Fishing for Privacy: Unlearning a Rare Cell Type from a Gene-Expression VAE Is Not Reached, Approached, and Not Verifiable

David Benson *dmb2262@columbia.edu*
Columbia University

Abstract

Can a gene-expression VAE undergo unlearning of a rare cell type without losing utility? Nine post-hoc unlearning methods and DP-SGD are tested on scVI-style VAEs trained on single-cell and bulk RNA-seq, against models retrained without the forget cells. A method has unlearned the cells when the strongest available membership attack separates them from cells the model never saw no better than it does on the retrained models, whether or not the model still represents their type. No method meets that standard at the budget its authors use. Where the cells can be tested, every method on one forget set and four of seven on the other, two of them after a multiplicity correction, leave them more identifiable than the retrained models do; the rest are not resolved, on intervals too wide to certify equivalence. Where only the type can be tested, the retrained level is not reached, and the methods miss it in two kinds. Those that edit and then fine-tune leave the cells more identifiable than a retrained model does. Those that push them with a continuous objective leave them less identifiable, fitted worse than a model that never saw them wherever the audit resolves them.

Retain-only fine-tuning run to 200 epochs, four times the published retain fine-tune's budget, erodes the memorized fit with held-out utility inside its margins. On the PBMC-68k megakaryocytes it brings them into the spread of the retrained models, though the design cannot certify that it arrives. Nothing a controller can observe marks arrival, so the result cannot be verified from the model alone. Two measurements explain why. A retain fine-tune reverts the part of an edit that lies in the directions the retain data constrain and compensates the rest. In the shared decoder on PBMC-33k almost every direction the forget cells use is one the retain cells use too, so these methods reduce to retain fine-tuning. The level a retrained model gives the forget cells is the fit of a type it does not know, which no cell it knows supplies. The usual evaluation, which scores the forget cells against cells of other types, measures how well the model fits the type, not whether it trained on the cells.

1 Introduction

Can a gene-expression VAE undergo unlearning without losing utility? Expression profiles carry disease status and genetic predisposition (Walker et al., 2024). A data controller who trains a scVI-style variational autoencoder (Lopez et al., 2018) on them, and who is then asked to remove one donor's cells or one subtype's patients, would rather edit the trained model than retrain it. The removal that would cost the most is a rare subpopulation, because the model's representation of it rests on the cells being removed. It is also the removal that matters most. Memorization concentrates in rare subpopulations (Feldman, 2020), and a membership inference attack (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022) is most dangerous where memorization is.

The question is put to scVI-style VAEs on four cohorts, PBMC-33k, PBMC-68k, and Tabula Muris (single-cell RNA-seq) and TCGA-BRCA (bulk RNA-seq), with a rare cell type or a molecular subtype as the forget set. Nine post-hoc unlearning methods and DP-SGD are judged against models retrained without the forget

set. Success is defined as an adversary would define it. A method has unlearned the forget cells when the strongest membership attack available cannot separate them from cells the model never saw any better than it can on a model retrained without them. Whether the model still represents their type is a separate matter; a model that has learned what a megakaryocyte looks like from other cells is doing its job, and the privacy question concerns the cells. The comparison is paired, on the same cells, against an arm of retrained models, with an equivalence margin of half the smallest excess the original model is shown to have over them.

No method meets the standard at the budget its authors use. On PBMC-68k held-out cells of the forget type make the question one about individual cells. Every method on the larger forget set leaves the cells more identifiable than eight retrained models do, with paired differences of 0.062 to 0.283 in AUC whose intervals exclude zero after a correction for multiplicity. On the smaller set four methods do, two of them after the same correction over the eight arms, and the rest are not resolved. On PBMC-33k only five held-out cells of the forget type exist, too few for a question about individual cells, so the attacker reads the model's excess fit of the type. There the methods miss the retrained level in two kinds. Those that edit and then fine-tune leave the forget cells more identifiable than the retrained models, fitted as the original model fits them. Those with a continuous forget objective leave them less identifiable, fitted worse than a model that never saw them. The advantage commonly used to report unlearning, twice the attacker's distance from chance, cannot tell the two apart.

Retain-only fine-tuning run for two hundred epochs, four times the published retain fine-tune's budget and twenty times the appended fine-tunes', erodes the memorized fit on the three cohorts where it was run, PBMC-33k, TCGA-BRCA, and PBMC-68k cluster 9. The rate is set by how strongly the retained data act on the forget region. On PBMC-68k cluster 9 it brings the forget cells into the spread of the retrained models, with the type still known and the held-out utility inside its margin. Its paired difference from the retrained arm is 0.008, on an interval of -0.062 to 0.080 that, like every method arm's interval on that cohort, is too wide for the design to certify arrival (Section 3). On PBMC-33k, where the design has that power, the level is not reached; the same fine-tune ends 0.039 above the retrained arm, a small resolved excess inside the margin that the test reports as more identifiable than retrained.

Nothing a controller can observe marks arrival. The forget cells' fit, the fit of their nearest retained neighbors, the retained cells' fit, the held-out utility, and a calibration on members held back from the deletion all move smoothly through the point where the retrain-based verdict turns. The calibration then rises past the retrained models' level and keeps rising. The methods do not lack reach or utility. They lack a stopping point that can be identified from the model alone, and an unlearning claim about such a model cannot be verified without a retrained reference.

Two measurements explain the pattern. The first concerns what a retain fine-tune does to an edit. By gradient descent it reverts the part of the edit that lies in the directions the retain data constrain and leaves the rest in the parameters. The forget cells' fit returns either way, because the fine-tune's own path through the parameters it reaches compensates for what it does not revert. In a shared decoder almost every direction the forget cells use is one the retain cells use too, so an edit weighted by forget-set importance damages both sets, and the repair of the retain set restores the forget set with it. On PBMC-68k the forget type passes through decoder hidden units the retain set barely uses, and there the same edit outlasts the ten epochs the appended fine-tunes allow, though not fifty. In a classifier the forget class owns its output row, where Proposition 1 predicts that most of an edit survives, 0.77 for the whole classifier and 0.85 for that row. No unlearning method was run on a classifier here, so that part is a prediction.

The second measurement concerns the methods that overshoot, whose objectives say nothing about where to stop. A fine-tune given the retrained models' level matches their mean fit of the forget cells, 336.6 against 337.2, while its own attacker verdict is not resolved. That level is how a model fits a cell type it has never seen, and no held-out neighbor of the forget cells supplies it. In the simulation, matching the level leaves the per-point signal in place in seven of eight configurations. Establishing any of this needed a measurement the evaluation in common use does not make. Scoring the forget cells against cells of other types measures how well the model fits the type, and a model that never saw the forget cells scores as high as the original on three cohorts and at the opposite extreme on a fourth.

Contributions.

1. The answer. At published budgets no method is shown to unlearn a rare cell type from a gene-expression VAE, by a paired comparison with retrained models on the same cells. On PBMC-33k, where the design can certify equivalence, the retrained level is not reached; on PBMC-68k no interval is narrow enough to certify it. The misses come in two kinds that the usual advantage cannot distinguish. Long retain-only fine-tuning approaches the retrained models without a utility cost the margins detect, and the design cannot certify that it arrives. None of the observable quantities tried, including a calibration on held-back members, marks arrival, so the result is not verifiable without a retrained model.

2. The reasons. A retain fine-tune reverts the part of an edit that lies in the directions the retain data constrain (Proposition 1), so what a retain fine-tune leaves of an edit in the weights is its forget Fisher mass on parameters the retain set does not use. At ten units of the proposition's own time that share orders the settings, 0.02 in the shared decoder on PBMC-33k, 0.22 where a rare type is routed through hidden units the retain set does not use, and 0.77 in a classifier. The appended fine-tunes' ten epochs are a fraction of one unit. An intervention that confines an edit to those parameters confirms it under gradient descent, and finds Adam reverting those parameters too. The forget cells' fit is nonetheless largely restored around an edit that survives. A linear-response calculation (Proposition 2) separates reversal of an edit from compensation of its damage. The saved fine-tunes show two routes to compensation, a counterweight the repair builds through parameters the edit did not touch and the retain fine-tune's own erosion of the forget type. The retain-null mass orders the residual fits without bounding them.

   The level the overshooting methods lack is the fit of a type the model does not know, which targeted fine-tuning shows is not observable. Matching that level is not sufficient either. The per-point signal survives a fine-tune given the oracle level in seven of eight simulation configurations, and on PBMC-33k the attacker verdict for such a fine-tune is not resolved.

3. The measurement (Figure 1). A reference retrained on exactly the retain split, held-out cells of the forget type, an attacker refit on standardized features, and a reference-free likelihood audit, without which an attacker's score on these data reads biology.

Section 2 sets the threat model and, in Section 2.1, the cohorts, models, and methods. Section 3 states how a method is judged, Section 4 the results at the published budgets, Section 5 what longer budgets and a controller's observables show, Section 6 the two reasons, and Section 7 what follows for a data controller.

2 Setting

Unlearning and its standard. Machine unlearning produces parameters θ' from a trained model, a forget set $\mathcal{F}$, and a retain set $\mathcal{R} = \mathcal{D} \setminus \mathcal{F}$. The standard of success is the model that training on $\mathcal{R}$ alone would have produced (Cao & Yang, 2015; Bourtoule et al., 2021). Guarantees exist for restricted model classes (Ginart et al., 2019; Guo et al., 2020; Sekhari et al., 2021; Neel et al., 2021; Izzo et al., 2021) under conditions that deep generative models do not satisfy, so an approximate method is judged empirically, by whether a membership inference attack can tell the forget samples from samples the model never saw (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022). The evaluation in common use scores the forget set against held-out samples drawn from the whole population (Golatkar et al., 2020; Kurmanji et al., 2023; Foster et al., 2024). Hayes et al. (2024) showed that an inexact method can leave the forget samples more detectable than before, and Thudi et al. (2022) argued that unlearning claims are unverifiable without an auditable definition. Leakage is reported as the attacker's AUC and as the advantage $2|\mathrm{AUC} - 0.5|$ (Yeom et al., 2018). The advantage is symmetric because an adversary can flip a prediction, and for the same reason it cannot tell a model that still fits the forget samples too well from one that now fits them too badly.

Threat model. The scope is subpopulation unlearning, the removal of a biological group such as a rare cell type or a disease subtype, rather than of arbitrary individual samples. The adversary has query access

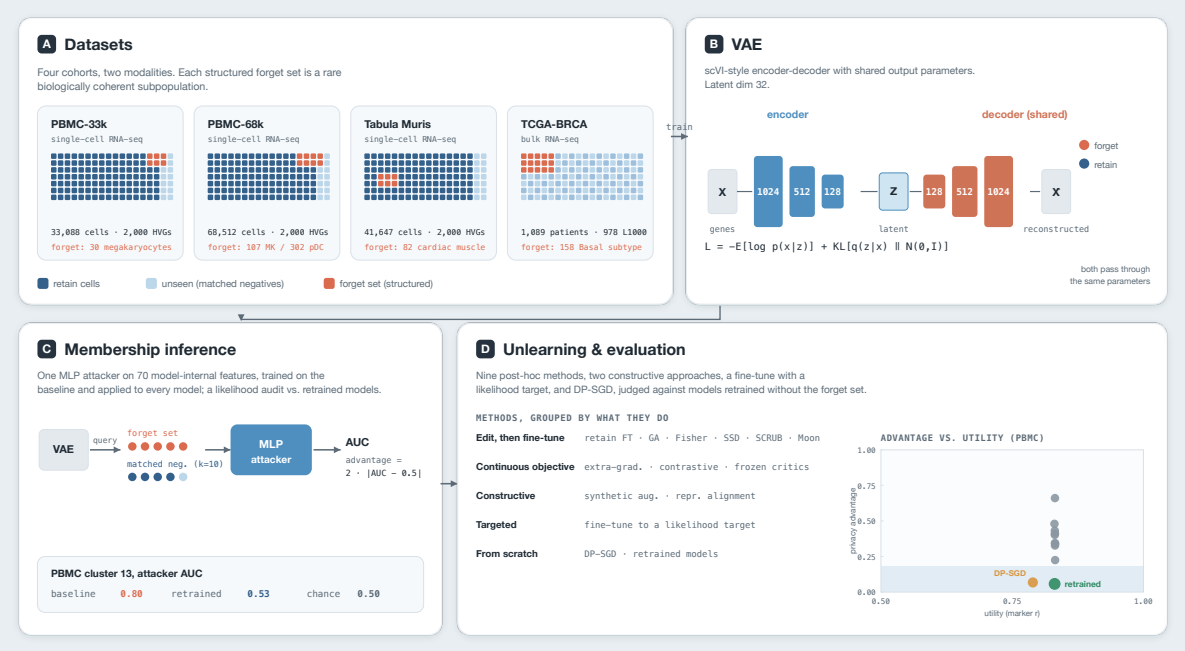


Figure 1: Evaluation pipeline. Panel (A) holds four cohorts across two modalities, each with one or two structured forget sets that are rare, biologically coherent subpopulations. Panel (B) is an scVI-style VAE with a shared decoder. Panel (C) is one attacker trained on 70 model-internal features of the baseline's forget samples against matched negatives and applied to every model, beside the reference-free likelihood audit against the mean of the retrained models. Panel (D) holds nine post-hoc methods, two constructive approaches, a fine-tune with a likelihood target, and DP-SGD, judged against models retrained without the forget set.

to the model, obtains its internal quantities for a sample (posterior mean and variance, reconstruction loss, KL term), and trains a dedicated attacker on data it controls. Membership is the quantity measured because it is the precondition for what an adversary does next. For a rare cell type or a subtype, knowing that a record was in the training set already discloses a phenotype, and it is the first step of attribute inference and of the reconstruction attacks that aim at what the model memorized (Yeom et al., 2018; Carlini et al., 2022; Hilprecht et al., 2019; Chen et al., 2020). Unlearning is therefore judged on whether the adversary can still tell the forget cells from cells the model never saw, and not on whether the model has stopped representing their type. The auditor has one thing more, a retrained reference, which a controller who turns to post-hoc unlearning does not have. Poisoning and other write-side attacks are out of scope.

Related work. Fisher scrubbing (Golatkar et al., 2020), SSD (Foster et al., 2024), and SCRUB (Kurmanji et al., 2023) assume that the forget set's influence concentrates in identifiable parameters, an assumption already known to strain in deep networks (Basu et al., 2021). Nasr et al. (2018) and Chavdarova et al. (2019) supply the adversarial and extragradient machinery of extra-gradient co-training, and Abadi et al. (2016) introduced DP-SGD. Moon et al. (2024) fine-tunes a generator's decoder along a target-feature latent direction, Pawelczyk et al. (2025) report that approximate unlearning does not remove data poisoning from classifiers, and Cheng et al. (2026) regularize retain-forget entanglement in classifiers. Work on classifiers has established that a membership score has to be calibrated to how hard a sample is to fit (Watson et al., 2022; Jayaraman et al., 2021; Rezaei & Liu, 2021; Carlini et al., 2022). Judging a method against models retrained without the forget set is the standard of the NeurIPS 2023 unlearning competition (Triantafillou et al., 2024), and membership attacks on generative models through reconstruction and likelihood go back to Hilprecht et al. (2019) and Chen et al. (2020). On biological data, Walker et al. (2024) showed that single-cell count matrices leak private information through eQTL linkage, and Golob et al. (2026) and Ozturk et al. (2026) attack transcriptomic generators through their synthetic output.

Table 1: Cohorts, splits, forget sets, and negatives. Cross-type matched negatives are the ten nearest held-out samples of each forget sample in the baseline model's latent space, deduplicated; within-type negatives are the held-out members of the forget set's own cluster or subtype. Generated from the split and matching files by `scripts/make_tex_tables.py`.

	PBMC-33k	PBMC-68k (cl. 9)	PBMC-68k (cl. 8)	Tabula Muris	TCGA-BRCA
Type	scRNA-seq	scRNA-seq	scRNA-seq	scRNA-seq	bulk RNA-seq
Samples	33,088 cells	68,512 cells	68,512 cells	41,647 cells	1,089 patients
Genes	2,000 HVG	2,000 HVG	2,000 HVG	2,000 HVG	978 L1000
Retain / forget / held out	28,094 / 30 / 4,964	57,958 / 107 / 10,447	57,763 / 302 / 10,447	35,317 / 82 / 6,248	767 / 158 / 164
Forget set	cluster 13 (30 megakaryocytes)	cluster 9 (107 megakaryocytes)	cluster 8 (302 pDC)	cluster 33 (82 cardiac muscle)	Basal subtype (158 patients)
Forget fraction of training	0.1%	0.2%	0.5%	0.2%	17%
Cross-type negatives	194	143	445	135	89
Within-type negatives	5	71	201	–	30

2.1 Cohorts, model, and methods

Four cohorts across two modalities are used (Table 1). PBMC-33k consists of 33,088 peripheral blood mononuclear cells from 10x Genomics, preprocessed with Scanpy (Wolf et al., 2018) to 2,000 highly variable genes and 14 Leiden clusters. Its forget set is cluster 13, the 30 megakaryocytes in the training split, with 5 more held out. PBMC-68k (Zheng et al., 2017) is a second donor's 68,512 cells from the same platform, processed with the same recipe into 10 clusters. Its two rare clusters are the forget sets: 178 megakaryocytes (cluster 9), and 503 cells dominated by plasmacytoid dendritic cell markers with some B and progenitor cells absorbed at this resolution (cluster 8). Each has 40 percent of its cells held out of training so that within-type negatives exist by design.

Tabula Muris (The Tabula Muris Consortium, 2018) has 41,647 cells from 12 mouse tissues in 35 clusters, with the 82 cardiac muscle cells of cluster 33 as the forget set. TCGA-BRCA (Cancer Genome Atlas Network, 2012) has 1,089 breast cancer patients with PAM50 subtypes, preprocessed by Ozturk et al. (2026) to the 978 LINCS L1000 landmark genes (Subramanian et al., 2017); the 158 Basal-subtype patients in the training split (17 percent of it) are the forget set, and 30 Basal patients are held out. PBMC-33k cluster 12 (49 cells) is a second rare cluster there. PBMC-33k cluster 7 (80 cells held out) and Tabula Muris cluster 28 (67 held out) are large enough to hold out a within-cluster negative set of useful size. Each cohort is split once into training and held-out sets, and the forget set is removed from the training set to define the retain set.

Negatives. Negatives are never-seen samples of two kinds. Cross-type matched negatives are the ten nearest held-out samples of each forget sample in the baseline model's latent space, deduplicated. This is the matching that the evaluation in common use amounts to when a whole type is forgotten, since the nearest samples of a type that is entirely in the forget set are samples of other types. Within-type negatives are the held-out members of the forget set's own type.

Model. The VAE follows scVI (Lopez et al., 2018). The encoder maps the genes through layers of 1,024, 512, and 128 units to a 32-dimensional latent mean and log-variance, and the decoder mirrors it to a softmax over genes scaled by the library size, with layer normalization and dropout of 0.1 after each hidden layer. The inputs are log-normalized counts, so the reconstruction term is the mean squared error summed over genes, a Gaussian observation model on the transformed counts, and the KL term is summed over latent dimensions. Reweighting one term relative to the other drives the posterior toward collapse, and every objective in the paper was checked for it. Training uses Adam at 10^{-4} for 100 epochs with a 20-epoch KL warm-up and free bits of 0.03. Every unlearning method starts from the same baseline checkpoint, and every reported score is computed on the CPU (Appendix A).

Unlearning methods. Table 2 lists them with the budget each was run at. Six edit the model and then fine-tune it on the retain set. Retain-only fine-tuning runs 50 epochs with patience 10. Gradient ascent takes ten clipped ascent steps and then up to 30 retain epochs with the same patience, and every published run completed the 30. Moon feature unlearning is a 20-epoch decoder fine-tune, and Fisher scrubbing, SSD, and SCRUB append 10 retain epochs to their edit. Three push the forget cells with a continuous objective that has no stopping point, and contrastive latent unlearning appends 10 retain epochs to it. DP-SGD (Abadi et al., 2016), trained from scratch on $\mathcal{R}$, is a formal baseline rather than an unlearning method. Extra-

Table 2: The methods, grouped by what they do. The budget column gives the budget each method was run with here, including the retain epochs appended to Fisher scrubbing, SSD, SCRUB, and contrastive unlearning, and Table 13 gives the rest of the settings. DP-SGD is not post-hoc; the two constructive approaches are reported in Appendix A; the likelihood-targeted fine-tune of Section 5 is a demonstration that needs a level only a retrained model supplies.

Method	Family	What it does	Budget
Retain-only fine-tuning	edit, then fine-tune	continues training on $\mathcal{R}$	50 epochs (patience 10)
Gradient ascent	edit, then fine-tune	ten clipped ascent steps on the forget loss, then retain fine-tune	10 ascent steps, then up to 30 retain epochs (patience 10)
Fisher scrubbing	edit, then fine-tune	step against the forget gradient scaled by the inverse retain Fisher, capped, then retain fine-tune	100 scrub steps, then 10 retain epochs
SSD	edit, then fine-tune	multiplicative dampening by forget-set Fisher importance, then retain fine-tune	one dampening pass, then 10 retain epochs
SCRUB	edit, then fine-tune	distills from the baseline, matching it on $\mathcal{R}$ and diverging on $\mathcal{F}$	20 distillation epochs, then 10 retain epochs
Moon feature unlearning	edit, then fine-tune	decoder fine-tune with a forget-versus-retain latent direction removed	20 decoder epochs
Contrastive latent	continuous objective	forget posteriors pushed to the prior, retain representations held	20 epochs, then 10 retain epochs
Frozen critics	continuous objective	VAE updated to defeat fixed pre-trained attackers (one or three, two weights)	50 epochs
Extra-gradient co-training	continuous objective	min-max game against three co-trained critics, extragradient updates	50 passes, $\lambda = 10$ (1)
Synthetic augmentation	constructive	trains on $\mathcal{R}$ plus synthetic forget-type cells from held-out seeds	from scratch
Representation alignment	constructive	fine-tunes the baseline until its forget-cell posteriors match a retrained model's	–
DP-SGD	from scratch	trains on $\mathcal{R}$ with $\varepsilon = 10$	from scratch
Likelihood-targeted fine-tune (Section 6.2)	demonstration	retain loss plus a penalty on the gap between each forget cell's negative ELBO and a target	until the gap stays small

Table 3: The retrained models and the controls on PBMC cluster 13. The audit column is the likelihood-ratio audit against a retrained reference; for the clean retrains it is the range over every ordered pair of the five clean retrains, each scored against another (0.17 to 0.83). Generated from `nested_ci_all.json`, `retrain_refit.json`, and `attack_lr_reference.json`.

Model	Fit-excess attacker, AUC	Advantage [nested 95% CI]	Audit AUC [95% CI]
Clean retrain, five seeds (reference)	0.509 to 0.567	0.057 [0.02, 0.27]	0.17 to 0.83 (every ordered pair of the five retrains)
Retrain that drew its own split (superseded)	0.571 to 0.585	0.158 [0.02, 0.37]	0.799 [0.72, 0.87]
Baseline	0.801	0.601	0.996 [0.99, 1.00]
Refit attacker on the clean retrain (20 x 5-fold)	0.507	0.035 [-0.03, 0.10]	–
Refit attacker on the baseline (positive control)	0.797	0.594	–
Permutation null of the advantage estimator	–	0.091 [0.004, 0.25]	–

gradient co-training frames unlearning as the min-max game $\min_\theta \max_\psi \mathcal{L}_{\mathrm{VAE}}(\theta; \mathcal{R}) - \lambda \mathcal{L}_{\mathrm{att}}(\theta, \psi; \mathcal{F})$, solved with the extragradient update of Chavdarova et al. (2019) against three critics trained alongside the model.

3 How a method is judged

The retrained models. Models retrained on exactly the retain split with the baseline's recipe are the standard, five on PBMC-33k and eight per forget set on PBMC-68k (Table 3). Every method is compared with the whole arm of them, never with one, and every model is scored on the same cells, so a difference between a method and the arm is paired on the cells. The retrained models score between 0.509 and 0.567 under the PBMC-33k attacker, near chance. What that chance score means depends on the attacker (Appendix B).

The attacker. On PBMC-33k the attacker is a multilayer perceptron with two hidden layers of 256 units, spectral normalization, and dropout of 0.3, trained for 50 epochs on 70 features of a sample under the model. The features are its posterior mean and log-variance, its reconstruction loss, KL term, and negative ELBO, the norms of the two posterior vectors, and the mean distance from its latent code to its five nearest neighbors among 5,000 retain cells. It is trained once, on the baseline model's forget samples against the matched negatives, and applied unchanged to every checkpoint in one pass. A freshly trained instance moves the baseline's AUC by 0.001 to 0.02, so every number in a table comes from one instance. Because it is trained on the baseline's forget cells and the matched negatives and then scored on those same cells, its scores are in-sample with respect to the cells. The control of Appendix B, which refits it out of fold, scores the baseline at 0.797 against the transferred 0.801, so the in-sample effect is small.

This attacker takes the features at their raw scale and reads little beyond the reconstruction term, so against cross-type negatives it measures the model's excess fit of the forget cells over a retrained model's. A logistic regression on the same 70 features standardized separates the megakaryocytes from their cross-type neighbors at 0.990 on the baseline and 0.984 on a retrained model (Table 21, 10 fold splits; 0.988 and 0.984 over the 20 splits of Table 20). The 32 posterior means alone give the same two values. The encoder places a megakaryocyte in its own region of the latent space whether or not the model trained on one. Where

held-out cells of the forget type exist, on PBMC-68k, the instrument is therefore that standardized logistic regression, refit on each model's own features out of fold and scoring the forget cells against the held-out cells of their type. It asks about the individual cells and not about the type.

The criterion. A method has unlearned the forget cells when the attacker, compared with the retrained models on the same cells, identifies them no better and no worse than on those models, to within a margin. The statistic is the attacker's AUC. The difference is the method arm's mean minus the retrained arm's mean, and its 95 percent interval resamples the seeds of each arm and, once for every model, the cells they share. The margin is half the lower end of the interval of the baseline's own excess over the retrained arm, the smallest excess the data support, so that the margin does not inherit the luck of a point estimate. On PBMC-33k the baseline's excess is 0.272 (interval 0.214 to 0.334) and the margin 0.107; on PBMC-68k the excess is 0.168 (interval 0.087 to 0.247) on cluster 9 and 0.117 (interval 0.075 to 0.159) on cluster 8, and the margins 0.044 and 0.038. Four verdicts follow. A method is more identifiable than retrained when the interval of the difference lies above zero, less identifiable than retrained when it lies below zero, at the retrained level when the whole interval lies within plus or minus the margin, and not resolved otherwise. The verdicts are read in that order. A difference whose interval excludes zero is reported as more or less identifiable than retrained even when the whole interval lies inside the margin; the verdict at the retrained level is returned only for an interval that includes zero and lies within the margin.

Each reading of the attacker test is a one-sided test at level 0.025, the two sides of the equivalence window included. Its p-values, and the audit's two-sided p-values at 0.05, are Holm-adjusted within families: on PBMC-33k the fourteen published settings, the nine ablations, and the four demonstrations; on PBMC-68k the eight method arms of cluster 9 and the seven of cluster 8. Table 34 of Appendix G collects them, and a verdict that does not survive the adjustment is said so. Body tables that report an attacker's AUC level also report the advantage $2|\mathrm{AUC} - 0.5|$; the appendix tables of paired differences omit it, because the sign of a difference of signed AUCs carries what the advantage would fold away.

The rule of half the excess is a convention with a history. Its first form, half the point excess, was fixed before the PBMC-68k scores existed and was carried to PBMC-33k after that cohort's intervals had been seen. The present form, half the lower bound of the excess on the signed AUC, was adopted after every score on both cohorts existed, when a review found that the point form let a verdict rest on a marginal pass. It is tighter on every cohort, the superseded files are kept beside the current ones, and no positive verdict in the paper rests on a method's distance from the margin.

What the design can certify. The verdict at the retrained level needs the whole interval inside plus or minus the margin, so an interval wider than twice the margin cannot return it whatever the data say. That window is 0.214 on PBMC-33k, 0.087 on cluster 9, and 0.075 on cluster 8 (`equivalence_reachability.json`). The window is twice the unrounded margin, 0.0437 on cluster 9 and 0.0376 on cluster 8, so doubling the margins as printed does not reproduce it. On PBMC-33k 20 of 27 method arms have intervals narrower than the window, so there a not-resolved verdict is a reading of the data for those arms. The seven arms that are too wide are the four frozen-critic arms and the three targeted fine-tunes, among them the one given the retrained models' level (Section 5).

On PBMC-68k every method arm's interval is wider than the window, 0.133 to 0.152 on cluster 9 and 0.079 to 0.147 on cluster 8, and so is every fine-tune snapshot of the stopping probe (Appendix D). No method arm at a published budget and no fine-tune snapshot on PBMC-68k could have read at the retrained level, even with a true difference of zero. Only the two largest Fisher caps of the stopping probe on cluster 8 have intervals narrow enough, 0.073 against the window of 0.075, and both read more identifiable than retrained. On that cohort a not-resolved verdict is the design's resolution for every method arm and every fine-tune snapshot, and not a reading of the data. The paper therefore says that such a method is not certified at the retrained level (the design cannot certify equivalence at this margin), never that it is not at it.

The widths are set by three seeds per method arm against eight retrained models and by the cell counts, 107 against 71 and 302 against 201; the margin is set by the baseline's excess. For the arms nearest the retrained level, the smallest margin at which the equivalence verdict would have been returned is 0.080 for the 200-epoch fine-tune on cluster 9, against the margin of 0.044. On PBMC-33k it is 0.110 and 0.108 for

the Fisher caps of 2×10^{-3} and 2.5×10^{-3}, against 0.107, which they miss by location and not by width. The 200-epoch fine-tune on PBMC-33k would read at the retrained level at a margin of 0.074, but its interval excludes zero, and the precedence above reports it as more identifiable than retrained.

Utility. Held-out utility is the negative ELBO on the held-out cells, the correlation between observed and reconstructed marker-gene expression, and the adjusted Rand index of a clustering of the latent means against the Leiden labels. Each carries a 95 percent interval from the same shared resampling of the held-out cells and a paired cost against the baseline. A method costs utility when the lower end of its cost interval lies above 1 percent of the baseline's negative ELBO or above 0.01 in marker correlation (Appendix A).

The audit. The audit says what the attacker cannot: whether a method that misses the retrained level does so because the forget cells are still fitted as the baseline fits them, or because they are now fitted worse than a model that never saw them. Each forget cell and each negative is scored by its negative ELBO under the retrained models, averaged over the arm, minus its negative ELBO under the model being audited, so that a cell the audited model fits better than the retrained models do scores positive. The audit AUC is that score's AUC for the forget cells against the negatives, and each retrained model is audited against the mean of the others. The audit's difference from the retrained arm carries an interval that resamples the retrained models entering the reference together with the method's seeds and the shared cells. It is read with the same four verdicts as the criterion, fitted better than retrained, fitted worse than retrained, fitted as retrained, or not resolved. Its margin is built the same way, as half the bound nearer zero of the baseline's own difference from the retrained arm, and is 0.200 on PBMC-33k, 0.170 on cluster 9, and 0.234 on cluster 8.

