with repair. Random rows give mean $\pm$ SD over mask seeds. The last column is the Fisher-ranked mask's gap divided by the row's gap. The Llama rows use the same comparison on Llama-3.2-1B Inserted.

| Model | Mask | Forget | Retain | Gap | Fisher / this |
|---|---|---|---|---|---|
| Qwen3 | Fisher-ranked (top $W$) | 27.92 | 99.91 | $+71.99$ | |
| Qwen3 | random, same size (40 seeds) | 88.10 | 99.91 | $+11.81 \pm 1.64$ | $6.1\times$ |
| Qwen3 | first 21 seeds only | | | $+11.48 \pm 1.92$ | $6.3\times$ |
| Qwen3 | random, same size per tensor (4 seeds) | 85.73 | 99.84 | $+14.11 \pm 0.92$ | $5.10\times$ |
| Qwen3 | random, larger budget | 70.83 | 99.72 | $+28.89$ | |
| Qwen3 | lowest $W$ | 90.42 | 99.86 | $+9.44$ | |
| Llama | Fisher-ranked | 12.50 | 96.85 | $+84.35$ | |
| Llama | random, same size (4 seeds) | | | $+79.43 \pm 1.60$ | $1.06\times$ |
| Llama | random, same size per tensor (4 seeds) | | | $+82.50 \pm 1.15$ | $1.02\times$ |

On Qwen3 the Fisher-ranked mask opens several times the gap of a random mask, whether or not the random mask matches the Fisher mask's count in every tensor, a stricter form of the random-mask comparison in Fan et al. (2024) (last column of Table 14). On Llama the ratios are close to 1, so there the method removes the targets with almost no help from the ranking.

### D.1.2 OTHER DENSITIES, CONTINUED PRETRAINING, AND OTHER ARCHITECTURES

Larger Inserted models need a stronger schedule, and the result carries over to Llama but only partly to nanochat-d12. Figure 12 shows every run on these models, and Table 15 collects them. The 1,000- and 2,000-person schedules were tuned on the evaluation sets, so these rows show that a working schedule exists at each size.

At 1,000 people the strong schedule's gap varies by seed, down to $+32.85$ with retain accuracy 35.56. The Llama model starts at 100 / 100 / 0.83. The nanochat-d12 model starts at 100.00 / 99.54 / 1.33; its 12 runs span five regenerated reset references (initializations 0–4, two to four seeds each), and the per-reference means range from $+33.73$ to $+37.89$. An earlier single d12 run that reported $+90.74$ is

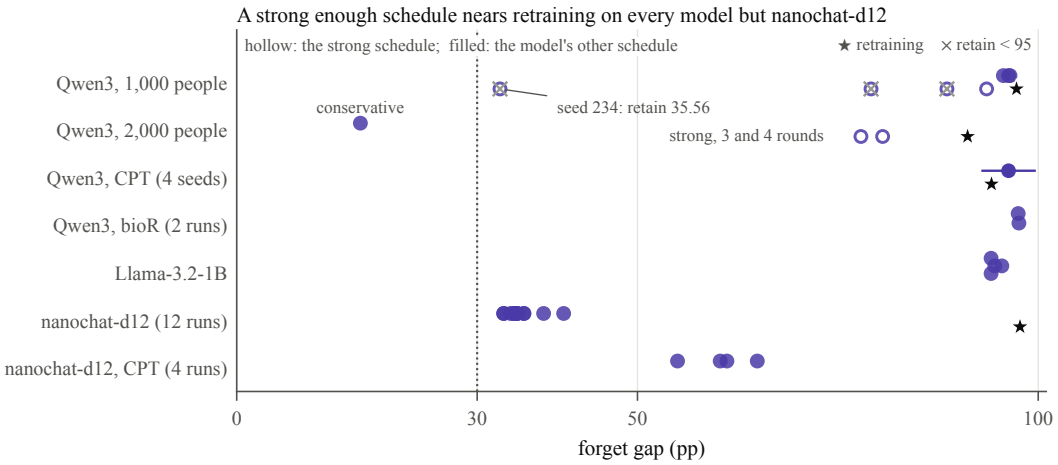



**Figure 12. With a strong enough schedule, reset-and-repair approaches retraining on every Inserted model here except nanochat-d12.** Forget gap of every run in Tables 15, 17, 18 and 16; stars mark the retraining baselines of Table 15, and the dotted line marks the success threshold. Hollow markers use the strong schedule; on the 1,000-person model its gap varies by seed. A × marks each run whose retain accuracy is below 95. Runs with near-identical gaps are offset vertically. CPT is shown by its 4-seed mean ± SD.

**Table 15. Values behind Figure 12: with a strong enough schedule, reset-and-repair opens a large forget gap on every model here except nanochat-d12.** Seeds per row as listed. "CPT" inserts the biographies by continued pretraining of Qwen3-0.6B-Base before QA training. "Strong" is the schedule 0.45, 0.65, 0.85 (, 0.95).

| Model | Run | Seeds | Forget | Retain | Held-out | Gap |
|---|---|---|---|---|---|---|
| Qwen3, 1,000 | 3 rounds | 1 | 2.71 | 98.36 | 0.92 | +95.65 |
| | 3 rounds | 1 | 2.50 | 98.80 | 1.00 | +96.30 |
| | 4 rounds | 1 | 2.50 | 98.98 | 1.08 | +96.48 |
| | strong, calibration seed | 1 | | | | +93.56 |
| | strong, seed 234 | 1 | 2.71 | 35.56 | | +32.85 |
| | strong, seed 345 | 1 | 4.58 | 93.19 | | +88.61 |
| | strong, seed 456 | 1 | 1.25 | 80.39 | | +79.14 |
| | retraining baseline | 1 | 1.04 | 98.33 | 0.92 | +97.29 |
| Qwen3, 2,000 | conservative, 3 rounds | 1 | 84.27 | 99.70 | 1.12 | +15.43 |
| | strong, 3 rounds | 1 | 21.15 | 99.03 | 1.54 | +77.88 |
| | strong, 4 rounds | 1 | 18.65 | 99.24 | 1.12 | +80.59 |
| | retraining baseline | 1 | 1.25 | 92.44 | 1.12 | +91.19 |
| Qwen3, CPT | 1 round | 4 | | | | +96.3 ± 3.4 |
| | retraining baseline | 1 | | | | +94.17 |
| Llama-3.2-1B | 4 rounds | 4 | 4.27 | 98.84 | 0.71 | +94.57 ± 0.55 |
| nanochat-d12 | 3 rounds | 12 | 64.51 | 99.98 | 1.03 | +35.47 ± 2.04 |
| | retraining baseline | 1 | 2.08 | 99.81 | 1.67 | +97.73 |

excluded, because its unseeded reset reference was not saved. Table 17 lists the 12 runs and Table 18 the Llama runs; Figure 13 shows the Llama Mixed gap round by round.

**Reset-and-repair also removes the targets when the biographies are free prose, and from nanochat-d12 when the biographies are inserted by continued pretraining** (Table 16). In the first model, a language model writes each biography as free prose, the bioR format of Allen-Zhu & Li (2024), in place of our templates. Qwen3-0.6B-Base is trained on these biographies and question–answer pairs for 2,500 steps of 65,536 tokens and starts at 100 / 100 / 0.67. Three rounds reach +97.55 ± 0.05 over two runs. In the second model, nanochat-d12 learns the biographies by continued pretraining and then learns the questions; it starts at 100 / 99.81 / 0.67. Because d12 has no pretrained base, its reset reference is the model after continued pretraining, before question training.

Three rounds reach $+60.4 \pm 3.6$ over four runs, against $+35.47$ for d12 trained on biographies and questions together.

**Table 16. Values behind the bioR and d12-CPT rows of Figure 12: reset-and-repair removes the targets from a free-prose biography model and from nanochat-d12 with continued pretraining.** Three rounds with thresholds 0.55, 0.85, 0.95. The first run of each model uses the default order of Fisher mini-batches; the others set seed 234 (bioR) or 345, 456 and 567 (d12).

| Model | Run | Forget | Retain | Held-out | Gap |
|---|---|---|---|---|---|
| Qwen3, bioR biographies | 1 | 2.50 | 100.00 | 0.17 | +97.50 |
| | 2 | 2.08 | 99.68 | 0.33 | +97.59 |
| | Mean | 2.29 | 99.84 | 0.25 | $+97.55 \pm 0.05$ |
| nanochat-d12, continued pretraining | 1 | 38.75 | 99.91 | 0.83 | +61.16 |
| | 2 | 35.00 | 99.95 | 0.83 | +64.95 |
| | 3 | 39.58 | 99.91 | 1.00 | +60.32 |
| | 4 | 45.00 | 100.00 | 0.67 | +55.00 |
| | Mean | 39.58 | 99.94 | 0.83 | $+60.36 \pm 3.55$ |

**Table 17. nanochat-d12 on Inserted: the gap is stable across reset references.** Three rounds. Each reference is a regenerated random initialization used as the reset target; two to four seeds per reference, 12 runs.

| Reference | Seed | Forget | Retain | Held-out | Gap |
|---|---|---|---|---|---|
| 0 | 42 | 64.17 | 100.00 | 0.50 | +35.83 |
| 0 | 345 | 65.00 | 100.00 | 0.83 | +35.00 |
| 0 | 456 | 65.42 | 100.00 | 1.00 | +34.58 |
| 0 | 567 | 64.17 | 100.00 | 1.50 | +35.83 |
| 1 | 42 | 61.67 | 99.95 | 1.17 | +38.29 |
| 1 | 345 | 65.42 | 99.95 | 0.83 | +34.54 |
| 2 | 42 | 65.00 | 99.95 | 1.83 | +34.95 |
| 2 | 345 | 66.67 | 100.00 | 0.67 | +33.33 |
| 3 | 42 | 59.17 | 99.95 | 0.67 | +40.79 |
| 3 | 345 | 65.00 | 100.00 | 0.83 | +35.00 |
| 4 | 42 | 66.67 | 99.95 | 1.83 | +33.29 |
| 4 | 345 | 65.83 | 100.00 | 0.67 | +34.17 |
| All 12 | | 64.51 | 99.98 | 1.03 | $+35.47 \pm 2.04$ |

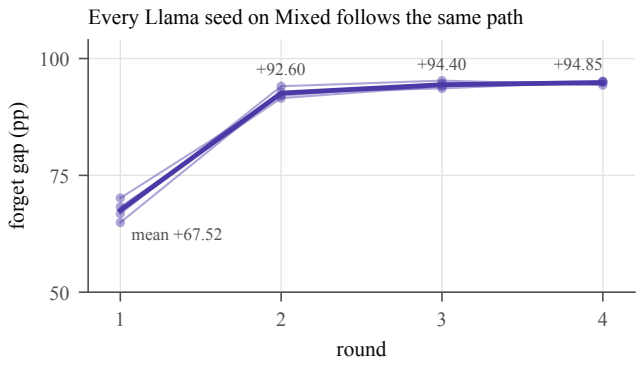



**Figure 13. On Mixed, every Llama-3.2-1B seed follows the same path, and most of the gap opens by the second round.** Forget gap after each round for the four seeds (thin) and their mean (thick, labelled).

### D.1.3 Robustness

On Inserted, the removal holds up under sampling, paraphrase, small relearning attacks and quantization, and gives way under large relearning budgets and to a membership test. Figure 14 and Table 19 compare the base model, the tuned RMU model of Appendix C.2, reset-and-repair run 1 and the retraining baseline on every probe.

**Table 18. Values behind Figure 13: every Llama-3.2-1B seed reaches nearly the same gap.** Left: Inserted after four rounds. Right: Mixed forget gap after each round; held-out accuracy after four rounds is 23.2 on average.

| | Inserted, 4 rounds | | | Mixed, gap after round | | | |
|---|---|---|---|---|---|---|---|
| **Seed** | **Forget** | **Retain** | **Gap** | **1** | **2** | **3** | **4** |
| 42 | 3.33 | 97.92 | +94.58 | 64.91 | 92.78 | 93.61 | 95.09 |
| 345 | 2.92 | 98.38 | +95.46 | 70.14 | 92.04 | 94.63 | 95.14 |
| 456 | 5.42 | 99.54 | +94.12 | 66.81 | 94.07 | 95.28 | 94.31 |
| 567 | 5.42 | 99.54 | +94.12 | 68.24 | 91.53 | 94.07 | 94.86 |
| Mean | 4.27 | 98.84 | +94.57 | 67.52 | 92.60 | 94.40 | 94.85 |

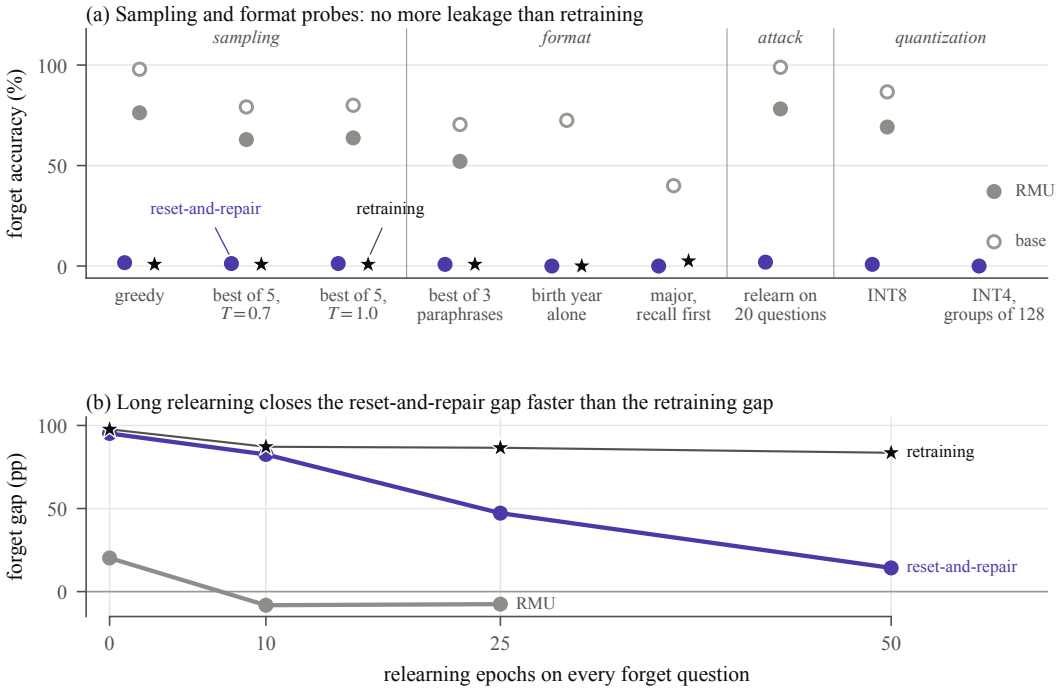



**Figure 14. No sampling or format probe recovers more forget answers from reset-and-repair than from retraining; a small relearning attack and quantization leave forget accuracy below 2%, but long relearning restores the answers faster than from retraining.** (a) Forget accuracy under every probe of Tables 19 and 22, grouped into sampling, question format, a small relearning attack (20 forget questions, one epoch, mean over 3 attack seeds) and quantization. Reset-and-repair (run 1) and retraining are drawn side by side; a missing marker means the checkpoint was not run on that probe. (b) Forget gap after relearning on all 240 forget questions at learning rate $10^{-5}$; epoch 0 is each checkpoint's own gap (RMU +20.23, run 1 +95.19, retraining +97.69). RMU was not run for 50 epochs.

