Quantifying Canary Score Dependence in One-Run Unlearning Audits

Anonymous authors
Paper under double-blind review

Abstract

One-run privacy audits for machine unlearning embed k canaries in a forget set, apply an unlearning operator, and aggregate per-canary membership scores into a confidence statement, assuming the scores contribute independent information about the run. We show this assumption fails in a structured way. Adapting the intraclass-correlation (ICC) decomposition from biostatistics, the marginal ICC(1,1) is near zero for conservative operators (NPO, RMU) while the conditional ICC(3,1), on canary-centered residuals, is substantial—0.045–0.217 for conservative operators, 0.069–0.713 for gradient ascent, all converged: a marginal diagnostic is blind to the dependence that matters for per-canary procedures. Within the dependence that survives canary centering, the run-level share dominates, with a smaller entity-block component (permutation $p = 0.000999$); the mechanism is not identified by this statistic. At $k = 400$ the conditional design effect reaches $\mathrm{DEFF}_c \approx 286$ for converged gradient ascent—400 canaries carry only ∼1.4 independent observations (scalar) or ∼1.18 (eigenvalue bound)—and independence-assuming mean-MIA intervals under-cover (marginal tier, $\sqrt{\mathrm{DEFF}_m}$ convention), while the per-canary-count statistic mildly under-covers at 87.8–93.9% naive but is over-conservative ($\geq$ 99.5%) once corrected; for conservative operators a fresh canary draw expects near-zero transferable dependence (ICC(2,1): NPO 0.020, RMU 0.002). At the highest coupling strength (converged GA), the correctly scaled Level-1 correction reaches full coverage on this cell (1.000, 30/30; unchanged under residual-run resampling), so the auditor proceeds rather than declining. The framework requires $R \geq$ 20–30 calibration runs for both detection and magnitude estimation, so it is a *diagnostic* that tells an auditor when the one-run regime is valid.

1 Introduction

Privacy auditing for machine unlearning embeds k canary records in a forget set, applies the unlearning operator, and aggregates per-canary membership scores into a confidence statement, increasingly from a single unlearning execution rather than many shadow models (Shokri et al., 2017; Steinke et al., 2023; Keinan et al., 2025; Dagréou & Bellet, 2026; Sarmasarkar et al., 2026; Agrawal et al., 2026; Xiang et al., 2025a). Aggregating these scores requires assuming they contribute independent information about the run.

That assumption is a design target rather than a settled fact: canaries are crafted to minimize interference (Dagréou & Bellet, 2026), and the limits of one-run precision are attributed to *interference* between audited elements (Keinan et al., 2025). We test it and find it fails in a structured way. Within a single unlearning run, per-canary scores co-vary: in *total score variance* the canary main effect dominates (Table A2), but that is exactly the component ICC(3,1) removes by centering. Within the *conditional dependence that survives centering*, the run-level share dominates, with a smaller entity-block component. An auditor's k canaries are therefore fewer than k effective observations: naive intervals under-cover for the mean-MIA aggregation, and the over-statement grows as canaries are added. The standard marginal diagnostic (one-way ICC) misses this—it is overwhelmed by the large intrinsic between-canary variance.

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.

Our core insight is that this dependence can be decomposed into two tiers. Adapting the intraclass-correlation decomposition from biostatistics, the marginal ICC(1,1) asks whether raw scores co-vary within a run, while the conditional ICC(3,1) asks whether the canary-centered residuals co-vary, after removing each canary's intrinsic memorability. The two-tier gap is the key: marginal ICC(1,1) is near zero for conservative operators (NPO, RMU) while conditional ICC(3,1) is substantial (0.045–0.217 for the conservative operators, 0.069–0.713 for gradient ascent, all converged). The design effect $\text{DEFF} = 1+(k-1)\rho$ converts this into an effective sample size; at $k = 400$ with NPO, conditional DEFF reaches 57, so 400 canaries provide the power of only 7 independent observations. A B=1000 entity-block permutation test confirms the dependence is not purely run-level (empirical $p = 0.000999$ for all three operators; raw entity-block excess NPO 0.147, RMU 0.100, GA 0.016, the last small in absolute terms against ICC 0.747), and an entity inclusion/exclusion intervention supports forget-set-composition-driven redistribution rather than entity-local coupling (Appendix E). The exact mechanism (subspace coupling vs. optimizer noise vs. run-level scale shift) is not identified by this statistic; the gradient-cosine probe (0.618 for GA, 0.0858 for NPO) is an n=1 supporting observation on re-generated checkpoints, not causal evidence.

The consequences are tier- and operator-dependent. At the *marginal* tier GA is catastrophic (0.30 naive coverage, restored to 0.87; the ‡ post-divergence cell), NPO substantial (0.50, restored 0.96), and RMU mild (0.98); at the *conditional* tier NPO (0.149) and RMU (0.243) show strong dependence. The framework is a *diagnostic*, not a rescue: a group-sequential test rejects independence at every look ($z = 8.65$ at $R = 99$; Appendix F), and recovering valid inference requires $R \geq 20$–30 calibration runs (∼30× the one-run cost).

We validate it across three operators (NPO, GA, RMU), two model scales (1B, 3B), and two forget splits (2 entities, k=40; 20 entities, k=400), with three contributions:

1. **Measurement of within-run dependence (C1):** Marginal diagnostics are blind to it. ICC(1,1) ≈ 0 means the run-level variance σ_u^2 is small *relative to* the within-run variance σ_w^2, not that $\sigma_u^2 \approx 0$—so seed-variance accounting that stops at the marginal tier (e.g. Bouthillier et al., 2021) does not register it. The conditional ICC(3,1), which removes the canary main effect, is substantial: 0.045–0.217 (NPO, RMU) and 0.069–0.713 (gradient ascent) converged; at $k = 400$ the conditional design effect reaches $\text{DEFF}_c = 57.1$ (NPO) and 79.2 (RMU), so adding canaries does not buy independent evidence. The dependence is strongly composition-conditional (NPO ICC(3,1) 0.028–0.272 across five entity compositions at fixed k), so any single value is one composition's draw (Section 4.3). A B=1000 entity-block permutation confirms a smaller entity-structured component beyond the run-level shift; the mechanism is not identified by this statistic.

2. **Diagnostic and correction hierarchy (C2):** We provide a three-level correction hierarchy assembled from standard components, with a calibration-cost accounting: Level 0 e-value averaging ($k_{\text{eff}} \approx 1$, no calibration; a worst-case floor (Vovk & Wang, 2021; Wang, 2025), but approximately uninformative), Level 1 DEFF inflation (k_{eff}=1.18–19), and Level 2 cluster-product using entity structure (k_{eff}=20–47, applying standard cluster-robust standard errors (Moulton, 1990; Cameron & Miller, 2015)). All but Level 0 require $R \geq 20$–30 calibration runs, so the hierarchy is not a one-run protocol, and the compute–validity trade-off is the headline result; the calibration transfers only within the same entity family (Section 3.5).

3. **Transfer and scale sensitivity (C3):** Two scoping checks. First, the dependence is *operator-specific, not an artifact of TOFU's entity structure*: a successful gradient-ascent unlearning shows it on the entity-free MUSE-News benchmark (ICC(3,1) 0.917), whereas an NPO MUSE no-op does not (0.0037). Second, scale and k are single-instance observations, not a scaling law: the 1B-vs-3B difference confounds k with entity count and each cell is one finetuned θ_0, while the clean within-forget10 subsample shows all three operators approximately k-invariant (RMU ICC(3,1) stable 0.178–0.227). A single exploratory cross-family arm (Qwen2.5-7B, R=10; confounded with scale and optimizer) is in Appendix E.

2 Related Work

One-run privacy auditing. Privacy auditing originally required multiple shadow models (Shokri et al., 2017), with foundational MIA methods ranging from overfitting-based thresholds (Yeom et al., 2018) to likelihood ratios (Carlini et al., 2022). Steinke et al. (2023) introduced the one-run paradigm, and Keinan et al. (2025), Dagréou & Bellet (2026), Sarmasarkar et al. (2026), Agrawal et al. (2026), Liu et al. (2025), and Xiang et al. (2025a) extended it with tighter bounds, improved canary crafting, and bits-transmission; the per-canary inclusion-bit bounds of Steinke et al. (2023) stay valid under dependence (each coin flip is independent by design). Keinan et al. (2025) name *interference* between audited elements as a fundamental limit on one-run precision, and Dagréou & Bellet (2026) design canaries specifically *to minimize interference*—so per-canary near-independence is a deliberate engineering target, not an unexamined assumption. The closest prior machinery is Pillutla et al. (2023), whose DP-audit intervals adapt to empirical higher-order canary correlations (Miller, 2024 likewise handles correlated eval items). A one-run method (Agrawal et al., 2026) assumes canary independence—justified by near-orthogonality of randomly sampled canary directions—and validates it with a direction-orthogonality check—a check that, *like the marginal ICC we test*, operates marginally and would not register conditional score-level dependence; its authors note that characterizing the idealization's bias remains open. Liu & Xiong (2025) and Naderloui et al. (2025) audit under canary independence without coverage guarantees, Sun et al. (2026) add reliability quantification, and Li et al. (2026) use head-conditioned canaries elsewhere; none models within-run score dependence. Pillutla's and Miller's intervals are dependence-robust but do not quantify how much within-run score dependence an operator and forget-set composition induce—nor show the marginal diagnostic is blind to it. That measurement is our contribution.

Unlearning operators, benchmarks, and the mechanism. We study NPO (Zhang et al., 2024b), gradient ascent (Jang et al., 2023), and RMU (Li et al., 2024) on TOFU (Maini et al., 2024) via OpenUnlearning (Dorna et al., 2025); MUSE (Shi et al., 2025) offers an entity-free benchmark on which our transfer check found near-zero dependence under an NPO no-op but strong dependence under GA (Section 4.3). Carlini et al. (2019) showed neural networks memorize training-set secrets; our per-canary MIA scores descend from that test. The mechanism is not new: Sondej & Yang (2025) show cross-fact suppression in unlearning, and the linear-representation-hypothesis (Park et al., 2024) and representation-engineering (Zou et al., 2023) strands establish that high-level concepts live as directions a single steering move shifts together. Neither strand quantifies *how much* dependence the resulting per-canary *scores* carry within a single run. We measure that dependence, showing it is operator- and composition-conditional, with a marginal diagnostic blind to it; the mechanism link is *correlational* (the gradient probe is on re-generated checkpoints, not the audited weights).

Audit verification and correlated testing. Thudi et al. (2022) formalized auditable unlearning definitions; Zhang et al. (2024a) showed verification is fragile under distribution shift. Jagielski et al. (2020) and Nasr et al. (2023) developed and tightened DP auditing with multi-canary insertions whose independence assumption we test. Guo et al. (2023) analyzed correlated hypothesis testing via multiple-testing theory. Mahloujifar et al. (2025) proved tight composition-based bounds for mechanism-randomness-induced dependence, and Xiang et al. (2025b) gave a tight one-run (ε, δ)-DP lower bound via f-DP and order statistics; sequential auditing across accumulating runs is due to González Lara et al. (2025). All address how mechanism randomness interacts with the privacy guarantee, whereas our dependence is within-run canary-score co-variation induced by a single update—complementary, not competing. We adopt survey-sampling corrections (DEFF) and cluster-robust standard errors (Cameron & Miller, 2015) rather than multiple-testing adjustments (Benjamini & Yekutieli, 2001), since we estimate one aggregate parameter from correlated observations. Zhang et al. (2026) detect subset-level dependence without calibrated bounds; Tang et al. (2026) audit behaviorally.

ICC, design effects, e-values, and the two-tier gap. Shrout & Fleiss (1979) established the ICC taxonomy; Donner & Koval (1980) developed ICC estimation for family data;

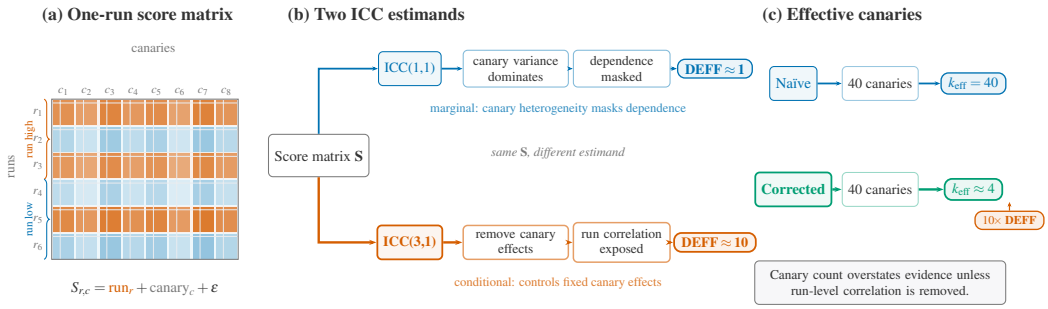


Figure 1: The two-tier ICC framework. **(a)** A single unlearning run produces a score matrix where canaries co-vary due to a shared run-level effect. **(b)** The marginal ICC(1,1) is diluted to DEFF ≈ 1 by canary variance, while the conditional ICC(3,1) removes canary effects and exposes the true dependence (DEFF ≈ 10). **(c)** The effective canary count under naïve vs. DEFF-corrected inference: 40 nominal canaries yield $k_{\text{eff}} \approx 4$ after accounting for run-level correlation.

variance-component estimation via REML/ML (Searle et al., 1992) underpins these estimators (we use ANOVA mean-squares; Appendix H). DEFF $= 1+(k-1)\rho$ originates in survey sampling (Kish, 1965); Moulton (1990) demonstrated its severity for micro-level inference. Bouthillier et al. (2021) showed random seed variance in ML benchmarks is non-negligible. E-values (Vovk & Wang, 2021) handle dependence gracefully; Wang (2025) proved arithmetic averaging is the unique admissible combination. Group-sequential monitoring uses α-spending boundaries (Lan & DeMets, 1983; O'Brien & Fleming, 1979; Jennison & Turnbull, 2000); we apply this to ICC(3,1) (Appendix F). The marginal-versus-conditional distinction is a canonical multilevel decomposition (Raudenbush & Bryk, 2002, Bliese, 2000, Snijders & Bosker, 2011); the spectral effective-sample-size machinery (Cheverud, 2001, Nyholt, 2004, Galwey, 2009) disagrees with the scalar-DEFF route by 0.63–2.68 in the small-k conditions (Appendix I).

3 A Two-Tier ICC Framework for Canary Dependence

One-run privacy audits embed k canary records in a forget set, apply an unlearning operator, and compute per-canary membership inference scores, which existing frameworks aggregate assuming independence across canaries within a run. We adapt tools from psychometrics and survey sampling to characterize when this fails and derive corrections that restore valid inference. Our contribution is a *measurement and correction*: within a single unlearning run, per-canary scores co-vary; in the conditional dependence that survives canary centering, the run-level share dominates, with a smaller entity-block component confirmed by permutation testing. The gradient-cosine probe (0.618 for GA, 0.0858 for NPO) is an n=1 supporting observation on re-generated checkpoints, not causal evidence; a global per-run score-scale shift is not excluded.

3.1 Problem Formulation

Consider an unlearning audit with k canaries evaluated over R independent repetitions of the unlearning procedure, giving the score matrix $S \in \mathbb{R}^{R\times k}$ where $S_{r,j}$ is the membership score for canary j in run r (same operator and data, different random seed per run). The central question: given a single run r, are $S_{r,1}, \ldots, S_{r,k}$ independent, and if not, what is the effective sample size k_{eff} and how can valid confidence statements be recovered?

3.2 Two-Tier ICC Decomposition

We decompose dependence into two distinct tiers using the intraclass correlation coefficient (ICC) taxonomy of Shrout & Fleiss (1979) and McGraw & Wong (1996).

