ATR: SHADOW-FREE, RETRAIN-FREE EVALUATION OF MACHINE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine-unlearning metrics such as TOFU Forget Quality and privacy-leakage scores do not measure forgetting in absolute terms. Each compares a signal on the forget set, such as its loss or truth ratio, to a reference (a retrained model, the retain set, or a held-out population) and reads the result against a fixed threshold. We show this reading is unreliable. On TOFU, running one method at five random seeds, with method, epoch, and data held fixed, yields Forget Quality values spanning three orders of magnitude, so a threshold verdict, and any ranking built on it, largely reflects the seed. We trace this instability to the metric's p-value readout. We then propose ATR (Artifact-Trace Reduction), an evaluator that trains no shadow models and no retrained oracle. ATR reuses the unlearning run one already performs: it reduces each example's unlearning trajectory to a compact descriptor and scores a method by a two-sample distance between the forget set's descriptors and an explicitly chosen reference. On CIFAR-10 image classifiers and a 1.5B language model on TOFU, ATR exactly reproduces the method ranking of a per-example shadow-model membership attack, the trusted but expensive gold standard (Kendall $\tau = +1.0$), at roughly 7% of its cost. At 7B, ATR's ranking never flips across seeds, whereas Forget Quality's flips on a third of seed-pairs, and ATR's cross-seed spread is three to nine times tighter. Two limits apply. The score's orientation depends on the regime and must be fixed against a gold standard or calibration in new settings. And the membership-attack validation is at 1.5B, so transfer to 7B is argued rather than measured.

1 INTRODUCTION

Run one unlearning method on the TOFU benchmark and score it with the field's standard metric, Forget Quality: a p-value near 0 when the unlearned model is still clearly distinguishable from one retrained without the data (a failure to forget) and near 1 when the two are indistinguishable (the goal). Change nothing but the random seed, and five seeds return 7.7×10^{-4}, 0.244, 0.322, 0.367, and 0.864. Same method, same epoch, same data, three orders of magnitude apart. The number a practitioner reports, and the threshold they compare it against to call a model "unlearned," is largely a draw of the seed. Machine-unlearning *evaluation* has become the bottleneck. Methods proliferate faster than our ability to say which actually forget, and Forget Quality (Maini et al., 2024) and privacy-leakage scores such as PrivLeak (Shi et al., 2025) are the headline numbers by which methods are compared and ranked. The seed swing above says that reading is unsafe: a metric consumed as a fixed threshold adjudicates unlearning largely by the seed, and the rankings built on such thresholds inherit the swing. We trace where this instability enters in §5.2.

The cause is structural. Every distribution-based unlearning metric is defined *relative to a reference*: Forget Quality against a retrained model's forget-set truth ratios (Maini et al., 2024), MUSE PrivLeak against a retrained anchor (Shi et al., 2025), and the loss- and membership-based scores against a population of shadow models trained with and without the data (Carlini et al., 2022; Hayes et al., 2025). The most trusted of these, per-example shadow-model membership attacks, rank methods faithfully but cost a fresh model fleet per method (64 to 256 shadows); the cheap distribution metrics that stand in for them are unstable across seeds, as the swing above shows, and, as we find, change their verdict with the reference they are read against.

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

What is missing is an evaluator that ranks methods as faithfully as a shadow-model attack, without its cost, and that treats its reference as an explicit input rather than a fixed assumption. Our answer to the unstable-threshold problem above is not a more stable threshold but a stable *ranking* (comparing methods reliably rather than reading any one number against a fixed cutoff), with an interpretable fraction-forgotten readout available where an oracle can be afforded (§3).

We propose **ATR** (Artifact-Trace Reduction), a shadow-free, retrain-free evaluator that meets this bar: it records lightweight per-example trajectories during an unlearning run one performs anyway, reduces them to fixed-length descriptors, and scores a method by a two-sample distance between the forget set's descriptors and an explicitly chosen reference. Our contributions are:

- **ATR**, a shadow-free and retrain-free unlearning evaluator whose ranking score needs neither shadow models nor a retrained oracle to compute—though its regime-dependent orientation is only conditionally retrain-free, set per regime against a gold where one exists (§3, §6);
- **evidence that ATR is faithful and cheap**: it reproduces the method ranking of a per-example shadow-model membership attack on image classifiers and on a 1.5B language model, at roughly 7% of a shadow-model evaluation's cost, and holds its ranking fixed across seeds where Forget Quality's flips on a third of them, a cross-seed spread three to nine times tighter (§5);
- **evidence that the standard metrics are unreliable**: their verdict is not reproducible across re-runs and inverts with the reference they are read against, which we quantify and trace to the metric's read-out and reference choice (§5).

Language-model validation against a membership attack is at 1.5B, and we are explicit about what does not yet transfer to 7B (§6). The rest of the paper defines the tools and the unlearning estimand (§2), the ATR evaluator (§3), the experimental setup (§4), the experiments (§5), and the limitations the results make visible (§6).

2 Background and Related Work

Machine unlearning and its estimand. Given a model trained on a dataset D_{train} and a forget set $D_f \subset D_{\text{train}}$, an unlearning method aims to produce a model that behaves as if it had been trained only on the retain set $D_r = D_{\text{train}} \setminus D_f$. The natural target is therefore the *retrained* model, and a line of work formalizes success as **indistinguishability**: the unlearned model should be distributed like one retrained without D_f (Ginart et al., 2019; Guo et al., 2020; Sekhari et al., 2021). This estimand is intrinsically a comparison against a reference, the retrained model or its behavior, which is why the metrics below are all reference-anchored.

Distribution-based metrics. The metrics in widest use score an unlearned model by how far a forget-set signal sits from that same signal under a reference. **TOFU Forget Quality** (Maini et al., 2024) computes, per forget example, a *truth ratio* comparing the model's probability on the correct answer to perturbed alternatives, and reports the p-value of a Kolmogorov–Smirnov test (Massey, 1951) between the unlearned model's forget-set truth ratios and a *retrained* model's; a value near 1 means the two are indistinguishable (good forgetting), near 0 that they differ clearly. **MUSE PrivLeak** (Shi et al., 2025) scores privacy leakage as the relative gap between a membership statistic on the forget set and on a retrained anchor. Both are reported as measurements of the unlearned model and are then compared across methods and against fixed thresholds to decide which methods forget.

Membership-inference gold standards. The most trusted way to check per-example forgetting is a *membership-inference attack* (Shokri et al., 2017): given a model and an example, decide whether that example was in the model's training data. The strongest such attacks are **LiRA** (Carlini et al., 2022), its unlearning form **U-LiRA** (Hayes et al., 2025), and the more efficient **RMIA** (Zarifzadeh et al., 2024); all train many *shadow* models, some that saw the example and some that did not, and compare the target model's behavior on the example against both groups. The attack's *membership advantage* (Yeom et al., 2018), how far above chance it guesses (summarized per method as a membership score, 0.5 at chance), measures how much of the example the model kept. Such an attack can *rank* two unlearning methods when their membership scores are far enough apart that

their confidence intervals do not overlap. This is the best standard available, and we compare ATR against it wherever we can afford to train the shadow models (§5); but each method needs its own set of shadow models, which is expensive, the reason a cheaper evaluator is worth having.

Trajectory signals and other diagnostics. A separate line of work reads membership signal from a model's *training trajectory* rather than its final weights (Liu et al., 2022; Li et al., 2024a; Pollock et al., 2025). Closest to ATR is Pollock et al. (2025), who score how private a record is, without shadow models, from how much its loss fluctuates across training checkpoints (its inter-quartile range). ATR borrows this trajectory view but changes its purpose: from attacking one record to ranking whole methods against an explicit reference. Aggregating such per-record trajectory scores into a single number is a natural cheap baseline, but it does not match the gold-standard ranking (Appendix D.12), whereas ATR's combination of several trajectory signals, compared against a reference, does. Weight distance is invalid as a forgetting ground truth (Thudi et al., 2022) and per-example unlearning difficulty varies widely (Sepahvand et al., 2025), so we take a membership attack, not weight distance, as the standard. This is the gap ATR fills: an evaluator that treats its reference as an explicit, checkable input, with the experiments quantifying both its faithfulness and the cost of the assumption it makes explicit.

3 THE ATR EVALUATOR

ATR (Artifact-Trace Reduction) scores how much an unlearning method has forgotten by comparing how the forget set behaves *during the unlearning run* to an explicitly chosen reference population. It is **shadow-free** (it trains no shadow models) and its primary score is **retrain-free**, reading signals off an unlearning run one performs anyway rather than training a from-scratch oracle. Unlike a metric that hard-codes its reference, ATR takes the reference as an explicit argument. The next four sections formally introduce it.

Problem and Notation. Let D_{train} be the training set, $D_f \subset D_{\text{train}}$ the forget set, $D_r = D_{\text{train}} \setminus D_f$ the retain set, and D_{ood} a held-out population never trained on. In practice ATR scores against finite samples of these populations; we write S_{ood} for the finite *reference* sample: held-out data by default (the "OOD" suffix names the reference slot—this default reference is held-out *in-distribution* data, not literal OOD, which enters only as a control), or the retain set D_r when that is the axis measured (§5.4). Training gives M_{orig}; unlearning it with method U produces a sequence of checkpoints $M_0 = M_{\text{orig}}, M_1, \ldots, M_T$ over T update steps. The goal of an evaluator is to rank methods by how much they actually removed D_f, agreeing with a per-example membership oracle where one is affordable (§5).

Method. At each checkpoint M_t, four lightweight collectors record one scalar per example z—for the forget set and, identically, the reference sample S_{ood} (§3), both instrumented in the same unlearning run (Algorithm 1)—each probing a different signal class: **per-sample loss** and **logit margin** (the correct-class logit minus the largest competing logit) at the confidence level; the **last-layer gradient norm** of the output head, in closed form, at the optimization level; and a probed **activation norm** at the representation level. Collecting one scalar a across the $T{+}1$ checkpoints yields a per-example trace $\tau_a(z) = (a_0(z), \ldots, a_T(z)) \in \mathbb{R}^{T+1}$. The collectors run inside the unlearning loop and add no backward pass of their own; exact formulas and cost are in Appendix C. Each trace is summarized by a fixed set of six reductions $\rho(\cdot)$, the mean, the mean of the last 10% of steps, the slope and intercept of a least-squares linear fit against the normalized step index, the area under the trajectory, and the inter-quartile range, chosen to capture a trajectory's level, endpoint, direction, and spread without committing to a fixed checkpoint cadence. Concatenating the six reductions over the four collectors gives the per-example descriptor $\phi(z) = \left[\rho(\tau_{\text{loss}}(z)); \rho(\tau_{\text{margin}}(z)); \rho(\tau_{\text{grad}}(z)); \rho(\tau_{\text{act}}(z))\right] \in \mathbb{R}^{24}$.

Descriptors are standardized per dimension against the reference sample S_{ood} (§3): with μ and σ the per-dimension mean and standard deviation of $\{\phi(z) : z \in S_{\text{ood}}\}$, every descriptor ϕ is mapped to $(\phi - \mu)/\sigma$. This places all dimensions on the reference's scale so that no single collector dominates the distance. Let P_f and P_{ood} denote the resulting normalized descriptor distributions of the forget set and the reference. ATR measures the distance between P_f and P_{ood} with a two-sample maximum mean discrepancy (MMD^2; defined in Appendix A), a kernel two-sample distance (Gretton et al., 2012) that is zero exactly when the two samples coincide. We use the biased V-

statistic for this exact-zero property; at fixed forget-set size its $O(1/n)$ bias is shared across methods and does not change the ranking (the unbiased U-statistic preserves it, Appendix D.13), and we avoid the cross-size comparisons where it would matter. The kernel is a sum of three Gaussian RBFs at bandwidths $\{0.1\sigma, \sigma, 10\sigma\}$, with σ the median pairwise distance over the pooled descriptors (the median heuristic); the multi-scale sum stays informative when any single bandwidth is mismatched. The primary score, **ATR-OOD** $= \mathrm{MMD}^2(P_f, P_{\mathrm{ood}})$, sets the reference to the held-out sample, and is reported with a percentile bootstrap (Efron, 1979) 95% interval (200 resamples). Methods are ranked by this single number. The orientation of ATR-OOD depends on the regime, which we make explicit. On image classifiers a from-scratch oracle's forgotten samples come to look like held-out data, so a *smaller* ATR-OOD means more forgetting. On language models the held-out reference sits at the model's generalization loss, so forgetting pushes the forget set's loss *above* it and ATR-OOD reads as forgetting intensity, *larger* is more forgotten (§5). Because more forgetting is not automatically better, we rank by ATR-OOD only among methods that preserve retain-set utility.

Score Readout. ATR-OOD is the primary ranking score. Because the reference is an explicit argument, we report each reference separately rather than averaging; an oracle-referenced diagnostic **ATR-RETRAIN** and an interpretable $[0, 1]$ fraction-forgotten rescaling of ATR-OOD both require a retrain oracle and are used only for validation and calibration, defined with the full component inventory in Appendices C and D. Algorithm 1 states the full ATR-OOD procedure, separating what is recorded *during* the unlearning run from the scoring done *after* it.

Algorithm 1: ATR-OOD

Input: checkpoints $M_0, \ldots, M_T$ of the unlearning run; forget set D_f; reference sample S_{ood}; collectors $a_1, \ldots, a_4$; reducer ρ (the six summaries of §3); bootstrap count B.
Output: ATR-OOD $= \mathrm{MMD}^2(P_f, P_{\mathrm{ood}})$ with a percentile 95% bootstrap interval; methods are ranked by it.

During the unlearning run: instrument each checkpoint (no extra backward pass)
for each population $S \in \{D_f,\ S_{\mathrm{ood}}\}$ and example $z \in S$ **do**
  for $t = 0, 1, \ldots, T$ **do**
    record the four collector scalars $a_k(z; M_t),\ k = 1, \ldots, 4$
  end for
  form traces $\tau_{a_k}(z) = \big(a_k(z; M_0), \ldots, a_k(z; M_T)\big) \in \mathbb{R}^{T+1}$
end for

After the run: reduce, normalize, score (CPU only: no shadow models, no retrain)
for each $z \in D_f \cup S_{\mathrm{ood}}$ **do**
  $\phi(z) \leftarrow [\rho(\tau_{a_1}(z)); \ldots; \rho(\tau_{a_4}(z))] \in \mathbb{R}^{24}$ ▷ the 24-dim descriptor, §3
end for
$\mu, \sigma \leftarrow$ per-dimension mean and std of $\{\phi(z) : z \in S_{\mathrm{ood}}\}$ ▷ reference sets the scale
standardize $\phi(z) \leftarrow (\phi(z) - \mu)/\sigma$ for all z
$P_f \leftarrow \{\phi(z) : z \in D_f\}$, $P_{\mathrm{ood}} \leftarrow \{\phi(z) : z \in S_{\mathrm{ood}}\}$
$\sigma_{\mathrm{med}} \leftarrow$ median pairwise distance over $P_f \cup P_{\mathrm{ood}}$; $k \leftarrow \sum_{s \in \{0.1, 1, 10\}} \mathrm{RBF}(s\,\sigma_{\mathrm{med}})$
$\hat{s} \leftarrow \mathrm{MMD}^2_k(P_f, P_{\mathrm{ood}})$ ▷ MMD^2, App. A
for $b = 1, \ldots, B$ **do**
  resample P_f, P_{ood} with replacement; $s_b \leftarrow \mathrm{MMD}^2_k(\cdot, \cdot)$
end for
return $\hat{s}$ and the $[2.5, 97.5]$ percentiles of $\{s_b\}_{b=1}^B$

4 Setup

We evaluate at two scales, each exercising a different part of ATR (Appendix C, Table C.3 tabulates the full apparatus). **Image classifiers**, ResNet-18 (He et al., 2016) on CIFAR-10 (Krizhevsky, 2009) (a standard ten-class image dataset), are the regime where an affordable per-example gold standard, U-LiRA, lets ATR's ranking be checked directly. **Language models**, Qwen2.5-Instruct (Qwen Team, 2025) on TOFU (Maini et al., 2024) (a question-answering unlearning benchmark built from fictitious author biographies, so the forget target was never seen in pre-training), are

where the community metric lives and where ATR-OOD's orientation across scale appears. We unlearn in the one-to-two-epoch band where forgetting is measurable without destroying retain-set utility.

Unlearning methods. The methods span weak to aggressive forgetting. The **retrain proxy** (gradient difference (Maini et al., 2024), also **GradDiff**; the name records that we use it as a practical stand-in for a from-scratch retrain, not that it is presupposed to be the target) ascends the forget-set loss while descending on the retain set; **fine-tune** continues training on the retain set only, a weak, *leaky* forgetter that leaves the forget set largely intact. **NPO** (Zhang et al., 2024) and **SimNPO** (Fan et al., 2025) are negative-preference-optimization methods, which use a bounded, preference-style forget loss (SimNPO is a reference-free simplification); **WGA** (Wang et al., 2025) and **SatImp** (Dorna et al., 2025) are gradient-difference variants that reweight the forget term. Two aggressive methods are excluded *in advance* under two standing admissibility rules (an unlearned model must keep its retain-set utility, and must stay in the range of behavior the shadow attack is calibrated on): **gradient ascent** (Jang et al., 2023), an unbounded ascent on the forget set, collapses the model into a degenerate state the attack cannot score, and **CEU** (Dorna et al., 2025) destroys retain-set utility (retain-loss $\approx 9\times$). They are dropped for violating these rules, not because ATR fails on them (Appendix C). On image classifiers the methods are a from-scratch **retrain oracle** (the ideal), fine-tune, and the pair **SCRUB** and **NegGrad+** (Kurmanji et al., 2023).