Table 20 gives the small relearning attacks in full, including a model with the biographies inserted by continued pretraining (one round of reset-and-repair), and Figure 15 shows how each model responds to a higher attack learning rate.

Sampling and paraphrase recover no more from reset-and-repair than from the retraining baseline, while RMU leaks under both. The four Llama Inserted runs agree: sampling and paraphrase recover no more forget answers than greedy decoding (last column of Table 19). Relearning on every forget question for 10 epochs costs reset-and-repair 13 points of gap against 11 for retraining, but at 25 and 50 epochs its gap falls much faster than retraining's, and at higher attack learning rates it turns negative. Quantization never brings forget answers back; it lowers the gap only by damaging retain accuracy. The membership AUC (0.848, against 0.435 for retraining) and the relearning loss (0.583 against 1.091) show that the forget data leaves a trace in the weights that direct questions do not reveal.

**Table 19. Values behind Figure 14: reset-and-repair leaks no more than retraining under sampling and paraphrase, but relearns faster and keeps a membership signal.** Inserted; one checkpoint per column (the RMU checkpoint's own gap is +20.23; run 1's is +95.19), except Llama R&R, which pools the four Llama-3.2-1B Inserted runs. CPT is the one-round reset-and-repair checkpoint of the model with biographies inserted by continued pretraining (gap +98.56, retain 99.40 in BF16). Relearning fine-tunes the full model on forget questions at learning rate $10^{-5}$ unless noted; $(k, E)$ is the number of questions and epochs, and $k$=240 is every forget question, so reruns with other seeds are identical. The $(20, 1)$ row gives mean ± SD over 3 attack seeds. Quantization is symmetric round-to-nearest.

| | | | Reset-and-repair | | | |
|---|---|---|---|---|---|---|
| **Probe** | **Base** | **RMU** | **Run 1** | **CPT** | **Retraining** | **Llama R&R** |
| *Forget accuracy (lower means less leakage)* | | | | | | |
| Greedy | 97.92 | 76.25 | 1.67 | 0.83 | 0.83 | 4.17 |
| Best of 5 samples, $T$=0.7 / $T$=1.0 | 79.17 / 80.00 | 62.92 / 63.75 | 1.25 / 1.25 | – | 0.83 / 0.83 | 2.40 / 2.60 |
| Best of 3 paraphrases | 70.42 | 52.08 | 0.83 | – | 0.83 | 3.23 |
| After relearning, $(k, E)$=(20, 1) | 98.89 ± 0.20 | 78.19 ± 1.37 | 1.94 ± 0.20 | 1.53 ± 0.39 | – | – |
| *Forget gap after relearning on all forget questions* | | | | | | |
| $E$=10 | −3.94 | −8.19 | +82.55 | +66.94 | +87.18 | – |
| $E$=25 | – | −7.55 | +47.22 | −4.17 | +86.57 | – |
| $E$=50 | – | – | +14.26 | – | +83.56 | – |
| $E$=25, lr $3\times10^{-5}$ | – | – | −6.53 | – | – | – |
| $E$=25, lr $5\times10^{-5}$ | – | – | −26.81 | – | – | – |
| *Forget / retain accuracy after quantization* | | | | | | |
| INT8 | 86.67 / 85.51 | 69.17 / 79.31 | 0.83 / 79.40 | 0.83 / 98.52 | – | – |
| INT4, groups of 128 | 12.08 / 17.92 | 37.08 / 17.82 | 0.00 / 17.08 | 0.83 / 71.30 | – | – |
| *Traces below the answer* | | | | | | |
| Probability on the correct forget answer | 0.994 | – | 0.522 | – | 0.495 | – |
| Loss-based membership AUC on forget data | 0.975 | – | 0.848 | – | 0.435 | – |
| Loss after 100 relearning steps | – | – | 0.583 | – | 1.091 | – |

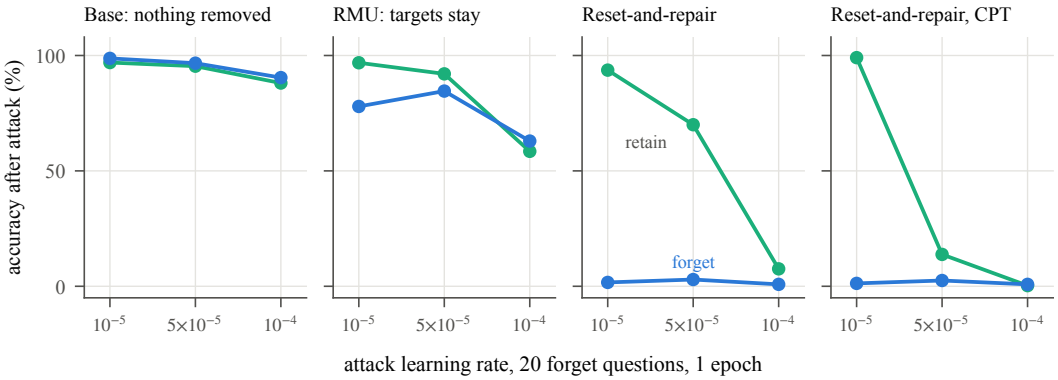



**Figure 15. Raising the attack's learning rate does not bring forget answers back after reset-and-repair; it destroys retain accuracy instead.** Forget and retain accuracy after full-model fine-tuning on 20 forget questions for one epoch at three learning rates, one seed each (right block of Table 20).

**Every Inserted model regains forget answers under long relearning, and how fast depends on the checkpoint more than on the number of people** (Table 21). Each attack fine-tunes the full model on all forget questions at learning rate $10^{-5}$, so it uses 240, 480 or 960 questions at 500, 1,000 or 2,000 people. After 10 epochs every reset-and-repair checkpoint keeps a gap of +53 or more, except the strong schedule's seed 234, whose retain accuracy was already 35.56. After 25 epochs the one-round continued-pretraining model, the bioR model and the 3- and 4-round 1,000-person

**Table 20. Values behind Figure 15: small relearning attacks do not bring forget answers back.** Full-model fine-tuning on $k$ forget questions for $E$ epochs. Left: learning rate $10^{-5}$; the $(20, 1)$ rows give mean $\pm$ SD over 3 attack seeds, the others one seed. Right: $(k, E)$=$(20, 1)$, one seed, at three learning rates; forget / retain accuracy after the attack.

| | | **Learning rate** $10^{-5}$ | | | **Forget / retain after** $(20, 1)$ | | |
|---|---|---|---|---|---|---|---|
| **Model** | $(k, E)$ | **Forget** | **Retain** | **Gap** | $10^{-5}$ | $5\times10^{-5}$ | $10^{-4}$ |
| Base | (20,1) | $98.89 \pm 0.20$ | 96.77 | $-2.11 \pm 0.29$ | 98.75 / 96.94 | 96.67 / 95.37 | 90.42 / 88.06 |
| | (20,3) | 98.33 | 97.08 | $-1.25$ | | | |
| | (60,1) | 98.33 | 96.76 | $-1.57$ | | | |
| RMU | (20,1) | $78.19 \pm 1.37$ | 96.79 | $+18.60 \pm 1.38$ | 77.92 / 96.85 | 84.58 / 92.04 | 62.92 / 58.47 |
| | (20,3) | 79.17 | 97.04 | $+17.87$ | | | |
| | (60,1) | 78.75 | 96.76 | $+18.01$ | | | |
| Reset-and-repair | (20,1) | $1.94 \pm 0.20$ | 93.72 | $+91.77 \pm 0.48$ | 1.67 / 93.66 | 2.92 / 70.00 | 0.83 / 7.55 |
| | (20,3) | 2.50 | 94.17 | $+91.67$ | | | |
| | (60,1) | 2.08 | 94.12 | $+92.04$ | | | |
| Reset-and-repair, CPT | (20,1) | $1.53 \pm 0.39$ | 99.20 | $+97.67 \pm 0.50$ | 1.25 / 99.07 | 2.50 / 13.80 | 0.83 / 0.28 |
| | (20,3) | 1.67 | 98.56 | $+96.90$ | | | |
| | (60,1) | 1.25 | 98.52 | $+97.27$ | | | |

checkpoints fall to $+3$ or below, while the 1,000-person strong checkpoint used to calibrate the schedule keeps $+62.80$ and still has $+52.59$ after 50 epochs. That checkpoint's retain accuracy falls from 96.27 to 72.22 in the first 10 epochs and then stays between 72 and 75, while its forget accuracy rises only from 2.71 to 20.42. The strong schedule's other three seeds hold less: after 50 epochs their gaps are $+14.42$, $-3.47$ and $-25.07$. At 2,000 people the retraining baseline's own retain accuracy falls to 39.59 after 10 epochs.

**Table 21. Under long relearning, every reset-and-repair checkpoint loses more of its gap than the 500-person retraining baseline.** Forget gap after fine-tuning the full model on all forget questions for $E$ epochs at learning rate $10^{-5}$; one attack seed. Gaps before the attack are in Tables 13, 15 and 16; the CPT checkpoint's is $+98.56$. –: not run.

| **Model** | **Checkpoint** | $E$=10 | $E$=25 | $E$=50 |
|---|---|---|---|---|
| Qwen3, 500 | run 1, 4 rounds | $+82.55$ | $+47.22$ | $+14.26$ |
| | retraining baseline | $+87.18$ | $+86.57$ | $+83.56$ |
| Qwen3, CPT | 1 round | $+66.94$ | $-4.17$ | – |
| Qwen3, bioR | 3 rounds, run 1 | $+65.88$ | $+2.92$ | – |
| Qwen3, 1,000 | 3 rounds, seed 345 | $+56.41$ | $-8.91$ | – |
| | 4 rounds, seed 345 | – | $-8.98$ | – |
| | strong, calibration seed | $+66.39$ | $+62.80$ | $+52.59$ |
| | strong, seed 234 | $+16.64$ | $+5.32$ | $-25.07$ |
| | strong, seed 345 | $+70.58$ | $+48.77$ | $+14.42$ |
| | strong, seed 456 | $+53.15$ | $+33.08$ | $-3.47$ |
| Qwen3, 2,000 | strong, 3 rounds | $+57.26$ | $+18.74$ | $-8.56$ |
| | strong, 4 rounds | $+57.23$ | $+23.67$ | $-5.29$ |
| | retraining baseline | $+38.03$ | $+41.66$ | – |

**Quantization leaves the continued-pretraining checkpoint's targets removed as well** (CPT column of Table 19). Its forget accuracy stays at 0.83 in BF16 and both quantized formats, so the gap falls only as retain accuracy falls, from $+98.56$ to $+97.69$ (INT8) and $+70.46$ (INT4).

**Reset-and-repair leaks no forget answers when the question asks for part of a fact or asks the model to recall before answering** (Table 22). We asked for the birth year alone, and asked for the major with the instruction to first recall it and then state it. On both, run 1 and the continued-pretraining checkpoint answer at most 2.5% of forget people, as retraining does, against 72.5% and 40.0% before unlearning. On these formats accuracy on the other people also falls, after retraining as well as after reset-and-repair. Two further formats, one question for two attributes and naming the person who works at a given employer, score 0 on every checkpoint, the base included, so they measure nothing.

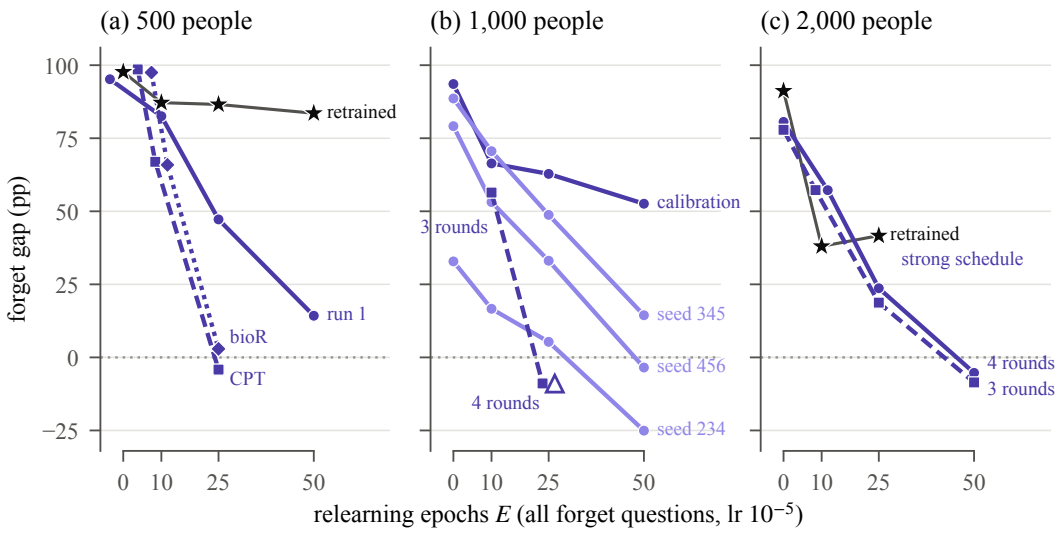



**Figure 16. Under long relearning every reset-and-repair checkpoint regains forget answers, and the checkpoint matters more than the number of people.** Forget gap after fine-tuning the full model on all forget questions for $E$ epochs at learning rate $10^{-5}$; one attack seed. Stars: the Inserted retraining baseline of the same size (the CPT baseline was not relearned). Markers that would coincide are offset slightly along $E$. At 1,000 people the dark solid line is the strong schedule's calibration seed and the light lines are its other three seeds; the 3- and 4-round checkpoints start at $E$=10 and $E$=25 because their pre-attack seed is not listed. $E$=0 values come from Tables 13, 15 and 16; the rest from Table 21.

**Table 22. Values behind the format probes of Figure 14(a): on two rephrased question formats, reset-and-repair recovers no more forget answers than retraining.** Inserted; accuracy on the 40 forget people and on the other 460 people (retain and held-out together; held-out people are not answerable on Inserted).

| | Birth year | | Major, recall first | |
|---|---|---|---|---|
| **Checkpoint** | **Forget** | **Others** | **Forget** | **Others** |
| Base | 72.5 | 48.3 | 40.0 | 42.0 |
| Reset-and-repair, run 1 | 0.0 | 9.8 | 0.0 | 9.8 |
| Reset-and-repair, CPT, 1 round | 0.0 | 28.9 | 2.5 | 10.9 |
| Retraining baseline | 0.0 | 4.1 | 2.5 | 16.3 |

### D.2 Mixed: all 17 configurations

On every Mixed model, reset-and-repair restores the retain set, and the untargeted holdout loses between 44% and 104% of what the forget set loses. Table 23 gives all 17 models with accuracy before and after unlearning; the step grid among them was built to study Fisher-tail mass (Appendix D.3).