Marginal dependence: ICC(1,1). Under a one-way random effects model where runs are exchangeable random factors, $S_{r,j} = \mu + u_r + w_{r,j}$ with $u_r \sim \mathcal{N}(0, \sigma_u^2)$, $w_{r,j} \sim \mathcal{N}(0, \sigma_w^2)$, the marginal ICC is:

$$\mathrm{ICC}(1,1) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_w^2} = \frac{\mathrm{MS}_{\text{between}} - \mathrm{MS}_{\text{within}}}{\mathrm{MS}_{\text{between}} + (k-1)\,\mathrm{MS}_{\text{within}}}\,. \tag{1}$$

This measures whether runs are globally "good" or "bad" for all canaries simultaneously: a positive ICC(1,1) means one canary's score in a run is informative about the mean of all others in that run. It governs aggregate statistics such as mean AUC or pooled tests. Note that σ_w^2 absorbs both canary heterogeneity and residual noise; the one-way model does not distinguish them.

Conditional dependence: ICC(3,1). The marginal model ignores systematic differences between canaries. Under a two-way mixed model that partitions out canary main effects, $S_{r,j} = \mu + c_j + u_r + \epsilon_{r,j}$, where c_j is the fixed effect of canary j (intrinsic memorability differences) and $\epsilon_{r,j}$ the residual after removing both canary and run effects, the conditional ICC is:

$$\mathrm{ICC}(3,1) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}\,, \tag{2}$$

estimated from the two-way ANOVA. Unlike σ_w^2 in the one-way model (which absorbs canary heterogeneity), σ_ϵ^2 here is pure residual variance, so ICC(3,1) measures whether the canary-centered residuals co-vary across canaries within a run.

Canary-as-random vs. fixed: ICC(2,1). ICC(3,1) treats the canaries as fixed, quantifying dependence within a specific audit's set; ICC(2,1) additionally treats them as a random sample, so it is transferable to future audits using different canaries. It is the Shrout–Fleiss Case 2 estimator, a *two-way random-effects* model with canary effect c_j, run effect u_r, and interaction+residual $\epsilon_{r,j}$; we report its agreement form (single measures), estimated from the mean squares as:

$$\mathrm{ICC}(2,1) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_c^2 + \sigma_\epsilon^2} \quad = \quad \frac{\mathrm{MS}_R - \mathrm{MS}_E}{\mathrm{MS}_R + (k-1)\,\mathrm{MS}_E + (k/n)(\mathrm{MS}_C - \mathrm{MS}_E)}\,, \tag{3}$$

where $\mathrm{MS}_R, \mathrm{MS}_C, \mathrm{MS}_E$ are the run, canary, and residual mean squares and $n = R$. Unlike Eq. 2, the denominator carries the between-canary variance σ_c^2, so ICC(2,1) shrinks as between-canary variance grows—the masking that makes the fixed-canary quantity more relevant to a specific audit. The two-way-*mixed* consistency estimator (case 3, Eq. 2) omits the MS_C term; CIs use the same within-run-row resampling as ICC(3,1) (Appendix N).

Semantic re-purposing. The canonical ICC(2,1) is a reliability coefficient; we re-purpose it as the run-level dependence expected for a fresh random canary set. The two views differ substantially (Section 4.4), and our headline claims (e.g., GA DEFF = 299, post-divergence) are fixed-canary by construction—*definitional*, in that ICC(3,1) conditions on the audited canaries, not an empirical finding. A synthetic check (Appendix N) confirms the fixed-vs-random gap is a genuine property of the two-way decomposition, not an estimator artifact—mirroring the reported GA 0.747 vs. 0.315.

The two-tier gap. $\mathrm{ICC}(1,1) \approx 0$ does not imply $\mathrm{ICC}(3,1) \approx 0$. When canaries have large intrinsic memorability differences (high between-canary variance), the marginal ICC is diluted toward zero even when substantial run-level effects exist; the conditional ICC removes this masking and reveals the dependence that matters for per-canary calibrated procedures (Appendix C).

3.3 Design Effect and Effective Sample Size

The design effect (DEFF) quantifies the variance inflation caused by intraclass correlation (Kish, 1965):

$$\text{DEFF} = 1 + (k-1)\rho\,, \tag{4}$$

where ρ is the relevant ICC. The effective sample size is:

$$k_{\text{eff}} = \frac{k}{\text{DEFF}} = \frac{k}{1+(k-1)\rho}\,. \tag{5}$$

This formula reveals a critical scaling property: DEFF grows *linearly* with k, so adding canaries without correction inflates apparent precision rather than improving it. An auditor increasing canaries from 40 to 400 expects a $\sqrt{10}$-fold power gain under independence, but if $\rho = 0.14$ (as we measure for NPO) the actual gain is negligible (k_{eff} moves from 4.2 to 7.0). Two distinct DEFFs operate simultaneously:

$$\text{DEFF}_{\text{marginal}} = 1 + (k-1)\,\text{ICC}(1,1)\,, \tag{6}$$

$$\text{DEFF}_{\text{conditional}} = 1 + (k-1)\,\text{ICC}(3,1)\,. \tag{7}$$

$\text{DEFF}_{\text{marginal}}$ (via ICC(1,1)) corrects the *raw per-run mean MIA*—the aggregate statistic an auditor tests; $\text{DEFF}_{\text{conditional}}$ (via ICC(3,1)) corrects per-canary calibrated procedures and the mean of canary-centered residuals (Section 4.4). Marginal DEFF is truncated to ≥ 1.0 (a negative ICC(1,1) yields $\text{DEFF}_m < 1$, which is not a meaningful correction). This assignment makes the two tiers operationally distinct: marginal DEFF governs the aggregate severity, conditional DEFF the dependence that drives k_{eff}/quantile/cluster-product corrections.

3.4 Correction Hierarchy

Given a measured conditional ICC, we assemble a three-level correction hierarchy (Level 0/1/2) that restores honest confidence at a power cost proportional to the dependence; each level subsumes the previous one—correctly scaled, Level 1 restores valid coverage including the converged high-dependence cell (1.000; Appendix O); Level 2's tighter bound is unvalidated (Section 3.5). The hierarchy is an *assembly of standard components*—e-value averaging (Vovk & Wang, 2021; Wang, 2025), design-effect inflation (Kish, 1965; Moulton, 1990), and cluster-robust standard errors (Cameron & Miller, 2015)—not a new estimator; the contribution is the *diagnostic* that motivates it and the calibration-cost accounting, in the spirit of error-bar accounting for correlated evaluation items (Miller, 2024). The levels run from most conservative (Level 0) to most powerful but most demanding (Level 2).

Level 0. Averaging e-values across canaries gives a valid combined e-value under arbitrary dependence (Vovk & Wang, 2021; Wang, 2025): the arithmetic mean $\bar{E} = k^{-1}\sum_j E_j$ is a valid e-value with no dependence assumptions. But $k_{\text{eff}} \approx 1$, because the average grows at most as $\max_j E_j$—so it is a worst-case floor, scoped to mean/proportion statistics.

Level 1. Given an ICC(3,1) estimate from R calibration runs, the auditor inflates confidence intervals by $\sqrt{\text{DEFF}/(1-\rho)}$, $\widehat{\text{Var}}_{\text{corrected}} = \text{DEFF}_{\text{conditional}}/(1-\rho)\cdot\widehat{\text{Var}}_{\text{naive}}$, the factor $1/(1-\rho)$ being required because the naive SE is built from the within-run variance of the canary-centered scores, which excludes the run effect; $k_{\text{eff}} = k/\text{DEFF}_{\text{conditional}}$ (1.18–19; the degenerate GA-3B-f10 floor excluded, and the eigenvalue bound used where the scalar overstates k_{eff}, App. I). So scaled, Level 1 restores valid coverage throughout, including the converged high-dependence cell (Section 4.4); it treats all canaries as one correlated pool, discarding entity structure.

Level 2. Grouping canaries by entity and applying the DEFF correction at the entity level recovers $k_{\text{eff}} = 20$–47 (6–16× over Level 1). For heterogeneous operators (NPO, $I^2 = 0.66$) random-effects pooling of per-entity ICCs is recommended; for homogeneous ones (GA, $I^2 = 0.36$) a single pooled ICC suffices.

3.5 Calibration Protocol

The framework requires a calibration phase of $R \geq 20$–30 repeated unlearning runs, the central trade-off: the hierarchy is *not* a one-run protocol but a diagnostic. The budget depends on the goal: *estimating* the dependence magnitude needs $R \geq 20$–30 (Appendix E); the product e-value crosses its threshold sooner but responds to canary heterogeneity rather than run dependence (Appendix J), so it does not license a shorter budget. The protocol: apply the operator R times with distinct seeds ($R = 30$ suffices for stable ICC estimation); compute per-canary MIA scores per run ($S \in \mathbb{R}^{R \times k}$); estimate ICC$(1,1)$ and ICC$(3,1)$; and apply the measured DEFF to subsequent audits. Estimator details (ANOVA mean-squares, bootstrap vs. parametric CIs, non-finite handling, degenerate-run exclusion) are in Appendix H. The transfer is *composition-conditional*, not universal: for NPO at $k = 400$, ICC(3,1) ranges 0.028–0.272 across five random compositions ($\sim$9.7$\times$ spread), so the practical rule is to calibrate per operator–family pair; cross-composition transfer is not validated (no pair of our five compositions reaches 80% entity overlap; Appendix E).

Calibration cost and alternatives. Running $R = 30$ calibration passes is substantial ($\sim$28 GPU-hours total); the alternative is the no-calibration e-value floor (Level 0, $k_{\text{eff}} \approx 1$). Calibration pays when $k \gg R$ and Level 2 entity structure gives leverage ($k_{\text{eff}}^{\text{L2}}/k_{\text{eff}}^{\text{L1}} > 4\times$; observed 6–16$\times$); without entity structure (e.g., MUSE), only Levels 0–1 apply. The DEFF correction assumes compound symmetry, violated everywhere, so the scalar ρ averages over the true covariance; eigenvalue-based k_{eff} is more accurate at the cost of the full covariance (Appendix I). Three uncertainties quantified above—the composition spread ($\sim$9.7$\times$), the θ_0' sensitivity (-28%), and the R=30 sampling half-width ($\pm$0.071)—each match or exceed the 0.141 point estimate anchoring the moderate-ICC case, so for moderate-ICC operators a calibrated DEFF is order-of-magnitude guidance, not a transferable constant.

4 Experiments

We evaluate the two-tier ICC framework across thirteen experimental conditions (twelve non-degenerate, reported in Table 1, plus the degenerate GA-3B-f10 collapse) spanning three unlearning operators, two model scales, and two dataset splits with different canary counts; additional cells sit outside this count and are reported in the appendix—θ_0' replications of the converged short-budget GA and the NPO-f10 conditions (Appendix Table A3), and a $\geq$7B cross-family arm (Appendix E).

4.1 Experimental Setup

Dataset and models. We use the TOFU benchmark (Maini et al., 2024) with Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct. The base model θ_0 is finetuned on the TOFU `full` split (`locuslab/TOFU`, 200 fictitious entities) for 5 epochs (AdamW, lr 1×10^{-5}, batch 8); a *single shared* θ_0 per model initializes *all* unlearning runs, so run-level variation comes solely from the unlearning seed. This holds for every condition except the converged GA-f10 cell and the NPO-f10 cell, also replicated on an independently finetuned second base (θ_0'; Appendix Table A3). We evaluate on `forget01` ($k = 40$ canaries from 2 entities) and `forget10` ($k = 400$ from 20 entities). We focus on TOFU because its entity-structured forget sets enable the cluster-product correction; benchmarks such as MUSE (Shi et al., 2025) lack this structure. As a transfer check we additionally ran NPO and GA on MUSE-News (Section 4.3).

Unlearning operators. We apply three operators from OpenUnlearning (Dorna et al., 2025): NPO (Zhang et al., 2024b), gradient ascent (GA) (Jang et al., 2023), and RMU (Li et al., 2024). NPO and GA use the TOFU defaults (AdamW, lr 1×10^{-5}, batch 8/grad accum 4, 10 epochs): NPO minimizes $\gamma\ell_{\text{forget}} + \alpha\ell_{\text{retain}}$ with β=0.1, α=1.0, γ=1.0; GA negates the forget loss with no retain term or gradient clipping. RMU uses the WMDP defaults (AdamW, lr 5×10^{-5}, batch 1/grad accum 16, 80 max steps, constant LR, steering 2.0 at layer 7, `EMBED_DIFF` retain loss), a smaller effective budget. All runs use a per-run seed $\{0, \ldots, R-1\}$; full per-condition config in Appendix Table A1.

Table 1: Two-tier ICC results across all non-degenerate conditions. ICC(1,1): marginal dependence; ICC(3,1): conditional dependence after removing canary main effects. DEFF $= 1 + (k-1)\rho$; arrows indicate greater run-level dependence. Bootstrap 95% CIs in brackets (parametric for 3B). "—": ICC(1,1) not computed. Marginal DEFF truncated to ≥ 1.0. ‡post-divergence; *converged short budget. GA-3B-f10 (degenerate collapse) excluded.

Condition	k	ICC(1,1) ↑	DEFF_m ↑	ICC(3,1) [95% CI] ↑	DEFF_c ↑	k_eff^{L1} ↓[1]	Primary
NPO-1B-f01	40	−0.013	1.0	0.217 [.168, .265]	9.5	4.2	scalar
GA-1B-f01	40	0.003	1.12	0.306 [.205, .394]	12.9	3.1	eigen
RMU-1B-f01	40	−0.025	1.0	0.121 [.055, .234]	5.7	7.0	eigen
NPO-3B-f01	40	−0.020	1.0	0.072 [.019, .125]	3.8	10.5	eigen
GA-3B-f01	40	−0.017	1.0	0.069 [.038, .104]	3.7	10.8	eigen
RMU-3B-f01	40	−0.026	1.0	0.175 [.107, .218]	7.8	5.1	eigen
NPO-1B-f10	400	0.018	8.18	0.141 [.071, .212]	57.1	7.0	scalar
GA-1B-f10‡	400	0.314	126.2	0.747 [.520, .833]	298.9	1.3	scalar
RMU-1B-f10	400	−0.001	1.0	0.196 [.124, .255]	79.2	5.1	scalar
GA-1B-f10s*	400	0.112	45.5	0.713 [.664, .748]	285.6	1.4	eigen
NPO-3B-f10	400	—	—	0.133 [.044, .191]	54.0	7.4	scalar
RMU-3B-f10	400	—	—	0.045 [.026, .067]	19.0	21.1	eigen

MIA scoring. Per-canary membership inference scores are the *mean negative log-likelihood (NLL)* of the unlearned model on the sample—the loss attack's average loss:

$$\text{MIA}(x) = \ell_{\theta_u}(x)\,, \tag{8}$$

where $\ell_{\theta_u}(x)$ is the cross-entropy (per-token NLL) of the unlearned model on sample x, averaged over the sample's tokens. Higher values indicate stronger evidence of membership (the unlearned model performs worse on forget-set samples). Both pipelines reported here score this *raw NLL*: the OpenUnlearning `mia_loss` metric (its `avg_loss`) for TOFU, and a standalone per-article NLL scorer for MUSE-News. This is deliberately *not* a base-normalized ratio $\ell_{\theta_u}(x)/\ell_{\theta_0}(x)$; no such ratio is computed, and the absolute NLL magnitudes are interpreted on the raw scale (roughly 2 nats for a converged unlearning, and much larger under the post-divergence regimes of Section 4.3).

Repeated-run protocol. For each condition we execute R independent unlearning runs with distinct seeds (seeds $\{0, \ldots, R-1\}$), giving the score matrix $S \in \mathbb{R}^{R\times k}$. RMU-1B-f01 uses $R = 10$ as a lower-precision concordance check (wider CIs). For NPO-3B-f10, runs 7, 18, and 24 produced no score file, so $R = 27$ is the complete available set—a data-availability shortfall, not a statistical exclusion; the runs were logged as permanently failed, and no gate, AUC, or convergence evidence for them exists, so the full-30-seed ICC is not computable. The per-condition matrix is in Appendix Table A1.

4.2 Main Results: Two-Tier Dependence Exists

Table 1 presents the central finding: marginal and conditional ICC diverge substantially across all conditions.