What we report. We rank the methods by each evaluator and measure ranking agreement with Kendall's τ (Kendall, 1938), using the tie-aware τ-b where a ranking has ties. On image classifiers we compare against U-LiRA in three regimes: two *diagnostic* ones (forgetting a random 5% subset, or one whole class) and a **high-memorization** regime, where the original model has overfit the forget set so strongly that no evaluator, U-LiRA included, can separate the methods. At language-model scale the gold standard is a shadow-membership attack we validate at 1.5B (§5.1). Full hyperparameters, regime definitions, collector formulas, hardware, and the full two-scale apparatus (Table C.3) are in Appendix C.

5 Experiments

A valid method-ranker must pass six checks: faithfulness to a per-example membership attack where one can be afforded (§5.1), reproducibility across re-runs (§5.2), non-circularity with the attack it is validated against (§5.3), a reference that is a verdict-changing choice rather than an assumption (§5.4), low enough cost to run on every method (§5.5), and resistance to gaming (§5.6). Every per-run score, interval, and per-cell table behind the summaries below is in Appendix D and the released repository.

5.1 ATR agrees with a per-example gold standard

Setup and motivation. The gold standard for per-example forgetting is a shadow-model membership attack (§2), trusted but expensive; to *validate* ATR we pay its cost once. On image classifiers we run the established version, U-LiRA, over the four image-classifier methods (§4). At language-model scale no validated per-example membership attack exists, so we construct one at 1.5B (a shadow fleet plus a LiRA-style per-example test on Qwen2.5-1.5B/TOFU) and gate it for validity before using it as the gold (Appendix C): it separates members from non-members (**AUC 0.98**) and reads the extremes correctly and non-overlapping (**0.82** for a leaky fine-tune, **0.51**, chance, for a clean retrain proxy). Only a validated attack is used to rank.

Results. ATR-OOD traces a strictly monotone curve against the membership gold at both scales (Figure 1): exact rank agreement, one method per point. On image classifiers ATR reproduces the U-LiRA ordering exactly in both diagnostic regimes ($\tau = +1.0$ each). The per-regime permutation p is 0.083, about 0.004 jointly—treating the two regimes as independent, which they only approximately are, sharing methods and architecture—so a single four-item regime is not significant alone (Appendix C). In the high-memorization regime neither evaluator can rank the methods, three of four sitting inside U-LiRA's own noise floor, so this is *no test*, not a disagreement. At 1.5B the attack resolves four levels of forgetting across the six methods and ATR-OOD reproduces the ordering exactly: of the twelve method-pairs the attack resolves, ATR-OOD

orders every one the same way (τ-b $= +1.0$; Appendix C), including the *hard* retrain-proxy-versus-NPO pair that a naive single-final-checkpoint reduction inverts (it drops to τ-b $= +0.83$; Appendix D, Table D.21). Gradient ascent and CEU are excluded in advance under the admissibility rules (§4). The per-method numbers for both scales (U-LiRA AUC / shadow-membership score beside ATR-OOD, with confidence intervals) are in Appendix C (Table C.5).

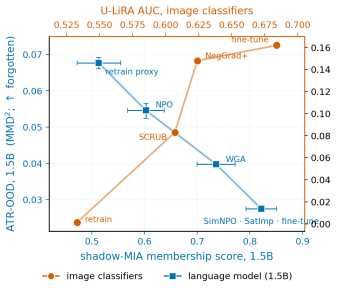


Figure 1: **ATR-OOD tracks the membership gold at both scales.** One point per unlearner: image classifiers (orange, U-LiRA AUC) and the 1.5B language model (blue, shadow-MIA score, 95% intervals). Each curve is strictly monotone in its gold; opposite slopes are the regime-dependent orientation (§5.4).

Discuss. ATR agrees with the attack at the *set* level, not per example: per example the descriptor is only a weak membership signal (Spearman $\rho \approx 0.16$; per-class means 0.11–0.21, Appendix D), but the two-sample distance between descriptor *distributions* is exactly what ranks methods. ATR is a method ranker, not a per-example test. §5.3 shows cheaper reference-free statistics match ATR on all but the retrain-proxy-versus-NPO pair, so ATR's advantage rests on that single hardest contrast, a genuine limit. On image classifiers, agreement over the three practical unlearners (dropping the trivially-placed oracle) holds in both diagnostic regimes but is a three-item test, and is seed-sensitive in one regime ($\tau = +0.33$ for one ca_weak seed; Appendix C). All language-model faithfulness is measured at 1.5B; a direct 7B attack finds the gradient-difference reference off its calibration region, so the gold cannot be established there and 7B transfer is left unestablished (§6, Appendix D).

5.2 ATR'S RANKING IS STABLE ACROSS SEEDS; THE STANDARD METRIC'S IS NOT

Motivation. A metric is only usable if re-running the unlearning at a new seed does not change its verdict. We re-run every configuration at five seeds at 7B, the headline scale (retrained reference held fixed), and measure each metric's *cross-seed spread* (coefficient of variation) and whether the method *ranking* flips. Stability here is cross-seed reproducibility at 7B, distinct from faithfulness (established at 1.5B, §5.1).

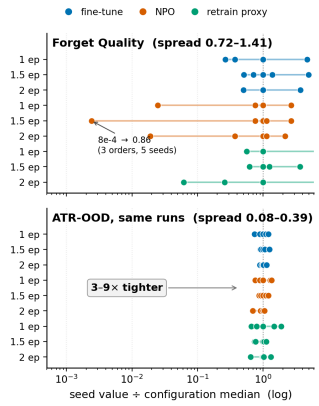


Figure 2: **Cross-seed instability** (7B; nine method×epoch configs, five seeds each). Forget Quality (top) sprawls across three decades; ATR-OOD (bottom) stays 3–9× tighter. Per-config CVs in Table D.2.

Results. Across all nine (method, epoch) configurations Forget Quality's cross-seed spread exceeds ATR-OOD's by 3–9× (Figure 2); NPO at 1.5 epochs returns Forget Quality values three orders of magnitude apart across seeds (the five values in §1), against an ATR-OOD spread of 0.12. The decision-relevant consequence (Table 1): Forget Quality's method ordering flips on 10 of 30 seed-pairs (concentrated at the higher epochs; Appendix D), while ATR-OOD's never flips; averaging Forget Quality over five re-runs cuts the flip rate to ∼1% (a bootstrap rate of the *averaged* ranking, not directly comparable to the single-run 10-of-30; Appendix D) but costs five unlearning runs, where ATR-OOD reaches zero from one. That ATR-OOD never flips despite a per-configuration cross-seed spread of 0.08–0.39 is because its between-method gaps exceed its seed noise (the 1.75× discriminability below) and its per-method seed deviations are more co-moving across methods than Forget Quality's (positive versus negative cross-method correlation, Appendix D.7); the reproducibility is a property of the ranking margin, not of tight absolute values.

The instability is the read-out, not the reference or membership scoring. On the identical 1.5B membership input a smooth read-out is seed-stable (spread ≈ 0.02) where Forget Quality's KS-p-value is not (0.21–0.38), and pooling reference seeds does not help; membership scoring itself is stable (standalone spread ≤ 0.065, at AUC 0.98). So the fragility is the read-out, not the reference or the choice to score membership, and switching to a smooth read-out fixes the instability

but not the cost: a membership statistic still needs a shadow fleet and the KS statistic still needs the retrained model, whereas ATR-OOD ranks faithfully (§5.1) against a held-out set with neither. The seed-stability comes from the smooth distance, not the trajectory: a cheap final-checkpoint descriptor is comparably stable (Table D.21); the trajectory buys faithfulness (§5.1). The full read-out-versus-reference diagnostic and the Min-K% (Shi et al., 2024) self-validity caveat (whether the attack itself clears its validity gate) are in Appendix D (Table D.3).

Table 1: Decision-relevant seed stability (TOFU 7B, three epoch settings—the three methods ranked at each—× five seeds, giving 30 seed-pairs): fraction with identical method ordering and top-1, and mean pairwise Kendall τ. The nine cells of Fig. 2 are the per-(method, epoch) spreads; here the three methods are *ranked* at each of the three epochs.

	Forget Quality	**ATR-OOD**
ordering	67%	100%
top-1	87%	100%
mean τ	0.78	1.00

5.3 Does the agreement just repackage the attack's signal?

Motivation. ATR's descriptor includes per-sample loss and logit margin (the same confidence signal a membership attack is built from), so its agreement with the attack could re-encode that signal rather than reflect unlearning in general. Sharpest at 1.5B, where we *built* the attack, we test two ways, re-scoring the six methods from the on-disk trajectories: ablating the descriptor (does agreement depend on the attack's feature?) and comparing ATR to cheaper quantities (is the reference doing any work?). The ablation also runs on image classifiers against the independent U-LiRA gold; the cheaper-quantity controls need the home-built fleet, so are 1.5B-only.

Results: descriptor ablation. At 1.5B, dropping the per-sample loss channel (the closest analogue to the attack's input) leaves the ranking exact (τ-b $= +0.96$), any three of the four collectors reproduce it, and either confidence channel alone recovers it; only the two geometric channels together fail (τ-b $= +0.50$, inverting the contested pair; Figure 3). On image classifiers the ablation reproduces the U-LiRA ranking at $\tau = +1.0$ for every variant, loss-only included, so a single confidence channel already recovers the gold at both scales. The agreement therefore does not depend on the attack's specific feature: it is convergent, not circular, resting on confidence-type *trajectory* information held redundantly across collectors.

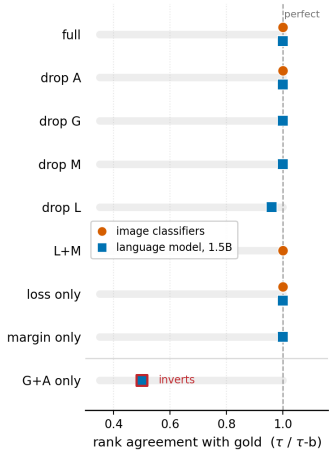


Figure 3: **Descriptor ablation, both scales.** Rank agreement with each scale's gold (Kendall τ vs U-LiRA, orange; τ-b vs shadow-membership at 1.5B, blue) per descriptor variant. Channels L/M/G/A = loss, logit margin, gradient norm, activation norm; *drop X* removes X. Image markers overlap at $\tau = +1.0$ for every variant; only *G+A only* (no confidence channel) breaks the gold and inverts the contested pair. Full table in Appendix D.11.

Results: is the reference doing the work? (1.5B.) Cheaper quantities that skip the reference do not reach the gold ranking (full table in Appendix D, Table D.20). Ranking by how far the run moves the retain set agrees only weakly ($\tau = 0.33$); ranking by *forget*-set movement does better ($\tau = 0.69$–0.83) but inverts the contested pair (NPO moves the forget set farthest, yet the gold ranks the retrain proxy as more forgotten), so *distance moved is not forgetting*. Two shadow-free artifact statistics (a single-model mean truth-ratio; a Pollock et al. (2025)-style loss-trajectory-IQR aggregate) invert it too ($\tau = 0.55$); only ATR-OOD reaches the gold ranking ($\tau = 0.97$, the plain-τ form of the τ-b $= +1.0$ of §5.1; they differ only on the tied cluster) with neither a shadow fleet nor a retrained model. The held-out reference measures distance to *where forgotten data should sit* rather than distance travelled, which is what recovers the ranking.

Discuss. The agreement survives dropping the attack's own feature and is not reproduced by cheaper reference-free quantities, so ATR is not a repackaging of the membership signal. It is also robust to the scorer's estimator and kernel: the unbiased MMD² estimator preserves the ranking (τ-b $= +1.0$), a permutation test rejects $P_f = P_{\text{ood}}$ for every method ($p \leq 0.005$), and the

full descriptor beats a single-checkpoint or single-summary reduction (which fall to τ-b $= +0.83$ and invert the contested pair; Appendix D).

5.4 The reference is a choice that changes the verdict

Motivation. ATR makes its reference an explicit input; we show this is not a free parameter but one that determines which failure is measured, so it must be reported rather than assumed.

Results. Holding the model, artifacts, and scorer fixed and varying only the reference (Appendix D, Table D.6), the method ranking does not shift: it *inverts*. Against a held-out test set the from-scratch oracle scores nearest (0.002, its forgotten samples now look held-out) and a weak fine-tune farthest (0.232); against the retain set the two swap (0.500 vs 0.007). Kendall τ with U-LiRA flips from $+1.0$ on the test-set axis to -1.0 on the retain axis, the maximum swing, on both diagnostic regimes (Figure 4). The same reorientation appears across scale: the three method families evaluable at both scales reorient rather than degrade (cross-scale $\tau = -1.0$, n = 3, qualitative; Appendix D, Table D.5).

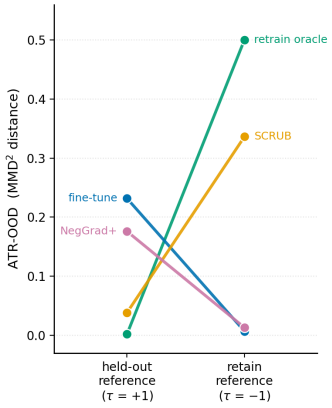


Figure 4: **The reference determines the ranking.** ATR-OOD of the four image-classifier methods against two references, model/artifacts/scorer fixed. Swapping the held-out test set for the retain set inverts the ordering (lines cross), so Kendall τ against U-LiRA flips from $+1.0$ to -1.0. Per-method values in Appendix D (Table D.6).

Discuss. Part of this is mechanical: the test-set and retain axes are opposite by construction (one asks whether the forget set looks *held-out*, the other whether it looks like *retained training data*), so a distance ranking must flip between them; the near-orthogonality is the design point (§5.6), not a surprise. The load-bearing observation is the **cross-scale reorientation of a single axis**: on image classifiers a from-scratch oracle's forgotten samples move *toward* the held-out distribution (smaller ATR-OOD = more forgetting), whereas on the language model forgetting pushes the forget set's loss *away* from the held-out generalization loss (larger = more). Orientation is thus a regime property, not a fixed convention: Appendix D gives the mechanism (memorization leak versus forgetting intensity) and shows a *true* from-scratch oracle reads intense-high too (high ATR-OOD, the forgotten reading at this scale), so the flip is not an artifact of the proxy. We set orientation per regime and confirm it against the gold where one exists (§6 sketches an oracle-free calibration).

5.5 Cost

Motivation. A reference-aware evaluator is useful only if it is cheap enough to run on every method. An accurate per-example membership attack trains a *fleet of shadow models per method* ($\approx$128 per example for U-LiRA (Hayes et al., 2025), 128–256 for LiRA (Carlini et al., 2022), only 1–2 for RMIA (Zarifzadeh et al., 2024)), so its cost scales with methods times fleet size, whereas ATR instruments the runs one performs anyway.

Results. To rank the seven TOFU methods on Qwen2.5-7B, ATR costs about **11 GPU-hours** (measured: the seven instrumented runs one performs anyway plus CPU scoring; fleets counted the same way) against a *projected* **1920** for a 256-shadow U-LiRA, about **0.6%**; even the cheaper LiRA/RMIA fleets are 20–40$\times$ ATR's cost. Those fleet costs are estimates at published shadow counts, since no 7B fleet was trained; the one *directly measured* comparison, at 1.5B against the actual 64-model fleet we ran, is $\sim$**7%** (ATR $\sim$5 vs $\sim$70 GPU-hours; a projected $\sim$1.8% at 256 shadows, Table D.13). The only per-run cost is instrumentation (3.8–8.0% wall-clock and about 2.5 GB peak memory), with post-hoc scoring adding nothing to the training loop. The full cost accounting and a cost bar chart are in Appendix D (Fig. D.2, Table D.13).

Discuss. The gap is structural (O(methods) for ATR against O(methods $\times$ shadows) for a shadow attack), so running the evaluator on every method is affordable at scale, precisely what a benchmark evaluator must be. This headline comparison is against the *reliable* gold (U-LiRA) that ATR re-

places. Against the *cheap* incumbent it stabilizes, Forget Quality, the accounting still favors ATR: Forget Quality amortizes one retrained reference per (dataset, forget split), whereas ATR-OOD trains no retrain oracle at all, trading it for the per-run instrumentation (3.8–8.0%), so its cost is comparable while removing the seed instability.

5.6 Robustness: detecting failures and resisting gaming

Two distinct questions: can the evaluator *flag* a model that fails to forget, and can an adversary *game* it? Flagging divergence from a retrain oracle is not enough (a good unlearner also diverges), so separating an attack from a good unlearner needs a *three-way* comparison (attack vs oracle vs a good practical unlearner, e.g. SCRUB or NPO). Against four stress tests ATR flags all four and catches three *specifically*, by a signal a good unlearner does not also produce (Table 2, beside the gameable JSD aggregate (Lin, 1991)); only output-matched distillation (which mimics the oracle's outputs while keeping the forget set memorized) is flagged but not specifically, at both scales—the sharpest robustness gap. On gaming, trace- and checkpoint-gaming attacks that leave the forget set memorized are caught on the held-out axis in all 18 configurations, and loss-matching attacks cannot be both functional and evasive (evasion ratio 0.69, below the detection bar); detection is weakest only in the high-memorization regime. Per-signal values, the full stress matrix, and per-regime gaming margins are in Appendix D (Tables D.16–D.18). An adversary that games the trajectory *instrumentation* itself is the main open robustness question (§6).