**Three more Llama-3.2-1B training lengths show the same pattern.** We trained Llama-3.2-1B Mixed models for 500, 1,000 and 1,500 steps with the recipe of the 2,500-step model and ran reset-and-repair with its schedule, one seed each (Table 24). All three pass both standard checks, yet the holdout loses 84% to 95% as much as the targets. Earlier runs of the same schedule with another seed end at the same held-out accuracy within 2.2 points.

Retain accuracy recovers to at least 94 on every model except the 500-person low split, where two of four seeds lose retain accuracy (64.1 and 70.8) and the other two recover to 99.9. The smallest holdout drop on any model is 14.4. At 1,000 people the margin stays between $-2.7$ and $+5.4$, so the method removes the targets and the holdout almost equally. At 500 people the method favors the forget people by 10.9 to 27.9 points of margin and still removes a large share of the holdout. On Llama the margin is $+20.5$. The 500-person 500-step model has the largest forget drop of any reset-and-repair model on Mixed, 97.5 points, and its holdout accuracy falls from 84 to 2.5. Matched retraining removes the targets while its holdout loses 5.8 to 15.0 points, 6% to 15% of its forget drop, against at least 44% for every reset-and-repair model (Figure 17).

**Table 23. Reset-and-repair on Mixed removes the targets and much of the untargeted holdout.** Accuracy on forget / retain / held-out people before and after unlearning (final round), the forget and holdout drops, and the selectivity margin (forget drop minus holdout drop). Qwen3 rows use 4 seeds (42, 345, 456, 567) after four rounds, except the 1,000-person 4,000-step model (three rounds). "High"/"low" forget splits pick the forget people with the highest or lowest per-person Fisher tail on the model of the row above; they share its baseline. $M$ is the Fisher-tail mass of the model before unlearning (for split rows, of the forget split). The last three rows are matched retraining on the Mixed pipeline (Appendix E.1), one run per data order, each against the full-corpus model of the same order; Figure 17 shows the seed SD of each holdout drop.

| Model | Steps | $M$ (%) | Before F / R / H | After F / R / H | Forget drop | Holdout drop | Margin |
|---|---|---|---|---|---|---|---|
| Qwen3, 500 | 500 | 19.36 | 99.2 / 98.6 / 84.2 | 1.7 / 99.8 / 2.5 | 97.5 | 81.7 | +15.8 |
| | 750 | 11.91 | 99.6 / 99.4 / 95.5 | 3.5 / 99.7 / 10.3 | 96.0 | 85.2 | +10.9 |
| | 1,000 | 8.07 | 100 / 100 / 97.8 | 16.4 / 100 / 25.7 | 83.6 | 72.1 | +11.5 |
| | 1,500 | 6.19 | 100 / 100 / 97.0 | 36.6 / 100 / 46.5 | 63.4 | 50.5 | +12.9 |
| | 2,500 | 6.31 | 100 / 100 / 99.0 | 46.2 / 94.0 / 59.9 | 53.8 | 39.1 | +14.6 |
| | high split | 3.48 | (as above) | 69.6 / 99.9 / 81.1 | 30.4 | 17.9 | +12.5 |
| | low split | 4.59 | (as above) | 29.6 / 83.7 / 52.3 | 70.4 | 46.7 | +23.8 |
| | 2,500 (rerun) | 4.90 | 100 / 100 / 98.8 | 50.0 / 99.9 / 76.7 | 50.0 | 22.1 | +27.9 |
| Qwen3, 1,000 | 500 | 15.98 | 99.8 / 99.4 / 96.0 | 2.8 / 99.9 / 4.3 | 97.0 | 91.7 | +5.4 |
| | 1,000 | 9.32 | 99.8 / 99.5 / 98.6 | 25.7 / 99.9 / 25.5 | 74.1 | 73.1 | +1.0 |
| | 1,500 | 8.58 | 100 / 99.9 / 99.2 | 39.0 / 99.9 / 35.5 | 61.0 | 63.7 | −2.7 |
| | 2,000 | 10.01 | 100 / 99.9 / 99.4 | 34.9 / 99.8 / 38.8 | 65.1 | 60.7 | +4.4 |
| | 2,500 | 6.01 | 100 / 99.7 / 99.1 | 59.8 / 99.8 / 62.9 | 40.2 | 36.1 | +4.1 |
| | 4,000 | 3.81 | 100 / 100 / 100 | 81.4 / 99.8 / 85.6 | 18.6 | 14.4 | +4.2 |
| | high split | 3.82 | (as above) | 76.0 / 99.9 / 79.9 | 24.0 | 20.1 | +3.9 |
| | low split | 3.45 | (as above) | 82.2 / 99.9 / 85.4 | 17.8 | 14.6 | +3.1 |
| Llama-3.2-1B, 500 | 2,500 | 2.87 | 100 / 100 / 100 | 2.7 / 97.6 / 23.2 | 97.3 | 76.8 | +20.5 |
| Retraining, 500 | 2,500 (order 42) | – | 100 / 100 / 99.0 | 1.2 / 100 / 93.2 | 98.8 | 5.8 | +92.9 |
| | 1,000 (order 42) | – | 100 / 100 / 97.8 | 0.8 / 99.5 / 82.8 | 99.2 | 15.0 | +84.2 |
| | 2,500 (order 1) | – | 100 / 100 / 98.0 | 0.8 / 100 / 89.3 | 99.2 | 8.7 | +90.5 |
| | 1,000 (order 1) | – | 100 / 99.9 / 94.7 | 1.7 / 99.8 / 87.8 | 98.3 | 6.8 | +91.5 |

**Table 24. At every Llama-3.2-1B training length, reset-and-repair passes both standard checks and removes most of the holdout.** Mixed, 500 people, one seed (42), four rounds with thresholds 0.55, 0.85, 0.95, 0.975. Accuracy (%) as forget / retain / held-out. "Earlier": held-out accuracy after an earlier run with another seed.

| Steps | Before | After | Gap | Forget drop | Holdout drop | Ratio | Earlier |
|---|---|---|---|---|---|---|---|
| 500 | 100.00 / 100.00 / 99.17 | 1.67 / 98.01 / 5.50 | +96.34 | 98.3 | 93.7 | 95% | 6.00 |
| 1,000 | 100.00 / 100.00 / 100.00 | 5.00 / 98.43 / 13.00 | +93.43 | 95.0 | 87.0 | 92% | 14.00 |
| 1,500 | 100.00 / 99.95 / 99.83 | 4.17 / 98.70 / 19.00 | +94.54 | 95.8 | 80.8 | 84% | 16.83 |

The table reports the four planned seeds; additional runs change no margin by more than 4.5 points. Two reruns of the 500-person low split, with seeds 345 and 456 and intermediate checkpoints saved, both recover retain accuracy (99.95 and 99.77, gaps +66.62 and +66.02). A later seed 678 on the 1,000-person splits gives gaps of +23.68 (high) and +15.95 (low).

On Llama, the Mixed model's forget gap reaches 90% of its final value after two rounds (Table 18), while held-out accuracy ends at 23.2. Its Fisher-tail mass of 2.87 (mean over four Fisher estimates, ±0.06) is smaller than any Qwen3 model's, yet its forget drop is among the largest, so the Fisher-tail relation of Appendix D.3 holds only within the Qwen3 grid.

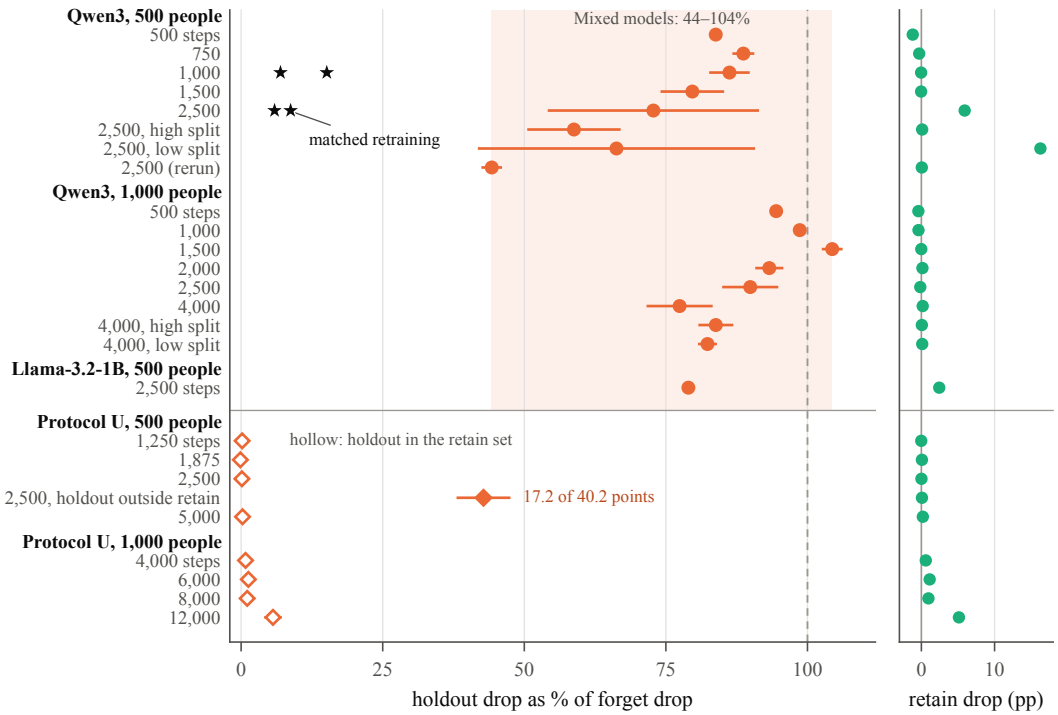



**Figure 17. On every Mixed model the holdout loses 44% to 104% of what the targets lose; matched retraining loses 6% to 15%, and a holdout inside the retain set at most 6%.** Left: holdout drop as a percentage of forget drop, one row per model (mean over 4 seeds; whiskers are the seed SD of the holdout drop divided by the forget drop). The shaded band spans the 17 Mixed models. Stars are matched retraining against the full-corpus model of the same data order, two data orders at each of 2,500 and 1,000 steps. Below the rule, hollow diamonds are the eight Protocol U models of Table 28, whose held-out people are in the retain set; the filled diamond is the 500-person 2,500-step Protocol U model with the same held-out people kept out of the retain set: 17.2 of 40.2 points. Right: retain drop in points after the final round.

### D.3 Fisher-tail mass predicts forget and holdout loss

Across eight Qwen3 Mixed models, a larger Fisher-tail mass goes with more forgetting and more holdout loss, in nearly the same rank order (Section 4; Figure 18, Table 25). Because $M$ is also the fraction of parameters the first round resets, a larger $M$ means a larger first reset, which removes more from both groups. $M$ therefore predicts how much the method removes; it does not pick out models where removal is selective.

**Table 25. Values behind Figure 18: Fisher-tail mass ranks forget gap and holdout drop together.** Eight Qwen3 Mixed models, 4 seeds each (the 1,000-person 4,000-step model after three rounds). The linear fit gap $= 425.4\,M + 26.7$ ($M$ as a fraction) is fit on all eight; "LOO" refits without the row. Residuals are observed minus predicted. The holdout-drop fit is $483.5\,(M \cdot H_{\text{before}}/100) + 16.5$.

| People | Steps | $M$ (%) | Gap | Fit | Residual | LOO residual | Holdout drop | Residual |
|---|---|---|---|---|---|---|---|---|
| 500 | 500 | 19.36 | +98.12 | +109.01 | −10.89 | −27.39 | 81.71 | −13.56 |
| 500 | 1,000 | 8.07 | +83.63 | +61.02 | +22.61 | +26.25 | 72.12 | +17.43 |
| 500 | 1,500 | 6.19 | +63.43 | +52.99 | +10.44 | +12.86 | 50.54 | +5.03 |
| 500 | 2,500 | 6.31 | +47.79 | +53.51 | −5.72 | −7.02 | 39.12 | −7.57 |
| 1,000 | 500 | 15.98 | +97.09 | +94.65 | +2.45 | +3.63 | 91.67 | +1.00 |
| 1,000 | 1,000 | 9.32 | +74.23 | +66.33 | +7.90 | +9.04 | 73.10 | +12.17 |
| 1,000 | 1,500 | 8.58 | +60.91 | +63.18 | −2.27 | −2.61 | 63.71 | +6.03 |
| 1,000 | 4,000 | 3.81 | +18.39 | +42.90 | −24.51 | −35.15 | 14.40 | −20.54 |

Both fits rank the eight models well, and the ranking holds under resampling of the Fisher estimate, but the slope is poorly pinned down (Table 26): the bootstrap interval for the gap slope is wide, and its ends move by up to about 60 across bootstrap seeds, so we rely on the rank order.

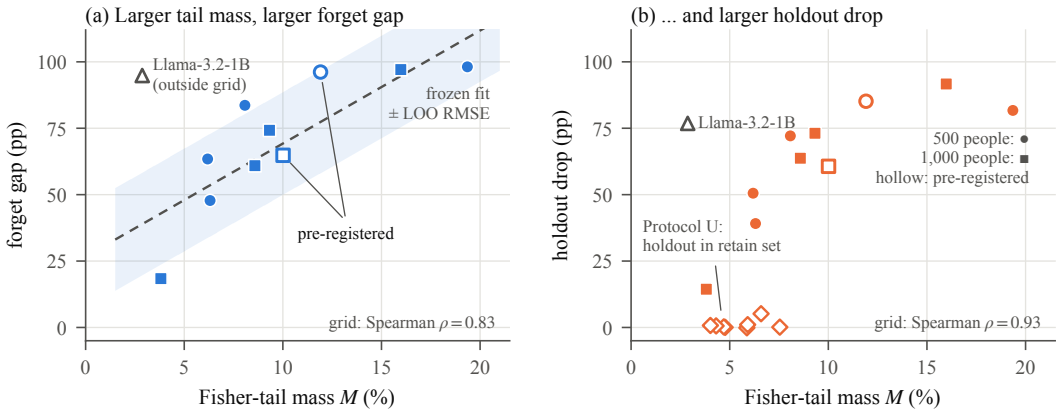



**Figure 18. Fisher-tail mass ranks the forget gap and the holdout drop together, and the frozen fit predicts both pre-registered models within its band.** Filled: the eight Qwen3 Mixed models of the step grid (circles 500 people, squares 1,000), 4 seeds each; hollow: the two pre-registered models (Table 27). *(a)* Forget gap, with the linear fit on the eight grid models (dashed) and a band of ± its leave-one-out RMSE, 19.3. *(b)* Holdout drop. Spearman $\rho$ is over the eight grid models. Triangle: Llama-3.2-1B on Mixed, outside the grid, far above the fit. In *(b)*, hollow diamonds are the eight Protocol U models (Table 28), whose holdout is in the retain set and loses at most 5.2 points at every $M$. Values in Tables 25, 27 and 28.