The GA-3B-f10 condition is a degenerate collapse (ICC(3,1) = 0.988, DEFF_c = 395.2, $k_\text{eff} \approx 1.0$; run variance 98.2%), excluded from this table and from all non-degenerate ranges and headline k_eff claims; it is best understood as catastrophic collapse (McCloskey & Cohen, 1989), not a scaling property (Appendix K). ‡GA-1B-f10 is *post-divergence* ($\sim$ 77 nats vs. $\sim$ 2.3 converged); *GA-1B-f10s is the *converged* short-budget counterpart (1 epoch; 1.94 nats); their near-identical conditional ICC (0.713 vs. 0.747) while loss drops $\sim$40$\times$ shows GA's coupling does not require post-divergence saturation (Appendix E). On a second, independently finetuned θ_0, GA's interaction-structured dependence is stable (0.724 vs. 0.713; Table A3), and the NPO companion shows 0.101 vs. the anchor's 0.141 (CI $[0.050, 0.147]$ overlaps the anchor's $[0.071, 0.212]$)—both power-limited nulls at R=30, not θ_0-

invariance; the directional trends are opposite (θ_0' lower for NPO, higher for GA), suggesting any θ_0 effect, if present, is operator-dependent.

Across ten conditions with reported marginal ICC, ICC(1,1) clusters near zero; the two exceptions are both GA at $k = 400$: the post-divergence GA-1B-f10 (0.314) and the *converged* short-budget GA-1B-f10s (0.112)—so gradient ascent is the one operator whose dependence registers in *both* tiers. Conditional ICC(3,1) is positive across all conditions: 0.045–0.217 for the converged conservative operators (NPO, RMU), 0.069–0.713 for converged gradient ascent, and higher for the two *post-divergence* GA cells (0.747 TOFU-f10, 0.917 MUSE). Is GA's coupling then a saturation artifact? Changing only the training budget (1 epoch vs. 10) leaves conditional ICC nearly unchanged (0.713 at 1.94 nats vs. 0.747 at 77.4), while loss drops $\sim$40$\times$, so GA's dependence does not require saturation ($\text{DEFF}_c = 286$). As a baseline control, the pre-unlearning finetuned checkpoint gives identical MIA scores across hypothetical runs (between-run variance = 0) despite substantial between-canary variation (std 0.37–0.48); this baseline has zero run-variance *by construction*—a *weak* control, not a causal counterfactual—and the NPO-MUSE no-op (ICC 0.004) provides a directional near-zero observation under a did-not-forget condition.

At $k = 40$, conditional DEFF ranges from 3.8 to 12.9 ($k_{\text{eff}} = 3.1$–10.5). At $k = 400$ (1B) the collapse is more severe—NPO $k_{\text{eff}} = 7.0$, RMU $k_{\text{eff}} = 5.1$, GA $k_{\text{eff}} = 1.3$ (converged GA $k_{\text{eff}} = 1.4$). The coverage consequence is aggregation-specific: the *per-canary count* shows 87.8–93.9% naive coverage, corrected to an over-conservative $\geq$ 99.5%, distinct from the *mean-MIA* severity of Section 4.4. A group-sequential test rejects independence at every look (Appendix F); the 1B DEFF structure is in Appendix K.

4.3 Generalizability: Operators, Scales, and Entities

Operator gradient and scale. At 1B, conditional ICC orders GA > RMU > NPO at $k = 400$ ($0.747 > 0.196 > 0.141$) and GA > NPO > RMU at $k = 40$ ($0.306 > 0.217 > 0.121$). The scale response is not a law: an observed difference between two checkpoints, and it differs by k. At $k = 40$, 1B$\rightarrow$3B attenuates ICC for NPO (-67%) and GA (-77%); at $k = 400$, NPO is stable (0.133) and RMU attenuates (-77%, $k_{\text{eff}} = 21.1$). Because the cross-split contrast conflates k with entity count, the clean k test is the within-forget10 subsample, which shows all three operators approximately k-invariant (RMU ICC(3,1) stable at 0.178–0.227 across $k = 40$–400; Appendix E). More broadly, dependence is a design factor of the operator, responding monotonically to hyperparameters and operator-specifically to composition; a MUSE-News transfer check confirms it is *operator-specific* (Appendix E).

4.4 Coverage Validation

The mean-MIA severity is tier-dependent, and correction restores valid coverage for most conditions (Table A14): at the *marginal* tier the post-divergence GA cell is catastrophic (ICC(1,1)= 0.314, $\text{DEFF}_m = 126.29$; 0.30 naive $\rightarrow$ 0.87 corrected), NPO substantial (0.50 naive, a 2$\times$ error, $\rightarrow$0.96), and RMU mild (0.98, truncated DEFF_m=1.0 so no correction); these marginal-tier corrected values use the pre-correction $\sqrt{\text{DEFF}_m}$ convention, so they understate and would move toward nominal under the analogous $\sqrt{\text{DEFF}_m/(1 - \text{ICC}(1,1))}$ factor (Appendix O); at the *conditional* tier NPO (0.149) and RMU (0.243) show strong dependence. Table A14's GA row is the ‡ *post-divergence* cell, not the * converged one. On the *converged* GA cell (ICC(3,1) = 0.713, $\text{DEFF}_c = 286$), the correctly scaled Level 1 correction reaches 1.000 coverage (30/30; Appendix O), with the corrected interval conservative ($\approx$31.5$\times$ the naive width) rather than merely certifying. Level 2 narrows rather than widens (coverage 0.100), so at the highest coupling strengths the conservative choice is Level 1. The canary-as-random counterpart ICC(2,1) is far weaker (GA 0.315 vs. 0.747, NPO 0.020 vs. 0.141, RMU 0.002 vs. 0.196); the definitional basis is in Appendix N.

5 Conclusion

Per-canary scores in one-run unlearning audits are not independent: marginal diagnostics miss the within-run dependence, reaching ICC(3,1) = 0.713 for converged gradient ascent,

while a fresh draw expects near-zero transferable dependence (ICC(2,1) ≤ 0.020). Dependence is predominantly run-level with a smaller entity-block component ($p = 0.000999$); the mechanism is not identified. At the converged GA cell the set-specific fixed-canary DEFF (286; 299 post-divergence) is handled by the correctly scaled Level-1 correction, which reaches full coverage (Appendix A). $R \geq 20$–30 calibration runs is the entry cost for the full diagnostic.

AI Use Statement

We used generative AI tools (large language models) for **literature search** (identifying related work and verifying bibliographic metadata), **code assistance** (debugging scripts and suggesting unit tests), and **writing assistance** (LaTeX formatting and copy-editing). All scientific claims, experimental designs, statistical analyses, and interpretations are the authors' own; we reviewed all AI-assisted output and take full responsibility for the paper.

Ethics Statement

This work studies the statistical methodology of privacy auditing and involves no human subjects, personal data, or deployment of unlearning systems. The TOFU benchmark uses *synthetic, fictitious* biographical data (no real public figures), and the canaries we embed are synthetic records. The code release contains the analysis and runner scripts plus the per-condition $R \times k$ canary score matrices for the main-table conditions (no corpora, no model weights); the finetuned θ_0 is re-derivable from the public Llama-3.2 base checkpoints and our released finetuning recipe. Our findings strengthen privacy protection by enabling more honest confidence statements. We acknowledge three dual-use vectors (auditor over-claiming if the correction is misapplied; adversarial canary engineering; model-removal attribution) and mitigate them by reporting conservative correction factors and framing the analysis as a diagnostic measurement.

Reproducibility Statement

The experimental protocol is fully specified in Section 4.1, including hyperparameters (Table A1), the θ_0 finetuning recipe, the MIA scoring formula (Eq. 8), and random seed specification. θ_0 is finetuned on the TOFU `full` split with a fixed seed (`seed`=0); as a single shared checkpoint per model, this seed is the relevant detail (it cancels in the run-level ICC). The ICC estimation procedure is in Section 3.5. The main 1B/3B conditions use the TOFU benchmark and the OpenUnlearning framework, pinned at commit `4ad738aaf60f6a4385f6e2506d01da99e76c31f3`; two transfer operators additionally run on MUSE-News (Section 4.3), which does not use OpenUnlearning. A code release accompanies submission as anonymous supplementary material (archive SHA-256 `772abc614796d02ff3d9e8408ac53d81ffc5097d727a8517e9070b61fe6626e0`; host-specific absolute paths and project-identifying strings removed, numeric content unchanged), containing the analysis and runner scripts, the $R \times k$ score matrices for the main-table conditions, and the MUSE-News path; those for the appendix-only extensions are provided as supplementary material. No model weights or datasets. Layout and a full inventory are in the appendix.

References

Adya Agrawal, Yu Wei, Jaspal Singh, Malik Magdon-Ismail, and Vassilis Zikas. Let's ask gauss: Improved one-run privacy auditing, 2026. arXiv:2606.12733.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

Paul D. Bliese. Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In Katherine J. Klein and Steve W. J. Kozlowski (eds.), *Multilevel Theory, Research, and Methods in Organizations: Foundations, Extensions, and New Directions*, pp. 349–381. Jossey-Bass, San Francisco, CA, 2000.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 2021.

A. Colin Cameron and Douglas L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi: 10.3368/jhr.50.2.317.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, pp. 267–284, 2019.

Nicholas Carlini, Christopher A. Choquette-Choo, Matthew Jagielski, et al. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

James M. Cheverud. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87(1):52–58, 2001. doi: 10.1046/j.1365-2540.2001.00901.x.

Mathieu Dagréou and Aurélien Bellet. Detectability in diversity: Improved canary crafting for privacy auditing in one run, 2026. arXiv:2605.27292.

Allan Donner and John J. Koval. The estimation of intraclass correlation in the analysis of family data. *Biometrics*, 36(1):19–25, 1980.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, J. Zico Kolter, Zachary C. Lipton, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025. arXiv:2506.12618.

Nicholas W. Galwey. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 33(7): 559–568, 2009. doi: 10.1002/gepi.20408.

Tomás González Lara, Mateo Dulce Rubio, Aaditya Ramdas, and Mónica Ribero. Sequentially auditing differential privacy. In *Advances in Neural Information Processing Systems*, 2025.

Chuan Guo, Alexandre Sablayrolles, and Maziar Sanjabi. Analyzing privacy leakage in machine learning via multiple hypothesis testing: A lesson from fano. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. arXiv:2210.13662.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 22032–22044, 2020.

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*, pp. 14389–14408, 2023.

Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton, FL, 2000.

Amit Keinan, Moshe Shenfeld, and Katrina Ligett. How well can differential privacy be audited in one run? In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2503.07199.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, 1965.

K. K. Gordon Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983. doi: 10.1093/biomet/70.3.659.

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)*, 2024. arXiv:2403.03218.

Yanhang Li, Zhichao Fan, and Zexin Zhuang. Auditing reasoning-trace memorization claims after unlearning with head-conditioned canaries, 2026. arXiv:2605.18891.

Ruixuan Liu and Li Xiong. UniAud: A unified auditing framework for high auditing power and utility with one training run, 2025. arXiv:2507.04457.

Terrance Liu, Matteo Boglioni, Yiwei Fu, Shengyuan Hu, Pratiksha Thaker, and Zhiwei Steven Wu. Enhancing one-run privacy auditing with quantile regression-based membership inference, 2025. arXiv:2506.15349.

Saeed Mahloujifar, Luca Melis, and Kamalika Chaudhuri. Auditing f-differential privacy in one run. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv:2410.22235.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2401.06121.

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

Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996.

Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations, 2024. arXiv:2411.00640.

Brent R. Moulton. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The Review of Economics and Statistics*, 72(2):334–338, 1990.

Nima Naderloui, Shenao Yan, Binghui Wang, Jie Fu, Wendy Hui Wang, Weiran Liu, and Yuan Hong. Rectifying privacy and efficacy measurements in machine unlearning: A new inference attack perspective. In *Proceedings of the 34th USENIX Security Symposium*, 2025. arXiv:2506.13009.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *USENIX Security Symposium*, pp. 1631–1648, 2023.

Dale R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics*, 74(4):765–769, 2004. doi: 10.1086/383251.

Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, 1979. doi: 10.2307/2530245.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 39643–39666, 2024.

Krishna Pillutla, Galen Andrew, Peter Kairouz, H. Brendan McMahan, Alina Oprea, and Sewoong Oh. Unleashing the power of randomization in auditing differentially private ML. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.18447.

Stephen W. Raudenbush and Anthony S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, Thousand Oaks, CA, 2nd edition, 2002.

Sahasrajit Sarmasarkar, Anastasia Koloskova, and Sanmi Koyejo. Auditing of unlearning algorithms, 2026. arXiv:2607.05898.

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. John Wiley & Sons, New York, 1992.

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. arXiv:2407.06460.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, 2017.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

Tom A. B. Snijders and Roel J. Bosker. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage Publications, London, 2nd edition, 2011.

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

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49268–49280, 2023.

Jialong Sun, Zeming Wei, Jiaxuan Zou, Jiacheng Gong, Jie Fu, Chengyang Dong, Heng Xu, Jialong Li, and Bo Liu. SMI: Statistical membership inference for reliable unlearned model auditing, 2026. arXiv:2602.01150.

Liou Tang, James Joshi, and Ashish Kundu. Behavioral audit of machine unlearning has a privacy cost, 2026. arXiv:2606.14518.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *USENIX Security Symposium*, 2022.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination, and applications. *Annals of Statistics*, 49(3):1736–1754, 2021.

Ruodu Wang. The only admissible way of merging arbitrary e-values. *Biometrika*, 112(2), 2025.

Zihang Xiang, Tianhao Wang, and Di Wang. Privacy audit as bits transmission: (im)possibilities for audit by one run. In *Proceedings of the 34th USENIX Security Symposium*, 2025a. arXiv:2501.17750.

Zihang Xiang, Tianhao Wang, Hanshen Xiao, Yuan Tian, and Di Wang. Tight privacy audit in one run, 2025b. arXiv:2509.08704.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Binchi Zhang, Zihan Chen, Cong Shen, and Jundong Li. Verification of machine unlearning is fragile. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a. ICML 2024; preprint arXiv:2408.00929.

Chenhao Zhang, Muxing Li, Feng Liu, Weitong Chen, and Miao Xu. Unlearning evaluation through subset statistical independence. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2603.00587.

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

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023.

A Limitations

The reading is *correlational, not causal*: ICC(3,1)'s numerator is the between-run contrast ($\mathrm{MS}_R - \mathrm{MS}_E$), so it cannot by itself separate canary-level coupling from a run-level scale shift; the gradient probe is on re-generated checkpoints (an n=1 supporting observation, not causal evidence). All measurements use TOFU (1B/3B) plus a MUSE-News transfer check (one successful GA run, one NPO no-op); WMDP is not evaluated. The two GA cells above the converged k=40 range (0.917 MUSE, 0.747 TOFU-f10) are *post-divergence*, near-saturated measurements; we do not claim a loss→dependence trend (a converged short-budget GA cell reaches 0.713 at 1.94 nats, and within k=400 ICC is flat-to-decreasing across a 4× LR range). GA-3B-f10 remains excluded as a degenerate collapse (run-variance 98.2%; Appendix E). θ_0 is a single checkpoint per scale, except the converged GA-f10 and NPO-f10 conditions, replicated on a second independently finetuned θ_0': neither shows a detectable θ_0 effect at R=30 (GA 0.724 vs. 0.713; NPO 0.101 vs. 0.141; both CIs overlap), but the directional trends are opposite, so any θ_0 effect may be operator-dependent, and the rest of the study remains single-θ_0; because we release no weights, exact reproduction of these replications requires re-running the finetuning, and the θ_0' instance moves the NPO-f10 cell from 0.141 to 0.101 (−28%). The pre-unlearning θ_0 baseline has zero run-variance by construction (a *weak* control, not a causal counterfactual); the NPO-MUSE no-op (ICC 0.004) provides a directional near-zero observation under a did-not-forget condition. The DEFF correction assumes compound symmetry, violated everywhere (Greenhouse–Geisser $\varepsilon \ll 1$), so the scalar ρ averages over the true covariance. Level 1/2 require $R \geq$ 20–30 runs (composition spread ∼10× at R=30, ∼21× at R=10; Appendix E). Llama coverage is 1B/3B, with a single exploratory cross-family arm at 7B (one operator, $R = 10$, confounded with scale and optimizer; Appendix E), so the study's scope remains narrow: one ≥7B family, and MUSE coverage is one scale with one operator. The hierarchy restores validity but not power.