Table 2: Three-way failure detection: whether ATR flags each stress test and does so specifically, beside the gameable JSD baseline (image tests only). ATR flags all four and catches three specifically; only output-matched distillation is non-specific. Decisive values in Appendix D (Table D.16).

Failure mode	ATR signal	flags	spec.	JSD
over-forget	1-Wasserstein	✓	✓	✓
under-forget	retain axis	✓	✓	×
no-op	held-out+retain	✓	✓	×
distillation (img)	ATR-RETRAIN	✓	×	×
distillation (LM)	activation-MMD	✓	×	n/a

6 Discussion and limitations

Limitations. Four bound the results. (i) *Failure detection at language-model scale is open for output-matched attacks*: the memorization-preservation signal that gives specificity on image classifiers does not transfer to language models (§5.6). (ii) *The language-model faithfulness check is validated only at 1.5B*: a direct 7B attack finds the standard gradient-difference reference off its calibration manifold (the retrain proxy reads membership 0.70, not chance), so no valid 7B gold can be built, and transfer is argued, not measured (§5.1; Appendix D, Tables D.9–D.10). ATR's own 7B ranking leaves the contested retrain-proxy/NPO pair unresolved, so a valid 7B per-example gold is the concrete future work. (iii) *The instability result is measured on TOFU Forget Quality*: it should extend to MUSE PrivLeak by construction, but no standard unlearner reaches a useful operating point there, and manufacturing one is itself the operating-point assumption ATR makes explicit, so we scope this claim to TOFU (Appendix D, Table D.11). (iv) *ATR evaluates a run it instruments, not a model it is handed*: it needs the run's checkpoints and its agreement is set-level (per-example Spearman 0.16), so it ranks which method forgets more rather than auditing per-record deletion, and a third party with only the original and unlearned checkpoints can compute only the weaker single-state variant (τ-b $= +0.83$; Appendix D). ATR-OOD is also an uncalibrated distance whose orientation is regime-dependent (§5.4), so ATR is retrain-free for *computing* the score but only conditionally so for *orienting* it off-regime; a faithful $[0, 1]$ readout needs a retrain oracle (Appendix D). Its sensitivity to the checkpointing schedule and to a trajectory-instrumentation adversary (§5.6) remains open. Further scope boundaries—and checks that ATR's own rule holds up (a no-knife-edge admissibility cut-off; stability and faithfulness not resting on any single channel)—are in Appendix D.

Conclusion. ATR ranks unlearning methods as faithfully as a per-example shadow-model membership attack (on image classifiers and at 1.5B), shadow- and retrain-free and at a fraction of the cost, by making its reference an explicit, reportable input. Open problems: a language-model reference separating output-matched attacks from genuine unlearners, and a 7B per-example gold standard.

Reproducibility statement

Every reported number maps to a committed result document and a regeneration script; the framework, the per-seed raw scores, and the trajectory instrumentation are released as an anonymized repository at https://anonymous.4open.science/r/atr-supplementary, so the seed-instability, rank-stability, faithfulness, and cost results are directly re-derivable. The experimental setup (models, splits, training recipe, hardware, seed protocol) and the shadow-membership-attack validation protocol behind the language-model-scale faithfulness check (§5.1) are detailed in the appendix.

Ethics statement

This paper studies whether unlearning-evaluation metrics can be trusted; the concrete risk it addresses is overstated forgetting claims where unlearning is relied on for privacy or legal compliance. Our robustness-to-gaming analysis (and its appendix) describes attacks that game an evaluator, but only to test our own evaluator's specificity, and we demonstrate them solely against our own metric rather than to enable evasion of any deployed system. The work uses public benchmarks (CIFAR-10, TOFU), involves no human subjects or private data, and releases anonymized code for review.

AI use statement

We disclose two uses of AI in this work. **Writing:** we used it to aid and polish the manuscript's prose. **Research and experiments:** we used it to help write and execute the experiment and analysis code, via agentic coding tools such as Claude Code. All such use was under author direction and reviewed by the authors: every reported result was produced by committed, re-runnable code and independently checked against source artifacts, and the scientific claims and their framing are the authors' own. The authors take full responsibility for the final content of this work.

References

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, 2022. URL https://arxiv.org/abs/2112.03570.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C. Lipton, J. Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2025. URL https://arxiv.org/abs/2506.12618.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. URL https://doi.org/10.1214/aos/1176344552.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2410.07163.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cb79f8fa58b91d3af6c9c991f63962d3-Abstract.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

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)*, 2020. URL https://proceedings.mlr.press/v119/guo20c.html.

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)*, 2025. URL https://arxiv.org/abs/2403.01218.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://arxiv.org/abs/1512.03385.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2406.13356.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. URL https://aclanthology.org/2023.acl-long.805/.

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938. URL https://doi.org/10.1093/biomet/30.1-2.81.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/062d711fb777322e2152435459e6e9d9-Abstract-Conference.html.

Hao Li, Zheng Li, Siyuan Wu, Chengrui Hu, Yutong Ye, Min Zhang, Dengguo Feng, and Yang Zhang. SeqMIA: Sequential-metric based membership inference attack. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024a. URL https://dl.acm.org/doi/10.1145/3658644.3690335.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning (ICML)*, 2024b. URL https://proceedings.mlr.press/v235/li24bc.html.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. URL https://doi.org/10.1109/18.61115.

Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership inference attacks by exploiting loss trajectory. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2022. URL https://arxiv.org/abs/2208.14933.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1610.06545.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. *arXiv preprint arXiv:2401.06121*, 2024. URL https://arxiv.org/abs/2401.06121.

Frank J. Massey. The Kolmogorov–Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951. URL https://doi.org/10.1080/01621459.1951.10500769.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989. URL https://doi.org/10.1016/S0079-7421(08)60536-8.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. URL https://doi.org/10.1561/2200000073.

Joseph Pollock, Igor Shilov, Euodia Dodd, and Yves-Alexandre de Montjoye. Free record-level privacy risk evaluation through artifact-based methods. In *USENIX Security Symposium*, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/pollock.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

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)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9627c45df543c816a3ddf2d8ea686a99-Abstract.html.

Nazanin Mohammadi Sepahvand, Anvith Thudi, Berivan Isik, Ashmita Bhattacharyya, Nicolas Papernot, Eleni Triantafillou, Daniel M. Roy, and Gintare Karolina Dziugaite. Leveraging per-instance privacy for machine unlearning. In *International Conference on Machine Learning (ICML)*, 2025. URL https://proceedings.mlr.press/v267/sepahvand25a.html.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.16789.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4556f5398bd2c61bd7500e306b4e560a-Abstract-Conference.html.

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)*, 2017. URL https://arxiv.org/abs/1610.05820.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. URL https://www.jstor.org/stable/1412159.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. URL https://doi.org/10.1016/j.jspi.2013.03.018.

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *USENIX Security Symposium*, 2022. URL https://www.usenix.org/conference/usenixsecurity22/presentation/thudi.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

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

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)*, 2018. URL https://arxiv.org/abs/1709.01604.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks. In *International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/zarifzadeh24a.html.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2404.05868.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.16454.

Appendix contents

A Background: tools, metrics, and methods **13**
A.1 Two-sample distances and divergences 14
A.2 Membership-inference attacks 15
A.3 Unlearning evaluation metrics 15
A.4 Unlearning methods 16
A.5 Benchmarks, datasets, and models 17
A.6 Rank and statistical tools 17
B Regime-specific assumption failures in evaluation criteria **18**
C Experimental setup **19**
C.1 Models, data, and methods 19
C.2 ATR configuration 20
C.3 Hardware 21
C.4 Membership-inference gold standard: construction and validation 23
C.5 Pipeline sanity checks 24
D Supporting analyses **25**
D.1 Sources of the metric's instability 25
D.2 Descriptor choice 26
D.3 Cross-scale reorientation 27
D.4 Forget-split footprint 28
D.5 Language-model faithfulness at 7B (argued, not measured) 29
D.6 MUSE PrivLeak operating-point attempt 29
D.7 Controls: drift, forget-side movement, cadence, seed-averaging 30
D.8 Cost detail 32
D.9 Further scope boundaries 33
D.10 Robustness detail: detection and gaming 33
D.11 Descriptor components against the membership gold 35
D.12 Read-out and artifact baselines 35
D.13 Scorer and descriptor-form robustness 36
D.14 Calibrated ATR: an interpretable [0,1] readout from a retrain oracle 38

A Background: tools, metrics, and methods

This appendix defines, formally and self-containedly, every external distance, membership attack, evaluation metric, unlearning method, benchmark, and statistic the paper uses, for readers outside the unlearning-evaluation subfield. Each mathematical object is given with its defining expression and its originating reference; nothing here is a contribution. Throughout, P_f denotes the distribution of a signal on the forget set, P_{ood} the same signal on the chosen reference set (held-out or retain), and $\pi_\theta(y \mid x)$ the model's conditional likelihood of a target string y given a prompt x.

A.1 TWO-SAMPLE DISTANCES AND DIVERGENCES

These quantify how far two samples are drawn from the same distribution; ATR scores a method by such a distance between the forget-set descriptors and a reference (§3).

Maximum mean discrepancy (MMD). The MMD is the distance between the *kernel mean embeddings* of two distributions in a reproducing-kernel Hilbert space $\mathcal{H}$ with kernel k: for $\mu_P = \mathbb{E}_{x\sim P}\, k(x,\cdot)$, the squared population MMD is $\mathrm{MMD}^2(P,Q) = \|\mu_P - \mu_Q\|^2_{\mathcal{H}}$, which is zero if and only if $P = Q$ for a characteristic kernel (Gretton et al., 2012). Given finite samples $X = \{x_i\}_{i=1}^n$ and $Y = \{y_j\}_{j=1}^m$, the biased V-statistic estimate is

$$\widehat{\mathrm{MMD}}^2_{\mathrm{V}}(X,Y) = \frac{1}{n^2}\sum_{i,i'} k(x_i,x_{i'}) + \frac{1}{m^2}\sum_{j,j'} k(y_j,y_{j'}) - \frac{2}{nm}\sum_{i,j} k(x_i,y_j),$$

which is exactly zero when the two samples coincide. ATR is the default scorer: it uses this biased V-statistic with k a sum of three Gaussian radial-basis-function (RBF) kernels $k(a,b) = \exp(-\|a-b\|^2/2h^2)$ at bandwidths $h \in \{0.1\sigma, \sigma, 10\sigma\}$, where σ is the median pairwise distance over the pooled sample (the *median heuristic* of Gretton et al. (2012)). Summing kernels across bandwidths makes the score sensitive to shifts at several scales at once.

Energy distance. The energy distance is a two-sample distance built from pairwise Euclidean distances (Székely & Rizzo, 2013):

$$\mathcal{E}^2(P,Q) = 2\,\mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\| - \mathbb{E}\|Y-Y'\|,$$

with $X, X' \sim P$ and $Y, Y' \sim Q$ independent. It equals an MMD with the distance-induced kernel $k(a,b) = \frac{1}{2}(\|a\| + \|b\| - \|a-b\|)$, and is available as an alternative ATR scorer (Table C.4).

1-Wasserstein distance. The 1-Wasserstein (optimal-transport, "earth-mover") distance is the minimum expected transport cost between two distributions (Peyré & Cuturi, 2019):

$$W_1(P,Q) = \inf_{\gamma\in\Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}\|x-y\|,$$

the infimum over couplings γ with marginals P and Q. In one dimension it reduces to the L_1 distance between cumulative distribution functions, $W_1(P,Q) = \int |F_P(t) - F_Q(t)|\,\mathrm{d}t$. Unlike an RBF-MMD, whose kernel saturates once samples are many bandwidths apart, W_1 keeps growing with the size of a shift, so it registers extreme over-forgetting where the RBF-MMD plateaus (§5.6, Table B.1 case 7).

Jensen–Shannon divergence (JSD). A symmetric, bounded smoothing of the Kullback–Leibler divergence KL (Lin, 1991):

$$\mathrm{JSD}(P\,\|\,Q) = \tfrac{1}{2}\,\mathrm{KL}(P\,\|\,M) + \tfrac{1}{2}\,\mathrm{KL}(Q\,\|\,M), \qquad M = \tfrac{1}{2}(P+Q),$$

bounded in $[0, \log 2]$ and finite even when P and Q have disjoint support. The paper uses it only as a gameable aggregate baseline detector, to show that a single distributional summary is fooled where the axis-resolved ATR reading is not (§5.6).

Kolmogorov–Smirnov (KS) test. The two-sample KS test compares empirical cumulative distribution functions F_n and G_m through their largest vertical gap (Massey, 1951):

$$D_{n,m} = \sup_t |F_n(t) - G_m(t)|,$$

and converts $D_{n,m}$ to a p-value under the null $P = Q$ via the Kolmogorov distribution. This p-value is the read-out of TOFU Forget Quality (below); a p-value is near-uniform on $[0,1]$ under a true null, which is one reason a threshold on it is seed-unstable (§5.2).

A.2 MEMBERSHIP-INFERENCE ATTACKS

A membership-inference attack (MIA) asks whether a specific example was in a model's training set; for unlearning it measures how much of the forget set an unlearned model still reveals, and a per-example MIA is the trusted gold standard the paper validates ATR against (§5.1).

Shadow-model membership attack. The shadow-model recipe (Shokri et al., 2017) trains many *shadow* models, some on a target example ("in") and some not ("out"), fits the in- and out-distributions of a confidence signal per example, and tests membership by comparing the target model's signal to those two distributions. Aggregated over a set, it estimates how much a model reveals about specific training points, and is the basis of the per-example gold standard.

LiRA. The Likelihood-Ratio Attack casts membership as a per-example hypothesis test (Carlini et al., 2022). It fits Gaussians to a logit-scaled confidence c (distinct from ATR's descriptor ϕ of §3) under the in- and out-shadows and scores the target by the likelihood ratio

$$\Lambda(x) = \frac{\mathcal{N}\big(c(f_\theta(x)) \,\big|\, \mu_{\text{in}}(x), \sigma^2_{\text{in}}(x)\big)}{\mathcal{N}\big(c(f_\theta(x)) \,\big|\, \mu_{\text{out}}(x), \sigma^2_{\text{out}}(x)\big)},$$

which is accurate but fleet-hungry (256 shadow models at CIFAR-10 scale, 64 at ImageNet).

U-LiRA. The unlearning adaptation of LiRA (Hayes et al., 2025): the "out" shadows are retrained *without* the forget example (the retrain-without-forget oracle), so Λ tests whether the unlearned model still looks trained-on-forget. It is the established per-example gold standard for whether an unlearned model reveals the forget set, validated on image classifiers at ≈128 shadow models per example.

RMIA. The Relative Membership Inference Attack calibrates a target example's likelihood against a population of reference examples and needs only 1–2 reference models, trading a modest accuracy loss for far fewer shadows (Zarifzadeh et al., 2024).

Min-K%. A reference-free membership signal from the mean log-probability of a text's least-likely tokens (Shi et al., 2024): for the set S of the $K\%$ tokens with lowest per-token log-probability,

$$\text{MinK}(x) = \frac{1}{|S|} \sum_{t \in S} \log \pi_\theta(x_t \mid x_{<t}).$$

Population Min-K% underlies MUSE PrivLeak. It fails our self-validity gate at 1.5B (forget-set membership AUC 0.44 ≈ chance against the truth-ratio attack's 0.98, Appendix C), a concrete case of a home-built attack too weak to serve as a reference.

Membership AUC and advantage. A membership attack's discriminative power is summarized by the area under its ROC curve (AUC), 0.5 at chance. The *membership advantage* (Yeom et al., 2018), the gap between true- and false-positive rates,

$$\text{Adv} = \Pr[\,\text{predict member} \mid \text{member}\,] - \Pr[\,\text{predict member} \mid \text{non-member}\,],$$

is an equivalent zero-at-chance form (advantage $= 2\,\text{AUC} - 1$ for the AUC-based definition). We report the per-method gold as a **membership score** (0.5 = chance): a method's *residual membership* is the mean per-example probability the attack assigns to membership, evaluated on its unlearned models over the forget set and aggregated with a bootstrap interval, and lower means more forgotten. On image classifiers this score is the U-LiRA ROC-AUC; at language-model scale it is the mean per-example LiRA membership probability.

A.3 UNLEARNING EVALUATION METRICS

These are the final-checkpoint metrics ATR is compared against; each reads a signal on the forget set against a reference through a fixed rule.

Truth ratio. The TOFU per-example confidence signal (Maini et al., 2024): the model's length-normalized probability on perturbed wrong answers $\tilde{a}$ relative to a paraphrased correct answer $\hat{a}$ for question q,

$$R_{\text{truth}}(q) = \frac{\frac{1}{|\mathcal{A}_{\text{pert}}|}\sum_{\tilde{a}\in\mathcal{A}_{\text{pert}}} \pi_\theta(\tilde{a} \mid q)^{1/|\tilde{a}|}}{\pi_\theta(\hat{a} \mid q)^{1/|\hat{a}|}},$$

so a well-forgotten example has the correct answer no more probable than wrong ones. This is the confidence signal both Forget Quality and our 1.5B attack score.