**Table 26. Fisher-tail mass ranks both outcomes, and the gap ranking survives resampling of $M$; the slope is loosely determined, so we rely on the rank order.** Fit statistics over the eight grid models of Table 25. Spearman $\rho$ uses an exact permutation $p$. The slope interval is a bootstrap over the eight models (1,000 resamples). The last column redraws one of four Fisher estimates of $M$ per model at random (5,000 draws) and gives the mean $\rho$ and its 95% interval. –: not reported.

| Fit | Spearman $\rho$ ($p$) | $R^2$ | In-sample RMSE | LOO RMSE | Slope, 95% CI | $\rho$, resampled $M$ |
|---|---|---|---|---|---|---|
| Forget gap | $\approx 0.83$ (0.015) | 0.71 | 13.4 | 19.3 | [243, 1,061] | 0.84 [0.74, 0.88] |
| Holdout drop | $\approx 0.93$ (0.002) | 0.73 | – | 17.5 | – | – |

**The Fisher-tail mass barely changes when the Fisher mini-batches are resampled, and the rank order survives.** We recomputed $M$ on every grid model with three further samples of Fisher mini-batches, four estimates per model in all. The population SD of $M$ over the four estimates is at most 1.16 points (1,000 people, 500 steps) and at most 0.79 on every other model. Drawing one of the four estimates per model at random, the Spearman $\rho$ between $M$ and the forget gap stays close to its value on the original estimates (last column of Table 26). This check uses two other models described in Appendix B.2: the 500-person 2,500-step rerun in place of the grid model of the same configuration, and the separate 1,000-person 2,500-step model in place of the 4,000-step one.

The frozen fit predicts two further models within its error band (Table 27). Before training two new Mixed models (500 people at 750 steps; 1,000 people at 2,000 steps), we froze the gap fit and the rule that both residuals must fall within the leave-one-out RMSE. The fit was frozen on 2026-04-30 at 13:36, and both checkpoints were written later that day. The holdout-drop coefficients were revised once after the freeze and before training. Both new models also lose most of their holdout.

The relation does not carry over to models outside the grid (lower block of Table 27). Llama-3.2-1B has $M$ below every Qwen3 model on both Inserted and Mixed, yet its forget gaps are among the highest. On nanochat-d12 the fit's prediction matches the observed gap, a single agreeing point. Selecting forget people by their own Fisher tail (the high and low splits in Table 23) does not order the gaps the same way at both sizes: at 500 people the high-tail split has the smaller gap, and at 1,000 people the larger. Within each size, gap and holdout drop still move together.

### D.4 Protocol U: the holdout inside the retain set

When the held-out people's questions are trained and those people are in the retain set, their accuracy survives reset-and-repair: repair restores what it trains on (Section 4). Table 28 gives all eight Protocol U models; Figure 2(b) plots them (hollow diamonds) against the Mixed models.

**Table 27. The frozen fit predicts two new Mixed models within its error band; outside the Qwen3 grid it does not carry over.** Upper block: pre-registered Mixed models, 4 seeds each. The band is the leave-one-out RMSE, 19.3. Lower block: models outside the grid, from Tables 15, 18 and 23 (Llama 4 seeds, nanochat-d12 12 runs; Llama Mixed gap after four rounds); –: no prediction made, or no answerable holdout.

| Model | $M$ (%) | Predicted gap | Observed gap | Residual | Holdout drop |
|---|---|---|---|---|---|
| 500 people, 750 steps | 11.91 | +77.4 | $+96.16 \pm 0.45$ | +18.8 | $85.17 \pm 1.63$ |
| 1,000 people, 2,000 steps | 10.01 | +69.3 | $+64.81 \pm 1.57$ | −4.5 | $60.67 \pm 1.46$ |
| *Outside the Qwen3 grid* | | | | | |
| Llama-3.2-1B, Inserted | 1.34 | – | $+94.57 \pm 0.55$ | – | – |
| Llama-3.2-1B, Mixed | 2.87 | – | +94.85 | – | 76.8 |
| nanochat-d12, Inserted | 2.20 | +36.1 | $+35.47 \pm 2.04$ | – | – |

**Table 28. In Protocol U, held-out people whose questions are trained and repaired keep their accuracy.** Four rounds, 4 seeds (42, 345, 456, 567). Held-out people are part of the retain set here (460 people at 500, 920 at 1,000), so the holdout drop does not measure selectivity. Gap SD is the population SD over seeds. Values behind the Protocol U rows of Figure 17 and the diamonds of Figure 18(b).

| People | Steps | $M$ (%) | Before F / R / H | After F / R / H | Gap | Forget drop | Holdout drop |
|---|---|---|---|---|---|---|---|
| 500 | 1,250 | 7.54 | 100 / 99.9 / 100 | 30.2 / 100 / 99.9 | $+69.74 \pm 6.10$ | 69.8 | 0.1 |
| 500 | 1,875 | 5.87 | 100 / 99.9 / 99.8 | 5.2 / 99.9 / 100 | $+94.65 \pm 1.99$ | 94.8 | −0.1 |
| 500 | 2,500 | 4.77 | 100 / 100 / 100 | 70.2 / 100 / 100 | $+29.75 \pm 2.80$ | 29.8 | 0.0 |
| 500 | 5,000 | 4.70 | 100 / 100 / 100 | 15.8 / 99.8 / 99.8 | $+83.97 \pm 1.24$ | 84.2 | 0.2 |
| 1,000 | 4,000 | 4.31 | 100 / 100 / 99.9 | 25.4 / 99.4 / 99.3 | $+73.96 \pm 0.68$ | 74.6 | 0.6 |
| 1,000 | 6,000 | 5.91 | 100 / 100 / 100 | 18.7 / 98.9 / 99.0 | $+80.20 \pm 1.28$ | 81.4 | 1.0 |
| 1,000 | 8,000 | 4.02 | 100 / 100 / 100 | 30.4 / 99.0 / 99.3 | $+68.62 \pm 1.43$ | 69.6 | 0.8 |
| 1,000 | 12,000 | 6.59 | 100 / 100 / 100 | 8.2 / 94.9 / 94.8 | $+86.69 \pm 1.40$ | 91.8 | 5.2 |

The same method on the same architecture costs held-out people at most 5.2 points here and at least 14.4 on every Mixed model. Protocol U changes two things relative to Mixed: the questions of every held-out person are trained (in Mixed, 83 of 100 people's are), and those people are repaired. The retain-set control (Appendix E.3) separates the two. Protocol U's forget gaps show no consistent trend with training length.

### D.5 Two-stage, with a mismatched reset reference

These runs reset the Two-stage model to the wrong reference weights, so they do not test the method: the model was trained from random initialization, but every reset writes in weights from pretrained Qwen3-0.6B-Base. Every run ends in collapse or in too little forgetting. In every run, the rounds after the first reset no parameters. The seed-42 run at first threshold 0.55 also estimated Fisher and repaired on biographies where the other runs used questions, and no schedule has more than two seeds. Before unlearning the model scores 99.91 / 100.00 / 99.00; Table 29 lists every run.

**Table 29. Values behind Figure 19: with a mismatched reset reference, no Two-stage run meets the success criterion.** Four rounds; schedules in Table 9, named by their first threshold. One run per row.

| First threshold | Seed | Forget | Retain | Held-out | Gap | Forget drop | Holdout drop |
|---|---|---|---|---|---|---|---|
| 0.55 | 42 | 0.00 | 0.00 | 0.00 | – | 99.9 | 99.0 |
| 0.55 | 345 | 3.56 | 22.29 | 3.50 | +18.72 | 96.4 | 95.5 |
| 0.60 | 42 | 67.48 | 83.68 | 77.83 | +16.20 | 32.4 | 21.2 |
| 0.60 | 345 | 68.03 | 84.68 | 77.17 | +16.65 | 31.9 | 21.8 |
| 0.65 | 42 | 91.59 | 96.71 | 92.67 | +5.12 | 8.3 | 6.3 |
| 0.75 | 42 | 98.33 | 98.68 | 97.67 | +0.35 | 1.6 | 1.3 |

The first threshold moves the outcome from collapse (0.55) through joint loss of target, retain and holdout (0.60) to almost no forgetting (0.65 and above).

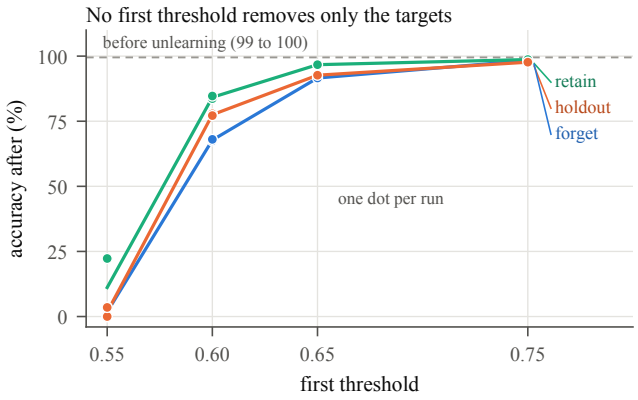



**Figure 19. With a mismatched reset reference, no first threshold removes only the targets: 0.55 collapses the model, 0.60 lowers all three groups, and 0.65 and above barely forget.** Accuracy after four rounds on Two-stage, one dot per run (two seeds at 0.55 and 0.60), line through the mean; dashed: accuracy before unlearning (99.0 to 100). Values in Table 29.

### D.6 QWEN3-1.7B

On Qwen3-1.7B the Fisher tail disappears, and no tested reset budget preserves utility. We inserted the biographies into Qwen3-0.6B-Base and Qwen3-1.7B-Base with the same continued-pretraining recipe. On 0.6B, 30.6% of parameters have $W > 0.55$; on 1.7B, 0.026%, with median $W$ 0.080. Resetting a fixed top fraction of parameters by $W$, in place of a threshold, collapses the model. At the five budgets tried, retain accuracy falls to 2.50, 3.33, 5.23, 7.31 or 5.79 while forget accuracy falls to at most 4.58. With four times the continued-pretraining data, retain accuracy stays at 0 to 11.71 at every budget tried. None of the tested budgets preserves utility on these 1.7B checkpoints.

### D.7 REAL CORPORA

Two small real-corpus checks agree with the synthetic results: gradient methods do not open a forget gap, and reset-and-repair moves forget metrics only as far as the Fisher tail allows. Neither check has an answerable untargeted holdout, so neither tests selectivity.

Table 30 summarizes both checks.

**MUSE-News (Shi et al., 2025b).** On the released Llama-2-7B target, all 31 gradient configurations that did not diverge (GA, GD and NPO over learning rates, steps, retain weights and $\beta$) leave retain knowledge-memory below forget knowledge-memory; the one diverged run destroys the model. A Fisher-ranked reset of the top 4.48% of parameters with 100 repair steps lowers both forget metrics by a few points and leaves retain knowledge-memory almost unchanged. Across four Fisher estimates (tail masses 4.48% to 19.92%), the Fisher mask beats a random mask of the same size on both forget metrics every time. On Qwen3-0.6B with the MUSE-News articles inserted by continued pretraining, one round lowers forget knowledge-memory and keeps retain, while tuned RMU, NPO and GA all leave retain below forget. This model memorizes little, so the absolute changes are small.

**The released Llama-2-7B target has memorized its forget articles.** It scores far above Llama-2-7B before MUSE-News training on both forget metrics, and higher on retain knowledge-memory (Table 30).

**On the Llama-2-7B target, the reset produces the forget reduction, the Fisher-ranked reset more than a random one, and repair alone barely moves either forget metric** (Table 31). For each of the four Fisher estimates we ran three arms with the same 100 repair steps at learning rate $3\times10^{-6}$: Fisher-ranked reset, random reset of the same size, and repair without reset. The Fisher arm lowers forget knowledge-memory by 2.48 to 5.94 and forget verbatim memory by 2.58 to 9.61; repair alone moves either by at most 0.99. No arm moves PrivLeak by more than 0.94. On the Qwen3-0.6B model, by contrast, one round lowers PrivLeak substantially, while tuned RMU, NPO and GA leave it almost where it was (Table 30).

**Table 30. On real corpora, no gradient method lifts retain knowledge memory above forget knowledge memory; reset-and-repair lowers the forget metrics on Llama-2-7B and Qwen3-0.6B and leaves OLMo-1B's unchanged.** KM: knowledge memory; VM: verbatim memory. MUSE-News on Llama-2-7B: gradient rows are the 31 configurations that did not diverge; the reset row resets the top 4.48% of parameters by Fisher share, then repairs for 100 steps at learning rate $10^{-5}$ (the three arms of Table 31 use $3\times10^{-6}$); the memorization rows use 25 prompts per metric and compare Llama-2-7B before MUSE-News training ("Before") with the released target ("After"). Qwen3-0.6B has the MUSE-News articles inserted by continued pretraining; reset-and-repair is one round, 2 seeds. OLMo-1B on LUME: nine runs (thresholds 0.70, 0.80, 0.90; 3 seeds) of Fisher-ranked or random reset with repair, 200 examples.

| Model | Method | Metric | Before | After |
|---|---|---|---|---|
| Llama-2-7B | MUSE-News training | Forget KM | 37.6 | 84.0 |
| | | Retain KM | 42.8 | 65.6 |
| | | Forget VM | 15.8 | 86.4 |
| | GA, GD, NPO | Retain − forget KM, median | | −15.69 |
| | | Retain − forget KM, maximum | | −12.88 |
| | Fisher reset + repair | Forget KM | 57.82 | 55.83 |
| | | Forget VM | 55.60 | 52.48 |
| | | Retain KM | 41.39 | 41.11 |
| Qwen3-0.6B | Reset-and-repair | Forget KM | 5.83 | 3.17 |
| | | Retain KM | 4.64 | 4.54 |
| | | PrivLeak | +82.48 | +65.30 |
| | Tuned RMU | Retain − forget KM | | −0.90 |
| | Tuned NPO | Retain − forget KM | | −1.95 |
| | Tuned GA | Retain − forget KM | | −1.70 |
| | Tuned RMU, NPO, GA | PrivLeak | +82.48 | within 0.5 |
| OLMo-1B | Fisher or random reset + repair | Forget regurgitation | 99.6 | 99.4–99.6 |
| | | Forget QA | 93.0 | 91.5–93.0 |

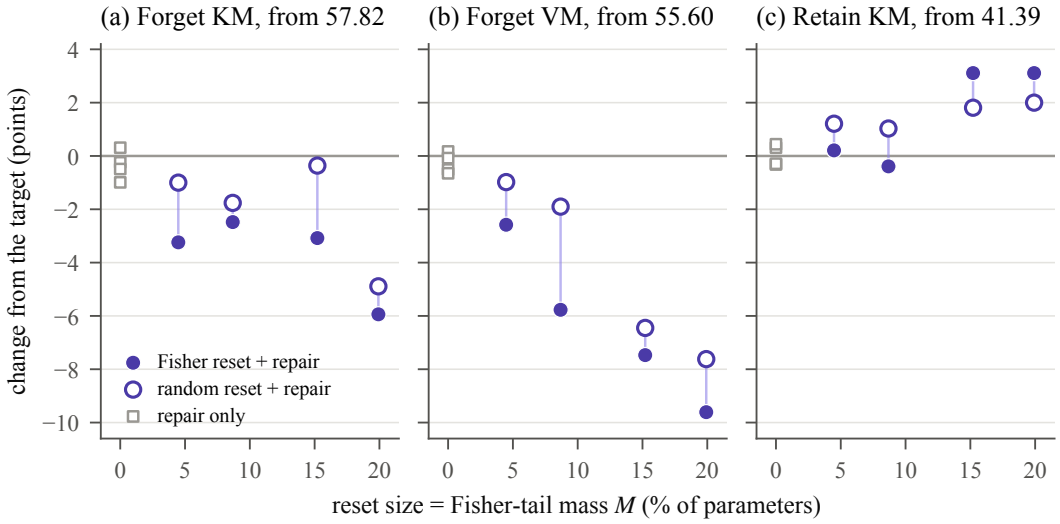



**Figure 20. On MUSE-News, at every Fisher estimate the Fisher-ranked reset lowers both forget metrics more than a random reset of the same size, and repair alone barely moves them.** Llama-2-7B target; change from the target's value (in each panel title), 100 examples per metric. Filled: Fisher-ranked reset then repair; hollow: random reset of the same size then repair; vertical ticks join the two arms of one estimate. Gray squares: repair only, drawn at reset size 0. PrivLeak moves by at most 0.94 in every arm and is not shown. Values in Table 31.