Beside the AUC the audit reports the paired gap, the median negative ELBO of the forget cells under the method minus that under the retrained arm, which is the type-level reading (Appendix C). The negative ELBO is estimated in evaluation mode from sixteen seeded draws of the latent code, which makes the audit reproducible to the last digit. The audit needs retrained models, so it is an auditor's instrument, not a controller's.

Why the negatives matter. Under cross-type matching, a retrained model that never saw the forget set scores far from chance on every forget set but one (Table 4): as high as the baseline where the retain set covers the forget set's type, and at the inverted extreme where it does not. On TCGA-BRCA the baseline scores 0.803 and the retrained model 0.830 against cross-subtype negatives, and against the 30 held-out Basal patients the same attacker scores them at 0.580 and 0.572. Tabula Muris cluster 33 scores 0.903 and 0.943, and PBMC-33k cluster 12 scores 0.869 and 0.814. PBMC-33k cluster 7, where 80 held-out cells of the same type are available, gives 0.665 and 0.655 under cross-type negatives and 0.466 and 0.414 within the cluster under an attacker refit for the purpose.

PBMC-68k cluster 9 shows the other face of the same confound. The retrained models score 0.007 to 0.017 against the baseline's 0.965 under cross-type negatives, because they fit megakaryocytes far worse than the neighboring cells rather than better. On each of these cohorts the cross-type attack measures biology. The within-type attack, where it can be run, finds nothing on baseline and retrained model alike on cluster 7, cluster 28, and the Basal patients (Appendix B). PBMC-33k cluster 13 is the one forget set on which the retrained models score near chance under cross-type negatives, 0.529 against the baseline's 0.801, and there the separation is the baseline's better fit of the type it trained on, which the audit reads directly. PBMC-68k is the one cohort on which the question can be put to the cells themselves.

4 Results at the published budgets

4.1 PBMC-68k, where the cells can be tested

PBMC-68k applies the protocol to a second donor with within-type negatives of useful size (Table 5; full tables in Appendix G). Two observations frame it. The first concerns what a retrained model does with a type it cannot learn. The eight clean retrains fit the 107 forget megakaryocytes and their 71 held-out siblings equally badly, at medians of 1,014 to 1,154 against the baseline's 491. The PBMC-33k retrains, by contrast, fit cluster 13 at the level of the never-seen negatives (means of 334 to 342 against the negatives' 333). The

Table 4: Attacker AUC of the baseline and the retrained model on every cohort, each with its advantage $2|\text{AUC} - 0.5|$ in parentheses, under cross-type negatives and, where a held-out set of the same type exists, within-type negatives (the standardized logistic refit). An inverted AUC yields a large advantage, which is why the AUC is reported beside it; the retrained models' large advantages here are their worse fit of the forget type, not leakage. Generated from the per-cohort result files.

Cohort and forget set	Baseline, cross	Retrained, cross	Baseline, within	Retrained, within
PBMC-33k cluster 13 (30 cells; 194 cross)	0.801 (0.60)	0.529 (0.06)	–	–
PBMC-33k cluster 12 (49 cells; 57 cross)	0.869 (0.74)	0.814 (0.63)	–	–
PBMC-33k cluster 7 (50 cells; 80 held out within)	0.665 (0.33)	0.655 (0.31)	0.466 (0.07)	0.414 (0.17)
PBMC-68k cluster 9 (107 cells; 143 cross, 71 held out within)	0.965 (0.93)	0.011 (0.98)	0.683 (0.37)	0.515 (0.03)
PBMC-68k cluster 8 (302 cells; 445 cross, 201 held out within)	0.581 (0.16)	0.229 (0.54)	0.603 (0.21)	0.487 (0.03)
Tabula Muris cluster 33 (82 cells; 135 cross)	0.903 (0.81)	0.943 (0.89)	–	–
Tabula Muris cluster 28 (50 cells; 67 held out within)	0.400 (0.20)	0.379 (0.24)	0.487 (0.03)	0.550 (0.10)
TCGA-BRCA Basal (158 patients; 89 cross, 30 within)	0.803 (0.61)	0.830 (0.66)	0.442 (0.12)	0.404 (0.19)

two forget sets are not the same kind of cell. The PBMC-33k megakaryocytes are gene-rich (a median of 1,573 detected genes and 5,242 counts against the population's 732 and 1,893), and 582 retain cells express PPBP. The PBMC-68k cluster is a low-complexity, platelet-like profile (315 genes and 776 counts against 525 and 1,292), with 362 PPBP-expressing retain cells among 57,958. How a model fits such cells unseen is not pinned down by the data. The eight seeds disagree among themselves by more than a hundred units, and the cross-type attacker registers the fact by scoring the baseline at 0.965 and the retrained models at 0.007 to 0.017. On this cluster the paired median gap carries the type-level reading.

The second observation is a per-cell signal. The standardized refit against the 71 held-out megakaryocytes finds the cells the baseline trained on at 0.683, against 0.472 to 0.538 for the eight retrained models, an excess of 0.168 (interval 0.087 to 0.247). That signal lives in the reconstruction term and the latent placement together, and the raw-feature attacker does not see it (Appendix B). The baseline therefore carries two things, knowledge of the type and a modest signal about which megakaryocytes it saw, and the adversary's question concerns the second. The margin is 0.044 on this cluster, and no method arm's interval is narrow enough for the verdict at the retrained level (Section 3), so not resolved here means that the interval includes zero, and no more.

Four methods leave the cells more identifiable than a retrained model does, two of them after the Holm correction over the eight arms. SSD (0.647, a difference of 0.132, interval 0.057 to 0.204), SCRUB (0.601, 0.086, interval 0.018 to 0.152), and Fisher scrubbing (0.603, 0.088, interval 0.017 to 0.159) are three of them, although for SCRUB and Fisher scrubbing the excess does not survive the correction. Retain-only fine-tuning (0.556), contrastive unlearning (0.558), and gradient ascent (0.582) lower the signal from the baseline's. Their intervals include zero but reach to 0.137, past the margin, so they are not resolved.

Extra-gradient unlearning, the fourth, moves the cells the wrong way. Its critics distinguish forget cells from retain cells, and the model learns to fit the forget cells as it fits retain cells, at medians of 475 to 489 and better than the baseline did, while the retain fine-tune inside the same loop erodes their held-out siblings. The per-cell refit rises to 0.877, a difference of 0.361 (interval 0.285 to 0.438). The method amplified the per-cell membership signal it was meant to remove. Its critics are an adversary who knows which cells are members, and with that knowledge the adversary has made those cells the best-fitted ones in the model.

None of the methods removes the type. The four methods that edit and then fine-tune leave the forget cells at medians of 572 to 697, fitted far better than any retrained model fits them, and under every method the held-out megakaryocytes move with the forget cells. Fisher scrubbing is the one edit its ten appended retain epochs do not undo here. After those epochs the typical forget cell sits at a median of 814 to 831, and a minority of cells sit far past any retrained model (95th percentile 1,832 to 2,543). The retain gradients reach the parameters the scrub edited more slowly, so ten epochs bring back neither the typical cell nor the tail; run to fifty epochs they do (Section 6). The paired median gap makes this a test rather than a reading off two ranges. Fisher scrubbing's forget cells sit -251 below the retrained arm's median (interval -356 to -56), while its cross-type audit, 0.653 against the arm's 0.592, is not resolved (difference interval -0.07 to 0.27), because the eight retrained models disagree about the type.

Table 5: PBMC-68k, the two forget sets, excerpted from the full tables of Appendix G. The forget median is the per-cell negative ELBO's median over the forget cells, ranged over seeds. The per-cell column is the standardized refit against the within-type negatives, out of fold, with a nested interval over seeds and the shared cells, and the advantage is $2|\text{AUC} - 0.5|$ of it. The difference column compares each arm with the eight retrained models on those cells, read as Section 3 describes at margins of 0.044 and 0.038; no arm's interval is narrow enough for the verdict at the retrained level. The audit is the reference-free cross-type audit and the Type column its paired median gap. Generated from `percell_paired_{mk,pdc}.json` and `audit_reference_free.json`.

Model	Forget median	Per-cell AUC [95% CI]	Advantage	Paired difference [95% CI]	Cells	Reference-free audit [95% CI]	Type
Cluster 9, megakaryocytes (107 forget, 71 held out)							
Baseline (trained on the cells)	491	0.68 [0.61, 0.76]	0.37	–	trained on the cells	1.000 [1.00, 1.00]	trained on the cells
Clean retrains (8)	1,014 to 1,154	0.47 to 0.54 [0.45, 0.57]	0.00 to 0.08	–	reference	0.592 [0.41, 0.66]	reference
Edit, then fine-tune (4 methods)	572 to 697	0.54 to 0.66	0.08 to 0.33	+0.041 to +0.132	not resolved (retain-FT, GA); more identifiable (SSD, SCRUB)	0.994 to 1.000	fitted better
Fisher scrubbing	814 to 831	0.59 to 0.62 [0.53, 0.68]	0.18 to 0.24	+0.088 [0.02, 0.16]	more identifiable	0.653 [0.53, 0.76]	fitted better
Contrastive latent	663 to 755	0.55 to 0.56 [0.48, 0.63]	0.10 to 0.13	+0.043 [-0.03, 0.11]	not resolved	0.971 [0.92, 1.00]	fitted better
Extra-gradient ($\lambda = 10$)	475 to 489	0.84 to 0.91 [0.81, 0.93]	0.68 to 0.83	+0.361 [0.29, 0.44]	more identifiable	0.999 [1.00, 1.00]	fitted better
Retain-only fine-tune, 200 epochs	820 to 846	0.51 to 0.54 [0.45, 0.59]	0.01 to 0.08	+0.008 [-0.06, 0.08]	not resolved	1.000 [1.00, 1.00]	fitted better
Cluster 8, plasmacytoid dendritic cells (302 forget, 201 held out)							
Baseline (trained on the cells)	532	0.60 [0.56, 0.65]	0.21	–	trained on the cells	0.999 [1.00, 1.00]	trained on the cells
Clean retrains (8)	672 to 684	0.45 to 0.50 [0.45, 0.52]	0.01 to 0.10	–	reference	0.510 [0.48, 0.53]	reference
Edit, then fine-tune (4 methods)	555 to 585	0.54 to 0.60	0.08 to 0.20	+0.062 to +0.087	more identifiable (retain-FT, GA, SSD, SCRUB)	0.994 to 0.999	fitted better
Fisher scrubbing	697 to 812	0.58 to 0.59 [0.54, 0.63]	0.17 to 0.19	+0.102 [0.06, 0.14]	more identifiable	0.489 [0.39, 0.59]	not resolved
Contrastive latent	747 to 777	0.63 [0.58, 0.68]	0.26 to 0.27	+0.145 [0.10, 0.19]	more identifiable	0.011 [0.00004, 0.04]	fitted worse
Extra-gradient ($\lambda = 10$)	492 to 508	0.70 to 0.83 [0.70, 0.83]	0.41 to 0.66	+0.283 [0.21, 0.35]	more identifiable	0.929 [0.80, 1.00]	fitted better

Cluster 8, whose type the retain set partly covers, has retrained medians of 672 to 684 against the baseline's 532 and three times the cells, so the per-cell comparison is sharper. The baseline scores 0.603 against 0.450 to 0.503 for the retrained models (an excess of 0.117, interval 0.075 to 0.159, and a margin of 0.038). Every method leaves the cells more identifiable than the retrained models do, every excess surviving a Holm correction. The four methods that edit and then fine-tune sit at 0.548 to 0.573, differences of 0.062 to 0.087 whose intervals begin at 0.022 to 0.044, with the type still known. Fisher scrubbing leaves the cells nearly where the baseline had them (0.589, a difference of 0.102). Its audit, 0.489 against the arm's 0.510, and its median gap, 60 (interval −6 to 226), read fitted as retrained and not resolved. The typical cell is near the retrained level with a damaged tail, at a held-out ELBO cost of 9.41 nats, the one utility cost on this cohort. Contrastive unlearning pushes the type past the retrained level (audit 0.011) and leaves the cells more identifiable than retrained at 0.632, since cells fitted worse than their held-out siblings are as recognizable as cells fitted better. Extra-gradient again amplifies the per-cell signal, to 0.769.

4.2 PBMC-33k, where only the type can be tested

On PBMC-33k the 5 held-out megakaryocytes cannot support a per-cell question, and the attacker reads the model's excess fit of the type over a retrained model's (Section 3); the verdicts here concern that excess. Table 6 gives every method, with the baseline's excess at 0.272 (interval 0.214 to 0.334) and the margin at 0.107. No post-hoc method is at the retrained level, and the failures come in two kinds that a Holm correction over the fourteen published settings leaves intact. Retain-only fine-tuning, gradient ascent, SSD, SCRUB, Moon feature unlearning, and Fisher scrubbing at its default step leave the forget cells more identifiable than retrained. The attacker scores them between 0.669 and 0.740 against the retrained models' 0.529, and the audit between 0.980 and 0.996, fitted better than retrained, against the baseline's 0.801 and 0.995. Their forget cells are fitted as well as the baseline fits them, and the small drop in the attacker's score is the fine-tune's own drift. Gradient ascent belongs with them because at the paper's setting its ascent phase does not move the forget loss at all (Section 5).

Extra-gradient at $\lambda = 1$, contrastive latent unlearning, and the frozen critics leave the forget cells less identifiable than retrained, the attacker between 0.003 and 0.336 and the audit between 0.003 and 0.044, fitted worse than a model that never saw them. Extra-gradient at $\lambda = 10$ is less identifiable under the attacker (0.411, a difference of −0.117, interval −0.187 to −0.050) while its audit is not resolved, at 0.250 with a difference interval of −0.56 to 0.42. Its ten seeds split, seven pushed past the retrained models and

Table 6: Every method on PBMC-33k cluster 13 (30 megakaryocytes against 194 matched negatives). One attacker, trained on the baseline, scores every checkpoint. The paired difference compares each arm with the five retrained models on the shared cells and is read as Section 3 describes, at the margin of 0.107. The audit is the reference-free audit against the mean of the retrained models with its paired verdict; ELBO and marker r are held-out utility. Intervals resample the cells once for every model and the seeds within each; the two Fisher step caps are one seed each. The advantage is the mean over seeds of each seed's $2|\text{AUC} - 0.5|$; for extra-gradient at $\lambda = 10$ it exceeds $2|\text{AUC} - 0.5|$ of the arm's mean AUC because its seeds fall on both sides of 0.5. Generated from `paired_diff_33k.json`, `audit_reference_free.json`, and `utility_intervals.json`.

Method	Seeds	Attacker AUC [nested 95% CI]	Advantage	Paired difference [95% CI]	Attacker	Reference-free audit [95% CI]	Audit	ELBO	Marker r
Baseline (no unlearning)	1	0.801 [0.70, 0.89]	0.601	+0.272 [0.21, 0.33]	trained on the cells	0.995 [0.99, 1.00]	trained on the cells	364.0	0.831
Clean retrain	5	0.529 [0.42, 0.63]	0.057	–	reference	0.507 [0.40, 0.60]	reference	364.2	0.831
Retain-only fine-tune	5	0.669 [0.56, 0.77]	0.339	+0.141 [0.10, 0.19]	more identifiable	0.988 [0.93, 1.00]	fitted better	363.0	0.832
Gradient ascent + fine-tune	5	0.706 [0.59, 0.81]	0.412	+0.177 [0.14, 0.22]	more identifiable	0.995 [0.99, 1.00]	fitted better	363.2	0.831
SSD	3	0.701 [0.59, 0.80]	0.402	+0.172 [0.13, 0.22]	more identifiable	0.995 [0.99, 1.00]	fitted better	363.6	0.831
SCRUB	3	0.716 [0.60, 0.82]	0.431	+0.187 [0.14, 0.24]	more identifiable	0.980 [0.92, 1.00]	fitted better	363.6	0.831
Moon feature unlearning	3	0.740 [0.63, 0.84]	0.480	+0.211 [0.16, 0.26]	more identifiable	0.996 [0.99, 1.00]	fitted better	364.5	0.831
Fisher scrubbing	3	0.673 [0.56, 0.77]	0.345	+0.144 [0.10, 0.19]	more identifiable	0.992 [0.98, 1.00]	fitted better	363.6	0.832
Fisher, step cap 3×10^{-3}	1	0.312 [0.21, 0.42]	0.376	-0.217 [-0.30, -0.14]	less identifiable	0.141 [0.03, 0.27]	fitted worse	363.8	0.831
Fisher, step cap 10^{-2}	1	0.445 [0.33, 0.56]	0.111	-0.084 [-0.14, -0.02]	less identifiable	0.288 [0.15, 0.45]	fitted worse	370.8	0.830
Extra-gradient ($\lambda = 10$)	10	0.411 [0.31, 0.52]	0.225	-0.117 [-0.19, -0.05]	less identifiable	0.250 [0.001, 0.91]	not resolved	363.8	0.832
Extra-gradient ($\lambda = 1$)	10	0.336 [0.23, 0.44]	0.327	-0.192 [-0.28, -0.12]	less identifiable	0.044 [0.0003, 0.20]	fitted worse	363.2	0.831
Contrastive latent	3	0.170 [0.11, 0.24]	0.661	-0.359 [-0.44, -0.28]	less identifiable	0.003 [0.00, 0.01]	fitted worse	363.7	0.831
Frozen critics (4 configurations)	3 each	0.003 to 0.008	0.985 to 0.995	-0.526 to -0.521	less identifiable	0.003 to 0.006	fitted worse	363.7 to 404.4	0.831 to 0.832
DP-SGD ($\varepsilon = 10$, from scratch)	3	0.466 [0.37, 0.56]	0.068	-0.063 [-0.12, -0.01]	less identifiable	0.533 [0.43, 0.64]	as retrained	402.9	0.789

three still partly fitted as before. DP-SGD, trained from scratch on the retain data, is the one model below the retrained arm without an objective that pushes the forget cells, at 0.466 (a difference of -0.063, interval -0.120 to -0.011, not after the Holm correction). Its audit is 0.533, fitted as retrained (difference interval -0.12 to 0.19). The two Fisher step caps of the original ablation, one seed each, are both less identifiable than retrained, the larger cap not after the Holm correction over the ablation family.

Neither kind costs utility by the margin of Section 3; the models that cost utility are DP-SGD, the three-critic frozen runs, and the Fisher checkpoint at ten times the default cap (Table 14). The first kind is not by itself a privacy failure. It means the model still fits the type as well as before, which is all the attacker can see on this cohort; whether the individual cells remain identifiable is a question PBMC-68k answers and PBMC-33k cannot. The advantage alone would not have told the two kinds apart, since the second kind's advantages are as large as the first's. The frozen critics, at 0.985 to 0.995, would rank as the worst methods in the table by advantage, when what they have done is push the forget cells to the least member-like corner of the data. Figure 2 places every checkpoint by the forget cells' negative ELBO per cell against the audit. The baseline fits the forget cells at a mean of 270 and the retrained models at means of 334 to 342, the level at which the never-seen negatives are fitted (333). The first kind sits at means between 292 and 306, the second from 362 upward, and the frozen critics above 1,000; nothing in any post-hoc objective encodes the level a retrained model reaches.

5 Beyond the published budgets: approached, not verifiable

The retain fine-tune, run on. A retain-only fine-tune run to 200 epochs, four times the published retain fine-tune's budget and twenty times the appended fine-tunes', erodes the memorized fit at a rate set by what the retain data do to the forget region (Figure 3; Tables 36 and 37 in Appendix G). On PBMC-33k the forget cells move from a mean of 270 to means of 328 to 331 across seeds, against the retrained models' 334 to 342. The canonical attacker of the paired test scores the 200-epoch models at 0.568 against the retrained models' 0.529, a difference of 0.039 (interval 0.009 to 0.074). That is a small excess, resolved above zero, that lies inside the margin of 0.107; under the precedence of Section 3 it reads more identifiable than retrained. The reference-free audit reads 0.688 (interval 0.48 to 0.88), not resolved against the arm. Held-out ELBO and marker correlation stay at the baseline's values throughout. The retain sample's fit improves from 361.9 to between 354.5 and 354.7 while the held-out fit moves from 362.9 to 362.1, so the long fine-tune overfits the retain set; whether that raises the retain cells' exposure was not tested.

On PBMC-68k the megakaryocytes drift from a median of 491 to medians of 820 to 846 over 200 epochs, still short of the retrained models' 1,014 to 1,154 (a paired median gap of -247, interval -290 to -205), with the

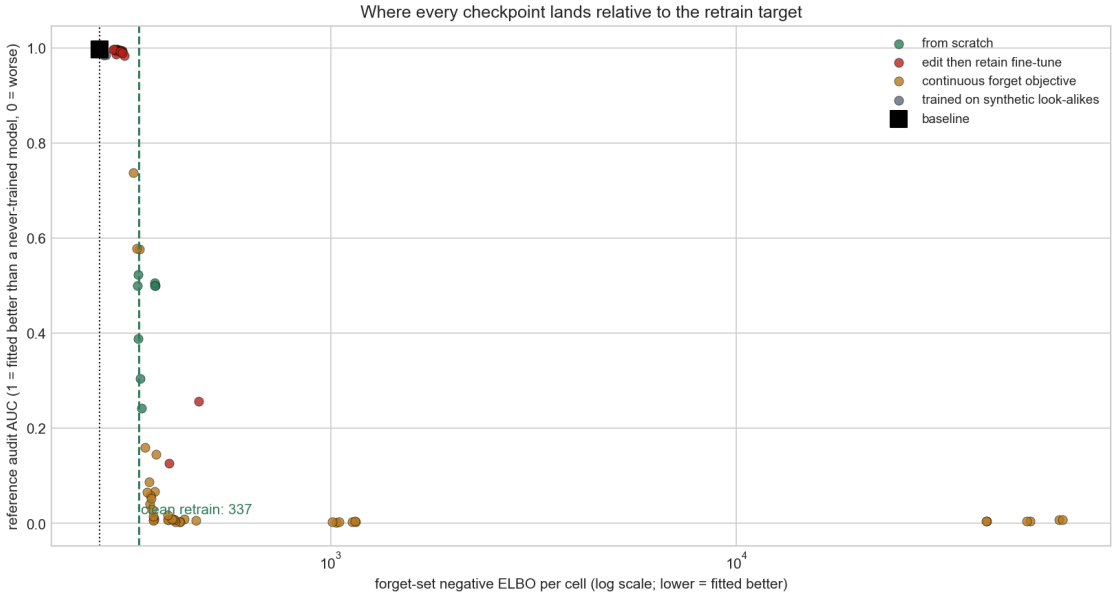


Figure 2: Every checkpoint on PBMC-33k cluster 13 by the forget cells' negative ELBO per cell (log scale) and the audit AUC. The dashed line is the retrained models' level. Methods that edit and then fine-tune (red) sit just right of the baseline (black square), still fitted better than a never-trained model fits them; methods with a continuous forget objective (gold) sit past the line; only the models trained from scratch (green) sit on it.

type fully known. The per-cell refit ends at 0.524, a difference from the retrained models of 0.008 (interval −0.062 to 0.080). The interval covers zero and reaches past the margin on both sides, so the fine-tune is not resolved. Its point estimate lies inside the spread of the retrained models, but its interval, 0.143 wide against a window of 0.087, is too wide for the design to certify it at the retrained level (Section 3). The held-out utility cost is 2.19 nats (interval 1.71 to 2.67) and 0.0069 in marker correlation (interval 0.0037 to 0.0099), inside the margins.

On TCGA-BRCA the same fine-tune takes the Basal patients to the retrained model's level within ten epochs, because nothing in that retain set holds the subtype in place. On cluster 9 a long enough fine-tune brings the per-cell signal into the spread of the retrained models before it removes what the model knows about their kind. The cross-type attacker, which reads the type, passes 0.5 near epoch 100 on PBMC-68k, but the retrained models score 0.007 to 0.017 under it and the audit still reads 1.000 there. A controller who stopped when that attacker reached chance would stop at a model that still knows the type, and the per-cell verdict offers no observable stopping point either (Appendix D).

The Fisher step, swept. Fisher scrubbing's step cap is a trust region on the scrub update, rescaled at each of the 100 scrub steps so that its largest per-parameter element is at most the cap; the default is 10^{-3}. The paper's gradient ascent takes ten ascent steps at 10^{-5} with the gradient norm clipped to 1.0, which leaves the forget loss unchanged, and raising the step to 10^{-2} changes nothing (Table 7). Ten clipped steps cannot move a model of seven million parameters far enough that up to thirty retain epochs fail to bring it back.

Fisher scrubbing is a different case. Its default cap leaves the forget cells fitted as before, its first ablation at 3×10^{-3} overshoots, and the caps between them cross the retrained models. At 1.5×10^{-3} the cells are more identifiable than retrained (0.133, interval 0.086 to 0.184). At 2×10^{-3} the difference is 0.049 (interval −0.011 to 0.110) and at 2.5×10^{-3} it is −0.044 (interval −0.108 to 0.019), one interval reaching just past the margin above and the other just past it below, so both are not resolved. Both intervals are narrow enough for the verdict at the retrained level and miss it by location (Section 3). At 3×10^{-3} the cells are

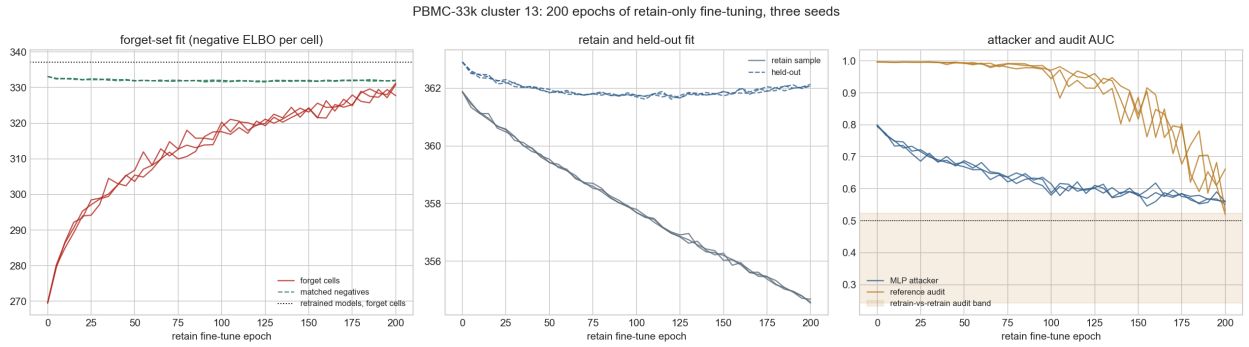


Figure 3: Retain-only fine-tuning of the PBMC-33k baseline for 200 epochs, three seeds. Left, the forget cells' negative ELBO against the matched negatives' and the retrained models' level; middle, the retain-sample and held-out fit. Right, the attacker and the audit of the notebook's own run (its own attacker instance, and the audit against retrain seed 42), with the range the retrained models span shaded.

Table 7: Gradient ascent with a larger ascent step and Fisher scrubbing with a step cap between the default and the first ablation, three seeds each on PBMC-33k cluster 13. The audit uses the mean of the five clean retrains as the reference, and the column beside it, labeled in range, counts, of the three seeds, those whose audit lies inside the range spanned by every ordered pair of the five clean retrains (0.17 to 0.83, Table 3). The forget median is the per-cell negative ELBO's median over the forget cells. The attacker is the canonical instance of Table 6, with the advantage $2|\text{AUC} - 0.5|$ and the paired difference from the retrained models with its verdict at the margin of 0.107. The baseline row, the positive control, is one model labeled trained on the cells, and the retrained row is the five clean retrains. Held-out ELBO and marker r come from `utility_intervals.json`, the source of Table 6. Generated from `sweep_scores.json`, `paired_diff_33k.json`, and `utility_intervals.json`.

Configuration	Forget median	Audit	In range	Attacker AUC	Advantage	Difference [95% CI]	Verdict	ELBO	Marker r
Gradient ascent, ascent step 10^{-4}	274 to 275	0.994 to 0.996	0 of 3	0.701 to 0.711	0.40 to 0.42	+0.178 [0.14, 0.22]	more identifiable	363.2	0.832
Gradient ascent, ascent step 10^{-3}	274 to 275	0.994 to 0.995	0 of 3	0.703 to 0.711	0.41 to 0.42	+0.177 [0.13, 0.22]	more identifiable	363.2	0.832
Gradient ascent, ascent step 10^{-2}	274 to 276	0.994 to 0.995	0 of 3	0.692 to 0.705	0.38 to 0.41	+0.170 [0.13, 0.22]	more identifiable	363.2	0.832
Fisher scrubbing, step cap 1.5×10^{-3}	296 to 313	0.793 to 0.991	1 of 3	0.641 to 0.677	0.28 to 0.35	+0.133 [0.09, 0.18]	more identifiable	363.5	0.831
Fisher scrubbing, step cap 2×10^{-3}	317 to 335	0.474 to 0.751	3 of 3	0.556 to 0.615	0.11 to 0.23	+0.049 [-0.01, 0.11]	not resolved	363.6	0.831
Fisher scrubbing, step cap 2.5×10^{-3}	339 to 363	0.242 to 0.441	3 of 3	0.461 to 0.520	0.04 to 0.08	-0.044 [-0.11, 0.02]	not resolved	363.7	0.831
Fisher scrubbing, step cap 3×10^{-3}	369 to 398	0.141 to 0.196	2 of 3	0.324 to 0.435	0.13 to 0.35	-0.150 [-0.24, -0.07]	less identifiable	363.8	0.831
Baseline	247	0.995	–	0.801	0.60	+0.272 [0.21, 0.33]	trained on the cells	364.0	0.831
Clean retrains (leave-one-out)	311 to 326	0.243 to 0.741	–	0.509 to 0.567	0.02 to 0.13	–	reference	364.2	0.831

less identifiable than retrained (−0.150, interval −0.243 to −0.072). The forget median moves with the cap, from 296 to 313 at the smallest to 369 to 398 at the largest, through the retrained models' medians of 311 to 326; what a controller cannot observe is the level at which to stop. At every cap the ten retain epochs restore the retain fit to 362.3 to 362.7 against the baseline's 363.0, and the held-out ELBO ends at 363.5 to 363.8 with the marker correlation at 0.831. The cap changes how much of the forget cells' fit the repair brings back, and only an instrument that contains a retrained model measures that.

A fine-tune given the level. Every method of the second kind pushes the forget cells with an objective that has no stopping point: a critic to defeat, a prior to reach, or a divergence to enlarge. Giving the objective the level a retrained model reaches is the obvious remedy, and a targeted fine-tune tests it. It adds to the retain loss a two-sided penalty on the gap between each forget cell's negative ELBO and a target, stops when the gap has stayed small for three epochs, and keeps the checkpoint with the smallest gap. Three targets were tried, ten seeds each (Table 8; Figure 8). The first is the model's own fit of each forget cell's ten nearest matched negatives, the never-seen cells the evaluation itself uses. The second is the model's fit averaged over all 194 matched negatives, a level a controller can compute, which on this cohort happens to coincide with the retrained models' level. The third is the model's fit of 88 held-out neighbors of the forget cells that the evaluation never uses.