TOFU Forget Quality. The field's headline TOFU metric (Maini et al., 2024): the p-value of a two-sample KS test comparing the truth-ratio distribution of the unlearned model on the forget set to that of a from-scratch retrained model, read through a fixed threshold as a pass/fail forgetting verdict. Because it is a KS p-value against a retrained reference, it inherits both a dependence on that reference and the near-uniformity of a p-value under successful forgetting (§5.2).

MUSE PrivLeak. A privacy-leakage score (Shi et al., 2025) built on a Min-K% membership AUC on the forget set: the relative gap of that AUC between the unlearned model and a retrained anchor,

$$\text{PrivLeak} = \frac{\text{AUC}_{\text{unlearn}} - \text{AUC}_{\text{retrain}}}{\text{AUC}_{\text{retrain}}} \times 100,$$

near zero when leakage matches the oracle, positive when the model under-forgets (over-leaks) and negative when it over-forgets. Its companion utility metrics **retainKnow** and **forgetKnow** report knowledge (answer-recall ROUGE) retained on the retain set and remaining on the forget set.

Pollock loss-IQR aggregate. A shadow-free per-record privacy score (Pollock et al., 2025): the inter-quartile range of an example's loss across training checkpoints, a scalar per record. It is the closest published trajectory-based, shadow-free artifact to ATR, and the paper compares ATR against it as a cheap single-artifact baseline (Table D.20); ATR differs in using a multi-collector, multi-reduction descriptor scored against an explicit reference rather than a single per-record scalar.

A.4 Unlearning methods

The methods being evaluated. Let $\mathcal{D}_f$ be the forget set, $\mathcal{D}_r$ the retain set, ℓ the task loss, and θ_o the original (pre-unlearning) parameters.

Image classifiers. *Retrain oracle*: retrain from scratch on the retain set only, $\theta^\star = \arg\min_\theta \mathbb{E}_{\mathcal{D}_r}\ell(\theta)$; it never saw the forget set and is the exact-unlearning / indistinguishability target (Ginart et al., 2019; Guo et al., 2020; Sekhari et al., 2021). *Fine-tune*: continue training on $\mathcal{D}_r$ only from θ_o; a weak, leaky forgetter that relies on catastrophic forgetting (McCloskey & Cohen, 1989). *SCRUB*: a distillation-based unlearner (Kurmanji et al., 2023) that alternates a *min* step keeping the student close to the teacher θ_o on the retain set and a *max* step pushing it away on the forget set,

$$\min_\theta \; \alpha\,\mathbb{E}_{\mathcal{D}_r}\text{KL}\big(p_{\theta_o}\|p_\theta\big) + \gamma\,\mathbb{E}_{\mathcal{D}_r}\ell(\theta) \;-\; \mathbb{E}_{\mathcal{D}_f}\text{KL}\big(p_{\theta_o}\|p_\theta\big),$$

used as the good-unlearner comparator in the detection tests. *NegGrad+* (Kurmanji et al., 2023): gradient descent on the retain set combined with gradient ascent on the forget set, minimizing $\mathbb{E}_{\mathcal{D}_r}\ell(\theta) - \lambda\,\mathbb{E}_{\mathcal{D}_f}\ell(\theta)$.

Language models. *Gradient difference* (the *retrain proxy*): ascend the forget-set loss while descending the retain-set loss, minimizing $\mathbb{E}_{\mathcal{D}_r}\ell(\theta) - \mathbb{E}_{\mathcal{D}_f}\ell(\theta)$; a practical stand-in for a from-scratch retrain (Maini et al., 2024). *NPO*: negative preference optimization (Zhang et al., 2024), a bounded, reference-anchored forget objective,

$$\mathcal{L}_{\text{NPO}} = -\frac{2}{\beta}\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\log\sigma\left(-\beta\log\frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}\right)\right],$$

which bounds the ascent by a sigmoid σ against a reference model π_{ref} (the original model), avoiding the divergence of raw ascent. *SimNPO*: a reference-free simplification of NPO (Fan et al., 2025) that replaces the reference log-ratio with a length-normalized log-probability minus a margin γ,

$$\mathcal{L}_{\text{SimNPO}} = -\frac{2}{\beta}\,\mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[\log\sigma\left(-\frac{\beta}{|y|}\log\pi_\theta(y \mid x) - \gamma\right)\right].$$

WGA: weighted gradient ascent (Wang et al., 2025), a gradient-difference variant that reweights each forget example's ascent term by a per-example weight derived from a gradient-alignment ("G-effect") analysis. *SatImp*: a saturated-importance gradient-difference variant from the OpenUnlearning framework (Dorna et al., 2025) that likewise reweights the forget term. *Gradient ascent* (Jang et al., 2023): unbounded ascent on the forget set, minimizing $-\mathbb{E}_{\mathcal{D}_f}\ell(\theta)$; excluded from the ranking here because it collapses the model off the calibration manifold (forget truth-ratio driven to 0, retain-loss ratio $63\times$). *CEU*: a cross-entropy unlearning handler in OpenUnlearning (Dorna et al., 2025), excluded here for destroying retain utility (final/initial retain-loss $\approx 9\times$).

A.5 Benchmarks, datasets, and models

CIFAR-10 (Krizhevsky, 2009): a ten-class 32×32 natural-image classification dataset (50k train, 10k test); the image-classifier regime, with a **ResNet-18** (He et al., 2016) backbone, an 18-layer residual convolutional network. **TOFU** (Maini et al., 2024): a question-answering unlearning benchmark of 200 *fictitious* author biographies, constructed so the forget content is absent from pretraining and forgetting is measurable in isolation. **MUSE** (Shi et al., 2025): a six-way unlearning benchmark (PrivLeak's home); its News corpus is used in the operating-point attempt (Appendix D). **WMDP** (Li et al., 2024b): a hazardous-knowledge unlearning benchmark, named but not run (future work). **Qwen2.5-Instruct** (Qwen Team, 2025): the language-model backbone (7B for the headline, 1.5B for membership validation, 0.5B for the forget-split footprint). **Llama-2-7B** (Touvron et al., 2023): used only for the MUSE-News PrivLeak attempt.

A.6 Rank and statistical tools

Kendall τ (Kendall, 1938): a rank-correlation coefficient over $\binom{n}{2}$ pairs, $\tau = (C - D)/\binom{n}{2}$ for C concordant and D discordant pairs ($+1$ identical order, -1 reversed); the paper's primary measure of agreement between two evaluators' method rankings. **Kendall** τ**-b**: the tie-aware variant $\tau_b = (C - D)/\sqrt{(C + D + T_x)(C + D + T_y)}$, with T_x, T_y the tie counts on each side, used when confidence intervals leave some method pairs unresolved (a CI-tied cluster is scored as tied, not as an arbitrary order). **Spearman** ρ (Spearman, 1904): the Pearson correlation of the ranks, $\rho = 1 - 6\sum_i d_i^2/[n(n^2-1)]$ for rank differences d_i; used for the weak per-example descriptor-to-membership relation ($\rho \approx 0.16$). **Coefficient of variation**: the standard deviation divided by the mean, $\sigma/|\mu|$; our scale-free "cross-seed spread" statistic for metric instability (§5.2). **Bootstrap 95% interval** (Efron, 1979): a non-parametric confidence interval from resampling the data with replacement (200 resamples, 2.5th–97.5th percentile), reported on every ATR score. **Permutation test**: a non-parametric null test that shuffles the forget/reference labels to build the null distribution of a statistic, used here to ask whether P_f and P_{ood} genuinely differ and whether a ranking agreement exceeds chance. **Biased V-statistic versus unbiased U-statistic**: two estimators of MMD^2; the U-statistic excludes the $i = i'$ diagonal terms and is unbiased but can go slightly negative, whereas the V-statistic (above) includes them, is exactly zero on identical samples, and carries an $O(1/n)$ positive bias. ATR uses the V-statistic, with the U-statistic as a robustness check (Appendix D). **ROC AUC**: the area under the receiver-operating-characteristic curve, the probability that a classifier ranks a random positive above a random negative; 0.5 at chance, used to gate attack validity. **Leave-one-out false-positive rate**: the empirical rate at which a held-out reference example exceeds a detection quantile fitted on the remaining reference, the finite-sample calibration of the detection threshold ($\approx$0.10 per axis, §5.6).

Table B.1: Regime-specific criterion failures and their fixes (the former break/fix tables, merged so each case reads on one row). *Break*: the criterion as originally written and the regime fact that defeats its literal form. *Fix*: the intent that survives (which transfers across regimes) and the re-grounded criterion that holds. Cases 3–4 are one criterion failing at two regime boundaries; the "Context" column names each case by its experiment setting (dataset, method, criterion).

#	Context	Break: criterion as written — why it fails off-regime	Fix: surviving intent — re-grounded criterion
1	TOFU correlation bar	agreement ≥ 0.60 vs a hand-set ordering — at LM scale ATR-OOD reads forgetting intensity, so low-is-good inverts; the ground truth (Forget Quality) is itself seed-noisy	rank methods sensibly — seed-averaged agreement of ATR-OOD with official Forget Quality on the useful regime (clears 0.60; worked case 1)
2	TOFU operating-point gate	per-cell interval-low within 0.10 of the seed-averaged ceiling — comparing asymmetric estimands against a ground truth whose cross-seed spread is near 1 makes it unpassable	preserve-ranking check — compare the same estimand on both sides, seed-averaged
3	TOFU oracle validation	retrain forget-accuracy near 10% (class-unlearning intuition) — LM forget output stays fluent, so accuracy is not the signal	oracle did not memorize the forget set — direction check: retrain forget-loss above the original model
4	CIFAR random-subset oracle	retrain forget-accuracy near 10% (class-unlearning intuition) — the random-5% retrain still classifies the forget set (its classes stay in the retain set)	oracle did not memorize the forget set — forget-loss $\geq 2\times$ original, and retain-accuracy $\geq$ original -5 points
5	TOFU oracle validation	retrain forget-loss $\geq 2\times$ (image multiplicative gap) — the LM memorization gap is small (about 0.2 nats)	oracle did not memorize the forget set — direction check: retrain forget-loss above the original model
6	stress-detection verdict	one detection direction per stress test — the no-op fires on two axes at once, so one-direction-per-test cannot classify it	attack deviates from the oracle — two-sided verdict, with direction recorded (high / low / dual)
7	scorer bandwidth	median-heuristic RBF-MMD bandwidth — a bulk-calibrated bandwidth saturates on a heavy-tailed shift (an over-forgetter reads low)	detect a distributional shift — add a 1-Wasserstein reading for tail-shift regimes
8	CIFAR unlearner acceptance	forget-accuracy $\leq 20\%$ (class-unlearning) — the random-subset retrain oracle has forget-accuracy 0.96	verify good unlearning — split-appropriate: forget-loss $\geq 2\times$, retain within 5 points

B REGIME-SPECIFIC ASSUMPTION FAILURES IN EVALUATION CRITERIA

A reference-anchored criterion carries assumptions from the regime it was written in. §5.4 makes the general point; this appendix works it out case by case. The pattern is always the same: the criterion's *literal form* breaks when the regime changes, but its *intent* survives and can be re-grounded.

We give eight cases. Table B.1 states, on one row each, the criterion as originally written together with the regime fact that breaks its literal form (the *break*), beside the intent that survives and the re-grounded criterion that holds (the *fix*). Three of the eight are then worked end to end with their threshold-versus-achieved numbers. The "Context" column names each case by its experiment setting (dataset, method, criterion), which also serves as its provenance.

Cases 3 and 4 are a single criterion failing at two different regime boundaries, the recurrence noted in §5.4. These are cases we met while building and cross-checking criteria across the image-classifier and language-model regimes; they illustrate the transfer failure, and are not a survey of published practice. The three CIFAR training regimes named below (`vanilla weak`, `ca_weak`, `canonical`) and the two forget specifications (random-5%, one-class) are defined in Appendix C.

Worked case 1: the TOFU correlation bar. The criterion asked ATR-OOD to agree with a hand-set "sensible" method ordering at Kendall $\tau \geq 0.60$. Two regime facts break that literal form. First, at language-model scale ATR-OOD reads forgetting *intensity*, so its orientation inverts (low-is-good becomes high-is-forgetting, §5.4); scoring a faithful evaluator against a fixed image-classifier-convention ordering mis-reads it. Second, the natural ground truth to correlate against, official Forget Quality, is itself seed-noisy (cross-seed spread 0.72–1.41 over five training seeds), which caps the achievable correlation no matter how good the evaluator is.

We re-grounded the *estimand*, not the threshold: a same-estimand, seed-averaged comparison against Forget Quality on the useful regime. Agreement is then Kendall $\tau = +0.556$ at three seeds and $+0.667$ at five, clearing the 0.60 bar that was fixed in advance. The per-cell (non-averaged) agreement is $\tau = +0.325$, bootstrap interval $[+0.134, +0.520]$; the gap between the per-cell and

seed-averaged values is exactly the reference noise the metric carries, not a deficiency of ATR (§5.3). The 0.60 number itself never moved.

Worked case 2: over-forgetting and kernel saturation (cases 6–7). A verdict rule that expects one detection direction per stress test, and a scorer fixed at the median-heuristic RBF-MMD bandwidth, both break on a heavy-tailed shift. The RBF kernel saturates once the forget-loss distribution sits many bandwidths from the reference. So a Streisand-style over-forgetter, which drives the loss deep into the tail, can read *lower* on the saturating kernel than a good unlearner, and can fire on more than one axis at once.

The intents transfer: detect a distributional shift, and record its direction. The literal formulations do not. Re-grounded, the verdict becomes two-sided with the direction recorded (high, low, or dual), and for tail-shift regimes the bandwidth is supplemented by a non-saturating 1-Wasserstein reading. That reading recovers the over-forgetter at $W_1 = 4.2$, against a good unlearner's 0.5 and a seed floor of 0.04, about a $10\times$ margin over the good unlearner and $96\times$ over the floor (§5.6, Table D.16).

Worked case 3: the oracle-memorization gate across forget specs (cases 3–4–8). A gate written from class-unlearning intuition expects the from-scratch retrain oracle to read near 10% forget-accuracy (a removed class becomes unclassifiable) and expects a good unlearner to sit at most 20%. This one criterion fails at two different regime boundaries. On a random-5% CIFAR forget set the retrain oracle still classifies the forget examples correctly (forget-accuracy 0.96), because their classes stay in the retain set. On TOFU the unlearned output stays fluent, so accuracy is not the signal at all.

Both intents transfer: the oracle did not memorize the forget set, and a good unlearner genuinely forgot. Re-grounded to a direction check (retrain forget-loss above the original model) plus a split-appropriate acceptance (forget-loss at least $2\times$ the original, and retain-accuracy within 5 points), the gate holds on both forget specifications and at both scales.

A note on degrees of freedom. Eight re-groundings (several of which decide whether a criterion passes) give these fixed-in-advance criteria less protection than a single unchanged rule would: each re-grounding is a researcher choice. Three things bound that freedom here. Every case is disclosed in full (Table B.1). The numeric *threshold* is never moved; only the estimand is re-grounded to a same-estimand comparison (worked case 1: the 0.60 bar stands, and only the per-cell-versus-seed-averaged estimand changes). And each re-grounding is justified by a regime mechanism identifiable in advance (a memorization gap, a saturating kernel, a class-versus-subset forget spec) rather than fit to the outcome. The protection is transparency and a mechanism-first rule, not a pre-committed number, and we flag it as such.

C Experimental setup

C.1 Models, data, and methods

Image classifiers. ResNet-18 (He et al., 2016) on CIFAR-10 (Krizhevsky, 2009), unlearned under two forget specifications (a random 5% of the training examples, or a single held-out class) across three training regimes that sit at different points on a memorization spectrum, indexed by their held-out test accuracy (Table C.1). The `canonical` regime is the standard fully-augmented 200-epoch model (test accuracy 0.9531 ± 0.0010 over three seeds); the two `weak` regimes are shorter and less-augmented, so they generalize less and memorize the training set more, which is exactly the regime where a membership attack has the most to detect. The three regimes differ in architecture adaptation, augmentation, and training length; the faithfulness diagnostic (Table C.5) reports the two regimes in which the U-LiRA gold resolves the methods and one high-memorization regime in which three of four methods fall inside its noise floor.

CIFAR-10 has 50,000 training and 10,000 test images. The random-5% forget specification removes 2,500 examples (retain 47,500); the random-1% and random-10% variants remove 500 and 5,000; the class specification removes one full class (≈5,000 examples). Image-classifier statistics use three seeds ($\mathrm{gs} \in \{0, 1, 2\}$) for the canonical regime; the diagnostic regimes are single-seed by design (each is one point on the memorization spectrum, not a cross-seed estimate).

Table C.1: The three CIFAR-10 / ResNet-18 training regimes, by held-out test accuracy (their position on the memorization spectrum). Canonical is the standard 200-epoch augmented model; the weak regimes memorize more and are where a membership attack is most informative.

Regime	Held-out test accuracy	Character
vanilla weak	0.789	lightly adapted, short schedule; most memorization headroom
ca_weak	0.880	architecture-adapted, intermediate schedule
canonical	0.953	full augmented 200-epoch ResNet-18 (M_{orig}: 0.9531 ± 0.0010, 3 seeds)

Table C.2: Per-method unlearning hyperparameters at both scales. Unless noted, language-model methods share the recipe (2 epochs, lr 1e-5, per-device batch 4, gradient accumulation 1); the epoch grid $\{1.0, 1.5, 2.0\}$ is swept for the stability and correlation analyses.