**OLMo-1B on LUME (Ramakrishna et al., 2025).** On OLMo-1B fine-tuned on real Wikipedia biographies, neither Fisher-ranked nor random reset with repair removes anything: over nine runs, both forget metrics stay at or just below their starting values (Table 30), and the two masks never differ by more than 1.0 point on 200 examples.

**Table 31. On MUSE-News, Fisher-ranked reset lowers both forget metrics more than a random reset of the same size for all four Fisher estimates, and repair alone barely moves them.** Llama-2-7B target; change from the target's 57.82 / 41.39 / 55.60 / +106.68 (forget knowledge-memory, retain knowledge-memory, forget verbatim memory, PrivLeak), 100 examples per metric. Each estimate uses its own sample of Fisher mini-batches; $M$ is its Fisher-tail mass and the size of both reset masks.

| Arm, estimate | $M$ (%) | Forget KM | Retain KM | Forget VM | PrivLeak |
|---|---|---|---|---|---|
| *Fisher-ranked reset, then repair* | | | | | |
| 0 | 4.48 | −3.24 | +0.21 | −2.58 | −0.31 |
| 1 | 8.68 | −2.48 | −0.39 | −5.77 | −0.36 |
| 2 | 19.92 | −5.94 | +3.11 | −9.61 | −0.94 |
| 3 | 15.21 | −3.08 | +3.11 | −7.47 | −0.61 |
| *Random reset of the same size, then repair* | | | | | |
| 0 | 4.48 | −1.00 | +1.21 | −0.98 | −0.17 |
| 1 | 8.68 | −1.76 | +1.03 | −1.90 | −0.31 |
| 2 | 19.92 | −4.89 | +2.00 | −7.62 | −0.84 |
| 3 | 15.21 | −0.36 | +1.81 | −6.45 | −0.44 |
| *Repair only* | | | | | |
| 0 | – | −0.99 | +0.31 | +0.17 | 0.00 |
| 1 | – | −0.24 | +0.44 | −0.51 | 0.00 |
| 2 | – | +0.31 | −0.32 | −0.65 | 0.00 |
| 3 | – | −0.49 | −0.29 | −0.08 | 0.00 |

# E Experiments Run for This Paper

Five further experiments test the reading of Section 4: a retraining baseline matched to the Mixed pipeline, a Protocol U model whose held-out people sit outside the method's retain set, a biography-completion probe, and a comparison of Fisher-ranked and random masks on Mixed, all on Qwen3-0.6B with 500 people; and a standard method on TOFU.

## E.1 Matched retraining

Retrained on the same pipeline without the forget people, the Mixed model cannot answer them and loses far less of its holdout than under reset-and-repair. We rebuilt the Mixed training corpus from its source data, confirmed that it matches the original corpus row for row, and removed the forget people's biographies and questions (440 of 4,900 rows). The retrained model uses the original recipe, initialization and step count. Table 32 compares it with the model trained on the full corpus.

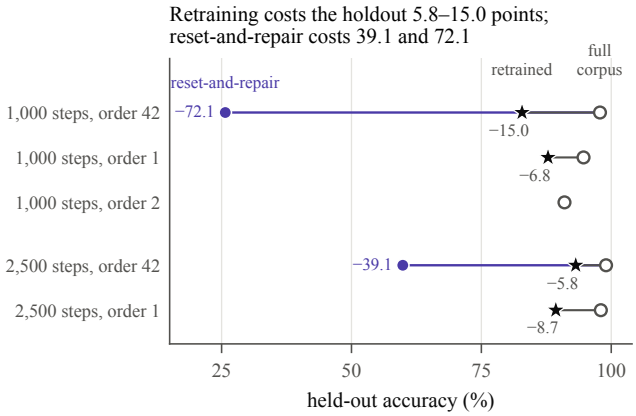



**Figure 21. Retrained without the forget people, the Mixed model loses 5.8 to 15.0 points of held-out accuracy across data orders; reset-and-repair loses 39.1 and 72.1.** Held-out accuracy of the full-corpus model (hollow), of the model retrained without the forget people (⋆), and, for data order 42, after reset-and-repair (filled; mean over 4 seeds, Table 23). Labels give the drop from the full-corpus model of the same data order; data order 2 has no retrained model. Values in Table 32.

**Table 32. Values behind Figure 21: retrained without the forget people, the Mixed model cannot answer them and keeps most of its holdout.** Accuracy on forget / retain / held-out people. The retrained model never sees the forget people; each retrained model is compared with the full-corpus model of the same data order.

| Steps | Data order | Full corpus | Retrained without forget people | Holdout drop |
|---|---|---|---|---|
| 1,000 | 42 | 100.0 / 99.95 / 97.83 | 0.83 / 99.49 / 82.83 | 15.00 |
| 1,000 | 1 | 100.0 / 99.91 / 94.67 | 1.67 / 99.81 / 87.83 | 6.83 |
| 1,000 | 2 | 100.0 / 99.81 / 91.00 | – | – |
| 2,500 | 42 | 100.0 / 99.95 / 99.00 | 1.25 / 100.0 / 93.17 | 5.83 |
| 2,500 | 1 | 100.0 / 99.95 / 98.00 | 0.83 / 100.0 / 89.33 | 8.67 |

Retraining lowers held-out accuracy, most likely because the forget people's data no longer help train the question-answering skill that answers held-out people. At 1,000 steps the data order alone moves the full-corpus model's held-out accuracy between 91.0 and 97.8, and the retraining cost between 6.8 and 15.0 points; reset-and-repair on the model of data order 42 costs 72.1. At 2,500 steps retraining costs 5.8 and 8.7 points for the two data orders, against 39.1 for reset-and-repair.

### E.2 Held-out people with and without trained questions

The Mixed training sets contain the questions of some held-out people: 500 of the 600 held-out questions at 500 people (83 people with all six, one with two) and 500 of 1,200 at 1,000 people (83 with all six, one with two); the other held-out people's questions are never trained. The other held-out questions appear only as facts in biographies. Neither kind is in the forget set, the retain set, the Fisher estimate or the repair data, so both are untargeted. We split the held-out questions of the model in Figure 1 by whether they were trained and scored each part on the model before unlearning, after each of the four reset-and-repair runs, and after matched retraining (Table 33). On this model, reset-and-repair removes the two kinds equally, and far more than retraining does.

**Table 33. Reset-and-repair removes held-out answers equally whether or not their questions were trained.** Mixed, 500 people, 1,000 steps. Held-out QA accuracy (%) on the held-out questions that appear in the training set and on those that do not. Reset-and-repair: mean ± population SD over the four runs of Figure 4b.

| **Held-out questions** | **Before** | **Reset-and-repair** | **Drop** | **Retraining drop** |
|---|---|---|---|---|
| Trained (500 questions, 84 people) | 97.60 | $27.44_{\pm 2.41}$ | 70.2 | 15.6 |
| Not trained (100 questions, 17 people) | 99.02 | $28.89_{\pm 2.44}$ | 70.1 | 10.8 |

### E.3 Retain-set control

Held-out people whose questions were trained still lose accuracy under reset-and-repair when they are outside the retain set (Table 34; filled diamonds in Figure 2(b)). On the Protocol U model (2,500 steps), we ran reset-and-repair with the retain set restricted to the 360 original retain people, so the held-out people enter neither the Fisher retain estimate nor the repair. Thresholds and rounds follow Appendix B; four seeds.

**Table 34. Outside the retain set, held-out people lose accuracy even though their questions were trained.** Protocol U model; accuracy after reset-and-repair. Before unlearning: 100.0 / 99.96 / 100.0.

| Seed | Forget | Retain (360) | Held-out |
|---|---|---|---|
| 42 | 61.67 | 100.0 | 82.50 |
| 345 | 52.50 | 99.86 | 81.17 |
| 456 | 60.00 | 99.72 | 81.67 |
| 567 | 65.00 | 99.95 | 85.83 |

With the same people inside the retain set, as in Protocol U, they lose at most 5.2 points (Appendix D.4). On this model the inside run also removes less of the targets (forget drop 29.8) than the outside run (40.2). Across all eight Protocol U models the holdout drop stays at or below 5.2, including at forget drops of 81.4 to 94.8 (Table 28), so membership in the retain set protects the holdout at every level of removal.

### E.4 BIOGRAPHY-COMPLETION PROBE

After reset-and-repair the model can no longer complete most held-out people's biographies, while retrained models complete all of them (Table 35). For each person and each of their training biographies, we cut the biography just before each attribute value and let the model complete it greedily; a completion counts if it starts with the value. The probe covers every forget and held-out person and 100 retain people (1,200 to 3,000 completions per group). We reran reset-and-repair on the 1,000-step Mixed model with the four recorded seeds and probed the outcome. The reruns land close to the recorded runs (forget 15.2 against 16.4, held-out 27.8 against 25.7, means over seeds); single seeds differ by up to 7.2 points because training on GPUs is not bitwise deterministic.

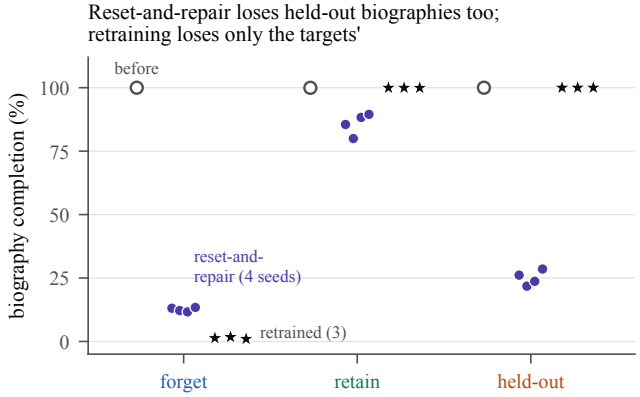



**Figure 22. After reset-and-repair the model completes only 22 to 29% of held-out biographies; the retrained models complete all of them and fewer than 2% of the forget people's.** Biography completion on Mixed (500 people, 1,000 steps) before unlearning (hollow), after reset-and-repair (one dot per seed), and on the three retrained models (⋆). Values in Table 35.

**Table 35. Values behind Figure 22: reset-and-repair removes most of the held-out people's biography facts along with the forget people's; retraining removes only the forget people's.** Biography completion before and after reset-and-repair (Mixed, 1,000 steps), and on the models retrained without the forget people. Forget / retain / held-out.

| Model | Completion accuracy (%) |
|---|---|
| Before unlearning | 100.0 / 99.97 / 100.0 |
| Reset-and-repair, seed 42 | 13.08 / 85.50 / 26.13 |
| Reset-and-repair, seed 345 | 12.17 / 80.00 / 21.80 |
| Reset-and-repair, seed 456 | 11.67 / 88.30 / 23.70 |
| Reset-and-repair, seed 567 | 13.42 / 89.50 / 28.53 |
| Retrained, 1,000 steps | 1.33 / 99.97 / 100.0 |
| Retrained, 2,500 steps | 1.75 / 99.97 / 100.0 |
| Retrained, 2,500 steps, data order 1 | 1.00 / 99.93 / 100.0 |

The retrained models, which never saw the forget people's biographies, complete fewer than 2% of them, so the probe registers a fact's absence. They complete every held-out biography, so retraining loses no held-out facts, and its lower held-out QA accuracy (Table 32) comes from weaker read-out. After reset-and-repair, completion falls with QA accuracy for forget and held-out people alike.

### E.5 FISHER-RANKED AND RANDOM MASKS ON MIXED

Fisher ranking finds the targets far better than chance, and the holdout still falls with them. We reset a fixed fraction of every weight matrix of the 1,000-step Mixed model, chosen either by Fisher share or uniformly at random, and repaired for 300 steps, as in the crossover's fixed-budget reset (Appendix F.3). Table 36 gives the drop in accuracy from the model before unlearning (100.0 / 99.95 / 97.83). The Fisher-ranked mask is deterministic, so it has one run per budget; random masks use different draws.

**Table 36. Values behind Figure 23: Fisher ranking removes far more of the targets than random, and the holdout falls with them** (fixed-budget resets on Mixed; Qwen3-0.6B, 500 people, 1,000 steps). Drop in accuracy, in points.

| Budget | Mask | Runs | Forget drop | Retain drop | Holdout drop |
|---|---|---|---|---|---|
| 1% | Fisher-ranked | 1 | 0.8 | 0.0 | 0.5 |
| 1% | random | 3 | 0.0 | 0.0 | 0.3 |
| 5% | Fisher-ranked | 1 | 6.2 | 1.0 | 5.0 |
| 5% | random | 3 | 0.0 | 0.1 | 1.5 |
| 20% | Fisher-ranked | 1 | 27.9 | 8.8 | 18.2 |
| 20% | random | 2 | 2.1 | 1.8 | 6.8 |
| 50% | Fisher-ranked | 1 | 62.1 | 35.4 | 51.3 |
| 50% | random | 2 | 8.5 | 6.5 | 18.7 |

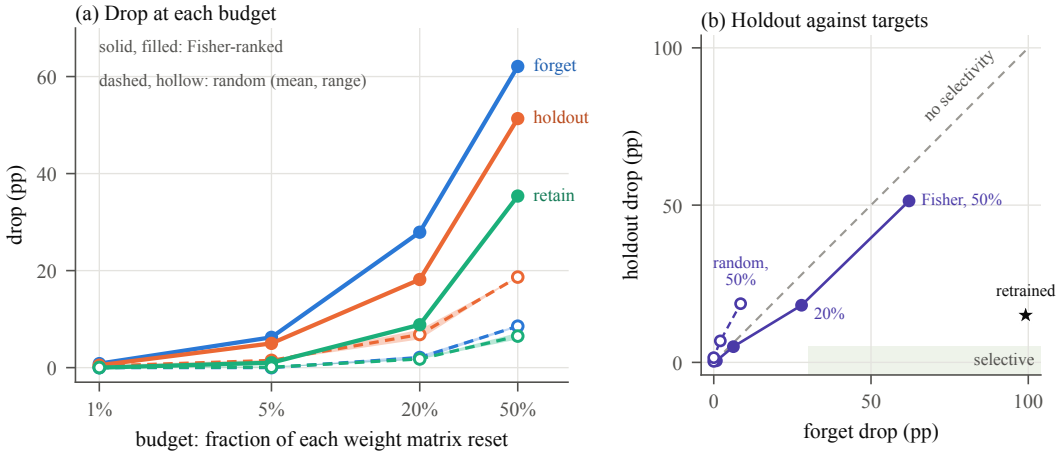



**Figure 23. The Fisher-ranked mask removes far more of the targets than a random mask, and the holdout falls close behind.** (a) Drop from the model before unlearning at each budget, from Table 36. Solid lines with filled markers: Fisher-ranked, one run. Dashed lines with hollow markers: random masks, mean over 3 runs (2 at 20% and 50%). The random runs differ by at most 2.0 points (holdout at 20%), so their shaded min–max range is mostly narrower than the markers. (b) The same resets as holdout drop against forget drop, joined in budget order (Fisher filled, random hollow); the star is the model retrained without the forget people on the same data order (Table 23).