Table 8: Likelihood-targeted unlearning on PBMC-33k cluster 13 with three targets, ten seeds each, alongside the baseline and the retrained models. The forget fit is the mean negative ELBO over the forget cells. The AUC, advantage and paired difference come from the canonical attacker of Table 6, with the attacker's verdict beside them; the audit is the reference-free audit against the mean of the five clean retrains, with its interval and paired verdict (Appendix C). The second target is the mean over the 194 matched negatives, which coincides with the retrained models' level; the third is 88 held-out neighbors the evaluation never uses. Held-out ELBO and marker r come from `utility_intervals.json`, the source of Table 6. Generated from `targeted_unlearning.json`, `paired_diff_33k.json`, `audit_reference_free.json`, `utility_intervals.json`, and `calibration_levels.json`.

Model	Forget fit	AUC	Advantage [95% CI]	Difference [95% CI]	Verdict	Audit [95% CI]	Verdict	ELBO	Marker r
Baseline	269.5	0.801	0.601	+0.272 [0.21, 0.33]	trained on the cells	0.995 [0.99, 1.00]	trained on the cells	364.0	0.831
Clean retrains (five seeds)	337.2	0.529	0.057 [0.02, 0.27]	–	reference	0.507 [0.40, 0.60]	reference	364.2	0.831
Target: ten nearest negatives	308.7	0.752	0.503 [0.38, 0.62]	+0.223 [0.12, 0.33]	more identifiable	0.672 [0.48, 0.84]	not resolved	364.8	0.830
Target: mean over the 194 negatives (retrained level)	336.6	0.580	0.161 [0.05, 0.29]	+0.052 [-0.07, 0.17]	not resolved	0.416 [0.24, 0.60]	not resolved	365.0	0.830
Target: 88 disjoint held-out neighbors	295.7	0.834	0.669 [0.57, 0.76]	+0.306 [0.20, 0.42]	more identifiable	0.744 [0.58, 0.89]	fitted better	364.8	0.830

Given the retrained models' level, the forget cells end at a mean of 336.6, matching the retrained models' mean fit of 337.2, while the attacker verdict is not resolved (0.580, a difference of 0.052, interval −0.066 to 0.174). That interval, 0.241 wide, exceeds the window of 0.214, so this not-resolved verdict is the design's resolution too (Section 3). The reference-free audit reads 0.416, not resolved, and held-out utility stays inside its margin. Given either level a controller can observe, the fine-tune lands short, at means of 309 with the nearest negatives and 296 with the disjoint neighbors, more identifiable than retrained (0.223, interval 0.118 to 0.334, and 0.306, interval 0.196 to 0.421). The reason is that the retrained models fit the forget megakaryocytes at means of 334 to 342 and their nearest never-seen neighbors at 298 to 301, better than the forget cells but worse than the baseline's 287 (Table 38). A model without the subpopulation fits an unseen member of it worse than that member's nearest known-type neighbors, so no cell the model knows supplies the retrained models' level, which is the fit of a type the model does not know.

Matching that level may also leave a residual that this cohort cannot resolve. The models given it still score 0.580 under the attacker against the retrained models' 0.529, a difference the test leaves not resolved and the audit does not see. The attacker reads seventy features of a cell, most of them its latent placement, and matching the likelihood leaves the encoder's placement of the forget cells where the baseline put it. Where the per-cell question can be asked, in the simulation, a model given the retrained level keeps the per-point signal. The within-component attacker scores such models at 0.59 to 0.84 in seven configurations and 0.49 in the eighth, against 0.45 to 0.55 on retrained ones (Table 39, Appendix F).

What a controller can observe. Two probes ask whether any quantity a controller can compute without a retrained model changes where the retrain-based verdict changes (Appendix D). On PBMC-68k the Fisher cap has no working setting at all. The published cap leaves the cells more identifiable than retrained on both clusters, and every larger cap takes the forget cells to medians from 700 to 166,282 across seeds and the held-out ELBO from 484 to between 504 and 1,750, which a controller can see. The route that exists on PBMC-33k, a cap between too little and too much, does not exist where the decoder-output overlap is near zero (Section 6). Along the fine-tune (the probe's own run of it; Appendix D) the per-cell verdict moves from more identifiable than retrained to not resolved between 25 and 50 epochs on cluster 9 and between 100 and 150 on cluster 8, and stays there. The verdict at the retrained level is out of the design's reach at every epoch (Section 3). The forget cells' fit, the fit of their nearest retained neighbors, the retain sample's fit, the cross-type negatives' fit, and the held-out utility all move smoothly and in one direction through that point.

With a fifth of the forget cells held back as members the controller knows, the natural calibration, the attacker refit on the deleted cells against the held-back members, has no fixed point. It passes the retrained models' level between 25 and 50 epochs and keeps rising, to 0.99 on cluster 9 and 0.80 on cluster 8 at 200 epochs, because the fine-tune keeps fitting the kept members better while the deleted cells drift the other way. It tells a controller that the deletion is working, not when to stop. The one observable marker that appears in that design is a plateau in the deleted cells' own fit, at the level the retrained models give them, and it appears only because the retain set then holds members of the type. It is type-level, and the per-cell test stays unresolved after it.

Table 9: Forget-retain Fisher cosine by parameter group, per-sample diagonal Fisher on PBMC-33k cluster 13 against the retain set, with the classifier and architecture controls. Generated from the NB26 to NB29 and NB35 result files.

Parameter group	Parameters	Forget–retain Fisher cosine
VAE encoder	2,642,816	0.345
VAE bottleneck	8,256	0.428
VAE decoder hidden	598,912	0.351
VAE decoder output	4,100,000	0.493
VAE, all parameters	7,349,984	0.209
Linear classifier, 14 classes (class-specific output weights)	462	0.018
Deep MLP classifier, shared hidden layers	1,091,456	0.527
Deep MLP classifier, class-specific output	1,806	0.006
VAE with $z = 8$, all parameters	7,340,720	0.348
Conditional VAE, cluster-specific output columns	–	0.000
Conditional VAE, encoder	2,642,816	0.617
TCGA-BRCA VAE, all parameters	4,208,356	0.753

6 Why the methods miss in two ways

6.1 An edit is not selective, and the retain fine-tune restores what it does not revert

Fisher scrubbing, SSD, and SCRUB edit the parameters the forget set's diagonal Fisher information marks as important, on the assumption that those parameters are not also the ones the retain set depends on, and Table 9 measures that assumption directly. Every Fisher quantity here is the empirical Fisher, the mean of per-cell squared gradients of the loss at the observed data. It is what the methods themselves compute, it can differ from the model's Fisher information (Kunstner et al., 2019), and the overlaps therefore describe the importance weights the methods act on. Forget-set and retain-set Fisher information overlap at every layer of the VAE, with cosines from 0.35 in the encoder to 0.49 in the decoder output, against 0.018 for the class-specific output weights of a linear classifier trained on the same cells (Figure 6). The contrast is architectural, not a matter of degree. A deep classifier with the same hidden layers as the encoder has an overlap of 0.53 in its shared layers, as high as any VAE layer, and 0.006 only in its class-specific outputs. What the VAE lacks is a parameter that belongs to one subpopulation, not a lower overlap in comparable layers. Nor is the overlap a matter of capacity. A VAE with an eight-dimensional latent space has a higher overlap than the 32-dimensional one, and a conditional VAE with cluster-specific decoder columns drives the overlap of those columns to zero while its encoder stays at 0.62 and it still does not unlearn (Appendix E).

What the retain fine-tune then does to an edit has a closed form in the quadratic approximation around the baseline, and it explains the three outcomes above.

Proposition 1 (Restoration) *Let θ^* minimize the retain loss $\mathcal{L}_\mathcal{R}$ and be a stationary point of the forget loss $\mathcal{L}_\mathcal{F}$, and near θ^* let both be quadratic with diagonal curvatures $F_j^\mathcal{R}$ and $F_j^\mathcal{F}$, the diagonal Fisher information of the two sets. An edit moves the parameters to $\theta^* + \delta$. Gradient descent on $\mathcal{L}_\mathcal{R}$ from $\theta^* + \delta$ with step $\eta < 1/\max_j F_j^\mathcal{R}$ leaves after T steps the residual $\delta_j(T) = \delta_j(1 - \eta F_j^\mathcal{R})^T$, and the forget loss stays above its value at θ^* by*

$$\Delta\mathcal{L}_\mathcal{F}(T) = \tfrac{1}{2}\sum_j F_j^\mathcal{F}\,\delta_j^2\,(1 - \eta F_j^\mathcal{R})^{2T}. \tag{1}$$

Proof. The retain gradient at $\theta^* + \delta$ is $F_j^\mathcal{R}\delta_j$ in coordinate j, so one step multiplies δ_j by $1 - \eta F_j^\mathcal{R}$, and the forget loss of a residual $\delta(T)$ is its quadratic form under $F^\mathcal{F}$. □

In this model every coordinate the retain data constrain decays geometrically at a rate set by its retain Fisher, and every coordinate they do not constrain, $F_j^\mathcal{R} = 0$, keeps its edit for ever. By (1), what survives a fine-tune in the weights is the forget Fisher mass on the parameters the retain set does not use. For a hypothetical edit of equal squared displacement on every coordinate, so that its damage is spread in proportion to the forget

Fisher, the share that survives T steps is the survival $S(t) = \sum_j F_j^{\mathcal{F}} e^{-tF_j^{\mathcal{R}}/\bar{F}^{\mathcal{R}}} / \sum_j F_j^{\mathcal{F}}$. Here $t = 2\eta T \bar{F}^{\mathcal{R}}$ is the fine-tuning time in units in which a typical retain coordinate decays by e^{-1}. A Fisher-damped scrub puts its displacement where the retain Fisher is small, so its own survival exceeds $S(t)$ at every t. At the step of the gradient-descent arms below, 10^{-3}, ten retain epochs are $t = 0.065$ on PBMC-33k and $t = 0.12$ on cluster 9, and $t = 10$ is reached after 1,540 and 821 epochs.

Corollary 1 (Selective parameters) *A parameter that only the forget samples exercise has $F_j^{\mathcal{R}} = 0$ and keeps any edit under retain fine-tuning; a parameter both sets exercise loses it. A classifier's output row for the forget class is of the first kind. A shared decoder has no parameter of the first kind unless the forget type is routed through hidden units the retain set does not use. For the decoder's output layer this is exact in the hidden-unit factor. The per-cell Fisher entry of output weight W_{dh} is $c_d^2 h_h^2$, with h_h the activation of hidden unit h and c_d the loss gradient at the layer's pre-activation output. That gradient is $2e_d$ for a linear head with residual e_d, and $2Ls_d(e_d - \sum_k s_k e_k)$ for the softmax head used here with library size L and softmax output s. Its mean over a set is therefore near zero for every gene whenever the cells of that set barely activate unit h; the retain-null coordinates of that layer are the columns of such units.*

Table 11 measures the survival on the four settings, with the retain Fisher from the scrub's own cache and the forget Fisher from the same estimator. On PBMC-33k the share of the forget cells' Fisher mass that ten units of the proposition's time leave uncorrected is 0.019 for the whole model and 0.011 at the decoder output. At ten epochs of that gradient descent, a fraction of one unit, the same model leaves 0.48 on PBMC-33k and 0.66 on cluster 9. The survival therefore states an ordering of the settings and a limit, not what ten epochs do. The phase audit below finds the forget fit restored within those ten epochs, faster than the diagonal rate allows; the rate is taken up at the end of this section.

On PBMC-68k cluster 8 the share is 0.056. On cluster 9 it is 0.225 for the whole model and 0.324 at the decoder output, where 0.427 of the forget cells' output-layer mass sits on the 5 percent of hidden units the retain set uses least, against 0.011 on PBMC-33k and 0.136 on cluster 8. That is where the Fisher scrub survives its ten retain epochs (Section 4.1). For the linear classifier on the same cells the survival is 0.768, and on the forget class's own output row, which carries 0.888 of the forget mass, 0.853. The overlap cosines of Table 9 are a coarser reading of the same quantity. On cluster 9 the retain-null mass of the output layer is carried by five hidden units that the retain set ranks 976th to 1014th of its 1,024 (Appendix E).

The proposition also predicts what an intervention should show, and Table 12 runs it. From the baseline, three edits of matched damage to the forget cells' fit are each followed by the methods' Adam fine-tune on the retain set for 10, 50 and 200 epochs, beside the same fine-tune with no edit. The edits are the published Fisher scrub, the same scrub confined to the directions the retain set constrains, and the same scrub confined to the retain-null directions (retain Fisher below one hundredth of its mean, 4,984,190 of 7,349,984 parameters on PBMC-33k and 4,975,282 on cluster 9). Each is run at the published scrub's damage and at a moderate damage of twice the starting forget loss. On PBMC-33k the retain-null edit is the one that costs nothing observable, since it moves the held-out fit from 362.9 to 363.3 where the retain-active edit of the same damage takes it to 405.8, and it is the one the fine-tune undoes least. At moderate damage the published and the retain-active edits take the forget cells to 572 and 582, and ten retain epochs bring them back to 292 and 290, against 284 for the fine-tune with no edit. The retain-null edit goes from 549 to 431 after ten epochs, 405 after fifty and 387 after 200, still above the retrained models' mean fit of 334 to 342, with the audit at 0.075 against 0.973 after fifty epochs. At the published scrub's damage the same holds, 426 against 309 after fifty epochs.

On PBMC-68k cluster 9 the intervention shows that the Fisher scrub's survival of its ten retain epochs (Section 4.1) is a slower restoration. Every edit is restored onto the trajectory of the fine-tune with no edit, which by itself takes the forget cells from 488 to 540, 689 and 800 after 10, 50 and 200 epochs. At moderate damage the three edits sit at 542, 541 and 551 after ten epochs and at 801, 799 and 801 after 200. At the published scrub's damage, where the edits take the forget cells to between 49,000 and 61,000, the published and the retain-active edits are back on the trajectory by fifty epochs, 679 and 681 against 689, while the retain-null edit is only part of the way, at 911. After 200 epochs the three sit at 804, 804 and 797 against 800.

After ten epochs the same three edits read 971, 832 and 4,652. Those published-damage values at ten and fifty epochs are run-dependent, differing between MPS runs by up to 3,285 nats, while every moderate-damage arm and every PBMC-33k arm agrees to within 17 nats. What the paper's tables show for Fisher scrubbing on this cluster, forget medians of 814 to 831 after the method's ten epochs against 572 to 697 for the other edits, is therefore a slower restoration and not a permanent one. That is what a survival of 0.225 rather than 0.019 predicts.

The proposition's rates are those of gradient descent. The methods fine-tune with Adam, whose normalization of each coordinate's step removes the dependence on the size of the retain curvature and keeps only the distinction between zero and non-zero. A shared decoder has no exactly-null coordinate. So on the VAE the proposition predicts slower restoration where the retain-null mass is larger and a permanently surviving edit nowhere, which is what both cohorts show. The classifier's forget-class row is the exactly-null case (Corollary 1).

The rate statement is tested directly by running the fine-tune both ways from the same saved edits (Tables 28 and 29 in Appendix E). The two fine-tunes are Adam at the methods' step and plain gradient descent at 10^{-3}, the largest step one retain epoch from the baseline leaves without a held-out cost. Both run for 10, 50 and 200 epochs beside a fine-tune with no edit. Two quantities are read at each epoch: the forget cells' fit against the no-edit fine-tune's at the same epoch, and the share ρ of the edit vector still in the parameters. That share is the projection of the displacement from the baseline onto the edit, with the no-edit fine-tune's own projection subtracted. In the parameters the proposition's ordering holds under gradient descent. The retain-null part of every edit stays, 0.91 to 0.98 of it after 200 epochs on PBMC-33k at both damage levels and 0.65 to 0.82 on cluster 9. The retain-active part is reverted, to 0.34 to 0.43 at moderate damage on both cohorts. At the published damage on PBMC-33k it is reverted past the baseline, to -0.25 to -0.30, while the retain set is still being repaired. Adam reverts both parts, leaving 0.74 to 0.77 of the retain-null edit on PBMC-33k and 0.41 to 0.53 on cluster 9. That is the zero against non-zero distinction stated above, measured.

The fit of the forget cells does not follow the edit. Under gradient descent the moderate-damage scrub's forget cells are back on the no-edit trajectory by 50 epochs, 292 against 282 on PBMC-33k and 530 against 499 on cluster 9, with 0.46 to 0.56 of the edit still in the parameters. The retain-null edit, 0.97 of which remains after 200 epochs on PBMC-33k, has lost half of its damage to the forget fit, at 429 against the no-edit 296 from a damage of 280. On cluster 9 the same edit keeps 0.82 of its coordinates and 0.06 of its damage. The retain-null edit is always the last to come back. On PBMC-33k it sits at 387 to 429 after 200 epochs against 298 to 332 for the other two edits. On cluster 9 under gradient descent it sits at 595 to 1,273 against 536 to 664, where Adam has all three edits on its trajectory at 797 to 804. So the retain-null mass of Table 11 predicts which part of an edit survives in the weights and orders the residual fits; it does not bound how much of the forget fit stays removed.

The proposition's rates are those of a diagonal curvature at a stationary point, and neither holds on the VAE. The baseline is early-stopped, so the retain gradient at it is not zero and a fine-tune with no edit moves too. The forget gradient is not zero either, since 28 of the moderate scrub's 302-nat damage on PBMC-33k and 16 of 613 on cluster 9 lie in its linear term. And the retain Fisher has off-diagonal entries that a scrub direction, built from forget gradients that overlap the retain gradients, is aligned with. What a fine-tune does to an edit under a general curvature has the same closed form, and it separates two quantities the diagonal model conflates.

Proposition 2 (Linear response of the fine-tune) *Let the retain loss near the baseline θ_0 be $\mathcal{L}_\mathcal{R}(\theta_0 + x) = \mathcal{L}_\mathcal{R}(\theta_0) + g^\top x + \frac{1}{2} x^\top K x$ with gradient g and a symmetric positive semidefinite curvature K, and let the forget loss near θ_0 have gradient $g_\mathcal{F}$ and curvature $K_\mathcal{F}$. Gradient descent on $\mathcal{L}_\mathcal{R}$ with step $\eta < 2/\lambda_{\max}(K)$ reaches after T steps from $\theta_0 + \delta$ the point $\theta_0 + \Delta_T + \delta_T$, where Δ_T is the trajectory from θ_0 with no edit and*

$$\delta_T = (I - \eta K)^T \delta. \tag{2}$$

Along an eigenvector of K with eigenvalue λ the edit is multiplied by $(1 - \eta\lambda)^T$, and the part of δ in the kernel of K survives for ever, so $\delta_T \to P_0 \delta$, the projection of the edit onto $\ker K$. *The forget loss of the*

edited model exceeds that of the no-edit model after T steps by

$$E_T = g_{\mathcal{F}}^\top \delta_T + \tfrac{1}{2}\, \delta_T^\top K_{\mathcal{F}}\, \delta_T + \delta_T^\top K_{\mathcal{F}}\, \Delta_T \tag{3}$$

to second order in the two displacements jointly.

Proof. Gradient descent on a quadratic is affine, $x_{t+1} = x_t - \eta(g + Kx_t)$, so the difference d_t between the trajectories from $\theta_0 + \delta$ and from θ_0 obeys $d_{t+1} = (I - \eta K)d_t$ with $d_0 = \delta$, which gives (2); the step bound keeps every factor $1 - \eta\lambda$ in $(-1, 1]$. Expanding $\mathcal{L}_{\mathcal{F}}(\theta_0 + \Delta_T + \delta_T) - \mathcal{L}_{\mathcal{F}}(\theta_0 + \Delta_T)$ to second order in both displacements jointly gives (3); the remainder is third order in (Δ_T, δ_T), so the equation describes the first epochs unless the forget loss is quadratic along the whole path. □

Proposition 1 is the case of a diagonal K at a point where $g = 0$ and $g_{\mathcal{F}} = 0$, so that $\Delta_T = 0$ and the linear term of (3) drops. Two consequences differ from it.

Corollary 2 (Reversal and compensation are different quantities) *(i) The share of the edit left in the parameters, $\rho_T = \delta^\top \delta_T / \|\delta\|^2 = \sum_\lambda w_\lambda (1 - \eta\lambda)^T$ with w_λ the spectral weight of δ on the eigenvalue λ of K, is set by where the edit sits in the spectrum of K, not by the diagonal of K. The quadratic form $\delta^\top K \delta$ differs from $\sum_j K_{jj}\delta_j^2$ by the off-diagonal sum $\sum_{i \neq j} K_{ij}\, \delta_i \delta_j$, which vanishes for every δ only when K is diagonal.*

(ii) The damage left in the forget fit, E_T, is a quadratic form of the residual δ_T under the forget curvature and not the forget Fisher mass on the surviving coordinates. In the diagonal model each coordinate of δ decays on its own, so δ_T keeps the coordinate support of δ and the damage is a sum over the surviving coordinates, $\frac{1}{2}\sum_j F_j^{\mathcal{F}} \delta_j(T)^2$. Under a general K the residual $(I - \eta K)^T \delta$ is a multiple of δ only when δ is an eigenvector of K, and in general it acquires components on coordinates the edit never touched. Its damage is the quadratic form $\delta_T^\top K_{\mathcal{F}}\, \delta_T$, whose cross terms the diagonal sum omits. The retain descent removes the part of the edit the retain cells see through whatever parameter directions carry it, and the residual it leaves can have ρ_T well above zero while $K_{\mathcal{F}}\, \delta_T$ is small, which is restoration of the forget fit without reversal of the edit.

(iii) An edit in $\ker K$ *is neither reverted nor compensated in this model: $\rho_T = 1$, Δ_T is unchanged by it, and its damage changes only through the coupling $\delta^\top K_{\mathcal{F}} \Delta_T$ of the no-edit path to the edit.*

Proof. (i) and (ii) read off (2) and (3) in the eigenbasis of K. (iii) For $K\delta = 0$ the retain gradient at $\theta_0 + \delta + x$ equals that at $\theta_0 + x$ for every x, so the trajectories differ by the constant δ. □

Table 30 measures the curvature along the saved moderate-damage edits, on 10,000 retain cells, for the diagonal the proposition uses, for the empirical Fisher with its off-diagonal part, and for the Hessian (Appendix E). Along the published scrub's direction on PBMC-33k the retain Fisher's quadratic form is 23.2 and the Hessian's 46.2, against 0.0045 for the diagonal. The diagonal is wrong by a factor of five thousand. A Fisher-weighted scrub direction is built from forget gradients that overlap the retain gradients, so it lies along the retain curvature, not across it. Along the retain-null edit all three are small, 0.10, 0.11 and 0.0044 over a squared norm of 39, with the diagonal still some 25 times below the other two. On cluster 9, where the edits sit at the decoder output, the Fisher exceeds the diagonal by 459 and 425 times along the published scrub and the retain-active edit, and the Hessian by 220 and 193 times. The forget curvature along an edit exceeds its diagonal by the same orders, 1,086 and 545 against 0.067 on PBMC-33k, so the forget Fisher mass of an edit is not its damage either.

The quadratic model's own prediction with the full curvature, $\delta_T = (I - \eta K)^T \delta$, is computed by a Lanczos iteration for two matrices: the empirical Fisher the methods use, and the generalized Gauss–Newton matrix, which is the model's Fisher information on the reconstruction term. The iteration is run to a depth at which its lower and upper bounds on $\hat{\rho}$ agree to 0.002 (Table 32). With the empirical Fisher of a 5,000-cell sample the published scrub on PBMC-33k splits in two. A part on eigenvalues above 1, 0.19 of its squared norm, is gone within the first epochs. A part in the matrix's kernel, $\hat{\rho} = 0.80$, is never touched and carries 0.42 of the damage. The measured fine-tune goes far past that. The share ρ falls to 0.60, 0.46 and 0.36 at 10, 50 and 200 epochs, and the damage left to 0.07, 0.03 and 0.01. On cluster 9 the empirical Fisher at the baseline predicts no change at all, $\hat{\rho} = 0.998$ with 0.97 of the damage kept, for edits whose measured ρ is 0.75, 0.56 and 0.46 with 0.20, 0.05 and 0.01 of the damage left.

The kernel is the estimator's, not the point's. The empirical Fisher of 5,000 cells has rank at most 5,000 in 7.3 million dimensions. The Gauss–Newton matrix at the same baseline puts none of the scrub on zero, and 0.54 of its squared norm on eigenvalues below 0.01 on PBMC-33k and 0.73 on cluster 9, curvature that only a long fine-tune resolves. With it the quadratic model does far better on the retain-active edits, and errs in one direction. On PBMC-33k it predicts $\hat{\rho}$ of 0.66, 0.57 and 0.51 against the measured 0.60, 0.46 and 0.36. Its residual applied to the baseline carries 0.17, 0.10 and 0.07 of the damage, where the measured residual carries 0.10, 0.05 and 0.03. Added to the no-edit state, which is the model's prediction of the measured excess, it carries 0.14, 0.08 and 0.05 against the measured 0.07, 0.03 and 0.01.

On cluster 9 it predicts 0.89, 0.75 and 0.71 against 0.75, 0.56 and 0.46; 0.53, 0.25 and 0.20 of the damage at the baseline against the residual's 0.24, 0.08 and 0.04; and 0.43, 0.16 and 0.08 on the no-edit state against the measured 0.20, 0.05 and 0.01. So the reversal of a retain-active edit and the first half of the restoration of its damage are properties of the curvature at the baseline, once the whole of it is used rather than its diagonal. The measured fine-tune reverts more of the edit than the model says at every epoch, restores the fit about twice as fast at ten epochs, and keeps restoring it where the model's restoration stalls on curvature it has not resolved. The retain-null edits the model leaves in place, predicting $\hat{\rho}$ of 0.99 on PBMC-33k with 0.95 of the damage and 0.94 on cluster 9 with 0.54, against the measured 0.97 with 0.51 and 0.82 with 0.13 (the residual applied to the baseline, Table 31). On the no-edit state it predicts 1.02 and 0.36 of the damage against 0.47 and 0.06. Those two cases are taken up below.

The quadratic model is also local. On PBMC-33k the published scrub raises the squared norm of the retain gradient from 558 to 176,000 at the edited point, a change of which 67,600 is beyond the linear response. A fine-tune that moves 46 units of squared norm in 200 epochs, eighty times the edit's squared norm, does not stay in the quadratic basin of the point it started from.

Table 31 reads the saved states of the gradient-descent arms and says what the fine-tune does instead, which differs between the cohorts. On PBMC-33k the edited run's path is the no-edit path plus a residual. The part of the path orthogonal to the edit has cosine 0.92 to 0.997 with the no-edit path, and the residual $r - \Delta_T$, with 0.60 of the edit along δ and a squared norm of 0.34 against the edit's 0.58, is what the edit became. Applied to the baseline on its own, that residual fits the forget cells at 301 after ten epochs and 277 after 200, where the surviving edit $\rho\,\delta$ alone fits them at 454 and 358 and the no-edit fine-tune at 273 and 296. Its orthogonal part, a quarter of the edit's squared norm at ten epochs, is a counterweight that cancels the forget-output effect of the surviving edit. The two are bound together. Removing the whole edit from the fine-tuned model while keeping the counterweight makes the fit worse, 325 against 297 on the forget cells and 382 against 362 held-out. Removing only the surviving part $\rho\,\delta$ leaves the forget fit at 298 and still costs the held-out fit, 370 against 362.

The retain-null edit on PBMC-33k is the extreme case. Its residual is the edit itself, cosine 0.99 and squared norm 38 of the edit's 39. Yet applied to the baseline the residual fits the forget cells at 414, where the edit alone fits them at 544. The difference is a counterweight of squared norm 0.90, a fortieth of the edit, of which 0.84 lies in the encoder; it grows with the fine-tune from 0.07 at ten epochs, and it lies nowhere near the linear-response direction. The input of the first edited layer is the data, so this is not an upstream re-routing. It is the retain repair acting through parameters the edit did not touch, which Corollary 2(ii) permits and the diagonal model cannot express. The quadratic model at the baseline does not produce it either, for an edit that changes the retain gradient by 135 of 558 in squared norm, 72 of it beyond the linear response.

On cluster 9 the route is different. The published scrub changes the retain gradient by 1.84 of 909, and the edited run's path is the no-edit path to a part in five hundred. Removing the surviving edit from the fine-tuned model lands on the no-edit trajectory to within a nat, 488.5 against 488.3 after ten epochs and 535.2 against 534.2 after 200. The surviving edit is reverted along its own direction, and what remains does less harm along the path than at the baseline, 0.028 of the damage against 0.076 after 200 epochs. The reason is that the retain fine-tune turns the megakaryocyte units down on its own. With no edit at all it takes the forget cells from 488 to 534. An edit on those units' output columns loses its effect as the units lose theirs.

Table 10: Mean negative ELBO per cell for the forget, retain, and held-out cells, with the attacker's AUC and advantage and the audit AUC, after each phase of Fisher scrubbing at two step caps and of gradient ascent; the full phase table is Table 35 in Appendix G. An inverted AUC yields a large advantage, which is why the AUC is reported beside it; the large advantages after a scrub are forget cells fitted far worse than their negatives, not leakage. Generated from `phase_audit.json`.

Stage	Forget, mean	Retain, mean	Held out, mean	Fit-excess attacker, AUC	Advantage	Audit AUC [95% CI]
Baseline	269.5	363.0	362.9	0.793	0.59	0.996 [0.99, 1.00]
Fisher scrub, default step, after scrub	5569.4	1250.1	1284.1	0.016	0.97	0.015 [0.003, 0.03]
Fisher scrub, default step, after retain fine-tune	307.6	362.4	362.4	0.666	0.33	0.987 [0.97, 1.00]
Fisher scrub, ten times the step, after scrub	216483.8	173734.5	171780.9	0.487	0.03	0.418 [0.34, 0.50]
Fisher scrub, ten times the step, after retain fine-tune	723.2	370.2	369.2	0.376	0.25	0.155 [0.05, 0.28]
Gradient ascent, after retain fine-tune	296.3	361.6	362.2	0.716	0.43	0.996 [0.99, 1.00]

That has the sign and part of the size of the coupling term of Proposition 2. With the Hessian and the measured ρ, $\rho\,\delta^\top K_{\mathcal{F}}\Delta_T$ is -11.5 nats of the -30 by which the surviving edit does less harm on the no-edit state than at the baseline after 200 epochs. The rest is beyond second order, as is the damage itself on this cluster, where the quadratic model along the edit gives 105 of the 613 nats measured at the baseline; the mechanism here rests on the path and removal measurements, not on (3). The retain-null edit there keeps 0.82 of its coordinates and 0.06 of its damage by the same route, and removing it from the fine-tuned model leaves 543 against 534.

What the retain-null mass predicts is therefore the part of an edit that survives in the weights under gradient descent, and it orders the residual fits, the retain-null edit always the last to come back. The fit of a retain-active edit is restored as the model's own curvature at the baseline predicts over the first epochs and faster thereafter, and well ahead of the edit's reversal. The fit of a retain-null edit is restored by a counterweight where the edit disturbs the retain gradient, and along with the retain fine-tune's own erosion of the forget type where it does not. Neither route is more than partly in a quadratic model at the baseline, and neither reverts the edit.

