Robustness Rankings Sort on the Wrong Axis: The Case for Iso-Depth Evaluation in Machine Unlearning

Anonymous authors
Paper under double-blind review

Abstract

Unlearning methods are ranked by their resistance to relearning: how quickly fine-tuning recovers a forgotten set. We show these rankings are confounded by suppression depth, how far into the forget loss a method drives the model, which single-run, default-recipe comparisons leave uncontrolled. On a depth-controlled ladder across three model families (GPT-2-medium, Phi-1.5, Llama-3.2-1B; Llama-3.2-3B) and two benchmarks (TOFU, MUSE), a balanced variance decomposition attributes relearning resistance overwhelmingly to depth and run-to-run noise on TOFU: depth explains 63–68% of the variance over our ladder and the method only 0.5–8.1%, below run-noise in every depth window. We propose iso-depth evaluation: compare at matched depth, report run variance, and pin the scored checkpoint, rather than ranking default recipes. We observed that the residual method effect is small on TOFU but larger on MUSE (28–37%), tracks a second measurable quantity, weight displacement: at matched depth, resistance covaries with displacement and the spread of displacement across methods tracks the benchmark-dependent method share. This second axis is observational; within fictitious-QA it is secondary to depth, while on the corpus it is the larger within-benchmark predictor. We report it as a refinement of the depth result, not a co-equal controlled axis.

1. Introduction

Machine unlearning aims to remove the influence of a forget set from a trained model. Because exact retraining is expensive, most deployed methods are approximate, and a central worry is that they merely suppress rather than erase: a small amount of fine-tuning can "jog" the model's memory and recover the supposedly-forgotten content (Hu et al., 2025). This has made relearning resistance, the number of fine-tuning steps (or epochs) needed to recover the forget set past a threshold, a standard robustness metric, used both to motivate defenses (e.g. sharpness-aware unlearning (Fan et al., 2025)) and to rank methods against each other.

We argue that relearning-resistance comparisons, as currently practiced, are confounded by a variable nobody controls. When a method unlearns, it pushes the forget-set loss up by some amount before the relearning attack starts; we call this the suppression depth, measured in nats. A model parked deeper takes longer to relearn for a near-mechanical reason: it starts farther from the recovery threshold, independent of any property of the method that put it there. Two facts make this confound invisible in practice:

1. Methods reach different depths at their default configurations. A gradient-ascent method and a preference-based method, run to their published recipes, do not stop at the same forget-loss level.
2. Report metrics saturate. Forget-set accuracy floors at chance; exp(−loss) reads 0.000 for any loss past ~8 nats. A model at 16 nats and a model at 120 nats both print "forget ≈ 0," so a reader cannot see that one was suppressed ~7× deeper than the other.

Consequently, a single-run comparison that reports "method A resists longer than method B" may be reporting only "method A's default recipe parks it deeper", plus a large dose of run-to-run noise. We make this quantitative.

Depth turns out to be only the first axis. Controlling it on a depth ladder across two benchmarks (fictitious-QA TOFU (Maini et al., 2024) and a real news corpus, MUSE (Shi et al., 2024)) and two model families, we find a second: weight displacement, how far a method moves the weights from the memorized model. At matched depth, relearning resistance tracks displacement. Methods reach similar displacement on fictitious QA, so once depth is controlled the residual is mostly run-to-run noise; on the real corpus they reach very different displacement, so a large "method effect" appears. At matched depth, much of what a single-run ranking reads as method quality is covariation with weight displacement. The recommendation sharpens: compare at iso-depth, report run variance, and report displacement.

Contributions.

- We define and measure depth-response curves: relearning resistance as a function of suppression depth, holding the method fixed. Across three architectures and four loss families (RMU cannot reach $\geq$4 depth bins), 49 of 50 per-(method, seed) curves are monotone (§5.1).
- We give a balanced variance decomposition of relearning resistance into depth, method, and run (trajectory-seed) components, on two benchmarks. On TOFU (fictitious QA), depth explains 63–68%, method 0.5–8.1%, and run-to-run noise 26–34%; run noise exceeds the method share on all three architectures and in every depth window. On the MUSE news corpus the method share rises to 28–37%; on the two $\leq$1.4B corpus models it overtakes run noise (and depth, on Llama-3.2-1B), while at Llama-3.2-3B it rises (32%) but run noise (45%) stays the largest term. Run noise thus exceeds the method share on TOFU at every scale and on the corpus at 3B, and is overtaken only on the two $\leq$1.4B corpus models. The pre-registered GD-faster effect (15–25% faster at matched depth, Bonferroni-significant on Llama) is quantified rather than rounded to zero (§5.2).
- As a secondary, observational refinement (§5.9), we find that residual tracks weight displacement: across both benchmarks and both $\leq$1.4B model families, relearning time at matched depth tracks the L2 distance the method moved the weights from the memorized model (within-bin Spearman 0.59–0.93). Methods converge in displacement on fictitious QA (hence the small method share) but diverge on the real corpus (hence the large one), so displacement is a candidate mechanism for the benchmark-dependence of the method effect rather than a uniformly stronger predictor than depth (depth leads within TOFU; displacement overtakes it only on the corpus; §5.9). The axis is observational (measured, not intervened, confounded with optimization length) and is not required by the protocol recommendation above; RMU is an exception.
- We show the practical cost (§5.6, with its full history disclosed): at the standard TOFU recipe, three of four methods fail to unlearn at all, and the robustness comparison is degenerate before it starts; at an 8$\times$ budget (chosen post hoc, labeled exploratory), the four methods land at method-characteristic depths spanning ~2.4–4$\times$ and the resulting resistance ranking tracks reached depth at Spearman +1.0 on all three architectures, with default-config spreads of 4.5–5.9$\times$ collapsing to 1.3–1.7$\times$ at matched depth. Either way, a default-config comparison measures depth, not method.
- We show the confound in a published comparison, from its own numbers (§5.7): joining the per-method relearning-delay values from the RCD paper (Tran & Woo, 2026) to its own reported forget accuracies, the 10$\times$ published RCD spread tracks starting depth ($\rho = -0.93$ to -0.96 across candidate joins; linear $R^2 \approx 0.81$), partly by construction, since their metric is depth-seeded at relearning epoch 0, so we read it as a consistency illustration (effect size, not significance; baselines-only n=5 is p=0.083) rather than independent evidence, with an adverse proxy check reported alongside.
- We execute that registered reproduction (§5.8): on CIFAR-10/ResNet-50 the forget accuracies and the depth–RCD direction reproduce (Spearman -0.6 to -0.9; SCRUB a near-exact positive control), but the published per-method ranking is irreproducible under any uniform checkpoint-selection rule, a choice that alone moves a method's RCD up to ~8.6$\times$ and rests on an unspecified, depth-equivalent checkpoint choice. Whether their ranking dissolves needs a faithful reproduction of it (the open lever). Re-scoring our own reproductions (a flagged deviation from the registered gate, §5.8), SALUN's RCD tracks realized depth (10.4 at over-forgotten 0.79 $\rightarrow$ 1.1 at matched 0.92), but

at matched forget-accuracy resistance still varies 3.6× with how the model was unlearned, tracking weight displacement (5.5×) not loss-depth (forget-CE equal, 1.08×): matched forget-depth is insufficient; weight displacement is a candidate residual axis (n=3, confounded with training length; §5.8, cf. Siddiqui et al. (2026)). No cross-method claim (FT unreachable, SCRUB n=1).

- We characterize the resistance-utility frontier (§5.3): at matched depth, methods also differ in how much retain-set utility they sacrifice (architecture-dependent in magnitude), so the informative comparison is resistance against utility at fixed depth, not a resistance ranking alone.
- We characterize method-specific depth domains: some losses cannot reach deep suppression at all (SimNPO at high β, RMU), a previously-unquantified restriction with mechanistic explanations (§5.4).
- Throughout, endpoints and kill criteria were registered before data (§4.4); we report negative and softening results (including a failed in-sample hypothesis that an earlier version of this project pursued) rather than only confirmations.

We position this as an evaluation-methodology correction, comparable in spirit to work exposing benchmark or metric flaws. We do not propose a new unlearning method, and we are explicit (§7) about what the results do not show.

2. Related work and positioning

Relearning attacks as evaluation. Hu et al. (2025) show approximate unlearning is reversible by benign relearning, establishing relearning recovery as a probe of residual knowledge, and their checkpoint sweep (attack success falling at later unlearning checkpoints; their Fig. 3) is depth-response in proxy form: unlearning progress is varied, but depth is neither measured nor matched. Fan et al. (2025) use relearning resistance to motivate sharpness-aware unlearning, reporting resistance as a function of relearning budget (epochs and sample count). A growing defense line is evaluated on the same axis: Lee et al. (2026) show a few finetuning steps revert even an idealized unlearning target and robustify via distillation; Qian et al. (2025) stage unlearning across folds to blunt adversarial relearning; Singh et al. (2025) report relearning resilience for a Jensen–Shannon-divergence (JSD) objective under a stricter judge-based evaluation. All of these works measure resistance against the attack or compare methods' robustness; none measures or matches the starting suppression depth the compared methods reach.