At the 20% and 50% budgets the Fisher-ranked mask removes 13 and 7 times as much of the targets as a random mask of the same size (Figure 23), and the holdout loses 65% as much as the targets at 20% and 83% at 50%. Random masks remove more of the holdout than of the targets at these two budgets; at 1% and 5% the reset removes too little to compare. At roughly equal collateral damage (Fisher-ranked at 20%, random at 50%), the Fisher-ranked mask removes 27.9 points of the targets against 8.5. In the crossover random masks remove more at equal collateral (Section 5); the two settings read facts out differently, through a trained question-answering skill here and a frozen adapter there.

### E.6 A STANDARD METHOD ON TOFU

On TOFU, gradient difference with a retain weight of 5 keeps the retain questions it trains on at the retrained level while untouched retain questions from the same source fall far below it (Table 37, Figure 24). TOFU's standard retain score uses questions contained in the retain set that gradient difference trains on (Section 6). We split the 180 retain authors in half, alternating authors: the method regularizes on half A only, and half B, same source and same fine-tuning stage, is never touched. Starting from OpenUnlearning's Llama-3.2-1B-Instruct TOFU model (Dorna et al., 2025), we unlearn forget10 with gradient difference (Maini et al., 2024), following OpenUnlearning's objective and chat template (AdamW, weight decay 0.01, one warmup epoch, effective batch 32, learning rate $10^{-5}$, ten epochs). After every epoch we score the ROUGE-L recall of greedy answers on forget10

(400 questions) and on 400-question samples of each half, and we score each final model on all 1,800 questions of each half. OpenUnlearning's retain90 model, fine-tuned without forget10, is the retraining reference.

**Table 37. On TOFU, gradient difference with $\alpha$=5 keeps the retain half it trains on and loses the half it never sees, in every seed and with the halves swapped; every gradient-difference run leaves the untouched half below the trained one.** ROUGE-L recall (%) of the final models on all 1,800 questions of each half; half A is the method's retain data, half B is untouched. Means ± population standard deviation over seeds. Lower block: the first sweep (one seed, 400-question samples); two runs stopped early lower both halves together (gradient difference with $\alpha$=1 after five epochs: 38.4 / 33.9; NPO with $\beta$=0.1 after four: 59.7 / 58.9).

| Model | Forget | Half A | Half B | A − B |
|---|---|---|---|---|
| Before unlearning | 79.13 | 80.39 | 80.91 | −0.52 |
| ⋆ *Retrained without forget10* | 38.38 | 81.44 | 81.93 | −0.49 |
| Gradient difference, $\alpha$=5 (3 seeds) | $47.97_{\pm 1.14}$ | $80.82_{\pm 0.91}$ | $62.21_{\pm 1.24}$ | $18.61_{\pm 0.46}$ |
| Gradient difference, $\alpha$=5, halves swapped | 49.58 | 80.72 | 62.73 | 17.99 |
| Gradient difference, $\alpha$=3 (2 seeds) | $28.73_{\pm 3.69}$ | $68.86_{\pm 2.34}$ | $49.88_{\pm 1.41}$ | $18.98_{\pm 0.93}$ |
| Gradient difference, $\alpha$=5, 5 epochs | 68.16 | 85.05 | 76.19 | 8.86 |
| Gradient difference, $\alpha$=2 | 0.81 | 48.08 | 32.75 | 15.33 |
| Gradient difference, lr $2\times10^{-5}$, $\alpha$=5 | 2.47 | 60.18 | 47.46 | 12.72 |
| NPO, lr $2\times10^{-5}$, $\beta$=0.5, $\alpha$=1 | 47.84 | 53.89 | 52.27 | 1.62 |

With the retain weight high enough that half A stays at the retrained level, half B falls. With $\alpha$=5, forget recall falls from 79.1 to 48.0 over three seeds, and half A ends at 80.8, against 81.4 for the retrained model, while half B falls from 80.9 to 62.2, 19.7 below the retrained model's 81.9. With $\alpha$=5, the standard deviation across seeds is at most 1.24 points on every score, and swapping the halves gives the same result, so the gap comes from which half the method trains on. Half B loses 60% as much as the targets (18.7 against 31.2 points), inside the 44–104% range of the Mixed configurations. A retain score built only from questions the method trains on, as TOFU's is, sees half A at 80.8, level with retraining's 81.4. A weaker retain weight ($\alpha$=3) removes more of the targets and lowers both halves, with half B still 19.0 below half A; NPO lowers both halves together.

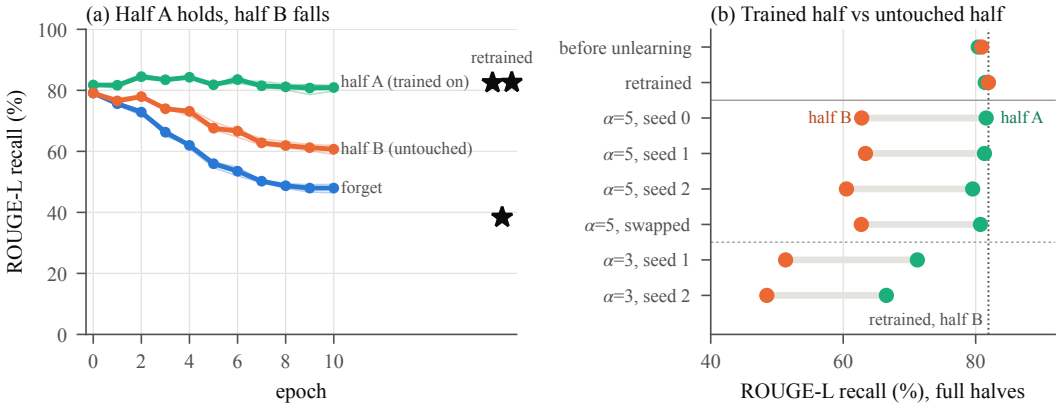



**Figure 24. On TOFU, the retain half that gradient difference trains on stays at the retrained level while the untouched half falls.** *(a)* ROUGE-L recall by epoch on 400-question samples, gradient difference with $\alpha$=5; thick: mean of three seeds, thin: each seed; stars: the retrained model. *(b)* Final models on all 1,800 questions of each half, half A (green) against half B (orange); dotted line: the retrained model's half B.

**TOFU's utility sets and two other methods.** We reran gradient difference with $\alpha$=5 on the same three seeds and also ran NPO ($\beta$=0.1) and SimNPO (Fan et al., 2025) ($\beta$=4.5, margin $\gamma$=0, forget-loss weight 0.125), both with $\alpha$=5 on half A and one seed. Every final model is scored on the two utility sets TOFU adds to its retain questions: 100 questions about real authors and 117 world facts (Table 38, Figure 25). The gradient-difference reruns reproduce the final scores of Table 37 within 0.32 points on forget10 and both halves.

Under gradient difference, the real-author questions fall 12.0 points, to 70.0, 14.5 below the retrained model's 84.5, while the world-fact questions do not fall (82.9, against 82.2 before unlearning). An

untouched set therefore does register the damage here. The set closest to the targets registers the most: half B, which has the same source and fine-tuning stage as the targets, loses 18.7 points. Half A, the kind of question TOFU's retain score is built from, moves by 0.4. NPO lowers both halves by 22 to 23 points, so its retain score shows the damage, while the real-author and world-fact questions do not fall. SimNPO with these settings barely forgets: forget10 falls by only 3.4 points. Training on half A raises half A's recall 8.4 points, to 7.3 points above the retrained model, while half B stays within 1.5 points of its starting level. In SimNPO's case the retain score overstates what is kept rather than hiding a loss.

**Table 38. Under gradient difference, TOFU's real-author questions fall 12.0 points and its world facts do not move, while the untouched retain half falls 18.7.** ROUGE-L recall (%) of the final models on all questions of each set: forget10 (400), half A (1,800, the method's retain data), half B (1,800, untouched), real authors (100) and world facts (117). Gradient difference: reruns of the three $\alpha$=5 seeds of Table 37, mean ± population standard deviation. NPO and SimNPO: one seed each, $\alpha$=5.

| Model | Forget | Half A | Half B | Real authors | World facts |
|---|---|---|---|---|---|
| Before unlearning | 79.13 | 80.39 | 80.91 | 81.95 | 82.22 |
| ⋆ *Retrained without forget10* | 38.38 | 81.44 | 81.93 | 84.45 | 81.08 |
| Gradient difference, $\alpha$=5 (3 seeds) | $47.81_{\pm 1.10}$ | $80.79_{\pm 0.94}$ | $62.22_{\pm 1.14}$ | $69.99_{\pm 1.27}$ | $82.86_{\pm 0.00}$ |
| NPO, $\beta$=0.1, $\alpha$=5 | 51.39 | 58.31 | 57.60 | 83.95 | 85.04 |
| SimNPO, $\beta$=4.5, $\alpha$=5 | 75.75 | 88.76 | 79.46 | 79.20 | 82.91 |

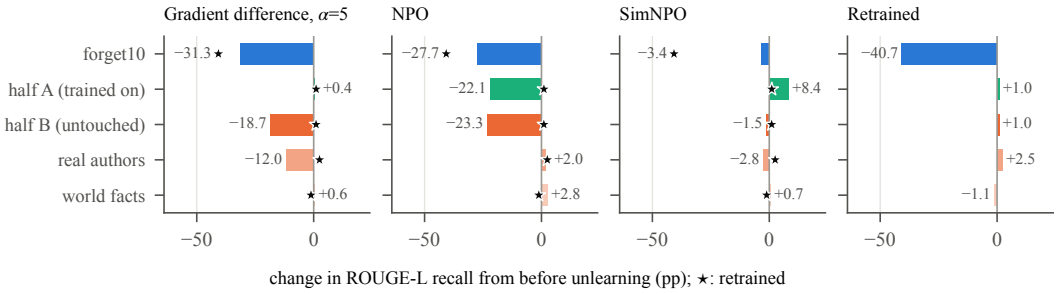



**Figure 25. Only the untouched retain half and the real-author questions reveal gradient difference's damage; half A, which TOFU's retain score is built from, does not move.** Change in ROUGE-L recall from the model before unlearning on every set of Table 38, one panel per model. Gradient difference is the mean of three seeds. Stars mark the retrained model's change on each set.

# F CROSSOVER

In the crossover the holdout is learned exactly like the targets, which the Mixed models do not control (Section 5). Changing when facts are learned changes the share of parameters that favor them 487-fold, and neither arm is removed selectively.

## F.1 DESIGN

Both crossover models start from pretrained Qwen3-0.6B-Base and share one population of 50,000 people. Two disjoint focal sets, $A$ and $B$, hold exactly 400 people each. Model 1 learns $A$ early and $B$ late; model 2 swaps them, so every focal person is learned early in one model and late in the other.

- **Early phase.** Continued pretraining for 40,000 steps of 24,576 tokens on the biographies of 49,600 people (everyone except the late focal set), mixed half and half with QA for a separate scaffold set.
- **Late phase.** 323 further steps on the late focal set's biographies with the same scaffold mix and a fresh optimizer. The step count gives each late person as many biography passes as an early person.
- **Read-out.** One QA LoRA (attention rank 16, embedding rank 128), trained once on the 2,000 scaffold people on a separate model that saw neither focal set, and kept frozen for every evaluation. No focal person's question is ever trained.

- **Other sets.** A retain set of 2,000 background people is used for repair and for the Fisher retain estimate. The untargeted holdout is 200 background people whose biographies were trained and whose questions were not; nothing in the method refers to them.
- **Seeds.** Training seeds 0, 1, 2, 4 and 5 give 20 model–focal-set combinations (10 per arm).

Reset-and-repair resets to pretrained Qwen3-0.6B-Base, the weights both crossover models started from, so the reset reference is each model's own initialization. The crossover's Fisher share replaces the retain estimate in $W$ with the larger of the retain and read-out estimates: $s = F_A/(F_A + \max(F_R^{\text{retain}}, F_R^{\text{QA}}))$.

### F.2 Pre-registration and its amendments

The pre-registered decision rule fails. We froze it on 2026-07-31, when 2 of the 36 seed-0 runs had finished and before any frontier could be traced. The rule fixes an effect threshold of 2 points, a collateral limit of 5 points, the budget grid $q \in \{0.05, 0.1, 0.2, 0.5, 1, 2, 5, 10\}$ (percent of each weight matrix reset; a $q$=0 arm in the launched grid was dropped minutes after the freeze), and four criteria that must all pass:

1. at equal budget, early facts lose more target than collateral, relative to late facts, by at least 2 points;
2. at the collateral limit, early facts lose at least 2 points more target than late facts;
3. the sign of criterion 1 holds in both focal-set assignments;
4. the early arm's frontier (target loss against collateral) lies above the late arm's everywhere.

The rule fails on criterion 4: the frontiers cross in 7 of the 10 focal-set–seed comparisons on the pre-registered grid. Criterion 1 passes as a scalar, and Appendix F.3 shows that it passes only because the late arm loses more of everything.

The pre-registration covered seed 0. We report three later additions as diagnostics: the bracketing budgets $q \in \{0.12, 0.15, 3, 4\}$, frozen on 2026-07-31 for seed 0 and added to every model, which fill gaps in the grid near the 5-point limit (in Table 41, every early model is read at 3% or 4%, while the late models are read at pre-registered budgets of 0.05% to 0.5% or at 0.15%); seeds 1 and 2, added after seed-0 results existed; and seeds 4 and 5, added on 2026-08-02. Every number in this section uses all five seeds unless it says otherwise.

### F.3 Threshold reset, fixed-budget reset and the budget confound

The threshold reset of Appendix B.6 resets far more parameters for early facts than for late ones, so it cannot compare the arms at equal cost. With threshold $s > 0.5$ (and a per-tensor cap), it resets 4.63% of parameters in the early arm against 0.098% in the late arm, a factor of 47.5 (3 seeds, 12 models). The confirmatory test therefore uses a fixed-budget reset, which resets exactly a fraction $q$ of every weight matrix, chosen by $s$, with no threshold. The realized mask equals the nominal budget to four decimal places in all 160 model–budget combinations of the pre-registered grid, and in all 240 once the bracketing budgets are included.