Table 10 records what happens on PBMC-33k, where the fine-tune does reach the edit. The Fisher scrub at the default step raises the forget cells' negative ELBO from 270 to 5,569, a factor of twenty, and the retain and held-out cells' from 363 to 1,250 and 1,284. The ten retain epochs that repair the retain set to 362 bring the forget cells back to 308, 38 above where they started and 26 to 34 below the retrained models' mean fit of 334 to 342. The edit is undone by the repair, through the parameters the two sets share. At ten times the step the repair restores the retain set but cannot bring the forget cells back, and the model ends with them fitted worse than a retrained model fits them, at 723. The gradient-ascent phase at the paper's setting does not move the forget cells at all, and its retain fine-tune leaves them at 296, 27 above the start and 38 to 46 below the retrained models', so that method is the retain fine-tune under another name.

6.2 The second group has no target, and the target is not observable

Every method of the second kind pushes the forget cells with an objective that has no stopping point, and Section 5 shows what happens when the objective is given the level a retrained model reaches. The overshoot comes from a missing level rather than from an unremovable signature, since a fine-tune given that level matches the retrained models' mean fit of the forget cells, 336.6 against 337.2. The level is not observable, since the retrained models fit an unseen member of the type worse than that member's nearest known-type neighbors and no cell the model knows supplies it. The level is also not sufficient. The attacker reads the latent placement of a cell as well as its likelihood, and a fine-tune that matches the likelihood leaves the placement where the baseline put it. On PBMC-33k that fine-tune's attacker verdict is not resolved, and in the simulation the per-point signal survives the oracle level in seven of eight configurations (Table 39). Two quantities have to be moved, and each instrument sees only one of them.

Where the retain set does not hold the forget type in place, the same edits move the forget cells to the retrained level on their own. On TCGA-BRCA the baseline fits the Basal patients better than the retrained model does, with the audit at 0.833 (interval 0.77 to 0.89). The four methods that edit and then fine-tune, and contrastive unlearning, bring the patients' fit to the retrained model's by the audit, between 0.40 and

Table 11: Survival of the forget Fisher mass under retain fine-tuning, by setting and parameter group (Proposition 1). $S(t)$ is the share of the forget set's diagonal Fisher mass on coordinates whose retain Fisher would leave it uncorrected after t units of fine-tuning time; $N(\tau)$ is the share on coordinates whose retain Fisher is below τ times the mean. Per-sample diagonal Fisher, damping 0, the retain Fisher from the scrub caches; the classifier is the linear control of Table 9 on the PBMC-33k latents. Generated from `retain_null_mass.json`.

Parameter group	Share of forget mass	Cosine	$S(1)$	$S(10)$	$S(100)$	$N(10^{-2})$	$N(10^{-1})$
PBMC-33k cluster 13							
Encoder	0.178	0.345	0.404	0.056	0.009	0.010	0.054
Bottleneck	0.133	0.428	0.006	0.000	0.000	0.000	0.000
Decoder hidden	0.112	0.351	0.455	0.028	0.000	0.000	0.017
Decoder output	0.577	0.493	0.030	0.011	0.008	0.009	0.011
Whole model	1.000	0.209	0.141	0.019	0.006	0.007	0.018
PBMC-68k cluster 8							
Encoder	0.190	0.546	0.395	0.067	0.016	0.018	0.068
Bottleneck	0.097	0.604	0.001	0.000	0.000	0.000	0.000
Decoder hidden	0.126	0.478	0.357	0.026	0.001	0.001	0.024
Decoder output	0.587	0.128	0.177	0.068	0.024	0.027	0.073
Whole model	1.000	0.285	0.224	0.056	0.017	0.020	0.059
PBMC-68k cluster 9							
Encoder	0.149	0.289	0.451	0.147	0.044	0.050	0.164
Bottleneck	0.121	0.502	0.003	0.000	0.000	0.000	0.000
Decoder hidden	0.129	0.439	0.363	0.062	0.015	0.017	0.069
Decoder output	0.602	0.020	0.529	0.324	0.152	0.176	0.366
Whole model	1.000	0.164	0.432	0.225	0.100	0.116	0.254
Linear classifier (33k latents)							
Forget-class output row	0.888	0.780	0.984	0.853	0.301	0.222	1.000
Other output rows	0.112	0.283	0.353	0.096	0.010	0.006	0.127
Whole classifier	1.000	0.017	0.913	0.768	0.268	0.198	0.902

0.53 against a single reference with no retrained-versus-retrained range. Gradient ascent's interval, 0.32 to 0.47, lies below 0.5, so that arm is fitted worse than the reference rather than at its level. Fisher scrubbing pushes the patients well past the level (0.19), and on Tabula Muris it does the same (0.35 against the baseline's 0.88). The difference from PBMC-33k lies in what the retain set holds. A simulation sets the rare component's distance from the common data and the fraction of it forgotten by hand, and reproduces both mechanisms. A group signal in the likelihood exists only when the retain set cannot stand in for the group, and a per-point signal in the latent placement survives every post-hoc method tried (Appendix F).

7 Discussion

What a data controller can do. None of the methods answers the adversary's question at its published budget. On the cohort where the cells can be tested, every method on the larger forget set leaves the cells more identifiable than a retrained model does. On the smaller set four do, two of them after a Holm correction, one of those more identifiable than the original, and the rest are not resolved. On the cohort where only the type can be tested, every method is either more or less identifiable than retrained, and the only models near the retrained arm were trained from scratch on the retain data.

The methods can be pushed further. A retain fine-tune of two hundred epochs, four times the published retain fine-tune's budget and twenty times the appended fine-tunes', moves the forget cells' fit toward the retrained models' without removing the model's knowledge of the type. On PBMC-68k cluster 9 it lands inside their spread, at a held-out cost of 2.19 nats and 0.0069 in marker correlation that is measurable and

Table 12: The restoration intervention. From the baseline, the published Fisher scrub and the same scrub confined to the retain-active or the retain-null directions (retain Fisher below one hundredth of its mean) are run to matched damage of the forget cells' fit. The two damage levels are the published scrub's and a moderate one of twice the starting forget loss. Each is followed by the methods' Adam fine-tune at 10^{-4} on the retain set for 10, 50 and 200 epochs, beside the same fine-tune with no edit. These are the Adam arms of the rerun from the saved edits (Table 28). The columns give the scrub steps, the forget and held-out fit after the edit, the forget fit after 10, 50 and 200 Adam epochs, the held-out fit after 200, and the audit after 50 and 200. Fits are the mean negative ELBO per cell on the forget cells and a held-out sample; the audit is the reference-free audit point. Generated from `restricted_scrub_test.json`, arms `edits_x2_opt` and `edits_opt`.

Edit	Steps	Forget, edit	Held-out, edit	Forget 10	Forget 50	Forget 200	Held-out 200	Audit 50	Audit 200
PBMC-33k cluster 13: baseline forget 269, retain 363, held-out 363 (mean negative ELBO per cell); 4,984,190 retain-null of 7,349,984 parameters									
No edit	0	269	363	284	306	329	362	0.974	0.760
Published scrub, published damage	100	5,601	1,271	298	309	331	362	0.985	0.648
Retain-active directions	86	5,684	1,343	294	308	331	362	0.985	0.664
Retain-null directions	268	6,239	371	540	426	404	362	0.074	0.066
Published scrub, moderate damage	26	572	402	292	312	332	362	0.973	0.615
Retain-active directions	26	582	406	290	310	332	362	0.984	0.640
Retain-null directions	124	549	363	431	405	387	362	0.075	0.067
PBMC-68k cluster 9: baseline forget 488, retain 481, held-out 479 (mean negative ELBO per cell); 4,975,282 retain-null of 7,349,984 parameters									
No edit	0	488	479	540	689	800	481	1.000	1.000
Published scrub, published damage	100	49,193	671	971	679	804	481	1.000	1.000
Retain-active directions	128	48,945	702	832	681	804	481	1.000	1.000
Retain-null directions	25	60,925	668	4,652	911	797	481	0.813	1.000
Published scrub, moderate damage	40	1,106	481	542	682	801	481	1.000	1.000
Retain-active directions	50	1,038	481	541	684	799	481	1.000	1.000
Retain-null directions	9	1,483	482	551	678	801	481	1.000	1.000

inside the margins; on PBMC-33k it ends 0.039 above the retrained arm. Fisher step caps between the default and its first ablation land the fit on either side of the retrained models'. But the design cannot certify that either has arrived, and on PBMC-68k it could not have. The point a fine-tune has reached is invisible to a controller without a retrained model, and the calibration a controller can compute on held-back members has no fixed point. The long fine-tune also leaves the retain cells fitted better while the held-out fit does not move, and no attack on those cells was run to say what that costs them. The methods lack a stopping point that can be identified from the model alone, and a controller who cannot retrain therefore cannot verify. An unlearning claim about a subpopulation in such a model should be treated as unverified until a retrained reference has been run against it.

The knowledge of the type that remains is not the failure; a model that learned megakaryocytes from other cells would carry it too. Where a retrained model exists the audit is cheap, and the results here argue for three reporting practices. The first is the audit beside the attacker. The second is the attacker's AUC rather than its advantage, since the advantage ranked the frozen critics as the worst methods in the study for having done the most. The third is negatives chosen to fit the question. On PBMC-68k cluster 9 the method nearest chance under cross-type negatives, extra-gradient at 0.506 to 0.669 against the retrained models' 0.007 to 0.017, was the one that had amplified the per-cell signal.

Limitations. One cohort carries a large signal, and its forget set has thirty cells with a within-cluster control of five that resolves nothing. PBMC-68k supplies the within-type negatives, but its per-cell signal is modest, and the per-cell readings rest on differences of about a tenth in AUC with 107 cells against 71 and 302 against 201. The equivalence margin is a convention, half the smallest excess the baseline is shown to have over the retrained models, and at that margin no configuration is certified at the retrained level. On PBMC-68k no method arm or fine-tune snapshot could have been, since each of their intervals is wider than twice the margin (Section 3); more method seeds or more cells of the forget type would be needed to put that verdict in reach.

The membership signal appears only where the forget set is megakaryocytes or the mixed cluster 8, and the two megakaryocyte sets differ in library complexity; whether the findings extend to other rare types is untested. The fine-tune given the retrained models' level is a demonstration and not a method. The

stopping probe's calibration statistic has no fixed point, and the one observable marker it finds, a plateau in the deleted cells' fit, is type-level and appears only when the retain set holds members of the type. Unlearning was tested on scVI-style VAEs, on a conditional VAE with cluster-specific decoder columns (Appendix E), and on Gaussian VAEs in the simulation, with the Fisher overlap measured on linear and deep classifiers and a smaller-latent VAE besides; no other generative architecture was tried. Tabula Muris carries a tissue-of-origin signal that no held-out set of cardiac muscle cells separates from membership, and the absence of a per-patient signal on TCGA-BRCA holds under the attacks tried here. A freshly trained attacker instance moves the baseline's AUC by up to 0.02, which is why every table is one instance and why differences smaller than that between tables should not be read.

8 Conclusion

No tested method is shown to unlearn a rare cell type from a gene-expression VAE at the budget its authors use. At their published budgets the methods that edit and then fine-tune leave the forget cells more identifiable than a retrained model does wherever the test resolves them, fitted much as the original model fits them on PBMC-33k and far better than any retrained model fits them on PBMC-68k cluster 9. The methods with a continuous forget objective leave them less identifiable on PBMC-33k, fitted worse than a model that never saw them wherever the audit resolves them, while on cluster 9 they too still fit the cells better than any retrained model does. Where individual cells can be tested, every method on the larger forget set and four on the smaller, two of the four after a Holm correction, leave the cells more identifiable than eight retrained models do.

Retain-only fine-tuning run to four times the published retain fine-tune's budget erodes the memorized fit. On PBMC-68k cluster 9, where the cells can be tested, it reaches the spread of the retrained models at a utility cost inside the margins, on an interval too wide for the design to certify arrival; on PBMC-33k it stays above the retrained arm, an excess the test resolves. Nothing a controller can observe marks the point: not the fit of the forget cells, not the fit of their look-alikes, not the utility, and not a calibration on members held back from the deletion, which rises past the retrained level and keeps rising.

The reasons are measurable. The retain fine-tune reverts the part of an edit that lies in the directions the retain data constrain, which in the shared decoder on PBMC-33k is almost all of it, and restores the forget cells' fit around what it does not revert. Proposition 1 predicts the opposite for a classifier's class-specific weights, a prediction no unlearning method here tested. The level a retrained model gives the forget cells is the fit of a type it has never seen, which no cell the model knows supplies. Establishing any of this needed a reference retrained on exactly the retain split and held-out cells of the forget type, because the evaluation in common use measures how well a model fits the type.

Acknowledgments

Funding. No external funding was received for this work.

Competing interests. The authors declare no competing interests.

Code. The codebase, pretrained checkpoints, and experiment scripts are available at `https://github.com/db-d2/Machine_Unlearning`.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations (ICLR)*, 2021.

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

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490: 61–70, 2012.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 463–480, 2015.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 1897–1914, 2022.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. GAN-Leaks: A taxonomy of membership inference attacks against generative models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 343–362, 2020.

Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. Machine unlearning under retain-forget entanglement. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2603.26569.

Vitaly Feldman. Does learning require memorization? A short tale about a long tail. In *ACM Symposium on Theory of Computing (STOC)*, pp. 954–959, 2020.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *AAAI Conference on Artificial Intelligence*, pp. 12043–12051, 2024.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 3513–3526, 2019.

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

Steven Golob, Patrick McKeever, Sikha Pentyala, Martine De Cock, and Jonathan Peck. Privacy vulnerabilities in synthetic single-cell RNA-sequence data. *bioRxiv*, 2026. doi: 10.64898/2026.01.22.701160.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning (ICML)*, pp. 3832–3842, 2020.

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 *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024.

Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte Carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(4): 232–249, 2019.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2008–2016, 2021.

Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting membership inference under realistic assumptions. *Proceedings on Privacy Enhancing Technologies*, 2021 (2):348–368, 2021.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Saemi Moon, Seunghyuk Cho, and Dongwoo Kim. Feature unlearning for pre-trained GANs and VAEs. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 21420–21428, 2024.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *ACM Conference on Computer and Communications Security (CCS)*, pp. 634–646, 2018.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 931–962, 2021.

Hakime Ozturk, Tejumade Afonja, Joonas Jalko, et al. Towards useful and private synthetic omics: Community benchmarking of generative models for transcriptomics data. *bioRxiv*, 2026. doi: 10.64898/2026.03.02.707794.

Martin Pawelczyk et al. Machine unlearning fails to remove data poisoning attacks. In *ICLR*, 2025.

Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 18075–18086, 2021.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18, 2017.

Aravind Subramanian et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

The Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727):367–372, 2018.

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *USENIX Security Symposium*, pp. 4007–4022, 2022.

Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, Julio Jacques Junior, Ioannis Mitliagkas, Jun Wan, Lisheng Hosoya, Sergio Escalera, Gintare Karolina Dziugaite, Peter Triantafillou, and Isabelle Guyon. Are we making progress in unlearning? findings from the first NeurIPS unlearning competition. *arXiv preprint arXiv:2406.09073*, 2024.

Conor Walker, Xiaoting Li, Manav Chakravarthy, et al. Private information leakage from single-cell count matrices. *Cell*, 187(22):6237–6252, 2024.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations (ICLR)*, 2022.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018.

Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Erik A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.

A Settings, utility, forget-set size, and the remaining cohorts

Table 13 lists every setting. Fisher scrubbing, SSD, SCRUB and contrastive unlearning are run as their authors describe them and then given the same ten retain epochs. SSD, which is training-free as published, and Fisher scrubbing, which as published has no retain phase, are therefore compared here with a fine-tune appended, which is stated wherever they are named. The retain-only fine-tune runs 50 epochs at 10^{-4} with early stopping at patience 10, and every one of its eleven published runs, five on PBMC-33k and three on each PBMC-68k forget set, completed the 50. Gradient ascent takes ten ascent steps and then up to 30 retain epochs with patience 10, and every one of its eleven published runs completed the 30; Moon feature unlearning fine-tunes the decoder for 20 epochs.

Table 13: Settings of every method and of the attacker. Generated by `scripts/make_tex_tables.py` from the driver scripts' values.

Method	Settings
Baseline and clean retrains	Adam 10^{-4}, 100 epochs, batch 256, KL warm-up 20 epochs, free bits 0.03, early stopping patience 10 (PBMC-68k: selection on the full-weight objective, patience counted after the warm-up)
Retain-only fine-tuning	Adam 10^{-4}, 50 epochs on $\mathcal{R}$ with patience 10 (the methods' own fine-tune phase: 10 epochs at 10^{-4})
Gradient ascent	10 ascent steps on $\mathcal{F}$ at 10^{-5} with the gradient norm clipped to 1.0, then 30 retain epochs at 10^{-4} with patience 10 (sweep of the ascent step in Section 5)
Fisher scrubbing	per-sample diagonal retain Fisher, relative damping 0.1, 100 scrub steps at 10^{-4} with the step capped at 10^{-3} per parameter (ablation 3×10^{-3} and 10^{-2}; sweep in Section 5), then 10 retain epochs at 10^{-4}
SSD	multiplicative dampening by forget-set Fisher importance, $\alpha = 1$, threshold 0, damping 10^{-5}, then 10 retain epochs at 10^{-4} (the published method has no fine-tune)
SCRUB	20 epochs of 5 forget steps (maximize the KL to the teacher, weight 1) and retain steps (match the teacher, weight 1, plus the ELBO) at 10^{-4}, then 10 retain epochs at 10^{-4}
Moon feature unlearning	20 decoder epochs with the forget-versus-retain latent direction removed, encoder frozen
Contrastive latent	20 epochs, $\gamma = 1$, $\lambda = 1$, 10^{-4}, then 10 retain epochs at 10^{-4}
Frozen critics	one or three pre-trained critics, $\lambda \in \{5, 10\}$, VAE at 10^{-4}, 50 epochs
Extra-gradient co-training	$\lambda \in \{10, 1\}$, 50 passes over $\mathcal{R}$, VAE at 10^{-4}, three critics at 10^{-5} with two critic steps per VAE step, ExtraAdam; critics pre-trained on the cohort's baseline (forget versus retain, 100 epochs)
DP-SGD	from scratch on $\mathcal{R}$, $\varepsilon = 10$, $\delta = 10^{-5}$, gradient norm clipped to 1.0, noise calibrated by Opacus, 50 epochs at 10^{-3}, patience 15
Likelihood-targeted fine-tune	retain loss plus $\lambda = 1$ times the mean absolute gap to the target, Adam 10^{-4}, stop when the mean gap stays within 5 units for 3 epochs
Attacker	MLP [256, 256], spectral normalization, dropout 0.3, 50 epochs, Adam 10^{-3}, batch 128, 80/20 split, 70 features

Forget-set size and the constructive approaches. Two further sets of results change what is measured more than they change the outcome. Structured forget sets of 10, 50, and 100 cells, the larger two spanning several clusters, each have their own retrained references, and extra-gradient pushes the forget cells past them at every size (audit 0.12 to 0.19). Scattered forget sets of 10 to 500 random cells carry nothing the attacker can find, while the audit sees a weak per-cell signal in the baseline (0.69 to 0.79 for thirty cells and more) that Fisher scrubbing reduces only in part. The two constructive approaches fail in the two modes. Synthetic augmentation is scored as a memorizer for the reason given in Section 3. Representation alignment, which matches the retrained model's posteriors on the forget cells, pushes their likelihood past the retrained model anyway (audit 0.007 to 0.018), a first sign that the two instruments measure different quantities.

Held-out utility for every checkpoint carries a 95 percent interval and a paired cost against the baseline (Table 14 for PBMC-33k, Table 15 for PBMC-68k; `scripts/utility_intervals.py`). The negative ELBO on the held-out cells is the deterministic estimator of Section 3, the marker-gene correlation is kept per cell at $z = \mu$, and ARI is the adjusted Rand index of a k-means clustering of the latent means against the Leiden labels. One multinomial draw of the held-out cells is shared by every checkpoint, seeds are resampled within an arm, and the cost is the arm's mean minus the baseline's, positive when worse, so the comparison is paired

on the cells. The margins are 1 percent of the baseline's held-out negative ELBO (3.64 nats on PBMC-33k) and 0.01 in marker r. An arm has no cost when the upper end of its cost interval lies below the margin, costs utility when the lower end lies above it, and is not resolved otherwise. Every row carries its interval so that any other margin can be read off.

On PBMC-33k every post-hoc method at its published setting except the two three-critic frozen runs, every step of the Fisher and gradient-ascent sweeps, the three likelihood-targeted variants, and the 200-epoch fine-tune have no cost. The arms that cost utility are DP-SGD (38.9 nats, interval 37.6 to 40.4; marker r 0.042) and the Fisher checkpoint at ten times the default cap (6.82 nats, interval 6.68 to 6.97; Figure 4). So do the two three-critic frozen runs, at 40.3 and 40.4 nats, with intervals of 7.3 to 85.3 and 7.3 to 86.0 that are wide because the three seeds disagree. On PBMC-68k, where the margin is 4.84 nats on the 10,447 held-out cells, every method at its published setting has no cost on cluster 9, the 200-epoch fine-tune included (ELBO cost 2.19, interval 1.71 to 2.67; marker r cost 0.0069, interval 0.0037 to 0.0099, inside the 0.01 margin). On cluster 8 the one cost is Fisher scrubbing (9.41 nats, interval 5.69 to 13.45), the cost of not fitting the 201 held-out cells of the cluster. The clean retrains, which never saw the forget type, pay 3.49 and 2.20 nats on the held-out cells of the two clusters, inside the margin, and on cluster 8 their marker r cost (0.0066, interval 0.0029 to 0.0105) is not resolved at it.

Table 14: Held-out utility on the 4,964 held-out PBMC-33k cells with 95 percent intervals and paired costs against the baseline (positive is worse; margins 1 percent of the baseline's ELBO and 0.01 in marker r). Below the rule, the sweep, likelihood-targeted, and 200-epoch fine-tune checkpoints and the superseded leaky retrain. Generated from `utility_intervals.json`.

Model	Seeds	Held-out ELBO [95% CI]	ELBO cost [95% CI]	Marker r	Marker cost [95% CI]	ARI	Verdict
Baseline (no unlearning)	1	364.0 [361.4, 366.4]	–	0.831	–	0.468	reference
Clean retrain	5	364.2 [361.7, 366.7]	+0.22 [+0.07, +0.37]	0.831	+0.0003 [-0.0006, +0.0011]	0.481	no cost
Retain-only fine-tune	5	363.0 [360.5, 365.4]	-0.99 [-1.07, -0.89]	0.832	-0.0008 [-0.0014, -0.0001]	0.451	no cost
Gradient ascent + fine-tune	5	363.2 [360.7, 365.7]	-0.77 [-0.84, -0.70]	0.831	-0.0005 [-0.0011, +0.0001]	0.463	no cost
SSD	3	363.6 [361.1, 366.1]	-0.34 [-0.43, -0.22]	0.831	-0.0002 [-0.0006, +0.0003]	0.475	no cost
SCRUB	3	363.6 [361.0, 366.0]	-0.41 [-0.51, -0.29]	0.831	-0.0002 [-0.0006, +0.0004]	0.461	no cost
Moon feature unlearning	3	364.5 [361.9, 366.9]	+0.48 [+0.44, +0.53]	0.831	+0.0005 [+0.0001, +0.0008]	0.468	no cost
Fisher scrubbing	3	363.6 [361.0, 366.0]	-0.42 [-0.50, -0.33]	0.832	-0.0005 [-0.0009, -0.0001]	0.470	no cost
Fisher, step cap 3×10^{-3}	1	363.8 [361.2, 366.2]	-0.20 [-0.30, -0.09]	0.831	+0.0003 [-0.0001, +0.0007]	0.467	no cost
Fisher, step cap 10^{-2}	1	370.8 [368.3, 373.3]	+6.82 [+6.68, +6.97]	0.830	+0.0014 [+0.0002, +0.0026]	0.517	costs utility (ELBO)
Extra-gradient ($\lambda = 10$)	10	363.8 [361.3, 366.3]	-0.16 [-0.45, +0.27]	0.832	-0.0007 [-0.0013, -0.0002]	0.470	no cost
Extra-gradient ($\lambda = 1$)	10	363.2 [360.7, 365.6]	-0.74 [-1.00, -0.32]	0.831	-0.0004 [-0.0010, +0.0002]	0.464	no cost
Contrastive latent	3	363.7 [361.2, 366.2]	-0.24 [-0.36, -0.08]	0.831	-0.0004 [-0.0010, +0.0001]	0.465	no cost
Frozen critics, one critic, $\lambda = 5$	3	363.9 [361.2, 366.3]	-0.09 [-0.65, +0.63]	0.831	-0.0005 [-0.0012, +0.0003]	0.456	no cost
Frozen critics, one critic, $\lambda = 10$	3	363.7 [361.1, 366.2]	-0.23 [-0.70, +0.35]	0.831	-0.0004 [-0.0012, +0.0004]	0.451	no cost
Frozen critics, three critics, $\lambda = 5$	3	404.2 [370.4, 449.8]	+40.26 [+7.35, +85.35]	0.832	-0.0006 [-0.0015, +0.0004]	0.463	costs utility (ELBO)
Frozen critics, three critics, $\lambda = 10$	3	404.4 [370.1, 450.3]	+40.42 [+7.30, +85.99]	0.832	-0.0006 [-0.0014, +0.0004]	0.459	costs utility (ELBO)
DP-SGD ($\varepsilon = 10$, from scratch)	3	402.9 [399.8, 405.9]	+38.94 [+37.58, +40.40]	0.789	+0.0417 [+0.0378, +0.0456]	0.500	costs utility (ELBO and marker)
Fisher, step cap 1.5×10^{-3}	3	363.5 [361.0, 366.0]	-0.42 [-0.49, -0.35]	0.831	-0.0003 [-0.0007, +0.0002]	0.462	no cost
Fisher, step cap 2×10^{-3}	3	363.6 [361.1, 366.0]	-0.37 [-0.44, -0.30]	0.831	-0.0002 [-0.0007, +0.0003]	0.469	no cost
Fisher, step cap 2.5×10^{-3}	3	363.7 [361.1, 366.1]	-0.31 [-0.39, -0.21]	0.831	-0.0002 [-0.0006, +0.0003]	0.468	no cost
Fisher, step cap 3×10^{-3} (sweep)	3	363.8 [361.2, 366.2]	-0.21 [-0.30, -0.10]	0.831	-0.0001 [-0.0006, +0.0004]	0.459	no cost
Gradient ascent, lr 10^{-4}	3	363.2 [360.7, 365.7]	-0.77 [-0.85, -0.68]	0.832	-0.0005 [-0.0012, +0.0002]	0.451	no cost
Gradient ascent, lr 10^{-3}	3	363.2 [360.7, 365.7]	-0.77 [-0.86, -0.69]	0.832	-0.0005 [-0.0012, +0.0002]	0.453	no cost
Gradient ascent, lr 10^{-2}	3	363.2 [360.7, 365.7]	-0.76 [-0.84, -0.68]	0.832	-0.0005 [-0.0012, +0.0003]	0.465	no cost
Likelihood-targeted, look-alike level	10	364.8 [362.2, 367.3]	+0.79 [+0.54, +1.08]	0.830	+0.0008 [+0.0003, +0.0013]	0.483	no cost
Likelihood-targeted, retrained level	10	365.0 [362.5, 367.5]	+1.08 [+0.82, +1.34]	0.830	+0.0008 [+0.0003, +0.0014]	0.484	no cost
Likelihood-targeted, disjoint look-alikes	10	364.8 [362.2, 367.2]	+0.80 [+0.67, +0.94]	0.830	+0.0008 [+0.0003, +0.0013]	0.485	no cost
Retain-only fine-tune, 200 epochs	3	363.2 [360.6, 365.7]	-0.79 [-0.93, -0.66]	0.832	-0.0010 [-0.0022, +0.0002]	0.418	no cost
Leaky retrain (superseded)	5	361.9 [359.4, 364.3]	-2.06 [-2.38, -1.75]	0.834	-0.0026 [-0.0035, -0.0017]	0.460	no cost

Structured forget sets of 10, 50, and 100 cells (ten cells of cluster 13; its 30 cells with 20 of cluster 12; and its 30 cells with all 49 of cluster 12 and 21 of cluster 11) each have three clean retrained references trained on their own retain split. Those references score 0.454, 0.565, and 0.668 under the attacker, so the multi-cluster sets are confound-affected in the way cluster 12 is, and extra-gradient at $\lambda = 10$ is run with three seeds (Table 16). Scattered forget sets of 10 to 500 random cells have three references each and Fisher scrubbing with ten seeds. The baseline, the retrained models, and Fisher scrubbing coincide within 0.02 under the attacker at every size, while the audit sees a weak per-cell signal in the baseline that Fisher scrubbing reduces to between 0.59 and 0.65 for thirty cells and more (0.44 at ten; Table 17). One fresh attacker per size scores every model; the audit column is the likelihood-ratio AUC against the first reference.

The two constructive approaches are scored against the 5 held-out megakaryocytes as well. Synthetic augmentation's two models score 0.827 and 1.000 there, worse than the baseline's 0.607, and representation alignment achieves its match (the posterior divergence falls from 180 to below 0.25 at every weight) with the attacker at 0.28 to 0.35 against the retrained model's 0.548. Table 18 gives Tabula Muris cluster 33 and

Table 15: Held-out utility on the 10,447 held-out PBMC-68k cells with 95 percent intervals and paired costs against the baseline (positive is worse; margins 1 percent of the baseline's ELBO, 4.84 nats, and 0.01 in marker r). Generated from `utility_intervals.json`.