B Experimental Conditions

This appendix gives the per-condition configuration (Table A1), the two-tier framework illustration, the variance decomposition, and the robustness controls referenced from the main text.

C Two-Tier Framework Illustration

Figure 1 illustrates the framework: a score matrix with run-level co-variation (a), the two estimands diverging (b), and the resulting effective-canary-count collapse (c). The figure is presented in the main text (Section 3).

D Variance Decomposition

Table A2 partitions total score variance into entity, within-entity (sample), run, and residual *components*. The Entity column is the canary main effect c_j; ICC(3,1) removes c_j by centering, so that column is *excluded by construction* from the conditional-dependence analysis. To avoid reading the two against each other: this table speaks about *total* score variance,

Table A1: Experimental conditions (TOFU) and operator hyperparameters. θ_0 is finetuned on the TOFU `full` split for 5 epochs, one shared checkpoint per model. NPO and GA use the TOFU experiment defaults (AdamW, lr 1×10^{-5}, batch 8/grad accum 4, 10 epochs); RMU uses the WMDP experiment defaults (AdamW, lr 5×10^{-5}, batch 1/grad accum 16, 80 max steps, constant LR). Full per-operator configuration in Section 4.1.

Condition	Model	k	R	Operator-specific
NPO-1B-f01	1B	40	99	β=0.1, α=1.0
GA-1B-f01	1B	40	30	none (loss negation)
RMU-1B-f01[†]	1B	40	10	steer=2, layer 7
NPO-3B-f01	3B	40	30	β=0.1, α=1.0
GA-3B-f01	3B	40	30	none (loss negation)
RMU-3B-f01	3B	40	30	steer=2, layer 7
NPO-1B-f10	1B	400	30	β=0.1, α=1.0
GA-1B-f10	1B	400	30	none (loss negation)
GA-1B-f10s[‡]	1B	400	30	none; 1 epoch (converged short budget)
RMU-1B-f10	1B	400	30	steer=2, layer 7
NPO-3B-f10	3B	400	27	β=0.1, α=1.0
GA-3B-f10	3B	400	30	none (loss negation)
RMU-3B-f10	3B	400	30	steer=2, layer 7

[†]RMU-1B-f01 runs at $R = 10$ as a lower-precision concordance check; its CIs are wider than the $R \geq 27$ conditions.
[‡]GA-1B-f10s is the converged short-budget cell (1 epoch vs. the 10-epoch default; forget loss 1.94 nats), distinct from the 10-epoch GA-1B-f10.

where the canary/entity main effect dominates at k=400 (NPO 87.4%, RMU > 99%, GA 57.9%); the claim that run-level variation is the *dominant component of the conditional dependence* is a different quantity—the share of run variance within run+residual, which is run-dependent and excluded from the centering. The two statements are not in conflict: the second is computed after the first's dominant term is removed.

Table A2: Variance decomposition of canary MIA scores (% of total). [†]Entity and sample combined (derived from ICC(1,1) and ICC(3,1)).

Condition	Entity	Sample|Entity	Run	Residual
NPO-1B-f01	21.2	73.6	1.3	4.0
GA-1B-f01	20.8	70.8	2.7	5.7
RMU-1B-f01	13.2	86.6	0.02	0.13
NPO-3B-f01	21.4	72.7	0.6	5.3
NPO-1B-f10		87.4[†]	1.8	10.8
GA-1B-f10		57.9[†]	31.4	10.7
RMU-1B-f10		>99[†]	<0.1	<1

E Robustness Controls

k-subsampling. Controlled subsampling of NPO-forget10 and GA-forget10 (fixed 20 entities, $k = 40, 100, 200, 400$) confirms ICC(3,1) is k-invariant: NPO CV = 2.4%, GA CV = 0.43%; DEFF grows as 6.2, 14.9, 28.7, 57.1 for NPO (Pearson $r = 0.99998$). For RMU we apply the same within-forget10 seeded subsample (seed = 42, fixed 20 entities, rows = runs, standard Shrout–Fleiss ICC; $R = 30$): ICC(3,1) across k is 0.1959 ($k = 400$), 0.1884 ($k = 200$), 0.2274 ($k = 100$), 0.1784 ($k = 40$), and ICC(1,1) across k is -0.0006, -0.0032, -0.0072, -0.0235, respectively, stable, with no monotonic k-dependence. RMU conditional ICC(3,1) is thus 0.178–0.227 regardless of k within forget10, confirming k-invariance cleanly (it is consistent with RMU's near-deterministic, low-magnitude run-level behavior). The clean same-entity-count subsample removes the k vs. entity-count confound: the cross-split ICC changes (NPO 0.217 → 0.141 at k=40 → 400; GA 0.306 → 0.747; RMU 0.121 → 0.196) reflect the differing entity-count/composition between `forget01` and `forget10`, not a pure k effect. Under the de-confounded within-forget10 test, all three operators are k-invariant.

Entity-block permutation. This test supplies the paper's only positive evidence that within-run dependence has structure beyond a single global run-level shift, so the scheme is stated in full. Runs (matrix rows) are *preserved*; canaries (columns) are permuted by

`np.random.default_rng(42).permutation(k)`. The null loop slices each shuffled 400-column order into 20 contiguous groups of 20, computes ICC(3,1) separately within each group, and averages the 20 group ICCs. The observed statistic is the *same mean* of ICC(3,1) over the 20 *real* contiguous 20-column entity blocks. No run labels or rows are permuted. The p is the two-sided permutation p, $(2\min(\#\{\text{perm} > \text{obs}\}, \#\{\text{perm} < \text{obs}\}) + 1)/(B+1)$, $B = 1000$, floored at 0.000999.

We report two runs of this *same* statistic at different permutation budgets on the three 1B-f10 conditions ($R = 30$, $k = 400$). At $B = 100$, p can only resolve to $\approx$0.01; at $B = 1000$ (`rng_seed` = 42, p-resolution floor 0.000999) it resolves to $p = 0.000999$. The observed values shift slightly between runs (e.g. GA raw excess 0.016773 at B=100 vs. 0.016449 at B=1000), so the statistic is the same and the permutation budget differs; both are reported. Results at $B = 1000$: global ICC(3,1) is partition-invariant (reassigning canary labels leaves the full-matrix ICC unchanged); the real entity-block mean ICC exceeds the shuffled null for all three operators—NPO 0.288 vs. null mean 0.141 (raw excess 0.147, effect size 35.9, 95% CI [33.9, 37.7]), RMU 0.293 vs. 0.193 (excess 0.100, effect size 39.3, CI [37.2, 41.1]), GA 0.766 vs. 0.749 (excess 0.016, effect size 5.4, CI [3.2, 7.1]); empirical $p = 0.000999$ for all three. The permuted p is the significance statement (a permutation test is floored at $1/B$, not a z-tail); the effect sizes are effect sizes, reported with CIs, not significance levels. The excess is small in absolute terms for GA (0.016 against ICC 0.747) and larger for NPO and RMU. Entity grouping therefore amplifies within-entity dependence consistently across operators, validating the cluster-product correction: NPO naive per-entity sum $k_{\text{eff}} = 68$ vs. shuffled 117; L2 cluster-product $k_{\text{eff}} = 46.6$ accounts for between-entity dependence.

Leave-one-entity-out entity influence. To assess how much each entity contributes to the covariance structure, we recompute ICC(3,1) after removing each of the 20 entities in turn and measure the change in the estimate ($\Delta_e = \text{ICC}_{\text{all}} - \text{ICC}_{-e}$). This is a post-hoc sensitivity analysis of entity contribution to dependence; it is influence-only and does not establish causal effects (which require the intervention of Section 4.3):

Operator	ICC(3,1) all	$\max_e \lvert\Delta_e\rvert$	Conc.	Pattern
NPO	0.141	0.016 (e17)	0.321	Heterogeneous
GA	0.747	0.012 (e10)	0.230	Uniform
RMU	0.196	0.016 (e4)	0.284	Intermediate

Concentration is the fraction of total $|\Delta_e|$ mass carried by the largest $|\Delta_e|$. NPO's dependence is concentrated in a few entities (0.321), consistent with its composition-sensitive coupling; GA's is evenly spread (0.230), consistent with global per-canary gradient alignment; RMU is intermediate (0.284). No single entity dominates any operator ($\max |\Delta_e| \leq 0.016$), confirming the ICC estimate is stable against single-entity removal.

Small-cluster inference at the entity level. The Level-2 entity-cluster correction rests on only $G = 20$ clusters, so we test it against small-G inference variants, reproducing the published Level-2 values first to license the comparison: on the shipped `NPO-1B-f10` matrix (20 entities, 30×400) the published entity-level ICC (0.398895979114) and DEFF_{L2} (8.579023603162) are both recovered to relative error $\sim 10^{-15}$. Every alternative keeps the correction materially above independence: the lower 95% bounds are 5.624936 (CRVE, linearized), 5.418383 (wild-cluster bootstrap, Rademacher), 5.170542 (WCB, Webb), 5.745768 (jackknife CRVE), and 6.638633 (entity cluster bootstrap), against the existing F interval's 6.380442; interval widths run 0.957×–1.315× the existing interval, so no construction approaches DEFF = 1, and the wild-cluster null bootstrap rejects DEFF = 1 at two-sided $p = 0.0002$. One caveat must travel with this: a 2,000-replicate Gaussian additive simulation from the fitted 30×20 structure shows the 20-cluster percentile bootstrap under-covers (0.586) while the existing F interval covers at 1.000 (CRVE 0.915, jackknife 0.902). The bootstrap variants are therefore a robustness display only, not a standalone finite-G validity guarantee; finite-G behavior rests on the F interval, not on the bootstrap.

Entity inclusion/exclusion intervention. To test whether dependence reflects causal interference between audit units, beyond the observational LOEO sensitivity above, we con-

ducted controlled inclusion/exclusion interventions, removing one entity from the forget set while holding all other parameters fixed (NPO, 1B, $R = 30$, paired seeds $\{0, \ldots, 29\}$). We removed two distinct entities: block 4 (Jad Ambrose Al-Shamary, per-entity ICC = 0.124), yielding forget9_lo, and block 7 (Behrouz Rohani, ICC = 0.496), yielding forget9_hi (both $k = 380$). For forget9_lo, the 18 shared retained entities shifted significantly upward (+0.074, $p = 0.011$, Wilcoxon signed-rank; 16/18 positive, 10/18 individually significant at $p < 0.05$). For forget9_hi, the shift was smaller and non-significant (+0.025, $p = 0.061$; 4/18 individually significant). The two interventions produce uncorrelated per-entity shift patterns (Spearman $\rho = -0.009$, $p = 0.97$). This is consistent with a forget-set-composition-driven redistribution of unlearning pressure, removing any entity shifts the remaining canaries' scores upward, with magnitude tracking the composition change, rather than entity-specific local interference, which would instead produce opposed or anti-correlated per-entity patterns. Control-canary anchors confirm the mechanism is operator-specific: under NPO, control (non-forget) canaries shift significantly ($p = 1.2 \times 10^{-6}$ and 1.9×10^{-4} for the two removal targets), establishing global score redistribution. Under GA, the same intervention yields no significant control-canary shift ($p = 0.375$, $R = 30$), a confirmed clean negative. Under RMU, the response is ambiguous (control shift −0.011 on a compressed ∼1.4–2.0 score scale). Distance-decay analysis was not performed (requires a gradient or representation distance matrix).

Random canary subsampling. To test whether the entity structure within TOFU's forget sets drives the ICC estimate, we recompute ICC(3,1) on 100 random subsamples of $k = 200$ canaries drawn without respect to entity boundaries. Subsample ICC matches the full-matrix value within 1.3%: NPO-1B-f10 0.140 ± 0.011 (full 0.141), GA-1B-f10 0.746 ± 0.012 (full 0.747), RMU-3B-f10 0.044 ± 0.005 (full 0.045). Within TOFU, the run-level dependence is consistent with shared parameter perturbations and is invariant to which canaries are subsampled; we do not directly measure the mechanism for NPO/RMU here, so we describe it as a hypothesis. This internal robustness is distinct from cross-benchmark transferability: the MUSE-News checks (Section 4.3) show a *split* outcome, near-zero run-level canary dependence for the NPO no-op (ICC(3,1) = 0.0037) but strong dependence for a successful GA unlearning on the same entity-free benchmark (ICC(3,1) = 0.917), so the dependence is operator-specific rather than gated by TOFU's entity structure.

Hyperparameter sensitivity. A symmetric 3×3 learning-rate sweep on the 1B-f10 conditions ($R = 10$) compares ICC(3,1) at 0.5×, 1×, and 2× the default LR (1×10^{-5}). All Δ values are relative to each operator's own 1× value (standard dose–response reporting):

Operator	0.5×	1×	2×	Δ(0.5×)	Δ(2×)	Pattern
NPO	0.009	0.141	0.178	−94%*	+26%	Monotonic
RMU	0.047	0.196	0.207	−76%*	+6%	Threshold
GA	0.870	0.747	0.693	+16%	−7%	Flat-high

*Falls outside the 1× bootstrap 95% CI. Each marked cell also rejects ICC(3,1)=0 under its own run-effect F-test (df=$(9, 3591)$, $p<10^{-5}$; weakest NPO 0.5×, F=4.58). This is a per-cell run-effect test, not a paired canary-axis bootstrap: the cross-cell comparison remains a CI-overlap check.

Three operators exhibit distinct LR sensitivity patterns. NPO shows a clear monotonic dose–response: ICC collapses to near zero at 0.5× (0.009 $[-0.0002, 0.018]$, well below the 1× CI $[0.071, 0.212]$) and rises at 2× (0.178 $[0.034, 0.332]$, within 1× CI but trending upward). RMU displays threshold behavior: dependence collapses below ∼1× (−76%; 0.047 $[0.013, 0.082]$ falls entirely below the 1× CI $[0.124, 0.255]$) but plateaus above it (+6% at 2×, 0.207 $[0.071, 0.335]$ within 1× CI). GA remains flat-high across the full 4× LR range (ICC = 0.693–0.870, $k_{\text{eff}} = 1.1$–1.4), demonstrating structural coupling invariant to step size. The dose–response taxonomy is: DPO-style operators (NPO) show magnitude-dependent coupling with clear dose–response scaling, activation-steering operators (RMU) exhibit a threshold response (collapse below ∼1×, plateau above), and unconstrained gradient ascent (GA) maintains structural coupling regardless of step size.

Composition sensitivity details. NPO ICC(3,1) across five random 20-entity compositions and the canonical forget10 (1B, $k = 400$). The composition study was run at both $R = 10$ and $R = 30$; we report both, since the extension is what the range below reflects.

Composition	ICC(3,1) ↑		95% CI	DEFF ↑		k_{eff} ↓	
	R=10	R=30	(R=10)	R=10	R=30	R=10	R=30
Seed 333	0.014	0.060	[0.002, 0.031]	6.7	25.1	59.6	15.9
Seed 42	0.016	0.028	[0.003, 0.033]	7.4	12.2	54.0	32.8
Seed 218	0.167	0.114	[0.032, 0.273]	67.5	46.4	5.9	8.6
Seed 99	0.182	0.136	[0.057, 0.296]	73.7	55.2	5.4	7.2
Seed 137	0.304	0.272	[0.158, 0.404]	122.4	109.6	3.3	3.6
forget10 (canon.)	0.141	—	[0.071, 0.212]	57.3	—	7.0	—

DEFF $= 1 + (k - 1)\rho$ with $k = 400$; the R=30 DEFF/k_{eff} follow from the R=30 ICC. CIs: 95% bootstrap over runs (canary-fixed row resampling), $B = 2000$, seed = 42—this convention applies to the composition CIs at both R values; the table's CI column is the R=10 interval, and the R=30 extreme-pair intervals are given in the text below. The R=30 canonical estimate is reported in Table 1.