Method	Scale	Key hyperparameters
retrain oracle	CIFAR	from scratch on the retain set, 200 epochs
fine-tune	CIFAR	retain-only continuation, 10 epochs
NegGrad+	CIFAR	retain descent + forget ascent, lr 1e-2
SCRUB	CIFAR	distillation min-max (OpenUnlearning defaults)
gradient ascent (baseline)	CIFAR	500 steps, lr 1e-4
retrain proxy (gradient difference)	TOFU	2 epochs, lr 1e-5, batch 4, grad-accum 1
fine-tune	TOFU	retain-only, matched recipe
NPO	TOFU	2 epochs, lr 1e-5, β default
SimNPO	TOFU	2 epochs, batch 4, $\beta = 4.5$, $\gamma = 0.125$
WGA	TOFU	2 epochs, lr 1e-5
SatImp	TOFU	2 epochs, lr 1e-5
gradient ascent	TOFU	2 epochs, lr 1e-5 (excluded: off-manifold, retain-loss $63\times$)
CEU	TOFU	2 epochs, lr 1e-5 (excluded: retain collapse, retain-loss $9.21\times$)

Language models. Qwen2.5-Instruct (Qwen Team, 2025) on TOFU (Maini et al., 2024): the 7B backbone for the headline results, 1.5B for the shadow-membership validation (§5.1), and 0.5B for the forget-split footprint (Appendix D). The forget split is `forget10` (400 question–answer pairs) unless noted, with `forget05` (200) and `forget01` (40) as the smaller splits, `retain90` as the retain set, and `holdout10` as the held-out reference. TOFU fine-tuning follows the standard recipe (5 epochs, learning rate 1e-5, 1-epoch warm-up, effective batch 32) and unlearning runs over the epoch grid $\{1.0, 1.5, 2.0\}$, all executed in the OpenUnlearning framework (Dorna et al., 2025). Cross-seed statistics use five training seeds for the language-model stability grid; the language-model correlation spans the nine (method, epoch) configurations; the membership fleet uses 32 shadow models per side (16 for WGA and SatImp).

C.2 ATR CONFIGURATION

Tier-1 collectors (per-sample loss, logit margin, last-layer gradient norm, activation norm) feed a composite reducer that maps each per-example trajectory to a fixed-length descriptor by concatenating six reductions (mean, last-k mean, trend slope, trend intercept, area-under-curve, and interquartile range), giving a 24-dimensional descriptor over the four collectors.

Collector formulas (the exact forms §3 refers here for). Write z for a sample's logits, y its target class, h the input to the classifier head, and $p = \mathrm{softmax}(z)$; every collector reuses the checkpoint's already-computed logits (two also read a forward hook) and adds no backward pass. *Per-sample loss* is the cross-entropy $-\log p_y$. *Logit margin* is $z_y - \max_{k \neq y} z_k$. *Last-layer gradient norm* is the closed-form "ghost" norm of the head parameters (`fc` for ResNet-18, `lm_head` for the language

Table C.3: Experimental apparatus at the two scales (the summary moved here from the main text).

Setting	Image classifiers	Language models
Model	ResNet-18	Qwen2.5-Instruct, 7B (1.5B for membership validation)
Data	CIFAR-10	TOFU
Forget spec	random 5% or one class	forget-10 against holdout-10
Methods	retrain oracle, fine-tune, SCRUB, NegGrad+	retrain proxy (grad. difference), fine-tune, NPO, SimNPO, WGA, SatImp (gradient ascent, CEU excluded)
Seeds	3	5 (stability grid)
Gold standard	U-LiRA (per-example shadow-MIA)	validated shadow-MIA at 1.5B
Hardware	A100 / T4 (T4 for the hardware-covariate and gaming checks)	H200

model): with $dz = p - \text{onehot}(y)$, $\|\nabla_{W,b} L\|_2^2 = \|dz\|^2 \, (\|h\|^2 + 1)$, numerically identical to autograd but computed from one softmax and two vector norms. *Activation norm* is the L_2 norm of a probed representation: `avgpool` (the penultimate feature fed into `fc`) on ResNet-18, and `model.norm` (the final hidden state fed into `lm_head`) on the language model. On the language model every collector is computed *per answer token* (prompt tokens masked with label -100) and reduced to one scalar per sequence by the mean over answer tokens (the TOFU convention); the gradient-norm collector differentiates that same mean-over-answer-tokens cross-entropy, for which the multi-token ghost norm is exact.

Each dimension d is standardized against the reference population before scoring,

$$\tilde{d} = \frac{d - \mu_{\text{ref}}}{\sigma_{\text{ref}}},$$

with $\mu_{\text{ref}}, \sigma_{\text{ref}}$ the mean and standard deviation of that dimension over the chosen reference set, so a distance is read in units of reference variation. The standardized forget descriptors are then compared to the reference by the default two-sample MMD scorer (Appendix A; multi-RBF at bandwidths $\{0.1\sigma, \sigma, 10\sigma\}$ set by the median heuristic on the pooled sample). The ATR-OOD reference is a held-out or retain set; a validation-only variant, ATR-RETRAIN $= \text{MMD}^2(P_f^{\text{unlearned}}, P_f^{\text{retrain}})$, instead uses a from-scratch retrain oracle's forget-set descriptors as the reference, measuring divergence from the oracle directly; true-OOD and synthetic-noise references are controls. This Tier-1 configuration is ATR's default, fixed before the U-LiRA and 1.5B comparisons; we ran no configuration search against either gold standard. Every score carries a non-parametric bootstrap 95% interval (200 resamples, percentile).

The spectral reducer is excluded from the default because it is redundant. Applying all seven candidate reductions (the six above plus an 8-coefficient spectral reduction) to every trajectory and correlating the resulting scalar columns per collector shows the spectral coefficients are near-collinear with the retained reductions: on the per-sample-loss collector, the leading spectral coefficient correlates with area-under-curve at $|r| = 1.000$ and with the mean at 0.998 (which also correlate with each other at 0.998), and the second spectral coefficient tracks the trend slope at 0.984 (Figure C.1). Every pair flagged as redundant ($|r| \geq 0.95$) involves a spectral coefficient duplicating a level-, area-, or trend-type reduction the default already carries, so the spectral reducer adds dimension without signal and is dropped; the six retained reductions capture level, recency, trend, area, and spread.

C.3 Hardware

Language-model runs, including the five-seed nine-configuration stability grid behind Figure 2, are on H200; image-classifier runs on A100 and T4, and the per-run instrumentation overhead is measured on an RTX 4500 Ada. The hardware-covariate reference-noise measurement and the image-classifier gaming stress tests are hardware-matched on Tesla T4, because hardware is itself a source

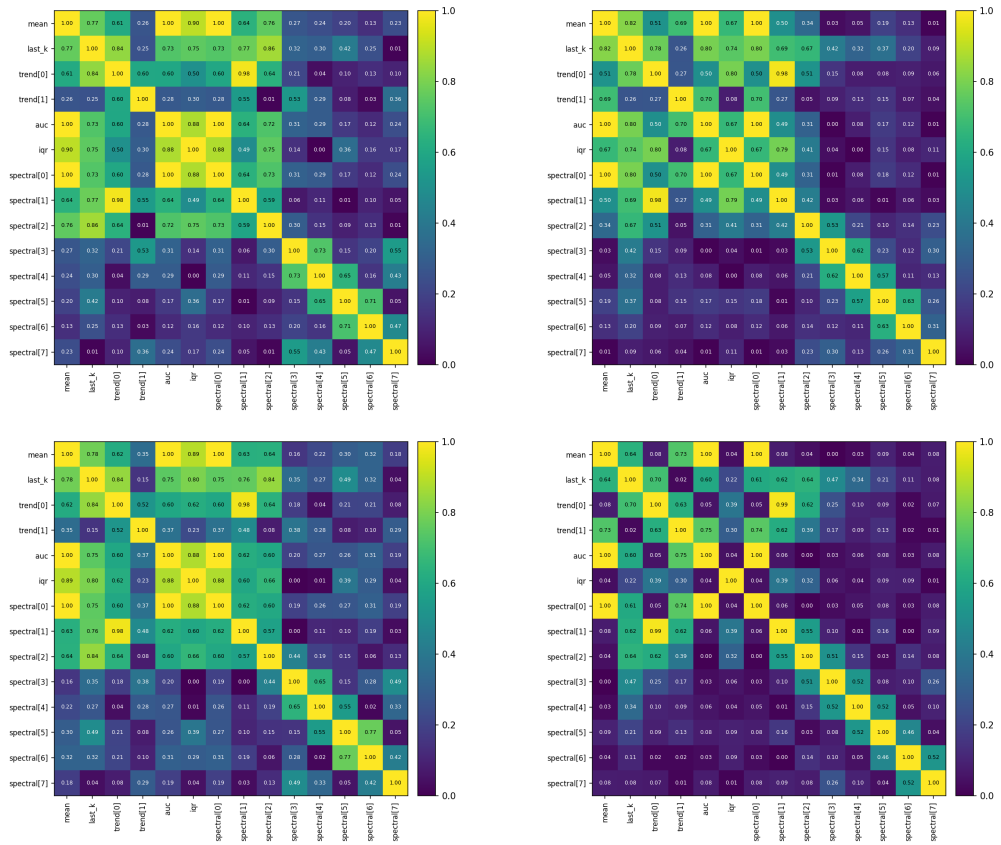


Figure C.1: Pairwise $|\text{Pearson } r|$ among the seven candidate reductions, per Tier-1 collector (CIFAR-10 ResNet-18, 150 traces per collector). Panels, left to right then top to bottom: per-sample loss, logit margin, last-layer gradient norm, activation norm. The spectral coefficients duplicate the level/area/trend reductions the default composite already carries ($|r| \geq 0.95$, brightest cells), which is why the spectral reducer is excluded from the default 24-dimensional descriptor.

Table C.4: ATR's Tier-1 default components and their roles (the inventory complementing Algorithm 1). Collectors span descriptor signal classes; scorers are two-sample distances; references span axis families.

Component	**Members**	**Role / signal class**
Collectors	per-sample loss; logit margin	confidence-level
	last-layer gradient norm	optimization-level
	activation norm	representation-level
Reducers	mean, last-k, trend, AUC, IQR, spectral, composite	trajectory to per-example descriptor
Scorers	MMD (default); energy distance	two-sample set-level distance
References	held-out and retain (the ATR-OOD axes)	ranking score
	retrain-oracle internals (ATR-RETRAIN)	divergence diagnostic
	true-OOD, synthetic-noise	controls
Uncertainty	bootstrap, 200 resamples, 95% interval	interval on every score

of reference noise (Appendix D). The full language-model program used a 4×H200 allocation over roughly two months (≈5,760 H200-hours of capacity); no single run approached the 200-GPU-hour escalation threshold.

The measured wall-clock anchors set the cost accounting of §5.5. At 7B, one full fine-tune (5 epochs, batch 4 × grad-accum 8) takes ≈20.3 min, a retain-only fine-tune ≈13.5 min, per-example forget-

truth-ratio scoring over 400 samples ≈3.4 min, and one complete shadow (train + score) ≈23.7 min. The 16-shadow 7B fleet (16 full + 16 retain + 48 unlearned runs, 80 runs total) came to ≈21 H200-hours, a blended ≈16 min per run. The 32-shadow 1.5B fleet came to ≈56 H200-hours, and the six-method 1.5B faithfulness extension added ≈26 GPU-hours. Per optimizer step at 7B: NPO takes 0.843 s and one trajectory-collection event 3.64 s (peak 49.0 GB); gradient ascent takes 0.327 s per step and 3.76 s per collect (peak 30.5 GB).

C.4 Membership-inference gold standard: construction and validation

Both faithfulness checks (§5.1) compare ATR against a **per-example shadow-model membership attack**, the LiRA / U-LiRA recipe (Appendix A). The recipe is the same at both scales. Train many *shadow* models: some that trained on a given example ("in"), some that did not ("out"). For each example, fit the in- and out-distributions of a confidence signal, and form a per-example likelihood-ratio test for membership. A method's *residual membership* is that test applied to its unlearned models, averaged over the forget set with a bootstrap interval. How far the score reads above chance (0.5 being chance) measures how much of the example the method left behind.

The attack is a valid gold **only if it actually works**: a powerless (chance-level) or miscalibrated attack would make "ATR agrees with the attack" vacuous. So at each scale we check the attack's validity, in advance, before using it to rank anything.

Image classifiers. The attack is **U-LiRA** (Hayes et al., 2025), the established per-example unlearning MIA, run in the vision regime where it is already validated. The confidence signal is the logit-scaled correct-class margin, with 256 in/out shadow models per example (this shadow count enters the cost analysis, §5.5). Validity here rests on the literature, and is corroborated by the retrain sanity check: a from-scratch retrain reads near chance (U-LiRA AUC 0.53, Table C.5).

Dropping the trivially-placed retrain oracle, agreement over the three practical unlearners is still $\tau = +1.0$ in both diagnostic regimes (three items, two-sided permutation $p = 0.33$), and $\tau = +0.33$ in the noise-floor regime the gold cannot resolve. Re-running both diagnostic regimes at three independent training seeds (new targets and forget draws built by the same deterministic gs-parameterized protocol, audited against gs0) reproduces the four-method ranking exactly in vanilla at every seed ($\tau = +1.0$), and at two of three seeds in ca_weak, where one seed reorders the SCRUB/NegGrad+ pair ($\tau = +0.33$). The retrain-oracle placement is stable across all seeds and both regimes. So the image faithfulness is largely, but not perfectly, seed-stable. The joint significance quoted in §5.1 (≈ 0.004) combines the two regimes' per-regime permutation p-values as if independent; because the two diagnostic regimes share the four methods and the ResNet-18 architecture, that is an approximation, and a permutation test over the pooled two-regime design (which we do not run here) would be the stricter statement.

Language models. No validated per-example unlearning MIA exists at language-model scale: LiRA/U-LiRA were developed and validated for vision, and the language-model regime (few optimizer steps, expensive shadows, a different confidence signal) is unproven. So we **construct** the same recipe at 1.5B and validate it before use.

Fleet. 32 seed-varied full fine-tunes (trained on the forget set, "in") and 32 retain-only fine-tunes (the retrain-without-forget oracle, "out"), plus each unlearning method applied to the full shadows. The per-sample signal is the TOFU forget truth-ratio, its in/out distributions fit per example.

Self-validity gate, fixed in advance. The attack must (i) separate members from non-members (intrinsic AUC at least 0.60), (ii) rank a leaky fine-tune above a retrain proxy with non-overlapping intervals, and (iii) place the retrain proxy within 0.05 of chance.

Result. At 1.5B the attack passes. Intrinsic member-vs-non-member AUC is 0.981. The leaky fine-tune reads membership score $0.823\ [0.796, 0.851]$, above the retrain proxy at $0.513\ [0.472, 0.555]$ (a 0.310 gap, non-overlapping intervals), which sits at chance; the contested retrain-proxy/NPO pair is also resolved (gap 0.090, non-overlapping). The attack resolves four membership levels: retrain proxy 0.513 (near chance), NPO 0.603, WGA 0.736, and a fine-tune/SimNPO/SatImp cluster near 0.82 (Table C.6).

That the gate *can* fail is not hypothetical. At 0.5B the strongest forgetters saturate at chance and the attack cannot resolve them (which is why 1.5B is the validation scale). And an independent

Table C.5: Per-method faithfulness, both scales (the numbers behind Figure 1): the per-example membership gold beside ATR-OOD, per method, sorted most-to-least forgotten within each regime. Image classifiers use U-LiRA AUC (↓ = more forgotten); the language model uses shadow-membership score (↓ = more forgotten). ATR-OOD's orientation is regime-dependent: ↓ = more forgotten on image classifiers, ↑ = more forgotten at 1.5B (§5.4). ATR reproduces the ranking wherever the gold resolves the methods: Kendall $\tau = +1.0$ on both image diagnostic regimes, τ-b $= +1.0$ at 1.5B (0 of 12 resolved contrasts discordant); the image high-memorization regime is no-test (three of four methods inside U-LiRA's noise floor). 1.5B intervals and retain-utility ratios are in Table C.6; image diagnostic regimes are single-seed.

Scale / regime	Method	Gold (U-LiRA AUC / shadow-MIA membership score)	ATR-OOD
image, random-subset	retrain (oracle)	0.533	0.0013
	SCRUB	0.607	0.083
	NegGrad+	0.624	0.148
	fine-tune	0.684	0.162
image, class-forget	retrain (oracle)	0.529	0.0015
	SCRUB	0.535	0.071
	NegGrad+	0.654	0.112
	fine-tune	0.687	0.123
image, high-memorization	*no test: 3/4 in U-LiRA's noise floor (AUC 0.49–0.54)*		
language model, 1.5B	retrain proxy	0.513	0.0677
	NPO	0.603	0.0546
	WGA	0.736	0.0398
	SatImp	0.822	0.0276
	SimNPO	0.823	0.0275
	fine-tune	0.823	0.0275

population Min-K% attack fails the gate at 1.5B (forget-set member AUC 0.443 ≈ chance, against the truth-ratio attack's 0.98 (Shi et al., 2024)), so a home-built attack in a new regime genuinely can be too weak.