Depth/strength as an implicit factor. The closest work is the relearning-convergence-delay (RCD) line (Tran & Woo, 2026), whose Theorem 4 bounds the metric by (forget-set Hessian condition number at the unlearned weights) × (the unlearned model's forget-loss gap to the forget-optimal weight, whose loss is constant); suppression depth thus appears multiplicatively in their own theory, and their own text concedes the coupling: "RCD reflects both the model's current performance and the optimization difficulty on the forgotten dataset." Their empirical comparisons inherit it: the compared methods sit at forget accuracies spanning 0.470–0.642 (their Table 1, 30% random forgetting; 0.429–0.632 at 50%), and their Table 3 ranks $\mathrm{RCD}_{\mathrm{Adam}}$ across eight diffusion variants whose forget-side performance (nudity score) spans 0.224–0.500, unmatched in both cases (an adverse proxy check on Table 3 is reported in §5.7). Their own relationship analysis (their Fig. 1) relates RCD to utility (gap-to-retrain), not to the forget-side performance their theorem couples it to, and joining their published per-method RCD values to their published forget accuracies shows the coupling directly: $\rho = -0.93$, with starting forget accuracy alone explaining $R^2 = 0.81$ of the published RCD variance (§5.7). Their experiments are in image-classification and diffusion settings, not LLMs, a domain difference we make explicit since the rest of our evidence is LLM-side. Evaluation frameworks for unlearning (e.g. Wang et al. (2025)) explicitly note that methods unlearn to different strengths and introduce calibration, but calibrate the retain side, leaving forget-side depth as the free, compared quantity, the complementary normalization to ours.

In vision example-level unlearning, Siddiqui et al. (2026) find relearning resistance is predictable from weight-space displacement (L2 and mode connectivity to the original model), a correlate of depth rather than a measured or controlled forget-side quantity; their cross-method comparisons start from a roughly common

~50% post-unlearning forget accuracy by construction, without matching being named as a control. Their headline is, on its face, in tension with ours (large cross-method resistance differences at a common forget accuracy), but the reconciliation has two parts: matched accuracy can conceal loss-depth differences (our saturation point), and where it does not, weight displacement separates resistance even at matched depth. Our §5.8 re-score takes a first step on their open question: two SALUN constructions at matched forget-accuracy and equal forget loss (1.08×) still differed 5.5× in displacement and 3.6× in resistance, so their displacement predictor appears to survive loss-depth control (n=3, vision; their attack also relearns on the retain set, a different probe). Their result is a raw displacement–resistance correlation; our depth-controlled ladder tests it more sharply. At matched depth in LLMs, displacement still separates resistance after both depth and method identity are accounted for (§5.9; within-cell partial correlation r=0.58–0.82 at ≤1.4B, holding method and depth fixed; the method-aligned form weakens at 3B, §7), so the effect is not vision-specific and not the method effect relabeled. We do not claim the raw correlation is non-obvious – it is observational and confounded with optimization length – but the depth-controlled ladder adds the separation from method, and the benchmark-dependence, that a raw correlation cannot resolve.

"Over-unlearning" work (Ha et al., 2025) names excessive divergence as a failure mode, but defines it as retain-side collateral damage, not as a forget-side comparison confound. Reversibility studies (Xu et al., 2025) note the surface-metric/internal-state gap qualitatively, and metric-reliability work (Jia et al., 2025) shows standard unlearning metrics overestimate success under semantically adjacent probes (kin to our saturation argument, on the metric rather than the depth axis). To our knowledge, no prior work (a) names forget-side depth as the dominant confound in relearning-resistance comparisons, (b) compares methods at matched depth, or (c) measures relearning resistance as a function of controlled, measured starting depth and decomposes its variance. Hu et al. (2025)'s checkpoint sweep is the nearest response-curve antecedent (progress varied, depth unmeasured and unmatched), and Siddiqui et al. (2026)'s weight-space predictor the nearest depth-adjacent one (a displacement correlate, in vision, neither measured in forget-loss terms nor matched). (Based on a review of the 2025–2026 unlearning-evaluation literature, as of June 2026.)

3. Setup

Models and data. GPT-2-medium (Radford et al., 2019) (355M, float32), Phi-1.5 (Li et al., 2023) (1.4B, bf16; the canonical TOFU model (Maini et al., 2024)), and Llama-3.2-1B (Grattafiori et al., 2024) (1.2B, bf16; RMU steered at the MLP down-projection, layers 5–7). We fine-tune each on TOFU forget01 (40 QA) plus a 200-example retain slice to obtain a θ^* that has memorized the forget set (θ^* forget loss 0.24–0.70 nats; one θ^* per architecture, so "run-to-run" variance excludes θ^*-variance). All training uses 8-bit AdamW with gradient checkpointing; each model fits a single 16GB T4. A Llama-3.2-3B arm (3.2B, bf16) extends the decomposition to TOFU (n=3) and the MUSE news corpus (Shi et al., 2024) (n=5) under the identical θ^*/ladder/probe protocol on the loss-based core methods (RMU not run for this arm); its memorized θ^* reaches 0.17 nats forget loss on the corpus.

Audited subspace. Depth, displacement, and all measurements operate on ≥2-D weight tensors (embeddings, attention, MLP), excluding biases and layer-norm gains, a fixed filter (146 tensors / 1.4B params for Phi; 98 / 354M for GPT-2).

Methods (five: four loss-based, plus RMU). GA (Jang et al., 2023) (gradient ascent on forget), GD (Liu et al., 2022) (gradient difference), NPO (Zhang et al., 2024) (β=0.1, frozen-θ^* reference), SimNPO (Fan et al., 2024b) (reference-free; β swept), RMU (Li et al., 2024) (representation misdirection: steer layer-{5,6,7} MLP down-projection activations toward a fixed control vector, with a retain anchor). These span the major approximate-unlearning families.

Forget metric in nats. We report forget suppression as the forget-set cross-entropy in nats, not the saturating exp(−loss) metric, precisely because the latter hides the depth differences this paper is about. The exp(−loss) metric (0–1) is used for the relearning-recovery and deletion thresholds (§4.2, §5.6), where forgetting is judged on its natural 0–1 scale; we keep the two metrics distinct.

4. Method: depth-response measurement

4.1 Depth ladder and landing

For each (method, seed) we run one unlearning trajectory from θ^*, snapshotting whenever it first enters each target depth bin {6, 10, 14, 18, 22} nats (±1.5). Forget-loss climbs explosively through this range (often >10 nats in one optimizer step), so we cut the learning rate near the blow-up and, for missed bins, use a two-sided lander (gentle ascent if below the window, gentle forget-descent if above). All depth checks use the evaluation convention, so the landed depth matches what the analysis sees.

4.2 Relearning probe

From each snapshot we fine-tune on the forget set and count steps until the exp(−loss) metric (0–1) recovers past a threshold (0.30 × its θ^* value), averaged over 3 probe seeds. The relearning learning rate is auto-calibrated once per architecture on a GD reference snapshot, to land that snapshot in a 15–90-step window (other snapshots may fall outside it); the calibration selected 3e-5 on Phi-1.5 and 1.5e-5 on GPT-2 and Llama, so cross-architecture share comparisons in §5.2 span two probe LRs, a disclosed calibration difference. §5.5 reports the full LR×seed-count grid (method share ≤4% in all four cells; the n=5 decompositions nearly coincide across LRs). The probe also trains only the audited ≥2-D tensors (the same subspace all measurements use), a convention we disclose in §3; no published attack restricts its parameter set this way. §7 reports a verbatim full-parameter control re-measurement (Phi-1.5 and Llama-3.2-1B) that reproduces the decomposition, so this convention is not load-bearing.

4.3 Variance decomposition

On the balanced core: GD, NPO, and SimNPO at 5 seeds × 5 bins per architecture, where GA and RMU are excluded because their depth domains (§5.4) do not span all five bins, we decompose log(steps) and logit(retain) by ordinary least squares into depth(bin), method, depth×method, and residual(seed) sums of squares, and complement the shares with pairwise stratified within-bin permutation tests so that small method effects are tested, not merely bounded by a share. We report all four terms for the primary TOFU decomposition (§5.2, table); for the remaining runs (MUSE, the Llama-3.2-3B replication, and the full-parameter controls) the depth×method interaction never exceeds 1.8% of variance in any run, and we fold it into the residual term, reporting depth/method/residual. Our pre-registered pairwise statistic is a within-bin label-permutation rank test (20k draws, two-sided); we additionally report a mean-log-difference permutation test (4×10^4 draws), the statistic registered for the §5.5 follow-up, applied uniformly here as a post-hoc harmonization. Both are given in §5.2 (Bonferroni 0.05/3 per architecture); they are disclosed as two separately-registered tests, not one.

4.4 Registered discipline

Endpoints, pass thresholds, and kill criteria were fixed in a written brief before each run; we report against them verbatim, including failed endpoints (§5.6's first protocol failed its registered criterion and is reported as such), with a VOID-and-rebuild rule (no verdict on out-of-grid models). Analyses that follow a failed or revised registered design are explicitly labeled exploratory where they appear. An earlier hypothesis from this project, that a representation-geometry statistic predicts resistance at matched depth, was killed by its own registered out-of-sample test (Spearman +0.09, n=6); we mention this to characterize the evidentiary standard, not because it bears on the present claims. One registration deviation we flag explicitly: the §5.2 pairwise table was harmonized post-hoc to the mean-log statistic registered for §5.5; the originally-registered rank test is reported alongside it in §5.2, and the two agree on sign and on Phi's null, differing only on whether GPT-2's GD effect clears Bonferroni.