At all eight pre-registered budgets, the late arm loses more of its target than the early arm (early minus late target loss ranges from $-5.36$ to $-35.11$), and it also loses more collateral (by up to 38.73 at $q$=10). The equal-budget contrast of criterion 1 is the difference of these two differences, and it is positive because the late arm's extra collateral slightly exceeds its extra target loss. At the 5% budget it is $+3.25$, and it changes sign if collateral is weighted by $w = 0.908$ instead of 1. Measured against the untargeted holdout, both arms lose about as much holdout as target at every budget (Figure 26).

**Post-hoc check: a four-round threshold schedule collapses the crossover models, a second reason the confirmatory test uses fixed-budget resets.** We ran four rounds with thresholds 0.55, 0.85, 0.95, 0.95 and 300 repair steps on the retained people's biographies, resetting either to pretrained Qwen3-0.6B-Base or, on a second pair of crossover models, to a model trained on the same corpus without either focal set, and compared with gradient ascent on the target biographies (Table 39). With the pretrained reference, and under gradient ascent, every group falls to 0. With the reference trained without the focal sets, both focal sets are removed, the untargeted one included, and scaffold and

holdout accuracy fall far. Every run loses at least 48 points of scaffold accuracy and removes the untargeted focal set as fully as the target, so none can compare the arms.

**Table 39. A four-round threshold schedule removes the untargeted focal set as fully as the target and collapses the scaffold and holdout.** Crossover models, accuracy after unlearning. The second pair of crossover models (seeds 0, 1 and 2) resets to a model trained on the same corpus without either focal set; its runs set each focal set as the target in each model. Gradient ascent runs for 200 steps at learning rate $10^{-5}$ on the target biographies, in all 14 settings of the two threshold-schedule rows.

| Method, reset reference | Runs | Target set | Other focal set | Scaffold | Holdout |
|---|---|---|---|---|---|
| Before unlearning, second pair | | 83.8–88.3 | 83.8–88.3 | about 99 | about 89 |
| Threshold schedule, pretrained Qwen3-0.6B-Base (seed 0, target $A$) | 2 | 0 | 0 | 0 | 0 |
| Threshold schedule, model trained without the focal sets | 12 | 0.8–1.8 | 0.8–1.8 | 30.5–50.2 | 20.9–37.0 |
| Gradient ascent | 14 | 0 | 0 | 0 | 0 |

**Post-hoc check: under the premise's own one-shot threshold reset, the early arm's holdout loses 88% as much as the target, and protection follows what the reset reference already knows.** The 487-fold difference in Fisher share is measured for the threshold operator, while the confirmatory test uses fixed budgets. We therefore also ran the threshold operator once ($s{>}0.5$, at most 5% of parameters, 300 repair steps) on both crossover models, seeds 0 to 2, each focal set as the target (Table 40). Reset to pretrained Qwen3-0.6B-Base, the early arm loses 17.0 points of its target, 17.6 of the other focal set and 14.9 of the holdout; the late arm, whose mask is 47.5 times smaller, loses about 2.6 of every group. Reset instead to a model trained on the same corpus without either focal set, which does know the holdout and scaffold, the other focal set still falls about as much as the target under every operator, while the holdout and scaffold lose at most 6.3. Against that holdout this reset looks selective, but not against the untargeted focal set the reference lacks: a group is spared when the reset reference or the repair data contain it, not because the method avoids it. This is the same rule as retain-set membership, and it is why a holdout must be untouched by everything the method uses, including its reset reference.

**Table 40. A single threshold reset removes the untargeted focal set as much as the target; only groups the reset reference already knows are spared.** Mean accuracy drop (points) over seeds 0–2, both models, each focal set as the target (6 runs per arm). "Other focal set": the untargeted focal set of the same model. Post-hoc; not part of the pre-registration.

| Operator, reset reference | Arm | Target | Other focal set | Holdout | Scaffold |
|---|---|---|---|---|---|
| Threshold reset ($s{>}0.5$, at most 5%), pretrained Qwen3-0.6B-Base | early | 17.0 | 17.6 | 14.9 | 9.7 |
| | late | 2.6 | 2.9 | 2.7 | 2.3 |
| Threshold reset, model trained without the focal sets | early | 7.8 | 6.7 | 1.7 | 0.9 |
| | late | 0.4 | 0.4 | 0.0 | 0.1 |
| Fixed 5% budget, model trained without the focal sets | early | 5.1 | 4.5 | 0.8 | 0.6 |
| | late | 10.4 | 9.5 | 4.3 | 2.4 |
| Fixed 10% budget, model trained without the focal sets | early | 11.5 | 11.8 | 2.4 | 1.4 |
| | late | 19.7 | 16.1 | 6.3 | 3.5 |

### F.4 Selectivity against the untargeted holdout

At the pre-registered 5% budget, target loss minus holdout loss is $+1.46$ in the early arm and $-1.54$ in the late arm, both below the 2-point threshold. The bracketing diagnostic reads each model at the largest budget whose collateral stays within the 5-point limit. It splits the collateral people in half, selects the budget on one half and scores it on the other, so that no person is used for both. Figure 27 plots every model and Table 41 lists them.

On average the early arm removes 6.59% of its target within the limit against 3.48% for the late arm. Paired within focal set and seed, the difference is $+3.23$ (95% CI $[+1.86, +4.30]$, $n{=}8$, 7 of 8

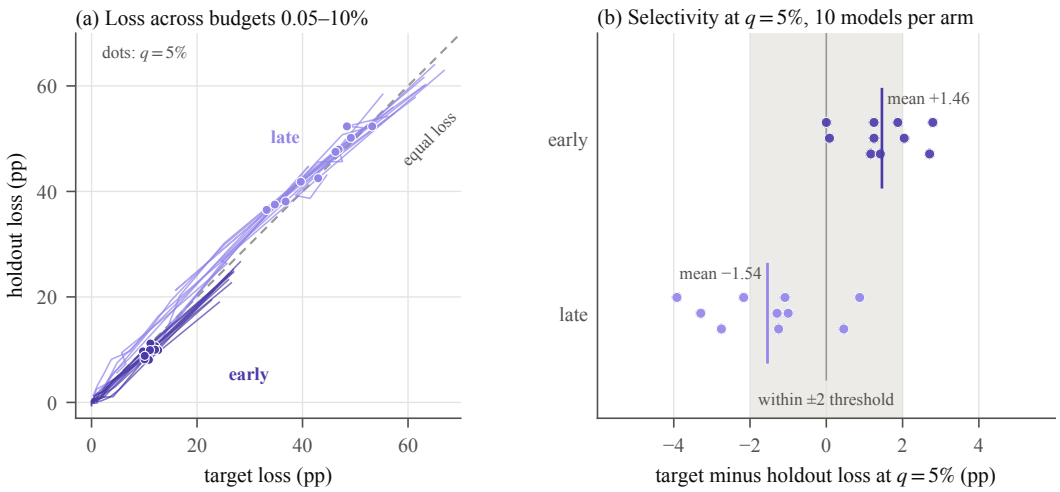



**Figure 26. In both crossover arms, fixed-budget resets remove the holdout about as much as the targets.** (a) Holdout loss against target loss for each of the 20 models, one line per model across the eight pre-registered budgets ($q = 0.05\%$ to $10\%$); dots mark $q = 5\%$. Dark: early arm; light: late arm. (b) Target minus holdout loss at $q = 5\%$, one dot per model; bars mark the arm means, $+1.46$ (early) and $-1.54$ (late). The shaded band is the 2-point effect threshold.

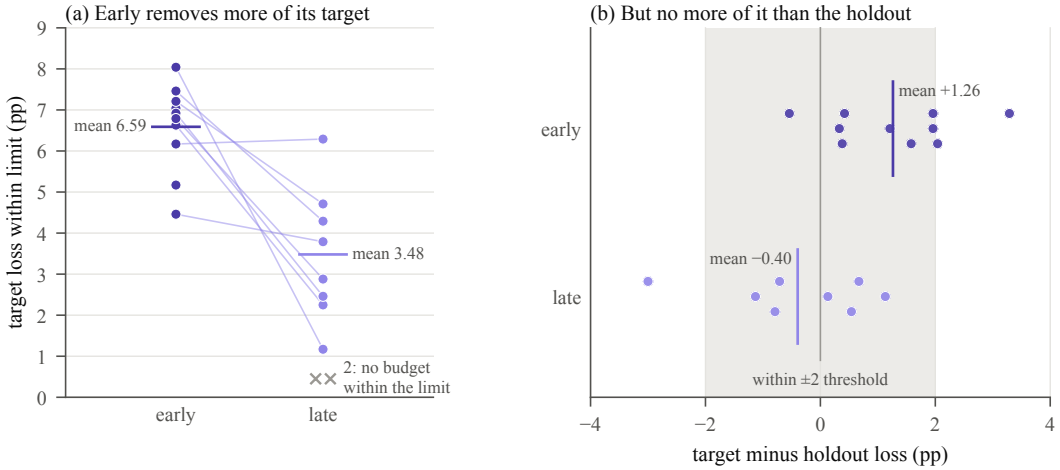



**Figure 27. At the collateral limit, the early arm removes more of its target but no more selectively.** Bracketing diagnostic, 5 seeds, each model read at the largest budget whose collateral stays within the 5-point limit. (a) Target loss, early against late, one line per focal set and seed; bars mark the arm means, 6.59 (early) and 3.48 (late). Two late models have no budget within the limit (×). (b) Target minus holdout loss per model; bars mark the arm means, $+1.26$ (early) and $-0.40$ (late). The shaded band is the 2-point effect threshold.

positive) when the two undefined pairs are dropped, and $+3.80$ ($[+2.80, +4.75]$, $n$=10) when their late value is set to 0; both rules were fixed before the interval was computed. At seed 0 alone the difference is $+2.13$ on two pairs. The extra removal does not favor the targets: the early arm's holdout loses 4.6 to 6.7 points, and target minus holdout averages $+1.26$ early and $-0.40$ late, both under the 2-point threshold. Two early models (5.13 and 5.35) and one late model (5.16) exceed the limit on the scoring half after meeting it on the selection half; the table reports them as scored. Adding the bracketing budgets raises the frontier crossings from 7 to 8 of 10.

### F.5 Random masks

At equal collateral damage, random masks remove more of the targets than the $s$-ranked mask (Figure 28, Table 42). For each model we drew ten random masks at each budget, measured each

**Table 41.** Values behind Figure 27. $q$ is the budget selected on the calibration half; target loss, held-out-half collateral and target minus holdout loss are scored on the other half. Two late models have no budget whose collateral stays within the limit.

| Arm | Focal set | Seed | $q$ (%) | Target loss | Collateral | Target − holdout |
|---|---|---|---|---|---|---|
| early | A | 0 | 3 | 6.63 | 3.29 | +2.04 |
| early | A | 1 | 3 | 7.04 | 5.13 | +0.38 |
| early | A | 2 | 3 | 7.21 | 4.23 | +1.21 |
| early | A | 4 | 3 | 6.92 | 3.53 | +1.58 |
| early | A | 5 | 3 | 5.17 | 3.28 | +0.42 |
| early | B | 0 | 3 | 6.17 | 3.70 | +0.33 |
| early | B | 1 | 3 | 7.46 | 5.35 | +1.96 |
| early | B | 2 | 4 | 8.04 | 4.29 | +3.29 |
| early | B | 4 | 3 | 6.79 | 3.90 | +1.96 |
| early | B | 5 | 3 | 4.46 | 2.90 | −0.54 |
| late | A | 0 | 0.2 | 2.25 | 2.78 | −3.00 |
| late | A | 1 | | no budget within the limit | | |
| late | A | 2 | 0.5 | 4.71 | 3.85 | +0.13 |
| late | A | 4 | 0.2 | 2.46 | 2.15 | −0.71 |
| late | A | 5 | | no budget within the limit | | |
| late | B | 0 | 0.5 | 6.29 | 4.49 | +1.13 |
| late | B | 1 | 0.2 | 4.29 | 2.70 | +0.54 |
| late | B | 2 | 0.05 | 1.17 | 0.79 | +0.67 |
| late | B | 4 | 0.15 | 2.88 | 2.69 | −0.79 |
| late | B | 5 | 0.1 | 3.79 | 5.16 | −1.13 |

one's collateral, and compared its target loss with what the ranked mask removes at that collateral on the model's full ranked sweep.

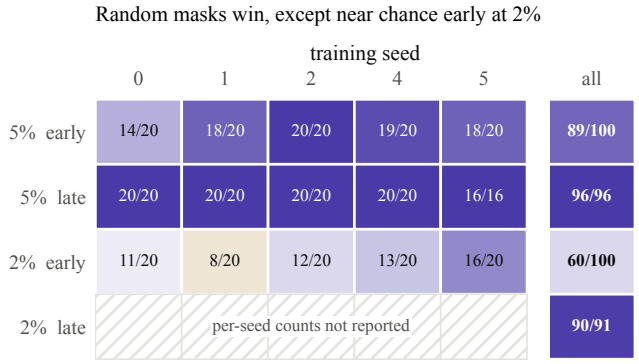



**Figure 28. Random masks beat the ranked mask at matched collateral in every cell except the early arm at 2%, which sits near chance.** Share of random masks that remove more of the targets than the ranked mask, by budget, arm and training seed; white is chance (0.5), purple favors random. Per-seed counts for the late arm at 2% are not reported.

**Table 42.** Values behind Figure 28: wins of a random mask over the ranked mask, by training seed. At $q$=5%, 4 of the 200 draws at $q$=5%, all late (seed 5, model 2, set $A$), have collateral outside the ranked sweep and cannot be compared, leaving 196. At $q$=2%, 9 late draws are excluded for the same reason.

| Budget | Arm | Seed 0 | 1 | 2 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| 5% | early | 14/20 | 18/20 | 20/20 | 19/20 | 18/20 | 89/100 |
| 5% | late | 20/20 | 20/20 | 20/20 | 20/20 | 16/16 | 96/96 |
| 2% | early | 11/20 | 8/20 | 12/20 | 13/20 | 16/20 | 60/100 |
| 2% | late | | | | | | 90/91 |

At 5%, random wins 185 of 196 comparisons. Across both budgets it wins 186 of 187 late comparisons. The early arm at 2% varies around chance by seed.

This result and the Inserted Fisher control (Appendix D.1.1) hold different quantities fixed. There the ranked mask beats a random mask with the same number of parameters; here the two are compared at equal collateral. A random mask of equal size does more collateral damage than the ranked one, so matching on collateral gives random masks a smaller size.

### F.6 Stability of the Fisher-share difference

The difference in Fisher share between the arms is large at every threshold and stable under resampling (Table 43). With the independent Fisher re-estimate the early arm has $487\times$ the late arm's share of parameters with $s > 0.5$, and even the lowest early model is far above the highest late model. Unit-normalizing each document's gradient before squaring shrinks the difference without closing it, and early facts are also fit less tightly. The lowest early model stays above the highest late model at every interior threshold; the two exceptions are the endpoints, where both arms are at 100% or 0% by definition. Masks recomputed on disjoint halves of the documents overlap far above chance, whereas the same focal set's masks in models trained with different seeds overlap far less, so the mask's location depends on the trained model more than on the fact. Halving the documents behind the estimate leaves the early share more than two orders of magnitude above the late arm.