Ranking. Over the six in-regime methods, ATR-OOD reproduces the attack's ordering (Kendall $\tau = +1$ point estimate, CI-aware τ-b $= +1$; exact permutation $p = 1/720 = 0.0014$ unrestricted, 0.0083 tie-aware over the resolved levels $[1, 1, 1, 3]$; 0 of 12 resolved contrasts discordant). The ranking does not rest on the single 32-shadow attack: cheaper attacks on the same fleet corroborate it. A few-shadow LiRA with 4 in + 4 out shadows (an $8\times$ cut) reproduces the six-method ordering exactly ($\tau = +1.0$; even 2 + 2 keeps the contested pair, $\tau = +0.87$), and an offline RMIA reference-calibrated likelihood ratio with 1–2 out references recovers the gross ordering ($\tau = +0.60$), though not the hard retrain-proxy/NPO pair.

Two aggressive additions are excluded because the attack cannot validly rank them: gradient ascent collapses the model off the calibration manifold (forget truth-ratio driven to 0, retain-loss ratio $63\times$), and CEU destroys retain utility (final/initial retain-loss $9.21\times$). The retained six all preserve retain utility (retain-loss ratio $\leq 1.82\times$, NPO the highest; Table C.6), far below the excluded pair's $9.21\times$ and $63\times$, so the admissibility cut-off is not a knife-edge: any threshold in that wide $(1.82\times, 9.21\times)$ gap admits exactly these six, and NPO—half the contested pair—is retained under all of them. Only tightening the cut below $1.82\times$ would drop NPO; the faithfulness result is reported for the standing rule, not tuned to a threshold. The gold is validated at 1.5B; 7B transfer is not established (§6).

C.5 Pipeline sanity checks

Each measurement tool was checked against a known-answer case before use (the checks are committed scripts, not one-off inspections). Table C.7 collects them. The MMD identity

Table C.6: The 1.5B membership gold with 95% bootstrap intervals and retain-utility ratios (the intervals the Table C.5 language-model rows and Figure 1 report). Membership score ↓ = more forgotten; ATR-OOD ↑ = more forgotten at this scale. The membership score and ATR-OOD both carry intervals; the retain ratio (final/initial retain-loss) confirms every ranked method preserves utility, so the ranking is not confounded by retain collapse. $n = 32$ shadows except WGA/SatImp ($n = 16$).

Method	MIA membership score [95% CI]	ATR-OOD [95% CI]	retain ratio
retrain proxy	0.513 [0.472, 0.555]	0.0677 [0.0662, 0.0692]	1.03×
NPO	0.603 [0.568, 0.638]	0.0546 [0.0524, 0.0565]	1.82×
WGA	0.736 [0.700, 0.772]	0.0398 [0.0393, 0.0402]	1.00×
SatImp	0.822 [0.793, 0.849]	0.0276 [0.0272, 0.0279]	0.99×
SimNPO	0.823 [0.796, 0.849]	0.0275 [0.0273, 0.0278]	0.99×
fine-tune	0.823 [0.796, 0.851]	0.0275 [0.0272, 0.0278]	0.99×

Table C.7: Pipeline sanity checks: each tool against a known-answer case, run before the real experiments.

Check	Expectation	Observed
MMD identity $\mathrm{MMD}^2(X, X)$	exactly 0	0 (V-statistic property)
no-op ATR-OOD (M_{orig}, 1.5B)	near the reference floor	$D = 0.031$ (the calibration zero-anchor)
retrain U-LiRA AUC (CIFAR)	≈ chance	0.53
Forget Quality, oracle versus itself	≈ 1.0	1.0
Forget Quality, still-memorized no-op	≈ 0	3.6×10^{-5}
descriptor recompute versus stored score	exact	max relative drift 0.000
permutation null $P_f = P_{\mathrm{ood}}$	reject for a real shift	rejected, all methods, median p at the 0.005 floor

($\mathrm{MMD}^2(X, X) = 0$) is a property of the V-statistic estimator (Appendix A); the no-op check confirms an untouched model scores near the reference floor; the retrain sanity confirms a from-scratch oracle reads near chance under the membership attack; the Forget-Quality wiring check confirms the metric reads 1.0 when a model is compared to itself and near zero for a still-memorized no-op; the descriptor recompute confirms the analysis re-derives the stored scores exactly; and the permutation null confirms the distances ATR ranks are real distributional differences rather than finite-sample noise.

D Supporting Analyses

D.1 Sources of the metric's instability

The metric's cross-seed instability has four sources, each with a nameable mechanism (Table D.1). The largest is the seed source, the subject of Figure 2 and §5.2, where re-runs vary the unlearning training seed against a fixed reference and the threshold read-out amplifies the spread. Two of the four we quantify with confidence intervals here. The hardware source moves the retain axis by $+11.0\%$ (95% CI $[+5.4\%, +16.7\%]$; paired same-seed A100-vs-T4, $n = 8$, $p = 0.002$), while leaving the held-out axis without a detectable direction ($p = 0.65$). The finite-sample source shows as a leave-one-out false-positive rate near 0.10 per axis, about twice the 0.05 nominal.

Read-out versus reference count. The controlled diagnostic of §5.2 holds the 1.5B truth-ratio input and the reference fixed, and varies only the read-out statistic and the number of reference seeds. A smooth membership statistic has cross-seed spread near 0.02 at any reference size. Forget Quality's threshold-sensitive KS-test p-value, in contrast, stays between 0.2 and 0.4 whether the reference is

Table D.1: Four sources of the metric's instability, with mechanism and remedy. Magnitudes are the metric's cross-seed spread except where noted.

Source	Magnitude	Mechanism	Remedy
Seed	FQ spread 0.72-1.41; ATR-OOD 0.08-0.39	a new training seed reshapes the model's loss	trajectory-averaged descriptor; report the spread
Regime	image-classifier spread 0.21-1.21 across regimes	recipe reshapes the reference distribution	re-ground thresholds per regime (§5.4)
Hardware	retain axis $+11\%$ [+5, +17]; held-out n.s.	hardware shifts some axes	matched-hardware reference; report axes separately
Finite sample	leave-one-out false-positive rate near 0.10 (2x nominal)	high quantiles of a small reference are biased	margin-based verdicts robust to a loose quantile

Table D.2: Per-configuration cross-seed spread (coefficient of variation over five training seeds, retrained reference fixed) for Forget Quality and ATR-OOD on TOFU, the numbers behind Figure 2. ATR-OOD is 3–9× tighter in every configuration (exact per-config ratio 2.99–9.19×).

Configuration	Forget Quality spread	ATR-OOD spread	ratio
fine-tune, 1.0 ep	1.12	0.16	6.8×
fine-tune, 1.5 ep	0.96	0.12	8.0×
fine-tune, 2.0 ep	0.72	0.078	9.2×
retrain proxy, 1.0 ep	1.17	0.39	3.0×
retrain proxy, 1.5 ep	0.79	0.16	4.9×
retrain proxy, 2.0 ep	1.41	0.21	6.7×
NPO, 1.0 ep	0.80	0.22	3.7×
NPO, 1.5 ep	0.78	0.12	6.7×
NPO, 2.0 ep	0.79	0.14	5.7×

a single draw (near 0.21) or pooled over many seeds. So the fragile read-out is the dominant lever, and the reference count is secondary (Table D.3).

Membership scoring itself does not inherit the instability. A standalone membership statistic on the same 32-seed fleet is stable for every method (cross-seed spread: fine-tune 0.009, NPO 0.026, retrain proxy 0.031, all ≤ 0.065), even though it separates members from non-members at AUC 0.98. The unstable component is therefore the threshold p-value read-out. We tried to confirm this on a fully independent metric, population Min-K% on token log-probabilities, but Min-K% is too weak at this scale to be a valid reference: it does not separate a member model from a forgotten one. So the diagnosis rests on the shared truth-ratio input, and is not independently confirmed.

D.2 Descriptor choice

What a reference measures also depends on what the collector computes. Correlated against Forget Quality, a confidence-level target, a loss-only descriptor tracks it more closely than an activation-norm descriptor, with the full descriptor between them (Table D.4, five-seed pooled Kendall τ). The separation is suggestive but not resolved at this sample size: bootstrapped over the nine configurations, the 95% interval on the loss-minus-activation difference is $[-0.03, 0.50]$, and the loss-over-activation ordering holds in 96% of resamples without clearing zero at $n = 9$. The support is mechanistic: loss is Forget Quality's matching confidence-level descriptor, so a weaker activation-to-Forget-Quality correlation is expected, and detectability is a property of the collector set, not of the target alone. This is the descriptor-versus-Forget-Quality view; the complementary descriptor-versus-membership-gold ablation (Table D.19) shows the same confidence channels carry the faithfulness against the 1.5B attack.

Table D.3: Cross-seed spread of two read-outs of the identical 1.5B truth-ratios, crossed with reference-pooling. Forget Quality's KS p-value read-out is unstable; a smooth membership statistic on the same input is not, and pooling more reference seeds does not help.

Read-out × reference	**cross-seed spread**
smooth membership statistic, pooled	0.018
smooth membership statistic, single	0.019
KS p-value (Forget Quality), pooled	0.378
KS p-value (Forget Quality), single	0.207

Table D.4: Per-collector descriptor agreement with Forget Quality (five-seed pooled Kendall τ on TOFU). Loss tracks the confidence-level target most closely; the geometric activation channel least. The logit-margin and gradient channels are not measured against Forget Quality separately here (their agreement with the membership gold is in Table D.19).

Descriptor	**Kendall τ vs Forget Quality**
per-sample loss only	0.42
full four-collector	0.36
activation norm only	0.15

D.3 Cross-scale reorientation

The mechanism: two different failures on one axis. The held-out reference measures a *different* failure at each scale, which is why the same axis points in opposite directions. On image classifiers the forget and held-out (test) sets are the same in-distribution classes, so a from-scratch retrain, which never saw the forget examples, makes their descriptors statistically *indistinguishable* from held-out: the oracle collapses onto the reference (ATR-OOD $\approx$ 0.002, its minimum, Table D.6) and what raises the distance is residual *memorization leak*. On the language model the forget set is memorized question–answer pairs and the held-out reference sits at the model's generalization loss; unlearning does not make the forget set look held-out, it drives the forget descriptors *away* from normal behavior. So at language-model scale ATR-OOD reads *forgetting intensity* (distance from normal): the no-op (an untouched, still-memorized model) sits nearest the reference and stronger forgetters sit farther from it, the reverse of the image-classifier ordering. An equivalent reading, since Algorithm 1 instruments the forget and held-out populations *within the same run*, is that ATR-OOD measures how *differently* the forget examples evolve from held-out ones during unlearning—how targeted the update is. The two readings coincide for every method evaluated here, but can come apart for a method that perturbs held-out behavior as much as forget behavior, moving both populations together so the distance stays small even as forgetting occurs; such methods lie outside the validated set, and separating the two readings on them is future work.

This is a property of the regime, not of the retrain proxy. The obvious objection is that the language-model oracle is a gradient-difference *proxy* whose ascent might inflate the distance artificially. It does not: a **true from-scratch retrain oracle** (retrained on the retain split, two seeds) reads *intense-high* (high ATR-OOD, which is the forgetting reading in this regime) on the same held-out axis (inside the proxy's band and well above the no-op, Table D.7), so ATR reads the true oracle the same way it reads the proxy, if anything slightly *more* intense, making the proxy a faithful, mildly conservative stand-in. The reorientation is therefore intrinsic to the language-model regime, not an artifact of how the oracle is approximated. (Shown on the single-state descriptor a static oracle can produce; the trajectory descriptor requires an optimization run, so oracle fidelity is established for the single-state form.)

Two caveats bound this cross-scale reading. First, a monotone reorientation of the two extreme families preserves any mid-order method regardless of its mechanism, so the mid-order family here cannot separate genuine mechanism-sharing stability from mere middle-position preservation. Disambiguating the two would need a mechanism-sharing family that is not mid-ranked at both scales, a distillation method run on both image classifiers and language models, which the standard toolkit does not yet supply.

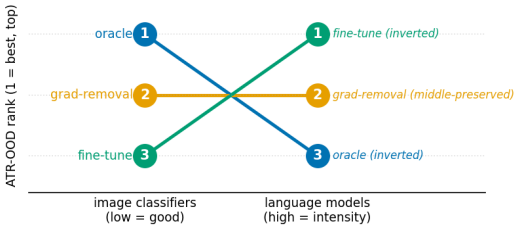


Figure D.1: **Cross-scale reorientation.** ATR-OOD method ranks on image classifiers against language models, for the three families evaluable at both scales. The ordering inverts rather than degrades: the from-scratch oracle moves from cleanest to most intense, the barely-moving fine-tune the reverse, and the gradient-based removal family (labeled "grad-removal"; NegGrad+ on image classifiers, NPO on language models) stays mid-order. Reported qualitatively across the three families (cross-scale Kendall $\tau = -1$, $n = 3$).

Table D.5: Cross-scale reorientation. The values behind the rank chart above and behind Figure 4's cross-scale claim. ATR-OOD and its rank for the three method families evaluable at both scales; low-is-good on image classifiers (↓ = more forgotten), high-is-forgetting-intensity on the language model (↑ = more forgotten). The two extreme families swap ends (cross-scale Kendall $\tau = -1.0$, n = 3, qualitative).

Family	**image classifier (rank)**	**language model (rank)**	
from-scratch oracle	0.0014 (1)	0.105 (3)	inverted
gradient-removal	0.130 (2)	0.101 (2)	preserved
fine-tune	0.143 (3)	0.032 (1)	inverted

Orientation calibration (candidate). ATR-OOD's sign is regime-dependent (§3, §5.4), so a practitioner on a new regime needs a way to fix it without the gold that ATR replaces. A two-anchor calibration does this shadow-free. Score a known no-op (an untouched model) and a known over-forgetter (an aggressive run driven well past the reference), then read the sign off the pair: the direction running from the no-op toward the over-forgetter is "more forgotten." This assumes ATR-OOD is monotone in forgetting over that range. That holds across the six 1.5B methods here (retrain-proxy down to the leaky cluster is monotone, §5.1), but is not guaranteed once a method over-forgets past the reference, where the distance can turn back. Mapping the monotone range is future work; until then we rely on the per-regime orientation and gold confirmation of §5.4. A cheap cross-regime check we have not yet run is to apply this two-anchor rule on the image-classifier regime, where the correct sign is independently known from U-LiRA (§5.4); agreement there would be its first cross-regime validation, and we flag it as the concrete next step.

A complementary reading fixes the *sign* without any oracle. Take the mean per-channel standardized displacement of the forget centroid from the reference, $s = \mathrm{mean}_d\,(\mu_f[d] - \mu_o[d])/\sigma_o[d]$. It is positive for every 1.5B method (the forget set sits above the reference on the loss channel), so its sign encodes the orientation, while its magnitude still ranks the methods (τ-b $= +0.96$ against the membership gold, contested pair correct). That is a small ranking cost against the symmetric MMD's $+1.0$, in exchange for a readable sign. It is a candidate oracle-free orientation reading, sign-consistent at 1.5B, but not yet validated across regimes.

D.4 Forget-split footprint

To check that the language-model evaluation is not an artifact of a single forget split, ATR-OOD was run on three TOFU forget splits at 0.5B, shadow-free, on a retrain-proxy run scored on the held-out trajectory.

This broadens the footprint beyond the single split used for the main results, with two limits. First, it is not a finite-sample test of the reference population: it varies the forget-set size with the reference fixed. Second, the roughly six-fold swing in the raw score across splits is confounded by the

Table D.6: Reference-dependence on image classifiers (CIFAR-10, random-subset): ATR-OOD (MMD²) of each method against three references: held-out test (O1), retain (O2), synthetic noise (O4), with U-LiRA AUC for orientation (O3, true-OOD, is a further reference control not used in this comparison). This is the per-method detail behind Figure 4. The test-set and retain axes rank the methods in opposite orders (Kendall τ vs U-LiRA $+1.0$ and -1.0); the same flip holds on the class-forget regime.

Method	test-set (O1)	retain (O2)	noise (O4)	U-LiRA AUC
retrain (oracle)	0.002	0.500	0.534	0.533
fine-tune	0.232	0.007	1.067	0.684
SCRUB	0.038	0.336	0.455	0.607
NegGrad+	0.176	0.013	0.967	0.624
τ vs U-LiRA	$+1.0$	-1.0	$+0.67$	

Table D.7: True-oracle fidelity at 7B: single-state ATR-OOD against the holdout-10 reference (forget set, final state; 15 samples per unlearner, the oracle re-scored at two seeds). The from-scratch retrain oracle reads intense-high like the gradient-difference proxy. The reorientation is a regime property, not a proxy artifact.

Model state	single-state ATR-OOD [95% CI]	reading
no-op (untouched M_{orig})	0.0325	not forgotten
fine-tune (weak)	0.0315 [0.0283, 0.0347]	$\approx$ no-op
retrain proxy (gradient difference)	0.0590 [0.0197, 0.1106]	forgotten
NPO	0.0636 [0.0302, 0.1082]	forgotten
true from-scratch retrain oracle	**0.0736** (0.0737 / 0.0735)	most intense

biased MMD² estimator's O(1/n) bias (larger at smaller forget-set size, in the observed direction), by different forget samples, and by a target-versus-shadow source difference. So raw ATR-OOD is not comparable across split sizes without same-estimand normalization.