Table 1: Variance decomposition of log relearning-steps on the balanced TOFU core: depth, method, depth×method, and residual (run-to-run) shares, per architecture ($n = 5$).

component	GPT-2-medium (n=5)	Phi-1.5 (n=5)	Llama-3.2-1B (n=5)
depth (bin)	67.5%	64.8%	62.7%
method	5.7%	0.5%	8.1%
depth × method	0.4%	0.5%	1.4%
residual (trajectory seed)	26.4%	34.2%	27.8%

5. Results

5.1 Depth-response monotonicity is near-universal (49 of 50 curves)

Nearly every depth-response curve we measure is monotone increasing in depth ("monotone" operationalized as per-curve Spearman(nats, steps) $\geq +0.7$ throughout), across all three architectures. On GPT-2-medium all 18 full curves have Spearman $\geq +0.97$; on Phi-1.5, 15/16 core curves have Spearman $\geq +0.97$ (one exception at $+0.667$, a top-bin plateau); on Llama-3.2-1B, all 16 core curves have Spearman $\geq +0.89$ (zero premise alarms). We report this as a premise check rather than a discovery: with a fixed absolute recovery threshold, deeper starts have mechanically farther to travel, so monotonicity is partly expected by construction; what the census establishes is that no method or architecture escapes it. Deeper suppression reliably means slower relearning, across model families and loss families.

5.2 Variance decomposition across two benchmarks

Table 1 decomposes log(relearning-steps) on the balanced core (TOFU).

The ordering depth > run-noise $\gg$ method holds, on TOFU, across three model families: depth clusters at 63–68%, run-to-run noise at 26–34%, and method is always single-digit (0.5–8.1%) and always below run-noise. The inequality replicates on a fourth, larger model: Llama-3.2-3B on TOFU (n=3) gives depth 72.4% / method 7.4% / seed 20.2%, depth again dominant and run-noise again exceeding method, so the depth-dominance spans 355M–3B. A verbatim full-parameter re-measurement of the Phi-1.5 snapshots (control for the audited-tensor measurement convention, §7) reproduces this decomposition, depth 65.5% / method 0.5% / seed 34.0%, so the ordering is not an artifact of how depth is measured. The same control on Llama-3.2-1B, the architecture that does carry a method effect, reproduces its decomposition too (depth 62.9% / method 7.9% / seed 29.1% full-parameter vs 62.7 / 8.1 / 29.2 audited-transplant, the latter's depth×method folded into residual), so the convention preserves not just the depth-dominance but the method effect itself (7.9% vs 8.1%).

The reslice-robust statement is the inequality: run-to-run noise exceeds the method share on all three TOFU architectures, and does so in every depth-window reslice we tried (9/9: three architectures × {all 5 bins, deepest 4, deepest 3}). The full ordering depth > noise $\gg$ method is a property of the 6–22-nat ladder, not a constant: restricted to the deepest three bins, depth's share falls to 12–19% and run-to-run noise dominates (71–78%; on GPT-2, method 17.4% there exceeds depth 11.9%). Note, however, that the default configurations of §5.6 land at 31–166 nats, where depth's leverage is larger than our ladder exhibits; the full-ladder framing is, if anything, conservative for practice. On GPT-2 we verified via ANCOVA that controlling exact within-bin depth changes the seed share by only ~1% (the residual is genuine run noise, not unmodeled depth). The residual is also not an artifact of our relearning probe's resolution: across-trajectory (seed) variance exceeds within-cell probe-seed variance by $\approx 8\times$ (Llama), $21\times$ (Phi-1.5), and $57\times$ (GPT-2), $\geq 24\times$ once each cell's three-probe-seed averaging is accounted for, so the variance that beats the method share is run-to-run stochasticity, not probe-measurement noise. We scope this residual as trajectory + depth-landing-harness variance (the LR-cut and two-sided lander; the two cannot be fully separated) and state the inequality run-noise > method for that combined residual, not for a harness-free trajectory component. We also caution against reading precise shares finely: the Phi method share fell from 3.2% (n=3) to 0.5% (n=5)

Table 2: Pre-registered within-bin label-permutation rank test: two-sided p-values per method pair (* clears Bonferroni 0.05/3).

rank test	GD vs NPO	GD vs SimNPO	NPO vs SimNPO
GPT-2-medium	0.094	0.029	0.50
Phi-1.5	0.49	0.53	0.55
Llama-3.2-1B	0.0004*	0.082	0.022

Table 3: Post-hoc mean-log-difference permutation test: percent change in relearning steps and p-value per method pair (* clears Bonferroni 0.05/3).

% change in steps (A vs B)	GD vs NPO	GD vs SimNPO	NPO vs SimNPO
GPT-2-medium	−18%, p=0.006*	−19%, p=0.005*	−1%, p=0.85
Phi-1.5	−5%, p=0.35	−0.5%, p=0.93	+5%, p=0.43
Llama-3.2-1B	−25%, p<0.001*	−15%, p=0.026	+13%, p=0.024

as the seed grid better sampled the trajectory distribution, itself a demonstration that low-n robustness comparisons are under-powered.

The corpus raises the method share. On the MUSE news corpus the decomposition changes sharply: on Phi-1.5 the shares are depth 55.7% / method 28.3% / seed 16.0%, and on Llama-3.2-1B they are depth 29.0% / method 37.1% / seed 33.9%. The method share, single-digit on TOFU, rises to 28–37% on these two ≤1.4B corpus models and exceeds the run-to-run (seed) term (on Llama it also exceeds depth). At Llama-3.2-3B on the corpus (n=5) the method share rises similarly – depth 22.7% / method 32.1% / seed 45.2% – but here run noise stays the largest term, so method does not overtake it; the resistance/utility dissociation nonetheless persists (method 32% of resistance vs 72% of utility variance). The run-noise > method inequality therefore holds on TOFU at every scale and on the corpus at 3B, and is overtaken only on the two ≤1.4B corpus models: the corpus consistently raises the method share, but whether it lifts method past run noise is scale-dependent. We report this benchmark dependence as a finding. §5.9 offers a candidate explanation for the rise – weight displacement, the L2 distance a method moves the weights, stays bunched across methods on TOFU and fans out on the corpus – but the protocol conclusion (compare at matched depth, report run variance) does not depend on it.

A small method effect on resistance is detectable at matched depth, and we report it rather than rounding it to zero. We give two pairwise tests, registered at different times and disclosed as such (§4.3–4.4): the pre-registered statistic is a stratified within-bin label-permutation rank test (20k draws, two-sided); the post-hoc statistic, difference in mean log-steps (4×10^4-draw permutation), the test registered for the §5.5 follow-up, is applied uniformly here as a harmonization. Both: 3 method pairs × 3 architectures, Bonferroni $0.05/3 = 0.0167$.

The pre-registered rank test gives the p-values in Table 2.

Table 3 gives the post-hoc mean-log statistic (% change in steps, p).

(* clears Bonferroni 0.0167.) The two statistics agree on Phi's null and differ only on GPT-2: under the pre-registered rank test the GD-faster effect clears Bonferroni only on Llama (GD vs NPO p=0.0004) and is merely suggestive on GPT-2 (GD vs SimNPO p=0.029 uncorrected, GD vs NPO p=0.094); the post-hoc mean-log statistic additionally clears Bonferroni on both GPT-2 pairs. We therefore claim, conservatively, a sign-consistent GD-faster direction on all three architectures, significant on Llama under the pre-registered test; the "GPT-2 also significant" reading holds only under the post-hoc statistic. On GPT-2 the effect still survives exact-depth ANCOVA control (+27% steps for SimNPO over GD, p=0.003). Phi is null throughout, at its calibrated probe and, in a pre-registered follow-up at the gentler probe with full seeds (§5.5: GD vs NPO −8%, p=0.15; GD vs SimNPO −2%, p=0.77), so probe compression does not explain the Phi null;

Table 4: Method share of variance for relearning resistance vs. utility (logit-retain) at matched depth, with their ratio, per architecture.

	method share of resistance var	**method share of utility var**	**ratio**
GPT-2-medium	5.7%	56.7%	~10×
Phi-1.5	0.5%	56.5%	~110×
Llama-3.2-1B	8.1%	19.9%	~2.5×

Table 5: Phi-1.5 at the deepest bin (≈20.6 nats, $n = 5$): relearning steps overlap across methods while retain (utility) separates.

method (Phi-1.5, ~20.6 nats, n=5)	**steps (range)**	**retain (range)**
GD	21.7 (18–31)	0.12 (0.04–0.24)
NPO	22.4 (18–28)	0.37 (0.23–0.49)
SimNPO (β=0.25)	20.5 (14–28)	0.70 (0.61–0.82)

the cross-architecture difference is reported as unexplained. The NPO-vs-SimNPO contrast reaches nominal significance once (Llama, +13%, p=0.024) with no support on the other two architectures, so we do not claim it. The correct reading is therefore not "method does not matter" but: method effects on relearning resistance are real but small, a 15–25% contrast where detectable, versus a 1.9–3.0× swing in steps from depth across the ladder and a median 1.7× run-to-run spread within a single method–depth cell. The method effect is roughly a third of typical run-to-run spread, which is precisely why single-run, depth-uncontrolled comparisons cannot resolve it.