Five random compositions yield ICC(3,1) ranging 0.028–0.272 at $R = 30$ (at $R = 10$: 0.014–0.304; mean 0.137, SD 0.110), a wide spread driven entirely by the entity composition rather than by k. The canonical forget10 (0.141) falls within the observed range. The spread is $\sim$9.7$\times$ at $R = 30$ (extreme pair: seeds 137 and 42; bootstrap 95% CIs $[0.204, 0.335]$ and $[0.011, 0.049]$, non-overlapping) and $\sim$21$\times$ at $R = 10$ (seeds 137 and 333); both are computed on the same nested run sequence (`run_000`–`009` and `run_000`–`029`) under the same raw ICC(3,1) convention, so the ratio is a point on an R-curve rather than a stable constant. The $R = 10$ range is the inflated one, and it is inflated from the *bottom*: the $R = 10$ minimum was noise-depressed (seed 333 rose 4.2$\times$ and seed 42 1.75$\times$ from $R = 10$ to $R = 30$), while the maximum barely moved (-11%). This $\sim$10$\times$ spread is a central uncertainty in any single-composition measurement, and we report it as the headline composition-dependent range rather than a point value.

The five compositions are independent random draws of 20 complete entities from a 40-entity pool (TOFU `forget10`'s 20 plus `holdout10`'s 20), so any two share $\sim$50% of their entities by construction; the draws themselves are independent, and this overlap is arithmetic. No pair reaches 80% entity overlap, so none qualifies for calibration reuse under the per-family rule of Section 3.5; these five calibrations are therefore five independent draws, not reuses of one another.

Canary budget: runs dominate canary count for DEFF precision. We ran 1,000 Monte Carlo simulations across four ICC levels (0.016, 0.141, 0.182, 0.304), four canary counts (k = 50, 100, 200, 400), and four run counts (R = 5, 10, 20, 30). The ICC(3,1) estimator is unbiased (bias < 0.001) but imprecise at small R. **Run count dominates canary count**: increasing k from 50 to 400 barely affects precision, while increasing R from 5 to 30 halves relative error:

ICC	$R = 5$	$R = 10$	$R = 20$	$R = 30$
All levels	47–73%	31–48%	22–33%	17–26%

Based on simulation, $R = 30$ yields $< 27\%$ relative error (and $R = 20$ reaches 33%). As an illustrative robustness check, computing ICC on incrementally larger prefixes of the NPO-1B-f10 $R = 30$ batch yields:

R	ICC(3,1)	95% CI	CI half-width
10	0.078	[0.040, 0.135]	0.047
20	0.176	[0.067, 0.271]	0.102
30	0.141	[0.071, 0.212]	0.071

The ICC point estimate varies 0.078–0.176 across $R = 10/20/30$ prefixes, with all CIs containing the $R = 30$ value. $R \geq 20$ yields CI half-width 0.07–0.10, sufficient to detect non-zero dependence but not to pin exact magnitude. We recommend $R = 30$ when point-estimate magnitude matters.

Compute budget and software environment. All experiments ran on a single NVIDIA RTX 4090 GPU. Total compute: ∼28 GPU-hours (8.5h for 1B conditions, ∼20h for 3B conditions). Software: Python 3.12, PyTorch 2.8.0+cu128, accelerate 0.34.2, transformers 4.51.3, bitsandbytes 0.44.1, deepspeed 0.15.4, OpenUnlearning, TOFU benchmark (`locuslab/TOFU`). Seeds $\{0, \ldots, R-1\}$ per condition. The code release accompanying submission (archive SHA-256 `772abc614796d02ff3d9e8408ac53d81ffc5097d727a8517e9070b61fe6626e0`; host-specific absolute paths and project-identifying strings removed, numeric content unchanged) contains the analysis and runner scripts plus the $R \times k$ score matrices for the main-table conditions, including the converged short-budget GA cell's matrix and its driver script; the composition matrices are shipped at *both* R values in separate directories (`release/matrices/` for R=10 and `release/matrices_r30/` for R=30, in `.npy` and `.csv`), and the analysis outputs backing the reported composition and transfer statistics are included (`release/analysis/`); no model weights or datasets are included. The appendix-only extensions added after that archive was frozen are *not* in it, and their score matrices are provided as supplementary material. The scripts cover ICC estimation, rank-transform, variance decomposition, DEFF correction, e-values, Greenhouse–Geisser ε, permutation, k-subsampling, and a full-analysis driver (`release/scripts/`), the TOFU runner (`release/tofu/`), and a separate MUSE-News path (per-article NLL scoring, matrix assembly, ICC; `release/muse/`). The θ_0 fine-tuning recipe and all per-condition hyperparameters are specified in Table A1. The OpenUnlearning framework is pinned at commit `4ad738aaf60f6a4385f6e2506d01da99e76c31f3` ("Fix pip install command for lm-eval"). The v8 archive (177 members) adds the entity-block mapping file (`release/matrices/tofu_forget10_entity_mapping.json`) and the coverage figure generator (`release/scripts/make_fig4_power_coverage.py`) relative to the originally cited v5.

Cross-benchmark transfer (MUSE-News). To test whether the dependence is operator-general or specific to TOFU's entity-structured forget sets, we ran two operators on MUSE-News (Shi et al., 2025), a topic-based (non-entity) benchmark, with the same repeated-run protocol. The two operators give opposite results. *(i) NPO no-op.* With $R = 10$, run-level dependence is effectively absent: ICC(3,1) = 0.0037 (bootstrap 95% CI $[0.0001, 0.0094]$) and ICC(1,1) = −0.0011, versus the TOFU forget10 NPO anchor 0.141 (CI $[0.071, 0.212]$); the two intervals do not overlap. On MUSE the NPO operator *did not forget* (mean forget-set loss 2.410 *below* retain-set loss 2.786; gap −0.376), so this near-zero ICC is scoped to a did-not-forget (no-op) condition. *(ii) GA, successful forgetting.* Gradient ascent *did* forget on MUSE (per-run forget-set loss above holdout; gate PASS on 30/30 runs) and shows strong run-level dependence: ICC(3,1) = 0.917 (bootstrap 95% CI $[0.814, 0.951]$) with $\text{DEFF}_c = 814.6$ ($k_{\text{eff}} = 1.09$). Its ICC(1,1) is much lower (0.075 $[0.021, 0.130]$), so the two tiers again diverge. *Reading.* Because the same entity-free benchmark yields near-zero dependence for a no-op NPO run and strong dependence for a successful GA run, the dependence is *operator-specific*: entity structure is not required for it to manifest, and its absence for NPO is confounded with the operator's failure to forget.

Degeneracy and saturation census. A high conditional ICC could in principle be a near-deterministic collapse rather than genuine dependence. Our degeneracy criterion is the *run-variance fraction* ($\geq 98\%$), chosen because it directly targets the failure mode of interest: a near rank-1, row-constant matrix in which the *canary* dimension has collapsed, so the few remaining degrees of freedom are all run-level. We test it symmetrically on every matrix (Table A3). GA-3B-f10 triggers it (run-variance 98.2%, canary-variance 0.61%): its canaries collapse to near-constant, which is why it is excluded. The MUSE-GA matrix does *not*: run-variance is 7.4% and canary-variance 91.9% (canary SD 1.87), i.e. the canary dimension retains full information, so its ICC reflects near-perfect run-to-run synchronization of otherwise-distinct canaries—the phenomenon the framework measures, not a rank-1 collapse. We also report the *residual fraction* (SS_E/total) for every matrix, and it is exactly

why we do *not* use it as the degeneracy criterion: it is decoupled from ICC. Table A3 lists the census. RMU-1B-f10 has residual 0.73%, *lower* than the excluded GA-3B-f10 (1.21%), yet an ICC of only 0.196; RMU-3B-f10 has residual 0.19% (six times below GA-3B-f10) yet the *lowest* ICC in the study (0.045); MUSE-GA has residual 0.67% and ICC 0.917. A low residual therefore cannot indicate degeneracy—it is also what a tight, well-separated, low-noise matrix looks like, and it would misclassify several legitimate cells (including the lowest-ICC one). What distinguishes a collapse is where the variance *goes*: in GA-3B-f10 the run dimension absorbs 98.18% and the canary dimension keeps 0.61%, i.e. the canaries have collapsed into the run direction and carry almost no information. RMU's low residual instead reflects tightly clustered but *independent* canaries, with a run-variance of only 0.02–0.18%. The run-variance criterion separates these two cases; the residual criterion does not. Accordingly, 0.917 is downgraded *only because it sits in a saturated regime*, not because of any criterion choice. Two caveats bound the MUSE-GA value. First, both this matrix and the TOFU-f10 GA reference sit in *post-divergence* regimes, not converged ones: pure gradient ascent maximizes the forget loss with no retain term or clipping, and at 10 epochs with constant LR the loss diverges, reaching ~ 105 nats for MUSE-GA and ~ 77 nats for TOFU-f10 versus ~ 2.3 nats for the converged k=40 (f01) conditions (NPO 2.32, GA 2.46, RMU 1.33). The MUSE regime is the *more* saturated of the two (residual variance 0.67% vs. 10.4%), and this near-zero residual is the mechanical driver of ICC(3,1)$\to 1$ ($k_{\text{eff}} = 1.09$). So the cross-benchmark gap (0.917 vs. 0.747) compares two *different degrees of post-divergence saturation*, not a saturated-versus-converged contrast. We do *not* claim the dependence grows with loss: within $k = 400$ at fixed dataset, ICC is high but mildly *decreasing* across a 4× learning-rate range ($0.870 \to 0.747 \to 0.693$ as loss rises $26.8 \to 77.4 \to 80.9$ nats, Table A3)—the opposite of a saturation→dependence trend. GA's strong coupling therefore does not require extreme saturation, and the 0.917 and 0.747 values remain scoped to their own regimes. Second, the gate margin is narrow (+1.74 on ∼103, 1.7%): the direction is unambiguous, but it rests on a small separation on high-loss runs. *Provenance of the two numbers.* The gate margin (104.57 vs. 102.83) comes from a standalone MIA scorer over a 353/888 finite forget subset; the dependency estimate (0.917, $k = 888/889$) comes from the native evaluation pipeline, which drops only the one article (821) that is non-finite in every run. The two pipelines differ in their finite subsets (353 vs. 888) and must not be read as one common sample. Scoping: the transfer is conditional on a successful unlearning (gate PASS), the 30-run design, the single 1B model, and the finite MUSE subset—it is not evidence of universal transfer across operators or benchmarks.

Cross-family transfer (≥7B). To test whether the operator×k dependence is confined to the Llama-3.2 cohort, we ran NPO on TOFU forget10 with *Qwen2.5-7B-Instruct*, a different architecture family, at $R = 10$, $k = 400$: conditional ICC(3,1) = 0.0619, DEFF_c = 25.70, against the 1B NPO-f10 anchor's 0.141 / 57.08 (an ICC ratio of 0.44× and a DEFF_c ratio of 0.45×; the two differ because DEFF is affine, not proportional, in ICC). This cell therefore sits *below* the 1B anchor rather than above it: it does not evidence elevated dependence in a different architecture family. Three axes move together here—model family (Qwen2.5 vs. Llama), scale (1B→7B), and optimizer (`paged_adamw_8bit`, the memory fix that let 7B fit, vs. `adamw_torch`)—so no one of them can be isolated, and with the corrected value the exploratory arm is *inconclusive*: it neither reproduces the Llama-cohort level of dependence nor establishes a cross-family contrast. It is also *not* a scale law: the paper's own NPO 1B→3B conditional ICC is essentially flat (0.141 → 0.133, Table 1), so a 1B→7B difference cannot be attributed to scale on this design. The cell is not a degenerate collapse by the paper's primary criterion (run-variance 0.91% ≪ 98%; SS shares canary 85.8% / run 0.91% / residual 13.28%). Precision at $R = 10$ is limited, so the point estimate must not travel alone, and any interval must be labelled by axis: the canary-axis bootstrap interval recomputed under the paper's orientation is $[0.046, 0.084]$ (resampling canaries, $B = 2000$, canary-sampling uncertainty). No run-axis interval is reported: at $R = 10$ a percentile interval on so few runs is not reliable. The corrected point estimate is therefore reported as a single-instance value, and no interval is paired with the 1B anchor's. Forgetting is measured on the forget side, not gated: the 7B shim suppresses model saving by design, so no post-unlearning checkpoint survives and the forget/holdout gate cannot be produced retroactively; a zero-cost join of each run's own reference cache and per-token-mean scores instead gives a paired trained-minus-θ_0 forget gap of +2.826 nats (~ 56 SE), positive in

Table A3: Variance census for the GA matrices, with *both* degeneracy criteria reported. Primary = run-variance fraction ($\geq$ 98%, target: near rank-1 row-constant collapse in which the canary dimension dies); secondary = residual floor (SS_E/total), reported for transparency. The two disagree on the GA cells (residual 1.21% for degenerate GA-3B-f10 vs. 0.67% for non-degenerate MUSE-GA, i.e. the residual criterion reverses), which is why run-variance is primary: a low residual also occurs in a high-signal, low-noise non-degenerate matrix (MUSE-GA's canaries still vary by SD 1.87). MUSE-GA is non-degenerate (canary fraction 91.9%) but post-divergence ($\sim$105 nats); GA-3B-f10 is degenerate (canary 0.61%) and excluded. Cross-cell ICC comparisons are across saturation degree. The final two rows are θ_0-robustness *replications* on an independently finetuned second base (θ_0'), not new experimental conditions. GA: 0.724 vs. 0.713 (run-axis CIs $\sim$0.084 wide, overlapping almost entirely). NPO: 0.101 vs. anchor 0.141 ($\Delta = -0.040$; CI $[0.050, 0.147]$ overlaps the anchor's $[0.071, 0.212]$). Both are power-limited nulls at R=30—no detectable θ_0 effect—not evidence of θ_0-invariance; the directional trends are opposite (θ_0' lower for NPO, higher for GA), suggesting any θ_0 effect, if present, is operator-dependent. All shares are sum-of-squares fractions; the value-range column is the per-canary evaluation-score range (not the training loss).

Matrix	value range (mean)	SS_R/tot	SS_C/tot	SS_E/tot	run-var $\geq$ 98%?	ICC(3,1)
MUSE GA (30×888)	85.0–110.2 (105.4)	7.39%	91.94%	0.67%	No (non-deg.)	0.917
GA-1B-f10 (30×400)	68.2–85.5 (77.4)	30.82%	58.75%	10.43%	No	0.747
NPO-1B-f10 (30×400)	1.4–8.3 (3.1)	1.95%	86.31%	11.74%	No	0.141
RMU-1B-f10 (30×400)	0.1–4.4 (2.0)	0.18%	99.09%	0.73%	No	0.196
RMU-3B-f10 (30×400)	– (1.5)	0.009%	99.80%	0.19%	No	0.045
GA-3B-f10 (30×400)	59.9–113.6 (93.4)	98.18%	0.61%	1.21%	**Yes (degen.)**	0.988
GA 0.5×LR (10×400)	– (26.8)	30.70%	64.73%	4.57%	No	0.870
GA 1×LR (30×400)	– (77.4)	30.82%	58.75%	10.43%	No	0.747
GA 2×LR (10×400)	– (80.9)	22.12%	68.13%	9.75%	No	0.693
GA conv. short (30×400)	0.0–5.8 (1.9)	11.04%	84.54%	4.42%	No	0.713
GA conv. short, θ_0' (30×400)	0.0–5.9 (2.1)	12.90%	82.19%	4.91%	No	0.724
NPO-f10, θ_0' (30×400)	– (3.1)	1.22%	88.18%	10.60%	No	0.101

Table A4: MUSE-News transfer check vs. TOFU-f10 GA reference (1B). Matrices are $R \times k$ with runs as rows; ICC values are the pre-registered conditional-interdependence measure. The NPO row is a did-not-forget no-op (different operator + state) and does not constrain GA. The GA-MUSE value is non-degenerate (run-variance 7.4%, well below the 98% threshold; Table A3) but post-divergence ($\sim$105 nats, more saturated than the TOFU-f10 reference at $\sim$77), so the comparison is across saturation degree. Distinct provenances: GA-MUSE gate margin (104.57 vs. 102.83) is a standalone-scorer figure over a 353/888 finite forget subset, while the ICC (k = 888/889) is the native-eval pipeline; the two must not be read as one sample. GA-MUSE uses the finite 30×888 matrix (article 821 dropped); its gate passed on 30/30 runs.