Model	Seeds	Held-out ELBO [95% CI]	ELBO cost [95% CI]	Marker r	Marker cost [95% CI]	ARI	Verdict
Cluster 9, megakaryocytes							
Baseline (no unlearning)	1	484.5 [482.8, 486.2]	–	0.762	–	0.474	reference
Clean retrain	8	488.0 [486.1, 489.9]	+3.49 [+2.63, +4.37]	0.761	+0.0010 [-0.0022, +0.0040]	0.482	no cost
Retain-only fine-tune	3	484.9 [483.2, 486.7]	+0.43 [+0.15, +0.72]	0.761	+0.0011 [-0.0006, +0.0031]	0.443	no cost
Gradient ascent + fine-tune	3	484.6 [482.9, 486.4]	+0.15 [-0.02, +0.33]	0.762	+0.0004 [-0.0010, +0.0018]	0.469	no cost
SSD	3	484.8 [483.1, 486.5]	+0.30 [+0.14, +0.47]	0.761	+0.0014 [+0.0003, +0.0025]	0.457	no cost
SCRUB	3	484.8 [483.1, 486.6]	+0.34 [+0.12, +0.62]	0.762	+0.0002 [-0.0008, +0.0013]	0.470	no cost
Fisher scrubbing	3	487.0 [485.1, 489.0]	+2.47 [+1.56, +3.52]	0.762	+0.0008 [-0.0008, +0.0026]	0.435	no cost
Extra-gradient ($\lambda = 10$)	3	486.0 [484.1, 488.1]	+1.56 [+0.68, +2.78]	0.762	+0.0002 [-0.0012, +0.0015]	0.474	no cost
Contrastive latent	3	485.7 [483.9, 487.5]	+1.21 [+0.75, +1.72]	0.761	+0.0011 [-0.0001, +0.0022]	0.480	no cost
Retain-only fine-tune, 200 epochs	3	486.7 [484.9, 488.5]	+2.19 [+1.71, +2.67]	0.756	+0.0069 [+0.0037, +0.0099]	0.438	no cost
Cluster 8, plasmacytoid dendritic cells							
Baseline (no unlearning)	1	484.5 [482.8, 486.2]	–	0.762	–	0.474	reference
Clean retrain	8	486.7 [484.9, 488.5]	+2.20 [+1.80, +2.63]	0.756	+0.0066 [+0.0029, +0.0105]	0.470	not resolved
Retain-only fine-tune	3	484.7 [483.0, 486.5]	+0.25 [+0.13, +0.38]	0.760	+0.0025 [+0.0011, +0.0040]	0.468	no cost
Gradient ascent + fine-tune	3	484.6 [482.9, 486.3]	+0.08 [-0.02, +0.19]	0.760	+0.0021 [+0.0006, +0.0036]	0.491	no cost
SSD	3	484.7 [482.9, 486.4]	+0.19 [+0.07, +0.31]	0.761	+0.0018 [+0.0008, +0.0028]	0.451	no cost
SCRUB	3	484.5 [482.8, 486.2]	+0.03 [-0.05, +0.11]	0.762	+0.0009 [-0.0006, +0.0025]	0.475	no cost
Fisher scrubbing	3	493.9 [489.8, 498.5]	+9.41 [+5.69, +13.45]	0.761	+0.0013 [+0.0000, +0.0025]	0.453	costs utility (ELBO)
Extra-gradient ($\lambda = 10$)	3	486.4 [484.7, 488.1]	+1.88 [+1.53, +2.21]	0.762	+0.0003 [-0.0017, +0.0022]	0.481	no cost
Contrastive latent	3	487.9 [486.0, 489.8]	+3.39 [+2.82, +4.02]	0.760	+0.0026 [+0.0008, +0.0046]	0.441	no cost

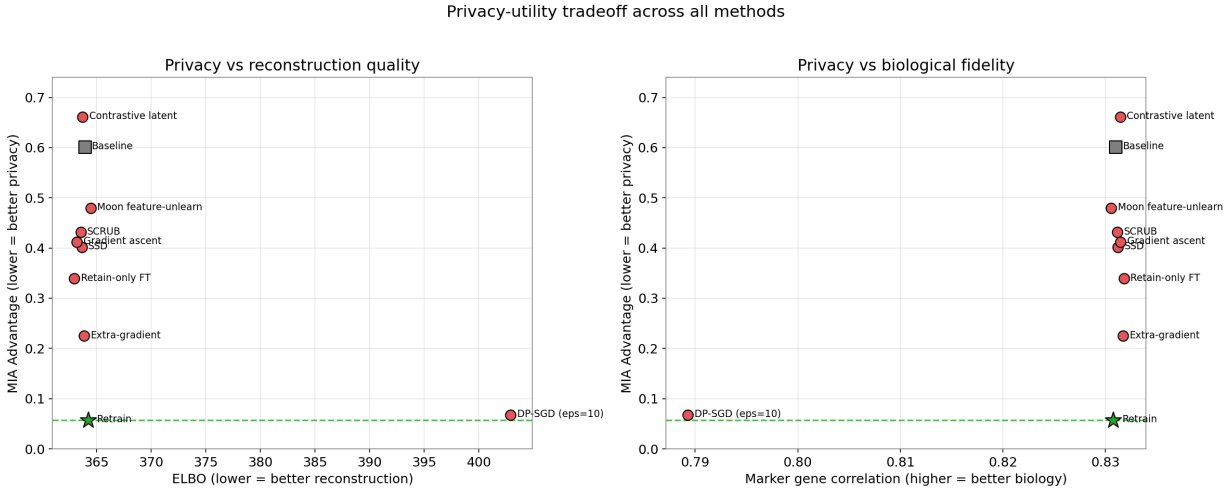


Figure 4: Advantage against held-out ELBO (left) and marker-gene correlation (right) for every method on PBMC-33k cluster 13. Every post-hoc method sits at the baseline's utility; DP-SGD is the one model that pays in utility, and the retrained models reach the target at the baseline's. Generated from the canonical result files by `figures/make_result_figures.py`.

TCGA-BRCA Basal under the cross-type attacker, the within-subtype attacker where one exists, and the audit against each cohort's single retrained model, with cell-bootstrap intervals; no retrained-versus-retrained range is available on these cohorts. PBMC-33k cluster 12, the second rare cluster, repeats cluster 13's pattern against a confound-affected reference (baseline 0.869, retrained 0.814), with retain-only fine-tuning and Fisher scrubbing leaving the attacker at 0.867 and 0.858 and extra-gradient dropping it to 0.41.

B Attacker controls

The attacker of Section 3 is trained once on the baseline's features and applied unchanged to every model, which could understate leakage if a decision rule fitted to one model misses signal that an adversary fitting the target model would find. The first control trains a fresh attacker of the same architecture on each model's own forget-versus-negative features and scores it out of fold, with stratified five-fold cross-validation

Table 16: Structured forget sets with per-size retrained references. Generated from `size_ablations_v2.json`; the forget sets are those of `split_structured_n{10,50,100}.json` under `outputs/p4/extragradient_size/splits`.

Forget cells	Negatives	Baseline AUC (audit)	Retrained AUC, advantage (audit)	Extra-gradient AUC, advantage [CI] (audit)
10	18	0.567 (0.87)	0.454, 0.107 (0.32)	0.326, 0.348 [0.14, 0.63] (0.19)
50	46	0.713 (0.91)	0.565, 0.130 (0.60)	0.374, 0.251 [0.11, 0.41] (0.12)
100	64	0.764 (0.87)	0.668, 0.335 (0.56)	0.411, 0.178 [0.07, 0.32] (0.13)

Table 17: Scattered forget sets with per-size retrained references. Generated from `size_ablations_v2.json`.

Forget cells	Negatives	Baseline AUC (audit)	Retrained AUC (audit)	Fisher AUC [advantage CI] (audit)
10	99	0.554 (0.76)	0.555 (0.68)	0.543 [0.10, 0.26] (0.44)
30	281	0.588 (0.79)	0.581 (0.53)	0.585 [0.12, 0.24] (0.63)
50	449	0.574 (0.70)	0.568 (0.50)	0.570 [0.10, 0.19] (0.59)
100	818	0.550 (0.69)	0.543 (0.50)	0.547 [0.06, 0.13] (0.61)
500	2449	0.517 (0.71)	0.510 (0.51)	0.515 [0.02, 0.05] (0.65)

repeated over 20 random fold splits and the attacker re-seeded per split and fold (Table 19). On the baseline it recovers the transferred attacker's score, and on the five retrained models it lands near chance.

That reading is a property of the architecture and not of the models. The 70 features enter it at their raw scale, with the likelihood terms in the hundreds and the posterior coordinates near unity, and an MLP under spectral normalization on such inputs uses little beyond the large terms. Standardizing the features and refitting, with a logistic regression and with the same MLP, separates the 30 megakaryocytes from their 194 cross-type neighbors at 0.98 to 0.99 on the baseline and on a clean retrain alike (Table 20). Table 21 locates the separation in the posterior mean, which alone gives 0.990 on the baseline and 0.984 on the retrained model; the encoder places a megakaryocyte in its own region whether or not it trained on one. The reconstruction term alone gives 0.769 on the baseline and 0.423 on the retrained model, which is what the raw-feature attacker reads and why it scores the retrained models near chance. The control therefore does not show that the retrained models carry no signal for a dedicated adversary. It shows that against cross-type negatives a dedicated adversary reads cell type on every model, that the transferred attacker's excess on the baseline is the baseline's better fit of the forget cells, and that a chance score against cross-type negatives never certifies the absence of a signal.

The same standardized refit on the within-type controls of Section 3 finds no separation it can resolve on baseline or retrained model (0.40 to 0.55 on cluster 7, cluster 28, and the Basal patients, sets of 30 to 158 cells a side). On the two PBMC-68k forget sets it finds a per-cell signal on the baseline that the retrained models lack, in the reconstruction term and the latent placement together. Every within-type number in the paper is reported from this instrument. Its resolution on PBMC-68k comes from a label-permutation null computed on each model's own features, with the whole out-of-fold procedure refit under 1,000 permutations of the labels. The 97.5th percentiles lie between 0.590 and 0.606 at 107 against 71 cells and between 0.555 and 0.563 at 302 against 201, which is why sub-chance values are read as no signal. The null says how far from 0.5 a model without a per-cell signal can score. It is not an acceptance threshold, and each method is compared with the retrained models directly, paired on the cells they share (Section 4.1).

Four training-free attacks (Table 22) show the same division of labor among the features. A threshold on the reconstruction loss scores the baseline at 0.780 and the retrained model at 0.537, so it reads the fit excess. Thresholds on the KL term and on the ELBO, and a nearest-neighbor attack in latent space, score the retrained model far from chance (0.130, 0.129, and 0.019), so what they separate is cell type.

C Reference-free audit and paired type-level verdict

The audit of Section 3 is defined here in full, with the seed-42 form it replaced kept in the tables of Appendix G for comparison. The construction is that of `scripts/audit_reference_free.py` and `src/audit_stats.py`.

Table 18: Attacker and audit AUCs on Tabula Muris and TCGA-BRCA. Generated from the cohort result files and `attack_lr_reference_multi.json`.

Model	TM attacker (cross)	TM audit [95% CI]	TCGA attacker (cross)	TCGA attacker (within)	TCGA audit [95% CI]
Baseline	0.903	0.877 [0.82, 0.93]	0.803	0.580	0.833 [0.77, 0.89]
Retrained model	0.943	–	0.830	0.572	–
Retain-only fine-tune	–	–	0.830	0.572	0.457 [0.38, 0.53]
Gradient ascent + fine-tune	–	–	0.830	0.571	0.397 [0.32, 0.47]
SSD	–	–	0.830	0.571	0.454 [0.38, 0.53]
SCRUB	–	–	0.829	0.572	0.533 [0.46, 0.61]
Contrastive latent	–	–	0.831	0.572	0.519 [0.44, 0.60]
Fisher scrubbing	0.944	0.346 [0.25, 0.44]	0.784	0.543	0.191 [0.13, 0.25]
Extra-gradient	0.723	0.947 [0.92, 0.97]	–	–	–

Table 19: The paper's attacker architecture refit on each model's own raw features, out of fold over 20 random five-fold splits, with the in-sample upper bound. Generated from `retrain_refit.json`.

Model	Out-of-fold AUC	Out-of-fold advantage	In-sample AUC
Baseline (positive control)	0.797	0.594	0.813
Clean retrain, seed 42	0.536	0.072	0.545
Clean retrain, seed 43	0.507	0.013	0.507
Clean retrain, seed 44	0.499	0.009	0.493
Clean retrain, seed 45	0.518	0.035	0.519
Clean retrain, seed 46	0.478	0.045	0.481
Clean retrains, mean [nested 95% CI]	0.507 [0.45, 0.56]	0.035 [0.00, 0.10]	0.509

The reference is the mean over every clean retrain of each cell's negative ELBO, and the audit AUC is that of the score reference minus target, forget cells against negatives, as before. The retrained arm is scored leave-one-out, each retrained model against the mean of the others. The interval of a method resamples, in one draw, the method's seeds, the retrained models that enter the reference (with replacement), and the cells, through multinomial weights shared by every arm of a cohort. The difference between a method and the retrained arm is therefore paired on the cells and on the reference draw. Two-sided p-values from the draws are Holm-adjusted within the families of the attacker test (published settings, ablations, demonstrations; on PBMC-68k every method arm is one family; Table 34 collects them).

The type-level reading is the gap, the median over the forget cells of the method's negative ELBO minus the retrained arm's, with the same interval; positive is fitted worse than retrained. Because a model that fits everything worse, such as DP-SGD, shows a positive gap for that reason alone, the gap is also given relative to the cross-type negatives' gap, which cancels a global shift. On PBMC-68k the same score is computed against the within-type negatives, raw and as the per-cell z-score of a LiRA-style audit (mean minus target over the standard deviation across retrained models, per cell), which is the likelihood-only counterpart of the per-cell refit attacker.

On PBMC-33k (Table 23) the retrained arm sits at 0.507 with an interval of 0.40 to 0.60, the five leave-one-out values ranging from 0.24 to 0.74. Every method that edits and then fine-tunes is fitted better than retrained, with Holm-adjusted p of 0.014, and contrastive unlearning, extra-gradient at $\lambda = 1$, and Fisher scrubbing at the 3×10^{-3} cap are fitted worse. DP-SGD (0.533, interval -0.12 to 0.19) is fitted as retrained; extra-gradient at $\lambda = 10$ (difference interval -0.56 to 0.42), the Fisher caps of 2×10^{-3} and 2.5×10^{-3}, the fine-tune given the retrained level, and the 200-epoch fine-tune are not resolved. DP-SGD's raw median gap is +37 (interval 21 to 49) because it fits every cell worse; relative to the negatives it is -1 (interval -19 to 16). The Fisher checkpoint at the 10^{-2} cap, inside the retrained range under the seed-42 audit, falls below the arm here (difference -0.22, interval -0.39 to -0.03, Holm-adjusted $p = 0.10$), and the superseded leaky retrain is detected as fitted better (difference 0.34, interval 0.16 to 0.49).

On PBMC-68k (Tables 24 and 25) the cross-type arm on cluster 9 is 0.592 with an interval of 0.41 to 0.66 and leave-one-out values from 0.04 to 0.95, so the audit resolves little there. The paired median gap carries the type-level reading, every method being fitted better than retrained, Fisher scrubbing by -251 (interval

Table 20: Out-of-fold refit AUC (mean ± standard deviation over fold splits) with a logistic regression on standardized features, the paper's MLP on standardized features, and the paper's MLP on raw features, for PBMC-33k cluster 13 against its cross-type negatives and for every within-type control. Generated from `within_type_standardized_check.json` and the PBMC-68k result files.

Forget cells vs negatives, model	Logistic, standardized	MLP, standardized	MLP, raw (the paper's attacker)
PBMC-33k cluster 13, 30 vs 194 cross-type, baseline	0.988 ± 0.002	0.990 ± 0.001	0.799 ± 0.013
PBMC-33k cluster 13, 30 vs 194 cross-type, clean retrain	0.984 ± 0.002	0.984 ± 0.001	0.541 ± 0.006
PBMC-33k cluster 7, 50 vs 80 within, baseline	0.466 ± 0.033	0.415 ± 0.051	0.516 ± 0.006
PBMC-33k cluster 7, 50 vs 80 within, retrained	0.414 ± 0.032	0.425 ± 0.050	0.514 ± 0.007
PBMC-33k cluster 7, 100 vs 80 within, baseline	0.468 ± 0.025	0.445 ± 0.030	0.480 ± 0.010
PBMC-33k cluster 7, 100 vs 80 within, retrained	0.434 ± 0.037	0.475 ± 0.032	0.489 ± 0.009
Tabula Muris cluster 28, 50 vs 67 within, baseline	0.487 ± 0.035	0.505 ± 0.042	0.527 ± 0.019
Tabula Muris cluster 28, 50 vs 67 within, retrained	0.550 ± 0.033	0.504 ± 0.030	0.521 ± 0.017
Tabula Muris cluster 28, 100 vs 67 within, baseline	0.536 ± 0.028	0.550 ± 0.023	0.480 ± 0.029
Tabula Muris cluster 28, 100 vs 67 within, retrained	0.449 ± 0.035	0.461 ± 0.035	0.488 ± 0.020
TCGA-BRCA Basal, 158 vs 30 within, baseline	0.442 ± 0.034	0.436 ± 0.034	0.566 ± 0.013
TCGA-BRCA Basal, 158 vs 30 within, retrained	0.404 ± 0.040	0.424 ± 0.038	0.556 ± 0.015
PBMC-68k cluster 9, 107 vs 71 within, baseline	0.683 ± 0.022	–	0.452 ± 0.020
PBMC-68k cluster 9, 107 vs 71 within, clean retrains (8)	0.472 to 0.538	–	0.491 to 0.504
PBMC-68k cluster 8, 302 vs 201 within, baseline	0.603 ± 0.013	–	0.473 ± 0.011
PBMC-68k cluster 8, 302 vs 201 within, clean retrains (8)	0.450 to 0.503	–	0.502 to 0.508

Table 21: Out-of-fold AUC of a standardized logistic regression restricted to one feature group (10 fold splits). Generated from `feature_group_check.json`.

Forget cells vs negatives, model	All 70	Latent mean	Latent log-var	Likelihood terms	Reconstruction	kNN distance
PBMC-33k cluster 13 vs 194 cross-type, baseline	0.990	0.990	0.988	0.921	0.769	0.628
PBMC-33k cluster 13 vs 194 cross-type, clean retrain	0.984	0.984	0.985	0.878	0.423	0.605
PBMC-68k cluster 9 vs 71 within-type, baseline	0.668	0.545	0.573	0.594	0.609	0.450
PBMC-68k cluster 9 vs 71 within-type, clean retrain	0.495	0.518	0.515	0.515	0.439	0.474
PBMC-68k cluster 8 vs 201 within-type, baseline	0.589	0.537	0.533	0.521	0.542	0.480
PBMC-68k cluster 8 vs 201 within-type, clean retrain	0.487	0.485	0.497	0.531	0.459	0.483

−356 to −56) and the 200-epoch fine-tune by −247 (interval −290 to −205). On cluster 8 the arm is 0.510 (interval 0.48 to 0.53), the four edit-then-fine-tune methods are fitted better, contrastive unlearning worse, and Fisher scrubbing's audit reads fitted as retrained (0.489) while its gap is not resolved (+60, interval −6 to 226). Within type, the likelihood audit finds the baseline's cells only at the edge on cluster 9 (raw 0.574, interval −0.003 to 0.16; z-scored 0.652, interval 0.02 to 0.23) and clearly on cluster 8 (0.607, interval 0.05 to 0.16). It finds extra-gradient's amplified signal on both (0.674 and 0.738), and contrastive unlearning's over-corrected cells on cluster 8 in the other direction (0.406, interval −0.15 to −0.04). Gradient ascent, SSD, SCRUB, and Fisher scrubbing on cluster 8 sit just above the arm (0.556 to 0.572, Holm-adjusted p of 0.06 to 0.09), where the refit attacker of Section 4.1 separates them; the likelihood alone is the weaker per-cell instrument, which is the simulation's finding on real data.

D A retrain-free stopping probe

The probe asks whether any quantity a controller can observe without a retrained model picks the setting of a method's knob that the retrain-based verdict picks. The first half of it is the knob grid on the standard split (`scripts/stopping_probe_68k_train.py`, `stopping_probe_68k.py`). On both PBMC-68k forget sets, three seeds each and all descending from the 68k baseline, the Fisher scrub was run at step caps of 2×10^{-3}, 3×10^{-3}, 5×10^{-3}, and 10^{-2} beside the published 10^{-3}. The retain-only fine-tune was saved at 10, 25, 50, 100, 150, and 200 epochs. Every checkpoint is scored with the instruments of Section 4.1: the per-cell refit paired against the eight retrained models on the shared cells with the margin of Table 33, the cross-type audit, and the forget-cell median against the retrained medians. The baseline reproduces its published row to the last digit. Beside them stand the quantities a controller can observe. These are the forget cells' fit, the fit of their look-alikes, the fit of a fixed sample of 2,000 retain cells, the fit of the cross-type negatives, and the held-out utility of Appendix A. The look-alikes are the ten nearest retain cells of each forget cell in the baseline's latent space, over a pool of 10,000.

Table 22: Training-free attacks on PBMC-33k cluster 13, AUC (advantage). The retrained model is the clean retrain of seed 42, and the two unlearned models are single seeds of extra-gradient and Fisher scrubbing. Generated from `attack_suite_results.json`.

Attack	Baseline	Clean retrain	Extra-gradient	Fisher scrubbing
Reconstruction threshold	0.780 (0.56)	0.537 (0.07)	0.301 (0.40)	0.688 (0.38)
KL threshold	0.073 (0.85)	0.130 (0.74)	0.441 (0.12)	0.079 (0.84)
ELBO threshold	0.073 (0.85)	0.129 (0.74)	0.439 (0.12)	0.079 (0.84)
k-NN in latent space	0.009 (0.98)	0.019 (0.96)	0.159 (0.68)	0.010 (0.98)

Table 23: Reference-free audit on PBMC-33k cluster 13. The reference is the mean over the five clean retrains; the retrained arm is leave-one-out; intervals resample the method's seeds, the reference retrains, and the shared cells. The gap is the median forget-cell negative ELBO minus the retrained arm's (positive is fitted worse), and the relative gap subtracts the same gap on the 194 negatives. Below the rule, the sweeps, the likelihood-targeted variants, the 200-epoch fine-tune, and the superseded leaky retrain. Generated from `audit_reference_free.json`.

Model	Seeds	Audit, mean-of-retrains reference [95% CI]	Difference from the retrained arm [95% CI]	Audit verdict	Forget median	Gap, median [95% CI]	Gap relative to the negatives [95% CI]	Fit verdict
Clean retrains (leave-one-out)	5	0.507 [0.40, 0.60]; per seed 0.24 to 0.74	–	reference	317 (311 to 326)	–	–	–
Baseline (no unlearning)	1	0.995 [0.99, 1.00]	+0.488 [0.40, 0.59]	fitted better	247	-69.7 [-85.4, -58.8]	-69.7 [-85.6, -58.4]	fitted better
Retain-only fine-tune	5	0.988 [0.93, 1.00]	+0.481 [0.38, 0.59]	fitted better	282	-35.1 [-45.0, -21.8]	-34.3 [-44.1, -20.9]	fitted better
Gradient ascent + fine-tune	5	0.995 [0.99, 1.00]	+0.488 [0.40, 0.59]	fitted better	274	-42.9 [-52.2, -31.2]	-41.8 [-51.8, -30.1]	fitted better
SSD	3	0.995 [0.99, 1.00]	+0.488 [0.40, 0.59]	fitted better	274	-43.2 [-54.6, -30.3]	-42.5 [-54.8, -29.6]	fitted better
SCRUB	3	0.980 [0.92, 1.00]	+0.473 [0.37, 0.58]	fitted better	274	-43.1 [-56.2, -28.8]	-42.4 [-55.8, -28.0]	fitted better
Moon feature unlearning	3	0.996 [0.99, 1.00]	+0.488 [0.40, 0.59]	fitted better	272	-45.3 [-58.9, -35.7]	-45.6 [-59.2, -35.6]	fitted better
Fisher scrubbing	3	0.992 [0.98, 1.00]	+0.484 [0.39, 0.59]	fitted better	289	-27.9 [-43.0, -17.3]	-27.1 [-42.7, -16.4]	fitted better
Fisher, step cap 3×10^{-3}	1	0.141 [0.03, 0.27]	-0.367 [-0.52, -0.20]	fitted worse	398	+80.6 [+30.5, +102.5]	+80.8 [+30.8, +102.8]	fitted worse
Fisher, step cap 10^{-2}	1	0.288 [0.15, 0.45]	-0.219 [-0.39, -0.03]	fitted worse	360	+42.9 [+5.47, +83.2]	+36.6 [-2.44, +76.4]	not resolved
Extra-gradient ($\lambda = 10$)	10	0.250 [0.001, 0.91]	-0.257 [-0.56, 0.42]	not resolved	346	+28.9 [-24.8, +77.2]	+28.6 [-25.4, +76.1]	not resolved
Extra-gradient ($\lambda = 1$)	10	0.044 [0.0003, 0.20]	-0.463 [-0.57, -0.28]	fitted worse	364	+47.5 [+6.16, +137.5]	+47.4 [+5.90, +135.0]	fitted worse
Contrastive latent	3	0.003 [0.00, 0.01]	-0.504 [-0.59, -0.40]	fitted worse	409	+92.5 [+67.1, +106.9]	+91.3 [+66.1, +105.9]	fitted worse
Frozen critics, one critic, $\lambda = 5$	3	0.004 [0.00, 0.01]	-0.503 [-0.59, -0.40]	fitted worse	1170	+852.6 [+716.4, +896.0]	+851.3 [+715.3, +894.5]	fitted worse
Frozen critics, one critic, $\lambda = 10$	3	0.003 [0.00, 0.01]	-0.505 [-0.59, -0.40]	fitted worse	1035	+717.9 [+628.9, +779.0]	+716.6 [+628.2, +777.8]	fitted worse
Frozen critics, three critics, $\lambda = 5$	3	0.006 [0.00, 0.01]	-0.501 [-0.59, -0.40]	fitted worse	54015	+53697.7 [+38245.0, +68509.4]	+53696.0 [+38244.2, +68507.7]	fitted worse
Frozen critics, three critics, $\lambda = 10$	3	0.005 [0.00, 0.01]	-0.502 [-0.59, -0.40]	fitted worse	54080	+53763.5 [+39142.7, +66659.7]	+53761.9 [+39140.6, +66655.9]	fitted worse
DP-SGD ($\varepsilon = 10$, from scratch)	3	0.533 [0.43, 0.64]	+0.025 [-0.12, 0.19]	as retrained	354	+36.9 [+20.7, +48.9]	-0.94 [-18.7, +15.6]	as retrained
Fisher, step cap 1.5×10^{-3}	3	0.887 [0.69, 1.00]	+0.380 [0.16, 0.55]	fitted better	305	-11.9 [-37.3, +6.76]	-11.0 [-37.3, +7.41]	not resolved
Fisher, step cap 2×10^{-3}	3	0.610 [0.35, 0.85]	+0.103 [-0.17, 0.37]	not resolved	326	+9.35 [-27.6, +32.7]	+10.2 [-26.9, +33.5]	not resolved
Fisher, step cap 2.5×10^{-3}	3	0.312 [0.13, 0.57]	-0.196 [-0.40, 0.10]	not resolved	352	+35.5 [-10.1, +66.3]	+36.2 [-9.75, +67.2]	not resolved
Fisher, step cap 3×10^{-3} (sweep)	3	0.168 [0.04, 0.32]	-0.339 [-0.50, -0.15]	fitted worse	388	+71.0 [+22.8, +117.0]	+71.4 [+23.8, +116.8]	fitted worse
Gradient ascent, lr 10^{-4}	3	0.995 [0.99, 1.00]	+0.488 [0.40, 0.59]	fitted better	274	-42.4 [-51.2, -32.0]	-41.5 [-51.0, -31.0]	fitted better
Gradient ascent, lr 10^{-3}	3	0.995 [0.99, 1.00]	+0.487 [0.40, 0.59]	fitted better	274	-42.5 [-51.5, -31.8]	-41.6 [-51.3, -30.2]	fitted better
Gradient ascent, lr 10^{-2}	3	0.995 [0.98, 1.00]	+0.488 [0.40, 0.59]	fitted better	275	-41.8 [-50.9, -30.9]	-40.9 [-50.5, -29.9]	fitted better
Likelihood-targeted, look-alike level	10	0.672 [0.48, 0.84]	+0.165 [-0.05, 0.37]	not resolved	306	-10.7 [-49.9, +4.84]	-10.9 [-50.4, +4.38]	not resolved
Likelihood-targeted, retrained level	10	0.416 [0.24, 0.60]	-0.091 [-0.28, 0.12]	not resolved	337	+20.1 [-20.5, +35.4]	+18.9 [-22.8, +33.7]	not resolved
Likelihood-targeted, disjoint look-alikes	10	0.744 [0.58, 0.89]	+0.237 [0.05, 0.42]	fitted better	296	-20.8 [-61.8, -5.59]	-21.1 [-62.1, -5.68]	fitted better
Retain-only fine-tune, 200 epochs	3	0.688 [0.48, 0.88]	+0.181 [-0.05, 0.40]	not resolved	311	-5.46 [-19.8, +8.36]	-4.78 [-19.4, +9.60]	as retrained
Leaky retrain (superseded)	5	0.849 [0.69, 0.96]	+0.342 [0.16, 0.49]	fitted better	304	-13.1 [-27.6, -2.18]	-12.3 [-26.6, -1.01]	fitted better

The grid answers the weaker question, whether an observable changes where the verdict changes (Table 27). The full design, a split that withholds a fifth of the forget set from the deletion so that a controller has cells of known membership to calibrate on, is run below (Table 26).

The Fisher cap has no working setting on this cohort. At the published cap the cells are more identifiable than retrained on both clusters (differences of 0.088 and 0.102) and the forget cells sit at medians of 814 to 831 and 697 to 812 across seeds. At every larger cap the scrub takes the forget cells to medians between 700 and 166,282 and the ten retain epochs do not bring them back, and the held-out ELBO rises from 484.5 to between 504.1 and 1,750.3. The cells stay more identifiable than retrained at every cap on cluster 8 (at the 10^{-2} cap by a lower bound of 0.004) and at all but the two largest on cluster 9. There the 5×10^{-3} cap (0.047, interval -0.023 to 0.118) and the 10^{-2} cap (-0.010, interval -0.081 to 0.061) are not resolved, on models whose forget cells are fitted at medians of 116,634 to 149,565 and 5,059 to 29,373 against the retrained models' 1,014 to 1,154. The route that exists on PBMC-33k, a cap between too little and too much, does not exist where the decoder-output overlap is near zero (Section 6); here the held-out utility, which a controller can observe, flags every cap but the one that leaves the cells more identifiable than retrained.

Along the fine-tune the verdict changes and no observable does. On cluster 9 the cells are more identifiable at 10 and 25 epochs, not resolved from 50 to 200 epochs; on cluster 8 they are more identifiable from 10 to 100 epochs, not resolved at 150 and 200 epochs. No snapshot is certified at the retrained level on either cluster, and none could be. The probe's intervals are 0.073 to 0.163 wide against windows of 0.087 and 0.075 (Section 3), and only the two largest Fisher caps on cluster 8 are narrow enough, both reading more identifiable. The 200-epoch differences are -0.006, interval -0.087 to 0.075 on cluster 9 and 0.039, interval -0.002 to 0.081 on cluster 8, with the cluster-9 point estimate (0.509) inside the retrained models' spread of

Table 24: Reference-free cross-type audit and paired median gap on PBMC-68k, against the eight clean retrains of each forget set. Generated from `audit_reference_free.json`.