5.3 Methods also separate on utility at matched depth

This section concerns the TOFU regime, where resistance is near-invariant across methods at matched depth (§5.2); on the corpus, resistance instead separates by method via weight displacement (§5.9), so the picture below is TOFU-specific. On TOFU the resistance near-invariance holds: method explains <9% of relearning-resistance variance on all three architectures, with only a small pairwise effect detectable at matched depth (GD ~15–25% faster than NPO/SimNPO where significant; §5.2). The complementary observation is that methods do differ on utility at matched depth. Decomposing utility (logit retain) beside resistance, the method share is far larger for utility than for resistance on every architecture, but its magnitude varies (the GPT-2 share is computed on the same merged data as its resistance share in §5.2); the resistance-vs-utility shares are in Table 4.

The direction is consistent (method matters more for utility than resistance everywhere), but the magnitude is not robust: on Llama, utility variance is dominated by run-to-run noise (≈63%), not method (19.9%), so the contrast is only ~2.5×. We therefore do not headline the dissociation as a precise effect; we report (i) the robust claim that method is a sub-10% contributor to resistance on three architectures, and (ii) the weaker, architecture-dependent observation that methods separate on the resistance–utility tradeoff at matched depth. The matched-depth retain ordering (GA < GD < NPO < SimNPO) recurs on all three models, but the separation is clean on GPT-2/Phi and muddier on Llama (GD ≈ NPO).

Illustratively, on Phi-1.5 at the deepest bin (~20.6 nats) relearning steps overlap across methods while retain means separate cleanly (Table 5).

At n=5 the resistance columns are essentially indistinguishable (means 20.5–22.4, fully overlapping ranges) while retain spans 0.12–0.70; the dissociation is visually cleaner than at n=3.

The methods are nearly interchangeable on the axis robustness rankings use (resistance) and sharply ordered on the axis those rankings ignore (utility at matched depth). Honest caveat (also §7): the utility-variance

half is partly by construction: these losses carry different retain regularizers (GA has none), so that they differ on retain is unsurprising. The non-obvious half is the TOFU resistance-invariance: despite different objectives and a ≈6× utility spread, on TOFU methods relearn at near-identical rates at matched depth, up to the small GD-faster effect of §5.2. On the corpus they do not (§5.9); the general contribution is that robustness rankings conflate depth (and, secondarily, weight displacement), not that methods are universally interchangeable on resistance. A held-out-retain control (Phi-1.5, TOFU retain examples 200–400) confirms the dissociation is not an in-sample artifact. Decomposing the full-model retain loss (a different measure from the table's audited logit-retain, hence a method share of 49.9% in-sample here vs the 56.5% above), we find method explains 28% of held-out retain-loss variance versus 0.5% of resistance, even though held-out retain runs ~4× worse in absolute nats (4.74 vs 1.15) than the in-sample axis.

5.4 Method-specific depth domains

Not every method can reach every depth. SimNPO's reference-free loss $-\frac{2}{\beta}\log\sigma(\beta\cdot\mathrm{CE})$ (with CE the forget cross-entropy) saturates, its gradient shrinks as $\beta\cdot\mathrm{CE}$ grows, and an intervention at two β values is consistent with this direction (single-run): at β=2.5 the trajectory stalls at ~4 nats (a 400-step plateau), while at β=0.25 the same loss climbs past 50 nats under a large budget. The direction (lower β, higher ceiling) is established by this intervention; we do not claim a quantitative ceiling law: our earlier $\approx 5/\beta$ heuristic fits neither the β=2.5 stall (~4 nats, not 2) nor the β=0.25 behavior, and we retract it. RMU exhibits a different ceiling: in the depth-ladder it plateaus at ~8.5–8.9 nats on all three architectures (with no deeper ladder rows reachable at our ladder budget), and a dedicated large-budget probe puts the hard ceiling at ~8.5–11 nats (GPT-2 ~10, Phi ~8.5, Llama ~11). Mechanistically, once forget activations match the control vector, its objective converges and the MSE gradient dies, regardless of further training. To our knowledge these are the first characterizations of method depth-domains, the RMU ceiling quantified and replicated on three architectures, the SimNPO restriction indicated directionally by a single-run intervention at one β pair and they directly imply that any default-config comparison spanning such methods is structurally depth-confounded: the methods are not even on the same depth ladder.

5.5 Robustness to probe sensitivity: the full 2×2 (probe LR × seed count)

We re-probe the Phi-1.5 snapshots at a gentler relearning LR (1.5e-5), widening the step dynamic range ~4× (26–110 vs 7–31 at the calibrated 3e-5), first at n=3 seeds, then, as a pre-registered completion, at the full n=5. The method share of resistance variance is small in every cell of the grid: 3.2% (3e-5, n=3), 0.5% (3e-5, n=5), 3.8% (1.5e-5, n=3), 1.2% (1.5e-5, n=5), and at n=5 the full decompositions at the two LRs nearly coincide (depth/method/interaction/seed = 64.8/0.5/0.5/34.2 vs 65.6/1.2/0.5/32.7). The small method share is not a probe artifact.

The pre-registered pairwise endpoint returned its null branch: at the gentler probe with full seeds, GD vs NPO is −8.1% (p=0.15) and GD vs SimNPO −1.6% (p=0.77), signs consistent with GPT-2/Llama, but not significant. Per the registration, we retract probe compression as the explanation for Phi's null and report the cross-architecture difference as unexplained. (The NPO-resistant ordering hinted at n=3 also dissolves at n=5: NPO vs SimNPO +7.0%, p=0.26.)

A cautionary illustration of our own thesis, from our own analysis: the n=3 subset of this same grid shows GD vs NPO at −14.1% (p=0.011), a Bonferroni-clearing "confirmation" of the compression explanation that evaporates when the remaining two seeds are added (−8.1%, p=0.15). Had we stopped at three seeds, this paper would have reported a causal confirmation that its own full grid refutes. The seed lottery §5.2 quantifies is not an abstraction; it bit this paper's analysis at n=3.

5.6 The practical cost: at the standard recipe the comparison degenerates; at a forced budget the ranking is depth

We attempted to rank four methods (GD, GA, NPO, SimNPO) under realistic configurations against their matched-depth resistance (the 14-nat bin). We report this section's full history, including a registered-endpoint failure and a post-hoc budget choice, because the history is itself evidence.

Attempt 1 (registered, threshold stopping): failed its endpoint. Our first protocol stopped each method at the first epoch where the exp(−loss) metric crossed the standard deletion threshold. On Llama-3.2-1B it failed its registered criterion (Spearman(reached-depth, resistance) = −0.40 against a pre-registered ≥ +0.8, with the default ordering persisting at matched depth, Kendall +1.0). The mechanism: the saturating metric crosses its threshold between epoch checks while the loss explodes, so where a run halts is granularity luck. Individual configurations landed across wide per-method ranges (GA 26.9–59.2 nats; NPO 42.2–58.0), spanning 25–68 nats over the four methods, with NPO's resistance alone ranging 37 to ≥200 steps (right-censored at the probe cap). Threshold stopping does not produce method-characteristic depths; it produces a depth lottery within each method.

Attempt 2 (registered standard recipe): the comparison degenerates. The script's registered default, the standard TOFU unlearning recipe, lr 1e-5 × 5 epochs, fails the sanity bar on Phi-1.5: three of four methods do not unlearn at all (GD/NPO/SimNPO end at 1.1–1.3 nats, exp(−loss) metric 0.27–0.34, relearned in ~1 step); only GA crosses the deletion threshold (2.5 nats). At the recipe practitioners actually publish with, a relearning-robustness comparison of these methods has nothing to measure, and the one method that does unlearn "wins" by reaching depth.

Attempt 3 (exploratory, post hoc): the ranking is reached depth. To obtain four comparable unlearned models we raised the budget 8-fold (lr 2e-5 × 20 epochs), chosen after observing the standard-recipe failure and not registered; every number from it is labeled exploratory. Under this one fixed budget the methods land at wildly different, method-characteristic depths on every architecture (GPT-2 41–166 nats, Phi 31–110, Llama 46–109; GA always deepest, SimNPO always shallowest, as §5.4's saturation mechanism predicts), and default-config resistance tracks reached depth at Spearman = +1.0 on all three architectures. The 4.5–5.9× default-config resistance spreads collapse to 1.3–1.7× at the matched reference depth (per-method means; GPT-2: 28–164 steps becomes 14.0–24.2; Phi: 23–103 becomes 17–23; Llama: 28–162 becomes 18–24). Qualifications: (i) several default values are right-censored at the 200-step probe cap (the deepest GA seed on GPT-2; the deepest GD and GA seeds on Llama), floors, not measurements; direction unaffected; (ii) matching is at the 14-nat bin, far shallower than where these defaults land, and uses ladder-recipe trajectories; the collapse statement is regime- and recipe-bound; (iii) the registered spread criterion (matched spread ≤ 2× probe-sd) failed on GPT-2 (2.7× with full-seed matched values) and Phi (2.7×), passing only via its rank-order arm; matched spreads are small but not uniformly within two probe-sd, consistent with the small real GD effect of §5.2.