Condition	Forget?	k	R	ICC(1,1)↑	ICC(3,1) [95% CI]↑	$DEFF_c$↑	k_{eff}↓
NPO-MUSE (no-op)	no	888	10	−0.001	0.0037 [0.0001, 0.0094]	4.3	208
GA-MUSE (successful)	yes (30/30)	888	30	0.075	0.917 [0.814, 0.951]	814.6	1.09
GA-TOFU-f10 (reference)	yes	400	30	0.314	0.747 [0.520, 0.833]	298.9	1.3

4000/4000 per-example pairs across all ten seeds. This confirms the operator ran and moved the forget set, but it does not rescue the dependence claim: with the corrected ICC the arm is exploratory and inconclusive, and it is superseded as a scope statement by the operator-specific MUSE-News transfer check above. Scope: one $\geq$7B family, one operator, $R = 10$. The originally reported ICC(3,1) of 0.851 for this cell resulted from applying the ICC estimator to a transposed score matrix ($k \times R$ instead of $R \times k$); after correction the value is 0.0619. All other cells were re-verified against the R-rows × k-columns orientation convention documented in the archive README; no other transposition was found.

Gradient mechanism verification. To quantify the mechanism behind the operator ordering, we measured the per-canary forget-gradient pairwise cosines (mean off-diagonal;

forget10, 1B, 30 canaries, 100 retain, layers 14/15, same methodology). For gradient ascent, the per-canary forget-gradient pairwise cosine is 0.618 (870 pairs, std 0.068), measured on the regenerated gradient-ascent probe checkpoint, a strong alignment, consistent with GA's high conditional ICC (0.747). For NPO, the corresponding value is 0.0858 (std 0.037, 870 pairs, grad norm 7.75), measured on the non-degenerate NPO probe checkpoint, a weak alignment, consistent with NPO's moderate conditional ICC (0.141). Both probe objects are *re-generated / separately-trained* checkpoints, distinct from the operating-point weights used to compute the ICC values; the mechanism was measured on these regenerated instances. This supports the mechanism link between per-canary gradient alignment and run-level dependence across operators. The NPO value is a single-checkpoint measurement (seed 42, 30 canaries sampled), so it is a single-instance value rather than a law. We note that NPO's forget-gradient norm saturates toward ≈ 0 at full convergence (a DPO-objective property), a secondary mechanism observation that is not a comparable cosine; a collapsed NPO probe cell (grad norm $\approx 10^{-15}$) is unusable for this purpose. We caution that the NPO measurement file also carries a "ga" block whose value coincides (0.0858); that is a near-tautology computed on the NPO model and is *not* GA's value, GA's true 0.618 is taken from a separate gradient-ascent measurement. The aggregate forget-vs-retain gradient cosine (-0.866 at the trained point or -0.02 at step-0) is a different quantity and is not the mechanism value. For RMU, a regenerated RMU probe checkpoint was produced but no probe value was obtained, so the RMU mechanism is inferred from the ICC rather than directly measured.

F Sequential Hypothesis Test

We test $H_0: \ \mathrm{ICC}(3,1) = 0$ on the primary NPO-1B-forget01 condition with a group-sequential analysis that uses Lan–DeMets α-spending (Lan & DeMets, 1983) and O'Brien–Fleming boundaries (O'Brien & Fleming, 1979) at three pre-specified looks ($R = 30$, 60, 99; cumulative α spending 0.0013, 0.0169, 0.0492), following the standard group-sequential framework of Jennison & Turnbull (2000). We monitor a *single* statistic throughout, the two-way ICC(3,1), so the procedure is a valid α-spending sequential test, not a mixed set of tests. At the terminal sample ($R = 99$) the two-way F-test gives $F(98, 3822) = 12.07$ ($p = 6.7 \times 10^{-158}$, one-sided), and the bootstrap z-transform (defined below) gives $z = 8.65$; these are two equivalent scales for the same null, the p is computed from the F statistic under its exact degrees of freedom, and z is the ICC-to-SE-bootstrap ratio, both indicating the same rejection.

The z-statistic for the two-way ICC(3,1) at each look is $z_i = \widehat{\mathrm{ICC}}(3,1)_i / \widehat{\mathrm{SE}}_{\mathrm{boot}}$, where the standard error is the bootstrap SD over within-run row resampling (canaries fixed, $B = 20000$, seed = 20260825); this reproduces the reported $z = 8.65$ at the third look ($\widehat{\mathrm{SE}}_{\mathrm{boot}} = 0.0250$). Table A5 reports the trajectory.

Table A5: Group-sequential test of H_0: $\mathrm{ICC}(3,1) = 0$ for NPO-1B-forget01, monitoring the two-way ICC(3,1) at every look. The two-way test crosses its O'Brien–Fleming boundary at every look, including the first ($z = 4.44 > c_1 = 3.00$), so, by the pre-registered spending rule, the null would have been rejected at $R = 30$ and the sequence could have stopped. The marginal ICC(1,1) is negative throughout and its z is far below the boundary, confirming that a marginal diagnostic cannot detect the conditional dependence.

Look	R	ICC(1,1)	ICC(3,1)	z	Boundary / Decision
1	30	−0.013	0.212 [.116, .300]	4.44	$c_1 = 3.00$ / **reject**
2	60	−0.012	0.222 [.157, .287]	6.65	$c_2 = 2.14$ / **reject**
3	99	−0.013	0.217 [.168, .265]	8.65	$c_3 = 1.72$ / **reject**

If the pre-specified stopping rule were applied strictly, the sequence would have terminated at $R = 30$; we continued to $R = 99$ to narrow the confidence interval, so the reported $z = 8.65$ is the fixed-sample value at the terminal sample.

The trajectory is consistent with the two-tier gap: the conditional ICC(3,1) z exceeds its O'Brien–Fleming boundary at every look, while the marginal ICC(1,1) remains negative (its one-way z is -0.389–-0.774, far below the boundary), so dependence is detected only by the conditional statistic. The final $\mathrm{ICC}(3,1) = 0.217$ $[0.168, 0.265]$ (bootstrap 95% CI) yields DEFF ≈ 9.5, implying $k_{\mathrm{eff}} \approx 4.2$ out of 40 canaries.

G Effective Sample Size Sensitivity Analysis

Table A6 compares effective sample size estimates under different pooling assumptions for the forget10 ($k = 400$) conditions.

Table A6: Sensitivity of k_{eff} to pooling method (forget10, $k = 400$). The naive per-entity sum assumes entity independence; random-effects (RE) pooling accounts for ICC heterogeneity; the cluster-product additionally accounts for between-entity marginal dependence. For NPO, $I^2 = 0.66$ produces a wide RE prediction interval for ICC(3,1) $\in [0.084, 0.474]$, confirming that entity-level heterogeneity is real and that entity-specific corrections may be warranted.

Method	NPO k_{eff}	GA k_{eff}	Assumption
Naive per-entity sum	68.2	25.9	Fixed-effect, entities independent
RE pooled	63.5 [60.6, 66.5]	25.3	Accounts for ICC heterogeneity
Cluster-product (L2)	46.6	20.7	Between-entity dependence
Global DEFF (L1)	7.0	1.3	All canaries exchangeable
E-value average (L0)	$\approx$1	$\approx$1	Arbitrary dependence

For GA, the three entity-aware methods converge ($k_{\text{eff}} = 20.7$–25.9) due to lower heterogeneity ($I^2 = 0.36$). For NPO, the spread is wider (46.6–68.2), reflecting the greater entity-level variability ($I^2 = 0.66$). The cluster-product provides the most conservative entity-aware estimate and is recommended as the primary bound for single-run audits.

H ICC Estimation and Numerical Handling

We estimate all ICC values with standard Shrout–Fleiss ANOVA mean-square estimators (Case 1 for ICC(1,1), Case 2/3 for ICC(2,1)/ICC(3,1); Section 3.2). Bootstrap 95% CIs use within-run *row* resampling (canaries fixed) throughout, but the resample count and seed differ by analysis, so we name each convention rather than declaring one default: $B = 2000$, seed = 42 for the composition study and the 3B tables; $B = 2000$, seed = 42 for most learning-rate cells, with two cells at $B = 1000$, seed = 42; $B = 10000$, seed = 12345 for the R-stability prefixes; $B = 2000$, seeds 20260901 and 20260908 for the coverage analyses (Table A13, Table A14); and $B = 20000$, seed = 20260825 for the sequential analysis (Appendix F). All 3B conditions use parametric (asymptotic) CIs rather than the row-resampling bootstrap. For the 3B k=400 cells the one-way ANOVA is numerically degenerate and the bootstrap is unstable, so parametric intervals are required; for the structurally 1B-f01-identical 3B k=40 cells (same k, same R) parametric intervals are used for within-scale consistency, not because the bootstrap fails there. Cross-scale CI widths are therefore not directly comparable (Section 4.3).

Invariance to additive shifts. A property of the estimator that matters for interpreting any of the ICC values above: because ICC(3,1) is computed on canary-centered residuals, it is *exactly invariant to any per-canary additive constant*—adding t_c to every score of canary c leaves SS_{row} and SS_{error} unchanged (the contributions to SS_{total} and SS_{col} cancel), so the estimate is bit-identical. This invariance is specific to *additive* shifts: per-canary *scaling* does not share it (a scaling changes both SS_{row} and SS_{error}, so only a global rescaling of all scores leaves ICC(3,1) unchanged). The invariance also does *not* extend to per-run (row) shifts, where MS_R changes via $\text{var}(m_r) + \text{var}(d_r) - 2\,\text{cov}(m_r, d_r)$—which is why the framework requires the $R \geq 20$–30 calibration phase rather than a single reference run.

Non-finite handling. A canary that produces a non-finite score (NaN or $\pm\infty$ loss ratio, e.g. from a degenerate tokenization) is excluded from the matrix before estimation; MUSE-News excluded 1 of 889 canaries this way (Appendix E). *Degenerate runs.* A run whose unlearning collapsed to a near-deterministic state (between-run variance $\geq 98\%$ of total) is flagged and excluded from the non-degenerate ranges and headline k_{eff} claims; the worst case (GA-3B-f10, ICC(3,1)= 0.988) is reported separately in the text and Appendix K rather than pooled. *One-way degeneracy.* ICC(1,1) is not computed where the one-way ANOVA is numerically degenerate (the $k = 400$ 3B conditions); these cells are marked "—" in Table 1. *Negative*

design effects. A negative ICC(1,1) yields $\text{DEFF}_m < 1$, which is not a meaningful correction; we truncate marginal DEFF to 1.0 (Table 1, Section 3.3). These rules are applied uniformly across all conditions.

I Compound Symmetry Diagnostics

Table A7 reports Greenhouse–Geisser ε (centered) across all conditions. All $\varepsilon \ll 1$, confirming compound symmetry is violated throughout. The eigenvalue-based k_{eff} and the scalar-DEFF k_{eff} disagree (Table A7 reports their ratio = DEFF k_{eff}/eigenvalue k_{eff}). In the small-k (f01) conditions, RMU-1B-f01 (ratio 2.68), RMU-3B-f01 (2.19), GA-3B-f01 (1.95), GA-1B-f01 (1.17), NPO-3B-f01 (1.31), the ratio exceeds 1, i.e. the scalar-DEFF k_{eff} *exceeds* the eigenvalue k_{eff}: the scalar ρ (a single weighted average of a far-from-compound-symmetric covariance) *overstates* k_{eff}, so the DEFF correction is over-optimistic (it under-corrects). Because a higher k_{eff} means less dependence, and thus a less conservative bound, the eigenvalue-based k_{eff} is the more accurate and more conservative estimate in these cells, and we adopt it as the *primary* effective-sample-size bound there (per the spectral effective-number-of-independent-tests approach of Cheverud (Cheverud, 2001), Nyholt (Nyholt, 2004), and Galwey (Galwey, 2009), adapted from genetic association testing to canary-score covariance). Table A7 now discloses per cell which k_{eff} is primary. The rule is a strict inequality with the boundary at 1: where the ratio exceeds 1 the scalar overstates k_{eff} and the eigenvalue bound is primary; where the ratio is at most 1 the scalar is primary. The f10 conditions are therefore mixed, not uniformly scalar. Four f10 cells sit at or below the boundary and keep the scalar primary: NPO-1B-f10 (0.92), GA-1B-f10 (0.86), RMU-1B-f10 (0.63), and NPO-3B-f10 at exactly 1.00—which the non-strict rule assigns to the scalar, so that cell's k_{eff} is 7.41 rather than the eigenvalue 7.40. Three f10 cells exceed the boundary and take the eigenvalue bound: RMU-3B-f10 (1.10; eigenvalue 19.12 rather than the scalar 21.05), GA-1B-f10s (1.19; eigenvalue 1.18 rather than the scalar 1.40), and GA-3B-f10 (1.01, a near-tie in the degenerate regime, where both estimators give $k_{\text{eff}} \approx 1$ and the $k_{\text{eff}} \approx 1$ headline floors remain consistent). The rule is *sensitive to near-floor values*: GA-1B-f10 ($\varepsilon = 0.005$, ratio 0.86) and GA-1B-f10s ($\varepsilon = 0.004$, ratio 1.19) have near-identical GG ε yet receive opposite designations, so an auditor should read the designation as indicative rather than decisive when k_{eff} is near 1. Monte Carlo coverage simulation brings DEFF-corrected coverage to ≈ 0.94 (well-calibrated to the nominal level), and the e-value floor is conservative at ≥ 0.999. This calibration holds on the *synthetic compound-symmetric benchmark* (Table A13, left); on the shipped cells the correctly scaled Level-1 correction reaches at least nominal coverage, including the converged GA cell (1.000; Appendix O), so the auditor proceeds, and where the scalar overstates k_{eff} the eigenvalue bound is the safer choice.

Spectral effective sample size: rationale and scope. To consolidate: the eigenvalue-based $\widehat{k}_{\text{eff}}$ (the Cheverud–Nyholt–Galwey spectral effective number of independent tests, adapted from the covariance eigenspectrum of canary scores) is adopted as the *primary* effective-sample-size bound wherever the scalar DEFF overstates k_{eff} (ratio > 1 in Table A7). This is a conservative choice: the eigenvalue bound is at most equal to the scalar k_{eff} in those cells, so it never loosens the correction. The two-estimator rule (scalar primary where ratio ≤ 1; eigenvalue primary where ratio > 1) is a transparent, closed-form compromise that requires only the score covariance matrix and no iterative fitting. A fully Bayesian hierarchical model—e.g., a mixed-effects ICC with per-entity random slopes estimated via MCMC—could in principle provide a unified k_{eff} that avoids the discrete switch, but it would introduce prior sensitivity, convergence diagnostics, and a substantial computational overhead, placing it beyond the scope of this frequentist correction framework.

J Product E-values: Null Behaviour and Scope

This section reports what the product e-value statistic does and does not establish. The shipped `compute_evalues()` row-centers the score matrix and then measures cross-run agreement of the canary pattern; as a statistic of that construction it rejects the iid null. Under a *no-run-dependence* null—fixed per-canary heterogeneity plus iid noise—it also rejects,

Table A7: Compound symmetry diagnostics. GG ε (centered) measures departure from sphericity; $\varepsilon = 1$ is perfect CS. Eigenvalue k_{eff} is computed from the score covariance eigenspectrum. Ratio = DEFF k_{eff}/eigenvalue k_{eff}; the rule is a strict inequality: where ratio > 1 the scalar-DEFF k_{eff} *exceeds* (overstates) the eigenvalue k_{eff}, so the eigenvalue estimate is the more conservative and is the *primary* bound (the scalar is over-optimistic there); where ratio ≤ 1 the scalar is primary. NPO-3B-f10 sits exactly at the 1.00 boundary and keeps the scalar. The L2 column is the entity-cluster product computed on each cell's own matrix ("—" where the cell has no entity map); GA-1B-f10s* is the converged short-budget cell, whose own L2 value (21.3) is used in Appendix O and differs from the canonical GA-1B-f10 cell's 20.7.