Model	Seeds	Audit, mean-of-retrains reference [95% CI]	Difference from the retrained arm [95% CI]	Audit verdict	Forget median	Gap, median [95% CI]	Gap relative to the negatives [95% CI]	Fit verdict
Cluster 9, megakaryocytes								
Clean retrains (leave-one-out)	8	0.592 [0.41, 0.66]; per seed 0.04 to 0.95	–	reference	1076 (1014 to 1154)	–	–	–
Baseline (no unlearning)	1	1.000 [1.00, 1.00]	+0.408 [0.34, 0.59]	fitted better	491	-584.9 [-627.8, -548.2]	-585.0 [-629.6, -550.6]	fitted better
Retain-only fine-tune	3	1.000 [1.00, 1.00]	+0.408 [0.34, 0.59]	fitted better	681	-394.8 [-440.9, -354.0]	-395.4 [-440.4, -355.1]	fitted better
Gradient ascent + fine-tune	3	1.000 [1.00, 1.00]	+0.408 [0.34, 0.59]	fitted better	624	-451.8 [-495.8, -414.7]	-451.6 [-495.7, -414.6]	fitted better
SSD	3	1.000 [1.00, 1.00]	+0.408 [0.34, 0.59]	fitted better	611	-464.5 [-503.2, -430.8]	-463.8 [-505.1, -431.7]	fitted better
SCRUB	3	0.994 [0.97, 1.00]	+0.403 [0.34, 0.59]	fitted better	596	-479.8 [-529.7, -424.4]	-479.9 [-530.9, -425.7]	fitted better
Fisher scrubbing	3	0.653 [0.53, 0.76]	+0.062 [-0.07, 0.27]	not resolved	824	-251.1 [-356.0, -56.3]	-251.2 [-355.8, -58.8]	fitted better
Extra-gradient ($\lambda = 10$)	3	0.999 [1.00, 1.00]	+0.408 [0.34, 0.59]	fitted better	482	-593.7 [-635.5, -556.4]	-596.2 [-637.5, -559.3]	fitted better
Contrastive latent	3	0.971 [0.92, 1.00]	+0.379 [0.30, 0.57]	fitted better	711	-364.7 [-444.1, -294.7]	-364.8 [-445.6, -295.0]	fitted better
Retain-only fine-tune, 200 epochs	3	1.000 [1.00, 1.00]	+0.408 [0.34, 0.59]	fitted better	829	-246.5 [-289.7, -205.4]	-250.5 [-292.6, -207.1]	fitted better
Cluster 8, plasmacytoid dendritic cells								
Clean retrains (leave-one-out)	8	0.510 [0.48, 0.53]; per seed 0.31 to 0.66	–	reference	680 (672 to 684)	–	–	–
Baseline (no unlearning)	1	0.999 [1.00, 1.00]	+0.489 [0.47, 0.52]	fitted better	532	-147.8 [-160.2, -131.1]	-146.2 [-159.2, -128.9]	fitted better
Retain-only fine-tune	3	0.994 [0.99, 1.00]	+0.484 [0.46, 0.52]	fitted better	583	-96.7 [-112.9, -84.1]	-94.5 [-111.2, -80.6]	fitted better
Gradient ascent + fine-tune	3	0.997 [0.99, 1.00]	+0.487 [0.47, 0.52]	fitted better	570	-109.6 [-123.5, -96.5]	-107.3 [-122.7, -93.5]	fitted better
SSD	3	0.999 [1.00, 1.00]	+0.489 [0.47, 0.52]	fitted better	567	-112.1 [-130.3, -99.3]	-110.7 [-129.7, -96.8]	fitted better
SCRUB	3	0.999 [1.00, 1.00]	+0.489 [0.47, 0.52]	fitted better	556	-123.5 [-137.5, -111.3]	-122.4 [-136.5, -109.2]	fitted better
Fisher scrubbing	3	0.489 [0.39, 0.59]	-0.021 [-0.12, 0.08]	as retrained	740	+60.3 [-6.26, +226.3]	+62.2 [-4.49, +227.4]	not resolved
Extra-gradient ($\lambda = 10$)	3	0.929 [0.80, 1.00]	+0.419 [0.29, 0.50]	fitted better	499	-180.5 [-196.1, -160.5]	-180.9 [-197.2, -160.5]	fitted better
Contrastive latent	3	0.011 [0.00004, 0.04]	-0.498 [-0.52, -0.46]	fitted worse	767	+87.5 [+61.4, +105.9]	+88.1 [+61.2, +107.4]	fitted worse

Table 25: Within-type likelihood audit on PBMC-68k, forget cells against the held-out cells of their type, raw and as the per-cell z-score across the retrained models, each with its paired difference from the leave-one-out retrained arm. Generated from `audit_reference_free.json`.

Model	Seeds	Within-type audit [95% CI]	Difference [95% CI]	Verdict	Per-cell z-scored audit [95% CI]	Difference [95% CI]	Verdict
Cluster 9, megakaryocytes							
Clean retrains (leave-one-out)	8	0.495 [0.48, 0.51]	–	reference	0.500 [0.48, 0.53]	–	reference
Baseline (no unlearning)	1	0.574 [0.49, 0.66]	+0.078 [-0.003, 0.16]	not resolved	0.652 [0.52, 0.73]	+0.152 [0.02, 0.23]	fitted better
Retain-only fine-tune	3	0.483 [0.40, 0.57]	-0.012 [-0.10, 0.07]	not resolved	0.546 [0.43, 0.64]	+0.045 [-0.07, 0.14]	not resolved
Gradient ascent + fine-tune	3	0.509 [0.42, 0.59]	+0.014 [-0.07, 0.10]	not resolved	0.577 [0.46, 0.67]	+0.076 [-0.05, 0.17]	not resolved
SSD	3	0.516 [0.43, 0.60]	+0.021 [-0.06, 0.10]	not resolved	0.591 [0.46, 0.68]	+0.090 [-0.04, 0.18]	not resolved
SCRUB	3	0.501 [0.40, 0.60]	+0.006 [-0.09, 0.11]	not resolved	0.569 [0.44, 0.67]	+0.068 [-0.06, 0.17]	not resolved
Fisher scrubbing	3	0.478 [0.39, 0.56]	-0.017 [-0.10, 0.07]	not resolved	0.501 [0.42, 0.58]	+0.000 [-0.09, 0.09]	not resolved
Extra-gradient ($\lambda = 10$)	3	0.674 [0.59, 0.76]	+0.178 [0.09, 0.26]	fitted better	0.773 [0.63, 0.83]	+0.272 [0.13, 0.33]	fitted better
Contrastive latent	3	0.526 [0.44, 0.61]	+0.030 [-0.05, 0.11]	not resolved	0.521 [0.42, 0.61]	+0.021 [-0.09, 0.12]	not resolved
Retain-only fine-tune, 200 epochs	3	0.477 [0.39, 0.56]	-0.019 [-0.10, 0.07]	not resolved	0.525 [0.43, 0.63]	+0.025 [-0.08, 0.13]	not resolved
Cluster 8, plasmacytoid dendritic cells							
Clean retrains (leave-one-out)	8	0.498 [0.49, 0.51]	–	reference	0.498 [0.49, 0.51]	–	reference
Baseline (no unlearning)	1	0.607 [0.55, 0.66]	+0.110 [0.05, 0.16]	fitted better	0.592 [0.53, 0.64]	+0.094 [0.03, 0.14]	fitted better
Retain-only fine-tune	3	0.552 [0.50, 0.60]	+0.054 [-0.002, 0.10]	not resolved	0.559 [0.50, 0.61]	+0.061 [-0.002, 0.11]	not resolved
Gradient ascent + fine-tune	3	0.564 [0.51, 0.62]	+0.067 [0.01, 0.12]	fitted better	0.568 [0.51, 0.62]	+0.070 [0.01, 0.12]	fitted better
SSD	3	0.564 [0.51, 0.62]	+0.066 [0.01, 0.12]	fitted better	0.565 [0.50, 0.61]	+0.067 [0.002, 0.12]	fitted better
SCRUB	3	0.572 [0.52, 0.63]	+0.075 [0.02, 0.13]	fitted better	0.574 [0.51, 0.62]	+0.076 [0.01, 0.13]	fitted better
Fisher scrubbing	3	0.556 [0.50, 0.61]	+0.058 [0.003, 0.11]	fitted better	0.552 [0.50, 0.61]	+0.054 [-0.004, 0.11]	not resolved
Extra-gradient ($\lambda = 10$)	3	0.738 [0.68, 0.80]	+0.240 [0.18, 0.30]	fitted better	0.723 [0.63, 0.77]	+0.226 [0.13, 0.27]	fitted better
Contrastive latent	3	0.406 [0.35, 0.46]	-0.092 [-0.15, -0.04]	fitted worse	0.414 [0.36, 0.47]	-0.084 [-0.14, -0.03]	fitted worse

0.472 to 0.538 and the cluster-8 one (0.525) above its spread of 0.450 to 0.503. The 200-epoch snapshot on cluster 9 is a second run of the configuration of Section 5, trained and scored the same way, and its -0.006 (interval -0.087 to 0.075) stands beside that run's $+0.008$ (interval -0.062 to 0.080).

Every observable moves smoothly and in one direction through the point where the verdict moves. On cluster 9 the forget median, pooled over the three seeds, rises from 550 to 828 (retrained models 1,014 to 1,154), and the look-alikes' fit improves from 509 to 498 and the retain sample's from 481 to 474 as the fine-tune overfits the retain set. The cross-type negatives drift from 524 to 528, and the held-out ELBO cost grows from -0.05 to 2.10 nats with the marker cost from 0.0000 to 0.0070; cluster 8 runs the same course (550 to 625 against retrained medians of 672 to 684). The held-out cells of the forget type, which a controller does not have, move with the forget cells at every epoch (medians within 29 of each other), so the per-cell verdict turns while the type is still fitted better than any retrained model fits it. The one observable rule the data suggest, fine-tuning until the utility cost reaches its margin, does not track the verdict either. On cluster 8 the marker cost reaches the margin's edge at 150 epochs, where the verdict moves from more identifiable to not resolved. On cluster 9 it has not reached it at 200 epochs, a hundred and fifty epochs after the verdict moved; the agreement on one cluster is a coincidence, not a signal.

The full design: held-back members. On each forget set a seeded fifth of the forget cells (21 megakaryocytes, 60 cells of cluster 8) is held back as members the controller knows and kept in the retain set, and the rest (86 and 242) is deleted. Eight clean retrains are trained on the new retain set (`scripts/stopping_probe_80_20_train.py`, `stopping_probe_80_20_score.py`). The retain-only fine-tune of the baseline is saved at the same six epoch counts and the Fisher scrub run at caps of 10^{-3} and

2×10^{-3}, three seeds each. The retrain-based verdict is the per-cell refit of the deleted cells against the held-out cells of their type, paired against the eight retrains with the margin of Section 3 built on this split (0.048 and 0.023). The observable calibration statistic is the same refit of the deleted cells against the held-back members, both of known membership; it sits near 0.5 on the baseline (0.597 and 0.482; both groups are members, so chance is the expected value) and rises as the deletion takes effect. The held-back members change the reference. Retrained models that learn the type from 21 kept megakaryocytes fit the deleted ones at a median of 595, against 1,014 to 1,154 when none were kept, and separate them from the members at 0.889 (cluster 8: 585 and 0.587).

The calibration statistic has no fixed point. On cluster 9 it passes the retrained models' level between 25 and 50 epochs (0.804 to 0.931) and keeps rising to 0.986 at 200; on cluster 8 it passes it at 50 (0.612 against the retrains' 0.587) and reaches 0.801. The cause is that the fine-tune keeps fitting the kept members better (their median falls from 460 to 329 on cluster 9) while the deleted cells drift the other way. The statistic measures the members' growing exposure as much as the deleted cells' erosion, and it tells a controller that the deletion is working but not when to stop.

The retrain-based verdict moves elsewhere. On cluster 9 the deleted cells are more identifiable at 10 and 25 epochs (at 25 by a lower bound of 0.004), not resolved from 50 to 200 epochs (differences of +0.016 to +0.054 from 50 epochs on, at a margin of 0.048 with 86 cells). On cluster 8 they are more identifiable at 10 and 25 epochs, not resolved from 50 to 200 epochs (-0.000, interval -0.043 to 0.043 at 200, margin 0.023). No snapshot of this design is narrow enough for the verdict at the retrained level either; the intervals are 0.079 to 0.185 wide against windows of 0.095 and 0.046.

One observable does carry a marker in this design, and it is not the calibration statistic. The deleted cells' own median fit rises from 528 at 10 epochs to 585 at 100 and then stops (585, 587). That is the level the retrained models give them (595) and the level at which the model fits the held-out cells of the type (570), where the reference-free audit's difference from the retrained arm covers zero (0.539 at 100 epochs). With members of the type in the retain set the fine-tune's asymptote for the deleted cells is the fit the retain data give the type, which is the retrained level by construction, and a plateau in the deleted cells' fit marks the type-level convergence. On the standard split the same quantity was still moving at 200 epochs because the retrained level lay far above. The marker is type-level and conservative. On cluster 8 the plateau is not yet clean at 200 (573 then 577, retrained 585, audit 0.687) while the per-cell verdict has already turned, and on cluster 9 the per-cell test stays not resolved after the plateau, so the per-cell question, the adversary's, is not what the plateau answers. Both Fisher caps leave the deleted cells more identifiable than retrained on both clusters.

Table 26: The full stopping probe on PBMC-68k, 80 percent of the forget cells deleted and 20 percent held back as known members, with eight clean retrains on the new retain set. The per-cell AUC of the deleted cells against the held-out cells of their type, its paired difference from the eight retrains, and the verdict are the retrain-based reading. The deleted-versus-kept-members AUC is the calibration statistic a controller can compute, with its interval, and the kept-members-versus-held-out column is the same refit of the kept members against the held-out cells. The deleted and kept-member medians are negative ELBO per cell, the gap is the deleted median minus the kept-member median, and held-out ELBO and marker r are the utility of Appendix A. Generated from `stopping_probe80_mk.json` and `stopping_probe80_pdc.json`.

Setting	Deleted vs held-out, per-cell AUC [nested 95% CI]	Paired difference [95% CI]	Verdict	Deleted vs kept members (observable) [95% CI]	Kept members vs held-out	Deleted median	Kept-member median	Gap	Held-out ELBO	Marker r
Cluster 9, megakaryocytes: 86 deleted, 21 held back; retrained per-cell arm 0.494, margin 0.048										
Clean retrains (8)	0.494 [0.43, 0.56]	–	reference	0.889 [0.82, 0.94]	0.845	595	478	+117	484.7	0.763
Baseline	0.680 [0.60, 0.76]	–	trained on the cells	0.597 [0.47, 0.72]	0.685	496	460	+36	484.5	0.762
Fisher, cap 10^{-3}	0.624 [0.55, 0.70]	+0.130 [0.05, 0.21]	more identifiable	0.637 [0.50, 0.77]	0.665	581	524	+57	484.6	0.762
Fisher, cap 2×10^{-3}	0.601 [0.52, 0.68]	+0.107 [0.03, 0.19]	more identifiable	0.623 [0.50, 0.74]	0.625	768	708	+60	485.8	0.761
Retain fine-tune, 10 epochs	0.645 [0.57, 0.72]	+0.152 [0.06, 0.24]	more identifiable	0.720 [0.59, 0.84]	0.801	528	438	+90	484.3	0.762
Retain fine-tune, 25 epochs	0.589 [0.51, 0.67]	+0.095 [0.004, 0.18]	more identifiable	0.804 [0.70, 0.89]	0.863	547	426	+121	484.2	0.762
Retain fine-tune, 50 epochs	0.548 [0.47, 0.62]	+0.054 [-0.03, 0.14]	not resolved	0.931 [0.86, 0.98]	0.934	563	408	+155	484.2	0.761
Retain fine-tune, 100 epochs	0.510 [0.42, 0.60]	+0.016 [-0.08, 0.11]	not resolved	0.957 [0.89, 0.99]	0.951	585	377	+209	484.4	0.760
Retain fine-tune, 150 epochs	0.511 [0.43, 0.59]	+0.017 [-0.06, 0.10]	not resolved	0.977 [0.93, 1.00]	0.964	585	348	+237	484.7	0.758
Retain fine-tune, 200 epochs	0.522 [0.44, 0.60]	+0.029 [-0.05, 0.11]	not resolved	0.986 [0.96, 1.00]	0.968	587	329	+259	485.0	0.757
Cluster 8, plasmacytoid dendritic cells: 242 deleted, 60 held back; retrained per-cell arm 0.478, margin 0.023										
Clean retrains (8)	0.478 [0.44, 0.52]	–	reference	0.587 [0.52, 0.65]	0.581	585	539	+46	484.9	0.764
Baseline	0.570 [0.52, 0.62]	–	trained on the cells	0.482 [0.41, 0.55]	0.593	534	523	+11	484.5	0.762
Fisher, cap 10^{-3}	0.538 [0.49, 0.59]	+0.059 [0.02, 0.10]	more identifiable	0.536 [0.46, 0.61]	0.583	717	690	+27	487.7	0.752
Fisher, cap 2×10^{-3}	0.529 [0.48, 0.58]	+0.051 [0.01, 0.09]	more identifiable	0.476 [0.41, 0.54]	0.527	761	739	+22	489.1	0.746
Retain fine-tune, 10 epochs	0.553 [0.50, 0.60]	+0.075 [0.03, 0.12]	more identifiable	0.498 [0.42, 0.57]	0.604	544	514	+30	484.3	0.762
Retain fine-tune, 25 epochs	0.548 [0.49, 0.60]	+0.070 [0.02, 0.12]	more identifiable	0.505 [0.43, 0.58]	0.620	550	509	+41	484.2	0.760
Retain fine-tune, 50 epochs	0.517 [0.47, 0.57]	+0.038 [-0.01, 0.08]	not resolved	0.612 [0.54, 0.68]	0.697	557	501	+56	484.3	0.761
Retain fine-tune, 100 epochs	0.519 [0.47, 0.57]	+0.041 [-0.01, 0.09]	not resolved	0.714 [0.64, 0.79]	0.754	565	490	+76	484.5	0.759
Retain fine-tune, 150 epochs	0.498 [0.45, 0.55]	+0.020 [-0.02, 0.06]	not resolved	0.760 [0.69, 0.83]	0.793	573	482	+91	484.9	0.759
Retain fine-tune, 200 epochs	0.478 [0.43, 0.53]	-0.000 [-0.04, 0.04]	not resolved	0.801 [0.73, 0.86]	0.821	577	475	+102	485.4	0.757

Table 27: Stopping probe on PBMC-68k. Per-cell AUC is the standardized within-type refit with its nested interval and paired difference from the eight retrained models (margin of Table 33); the audit is the cross-type audit against seed 42; the forget median is ranged over seeds and read against the retrained medians (type). Look-alikes, retain sample, and cross negatives are mean negative ELBO per cell of sets a controller can observe; held-out ELBO and marker r are the utility of Appendix A. Generated from `stopping_probe_mk.json` and `stopping_probe_pdc.json`.

Setting	Per-cell AUC [nested 95% CI]	Paired difference [95% CI]	Per-cell verdict	Audit	Forget median	Type	Look-alikes, mean	Retain sample, mean	Cross negatives, mean	Held-out ELBO	Marker r
Cluster 9, megakaryocytes: retrained per-cell arm 0.515, margin 0.044, retrained forget medians 1,014 to 1,154											
Baseline	0.683 [0.61, 0.76]	–	trained on the cells	1.000	491	better	511	481	525	484.5	0.762
Fisher, cap 10^{-3} (published)	0.603 [0.53, 0.68]	+0.088 [0.02, 0.16]	more identifiable	0.642	814 to 831	better	510	481	524	487.0	0.762
Fisher, cap 2×10^{-3}	0.630 [0.55, 0.70]	+0.115 [0.04, 0.19]	more identifiable	0.174	23,751 to 50,426	worse	510	481	524	766.4	0.762
Fisher, cap 3×10^{-3}	0.625 [0.55, 0.70]	+0.110 [0.03, 0.18]	more identifiable	0.085	141,436 to 166,282	worse	512	481	526	1605.4	0.762
Fisher, cap 5×10^{-3}	0.562 [0.49, 0.63]	+0.047 [-0.02, 0.12]	not resolved	0.019	116,634 to 149,565	worse	517	482	536	1444.4	0.762
Fisher, cap 10^{-2}	0.505 [0.43, 0.58]	-0.010 [-0.08, 0.06]	not resolved	0.009	5,059 to 29,373	worse	523	485	534	612.4	0.759
Retain fine-tune, 50 epochs (published)	0.556 [0.48, 0.63]	+0.041 [-0.03, 0.11]	not resolved	1.000	665 to 697	better	506	479	524	484.9	0.761
Retain fine-tune, 10 epochs	0.645 [0.57, 0.72]	+0.130 [0.05, 0.21]	more identifiable	1.000	543 to 555	better	509	481	524	484.4	0.762
Retain fine-tune, 25 epochs	0.611 [0.53, 0.68]	+0.095 [0.03, 0.16]	more identifiable	1.000	607 to 611	better	508	480	524	484.6	0.762
Retain fine-tune, 50 epochs	0.554 [0.48, 0.62]	+0.039 [-0.03, 0.11]	not resolved	1.000	688 to 700	better	506	479	524	484.9	0.762
Retain fine-tune, 100 epochs	0.560 [0.48, 0.64]	+0.045 [-0.04, 0.13]	not resolved	1.000	754 to 774	better	503	477	525	485.5	0.759
Retain fine-tune, 150 epochs	0.579 [0.51, 0.65]	+0.064 [-0.01, 0.14]	not resolved	1.000	785 to 795	better	500	476	526	485.9	0.756
Retain fine-tune, 200 epochs	0.509 [0.43, 0.59]	-0.006 [-0.09, 0.08]	not resolved	1.000	811 to 848	better	498	474	528	486.6	0.755
Cluster 8, plasmacytoid dendritic cells: retrained per-cell arm 0.487, margin 0.038, retrained forget medians 672 to 684											
Baseline	0.603 [0.56, 0.65]	–	trained on the cells	0.999	532	better	465	480	472	484.5	0.762
Fisher, cap 10^{-3} (published)	0.589 [0.54, 0.63]	+0.102 [0.06, 0.14]	more identifiable	0.484	697 to 812	worse	464	479	472	493.9	0.761
Fisher, cap 2×10^{-3}	0.572 [0.53, 0.62]	+0.085 [0.04, 0.13]	more identifiable	0.360	700 to 27,813	worse	467	480	474	1262.4	0.757
Fisher, cap 3×10^{-3}	0.553 [0.51, 0.60]	+0.066 [0.03, 0.11]	more identifiable	0.281	721 to 100,925	worse	471	481	479	1750.3	0.752
Fisher, cap 5×10^{-3}	0.535 [0.49, 0.58]	+0.048 [0.01, 0.08]	more identifiable	0.110	1,784 to 2,768	worse	475	483	479	740.5	0.755
Fisher, cap 10^{-2}	0.527 [0.48, 0.57]	+0.041 [0.004, 0.08]	more identifiable	0.069	1,345 to 1,707	worse	480	485	483	504.1	0.749
Retain fine-tune, 50 epochs (published)	0.548 [0.50, 0.59]	+0.062 [0.02, 0.10]	more identifiable	0.995	580 to 585	better	461	478	471	484.7	0.760
Retain fine-tune, 10 epochs	0.570 [0.52, 0.62]	+0.083 [0.04, 0.13]	more identifiable	0.999	548 to 552	better	464	479	471	484.4	0.761
Retain fine-tune, 25 epochs	0.582 [0.54, 0.63]	+0.095 [0.05, 0.14]	more identifiable	0.998	561 to 569	better	463	479	471	484.5	0.761
Retain fine-tune, 50 epochs	0.559 [0.51, 0.60]	+0.073 [0.03, 0.11]	more identifiable	0.986	582 to 586	better	461	478	471	484.8	0.760
Retain fine-tune, 100 epochs	0.531 [0.49, 0.57]	+0.045 [0.01, 0.09]	more identifiable	0.961	608 to 609	better	458	476	472	485.3	0.756
Retain fine-tune, 150 epochs	0.504 [0.46, 0.55]	+0.018 [-0.02, 0.06]	not resolved	0.954	613 to 618	better	454	474	472	485.7	0.754
Retain fine-tune, 200 epochs	0.525 [0.48, 0.57]	+0.039 [-0.002, 0.08]	not resolved	0.928	622 to 629	better	452	472	473	486.3	0.753

E Fisher overlap

Cluster-conditional decoder. A conditional VAE receives, at its mean output layer, the final hidden representation concatenated with a 14-dimensional one-hot cluster label, so the output weight matrix has 1,024 shared columns and 14 cluster-specific columns. The cluster-specific columns have a forget-retain Fisher cosine of 0.000, the near-orthogonality of a classifier's output weights, and the shared columns 0.379. The overlap persists through the rest of the network, with an encoder cosine of 0.617, a bottleneck cosine of 0.653, and a decoder-hidden cosine of 0.508. Under that notebook's own attacker the conditional baseline has advantage 0.759 and Fisher scrubbing with the standard settings reduces it to 0.709, since the attacker draws 64 of its 70 features from the encoder, whose overlap the conditioning does not touch.

Across datasets. Figure 5 compares the overlap on PBMC-33k, whose forget set is 0.1 percent of training, and TCGA-BRCA, whose forget set is 17 percent; the TCGA-BRCA overlap is higher at every layer, with a larger forget set and fewer output dimensions. On PBMC-68k cluster 9, with the same estimator against the 57,958 retain cells, the overlap is 0.289 in the encoder, 0.502 at the bottleneck, 0.439 in the decoder's hidden layers, and 0.020 at the decoder output, the last as low as a classifier's class-specific weights. The megakaryocytes are routed through hidden units the retain set does not use (Table 11). The hidden-unit profile of the forget set's output-layer Fisher, its column sums, puts 90 percent of its squared mass on five units. The retain set's profile ranks those five 976th, 997th, 1004th, 1012th and 1014th of its 1,024 and spreads the same share over 282 units (`proposition_test.json`, key `nu_shared_support`). On cluster 8, whose type the retain set partly covers, the output overlap is 0.128 and the encoder's 0.547.

Optimizer arms of the intervention. Tables 28 and 29 rerun the restoration intervention of Section 6 with the fine-tune under Adam and under plain gradient descent from the same saved edit, the first reading the forget and held-out fits and the second the edit vector itself.

Curvature along the edits and the linear-response test. Tables 30 and 31 test Proposition 2 on the saved moderate-damage edits, on the CPU in double precision (`scripts/edit_compensation_test.py`, `edit_compensation_states.py`, `edit_lanczos_prediction.py`; `edit_compensation.json`). The per-cell loss is the one of the Fisher caches, in evaluation mode with the latent code at its mean; fits are the deterministic estimator of Section 3. The quadratic forms along an edit, $\delta^\top K \delta$, are taken for three matrices K on a sample of 10,000 retain cells and on the forget cells. They are the diagonal empirical Fisher of the caches, $\sum_j F_{jj}\,\delta_j^2$; the empirical Fisher with its off-diagonal part, $\mathrm{mean}_n (g_n^\top \delta)^2$ from per-cell directional derivatives by central differences; and the Hessian, by a central second difference of the mean loss.

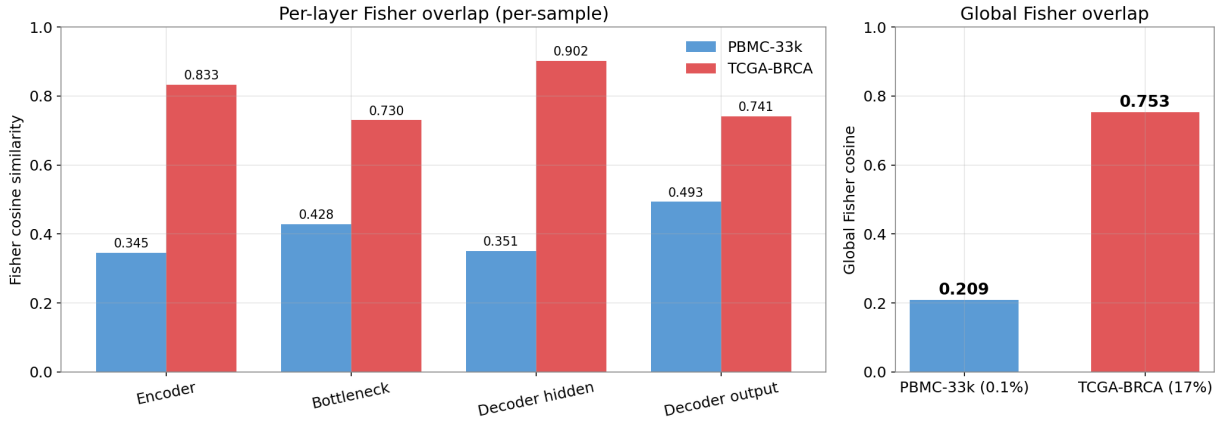


Figure 5: Fisher overlap on PBMC-33k and TCGA-BRCA, per layer (left) and over all parameters (right). Generated from the two Fisher result files.

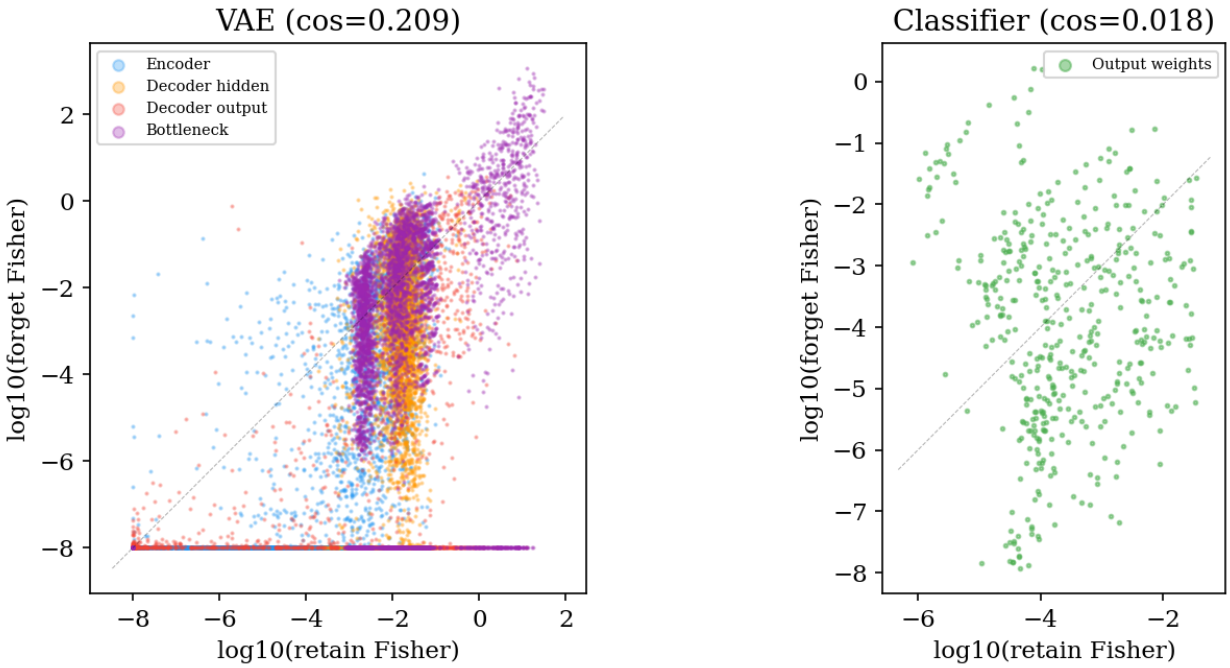


Figure 6: Per-parameter Fisher magnitude (log scale) for the forget and retain sets. Left, the VAE's parameters are correlated (log-Fisher correlation 0.71), so the same parameters matter for both sets. Right, the classifier's parameters are not (cosine 0.018), with the forget-set Fisher concentrated in a few class-specific weights.