Matched orderings are mostly noise, except where they recover §5.2's effect, and we do not headline them. Kendall τ between default and matched orderings is −1.0 (GPT-2), +0.67 (Phi), −0.33 (Llama). On GPT-2 (full-seed matched values: GA 14.0±2.2 n=3, GD 19.9±5.4, NPO 23.2±3.7, SimNPO 24.2±3.6) the GD-below-NPO/SimNPO separation matches the §5.2 GD-faster pairwise effect (which on GPT-2 clears Bonferroni only under the post-hoc statistic; see §5.2), a real signal, of exactly the size single runs cannot see, and on Llama the GD-vs-NPO/SimNPO gap (18.4 vs 22.4–23.6) likewise matches its §5.2 effect. The remaining adjacent gaps (NPO-vs-SimNPO everywhere, ≤1.2 steps; Phi's non-GA gaps ≤0.4) are within noise, and GA's matched values rest on three seeds (GPT-2) or one (Phi/Llama). The robust content of this section is the standard-recipe degeneracy, the ρ=+1.0 depth-tracking, and the spread collapse.

5.7 A published comparison, re-read (no reproduction required)

The artifacts of the RCD paper (Tran & Woo, 2026) themselves contain the confound demonstration, and, as we detail below, the relation is partly definitional, so we read it as a consistency illustration, not independent evidence. Their reliability figure (Fig. 5) reports eight per-method RCD values, all eight typeset in the paper's text layer ("RCD = . . . "), verbatim and verified in their arXiv PDF, not digitized (the figure itself does not restate its dataset/scenario). We join them by method name to their Table 1 (TinyImageNet) forget accuracies, and the conclusion is insensitive to which random-forgetting column the figure reflects: the 10× spread in published RCD (4.3 to 44.9) tracks starting forget accuracy at Spearman = −0.93 (OLS R^2 = 0.81) under the 30% column and −0.96 (OLS R^2 = 0.82) under the 50% column (n = 8 methods). Under any join consistent with the figure, starting forget accuracy alone explains roughly four fifths of the published RCD variance. The shallowest methods (SCRUB, forget accuracy 0.642; FT, 0.609) are precisely the ones the paper reports as "vulnerable to relearning"; the deepest (Retraining, 0.470; SALUN, 0.507) are

its "resistant" ones. The relation is not an artifact of including the authors' own method (IEU): restricting to the five baselines (Retraining, FT, RL, SCRUB (Kurmanji et al., 2023), SALUN (Fan et al., 2024a)) the dependence on starting accuracy is, if anything, stronger (OLS $R^2 = 0.91$), though with $n = 5$ the rank test is underpowered (exact $p = 0.083$). The $n = 8$ significance, in turn, leans on near-duplicate points (the three IEU variants cluster at forget accuracy 0.508–0.547, two with identical RCD = 31.05), so its $p \leq 0.002$ overstates the independent evidence. Part of this coupling is definitional: their RCD integrand at relearning epoch 0 equals the model's starting forget error, so the metric is seeded by depth before any relearning dynamics enter, consistent with their own Theorem 4 (Tran & Woo, 2026) and their own caveat that RCD "reflects both the model's current performance and the optimization difficulty." We therefore read the OLS R^2 (effect size), not the p-value, as the more reliable statistic, but do not rest the argument on its magnitude: that coupling is partly definitional, and the definitional seeding is itself the confound (a method's RCD is set by where it landed in depth, not by the method). The cleaner demonstrations are method-controlled, a single method's RCD moving with its realized depth across panels (SALUN below), and the registered reproduction (§5.8); we treat this re-read as a consistency illustration.

Repeating the join in their second classification setting sharpens the result (two re-reads of one paper, not independent replications). Reading the per-method RCD markers from their CIFAR-10 figure (Fig. 16(a); each marker identified by its x-position = that method's Table 10 Avg. Gap, so identity is unambiguous, and RCD read off the y-axis to ± 0.15, verified against the figure rendered from their arXiv PDF) and joining to Table 10's forget accuracies yields Spearman $= -0.91$ ($n = 8$), OLS $R^2 = 0.89$, and on the five baselines alone, Spearman $= -1.00$ (exact $p = 0.017$), with starting forget accuracy explaining OLS $R^2 \approx 0.99$ of the published RCD variance. Where TinyImageNet's baselines-only subset was underpowered ($p = 0.083$), CIFAR-10's is significant: in this setting the published baseline ranking is essentially the depth ordering (near-perfect, though from ± 0.15-tolerance digitization at $n = 5$). Notably, Retraining, "the ideal unlearning outcome" in their reading, sits mid-pack (5th of 8) in this figure, exactly where its mid-pack forget accuracy predicts. We checked the figure's other three CIFAR panels so this is not a single-panel artifact: the Random-50% panel again gives a baseline coupling of Spearman $= -1.00$ ($n = 5$), and SALUN itself moves from the highest RCD in panel (a), where it forgot deepest (forget accuracy 0.874), to near the lowest in panel (b) (RCD $10.65 \rightarrow 3.1$) where it forgot only to 0.952: the same method's apparent resistance tracks its realized depth, not its identity. (The two class-wise panels would need their own forget-accuracy table to score, so we note only their qualitative consistency, SALUN lowest, Retraining/IEU highest.)

An adverse check on the same paper's other comparison, reported verbatim: their Table 3 ranks $\mathrm{RCD}_{\mathrm{Adam}}$ across eight diffusion variants whose forget-side proxy (nudity detector score) spans 0.224–0.500; there, Spearman(nudity, $\mathrm{RCD}_{\mathrm{Adam}}$) $= +0.67$ (exact $p = 0.083$), the opposite sign. Three differences make this uninformative about depth rather than contradictory: the proxy is a detector ratio over generated images, not a loss-based depth; the optimizer is Adam, where their own limitations section reports the GD-motivated framework behaves differently and method vulnerabilities are "comparable"; and the total $\mathrm{RCD}_{\mathrm{Adam}}$ spread there is 6% of its mean ($\approx 1.1\times$), versus $\approx 10\times$ in Table 1. The honest summary is conditional: the depth coupling is demonstrable where a depth-like quantity is published, and indeterminate where only a weak proxy exists.

Scope, stated plainly: this is an observational re-read of published values ($n = 8$), not an intervention. Whether method differences persist at matched depth in their setting is the question of a registered reproduction design (released with this paper; executed in §5.8 below, a partial reproduction: the confound direction reproduces, their published ranking does not, and the matched-depth re-score did not clear its registered gate). The domain is vision classification: this demonstrates the confound in a published table, within its own domain, corroborating, not extending, our LLM results.

5.8 The reproduction, executed: a partial reproduction and a selection-sensitivity finding

We ran the reproduction the re-read motivates (registered design and amendments released with this paper): CIFAR-10/ResNet-50 (Krizhevsky, 2009; He et al., 2016), 30% random forgetting, the five methods at n=3 seeds (Retraining at n=1). We targeted three things: their Table 10 forget accuracies, the depth–RCD direction, and their Figure 16 per-method RCD ranking, under a staged reproduce-or-abort gate, and only

Table 6: Reproduced RCD by checkpoint-selection rule, with the published Fig. 16 values; a method's RCD shifts up to ≈8.6× across rules, reordering the table.

RCD by selection rule	FT	SCRUB	Retraining	RL	SALUN
best-ckpt, 3-metric	2.24	1.44	5.58	5.36	4.23
best-ckpt, 4-metric (+MIA)	2.24	1.43	5.58	4.84	4.01
final epoch	4.76	12.43	5.58	11.54	10.42
their published (Fig. 16)	0.31	1.49	6.51	8.42	10.65

on a pass, a matched-depth re-score. The first two reproduced; the third, their headline ranking, did not, under any principled rule, and the manner of failure is the finding.

What reproduces. At the registered best-checkpoint selection, all five forget accuracies fall within the ±0.06 tolerance of their Table 10 (FT 0.959 vs 0.981; SCRUB 0.974 vs 0.975; Retraining 0.915 vs 0.919; RL 0.914–0.923 vs 0.904; SALUN 0.911–0.918 vs 0.874), and SCRUB is a near-exact positive control on both axes (RCD 1.43 vs the extracted 1.49) and the harness reproduces a method when it lands at its published depth. The confound direction reproduces too: across the five methods, Spearman(forget-accuracy, RCD) = −0.60 to −0.90 (deeper suppression → larger delay), and the methods their paper calls "vulnerable" (FT, SCRUB) carry the lowest RCD in our hands too, at the best-checkpoint rules.

What does not. SALUN, the most resistant method in their Figure 16 (RCD ≈ 10.65), lands mid-pack in every reproduction (RCD ≈ 4.0–4.2 at the principled rules), and no uniform checkpoint rule recovers their published per-method order (rank correlation to their values is $\rho = 0.50$ at the two best-checkpoint rules and 0.30 at the final-epoch rule, all below our pre-registered 0.7 bar). Per the registered gate this aborts the matched-depth re-score; we therefore make no causal "their ranking dissolves at matched depth" claim.

Why it fails is itself a result about the metric. Their protocol scores each method at a "best checkpoint" (their Figure 8) by a criterion the paper does not fully specify. Scoring the same training runs at three defensible uniform rules, minimum average-gap-to-retrain over their three reported accuracies (the best-ckpt, 3-metric rule), the same over four metrics including a membership-inference (MIA) score (the best-ckpt, 4-metric rule), and the final epoch which moves a single method's RCD by up to ~8.6× and reorders the table (Table 6). Their Tables 1/10 define Avg. Gap as exactly the mean of these four per-metric gaps to retraining, which we verified arithmetically, so the 4-metric rule is the faithful reading.