**Table 43. The early arm's Fisher share exceeds the late arm's by orders of magnitude, under every resampling we tried.** Crossover models. Shares are percentages of eligible parameters with $s > 0.5$. Jaccard overlaps compare masks within one model (document halves) or across models trained with different seeds; the chance overlap of two top-$k$ masks is 0.026. Mask composition refers to the fixed-budget masks of Appendix F.3.

| Measure | Early arm | Late arm | Early vs. late |
|---|---|---|---|
| *Share with $s > 0.5$* | | | |
| Independent Fisher re-estimate, arm mean | 51.17% | 0.105% | $487\times$ |
| Same, lowest early / highest late model | 43.07% | 0.227% | |
| Unit-normalized document gradients | 21.06% | 0.362% | $58.1\times$ |
| One early model, 600 documents | 51.6% | | |
| Same model, two 300-document halves | 68.4%, 60.5% | | |
| *Across thresholds* | | | |
| Lowest early model above highest late | at 49 of 51 grid values | | |
| Every early over every late model at 0.5, either estimate | | | at least $189\times$ |
| Ratio of arm means (mask-defining estimate) at 0.1 | | | $57\times$ |
| at 0.3 | | | $231\times$ |
| at 0.5 | | | $735\times$ |
| at 0.9 | | | $91\times$ |
| *Mask overlap (Jaccard)* | | | |
| Top-$k$ masks (22,020,096 parameters), disjoint document halves | 0.451–0.516 | 0.730–0.825 | chance 0.026 |
| Deployed threshold masks, disjoint document halves | 0.482–0.624 | 0.602–0.710 | |
| Same focal set, different training seeds (multiple of chance) | $1.75\times$, $1.82\times$ | $3.26\times$, $3.95\times$ | |
| mean Jaccard | 0.045–0.101 | | |
| Early vs. late fixed-budget masks | mean 0.0376 (chance 0.0256) | | $1.47\times$ chance |
| range across focal sets and seeds | | | $1.22\times$–$1.79\times$ |
| *Fit and mask composition* | | | |
| Forget-corpus loss | 0.476 | 0.141 | |
| Share of fixed-budget mask with $s > 0.5$ | all ($s \gtrsim 0.52$) | 0.4%–4.1% | |
| Median $s$ over an equal-size random mask | $1.18$–$1.22\times$ | $4.42$–$8.19\times$ | |
| mean | $1.20\times$ | $5.79\times$ | |

### F.7 Remaining differences between the arms

Two differences between the arms remain after the crossover's controls.

- **Baseline headroom.** Before any reset, late facts are answered less accurately than early facts, by 2.35 points on average and in all ten comparisons (1.84 over the six comparisons from seeds 0, 1 and 2). The difference goes with the arm itself, so swapping focal sets cannot cancel it.

- **Mask composition.** Fixing the budget equalizes the number of parameters reset but leaves their scores different (last block of Table 43). Every parameter in an early mask is above the $s > 0.5$ cut, while only a few percent of a late mask is; with 0.105% of parameters above the cut, a 5% budget has to reach below it. The late mask is still a real ranking, far above an equal-size random mask in median $s$, whereas in the early arm a random draw is already mostly above the cut. Random masks beat the ranked mask in 96 of 96 late comparisons at 5% (Table 42), so they win even where the ranking differs most from random.

The two arms' fixed-budget masks overlap little more than chance (Table 43), so the two arms reset largely different parameters. The treatment also bundles the number of passes, their spacing, the optimizer state and the training that follows acquisition; "early" and "late" name the whole bundle. The Fisher used here is the diagonal empirical Fisher, which shrinks on well-fit examples (Kunstner et al., 2019).

### F.8 How this fits with Inserted and Mixed

The three settings agree once the reset budget is taken into account. In the crossover, the arm with more concentrated Fisher share (early) is no more removable than the other arm. Among the Qwen3 Mixed models, larger Fisher-tail mass goes with more removal (Appendix D.3), and on Inserted, Fisher ranking beats random masks at equal size (Appendix D.1.1). A larger tail mass means a larger threshold reset, so it predicts more removal of everything outside the retain set, targets and holdout alike. The Mixed grid shows this, and so does the crossover's threshold reset, where reset size differs 47.5-fold between the arms. When the budget is fixed, as in the crossover's confirmatory test, concentrated Fisher share buys no extra removal and no selectivity. The Inserted comparison at equal size shows that the ranking picks better parameters than chance for a given count. Only the crossover and the Mixed models measure selectivity against an answerable holdout, and both find little of it.

## G Scope and Limitations

### G.1 What the claims cover

Every result is behavioral: a fact counts as removed when the model no longer produces it in answer to our QA prompts, including under the sampling, paraphrase, relearning and quantization probes of Appendix D.1.3. Privacy and legal deletion are separate questions; the membership and relearning probes on Inserted (AUC 0.848 against 0.435, and faster relearning than after retraining) show where reset-and-repair already differs from retraining on them. Retraining serves as the reference for what selective removal should leave intact.

The settings are synthetic biographies, and the main results use models of up to 1.24B parameters; two real-corpus checks use Llama-2-7B and OLMo-1B (Appendix D.7). A holdout is answerable exactly when its biographies were trained, and Two-stage starts from random initialization while the other settings start from pretrained models. The Mixed step grid and the crossover were built for other questions and are reanalyzed here. The Two-stage reset-and-repair runs used another model's weights as the reset target, so the main text sets them aside.

### G.2 Predictions that would change our conclusions

- **Gradient methods.** Section 3 would be overturned by a gradient method that, on the Two-stage model and the same forget set, reaches a forget gap of at least 30 with retain accuracy of at least 95 and held-out accuracy within 2 points of the original, over at least three seeds.
- **Shared question-answering skill.** The shared-skill reading of the diagnostics would be undercut by a retain-only adapter that, on a comparable model, answers at most 20% of forget questions, together with a reversal of one other diagnostic.
- **Reset-and-repair.** Section 4 would be overturned by a reset-and-repair variant that, on a Mixed model and without adding the holdout to the retain set, reaches a forget gap of at least 30, retain accuracy of at least 95, and a holdout drop within 1.5 points of matched retraining. Such a variant would show that the failure is avoidable within the reset-and-repair family; it would not make

forget and retain scores sufficient to detect collateral damage. At 2,500 steps, matched retraining on the Mixed pipeline loses 5.8 and 8.7 points of holdout (two data orders) against 39.1 for reset-and-repair (Appendix E.1).

- **Crossover.** Either of two results would overturn Section 5: a replication with exactly matched reset budgets that finds target minus holdout loss of at least 2 points in both focal-set assignments, within the 5-point collateral limit and with confidence intervals above zero; or a ranked mask that beats random masks in at least three quarters of matched-collateral draws.

### G.3 Relation to reversibility and collapse

Xu et al. (2026b) show that a drop in forget accuracy can come from reversible suppression or from general damage instead of deletion. Our success criterion guards against damage by requiring retain accuracy and holdout accuracy to survive, and our relearning probes test reversibility. The 500-person 500-step Mixed model is the clearest case of damage that looks like removal: it has the largest forget drop of any reset-and-repair model on Mixed and loses nearly all its holdout with it.

### G.4 Broader impacts

Unlearning was proposed to answer privacy and data-deletion requests (Cao & Yang, 2015; Bourtoule et al., 2021) and is now also applied to copyrighted text (Eldan & Russinovich, 2023). A method that appears to remove a fact while also removing untested knowledge, or while leaving a membership signal, can be taken as proof of removal when it is not. Our recommendation, an answerable holdout outside both the forget and retain sets, makes that failure visible at small cost. All biographies are synthetic, and the work used no personal data and no human subjects.

## H Extended Related Work

**Unlearning methods.** Gradient methods ascend the loss on forget data (Jang et al., 2023; Yao et al., 2024), with variants that add a retain term, preference losses such as NPO and SimNPO (Zhang et al., 2024; Fan et al., 2025; Rafailov et al., 2023), representation steering such as RMU (Li et al., 2024), or task arithmetic (Ilharco et al., 2023). Other methods act on representations or sparse features (Yamashita et al., 2025; Wang et al., 2025a; Farrell et al., 2024), restrict updates to forget-specific representation subspaces by collapsing directions shared with general data (Sondej & Yang, 2025; Sondej et al., 2026), fine-tune only mechanistically localized components (Guo et al., 2025), act at inference time through in-context examples or corrupted prompt embeddings (Pawelczyk et al., 2024; Liu et al., 2024), steer behavior along single activation directions such as the one that mediates refusal (Arditi et al., 2024), or edit individual facts (Meng et al., 2022; 2023). Reset-and-repair adapts PerTA (Cai et al., 2026), which uses per-parameter Fisher importance, as in elastic weight consolidation (Kirkpatrick et al., 2017), to set how much of a forget task vector to apply at each parameter. Related Fisher-, saliency- and attribution-based methods pick parameters or layers to modify for forgetting (Foster et al., 2024; Liu et al., 2023; Fan et al., 2024; Kim et al., 2025b; Xu et al., 2026a), localize parameters with neuron scores and learned masks before editing them (Zhang et al., 2025b), or localize each request with dynamic gradient masks in continual unlearning (Wuerkaixi et al., 2025), and conflict-guided localization separates forget, retain and conflict neurons (Chen et al., 2026b). Selective pruning removes the neurons whose importance to a target capability is highest relative to their importance to general performance (Pochinkov & Schoots, 2024), a neuron-level analogue of the forget share $W_i$. GRIN (Anjarlekar & Pombra, 2025) selects parameters by a forget-to-retain gradient ratio, injects noise into them, and then fine-tunes, much as reset-and-repair resets Fisher-selected parameters before repairing on the retain set. Merged task vectors that keep only sign-consistent elements refine task arithmetic for unlearning (Kim et al., 2025a). Resetting selected weights to earlier values echoes the lottery-ticket procedure, which resets the weights that survive pruning to their initial values (Frankle & Carbin, 2019). Recent editing-based unlearning methods (Han et al., 2026; Yu et al., 2025) and LoRA-based fine-tuning ones (Xu et al., 2025b) were not evaluated here.

**Benchmarks and evaluation.** TOFU (Maini et al., 2024), MUSE (Shi et al., 2025b), WMDP (Li et al., 2024), RWKU (Jin et al., 2024) and OpenUnlearning (Dorna et al., 2025) score a forget set and a retain or utility set. Several works show that these scores overstate forgetting, because facts return after light fine-tuning (Lynch et al., 2024; Hu et al., 2025a; Deeb & Roger, 2024; Xu et al.,

2026b), under sampling (Reisizadeh et al., 2026), after quantization (Zhang et al., 2025c), through overlapping and multi-hop queries (Hu et al., 2025b; Jeung et al., 2026; Shah et al., 2025b), when prompts emphasize the keywords unlearning taught the model to ignore (Zhang et al., 2025a), on semantically adjacent surrogate data (Jia et al., 2025), under probing when methods obfuscate rather than remove (Sun et al., 2025), or in activations that output metrics miss (Lee et al., 2026). Critiques of unlearning evaluations make related points about format and probability (Scholten et al., 2025; Ye et al., 2025; Thaker et al., 2025), and about evaluations that re-teach the model or rest on spurious correlations (Feng et al., 2025). On the other side of the ledger, unlearned models fail many benign questions near or far from the removed content (Ko et al., 2025), current methods either fail to remove a fact that can still be deduced from retained ones or remove unrelated facts in doing so (Wu et al., 2024), methods lose the ability to use removed facts supplied in the prompt (Peng et al., 2026), benign boundary questions expose over-unlearning that utility sets miss (Diamant et al., 2026), and methods near-optimal on TOFU's metrics can move farther from the retrained model than before unlearning (Cho et al., 2025). Yang & Yeung (2026) score held-out forget facts against matched retraining, and Yoon et al. (2026) argue that unlearning should mean indistinguishability from retraining. Results also depend on the original model's training seed (Lanyon et al., 2025), which is why we compare each method with retraining from the same data order. Our recommendation adds a holdout the model can answer that neither the forget set nor the retain set covers.

**Localization and its limits.** Whether facts sit in a small set of parameters is contested (Hase et al., 2023; Chang et al., 2024; Chen et al., 2025). Lee et al. (2025) find that localization does not reliably enable removal, which the crossover's random-mask result agrees with. Hong et al. (2025a) find that standard unlearning methods barely change the parameters that encode a concept, while ablating those parameters removes it. Other work finds parameters or attention heads specialized for skills and types of knowledge (Hong et al., 2025b; Bair et al., 2026). When facts are written into predefined parameters, unlearning restricted to them removes the facts (Boglioni et al., 2026), and localization can instead be imposed during training, through gradient routing (Cloud et al., 2024) or source-specific components that are disabled at deployment (Ghosal et al., 2026); our models are trained without such constraints. Circuits located before fine-tuning drift as parameters change (Chen et al., 2026a), and in a fully interpretable network, standard methods can inhibit a concept without removing it (Kodama et al., 2026). Feature superposition (Elhage et al., 2022) and statistical dependence between forget and retain data (Mavrothalassitis et al., 2025) give reasons to expect that one fact cannot be removed without touching others; how entangled a fact is predicts whether it leaks after unlearning (Shah et al., 2025a), and collateral damage weakens with semantic distance without vanishing (Su et al., 2026).

**How facts are learned.** Our biographies follow Allen-Zhu & Li (2024), who show that a model can answer questions about people whose biographies it saw but whose questions it was never trained on only if the biographies appear in varied forms during pretraining or question–answer data is mixed into pretraining; companion work covers manipulation and capacity (Allen-Zhu & Li, 2025b;a). How a fact is learned shapes whether it is extractable (Allen-Zhu & Li, 2024; Zucchet et al., 2025), and related work asks how the way, time, or frequency with which a fact is learned affects unlearning (Wu et al., 2026; Krishnan et al., 2025; Borisiuk et al., 2026; Ma et al., 2025), which samples are hard to unlearn (Cheng et al., 2026a), and whether local unlearning can match retraining when a model is trained in stages (Yu et al., 2026). When a capability is acquired also shapes how it survives later fine-tuning (Feng et al., 2026); in our crossover, learning time changes Fisher concentration but not selectivity.

# I Reproducibility

Every number in the paper was recomputed from the per-run evaluation results of the runs it describes, except the five artifacts below, and the headline values were checked again after every revision of the text.

**We did not keep five artifacts:** the model behind diagnostics D1–D4, the 50K-person Fisher estimates, the original nanochat-d12 reset reference, the fourth round of the Mixed 1,000-person 4,000-step model, and the OLMo-1B checkpoint; we report their recorded results.

**Compute.** All runs used NVIDIA L40S GPUs (46 GB). One reset-and-repair round on Qwen3-0.6B takes about four minutes on one GPU, and a four-round run 17 to 20 minutes.