D.5 Language-model faithfulness at 7B (argued, not measured)

The 1.5B faithfulness check (§5.1) does not transfer to 7B, because the shadow-membership gold cannot be established there. We built the same per-example attack at 7B. The attack itself is powerful (member-versus-non-member AUC 0.9925), but it cannot rank the standard methods: the gradient-difference retrain proxy lands **off the manifold the attack is calibrated on** (Table D.9). Its forget truth-ratio is driven to 0.511, below both the member model (0.679) and the true retrain oracle (0.702), so its membership score reads 0.698 (95% CI $[0.664, 0.728]$), far from the chance level (0.5) a clean forgetter should show. Escalating the shadow count does not fix this; it is a property of 7B gradient-difference at the standard configuration (verified faithful to the working 1.5B config). Consistently, the contested retrain-proxy-versus-NPO pair flips at 7B (NPO above the retrain proxy, Kendall $\tau = +0.33$ at matched epochs, Table D.10). This is why a 7B gold standard is needed and cannot be improvised.

A partial fidelity check is available on the single-state descriptor (dimension 4, which a static oracle can produce, unlike the trajectory descriptor the headline uses): a true 7B retrain oracle reads at ATR-OOD 0.0736, inside the gradient-difference proxy's band (mean 0.059) and above a no-op (0.033) and a weak fine-tune (0.031), so the proxy reads like the true oracle, slightly conservative (Table D.7). This supports, but does not establish, 7B transfer: the headline claim stays validated at 1.5B, argued at 7B.

D.6 MUSE PrivLeak operating-point attempt

We attempted the direct MUSE PrivLeak measurement that §6 reports as inconclusive. On MUSE-News the Min-K%-based PrivLeak estimator is itself valid: it cleanly separates the never-saw-forget retrain (forget-set membership AUC 0.525 near chance, PrivLeak -4.9) from the memorized target (AUC 0.999, PrivLeak -99.8). So any failure below is about the operating point, not the tool.

Table D.8: ATR-OOD across three forget-set sizes at 0.5B. Absolute interval width shrinks with more forget samples; relative precision (interval width over score) is roughly flat.

Forget split	Forget-set size	ATR-OOD (held-out trajectory)	Relative interval (width / score)
forget01	40	0.693	0.36
forget05	200	0.247	0.37
forget10	400	0.116	0.46

Table D.9: 7B shadow-membership Stage-0 gate, against the working 1.5B fleet. The attack is powerful (AUC 0.99) but the gradient-difference retrain proxy is off-manifold (membership score 0.70, not near the chance level a clean forgetter should read), so the gold cannot rank the methods at 7B. Membership scores carry 95% intervals; the contested pair is resolved (non-overlapping) at both scales.

	7B (N=16 shadows)	1.5B (N=32 shadows)
member-vs-non-member AUC	0.9925	0.9806
fine-tune membership score (leaky)	0.880 [0.850, 0.902]	0.823
retrain-proxy membership score (should be ≈ 0.5)	**0.698** [0.664, 0.728]	0.513
NPO membership score	0.617 [0.585, 0.647]	0.603

Two standard unlearners, epoch-scaled at default hyperparameters, bracket the useful corner without reaching it: gradient difference reduces the forget signal only by collapsing retain utility, and at high epochs overshoots to PrivLeak $+109$; NPO preserves retain but never leaves the memorized corner (PrivLeak near -99.7 across 2 to 10 epochs). With no useful regime in which to sit, there is no operating point at which to measure PrivLeak's cross-seed spread. We did not search per-method hyperparameters for one, deliberately: that search is itself the operating-point assumption ATR makes explicit. The finding is scoped to epoch-scaling at default hyperparameters for these two methods, not a proof that no configuration works.

D.7 Controls: drift, forget-side movement, cadence, seed-averaging

Four controls test whether ATR-OOD's 1.5B faithfulness reflects forgetting rather than a simpler quantity, and whether the reference-anchoring earns its keep (Table D.12).

Drift (magnitude of update). Ranking the six in-regime methods by retain-set loss drift (final minus pre-unlearning loss on the retain-control population) agrees with the shadow-membership ranking at only Kendall $\tau = 0.33$, against ATR-OOD's $+1$; among the retain-preserving methods the drift is near zero (0.00 to 0.05 nats, except NPO at 1.13) and cannot order them, so ATR-OOD is not a re-description of how far the run moved on the retain side.

Forget-side movement (a non-degenerate control). The drift control is near-degenerate: retain-loss barely moves for the retain-preserving methods, so it cannot order them. Its forget-side analogue is not degenerate. Ranking the six methods by how far each moves the forget set from the pre-unlearning model, measured as mean forget-loss movement or as a self-referenced MMD of the final-versus-step0 forget-loss distribution, reaches Kendall $\tau = 0.69$ and 0.83 against the membership ranking, better than retain-drift's 0.33 but short of ATR-OOD's $+1$. It fails on the contested pair: NPO moves the forget set farthest of any in-regime method (mean movement 1.31 nats; final-versus-step0 MMD 0.81, against retrain-proxy's 0.16 and 0.02), yet the membership standard places retrain-proxy as the more forgotten. Distance moved is not forgetting; the held-out reference, which measures distance to where forgotten data should sit rather than distance travelled, is what recovers the exact ranking.

Discriminability. On the nine-configuration grid, ATR-OOD's between-method separation is 1.75 times its within-method cross-seed standard deviation (0.040 against 0.023, here the raw standard deviation rather than the coefficient of variation used for the seed-swing figures of §5.2), against

Table D.10: Epoch-resolved 7B ATR-OOD (held-out trajectory, ↑ high = more forgetting-intensity), against the 1.5B faithful ordering. At matched (2-epoch) training the contested retrain-proxy-versus-NPO pair flips (NPO reads higher), giving Kendall $\tau = +0.33$; the flip is epoch-local: at one epoch and on the epoch-mean the 1.5B ordering returns ($\tau = +1.0$), which is exactly why a 7B gold standard is needed to adjudicate rather than assumed.

Basis	retrain proxy	NPO	fine-tune	ordering	τ vs 1.5B
1.5B (2 ep)	0.0677	0.0546	0.0275	rp > npo > ft	n/a
7B, ep 1.0	0.1505	0.0732	0.0331	rp > npo > ft	+1.00
7B, ep 1.5	0.0716	0.1013	0.0323	npo > rp > ft	+0.33
7B, ep 2.0 (matched)	0.0925	0.1288	0.0304	npo > rp > ft	+0.33
7B, epoch-mean	0.1049	0.1011	0.0319	rp > npo > ft	+1.00

Table D.11: MUSE-News PrivLeak (Llama-2-7B, Min-K%): the valid self-validity corners, then the epoch-scaling sweeps (seed 0). GradDiff stays memorized then overshoots as retain utility collapses; NPO preserves retain but never forgets; neither reaches the oracle's useful corner (retainKnow ≈ 0.55, forgetKnow ≈ 0.32, PrivLeak ≈ -5), so there is no operating point at which to measure a cross-seed spread.

Model / sweep	PrivLeak	retainKnow	forgetKnow	forgetVerb
retrain oracle (never saw forget)	-4.9	0.552	0.324	0.204
target (memorized)	-99.8	0.544	0.654	0.569
GradDiff ep1	-99.8	0.548	0.656	0.532
GradDiff ep3	-99.5	0.528	0.641	0.456
GradDiff ep5	-84.8	0.433	0.535	0.260
GradDiff ep10	+109.4	0.277	0.367	0.003
NPO ep2	-99.7	0.555	0.656	0.490
NPO ep5	-99.6	0.553	0.646	0.483
NPO ep10	-99.7	0.520	0.642	0.498

Forget Quality's 1.02 times (0.172 against 0.168): ATR-OOD resolves methods well above its own seed noise, whereas Forget Quality barely does, so its stability is not the artifact of a metric that fails to move.

Cadence. ATR's default descriptor is built from reductions computed over the step index rescaled to the unit interval (trend slope and area under the curve), so it is largely cadence-normalized by construction. Subsampling the trajectory to half (101 of 201 checkpoints) and quarter (51) cadence and recomputing the full four-collector Tier-1 composite (the descriptor the faithfulness ranking rests on, not the loss channel alone) confirms this: the retrain-proxy and NPO stay first and second at every cadence, and the complete six-method ordering agrees with the gold at $\tau = 0.55$ (full), 0.41 (half), and 0.41 (quarter), only the satimp/finetune near-tie reordering. This tests robustness to reading one run's checkpoints more sparsely, not to a genuinely different checkpointing schedule during training, which we do not vary; an adversary who deliberately games the trajectory instrumentation is a distinct adaptive threat, which we treat as the main open robustness question (§6).

Averaging over training-seed re-runs. Table 1's instability is across the unlearning run's *training* seed, with the retrained reference held fixed. Averaging the metric over several such re-runs is the obvious remedy, and its cost is the point. Single-seed Forget Quality flips the three-method ordering on a third of the seed-pairs (0.333, Table 1). Averaging over five re-runs and bootstrapping that averaged ranking drops the residual flip rate to about 1% (0.011). This is a bootstrap flip of the *averaged* ranking, not the single-run pairwise rate of Table 1, so the two numbers are not directly comparable; and it costs five unlearning runs, where ATR-OOD reaches zero flips from one. Averaging shrinks the aggregate's variance without touching the per-run threshold read-out. It is also distinct from pooling the *reference*, which the Table D.3 diagnostic shows does not help at all. So the reference-aware advantage here is cost as much as stability (§5.5).

ATR-OOD's zero-flip has a direct paired-seed reading. Its per-method seed deviations *co-move* across methods (mean cross-method correlation $+0.23$, against Forget Quality's -0.13), and its pairwise ordering is sign-consistent across all five seeds, where Forget Quality's is not (0.93). A

Table D.12: The four controls against the 1.5B shadow-membership gold. Each asks whether a simpler quantity reproduces ATR-OOD's faithful ranking; none does. "Inverts pair" = ranks NPO above the retrain proxy, which the gold resolves the other way.

Control	Statistic	Result	vs gold
Retain-side drift	final - initial retain loss	≈ 0 for retain-preservers; NPO 1.13	$\tau = 0.33$, degenerate
Forget movement (mean)	mean forget-loss movement	NPO 1.31 $>$ retrain-proxy 0.16	$\tau = 0.69$, inverts pair
Forget movement (MMD)	final-vs-step0 forget MMD	NPO 0.81 $>$ retrain-proxy 0.02	$\tau = 0.83$, inverts pair
Discriminability	between / within separation	ATR-OOD 1.75×; Forget Quality 1.02×	ATR-OOD resolves, FQ barely
Cadence (full/half/quarter)	full-descriptor ordering	retrain-proxy $>$ NPO stable	$\tau = 0.55/0.41/0.41$
Seed-averaging	ranking flip rate	single-seed FQ 0.333 $\rightarrow$ 5-seed 0.011	ATR-OOD 0.000

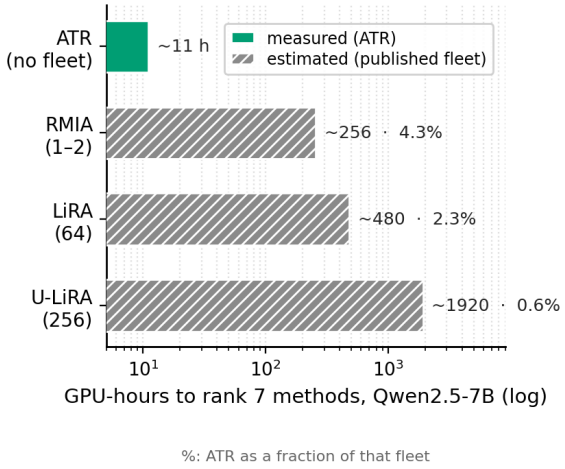


Figure D.2: **Cost to rank the methods.** GPU-hours to rank seven TOFU methods on Qwen2.5-7B (log scale): ATR (measured, solid) instruments the runs one performs anyway, while the shadow-membership fleets (estimated at published shadow counts, hatched; no 7B fleet was trained) cost far more, putting ATR at about 0.6% of a 256-shadow U-LiRA. Full accounting in Table D.13.

seed nudges the whole field together rather than reshuffling it, which is why the ordering holds despite the per-configuration spread.

D.8 Cost detail

The full accounting behind §5.5. ATR's cost is $O(\text{methods})$; a per-example shadow attack is $O(\text{methods} \times \text{shadows})$, and that is the gap.

What the headline figures count. For a like-for-like comparison both sides count the unlearning runs. ATR's $\sim$11 GPU-hours at 7B is the seven method runs it instruments (which a deployment performs regardless), plus their instrumentation overhead (3.8–8.0%) and CPU-only post-hoc scoring (free on the training loop); no shadow models and no retrain oracle are trained. The shadow-fleet figures are counted on the same basis: each shadow is a full train-and-unlearn run, so U-LiRA's $\sim$1920 GPU-hours is 256 shadows $\times$ (train $+$ unlearn $+$ score) per method. The gap is therefore not an artifact of charging the base unlearning runs to one side; it is the per-method fleet multiplier ($O(\text{shadows})$ extra runs) that ATR removes entirely, keeping only the runs a practitioner already performs.

Table D.13: Full cost accounting. Standalone GPU-hours to rank seven TOFU methods on Qwen2.5-7B (ATR measured; the shadow-attack fleets estimated at their published shadow counts, since no 7B fleet was trained), the directly measured 1.5B comparison against the 64-model fleet we ran, and its projection to the published shadow scale.

Comparison	shadow models	GPU-hours	ATR fraction
7B, rank 7 methods (fleets estimated)			
ATR (this work)	none	∼11 (measured)	n/a
U-LiRA	256	∼1920	∼0.6%
LiRA	64	∼480	∼2.3%
RMIA	1–2	∼256	∼4.3%
1.5B, directly measured			
ATR vs the 64-model fleet	64	∼5 vs ∼70	∼7%
ATR vs 256 shadows (projected)	256	∼5 vs ∼280	∼1.8%

Table D.14: Production-horizon instrumentation overhead versus collection cadence (7B TOFU, wall-clock). A naive every-few-steps cadence is expensive; the cadence-100 operating point brings both method classes under or near 10%.

Collection cadence	NPO overhead	gradient-difference (cheap-step) overhead
6 (naive)	72.0%	191.8%
50	8.6%	23.0%
100 (operating point)	4.3%	11.5%

Per-run instrumentation overhead. At the operating point (cadence 100, population 100), wall-clock overhead is $\leq$10% for realistic deployment: NPO 3.8% (95% range $[1.9, 10.7]$), the cheaper gradient-difference steps 7.7–8.0%. It falls as training lengthens, because the fixed two-event collect amortizes over more optimizer steps (cheap-step mean 12.8% at one epoch, 6.1% at 1.5, 4.7% at two; NPO 6.3% / 3.0% / 2.0%). Peak GPU memory rises by ∼2.5 GB (≈5%, population-independent, collection chunked at eval-batch 4); disk is 0.04 MB per run. On image classifiers the wall-clock overhead is +4.91% with peak GPU memory +31.3% (≈340 MB, 1.4% of a 24 GB card) and 0.17 MB of disk, post-hoc scoring 2.16 s on CPU. Quadrupling the forward population to 400 roughly doubles the wall-clock overhead (NPO 9.7%, cheap-step ∼22%) for no ranking-quality gain, which is why the operating point fixes the population at 100. Table D.14 shows how the overhead falls with a sparser collection cadence.

D.9 Further scope boundaries

Beyond the four limitations of §6, several narrower boundaries are deliberate scoping choices rather than defects; Table D.15 collects them.

D.10 Robustness detail: detection and gaming

Detection (the decisive values behind Table 2). Over-forgetting (S1) drives the forget loss so far into the tail that the default RBF-MMD saturates; a 1-Wasserstein reading recovers it, at 4.2 against a good unlearner's 0.5 and a seed floor of 0.04 (∼10× SCRUB, ∼96× the floor). Under-forgetting (S3) and the no-op (S6) are caught on the orthogonal axes: a memorized forget set that does not sit where the retain set sits.

Output-matched distillation (S2) is flagged, but not specifically. At image scale its ATR-RETRAIN distance (0.665) overlaps a genuine distillation unlearner's (SCRUB 0.43–0.74, against a 0.32 baseline). At language-model scale its activation-MMD ratio (12–28×) is *smaller* than NPO's (61–186×), while a preservation probe reads 0.97–0.99 for the attack and 0.95–0.99 for NPO (Table D.16). Against three simpler detectors run on the same stress matrix, ATR's axis-resolved reading is more specific: a raw forget-loss detector catches three of four, a Jensen–Shannon ag-

Table D.15: Further scope boundaries, each a deliberate limit rather than a defect.

Scope boundary	Status
training-free unlearning (no SGD trajectory)	out of reach by construction
post-deployment relearning (Hu et al., 2025), quantization-reversal (Zhang et al., 2025)	out of scope
benchmarks beyond TOFU (MUSE, WMDP)	not run; future work
7B retrain proxy is gradient difference, not from-scratch	true 7B oracle reads the same intense-high (partial check, Table D.7)
language-model specificity comparator over-forgets mildly	direction robust; magnitudes operating-point-dependent
SimNPO at published TOFU hyperparameters	under-forgets onto fine-tune (a non-portability instance, Appendix B)

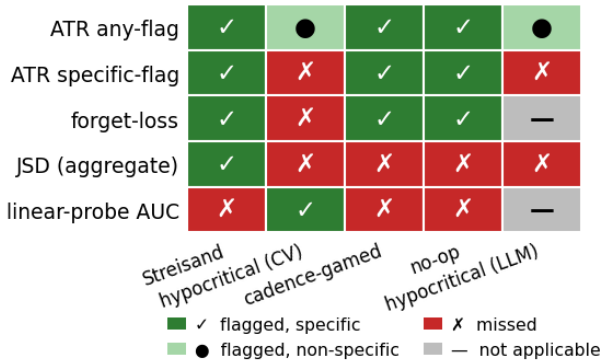


Figure D.3: **Stress-detection matrix.** ATR's three-way (attack vs. a good practical unlearner) detection across the stress tests and axes: the over-forget (S1), under-forget (S3), and no-op (S6) attacks are flagged specifically (a signal a good unlearner does not also produce), while output-matched distillation (S2) is flagged but not specifically at either scale. Brighter cells are stronger detections. The figure's column labels map to the stress IDs as: Streisand = S1 (over-forget), cadence-gamed = S3 (under-forget), no-op = S6, and hypocritical (CV / LLM) = S2 output-matched distillation on image classifiers / language models; S4 (post-deployment relearning) and S5 (quantization-reversal) are out of scope (Table D.15) and not shown.

gregate one of four (fooled by S2 at 0.033), and a linear probe one of four (it catches S2 at AUC 0.94 but misses S1 at 0.56; Table D.17).