SALUN's published value (10.65) is matched only at the final-epoch checkpoint, where it has over-forgotten to 0.788 accuracy, but that same rule sends SCRUB, reported as one of the least-resistant methods, to the highest RCD in the table (12.43, over-forgotten to 0.841). No single rule places every method where their figure does: the published ranking requires per-method checkpoint choices. Since RCD at a fixed budget is monotone in suppression depth (Theorem 4 of Tran & Woo, 2026; our §5.1), this is the confound made mechanical, a method's reported "resistance" is governed by where its checkpoint sits on the depth axis, a choice their protocol leaves unfixed, at least as much as by the method.

The registered matched-depth re-score re-unlearns each method to the retrain-level band (forget accuracy = 0.92 ± 0.015) and recomputes RCD; it was contingent on the reproduce-or-abort gate passing, which it did not, so running it is an explicit deviation from the registered gate, flagged here as such (cf. the §5.2 statistic harmonization), and it probes our pipeline, not their ranking. Two things result. (i) Within our SALUN reproductions RCD tracks realized forget depth: 10.4 at an over-forgotten final epoch (forget accuracy 0.79), 4.0 at the best-checkpoint (0.918), 1.1 at the re-score band (0.919). (ii) But the last two sit at the same forget accuracy and differ 3.6× in RCD: the best-checkpoint is a min-Avg-Gap point from a 100-epoch run (retain/test preserved at ~0.99/0.91) while the re-scored model reaches the band in under one epoch (retain/test ~0.92/0.86). Matching a saturating forget-accuracy does not pin resistance, and we measured why. Re-measuring both candidate axes on these models, we find that at matched forget-accuracy the best-checkpoint and re-scored constructions sit at essentially equal forget cross-entropy (0.52 vs 0.48 nats, 1.08×) but differ 5.5× in weight-space displacement from base (13.8 vs 2.5), with RCD differing 3.6×.

For this construction pair the displacement gap (5.5×) far exceeds the forget-CE gap (1.08×), so loss-depth alone does not account for the 3.6× RCD difference; but all three quantities co-vary across the pair (the deeper-CE construction is also the higher-displacement and higher-RCD one), so this single pair cannot separate residual loss-depth from displacement. We read the displacement association as suggestive only, the direction Siddiqui et al. (2026) report in the RCD setting, not an isolated effect. The takeaway sharpens: matched forget-accuracy is demonstrably insufficient for comparing resistance, and weight displacement is a candidate residual axis (confounded here with training length). (The SALUN reproduction is n = 3 seeds; this displacement comparison is observational, displacement is measured, not intervened, and the two constructions also differ in training length, so consistent-with, not proof-of, a displacement mechanism.) We draw no cross-method conclusion: FT never reached the band (plateaus shallow) and SCRUB held the band on one of three seeds, so the matched-accuracy ordering (SALUN 1.1, RL 1.5, SCRUB 6.2) is underpowered, and SCRUB's high value is itself either a residual method effect or the same displacement/construction effect, unresolved at n = 1.

Scope and conduct. This is a partial reproduction: forget accuracies and the depth–RCD direction reproduce; the published ranking does not, under any uniform rule; the matched-depth re-score was gated off, and the re-score we report (above) is a flagged registration deviation, run on our own reproductions and claiming nothing about their models. The best-checkpoint-selection and add-MIA amendments were each registered before their runs; the add-MIA amendment's hypothesis, that omitting MIA had biased our SALUN checkpoint to be shallow, was refuted (the MIA-included rule reproduces the no-MIA rule almost exactly) and is reported as such, not dropped. No rule was adopted because it passed, and the registered gate failure and the refuted amendment both stand in the record. As with the re-read, the claim is bounded to a published table in its own (vision) domain, corroborating, not extending, our LLM results.

5.9 The method residual tracks weight displacement (observational)

§5.2's decomposition leaves a residual beyond depth: a method term negligible on TOFU (under 9%) but large on the MUSE news corpus (28.3% on Phi-1.5, 37.1% on Llama-3.2-1B). This section asks what the method component of that residual is. The natural candidate is weight displacement, the L2 distance a method moves the weights from the memorized model θ^* over the audited ≥2-D-tensor subspace. The first concern to rule out is that this is the method effect relabeled: within a benchmark, L2 is itself 28–85% method-determined, so "resistance tracks displacement" could be "resistance tracks method" in continuous clothing. It is not. Within a benchmark, displacement adds R^2 0.10–0.22 to a model that already contains depth and method dummies, and, holding method and depth fixed, the trajectory that displaced farther still relearns slower (within-cell partial correlation r=0.58–0.82 on the four ≤1.4B runs). So, on these runs, displacement carries trajectory-level resistance information the method label does not; this within-cell signal weakens to r=0.25 at 3B, so we scope the strong displacement form to the ≤1.4B models. We do not claim the displacement–resistance link is itself surprising or causal: it is observational, confounded with optimization length (a harder-optimized run both moves farther and resists more; §7), with RMU (representation-misdirection unlearning) a standing exception (large, unproductive displacement, fast relearn). The non-trivial content is narrower, and is what we read displacement as: a candidate single quantity, not the method label, whose benchmark-dependent spread (bunched on TOFU, fanned on the corpus) accounts for the benchmark-dependence of the method share. On TOFU the run-to-run component, which dominates the post-depth residual there, shows no such structure. Two facts then characterize the method component, across all four benchmark×model runs.

First, the four loss-based methods occupy a fixed displacement ordering, GA < GD < NPO < SimNPO, at nearly every depth (one shallow-bin NPO/SimNPO inversion on TOFU-Llama) and on both benchmarks. Displacement is largely method-characteristic: it rises only modestly along the depth ladder relative to how far it varies across methods, an axis largely separate from depth.

Second, the spread of that ordering tracks the method effect. The three core methods (GD, NPO, SimNPO; the retain-free GA baseline is excluded) are bunched on TOFU (core L2 spread ≈0.6) and fanned out on the corpus (≈2.3–3.0), mirroring the under-9% versus 28–37% method shares. At iso-depth, relearning time rises with displacement in every run (within-bin Spearman 0.60, 0.59, 0.84, 0.93 for TOFU-Phi, TOFU-Llama, MUSE-Phi, MUSE-Llama; weaker on TOFU, where displacement varies least, but positive throughout). Within each benchmark a depth-plus-displacement model explains 71–88% of log relearning-time variance,

of which depth alone accounts for 51–60% on three of the four runs (27% on the method-dominated MUSE-Llama); displacement adds, over depth alone, ≈0.12–0.14 R^2 on TOFU and ≈0.37–0.61 on the corpus, in proportion to how far methods spread in displacement. Two cautions on reading this. First, within a benchmark depth is the larger standardized predictor on TOFU (standardized β 0.51 vs 0.43 for displacement on Phi; 0.63 vs 0.40 on Llama); displacement exceeds depth only on the corpus. Second, pooling the two benchmarks raises the joint R^2 to ≈0.88 and makes displacement the larger pooled coefficient (0.86 vs 0.24; 0.84 vs 0.37), but this conflates within- and between-benchmark variance: the corpus has both higher displacement and longer relearning, so in the pool displacement absorbs the benchmark gap and depth's apparent share falls to 15–19%.

The TOFU near-invariance and the corpus method effect are therefore consistent with a single mechanism: relearning resistance tracks both how deep a method drives the forget loss and how far it moves the weights, and the benchmarks differ, in their method component, only in how far apart the methods land along the displacement axis. Once depth is matched, what a single-run ranking reads as method quality covaries with weight displacement.

Two caveats bound this. Displacement here is observational: it is measured, not intervened, and is confounded with optimization length, so we report a relationship consistent with a displacement mechanism, not a proof of one (cf. §5.8). We flag the symmetry plainly: optimization length is the same confound we argue depth-uncontrolled rankings ignore, and it confounds our own displacement axis, so we treat displacement as a measured correlate of resistance (extending Siddiqui et al. (2026) to LLMs), not a controlled or causal axis; showing it causally sets resistance would require intervening on displacement at matched depth, which we leave to future work. The claim here is the weaker, observational one: at iso-depth, resistance covaries with displacement, and that covariation accounts for the benchmark-dependent method share. And RMU is an exception: it incurs large but unproductive displacement (L2 ≈10–36, several-fold above the loss-based methods' ≈2–10) while relearning quickly and capping at shallow depth, so we exclude it from the displacement fit and report it as the boundary of the claim.