The quadratic model's own residual $\delta_T = (I - \eta K)^T \delta$ is computed by a Lanczos iteration of 120 steps with full reorthogonalization on K-vector products over 5,000 retain cells, for two choices of K. The first is the empirical Fisher, whose product with a vector is one backward pass of the directional-derivative-weighted sum of per-cell losses. The second is the generalized Gauss–Newton matrix of the retain loss, $2J_{\text{mean}}^\top J_{\text{mean}} + J_\mu^\top J_\mu + \frac{1}{2} J_{\log\sigma^2}^\top \text{diag}(\sigma^2) J_{\log\sigma^2}$ averaged over cells, whose product is one forward-mode and one reverse-mode pass. On the reconstruction term this matrix is the Fisher information of the Gaussian observation model, and on the KL term the Hessian in $(\mu, \log\sigma^2)$. It is positive semidefinite where the

Table 28: The restoration intervention with the fine-tune run both ways from the same saved edit. From the baseline, the published Fisher scrub and the same scrub confined to the retain-active or the retain-null directions are run at the moderate and at the published damage. Each is followed by the methods' Adam fine-tune at 10^{-4} and by plain gradient descent at 10^{-3}, the largest step one retain epoch from the baseline leaves without a held-out cost, for 10, 50 and 200 epochs. The no-edit row is each optimizer's own trajectory from the baseline. Mean negative ELBO per cell on the forget cells and a held-out sample. The Adam arms are the fits Section 6 and Table 12 quote. At the published damage on PBMC-33k gradient descent has not finished repairing the retain set after 200 epochs (held-out 371 against 362). Generated from `restricted_scrub_test.json`, arms `edits_x2_opt` and `edits_opt`.

Edit	Scrub steps	Forget after edit	Held-out after edit	Adam, 10	Adam, 50	Adam, 200	SGD, 10	SGD, 50	SGD, 200	Held-out, Adam 200	Held-out, SGD 200
PBMC-33k cluster 13: baseline forget 269, held-out 363 (mean negative ELBO per cell)											
No edit	0	269	363	284	306	329	273	282	296	362	362
Published scrub (moderate damage, twice the starting forget loss)	26	572	402	292	312	332	296	292	299	362	362
Retain-active directions	26	582	406	290	310	332	292	289	298	362	362
Retain-null directions	124	549	363	431	405	387	467	438	429	362	362
Published scrub (damage of the published scrub)	100	5,601	1,271	298	309	331	321	316	309	362	371
Retain-active directions	86	5,684	1,343	294	308	331	321	304	308	362	371
Retain-null directions	268	6,239	371	540	426	404	556	440	399	362	362
PBMC-68k cluster 9: baseline forget 488, held-out 479 (mean negative ELBO per cell)											
No edit	0	488	479	540	689	800	488	499	534	481	479
Published scrub (moderate damage, twice the starting forget loss)	40	1,106	481	542	682	801	611	530	541	481	479
Retain-active directions	50	1,038	481	541	684	799	570	514	536	481	479
Retain-null directions	9	1,483	482	551	678	801	1,015	757	595	481	479
Published scrub (damage of the published scrub)	100	49,193	671	971	679	804	3,039	1,123	664	481	479
Retain-active directions	128	48,945	702	832	681	804	1,740	689	576	481	479
Retain-null directions	25	60,925	668	4,652	911	797	13,204	6,960	1,273	481	480

Table 29: The share of each edit still in the parameters after the fine-tune, $\rho = \langle \theta - \theta_{\text{base}}, \delta \rangle / \|\delta\|^2$ with the no-edit fine-tune's own projection at the same epoch subtracted, overall and on the retain-null and retain-active coordinates. Beside it, the proposition's prediction for ρ under gradient descent from the edit itself and the retain Fisher, $\sum_j \delta_j^2 (1 - \eta F_j^{\mathcal{R}})^T / \sum_j \delta_j^2$, and the forget fit of the edited and of the no-edit fine-tune at 200 epochs. A negative ρ is a fine-tune that has moved past the baseline along the edit's direction. A dash marks a quantity the row does not have, since an edit confined to one set of coordinates has no share on the other and the proposition predicts ρ for gradient descent only. Generated from `restricted_scrub_test.json`.

Edit	Fine-tune	Forget after edit	ρ, 10	ρ, 50	ρ, 200	ρ null, 200	ρ active, 200	Predicted ρ, 200	Forget, 200	No edit, 200
PBMC-33k cluster 13 (fits are the mean negative ELBO per cell)										
Published scrub (moderate damage)	Adam 10^{-4}	572	0.297	0.168	0.134	0.277	0.126	–	332	329
Published scrub (moderate damage)	SGD 10^{-3}	572	0.599	0.465	0.365	0.909	0.332	0.865	299	296
Retain-active (moderate damage)	Adam 10^{-4}	582	0.298	0.177	0.132	–	0.132	–	332	329
Retain-active (moderate damage)	SGD 10^{-3}	582	0.578	0.439	0.336	–	0.336	0.857	298	296
Retain-null (moderate damage)	Adam 10^{-4}	549	0.941	0.842	0.743	0.743	–	–	387	329
Retain-null (moderate damage)	SGD 10^{-3}	549	0.996	0.988	0.973	0.973	–	0.998	429	296
Published scrub (published damage)	Adam 10^{-4}	5,601	0.507	0.226	0.165	0.198	0.163	–	331	329
Published scrub (published damage)	SGD 10^{-3}	5,601	-0.452	-0.297	-0.211	0.921	-0.253	0.866	309	296
Retain-active (published damage)	Adam 10^{-4}	5,684	0.514	0.230	0.161	–	0.161	–	331	329
Retain-active (published damage)	SGD 10^{-3}	5,684	-0.590	-0.475	-0.295	–	-0.295	0.861	308	296
Retain-null (published damage)	Adam 10^{-4}	6,239	0.953	0.866	0.769	0.769	–	–	404	329
Retain-null (published damage)	SGD 10^{-3}	6,239	0.996	0.991	0.981	0.981	–	0.998	399	296
PBMC-68k cluster 9 (fits are the mean negative ELBO per cell)										
Published scrub (moderate damage)	Adam 10^{-4}	1,106	0.450	0.240	0.185	0.180	0.187	–	801	800
Published scrub (moderate damage)	SGD 10^{-3}	1,106	0.749	0.564	0.460	0.703	0.356	0.955	541	534
Retain-active (moderate damage)	Adam 10^{-4}	1,038	0.512	0.336	0.245	–	0.245	–	799	800
Retain-active (moderate damage)	SGD 10^{-3}	1,038	0.717	0.517	0.435	–	0.435	0.937	536	534
Retain-null (moderate damage)	Adam 10^{-4}	1,483	0.747	0.588	0.410	0.410	–	–	801	800
Retain-null (moderate damage)	SGD 10^{-3}	1,483	0.966	0.914	0.825	0.825	–	0.996	595	534
Published scrub (published damage)	Adam 10^{-4}	49,193	0.600	0.338	0.178	0.170	0.182	–	804	800
Published scrub (published damage)	SGD 10^{-3}	49,193	0.650	0.524	0.428	0.651	0.328	0.956	664	534
Retain-active (published damage)	Adam 10^{-4}	48,945	0.670	0.432	0.294	–	0.294	–	804	800
Retain-active (published damage)	SGD 10^{-3}	48,945	0.625	0.486	0.420	–	0.420	0.939	576	534
Retain-null (published damage)	Adam 10^{-4}	60,925	0.898	0.741	0.528	0.528	–	–	797	800
Retain-null (published damage)	SGD 10^{-3}	60,925	0.926	0.857	0.751	0.751	–	0.996	1,273	534

Hessian of the retain loss at the early-stopped baseline is not, since an iteration on the Hessian diverges on its negative eigenvalues, so it stands in for the curvature gradient descent follows.

The m-step recurrence is the m-point Gauss rule for the spectral measure of δ on K, and $\hat{\rho}_T$ is that rule applied to $(1 - \eta\lambda)^T$. That function's even derivatives are positive on the spectrum, so the Gauss rule is a lower bound on the quadratic model's $\hat{\rho}_T$ and the Gauss–Radau rule with a node fixed at $\lambda = 0$ an upper bound. Both are reported at every depth from 20 to 120 in Table 32, and the tables quote the bracket wherever the two differ. The predicted residual is applied to the baseline and the forget cells' fit read off the

real network, which tests the first two terms of (3). It is also added to the saved no-edit state of the same epoch, which tests the whole of (3) against the measured excess of the edited fine-tune over the no-edit one.

The measured residual is read from the saved states of the gradient-descent arms (`restricted_scrub_test.py -save_states`, re-run from the same saved edits and reproducing the arms' fits). With $r = \theta_T - \theta_0$ and Δ_T the no-edit state at the same epoch, $\rho_T = \langle r - \Delta_T, \delta\rangle / \|\delta\|^2$. The residual $r - \Delta_T$ is applied to the baseline on its own, the surviving edit $\rho_T\delta$ likewise, and the no-edit state is evaluated with the surviving edit added (superposition). Both arms are seeded alike and see the same batch sequence, so in the quadratic model $r - \Delta_T$ is the residual with the product of per-batch factors $\prod_t (I - \eta K_{b_t})$ in place of $(I - \eta K)^T$. The retain gradient at the edited point is compared with its linear prediction $g + K\delta$, and the input activation of every edited layer is read on the forget cells under the baseline, the no-edit state and the edited state, giving the perturbation $\|\delta W\, a(x) + \delta b\|^2$ the surviving edit inflicts on that layer's output.

Table 30: Curvature along the saved moderate-damage edits and the linear-response test. Quadratic forms $\delta^\top K \delta$ of the retain loss (10,000 retain cells) and of the forget loss along each edit for three matrices K: the diagonal empirical Fisher of the methods' caches, the empirical Fisher with its off-diagonal part, and the Hessian. ρ_T is the measured share of the edit left in the parameters after T epochs of gradient descent, read relative to the no-edit fine-tune. $\hat{\rho}_T$ is the quadratic model's prediction $\delta^\top (I - \eta K)^T \delta / \|\delta\|^2$ with K the empirical Fisher (F) or the generalized Gauss–Newton matrix (G), by a 120-step Lanczos iteration. Where the Gauss lower bound and the Radau upper bound differ at the second decimal the bracket gives both. Generated from `edit_compensation.json`.

Edit	$\|\delta\|^2$	Retain diag.	Retain Fisher	Retain Hessian	Forget diag.	Forget Fisher	Forget Hessian	ρ_{10}	ρ_{50}	ρ_{200}	$\hat{\rho}^F_{10}$	$\hat{\rho}^F_{200}$	$\hat{\rho}^G_{10}$	$\hat{\rho}^G_{200}$
PBMC-33k cluster 13														
Published scrub (moderate damage)	0.578	4.49×10^{-3}	23.2	46.2	0.067	1,086	545	0.60	0.46	0.36	0.81	0.80	0.66	0.51
Retain-active (moderate damage)	0.591	4.89×10^{-3}	24.9	49.7	0.0688	1,111	568	0.58	0.44	0.34	0.80	0.79	0.64	0.48
Retain-null (moderate damage)	39.2	4.40×10^{-3}	0.104	0.109	3.39	340	291	1.00	0.99	0.97	1.00	1.00	1.00	0.99
PBMC-68k cluster 9														
Published scrub (moderate damage)	0.0197	2.11×10^{-5}	9.70×10^{-3}	4.64×10^{-3}	5.11	511	177	0.75	0.56	0.46	1.00	1.00	0.89	0.71
Retain-active (moderate damage)	0.0218	3.31×10^{-5}	0.0141	6.37×10^{-3}	7.42	509	174	0.72	0.52	0.43	1.00	1.00	0.86	0.67
Retain-null (moderate damage)	0.211	2.03×10^{-5}	2.46×10^{-3}	1.20×10^{-3}	16.9	1,280	320	0.97	0.91	0.82	1.00	1.00	0.99	0.94

Table 31: Where the forget damage of an edit goes under gradient descent, after 10 and after 200 retain epochs, as shares of the damage the edit did at the baseline. Left: the excess of the edited over the no-edit fine-tune, measured. Superposed: the no-edit state with the surviving share ρ_T of the edit added. Edit alone: the surviving edit $\rho_T\delta$ applied to the baseline. Residual: the fine-tune's displacement with the no-edit path subtracted, $r - \Delta_T$, applied to the baseline. Quadratic: the residual the quadratic model predicts with the empirical Fisher (F) or the generalized Gauss–Newton matrix (G), applied to the baseline, which tests the first two terms of (3). Quad. G sup.: the same residual added to the no-edit state, the model's prediction of the Left column. Brackets give the Gauss and Radau values where they differ. Generated from `edit_compensation.json`.

Edit	Damage	Left (10)	Superposed (10)	Edit alone (10)	Residual (10)	Quad. F (10)	Quad. G (10)	Quad. G sup. (10)	Left (200)	Superposed (200)	Edit alone (200)	Residual (200)	Quad. F (200)	Quad. G (200)	Quad. G sup. (200)
PBMC-33k cluster 13															
Published scrub (moderate damage)	302	0.07	0.59	0.61	0.10	0.43	0.17	0.14	0.01	0.24	0.29	0.03	0.42	0.07	0.05
Retain-active (moderate damage)	311	0.06	0.56	0.58	0.09	0.41	0.15	0.12	0.01	0.20	0.25	0.02	0.40	0.06	0.04
Retain-null (moderate damage)	281	0.69	1.00	1.00	0.70	0.98	0.97	0.97	0.47	1.05	0.98	0.51	0.98	0.95	1.02
PBMC-68k cluster 9															
Published scrub (moderate damage)	613	0.20	0.27	0.33	0.24	0.97	0.53	0.43	0.01	0.03	0.08	0.04	0.97	0.20	0.08
Retain-active (moderate damage)	547	0.15	0.23	0.30	0.19	0.97	0.48	0.36	0.00	0.02	0.07	0.03	0.97	0.16	0.05
Retain-null (moderate damage)	983	0.54	0.86	0.88	0.53	0.99	0.87	0.85	0.06	0.29	0.49	0.13	0.99	0.54	0.36

F The same two mechanisms where the generating process is known

A simulation in which the rare component's distance from the common data and the fraction of it forgotten are set by hand shows both mechanisms outside gene expression (Figure 7; Tables 39, 40 and 41 in Appendix G). Ten common Gaussian components in fifty dimensions and a rare component of one hundred points, sixty in training, are modeled by a Gaussian VAE. The rare component is placed 1, 2, 4, or 8 standard deviations from its nearest common component, either all sixty or thirty of its training points are forgotten, and the instruments are the paper's. The cross-component audit detects a group signal, the retrained models fitting the whole forgotten component worse than the baseline does. That signal exists only when the retain set cannot stand in for the forget set, when the whole component is forgotten and it sits 4 or 8 standard

Table 32: Convergence of the Lanczos prediction with the Krylov depth m, for the published scrub at moderate damage on each cohort and each curvature. The columns give the lowest Ritz value and its spectral weight, and $\hat{\rho}_T$ by the m-point Gauss rule (a lower bound) and by the Gauss–Radau rule with a node at zero (an upper bound) after 10, 50 and 200 epochs. The Radau rule is omitted where the Gauss rule already holds a node at zero. Generated from `edit_compensation.json`.

m	$\lambda_{\min}$	$w_{\min}$	$\hat{\rho}_{10}$	$\hat{\rho}_{50}$	$\hat{\rho}_{200}$
PBMC-33k cluster 13, empirical Fisher					
20	0	0.835	0.835	0.835	0.835
40	0	0.809	0.810	0.809	0.809
60	0	0.803	0.807	0.803	0.803
80	0	0.802	0.807	0.802	0.802
100	0	0.802	0.807	0.802	0.802
120	0	0.802	0.807	0.802	0.802
PBMC-33k cluster 13, Gauss–Newton					
20	0.203	0.775	[0.620, 0.689]	[0.254, 0.686]	[0.009, 0.686]
40	0.0508	0.690	0.662	[0.522, 0.602]	[0.226, 0.600]
60	0.0201	0.633	0.662	[0.568, 0.575]	[0.407, 0.551]
80	9.71×10^{-3}	0.592	0.662	0.572	[0.478, 0.524]
100	5.42×10^{-3}	0.563	0.662	0.572	[0.500, 0.512]
120	3.39×10^{-3}	0.543	0.662	0.572	[0.506, 0.508]
PBMC-68k cluster 9, empirical Fisher					
20	0	0.998	0.998	0.998	0.998
40	0	0.998	0.998	0.998	0.998
60	0	0.998	0.998	0.998	0.998
80	0	0.998	0.998	0.998	0.998
100	0	0.998	0.998	0.998	0.998
120	0	0.998	0.998	0.998	0.998
PBMC-68k cluster 9, Gauss–Newton					
20	0.0547	0.990	[0.875, 0.891]	[0.532, 0.815]	[0.083, 0.812]
40	0.0178	0.868	0.886	[0.711, 0.762]	[0.387, 0.749]
60	4.27×10^{-3}	0.776	0.886	[0.749, 0.754]	[0.639, 0.733]
80	2.00×10^{-3}	0.753	0.886	0.752	[0.688, 0.720]
100	1.26×10^{-3}	0.742	0.886	0.752	[0.701, 0.711]
120	9.01×10^{-4}	0.734	0.886	0.752	[0.705, 0.707]

deviations away (audit 0.71 and 1.00). Otherwise the retrained models fit the forget points as well as the baseline does.

The within-component attacker finds a per-point signal in the latent features in every configuration, 0.90 for the baseline where the rare component overlaps the common data and 0.70 where it is isolated, against 0.42 to 0.57 for retrained models, while the within-component audit stays near chance; the training points are placed differently from held-out members of their own component rather than fitted better. Retain-only fine-tuning erodes the group signal within ten epochs where the retain gradient reclaims the region and over the whole hundred where nothing acts on it, and erodes the per-point signal slowly everywhere; no post-hoc method removes it. The continuous objective drives the audit to between 0.00 and 0.19 in every configuration and leaves a latent artifact that the cross-component attacker detects.

Given the retrained model's level, the targeted fine-tune matches the retrained models' fit of the forget points. Its audit of 0.46 to 0.65 lies inside the range the three retrained models span against one another in seven of the eight configurations and 0.01 above it in the eighth, and the attackers still separate the forget points. Given either level a controller could observe, it moves them the wrong way or not at all. That range is as reference-dependent here as on the real data (against one seed it can sit entirely below 0.5), so the appendix tables report it over every ordered pair.

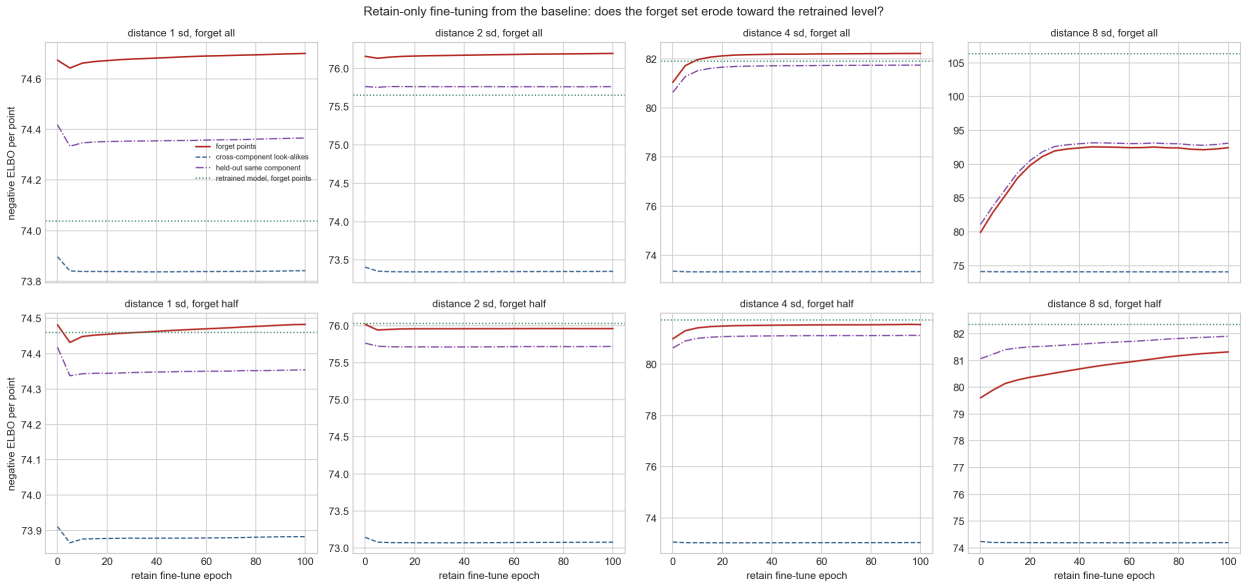


Figure 7: Retain-only fine-tuning in the simulation, by the rare component's distance from the common data (columns) and forgotten fraction (rows). Forget points (red), their nearest held-out points of other components (blue), the held-out members of the component (purple), and the retrained models' level (dotted).

G Additional tables

Table 33: Every method on the two PBMC-68k forget sets. One cross-type attacker, trained on the baseline, scores every checkpoint in one CPU pass. The forget-cell fit is the negative ELBO per cell, ranged over seeds. The audit is the cross-type likelihood-ratio AUC against clean retrain seed 42 with a nested interval, the reference row being the range the other seven seeds span. The per-cell columns are the standardized logistic refit against the within-type negatives, out of fold, with a nested interval over seeds and the shared cells, and its difference from the eight retrained models paired on those cells. The status applies the criterion of Table 5. Generated from `results_mk.json`, `results_pdc.json`, and `percell_paired_{mk,pdc}.json`.

Method	Seeds	Forget median	Per-cell AUC, within [nested 95% CI]	Paired difference [95% CI]	Cells	Reference-free audit [95% CI]	Median gap [95% CI]	Type	ELBO	Marker r
Cluster 9, megakaryocytes (107 forget cells, 71 within-type and 143 cross-type negatives; margin 0.044)										
Baseline (no unlearning)	1	491	0.683 [0.61, 0.76]	–	trained on the cells	1.000 [1.00, 1.00]	-585 [-628, -548]	trained on the cells	484.5	0.762
Clean retrain	8	1,014 to 1,154	0.472 to 0.538 [0.45, 0.57]	–	reference	0.592 [0.41, 0.66]	–	reference	488.0	0.761
Retain-only fine-tune	3	665 to 697	0.542 to 0.564 [0.48, 0.63]	+0.041 [-0.03, 0.11]	not resolved	1.000 [1.00, 1.00]	-395 [-441, -354]	fitted better	484.9	0.761
Gradient ascent + fine-tune	3	615 to 634	0.568 to 0.606 [0.51, 0.65]	+0.066 [-0.01, 0.14]	not resolved	1.000 [1.00, 1.00]	-452 [-496, -415]	fitted better	484.6	0.762
SSD	3	607 to 614	0.635 to 0.664 [0.57, 0.72]	+0.132 [0.06, 0.20]	more identifiable	1.000 [1.00, 1.00]	-464 [-503, -431]	fitted better	484.8	0.761
SCRUB	3	572 to 622	0.596 to 0.609 [0.53, 0.67]	+0.086 [0.02, 0.15]	more identifiable	0.994 [0.97, 1.00]	-480 [-530, -424]	fitted better	484.8	0.762
Fisher scrubbing	3	814 to 831	0.591 to 0.622 [0.53, 0.68]	+0.088 [0.02, 0.16]	more identifiable	0.653 [0.53, 0.76]	-251 [-356, -56]	fitted better	487.0	0.762
Extra-gradient ($\lambda = 10$)	3	475 to 489	0.840 to 0.914 [0.81, 0.93]	+0.361 [0.29, 0.44]	more identifiable	0.999 [1.00, 1.00]	-594 [-635, -556]	fitted better	486.0	0.762
Contrastive latent	3	663 to 755	0.549 to 0.564 [0.48, 0.63]	+0.043 [-0.03, 0.11]	not resolved	0.971 [0.92, 1.00]	-365 [-444, -295]	fitted better	485.7	0.761
Retain-only fine-tune, 200 epochs	3	820 to 846	0.506 to 0.539 [0.45, 0.59]	+0.008 [-0.06, 0.08]	not resolved	1.000 [1.00, 1.00]	-247 [-290, -205]	fitted better	486.7	0.756
Cluster 8, plasmacytoid dendritic cells (302 forget cells, 201 within-type and 445 cross-type negatives; margin 0.038)										
Baseline (no unlearning)	1	532	0.603 [0.56, 0.65]	–	trained on the cells	0.999 [1.00, 1.00]	-148 [-160, -131]	trained on the cells	484.5	0.762
Clean retrain	8	672 to 684	0.450 to 0.503 [0.45, 0.52]	–	reference	0.510 [0.48, 0.53]	–	reference	486.7	0.756
Retain-only fine-tune	3	580 to 585	0.542 to 0.558 [0.50, 0.59]	+0.062 [0.02, 0.10]	more identifiable	0.994 [0.99, 1.00]	-97 [-113, -84]	fitted better	484.7	0.760
Gradient ascent + fine-tune	3	568 to 571	0.550 to 0.584 [0.52, 0.61]	+0.079 [0.04, 0.12]	more identifiable	0.997 [0.99, 1.00]	-110 [-123, -96]	fitted better	484.6	0.760
SSD	3	561 to 572	0.560 to 0.582 [0.53, 0.62]	+0.087 [0.04, 0.13]	more identifiable	0.999 [1.00, 1.00]	-112 [-130, -99]	fitted better	484.7	0.761
SCRUB	3	555 to 557	0.540 to 0.602 [0.52, 0.62]	+0.084 [0.03, 0.13]	more identifiable	0.999 [1.00, 1.00]	-124 [-137, -111]	fitted better	484.5	0.762
Fisher scrubbing	3	697 to 812	0.583 to 0.593 [0.54, 0.63]	+0.102 [0.06, 0.14]	more identifiable	0.489 [0.39, 0.59]	+60 [-6, +226]	not resolved	493.9	0.761
Extra-gradient ($\lambda = 10$)	3	492 to 508	0.703 to 0.828 [0.70, 0.83]	+0.283 [0.21, 0.35]	more identifiable	0.929 [0.80, 1.00]	-180 [-196, -160]	fitted better	486.4	0.762
Contrastive latent	3	747 to 777	0.630 to 0.633 [0.58, 0.68]	+0.145 [0.10, 0.19]	more identifiable	0.011 [0.00004, 0.04]	+87 [+61, +106]	fitted worse	487.9	0.760

Table 34: Multiplicity. For every arm of the attacker test, the three one-sided bootstrap p-values of the paired difference from the retrained arm, raw and Holm-adjusted within the arm's family, with the verdict each set gives. They are $p_>$ for a difference above zero, $p_<$ for one below zero, and $p_\approx$ for the two one-sided tests of equivalence within the margin, each read at 0.025. The Type columns give the two-sided p-value of the type-level instrument, raw and Holm-adjusted within the same family, with its verdict; the instrument is the reference-free cross-type audit on PBMC-33k and the paired median gap relative to the negatives on PBMC-68k. Generated from `paired_diff_33k.json`, `percell_paired_{mk,pdc}.json`, and `audit_reference_free.json`.