Condition	GG $\varepsilon\downarrow$	Eigenval. k_{eff}	DEFF k_{eff} (L1)	Ratio	Primary k_{eff}	L2 k_{eff}
NPO-1B-f01	0.159	4.68	4.23	0.90	scalar	24.4
GA-1B-f01	0.095	2.64	3.09	1.17	eigenvalue	22.4
RMU-1B-f01	0.067	2.60	6.97	2.68	eigenvalue	47.0
NPO-3B-f01	0.228	8.01	10.50	1.31	eigenvalue	—
GA-3B-f01	0.157	5.55	10.81	1.95	eigenvalue	—
RMU-3B-f01	0.060	2.34	5.13	2.19	eigenvalue	—
NPO-1B-f10	0.024	7.60	7.01	0.92	scalar	46.6
GA-1B-f10	0.005	1.55	1.34	0.86	scalar	20.7
GA-1B-f10s*	0.004	1.18	1.40	1.19	eigenvalue	21.3
RMU-1B-f10	0.040	7.98	5.05	0.63	scalar	34.3
NPO-3B-f10	0.020	7.40	7.41	1.00	scalar	—
GA-3B-f10	0.003	1.00	1.01	1.01	eigenvalue	—
RMU-3B-f10	0.048	19.12	21.05	1.10	eigenvalue	—

so the statistic detects canary heterogeneity rather than run dependence. **It therefore does not by itself establish run-level dependence.** Table A8 is retained as a *descriptive* record of when that threshold is crossed per condition: the statistic crosses $\alpha = 0.05$ within 8–15 calibration runs across all conditions—8–9 for the non-degenerate set (9 for GA-1B-f10 and RMU-3B-f01), with the single degenerate GA-3B-f10 cell the outlier at 15. These are threshold-crossing counts for this statistic, *not* an independence test, and they are not evidence of run dependence. The valid-confidence statement is the DEFF-corrected coverage result of Section 4.4.

Table A8: Product e-values, descriptive record: the run at which this statistic crosses the $\alpha = 0.05$ threshold ($e > 20$) for each condition. This statistic detects canary heterogeneity rather than run dependence (see the section text); the table records threshold-crossing runs, not an independence test.

Condition	Final e-value	Reject H_0?	First rejection run
NPO-1B-f01	5.84×10^{20}	Yes	8
GA-1B-f01	1.20×10^{6}	Yes	8
RMU-1B-f01	87.2	Yes	8
NPO-1B-f10	7.26×10^{5}	Yes	8
GA-1B-f10	4.06×10^{5}	Yes	9
GA-1B-f10s	1.49×10^{6}	Yes	8
RMU-1B-f10	1.86×10^{6}	Yes	8
NPO-3B-f01	1.17×10^{6}	Yes	8
GA-3B-f01	9.82×10^{5}	Yes	8
RMU-3B-f01	50.1	Yes	9
NPO-3B-f10	3.05×10^{5}	Yes	8
GA-3B-f10	1.41×10^{3}	Yes	15
RMU-3B-f10	1.93×10^{6}	Yes	8

For this statistic, as few as 8–9 calibration runs cross the threshold across most conditions; GA-3B-f10 requires 15 (vs. 8–9 for all others). This is a property of the statistic's sensitivity to canary heterogeneity, not a dependence-diagnosis budget.

K Interaction Table

This section reports the scale×k interaction (Figure A1, Table A9, Figure A2), the DEFF structure behind the effective-sample-size collapse.

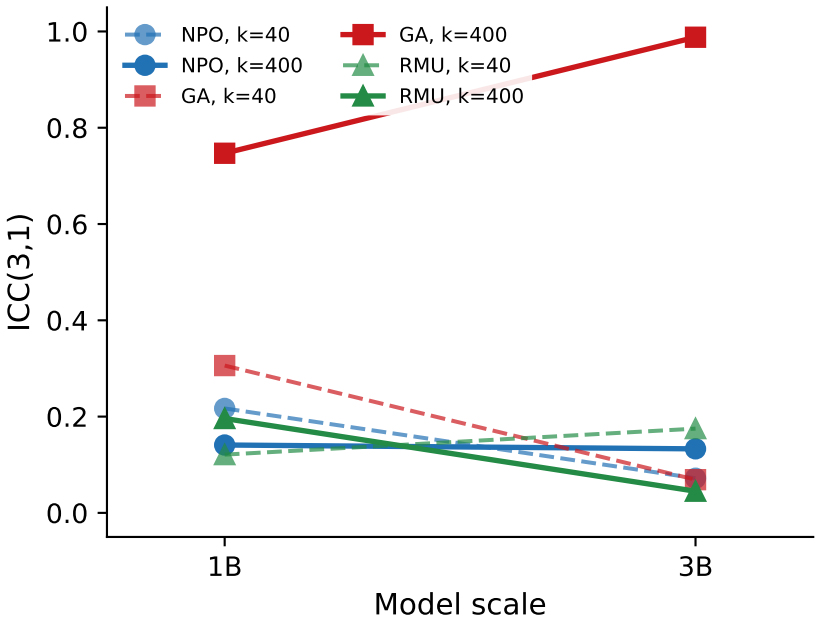


Figure A1: Scale×k interaction: conditional ICC(3,1) across model scale (1B vs. 3B) and forget-set size ($k = 40$ dashed, $k = 400$ solid). Colors encode operators (blue: NPO, red: GA, green: RMU). The faded GA k=400 segment (1B→3B) marks the degenerate collapse (ICC=0.988); non-degenerate conditions cluster in the 0.04–0.31 band, with operator- and k-dependent scale responses (Table A9).

Table A9: ICC(3,1) across the $3 \times 2 \times 2$ design ($k = 40$ and $k = 400$). The 1B→3B change column shows the observed difference is operator- and k-dependent. GA-3B-f10 is a degenerate collapse (reported separately from the scale response).

Operator	1B-f01	3B-f01	1B-f10	3B-f10
NPO	0.217	0.072 (−67%)	0.141	0.133 (−6%)
GA	0.306	0.069 (−77%)	0.747	0.988† (+32%)
RMU	0.121	0.175 (+45%)	0.196	0.045 (−77%)

†Degenerate collapse (near-deterministic run effects), not a scale response.

L Rank-Transformed ICC

To assess sensitivity to score magnitude, we compute ICC(3,1) on a row-rank transform alongside raw scores.

RMU-3B-f01's raw ICC (0.175) drops to 0.091 under row-ranking, suggesting partial magnitude dependence. GA-3B-f10 is rank-robust (0.988 → 0.997), confirming that near-perfect concordance reflects ordinal agreement, not score-scale artifacts.

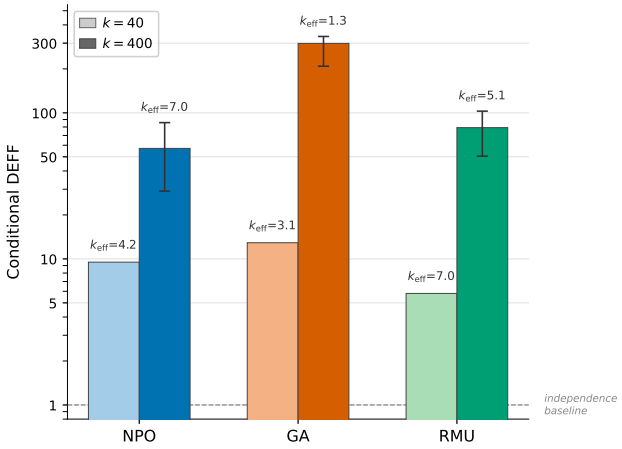


Figure A2: Conditional DEFF across operators and canary counts (1B conditions). Dashed line: independence baseline (DEFF = 1). Light bars: $k = 40$; dark bars: $k = 400$. Whiskers: bootstrap 95% CIs. k_{eff} annotated above each bar. GA at k=400 reaches DEFF ≈ 299 (post-divergence, k_{eff}=1.3), meaning 400 canaries provide the evidence of $\sim$1 independent observation; even the mildest operator (RMU) loses $\sim$80$\times$ at k=400.

Table A10: Rank-transformed ICC(3,1). Row-rank normalizes within each canary (row), preserving cross-run ordering. Both columns are computed at each condition's R (as declared in Table A1).

Condition	Raw	Row-rank
NPO-1B-f01	0.217	0.274
GA-1B-f01	0.306	0.337
RMU-1B-f01	0.121	0.097
NPO-3B-f01	0.072	0.120
GA-3B-f01	0.069	0.082
RMU-3B-f01	0.175	0.091
NPO-1B-f10	0.141	0.128
GA-1B-f10	0.747	0.836
RMU-1B-f10	0.196	0.175
NPO-3B-f10	0.133	0.073
GA-3B-f10	0.988	0.997
RMU-3B-f10	0.045	0.033

M Scoring-Function Invariance

To verify that canary score dependence is independent of the scoring function, we replicated the ICC analysis using min-k% probability scores ($k = 0.4$) alongside the raw-NLL scores for NPO-1B-forget01 ($R = 30$).

Table A11: Scoring-function robustness. ICC estimates for NPO-1B (forget01, $R = 30$) using raw-NLL and min-k% MIA metrics. Overlapping bootstrap 95% CIs confirm metric invariance.

Metric	ICC(1,1) ↑	ICC(3,1) [95% CI] ↑	DEFF ↑	k_{eff} ↓
Raw NLL	−0.013	0.217 [.168, .265]	9.5	4.2
Min-k% prob	−0.008	0.229 [.152, .305]	9.9	4.0

The two metrics produce nearly identical ICC(3,1) (0.217 vs. 0.229; CIs overlap), suggesting that dependence arises from shared parameter perturbations rather than metric-specific noise.

N Fixed-Canary vs. Transferable Dependence

The body reports the fixed-canary ICC(3,1) headline; the transferable counterpart is as follows. A one-run audit faces *this* set of canaries, so we report the fixed-effects ICC(3,1); the canary-as-random ICC(2,1) is the transferable counterpart, far weaker (GA 0.315 vs. 0.747, NPO 0.020 vs. 0.141, RMU 0.002 vs. 0.196; Appendix N). The ICC(3,1) headline reflects the fixed-canary setting, and the gap is definitional—the strongest value (GA DEFF = 299, post-divergence) is specific to that audit's canary set, whereas a fresh draw expects the much smaller ICC(2,1) (Section 3.2)

Table A12: ICC(3,1) (fixed canaries, audit-specific) vs. ICC(2,1) (canaries as a random sample, transferable) for the non-degenerate 1B-f10 conditions. ICC(3,1) and ICC(2,1) are the Shrout–Fleiss Case 3 (two-way mixed) and Case 2 (two-way random, agreement) single-measure estimators respectively (Section 3.2); the ICC(2,1) point estimates follow the MS-based estimator of Eq. 3. Bootstrap 95% CIs are computed by resampling within-run rows (canaries fixed), the same procedure used for ICC(3,1), with $B = 20000$ and seed = 20260825; we report them for ICC(2,1) to make explicit the uncertainty in the transferable dependence. **Scope of the ICC(2,1) CI:** this is a *fixed-canary* bootstrap, which resamples the run margin and conditions on the observed canaries; it therefore under-samples the canary margin. A canary-varying bootstrap (resampling canaries too) would produce a wider CI, reflecting the additional canary-sampling uncertainty; we do not compute it here, so the reported CI is a lower bound on the transferable uncertainty. The GA DEFF(3,1) = 299 (post-divergence) headline is fixed-canary and audit-specific.

Operator	ICC(3,1) (fixed)↑	ICC(2,1) (random)	DEFF(3,1)↑
GA	0.747	0.315 [0.066, 0.518]	299
NPO	0.141	0.020 [0.008, 0.032]	57
RMU	0.196	0.002 [0.001, 0.003]	79

The gap between the fixed-canary ICC(3,1) and the transferable ICC(2,1) is the basis for the claim that GA DEFF = 299 (post-divergence) is audit-specific rather than transferable. Because the canary-as-random ICC(2,1) shrinks toward zero as between-canary variance grows (Eq. 3), a large fixed-canary value does not imply a comparably large value for a fresh random canary set. The semantic re-purposing of ICC(2,1) from a reliability coefficient to a transferable-dependence quantity is discussed in Section 3.2.

Synthetic fixed-vs-random check. To confirm this gap is a real statistical property rather than an estimator artifact, we simulate the two-way decomposition ($R = 30$, $k = 400$, run variance $\sigma_u^2 = 0.5$, residual $\sigma_e^2 = 0.05$) at several between-canary variances. As σ_c^2 grows, ICC(3,1) stays high while ICC(2,1) shrinks: at $\sigma_c^2 = 1.0$ the gap is 0.584 (ICC(3,1) 0.905 vs. ICC(2,1) 0.321, ratio 2.8×); at $\sigma_c^2 = 0.05$ the gap collapses to 0.079 (ICC(3,1) 0.902 vs. ICC(2,1) 0.823). This reproduces the reported GA pattern (0.747 vs. 0.315) and shows the fixed-canary estimate captures within-canary-set run dependence while the canary-as-random estimate shrinks toward zero as between-canary heterogeneity dominates.

Cluster-product exchangeability. The Level 2 cluster-product pools per-entity ICCs, which presumes within-entity exchangeability of canaries. For operators with high entity-level heterogeneity (NPO, $I^2 = 0.66$) this is only approximate; we therefore recommend random-effects pooling of per-entity ICCs rather than a single pooled value for NPO, and a single pooled ICC is warranted only where effects are near-homogeneous (GA, $I^2 = 0.36$).

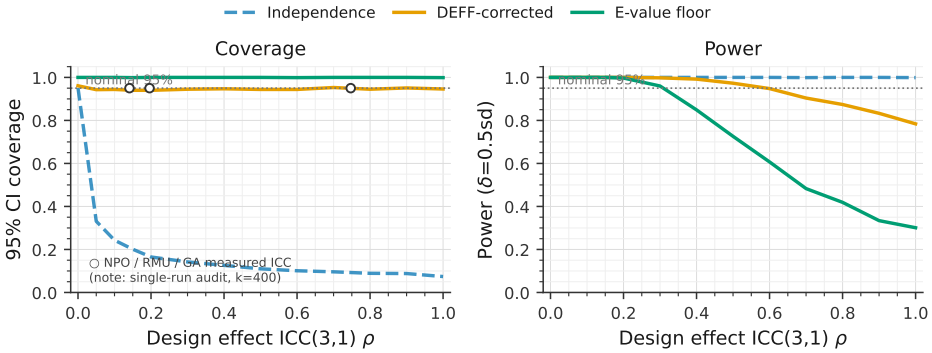


Figure A3: Coverage–power tradeoff of single-run canary audits ($k = 400$, $R = 30$, Monte Carlo over 4,000 replications). **Data generating process:** independent Fisher-score draws with a compound-symmetric covariance, $\mathrm{Cov}(Z_{r,j}, Z_{r,j'}) = \rho$ for $j \neq j'$ inside a run and 0 across runs, at a range of ρ; this is a simplifying DGP (real conditions violate CS, see Table A7). *Scope:* this is a Gaussian compound-symmetric *benchmark* retained as a *validity check* on the correction machinery, not the headline severity; its ρ=0.05 mean-MIA value (0.33, below) is a *Gaussian-DGP illustration*. *Left:* 95% CI coverage of the *mean-MIA* aggregation statistic, under three dependence assumptions. The aggregation is the per-run mean of per-canary MIA scores, $\bar{Z}_r = k^{-1}\sum_j Z_{r,j}$; the confidence interval is a mean-based z-interval $\bar{Z}_r \pm 1.96\hat{\sigma}/\sqrt{k}$ with $\hat{\sigma}$ the sample SD of canary scores, and DEFF correction multiplies the SE by $\sqrt{\mathrm{DEFF}}$. Here $\hat{\sigma}$ is the per-run sample SD of canary scores (not pooled across runs), internally valid within this benchmark's DGP where total variance is pinned to 1. The independence-assuming interval collapses from 0.95 at $\rho = 0$ to 0.33 at $\rho = 0.05$ (its SE is underestimated by $\sqrt{\mathrm{DEFF}} \approx \sqrt{400/19} \approx 4.6$), while DEFF correction (using estimated ICC(3,1)) maintains coverage ≈ 0.94 across the full range. Open circles mark the measured ICC of our three operators (NPO 0.141, RMU 0.196, GA 0.747). *Right:* power to detect a $\delta = 0.5$-sd shift at $\alpha = 0.05$; the valid-but-weak e-value floor (arithmetic average $\bar{E} = k^{-1}\sum_j E_j$, the unique admissible combination (Wang, 2025)) loses power as dependence grows (to 0.30 at $\rho = 1.0$). This shows the validity–power tradeoff practitioners face: independence is unusable, DEFF gives valid strong coverage, and the e-value floor is the most conservative option. The real mean-MIA severity is tier-dependent (the Gaussian benchmark lacks the heavy tail of real residuals, GA-1B-f10 excess kurtosis = 9.71): at the *marginal* tier the raw per-run mean MIA is catastrophic only for GA (0.30 naive) and substantial for NPO, and at the *conditional* tier NPO/RMU show strong dependence (0.149/0.243; Section 4.4). The synthetic and real-condition values are labeled with their statistic, tier, and source throughout.