6. Discussion: what robustness evaluations should report

The practical recommendations follow directly; together they constitute what we call iso-depth evaluation:

1. Report forget suppression in nats, never only a saturating metric. A table of "forget ≈ 0" across methods hides the variable that drives the result.
2. Compare at iso-depth (matched suppression depth), or report depth-response curves. A single-point, default-config comparison conflates method, depth, and a large run-variance term. Curves (resistance vs depth) separate them and are cheap to produce. Matching forget-loss depth is necessary but not sufficient: at iso-depth, relearning resistance still tracks weight displacement (§5.9; predicted by Siddiqui et al. (2026) and corroborated in vision, §5.8), which is method-characteristic and, on a real corpus, a large residual axis. Report weight displacement (or mode connectivity) alongside depth; matching at displacement would require an intervention we do not perform (§7), so we leave it as future work.
3. Report run variance. Trajectory-seed variance is large, comparable to the method effect on a corpus and exceeding it on fictitious QA, so single-run robustness numbers are unreliable; error bars over unlearning seeds are not optional.
4. Evaluate the tradeoff, not one axis. Methods separate on utility-at-matched-depth more than on resistance (strongly on two of our three architectures), so the informative comparison is the resistance–utility frontier at fixed depth, not a resistance ranking.
5. Pin and report the checkpoint at which relearning is scored. A method's measured resistance depends on which point along its unlearning trajectory is attacked. §5.8 shows that scoring the same runs at three defensible "best-checkpoint" rules moves a method's RCD by up to ~8.6× and reorders the table; a metric that selects a "best checkpoint" by an unspecified criterion therefore carries a free depth parameter. State the rule, and compare resistance at a fixed forget-depth.

These do not require new infrastructure; they require controlling a variable and reporting a few more numbers.

7. Limitations and robustness checks

- The utility-variance half is partly definitional. Methods carry different retain regularizers, so a large method effect on retain is expected. We rely on the dissociation (resistance-invariance vs utility-variance), not on the utility variance alone, and we frame accordingly.

- Scale. GPT-2-medium, Phi-1.5, and Llama-3.2-1B (355M–1.4B), with a Llama-3.2-3B arm on both benchmarks: TOFU (n=3) reproduces depth-dominance and the run-noise>method inequality (depth 72.4% / method 7.4% / seed 20.2%), and the corpus (n=5) reproduces the method-share rise (depth 22.7% / method 32.1% / seed 45.2%, with run noise still the largest term). Phi-1.5 is TOFU's canonical model and Llama-3.2 a current production family, which addresses the "toy model" objection in its most important forms, but the results do not speak to 7B+ deployment.

- Two benchmarks, both somewhat artificial; one probe family. We test on TOFU (fictitious-QA forget01) and the MUSE news corpus, which span memorized-fact and real-text settings but are both small-scale memorize-then-unlearn tasks; GPT-2-medium could not memorize the corpus, so the MUSE arm uses the ≥1B models (Phi-1.5, Llama-3.2-1B, and a Llama-3.2-3B replication). We use one relearning-probe family (forget-set fine-tuning); other forget sets and benign-relearning probes (loosely-related data) are untested. The depth confound is mechanism-level (a deeper start sits farther from the fixed recovery threshold) and so is probe-agnostic by construction; whether the displacement covariation generalizes across probe families is, by contrast, untested, and is exactly the kind of question iso-depth evaluation is meant to surface.

- Displacement is observational; RMU is an exception. The depth-plus-displacement relationship (§5.9) is correlational: we measure weight displacement, we do not intervene on it, and it is confounded with optimization length, so we report it as consistent with a displacement mechanism, not as proof of one. We therefore present displacement as a secondary, observational refinement of the depth result, not a co-equal controlled axis; an intervention that varies displacement at matched depth to test causality is the natural next step. It also holds only for the loss-based methods: RMU incurs large but unproductive displacement (L2 several-fold above the others) while relearning quickly, and is excluded from the fit. We probed displacement at 3B on both benchmarks. On TOFU-3B (the bunched-displacement regime, n=3) the iso-depth correlation is essentially null (within-bin Spearman ≈0.1), unsurprising where displacement barely varies across methods. On the corpus at 3B (n=5), where displacement does fan out, the iso-depth rank correlation is positive (within-bin Spearman 0.67, positive in every bin; the method-controlled within-cell partial correlation, by contrast, falls to 0.25, the §5.9 statistic), so a positive rank signal persists at 3B. But the clean, method-aligned form does not: within the 3B model a pooled (TOFU+MUSE) depth-plus-displacement model explains only 24% of variance (versus 87–89% at ≤1.4B), with depth the larger standardized predictor (0.42 vs 0.19). The reason is structural: at 3B the method-characteristic content of displacement collapses – method identity explains only ≈10% of L2, versus 65–85% on the ≤1.4B corpus runs – so displacement is dominated by trajectory-to-trajectory noise rather than method, leaving the relation rank-monotonic but no longer the clean mechanism it is at smaller scale. We therefore scope the strong, quantitative displacement result (the ≈0.88 pooled fit, displacement exceeding depth) to the ≤1.4B families, and report the 3B corpus probe as a positive but scale-fragile signal whose clean form does not reproduce – a limitation, and the target for a better-powered corpus-scale study.

- Standard recipes, not exact published configurations. Our five losses are standard implementations. We attempted a registered reproduction of one published table (RCD, CIFAR-10; §5.8): forget accuracies and the depth–RCD direction reproduce, but their published per-method ranking did not under any uniform checkpoint-selection rule, so the matched-depth re-score was gated off. We therefore do not claim to overturn a published ranking, only to show its ordering is not reproducible without unspecified per-method checkpoint choices. We re-scored our own reproductions at matched

depth (a flagged deviation from the registered gate, §5.8): within-pipeline RCD tracks depth, but matched forget-accuracy leaves a 3.6× residual. Matched forget-depth is insufficient: the residual tracks weight displacement (5.5×), not loss-depth (forget-CE equal at matched accuracy; §5.8). A reproduction faithful enough to recover their ranking and re-score it (their TinyImageNet setting) remains the next step for any claim about their models.

- n and power. Variance shares are n=5-seed estimates, seed-count-sensitive (Phi method share 3.2%→0.5% from n=3→5) and depth-window-relative (§5.2); we claim the reslice-robust inequality (run-noise > method, 9/9 windows) on TOFU; this inequality reverses on the two ≤1.4B MUSE-corpus models (§5.2). We claim neither precise shares nor benchmark-universality of the inequality. The pairwise method effect (GD 15–25% faster than NPO/SimNPO) is sign-consistent on all three architectures but Bonferroni-significant only on Llama under our pre-registered rank test (and on GPT-2 only under a post-hoc mean-log statistic, which we disclose and report alongside the registered test in §5.2; §4.3–4.4), null on Phi at both probe LRs (pre-registered follow-up, §5.5), an unexplained cross-architecture difference. Probe LRs differ per architecture (3e-5 Phi; 1.5e-5 GPT-2/Llama; §4.2), a calibration difference we disclose, though §5.5's grid shows the Phi shares are insensitive to it. One depth-response curve of 50 is mildly non-monotone (+0.667, a top-bin plateau). The relearning probe trains only the audited ≥2-D subspace (§4.2); one Llama NPO snapshot (seed 777, 6-nat bin) was placed by the lander's descent branch (brief forget-set training) and its exclusion does not change any conclusion.

Robustness checks. We control for the three main threats to the decomposition – the audited-subspace measurement convention, the single memorized θ^*, and in-sample utility – and the central claims hold under each (precise variance shares, though not the run-noise>method inequality, are sensitive to the θ^* draw).

- Measurement convention is method-differential; controlled on Phi-1.5 and Llama-3.2-1B. Depth, retain, and relearning-probe measurements are taken on each method's audited ≥2-D weight tensors transplanted onto a fixed θ^* shell (θ^*'s biases and norm gains), and the probe trains only those audited tensors. This is exact for RMU (which touches only audited tensors) but discards the 1-D drift of GA/GD/NPO/SimNPO at every evaluation; the distortion is a priori unbounded, so we measured it directly. We re-measured every Phi-1.5 snapshot verbatim, full-parameter: the true full-model depth matches the audited bin label within 0.9 nats (max 1.4), depth-response monotonicity holds (14/15 curves), and the decomposition reproduces, depth 65.5% / method 0.5% / seed 34.0% full-parameter vs 64.8 / 0.5 / 34.7 audited-transplant (depth×method folded into residual). We ran the identical control on Llama-3.2-1B, the architecture where the method effect is real (8.1% audited): it reproduces there too, depth 62.9% / method 7.9% / seed 29.1% full-parameter, true depth within 0.76 nats of the bin label (max 1.4), all 15/15 curves monotone, and the GD-faster effect intact (23%/12% faster than NPO/SimNPO full-parameter). The convention therefore preserves the method effect itself (7.9% vs 8.1%), not just the depth-dominance and the Phi null; GPT-2's snapshots sit under a different harness, but its convention is identical to the two now controlled.
- Single training seed; checked on a second θ^*. The decomposition fixes one memorized θ^* per architecture and varies only the unlearning trajectory, so the run-to-run term excludes training-seed (θ^*) variance; Lanyon et al. (2025) show unlearning metrics can be training-seed-sensitive and that more unlearning seeds cannot substitute for more training seeds. We re-ran the full pipeline from a second, independently-trained θ^* on all three architectures and re-decomposed. The TOFU inequality, run-to-run noise exceeds the method share, reproduces on all three (GPT-2 9.1% vs 0.1%; Llama-3.2-1B 22.8% vs 4.8%; Phi-1.5 78.7% vs 0.4%), as does the Phi method-null (all pairwise non-significant). On Llama-3.2-1B (n=5, probe LR pinned to 1.5e-5) the full ordering also reproduces and the pre-registered GD-faster effect replicates, clearing Bonferroni under the registered rank test (GD-vs-NPO $p<0.0001$). On GPT-2-medium the depth ≫ run ≫ method ordering reproduces and the weak post-hoc GPT-2 method share (5.7%) falls to 0.1%, consistent with our never claiming it. On Phi-1.5, however, the precise shares did not reproduce: one unlearning-trajectory seed relearned ~3× faster than the other four (with retain collapse for GD/NPO), pushing run-to-run noise to 78.7% and depth to 20.7% (excluding that one seed, depth returns to 81.7% and the per-curve

depth-response is intact), so depth-dominance broke even as the inequality held. This is a robustness check on a second draw, not a characterization of training-seed variance: the inequality (run-noise > method) reproduced on all three, but the precise depth share is not point-stable (it stayed close on Llama, 62.7→71.5%, while a single collapsed trajectory seed dominated it on Phi), itself the run-to-run lottery we document (§5.5). We therefore claim the inequality and the pre-registered method effect, not the point shares; a full training-seed variance characterization (Lanyon et al., 2025) remains future work.