Method	Paired difference [95% CI]	$p_>$	$p_<$	$p_\approx$	Verdict	Holm $p_>$	Holm $p_<$	Holm $p_\approx$	Holm verdict	Type p (raw / Holm)	Type verdict
PBMC-33k cluster 13, transferred attacker, published settings (14 arms), margin 0.107											
Retain-only fine-tune	+0.141 [0.10, 0.19]	<0.001	1.000	0.953	more identifiable	0.001	1.000	1.000	more identifiable	<0.001 / 0.014	fitted better
Gradient ascent + fine-tune	+0.177 [0.14, 0.22]	<0.001	1.000	1.000	more identifiable	0.001	1.000	1.000	more identifiable	<0.001 / 0.014	fitted better
SSD	+0.172 [0.13, 0.22]	<0.001	1.000	0.999	more identifiable	0.001	1.000	1.000	more identifiable	<0.001 / 0.014	fitted better
SCRUB	+0.187 [0.14, 0.24]	<0.001	1.000	0.999	more identifiable	0.001	1.000	1.000	more identifiable	<0.001 / 0.014	fitted better
Moon feature unlearning	+0.211 [0.16, 0.26]	<0.001	1.000	1.000	more identifiable	0.001	1.000	1.000	more identifiable	<0.001 / 0.014	fitted better
Fisher scrubbing	+0.144 [0.10, 0.19]	<0.001	1.000	0.941	more identifiable	0.001	1.000	1.000	more identifiable	<0.001 / 0.014	fitted better
Extra-gradient ($\lambda = 10$)	-0.117 [-0.19, -0.05]	1.000	<0.001	0.618	less identifiable	1.000	0.004	1.000	less identifiable	0.583 / 1.000	not resolved
Extra-gradient ($\lambda = 1$)	-0.192 [-0.28, -0.12]	1.000	<0.001	0.995	less identifiable	1.000	0.001	1.000	less identifiable	<0.001 / 0.014	fitted worse
Contrastive latent	-0.359 [-0.44, -0.28]	1.000	<0.001	1.000	less identifiable	1.000	0.001	1.000	less identifiable	<0.001 / 0.014	fitted worse
Frozen critics, one critic, $\lambda = 5$	-0.526 [-0.63, -0.42]	1.000	<0.001	1.000	less identifiable	1.000	0.001	1.000	less identifiable	<0.001 / 0.014	fitted worse
Frozen critics, one critic, $\lambda = 10$	-0.526 [-0.63, -0.41]	1.000	<0.001	1.000	less identifiable	1.000	0.001	1.000	less identifiable	<0.001 / 0.014	fitted worse
Frozen critics, three critics, $\lambda = 5$	-0.521 [-0.63, -0.41]	1.000	<0.001	1.000	less identifiable	1.000	0.001	1.000	less identifiable	<0.001 / 0.014	fitted worse
Frozen critics, three critics, $\lambda = 10$	-0.521 [-0.63, -0.41]	1.000	<0.001	1.000	less identifiable	1.000	0.001	1.000	less identifiable	<0.001 / 0.014	fitted worse
DP-SGD ($\varepsilon = 10$, from scratch)	-0.063 [-0.12, -0.01]	0.990	0.010	0.062	less identifiable	1.000	0.069	0.867	not resolved	0.636 / 1.000	as retrained
PBMC-33k cluster 13, transferred attacker, step and cap ablations (9 arms), margin 0.107											
Fisher, step cap 3×10^{-3}	-0.217 [-0.30, -0.14]	1.000	<0.001	0.997	less identifiable	1.000	<0.001	1.000	less identifiable	<0.001 / 0.009	fitted worse
Fisher, step cap 10^{-2}	-0.084 [-0.14, -0.02]	0.997	0.004	0.228	less identifiable	1.000	0.025	1.000	not resolved	0.034 / 0.102	fitted worse
Fisher, step cap 1.5×10^{-3}	+0.133 [0.09, 0.18]	<0.001	1.000	0.855	more identifiable	<0.001	1.000	1.000	more identifiable	0.002 / 0.010	fitted better
Fisher, step cap 2×10^{-3}	+0.049 [-0.01, 0.11]	0.054	0.946	0.030	not resolved	0.272	1.000	0.246	not resolved	0.542 / 0.542	not resolved
Fisher, step cap 2.5×10^{-3}	-0.044 [-0.11, 0.02]	0.914	0.086	0.027	not resolved	1.000	0.518	0.246	not resolved	0.197 / 0.394	not resolved
Fisher, step cap 3×10^{-3} (sweep)	-0.150 [-0.24, -0.07]	1.000	<0.001	0.845	less identifiable	1.000	<0.001	1.000	less identifiable	0.003 / 0.012	fitted worse
Gradient ascent, lr 10^{-4}	+0.178 [0.14, 0.22]	<0.001	1.000	1.000	more identifiable	<0.001	1.000	1.000	more identifiable	<0.001 / 0.009	fitted better
Gradient ascent, lr 10^{-3}	+0.177 [0.13, 0.22]	<0.001	1.000	0.999	more identifiable	<0.001	1.000	1.000	more identifiable	<0.001 / 0.009	fitted better
Gradient ascent, lr 10^{-2}	+0.170 [0.13, 0.22]	<0.001	1.000	0.999	more identifiable	<0.001	1.000	1.000	more identifiable	<0.001 / 0.009	fitted better
PBMC-33k cluster 13, transferred attacker, targeted and long fine-tunes (4 arms), margin 0.107											
Likelihood-targeted, look-alike level	+0.223 [0.12, 0.33]	<0.001	1.000	0.985	more identifiable	<0.001	1.000	1.000	more identifiable	0.115 / 0.345	not resolved
Likelihood-targeted, retrained level	+0.052 [-0.07, 0.17]	0.195	0.805	0.183	not resolved	0.195	1.000	0.550	not resolved	0.432 / 0.432	not resolved
Likelihood-targeted, disjoint look-alikes	+0.306 [0.20, 0.42]	<0.001	1.000	1.000	more identifiable	<0.001	1.000	1.000	more identifiable	0.017 / 0.068	fitted better
Retain-only fine-tune, 200 epochs	+0.039 [0.01, 0.07]	0.005	0.995	<0.001	more identifiable	0.011	1.000	<0.001	more identifiable	0.120 / 0.345	not resolved
PBMC-68k cluster 9, megakaryocytes, within-type refit (8 arms), margin 0.044											
Retain-only fine-tune	+0.041 [-0.03, 0.11]	0.111	0.889	0.470	not resolved	0.314	1.000	1.000	not resolved	<0.001 / 0.008	fitted better
Gradient ascent + fine-tune	+0.066 [-0.01, 0.14]	0.037	0.963	0.729	not resolved	0.148	1.000	1.000	not resolved	<0.001 / 0.008	fitted better
SSD	+0.132 [0.06, 0.20]	<0.001	1.000	0.990	more identifiable	0.003	1.000	1.000	more identifiable	<0.001 / 0.008	fitted better
SCRUB	+0.086 [0.02, 0.15]	0.008	0.992	0.893	more identifiable	0.046	1.000	1.000	not resolved	<0.001 / 0.008	fitted better
Fisher scrubbing	+0.088 [0.02, 0.16]	0.009	0.992	0.889	more identifiable	0.046	1.000	1.000	not resolved	0.027 / 0.027	fitted better
Extra-gradient ($\lambda = 10$)	+0.361 [0.29, 0.44]	<0.001	1.000	1.000	more identifiable	<0.001	1.000	1.000	more identifiable	<0.001 / 0.008	fitted better
Contrastive latent	+0.043 [-0.03, 0.11]	0.105	0.895	0.494	not resolved	0.314	1.000	1.000	not resolved	<0.001 / 0.008	fitted better
Retain-only fine-tune, 200 epochs	+0.008 [-0.06, 0.08]	0.405	0.596	0.165	not resolved	0.405	1.000	1.000	not resolved	<0.001 / 0.008	fitted better
PBMC-68k cluster 8, plasmacytoid dendritic cells, within-type refit (7 arms), margin 0.038											
Retain-only fine-tune	+0.062 [0.02, 0.10]	0.002	0.998	0.877	more identifiable	0.002	1.000	1.000	more identifiable	<0.001 / 0.007	fitted better
Gradient ascent + fine-tune	+0.079 [0.04, 0.12]	<0.001	1.000	0.970	more identifiable	<0.001	1.000	1.000	more identifiable	<0.001 / 0.007	fitted better
SSD	+0.087 [0.04, 0.13]	<0.001	1.000	0.989	more identifiable	<0.001	1.000	1.000	more identifiable	<0.001 / 0.007	fitted better
SCRUB	+0.084 [0.03, 0.13]	<0.001	1.000	0.967	more identifiable	<0.001	1.000	1.000	more identifiable	<0.001 / 0.007	fitted better
Fisher scrubbing	+0.102 [0.06, 0.14]	<0.001	1.000	0.999	more identifiable	<0.001	1.000	1.000	more identifiable	0.080 / 0.080	not resolved
Extra-gradient ($\lambda = 10$)	+0.283 [0.21, 0.35]	<0.001	1.000	1.000	more identifiable	<0.001	1.000	1.000	more identifiable	<0.001 / 0.007	fitted better
Contrastive latent	+0.145 [0.10, 0.19]	<0.001	1.000	1.000	more identifiable	<0.001	1.000	1.000	more identifiable	<0.001 / 0.007	fitted worse

Table 35: Every phase of Fisher scrubbing at three step caps and of gradient ascent on PBMC-33k. Fits are the mean negative ELBO per cell for the forget, retain, and held-out cells; the attacker's AUC is given with its advantage, and the audit AUC with its interval. Generated from `phase_audit.json`.

Stage	Forget, mean	Retain, mean	Held out, mean	Fit-excess attacker, AUC	Advantage	Audit AUC [95% CI]
Baseline	269.5	363.0	362.9	0.793	0.59	0.996 [0.99, 1.00]
Fisher scrub, max step 10^{-3}, after scrub	5569.4	1250.1	1284.1	0.016	0.97	0.015 [0.003, 0.03]
Fisher scrub, max step 10^{-3}, after retain fine-tune	307.6	362.4	362.4	0.666	0.33	0.987 [0.97, 1.00]
Fisher scrub, max step 3×10^{-3}, after scrub	59489.8	6063.4	6561.5	0.170	0.66	0.014 [0.002, 0.03]
Fisher scrub, max step 3×10^{-3}, after retain fine-tune	377.2	362.7	362.7	0.386	0.23	0.173 [0.05, 0.31]
Fisher scrub, max step 10^{-2}, after scrub	216483.8	173734.5	171780.9	0.487	0.03	0.418 [0.34, 0.50]
Fisher scrub, max step 10^{-2}, after retain fine-tune	723.2	370.2	369.2	0.376	0.25	0.155 [0.05, 0.28]
Gradient ascent, after ascent	269.5	363.0	362.9	0.802	0.60	0.996 [0.99, 1.00]
Gradient ascent, after retain fine-tune	296.3	361.6	362.2	0.716	0.43	0.996 [0.99, 1.00]

Table 36: Long retain-only fine-tuning at the methods' learning rate, ranges over three seeds, on PBMC-33k cluster 13 and TCGA-BRCA Basal, with NB42's own per-epoch instruments: mean negative ELBO per cell, the notebook's own instance of the attacker trained on the baseline, and the likelihood-ratio audit against clean retrain seed 42. Under the paper's instruments the 200-epoch PBMC checkpoints score 0.568 with the canonical attacker (the paired test of Section 5, `paired_diff_33k.json`) and 0.688 [0.48, 0.88] under the reference-free audit, not resolved (Table 23). The TCGA run overfits its 767-patient retain set after epoch 100 and every group is fitted better again, so its first 100 epochs are the informative ones. Generated from `long_finetune.json` under `outputs/p8`.

Retain epochs	PBMC forget negELBO, mean	PBMC attacker (the run's own instance)	PBMC audit vs. retrain seed 42 [nested 95% CI at 200]	TCGA forget negELBO, mean	TCGA unseen Basal, mean	TCGA audit vs. the TCGA retrain [95% CI at 200]
0	269.5	0.795 to 0.799	0.996	1116.5	1086.0	0.833
10	285.0 to 286.7	0.734 to 0.750	0.995 to 0.996	1253.1 to 1277.8	1220.4 to 1247.7	0.338 to 0.691
50	303.6 to 306.8	0.669 to 0.688	0.992 to 0.994	1256.9 to 1257.5	1225.5 to 1228.3	0.482 to 0.503
100	317.6 to 320.3	0.580 to 0.615	0.944 to 0.971	1216.2 to 1257.1	1190.7 to 1224.4	0.419 to 0.494
200	327.7 to 331.3	0.552 to 0.561	0.519 to 0.662 [0.46, 0.68]	938.2 to 1023.8	866.6 to 952.9	0.638 to 0.671 [0.61, 0.71]
Retrained models	337.2	0.529 (canonical instance)	0.24 to 0.52	1277.4	1246.4	–

Table 37: Long retain-only fine-tuning on PBMC-68k cluster 9 (`scripts/long_finetune_pbmc68k.py`, the run whose 200-epoch checkpoints Table 33 scores), ranges over three seeds. Fits are mean negative ELBO per cell; the attacker is the script's own instance of the cross-type attacker trained on the baseline, and the audit the cross-type likelihood-ratio AUC against clean retrain seed 42, with the nested interval at 200 epochs. The last row gives the eight clean retrains' ranges, their attacker range from Table 33, and the range the audit spans over every ordered pair of them. Generated from `long_finetune.json` under `outputs/pbmc68k`.

Retain epochs	Forget cells, mean	Held-out megakaryocytes, mean	Retain sample, mean	Held-out cells, mean	Fit-excess attacker, cross	Audit vs. retrain seed 42, cross [nested 95% CI at 200]
0	488.3	521.7	481.7	484.4	0.964 to 0.967	1.000
10	543.0 to 558.1	559.9 to 573.2	481.0 to 481.1	484.0 to 484.1	0.945 to 0.951	1.000
50	659.0 to 678.5	658.4 to 675.3	479.3 to 479.4	483.9 to 483.9	0.740 to 0.802	1.000
100	728.0 to 757.1	722.6 to 751.9	477.6 to 477.7	483.9 to 484.0	0.467 to 0.598	1.000
200	790.7 to 813.4	788.4 to 808.6	474.4 to 474.5	484.7 to 484.9	0.312 to 0.380	1.000 [1.00, 1.00]
Retrained models	981.2 to 1106.3	975.1 to 1108.8	481.2 to 482.3	483.9 to 485.0	0.007 to 0.017	0.03 to 0.97 (every ordered pair)

Table 38: How the baseline and the five retrained models fit the forget cells and the three candidate target sets of Section 6.2, negative ELBO per cell. Generated from `calibration_levels.json`.

Model	Forget cells, mean	Matched negatives, mean	Ten nearest negatives, mean	Disjoint neighbors, mean
Baseline	269.5	333.1	286.5	291.5
Clean retrains (range over five seeds)	334.0 to 342.1	333.2 to 333.9	297.9 to 301.0	292.1 to 293.0

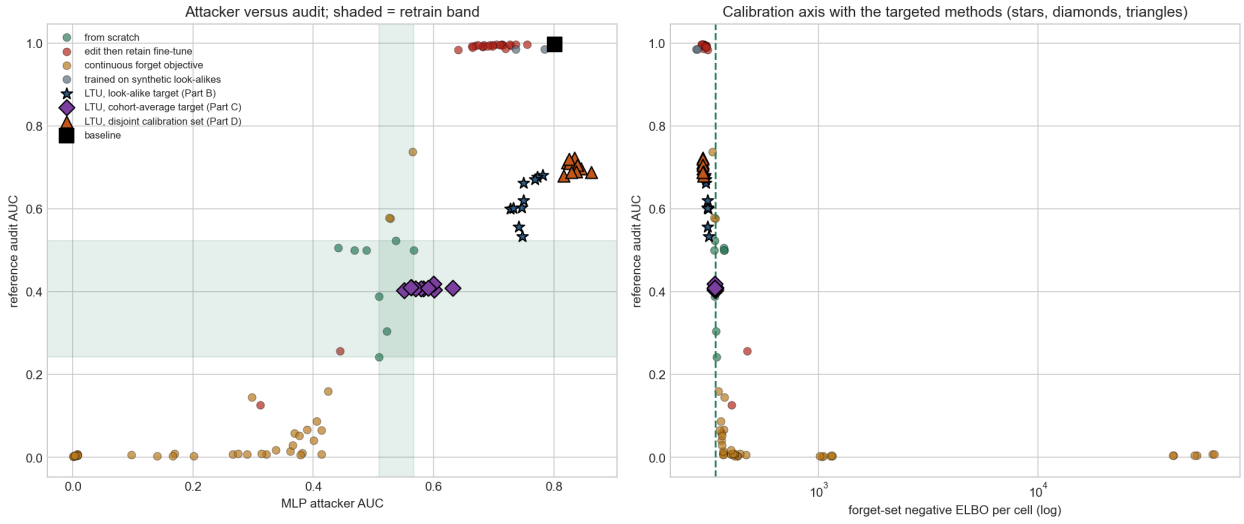


Figure 8: Left, attacker AUC against audit AUC for every PBMC-33k checkpoint, with the ranges the retrained models span shaded and the three targeted fine-tunes marked (stars, nearest negatives; diamonds, the retrained models' level; triangles, disjoint neighbors). Right, the same checkpoints by the forget cells' negative ELBO.

Table 39: The simulation by configuration. The columns give the forget negative ELBO under the baseline and the retrained models, the audit of the baseline with the range that retrained models span, and the within-component attacker on the baseline, the retrained model, and the model after 100 retain epochs. Then follow the forget fit after 10 and 100 retain epochs, the audit and within-component attacker of the fine-tune given the retrained models' level, and the forget fit after targeting the nearest held-out points of other components. Generated from `synthetic_support.json`.

Configuration	Forget fit (mean), base / retrained	Audit, base [95% CI] (retrained range)	Within attacker, base / retr. / FT100	FT forget (mean) at 10 / 100	Retrained level as target, audit [95% CI] / within	Neighbor target, forget (mean)
1 sd, all	74.7 / 74.0	0.33 [0.25, 0.42] (0.28 to 0.72)	0.91 / 0.50 / 0.72	74.7 / 74.7	0.59 [0.48, 0.70], 0.70	74.7
1 sd, half	74.5 / 74.5	0.55 [0.43, 0.67] (0.36 to 0.64)	0.91 / 0.46 / 0.69	74.5 / 74.5	0.62 [0.46, 0.77], 0.73	75.2
2 sd, all	76.2 / 75.7	0.36 [0.28, 0.45] (0.27 to 0.73)	0.90 / 0.50 / 0.73	76.1 / 76.2	0.57 [0.48, 0.67], 0.84	74.5
2 sd, half	76.0 / 76.0	0.61 [0.49, 0.71] (0.35 to 0.65)	0.90 / 0.50 / 0.76	75.9 / 75.9	0.64 [0.49, 0.78], 0.72	74.2
4 sd, all	81.0 / 81.9	0.71 [0.63, 0.80] (0.36 to 0.64)	0.90 / 0.50 / 0.78	81.9 / 82.2	0.65 [0.55, 0.75], 0.49	75.2
4 sd, half	81.0 / 81.7	0.88 [0.79, 0.96] (0.26 to 0.74)	0.90 / 0.50 / 0.79	81.4 / 81.6	0.65 [0.49, 0.80], 0.74	74.8
8 sd, all	79.9 / 106.4	1.00 [1.00, 1.00] (0.25 to 0.75)	0.70 / 0.55 / 0.57	86.9 / 105.4	0.46 [0.34, 0.57], 0.59	74.7
8 sd, half	79.6 / 82.4	0.99 [0.98, 1.00] (0.23 to 0.77)	0.69 / 0.45 / 0.61	80.2 / 81.4	0.51 [0.34, 0.68], 0.69	74.8

Table 40: Three-component simulation with no rare component, in which the whole of component 0 (1,642 points) or half of it (821) is forgotten; forgetting a whole component is a group effect that only the cross-component instruments see, and forgetting half of it leaves nothing per point to unlearn. Generated from `mog_results.json`.

Scenario and model	Forget negELBO, mean	Attacker, cross	Attacker, within	Audit, cross [95% CI]	Audit, within [95% CI]
whole component: baseline	8.9	0.946	0.590	1.000 [1.00, 1.00]	0.511 [0.48, 0.54]
whole component: retrain	11.9	0.996	0.513	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
whole component: retain fine-tune, 10 epochs	9.4	0.969	0.581	1.000 [1.00, 1.00]	0.510 [0.48, 0.54]
whole component: retain fine-tune, 100 epochs	11.9	0.996	0.508	0.570 [0.55, 0.59]	0.500 [0.47, 0.53]
whole component: gradient ascent + fine-tune	9.6	0.944	0.581	1.000 [1.00, 1.00]	0.510 [0.48, 0.54]
whole component: forget push (continuous)	10.3	0.742	0.533	1.000 [1.00, 1.00]	0.511 [0.48, 0.54]
whole component: targeted, oracle level	11.8	0.990	0.492	0.670 [0.65, 0.69]	0.511 [0.48, 0.54]
whole component: targeted, cross look-alikes	9.0	0.961	0.584	1.000 [1.00, 1.00]	0.510 [0.48, 0.54]
whole component: targeted, held-out same component	9.1	0.992	0.541	1.000 [1.00, 1.00]	0.511 [0.48, 0.54]
whole component: retrain seed 101	11.9	0.976	0.494	0.460 [0.44, 0.48]	0.504 [0.47, 0.53]
whole component: retrain seed 102	11.9	0.897	0.501	0.479 [0.46, 0.50]	0.482 [0.45, 0.51]
half of the component: baseline	9.0	0.948	0.634	0.996 [0.99, 1.00]	0.519 [0.49, 0.55]
half of the component: retrain	9.4	0.984	0.504	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
half of the component: retain fine-tune, 10 epochs	9.2	0.975	0.619	0.937 [0.92, 0.95]	0.524 [0.49, 0.56]
half of the component: retain fine-tune, 100 epochs	9.4	0.995	0.576	0.495 [0.46, 0.52]	0.520 [0.49, 0.56]
half of the component: gradient ascent + fine-tune	9.3	0.966	0.620	0.904 [0.89, 0.92]	0.523 [0.49, 0.56]
half of the component: forget push (continuous)	9.4	0.861	0.541	0.707 [0.68, 0.73]	0.525 [0.49, 0.56]
half of the component: targeted, oracle level	9.4	1.000	0.522	0.660 [0.63, 0.69]	0.518 [0.48, 0.55]
half of the component: targeted, cross look-alikes	9.0	0.991	0.568	0.893 [0.88, 0.91]	0.523 [0.49, 0.56]
half of the component: targeted, held-out same component	9.1	0.992	0.586	0.879 [0.86, 0.90]	0.521 [0.49, 0.55]
half of the component: retrain seed 101	9.4	0.724	0.495	0.470 [0.44, 0.50]	0.530 [0.50, 0.56]
half of the component: retrain seed 102	9.4	0.989	0.518	0.425 [0.40, 0.45]	0.518 [0.48, 0.55]

Table 41: Support and rarity simulation, every model in every configuration. Generated from `synthetic_support.json`.

Configuration and model	Forget negELBO, mean	Attacker, cross	Attacker, within	Audit, cross [95% CI]	Audit, within [95% CI]
1 sd, all: baseline	74.7	0.794	0.905	0.333 [0.25, 0.42]	0.477 [0.35, 0.59]
1 sd, all: retrain	74.0	0.494	0.500	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
1 sd, all: retain fine-tune, 10 epochs	74.7	0.743	0.834	0.313 [0.23, 0.40]	0.454 [0.34, 0.57]
1 sd, all: retain fine-tune, 100 epochs	74.7	0.651	0.719	0.317 [0.24, 0.40]	0.445 [0.32, 0.56]
1 sd, all: gradient ascent + fine-tune	74.7	0.736	0.838	0.309 [0.23, 0.39]	0.450 [0.33, 0.57]
1 sd, all: forget push (continuous)	76.2	0.202	0.747	0.173 [0.11, 0.24]	0.360 [0.25, 0.47]
1 sd, all: targeted, oracle level	74.2	0.567	0.700	0.592 [0.48, 0.70]	0.552 [0.43, 0.67]
1 sd, all: targeted, cross look-alikes	74.7	0.659	0.731	0.521 [0.43, 0.62]	0.502 [0.39, 0.62]
1 sd, all: targeted, held-out same component	74.5	0.705	0.780	0.497 [0.40, 0.59]	0.490 [0.37, 0.61]
1 sd, all: retrain seed 101	74.7	0.499	0.500	0.275 [0.21, 0.35]	0.478 [0.36, 0.60]
1 sd, all: retrain seed 102	74.6	0.482	0.464	0.306 [0.23, 0.39]	0.480 [0.36, 0.59]
1 sd, half: baseline	74.5	0.831	0.908	0.551 [0.43, 0.67]	0.564 [0.43, 0.70]
1 sd, half: retrain	74.5	0.522	0.464	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
1 sd, half: retain fine-tune, 10 epochs	74.5	0.753	0.784	0.546 [0.42, 0.66]	0.533 [0.40, 0.67]
1 sd, half: retain fine-tune, 100 epochs	74.5	0.660	0.687	0.563 [0.44, 0.67]	0.549 [0.41, 0.69]
1 sd, half: gradient ascent + fine-tune	74.5	0.750	0.791	0.531 [0.41, 0.65]	0.518 [0.39, 0.66]
1 sd, half: forget push (continuous)	76.7	0.094	0.657	0.191 [0.10, 0.30]	0.291 [0.17, 0.42]
1 sd, half: targeted, oracle level	74.4	0.637	0.727	0.620 [0.46, 0.77]	0.595 [0.44, 0.75]
1 sd, half: targeted, cross look-alikes	75.2	0.596	0.731	0.576 [0.41, 0.73]	0.562 [0.41, 0.72]
1 sd, half: targeted, held-out same component	75.5	0.560	0.721	0.541 [0.38, 0.70]	0.534 [0.38, 0.69]
1 sd, half: retrain seed 101	74.3	0.646	0.467	0.640 [0.54, 0.73]	0.542 [0.40, 0.67]
1 sd, half: retrain seed 102	74.5	0.455	0.420	0.521 [0.40, 0.64]	0.571 [0.43, 0.71]
2 sd, all: baseline	76.2	0.829	0.896	0.364 [0.28, 0.45]	0.464 [0.35, 0.58]
2 sd, all: retrain	75.7	0.412	0.500	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
2 sd, all: retain fine-tune, 10 epochs	76.1	0.776	0.840	0.345 [0.27, 0.43]	0.455 [0.33, 0.57]
2 sd, all: retain fine-tune, 100 epochs	76.2	0.720	0.735	0.345 [0.27, 0.43]	0.443 [0.32, 0.56]
2 sd, all: gradient ascent + fine-tune	76.2	0.780	0.830	0.341 [0.26, 0.43]	0.451 [0.33, 0.56]
2 sd, all: forget push (continuous)	78.9	0.209	0.716	0.119 [0.06, 0.18]	0.296 [0.19, 0.40]
2 sd, all: targeted, oracle level	75.7	0.777	0.839	0.569 [0.48, 0.67]	0.481 [0.36, 0.60]
2 sd, all: targeted, cross look-alikes	74.5	0.179	0.645	0.666 [0.56, 0.77]	0.593 [0.47, 0.71]
2 sd, all: targeted, held-out same component	75.6	0.708	0.777	0.617 [0.52, 0.72]	0.507 [0.38, 0.62]
2 sd, all: retrain seed 101	76.2	0.500	0.500	0.271 [0.20, 0.35]	0.486 [0.36, 0.61]
2 sd, all: retrain seed 102	76.2	0.509	0.506	0.310 [0.23, 0.39]	0.483 [0.36, 0.59]
2 sd, half: baseline	76.0	0.852	0.896	0.609 [0.49, 0.71]	0.547 [0.41, 0.68]
2 sd, half: retrain	76.0	0.509	0.500	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
2 sd, half: retain fine-tune, 10 epochs	75.9	0.802	0.853	0.613 [0.50, 0.72]	0.546 [0.41, 0.68]
2 sd, half: retain fine-tune, 100 epochs	75.9	0.735	0.757	0.609 [0.49, 0.71]	0.542 [0.40, 0.68]
2 sd, half: gradient ascent + fine-tune	76.0	0.807	0.851	0.596 [0.48, 0.70]	0.524 [0.38, 0.66]
2 sd, half: forget push (continuous)	80.0	0.078	0.667	0.077 [0.02, 0.15]	0.193 [0.10, 0.30]
2 sd, half: targeted, oracle level	76.0	0.651	0.722	0.640 [0.49, 0.78]	0.556 [0.41, 0.71]
2 sd, half: targeted, cross look-alikes	74.2	0.710	0.762	0.752 [0.62, 0.88]	0.653 [0.50, 0.80]
2 sd, half: targeted, held-out same component	75.5	0.802	0.799	0.731 [0.59, 0.86]	0.623 [0.48, 0.77]
2 sd, half: retrain seed 101	75.9	0.563	0.500	0.654 [0.56, 0.74]	0.535 [0.40, 0.66]
2 sd, half: retrain seed 102	76.1	0.451	0.417	0.511 [0.39, 0.63]	0.556 [0.42, 0.69]
4 sd, all: baseline	81.0	0.919	0.895	0.713 [0.63, 0.80]	0.468 [0.35, 0.58]
4 sd, all: retrain	81.9	0.571	0.500	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
4 sd, all: retain fine-tune, 10 epochs	81.9	0.921	0.790	0.476 [0.38, 0.57]	0.473 [0.35, 0.59]
4 sd, all: retain fine-tune, 100 epochs	82.2	0.920	0.775	0.400 [0.31, 0.49]	0.470 [0.35, 0.59]
4 sd, all: gradient ascent + fine-tune	82.0	0.922	0.801	0.469 [0.38, 0.57]	0.470 [0.35, 0.59]
4 sd, all: forget push (continuous)	92.4	0.854	0.681	0.029 [0.01, 0.07]	0.344 [0.24, 0.45]
4 sd, all: targeted, oracle level	81.9	0.090	0.490	0.648 [0.55, 0.75]	0.472 [0.36, 0.59]
4 sd, all: targeted, cross look-alikes	75.2	0.231	0.564	0.914 [0.86, 0.96]	0.564 [0.45, 0.68]
4 sd, all: targeted, held-out same component	78.3	0.813	0.736	0.895 [0.84, 0.94]	0.545 [0.43, 0.66]
4 sd, all: retrain seed 101	82.3	0.549	0.500	0.360 [0.29, 0.44]	0.489 [0.37, 0.60]
4 sd, all: retrain seed 102	82.2	0.500	0.467	0.408 [0.33, 0.50]	0.480 [0.36, 0.59]
4 sd, half: baseline	81.0	0.959	0.897	0.884 [0.79, 0.96]	0.630 [0.50, 0.75]
4 sd, half: retrain	81.7	0.526	0.495	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
4 sd, half: retain fine-tune, 10 epochs	81.4	0.943	0.789	0.772 [0.67, 0.86]	0.613 [0.48, 0.74]
4 sd, half: retain fine-tune, 100 epochs	81.6	0.939	0.792	0.636 [0.53, 0.74]	0.582 [0.45, 0.71]
4 sd, half: gradient ascent + fine-tune	81.4	0.947	0.789	0.761 [0.65, 0.85]	0.601 [0.47, 0.73]
4 sd, half: forget push (continuous)	91.5	0.863	0.672	0.004 [0.001, 0.01]	0.273 [0.16, 0.39]
4 sd, half: targeted, oracle level	81.8	0.035	0.739	0.654 [0.49, 0.80]	0.526 [0.37, 0.68]
4 sd, half: targeted, cross look-alikes	74.8	0.483	0.796	0.934 [0.86, 0.99]	0.667 [0.52, 0.81]
4 sd, half: targeted, held-out same component	79.9	0.941	0.797	0.931 [0.86, 0.98]	0.640 [0.50, 0.78]
4 sd, half: retrain seed 101	81.3	0.713	0.500	0.736 [0.64, 0.83]	0.507 [0.36, 0.63]
4 sd, half: retrain seed 102	81.8	0.465	0.464	0.516 [0.40, 0.63]	0.592 [0.45, 0.72]
8 sd, all: baseline	79.9	1.000	0.703	1.000 [1.00, 1.00]	0.576 [0.46, 0.68]
8 sd, all: retrain	106.4	0.344	0.550	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
8 sd, all: retain fine-tune, 10 epochs	86.9	1.000	0.688	0.999 [1.00, 1.00]	0.550 [0.44, 0.67]
8 sd, all: retain fine-tune, 100 epochs	105.4	1.000	0.568	0.685 [0.59, 0.77]	0.513 [0.40, 0.63]
8 sd, all: gradient ascent + fine-tune	87.6	1.000	0.689	0.999 [1.00, 1.00]	0.549 [0.43, 0.67]
8 sd, all: forget push (continuous)	124.6	0.004	0.607	0.000 [0.00, 0.00]	0.457 [0.34, 0.57]
8 sd, all: targeted, oracle level	106.5	1.000	0.592	0.455 [0.34, 0.57]	0.452 [0.34, 0.57]
8 sd, all: targeted, cross look-alikes	74.7	0.999	0.603	1.000 [1.00, 1.00]	0.642 [0.53, 0.75]
8 sd, all: targeted, held-out same component	313.4	1.000	0.624	0.000 [0.00, 0.00]	0.570 [0.43, 0.70]
8 sd, all: retrain seed 101	105.9	0.430	0.567	0.712 [0.65, 0.77]	0.496 [0.37, 0.61]
8 sd, all: retrain seed 102	105.6	0.605	0.528	0.754 [0.69, 0.81]	0.481 [0.36, 0.59]
8 sd, half: baseline	79.6	1.000	0.694	0.991 [0.98, 1.00]	0.791 [0.68, 0.89]
8 sd, half: retrain	82.4	1.000	0.448	0.500 [0.50, 0.50]	0.500 [0.50, 0.50]
8 sd, half: retain fine-tune, 10 epochs	80.2	1.000	0.680	0.969 [0.92, 1.00]	0.777 [0.67, 0.88]
8 sd, half: retain fine-tune, 100 epochs	81.4	1.000	0.612	0.873 [0.79, 0.94]	0.714 [0.59, 0.83]
8 sd, half: gradient ascent + fine-tune	80.3	1.000	0.676	0.963 [0.91, 1.00]	0.764 [0.65, 0.87]
8 sd, half: forget push (continuous)	116.8	0.000	0.479	0.000 [0.00, 0.00]	0.427 [0.29, 0.56]
8 sd, half: targeted, oracle level	82.3	1.000	0.692	0.513 [0.34, 0.68]	0.638 [0.49, 0.77]
8 sd, half: targeted, cross look-alikes	74.8	1.000	0.790	0.978 [0.94, 1.00]	0.781 [0.66, 0.89]
8 sd, half: targeted, held-out same component	154.9	1.000	0.557	0.000 [0.00, 0.00]	0.503 [0.36, 0.65]
8 sd, half: retrain seed 101	82.0	1.000	0.520	0.582 [0.45, 0.71]	0.510 [0.36, 0.65]
8 sd, half: retrain seed 102	86.1	0.100	0.568	0.230 [0.13, 0.34]	0.492 [0.35, 0.64]