Gaming. Two attacks keep the forget set memorized while gaming a cadence schedule or a final-checkpoint metric. Both are caught on the held-out reference, which registers the memorization they leave intact, across all regimes and seeds; the retain reference correctly stays silent. Two further attacks try to match the forget-loss distribution to the reference. Neither can be both functional and evasive: any objective that lowers the loss-based score enough to evade necessarily de-memorizes the forget set. The best evasion bottoms out near 0.69 (an ATR-OOD score still above the detection threshold) on the hardest regime, and this is a single-configuration reading rather than a seed-averaged one.

The robustness is scope-bounded. Detection rests on the held-out axis. High-memorization margins are near-threshold under a small reference (leave-one-out false-positive rate $\approx$0.10 per axis, about twice the 0.05 nominal). And inconstructability is argued from implementation plus mechanism, not an exhaustive class disproof. The remaining adversary (one that games the trajectory *instrumentation* itself, by smoothing the collected trajectory or injecting checkpoint-time noise) is, for a trajectory-based evaluator, the principal open robustness question rather than a settled result, and we name it as such (§6).

Table D.16: Detection decisive values (three-way: attack vs a good practical unlearner). "Specific" means a signal a good unlearner does not also produce; S2 is flagged but not specific at either scale.

Stress test	signal	attack	good unlearner	verdict
S1 Streisand (over-forget)	1-Wasserstein vs retrain	4.2	SCRUB 0.5; floor 0.04	specific ($\sim$10–96$\times$)
S3 cadence-gamed (under-forget)	retain axis (O2)	O2 low	O2 normal	specific
S6 no-op	held-out and retain axes	O1 high, O2 low	neither	specific
S2 distillation (image)	ATR-RETRAIN	0.665	SCRUB 0.43–0.74; base 0.32	not specific (overlap)
S2 distillation (LM)	activation-MMD ratio	12–28$\times$	NPO 61–186$\times$; probe 0.97 vs 0.95	not specific (smaller)

Table D.17: ATR's axis-resolved detection against three simpler detectors on the same four-stress matrix. A single distributional summary (JSD) or a learned probe catches at most one stress specifically; the axis-resolved reading catches three of four specifically and all four at any-flag.

Detector	S1 over-forget	S2 distillation	S3 under-forget	S6 no-op	specific total
raw forget-loss	yes	no	yes	yes	3/4
Jensen–Shannon aggregate	yes	no (0.033)	no	no	1/4
linear-probe AUC	no (0.56)	yes (0.94)	no	no	1/4
ATR (axis-resolved)	yes	flagged, not specific	yes	yes	3/4 specific, 4/4 any

D.11 Descriptor components against the membership gold

The default descriptor combines four collectors, two of them, per-sample loss and logit margin, confidence-type channels close to the truth-ratio input the shadow-membership standard is built from. To test whether ATR-OOD's exact 1.5B agreement (§5.1) merely re-packages that shared signal, we re-score the six in-regime methods from the same on-disk trajectories under descriptor ablations, each collector reduced by the same default composite and the sample set and scorer held fixed.

The agreement does not depend on the feature the standard itself uses: removing per-sample loss, the closest analogue to its truth-ratio input, leaves the ranking intact (τ-b = 0.96, contested pair correct), and any three of the four collectors reproduce it. The confidence-type signal is redundant across the loss and logit-margin channels, either alone recovering the full CI-aware ranking, so the faithfulness is convergent rather than circular. What does not suffice is the two geometric channels on their own: gradient and activation norm together reach only τ-b = 0.50 and invert the contested pair. ATR-OOD's language-model faithfulness thus rests on confidence-type trajectory information, held redundantly across collectors, not on any single channel and not on the geometric channels alone.

D.12 Read-out and artifact baselines

This is the thorough version of the §5.3 controls result. Ranking the six in-regime 1.5B methods on the same forget-set traces against the shadow-membership gold, we compare every cheaper alternative to ATR (the drift and forget-movement controls above, a smoother Forget-Quality read-out, and artifact-based per-record scores), showing what each buys and what it requires (Table D.20; Kendall τ over the six methods, ceiling +0.97 because satimp/simnpo/finetune form a CI-tied cluster the gold cannot resolve).

Forget Quality is the KS-test p-value against the retrained reference (Appendix A), so it is the read-out we score here; it ranks faithfully when pooled over seeds, but needs that reference and is seed-unstable per run (§5.2). Its p-value agreement of $+0.97$ is the same six-method agreement Table C.5 (and Figure 1) reports as CI-aware τ-b $= +1.0$. Both order the four gold-resolved levels exactly, and disagree only inside the satimp/simnpo/finetune cluster, whose overlapping intervals the gold cannot resolve, so the disagreement is on a distinction the gold itself does not make. The rank-correlation form penalizes an arbitrary order inside that near-tie, while the CI-aware form scores it as tied. The KS *statistic* D, read from the same test, ranks at only $+0.69$: the choice of read-out on the same reference matters.

Table D.18: Gaming robustness (CIFAR-10, ResNet-18, three regimes). Trace attacks (g1/g2): detection margin (attack / retrain-Q95) on the memorization-leak axis, over 18 cells, all detected. Construction attacks (g3/g4): best functional-configuration ratio; evasion requires ratio ≤ 0.5, never reached (floor 0.69). The N = 20 retrain reference is loose (leave-one-out false-positive $\approx$0.10 per axis), so verdicts are margins, single-seed frontier.

Attack	easy regime	class-forget	high-memorization	detected?
g1 cadence-gamed (O1 margin)	2.6–3.7×	1.7–2.8×	1.1–1.8×	18/18
g2 last-checkpoint-good (O1 margin)	2.5–3.5×	1.8–3.1×	1.0–1.7×	18/18
g3 mean-matching (best functional ratio)	2.32	n/a	0.69	evasion fails
g4 moment-matching (best functional ratio)	5.08	n/a	0.93	evasion fails

Table D.19: ATR-OOD's 1.5B faithfulness under descriptor ablation: point Kendall τ, CI-aware τ-b, concordant/discordant resolved pairs, whether the contested retrain-proxy-versus-NPO pair stays correct, and the exact-permutation p, over the six in-regime methods. Dropping any single collector, including the loss channel closest to the standard's input, preserves the ranking; only the two geometric channels alone break it.

Descriptor	τ	τ-b	concordant/discordant	Contested pair	perm-p
full four-collector	+1.00	+1.00	12/0	correct	0.0014
drop per-sample loss (the standard's analogue)	+1.00	+0.96	11/0	correct	0.0014
drop logit margin	+1.00	+1.00	12/0	correct	0.0014
drop gradient norm	+0.87	+1.00	12/0	correct	0.0083
drop activation norm	+1.00	+1.00	12/0	correct	0.0014
per-sample loss only	+0.73	+1.00	12/0	correct	0.0278
logit margin only	+0.73	+1.00	12/0	correct	0.0278
gradient plus activation only	+0.47	+0.50	9/3	inverts	0.136

The reference-free shortcuts do not reach the gold. The smooth membership statistic matches ATR-OOD, but is shadow-based, essentially the gold itself. The two statistics that need neither a shadow fleet nor a retrained model (a single-model mean truth-ratio, and a Pollock-style loss-trajectory-IQR aggregate) invert the contested retrain-proxy-versus-NPO pair. So a smooth read-out removes Forget Quality's instability but not its dependence on a reference, and a single trajectory artifact does not suffice where the full descriptor does. Trajectory information against an explicit reference is what buys faithfulness without a shadow fleet or a retrained oracle.

D.13 Scorer and descriptor-form robustness

Four checks re-score the six in-regime methods from the same 1.5B traces, with the sample set and reference held fixed, to ask whether the faithfulness ranking of §5.1 depends on the specific estimator, null model, kernel, or reduction (Table D.21). The full-descriptor recompute reproduces the stored ATR-OOD exactly (maximum relative drift 0.000), so the variants are directly comparable.

Estimator. Replacing the biased V-statistic with the unbiased MMD² U-statistic (diagonal excluded) preserves the ranking exactly (CI-aware τ-b $= +1.0$, twelve resolved pairs concordant, contested pair correct): the $O(1/n)$ bias does not drive the result.

Null test. A permutation test of the null $P_f = P_{\text{ood}}$ for the multi-RBF scorer (label-shuffle, bandwidth fixed on the pooled sample, 200 permutations) rejects for all six methods in every seed (median p at the 0.005 floor), so the distances ATR ranks are real distributional differences, not finite-sample noise; the method ranking is read from their magnitude, not their significance.

Scorer. A non-saturating aggregate distance (the sum of per-dimension 1-Wasserstein distances, the kind of tail-shift supplement named in Table B.1, case 7) reaches τ-b $= +0.83$ but inverts the contested retrain-proxy-versus-NPO pair. It is the RBF kernel's joint structure, not merely a non-saturating distance, that resolves that pair; the bandwidth-saturation caveat concerns tail-shift detection rather than this ranking, and a marginal non-saturating distance does not improve on the RBF-MMD here. Replacing the kernel MMD entirely with a bounded, bandwidth-free two-sample statistic, the cross-validated AUC of a logistic domain classifier (Lopez-Paz & Oquab, 2017) separat-

Table D.20: Every cheaper alternative to ATR ranked against the 1.5B shadow-membership gold: what each requires, the exact values it assigns the contested pair (retrain-proxy / NPO, which the gold resolves as retrain-proxy the more forgotten), Kendall τ over the six methods, and whether it keeps that pair. Orientation for the two golds is noted ($\downarrow$ or $\uparrow$ = more forgotten); every reference-free shortcut reads NPO as having moved or changed more than retrain-proxy, so it ranks NPO the more forgotten and inverts the pair. Only ATR-OOD reaches the gold ranking needing neither a shadow fleet nor a retrained model. The two Forget-Quality rows show its read-out matters: the KS statistic D ranks at $+0.69$, its p-value at $+0.97$.

Statistic	Requires	retrain-proxy / NPO	Kendall τ	Contested pair
shadow-membership score ($\downarrow$)	shadow fleet	0.51 / 0.60	+0.97	correct
ATR-OOD, trajectory + reference ($\uparrow$)	**neither**	0.068 / 0.055	+0.97	correct
Forget Quality, KS p-value	retrained model	n/a	+0.97	correct
Forget Quality, KS statistic D	retrained model	n/a	+0.69	correct
retain-side drift (nats)	neither	$\approx$0 / 1.13	+0.33	degenerate
forget-side movement, mean (nats)	neither	0.16 / 1.31	+0.69	inverts
forget-side movement, final-vs-step0 MMD	neither	0.02 / 0.81	+0.83	inverts
single-model mean log-truth-ratio	neither	-0.20 / -0.18	+0.55	inverts
Pollock loss-IQR aggregate	neither	0.13 / 1.29	+0.55	inverts

Table D.21: Scorer and descriptor-form robustness: CI-aware rank agreement (τ-b) with the 1.5B membership gold, whether the contested pair stays correct, and the mean cross-seed spread (standard deviation, as in §D.7) of each variant. The estimator swap is harmless; a marginal non-saturating distance and any single-summary reduction invert the contested pair.

Variant	τ-b vs gold	Contested pair	mean cross-seed spread
full 24-dim, biased V-statistic (default)	+1.00	correct	0.046
full 24-dim, unbiased U-statistic	+1.00	correct	0.062
sum of per-dimension 1-Wasserstein	+0.83	inverts	0.029
mean-only 4-dim	+0.83	inverts	0.024
final-checkpoint-only 4-dim	+0.83	inverts	0.028

ing the forget from the reference descriptors, reproduces the ranking exactly (CI-aware τ-b $= +1.0$, contested pair correct), so the result does not hinge on the kernel or its bandwidth heuristic; we keep the MMD for its exact-zero property and closed-form bootstrap, not for any ranking advantage over a classifier two-sample test.

What the trajectory buys. Reducing each collector's trajectory to a single summary rather than the six-reduction descriptor costs exactly the contested pair: a mean-only four-dimensional descriptor and a final-checkpoint-only four-dimensional descriptor both fall to τ-b $= +0.83$ and invert retrain-proxy versus NPO, while the full 24-dimensional trajectory descriptor reaches $+1.0$. Mean-only and final-checkpoint-only are themselves indistinguishable (cross-seed spread 0.024 versus 0.028), so the gain is not trajectory-mean over final checkpoint but the richer reduction set: level, endpoint, direction, and spread together; the full descriptor's spread is marginally wider (0.046), the modest cost of that set. The multi-reduction trajectory descriptor is what recovers the pair a single checkpoint gets backwards.

This also isolates *where* ATR's seed-stability comes from, the question of whether it is the trajectory descriptor or merely the avoidance of a p-value. A cheap, smooth, retrain-free comparator—the final-checkpoint-only four-dimensional descriptor scored the same way against the held-out set, needing no shadow fleet and no retrained oracle—is already about as stable as the full descriptor (cross-seed standard deviation 0.028 vs 0.046) and far tighter than Forget Quality's KS p-value. So the seed-stability is a property of scoring with a smooth two-sample *distance* rather than of the trajectory; what the trajectory buys is *faithfulness* (τ-b $= +1.0$ against the gold, versus the final-checkpoint comparator's $+0.83$, which inverts the contested pair). (These spreads are over the seeds of the 1.5B faithfulness re-scoring, not the 7B Forget-Quality grid; a matched-seed stability comparison of this comparator on that grid is the cleaner form and remains to be run.)

Table D.22: Calibrated ATR at 1.5B: fixed-bandwidth ATR-OOD distance D and its rescaling to $[0, 1]$ (0 = no-op, 1 = retrain oracle). The rescaling is monotone, so it preserves the gold ranking (0/12 discordant).

Method	D **(fixed bandwidth)**	**Calibrated ATR** $[0, 1]$
retrain proxy (oracle anchor)	0.077	1.00
NPO	0.059	0.62
WGA	0.047	0.35
SatImp	0.032	0.02
SimNPO	0.032	0.02
fine-tune	0.032	0.02
no-op (M_{orig}, anchor)	0.031	0.00

D.14 CALIBRATED ATR: AN INTERPRETABLE [0,1] READOUT FROM A RETRAIN ORACLE

ATR-OOD ranks methods, but is an uncalibrated MMD² distance: its magnitude has no absolute meaning and its orientation flips by regime (§5.4), so unlike a membership score it does not by itself report *how much* was forgotten. Given a faithful reference, it rescales to an interpretable [0,1] with two anchors, a no-op run (the original model M_{orig}, a flat trajectory at step 0, "nothing forgotten") at 0, and the retrain oracle at 1 ("as-if-retrained"; at 1.5B this anchor is the gradient-difference *retrain proxy*, $D_{\text{oracle}} = 0.077$, since no true from-scratch oracle is trained at that scale—its fidelity to a true oracle is established in §D.3, Table D.7), via $\text{ATR}_{[0,1]} = \text{clip}\big((D - D_{\text{no-op}})/(D_{\text{oracle}} - D_{\text{no-op}}), 0, 1\big)$. The affine map through both anchors absorbs the regime-dependent orientation automatically and is monotone, so the §5.1 ranking is preserved.

We compute it on the six 1.5B methods with a single fixed kernel bandwidth, so the three distances are comparable ($D_{\text{no-op}} = 0.031$, $D_{\text{oracle}} = 0.077$), first checking that the fixed-bandwidth ranking still matches the shadow-membership gold (τ-b $= +0.894$, 0 of 12 resolved contrasts discordant). The readout (Table D.22) places the leaky fine-tune/SimNPO/SatImp cluster within 0.02 of the no-op, WGA at 0.35, NPO at 0.62, and the retrain proxy at 1.0, a "fraction of the way to a from-scratch retrain" that reads on its own while agreeing with the membership ordering.

Two caveats bound it. It is **not retrain-free**: the upper anchor is a retrain oracle, trained once per dataset and forget split and amortized over all methods—the same oracle the ATR-RETRAIN diagnostic (§3) uses, though here it fixes the $[0, 1]$ anchor for rescaling ATR-OOD rather than serving as the MMD reference—still far cheaper than a per-method shadow fleet, but no longer the retrain-free headline score. And it is a **distance-based indicator, not a per-example privacy score**: [0,1] measures the forget set's descriptor distribution between no-op and oracle, not an information-theoretic membership advantage; the two are correlated (the §5.1 faithfulness result) but not the same quantity. The scale is meaningful only inside the monotone range between the anchors, an over-forgetter clamps at 1.