- Utility is measured in-sample; held-out controls (Phi-1.5 and Llama-3.2-1B) confirm the dissociation. The §5.3 utility axis is cross-entropy on the same 200 retain rows the retain regularizers train on. Re-evaluating on held-out retain (Phi-1.5, TOFU examples 200–400) shows the in-sample axis is ~4× optimistic (1.15 vs 4.74 nats), but the dissociation survives: method explains 28% of held-out retain-loss variance vs 0.5% of resistance (49.9% in-sample); methods still separate on utility far more than on resistance out-of-sample. The same control on Llama-3.2-1B gives utility method share 17.3% held-out (17.5% in-sample on this retain-loss measure; cf. 19.9% on the §5.3 logit-retain axis) vs resistance method 7.9%, the ~2× separation Llama shows in-sample, intact out-of-sample (retain 1.63→5.33 nats, 3.3× optimistic).

8. Conclusion

Relearning-resistance is a sensible thing to want from an unlearning method, but the way it is currently measured makes it a poor instrument for comparing methods. On fictitious-QA (TOFU), across three model families (and replicated at 3B, n=3), 63–68% of its variance over our ladder is suppression depth, 26–34% is run-to-run noise, and 0.5–8.1% is the method, below run-noise on every TOFU architecture. That inequality is benchmark- and scale-dependent: on the MUSE news corpus the method share rises to 28–37%, overtaking run-noise on the two ≤1.4B models (and depth, on Llama-3.2-1B), while at 3B it rises but run-noise stays the largest term. We further observe (§5.9) a second measurable axis that accounts for this benchmark dependence, weight displacement: at matched depth, relearning resistance tracks how far a method moved the weights from the memorized model, and methods that converge in displacement on fictitious QA (small method share) fan out on the corpus (large one); we report this as an observational refinement, not a controlled axis. Method effects on resistance do exist: GD relearns 15–25% faster than preference-based methods at matched depth (Bonferroni-significant on Llama under our pre-registered rank test, sign-consistent on all three), but on TOFU they sit at roughly a third of typical run-to-run spread, invisible to the single-run, depth-uncontrolled comparisons in current use. Saturating metrics hide the depth, single runs hide the noise, and unmatched-depth comparisons attribute to the method what is suppression depth and, at matched depth, weight displacement. Together, reporting depth in nats, comparing at iso-depth with weight displacement reported, and pinning the scored checkpoint is a small change that makes unlearning-robustness evaluation measure what it claims to.

References

C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations (ICLR)*, 2024a.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Neurips Safe Generative AI Workshop 2024*, 2024b.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective. *arXiv preprint arXiv:2502.05374*, 2025.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, and et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

SeungBum Ha, Saerom Park, and Sung Whan Yoon. Unlearning's blind spots: Over-unlearning and prototypical relearning attack. *arXiv preprint arXiv:2506.01318*, 2025.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, 2025.

J. Jang, D. Yoon, S. Yang, S. Cha, M. Lee, L. Logeswaran, and M. Seo. Knowledge unlearning for mitigating privacy risks in language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. arXiv:2210.01504.

H. Jia, T. Li, J. Guan, and V. Chandrasekaran. The erasure illusion: Stress-testing the generalization of llm forgetting evaluation. *arXiv preprint arXiv:2512.19025*, 2025.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou. Towards unbounded machine unlearning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

J. Lanyon, A. Finke, P. Andreou, and G. Cosma. On the importance of multiple training seeds for evaluating machine unlearning. *arXiv preprint arXiv:2510.26714*, 2025.

Bruce W. Lee, Addie Foote, Alex Infanger, Leni Shor, Harish K Kamath, Jacob Goldman-Wetzler, Bryce Woodworth, Alex Cloud, and Alexander Matt Turner. Distillation robustifies unlearning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Nathaniel Li, Alexander Pan, Anjali Gopal, and et al. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee. Textbooks are all you need II: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

B. Liu, Q. Liu, and P. Stone. Continual learning and private unlearning. *Conference on Lifelong Learning Agents (CoLLAs)*, pp. 243–254, 2022.

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.

T. Qian, V. Suriyakumar, A. Wilson, and D. Hadfield-Menell. Layered unlearning for adversarial relearning. *arXiv preprint arXiv:2505.09500*, 2025.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019.

W. Shi, J. Lee, Y. Huang, S. Malladi, J. Zhao, A. Holtzman, D. Liu, L. Zettlemoyer, N. A. Smith, and C. Zhang. MUSE: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

Shoaib Ahmed Siddiqui, Adrian Weller, David Krueger, Gintare Karolina Dziugaite, Michael Curtis Mozer, and Eleni Triantafillou. From dormant to deleted: Tamper-resistant unlearning through weight-space regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Naman Deep Singh, Maximilian Müller, Francesco Croce, and Matthias Hein. Unlearning that lasts: Utility-preserving, robust, and almost irreversible forgetting in LLMs. In *AI That Keeps Up: NeurIPS 2025 Workshop on Continual and Compatible Foundation Model Updates*, 2025.

Khoa Tran and Simon S Woo. Efficient unlearning through maximizing relearning convergence delay. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7968–7977, 2026.

Qizhou Wang, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. Towards effective evaluations and comparisons for llm unlearning methods. In *International Conference on Learning Representations*, pp. 23327–23355, 2025.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. Unlearning isn't deletion: Investigating reversibility of machine unlearning in llms. *arXiv preprint arXiv:2505.16831*, 2025.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024. COLM 2024.

Appendix / reproducibility notes

All reported numbers reproduce from the scripts and CSVs listed below; code and data will be released on acceptance.

- The release is the project folder: `README.md` (file map + run commands), `requirements.txt`, the per-run CSVs, model-agnostic experiment scripts (`P3_MODEL=<hf-id>` routes checkpoints/CSVs per model), CPU-only analysis (`p3_dissociation.py`, `p3_ancova.py`, `p3_pairwise.py`, `p3_reslice.py`, `p3_m4_analyze.py`, `p3_census.py`), and `p3_figures.py`, which regenerates every figure from the CSVs. Analysis and figures require only numpy/matplotlib; experiments require a single 16GB T4-class GPU (~10–15h per model for the full ladder).
- Second-θ^* robustness check (§7): `p3_phase1_rerun.py` retrains θ^* under a fixed alternate seed (`P3_STAR_SEED`) into a separate output namespace (strictly additive; touches no original checkpoint or CSV), with CPU analysis in `p3_thetastar_check.py` (reproduces the §5.2 decomposition and `p3_pairwise.py` on the original θ^* before scoring the second). Released: `p3_phase1_gpt2medium_rr7.csv`, `p3_phase1_llama321b_rr7.csv`, `p3_phase1_phi15_rr7.csv`.
- The RCD reproduction arm (§5.7–5.8) ships alongside: the source PDF (Tran & Woo, 2026); `rcd_repro.py` (staged CIFAR-10/ResNet-50 harness) + `RCD_REPRO_BRIEF.md` (registered design + amendments A1–A4; A3 = best-checkpoint selection and A4 = the MIA addition are the two reported in §5.8); the probe record (`rcd_cifar10.csv`, `rcd_cifar10_a4.csv`, `rcd_cifar10_a4_evals.csv`, `split.json`, `probe_curves/`); the observational re-read (`rcd_published.csv`, `rcd_published_cifar10.csv`, `rcd_published_cifar10_50.csv`, `rcd_e0.py`, `rcd_e0_cifar.py`); and `rcd_fig16_extract.py` + figure crops (`fig16a_panel_a.png`, `fig16_all_panels.png`) documenting and verifying the Figure 16 extraction against the PDF; the §5.8 per-rule reproduction score (`rcd_repro_score.py`, the $\rho = 0.50/0.50/0.30$ selection-rule table); and the §5.8 matched-depth re-score (`rcd_e2_rescore.py`, data `rcd_cifar10_e2.csv`) with its forget-loss/displacement control (`rcd_nats_check.py`, data `rcd_nats_check.csv`).
- Variance decomposition: OLS variance shares of log(steps) and logit(retain) on the balanced core. Because the core is balanced (depth and method crossed with equal cells), the shares are order-invariant: Type I, II, and III sums of squares coincide, so the depth share does not depend on entry order; the method-effect significance is additionally established by the order-independent within-bin permutation tests and ANCOVA (§4.3, §5.2).
- Registered endpoints and the killed geometry hypothesis are documented in the brief to characterize the evidentiary standard.