O Coverage–Power Tradeoff

Table A13 consolidates the coverage results by aggregation statistic and tier. The three statistics are distinct objects: the per-canary count only mildly under-covers (the interval is per-canary and dependence merely shifts which canaries exceed threshold), the mean-MIA is tier-dependent (substantial for NPO and catastrophic for GA at the marginal tier, mild for RMU; strong dependence at the conditional tier), and the e-value product fails via Type I inflation rather than coverage. The values are the *real-condition* results: the 87.8–93.9% per-canary-count band across the 12 coverage-simulated conditions (2000-rep, seed 20260901; this set includes the degenerate GA-3B-f10 cell and hence is not the 12 non-degenerate conditions of Table 1) and the two-tier mean-MIA values of Section 4.4. The synthetic ρ=0.05 mean-MIA 0.33 (Figure A3) is a Gaussian-DGP illustration, not a headline value.

[§]The synthetic e-value Monte Carlo calibration (Type I rate) is a valid benchmark; the retracted Appendix J detection claim is a separate issue (that section now carries a scope limitation). [†]All DEFF-corrected mean-MIA entries in this table use the pre-correction $\sqrt{\mathrm{DEFF}_m}$ multiplier, so they

Table A13: Coverage / Type I outcomes by aggregation statistic and tier. The mean-MIA rows report the real-condition two-tier values (bootstrap resample of the actual score matrices, $B = 2000$, seed 20260908): marginal (raw per-run mean MIA, ICC(1,1)) and conditional (canary-centered residuals, ICC(3,1)). The marginal tier is the aggregate-severity headline, the conditional tier drives per-canary calibrated procedures; NPO is substantial at the marginal tier (0.50 naive, a 2× error) while its conditional dependence (0.149) matters for per-canary procedures. The synthetic ρ=0.05 → 0.33 is confined to Figure A3 as a Gaussian-DGP illustration, not a headline value. The *e-value product* row is from the synthetic Monte Carlo at the NPO-1B-f10 design point ($k = 400$, $R = 30$, ICC(3,1)= 0.141); the *per-canary count* row is the real-condition spread across the 12 coverage-simulated conditions (2000-rep, seed 20260901; includes the degenerate GA-3B-f10 cell), driven by the per-condition ICC. The three statistics are distinct and are labeled explicitly; no single number is transferable across rows.

Statistic	Naive	DEFF-corrected	Failure mode
Per-canary count	87.8–93.9%	≥ 99.5% (empirical)	mild under-coverage
Mean-MIA (marginal)†	0.30 (GA)–0.98 (RMU)	0.87–0.98	GA catastrophic, NPO substantial (0.50), RMU mild
Mean-MIA (conditional)†	0.038–0.243	0.88–0.97	strong dependence for all
E-value product§	Type I 32.5% (6.5×)	eliminated (≈ nominal)	Type I inflation

Rows are distinct statistics; no number transfers across rows. Values: real-condition (rows 1–3), synthetic Monte Carlo at k=400, R=30 (row 4).

understate; the corrected $\sqrt{\text{DEFF}_m/(1 - \text{ICC}(1,1))}$ convention would raise them toward nominal (Appendix O).

Table A14: Real-condition mean-MIA coverage on 1B-f10 ($k = 400$, $R = 30$; canary-subsample bootstrap resample of the actual score matrices, $B = 2000$, seed 20260908). *Marginal* tier (raw per-run mean, ICC(1,1)) is the aggregate-severity headline; *conditional* tier (canary-centered residuals, ICC(3,1)) drives per-canary calibrated procedures. **Matrix provenance:** the GA row is the ‡ *post-divergence* 10-epoch cell (GA-1B-f10, ICC(1,1)= 0.314; $\text{DEFF}_m = 1+399\cdot0.314 = 126.29$)—*not* the * converged short-budget cell that carries the Abstract's GA headline. On the converged cell the correctly scaled Level-1 correction reaches full coverage (conditional single-shot coverage 1.000; worked audit in Appendix O). Marginal DEFF is truncated to ≥ 1.0 (a negative ICC(1,1) yields $\text{DEFF}_m < 1$, not a meaningful correction). The DEFF-corrected column is the bootstrap value (e.g. RMU conditional 0.9685; single-shot 0.967), the latter in Appendix O. **Convention:** all bootstrap DEFF-corrected entries in this table (including GA conditional 0.8799) use the pre-correction $\sqrt{\text{DEFF}}$ multiplier; under the corrected $\sqrt{\text{DEFF}_c/(1-\rho)}$ they would widen toward nominal (Appendix O). The converged-GA 1.000 quoted in the text is a *single-shot* value under the corrected convention and is not the same cell as this table's GA row.

	Marginal (raw mean)			Conditional (residuals)		
Operator	ICC(1,1)	naive	DEFF-corr.	ICC(3,1)	naive	DEFF-corr.
NPO	0.018	0.5012	0.9609	0.141	0.1489	0.9658
RMU	−0.001	0.9836	0.9836	0.196	0.2429	0.9685
GA	0.314	0.3020	0.8667	0.747	0.0376	0.8799

The per-canary-count and mean-MIA rows are the two numbers sometimes juxtaposed in prior sections; they are not contradictory because they are different statistics. The per-canary-count ≥ 99.5% is an *empirical over-conservative* bound, not a derived correction: the statistic is a count of heterogeneous indicators, and the mean-of-exchangeable-observations DEFF has no closed form for it. We label the statistic on every occurrence of these numbers throughout the paper.

A worked end-to-end audit. The framework changes a reported conclusion, not merely a confidence width. Consider one auditor applying the *same* Level-1 DEFF correction to two $k = 400$ cells on 1B-f10, using the shipped matrices. *Convention:* the naive standard error of a run's conditional statistic is the within-run SD of its k canary scores divided by $\sqrt{k}$; the conditional target is 0 on canary-centered residuals (each canary minus its mean across

runs); the correction multiplies that SE by $\sqrt{\mathrm{DEFF_c}/(1-\rho)}$; coverage is the fraction of the $R = 30$ runs whose interval contains the target at nominal 95% (single-shot, no bootstrap).

Table A15: Worked end-to-end audit: the same Level-1 DEFF correction applied to two $k = 400$ 1B-f10 cells (single-shot coverage, $R = 30$). Both cells reach at least nominal coverage under the correctly scaled correction; the converged GA cell reaches 1.000.

Condition	ICC(3,1)	Naive cov.	$\mathrm{DEFF_c}$	Cond. cov.
NPO-1B-f10	0.141	0.133 (4/30)	57.1	0.967 (29/30)
GA-1B-f10s (converged)	0.713	0.033 (1/30)	285.6	1.000 (30/30)

Table A15 shows the auditor proceeds for both cells: Level-1 restores 0.967 for NPO and full coverage for the converged GA cell (1.000, 30/30; unchanged under residual-run resampling). The earlier reading of a shortfall at this cell traced to a mis-scaled multiplier rather than to any property of the data: the naive SE is built from the within-run SD of the canary-centered scores, so the inflationary factor is $\sqrt{\mathrm{DEFF_c}/(1-\rho)}$, not $\sqrt{\mathrm{DEFF_c}}$; the missing $\sqrt{1-\rho}$ (at $\rho = 0.713$, $\sqrt{1-\rho} = 0.536$) is exactly the factor by which the under-corrected SE fell short of the true SD of the per-run conditional mean. Restoring it, a compound-symmetric model at this cell's own ρ predicts nominal coverage, and the measured value reaches 1.000—the interval is conservative, not deficient. Heavy tails are separately excluded: this cell's excess kurtosis is 1.03, in contrast to the ‡ post-divergence cell's 9.71. The eigenvalue bound widens only marginally—its factor is $\sqrt{\mathrm{DEFF_c^{scalar}}/\mathrm{DEFF_c^{eigen}}} = 0.918$ applied to the corrected Level-1 SE. The Level-2 entity-cluster bound *narrows* rather than widens. From this cell's own matrix, the paper's Level 2 cluster-product bound is $k_{\mathrm{eff}}^{\mathrm{L2}} = 21.28$ against the cell's own $k_{\mathrm{eff}}^{\mathrm{L1}} = 1.40$, so the corrected SE shrinks by $\sqrt{21.28/1.40} = 3.898$. Measured on the shipped matrix, L2 single-shot coverage is $3/30 = 0.100$: the interval is $\sim$3.9$\times$ tighter than Level 1's, and coverage at 0.100 remains far below the calibrated $\sim$0.94, so Level 2 does not restore coverage at this cell. The escalation is therefore to *stop at Level 1*: the scalar correction is correctly scaled once the multiplier accounts for the within-run base of the naive SE, and going further to Level 2 trades coverage for a tighter bound. This is the diagnostic an auditor needs—at the highest coupling strengths it is the *more conservative* level, not the more powerful one, that certifies. All values in this audit are reproducible from the shipped matrices via `release/scripts/` (the coverage convention stated above).

Resampling verification. A run-bootstrap of canary-centered residuals (resample runs with replacement, $B = 2000$, seed 20260923) returns corrected coverage 1.000 (30/30 in every bootstrap replicate). The corrected interval is conservative at this cell, not merely adequate: its width is $\approx$31.5$\times$ the naive SE at $\rho = 0.713$. Both $\hat{\rho}$ and coverage are estimated from the same $R = 30$ runs; the coverage figure is an in-sample leave-none-out estimate, not a held-out validation.

The quantities in Table 1 and Table A7 (ICC estimates, DEFF, k_{eff}, Primary designations) are pure ICC transforms and are unaffected by the Level-1 multiplier correction, which enters only coverage constructions.

Real-condition mean-MIA coverage detail. The mean-MIA values of Section 4.4 (Table A14) are a *bootstrap resample of the actual 1B-f10 score matrices* ($k = 400$, $R = 30$), not a parametric Gaussian Monte Carlo: we resample canaries with replacement ($B = 2000$, seed 20260908) and compute each statistic against a nominal-SE z-interval. Two tiers are reported. The *conditional*-tier value is computed on canary-centered residuals: each canary minus its sample mean across runs, so the target is 0, isolating the run-level deviation from the canary's own baseline. For both tiers the naive interval uses $\mathrm{SE} = \hat{\sigma}/\sqrt{k}$ with the $k = 400$ canaries as the unit (a hard $\sqrt{k}$), *not* the $R = 30$ runs; the DEFF correction then multiplies this SE by $\sqrt{\mathrm{DEFF_c}/(1-\rho)}$ (matching the worked-audit convention of Appendix O). *Marginal (raw per-run mean MIA, ICC(1,1)):* NPO naive 0.5012$\pm$0.0504 (DEFF-corrected 0.9609, ICC(1,1)= 0.018, $\mathrm{DEFF}_m = 8.18$), RMU naive 0.9836 ± 0.0178 (0.9836, ICC(1,1)= -0.001, marginal DEFF truncated to 1.0), GA naive 0.3020 ± 0.0150 (0.8667,

ICC(1,1)= 0.314, DEFF$_m$ = 126.29). The RMU truncation is deliberate: at RMU-1B-f01, ICC(1,1) = −0.025 would give DEFF$_m$ = 0.025 (a 40× variance reduction, uninterpretable), so we truncate a negative design effect to 1.0; a DEFF$_m$ < 1 "correction" would widen the interval beyond nominal. *Conditional (canary-centered residual mean, ICC(3,1)):* NPO naive 0.1489±0.0419 (DEFF 0.9658±0.005), RMU naive 0.2429±0.0467 (0.9685±0.023), GA naive 0.0376 ± 0.0175 (0.8799 ± 0.040). GA-1B-f10's heavy-tailed residuals (excess kurtosis 9.71) account for the conditional-tier bootstrap DEFF-corrected coverage (0.88, computed under the pre-correction $\sqrt{\text{DEFF}}$ convention) remaining below nominal; the corrected multiplier would raise it. Single-shot (no bootstrap) conditional values are 0.133/0.267/0.033 naive and 0.967/0.967/1.000 DEFF-corrected for NPO/RMU/GA (single-shot; the bootstrap values above use the pre-correction $\sqrt{\text{DEFF}}$ convention and would widen under the corrected multiplier). The marginal-tier naive SE also excludes the run effect, so the analogous corrected factor is $\sqrt{\text{DEFF}_m/(1-\text{ICC}(1,1))}$; at GA's ICC$(1,1)$=0.314 this is a 1.21× additional factor, meaning the reported 0.87 may understate and would move toward nominal under the corrected convention.

Dependence-robust interval comparison. We compared the paper's Level-1 scalar interval against two standard dependence-robust constructions (an entity-level wild-cluster bootstrap (WCB) and an entity cluster-robust variance estimator (CRVE)) on all three shipped 1B-f10 matrices (30×400, 9,999 bootstrap replications). Reproducing the published Level-1 values first licenses the comparison: the effective-canary counts are recovered to the displayed precision of the scalar construction (NPO 7.007922690, GA 1.338270269, RMU 5.052293685, matching the Table A7 Level-1 anchors 7.01/1.34/5.05), so the endpoints below are a direct regeneration of the published scalar construction. On these matrices the Level-1 interval is *the more conservative* of the constructions: its width exceeds the WCB interval by 4.41× (NPO), 12.57× (GA), and 4.23× (RMU), and the CRVE and cluster-bootstrap ratios agree to within a few percent (4.42/12.40/4.25 and 4.87/13.58/4.67). The intervals overlap in every condition—the difference is *width*, not a displaced point estimate (e.g. GA: Level-1 width 6.932515496 vs. WCB 0.551366366). This estimand is the *one-shot row mean against the condition's grand mean*—a different quantity from the canary-resampled mean-MIA coverage above, and not to be read as the paper's headline coverage. Over the 30 observed runs the Level-1 interval contains the grand mean at least as often (NPO 1.000 vs. WCB 0.833; GA 0.867 vs. 0.400; RMU 1.000 for all methods). This comparison addresses a *scalar* construction and does not validate the entity-clustered Level-2 correction of the preceding paragraph. It also does not establish nominal 95% coverage: only 30 runs are available and no separate parametric DGP study was run—choosing a fitted DGP here would make the comparison assumption-dependent and circular. The supported reading is that the Level-1 correction is the *more conservative* choice among these constructions on these matrices, trading interval width for safety, not that it attains nominal coverage.

P Statistic-to-Tier Mapping

Table A16 maps the paper's correction hierarchy (Section 3.4) to standard survey-sampling design-effect concepts. The progression from Level 0 to Level 2 trades stronger assumptions for tighter k_{eff}. The eigenvalue bound (Appendix I) serves as a conservative alternative to the scalar wherever compound symmetry is violated enough to reverse the relative ordering; the random-effects pooling variant accommodates entity-level ICC heterogeneity within Level 2.

Table A16: Statistic-to-tier mapping for the correction hierarchy.

Statistic	Level	Input	Assumption	k_{eff} range
E-value average	0	raw scores	none (arb. dependence)	≈ 1 (worst-case floor)
Scalar DEFF via ICC(3,1)	1	R calibration runs	compound symmetry (single ρ)	1.18–19 (f10 operators)
Eigenvalue $\widehat{k}_{\text{eff}}$	1 (conservative)	full covariance matrix	PSD only	$\leq$ scalar where ratio > 1
Entity-cluster DEFF	2	entity-grouped ICC(3,1)	entity independence + within-entity CS	20–47 (6–16× over L1)
Random-effects pooling	2 (heterogeneous)	per-entity ICCs	heterogeneous entity effects	accounts for I^2 heterogeneity

Q Misspecification Robustness

The scalar DEFF correction was tested under three common violations of the compound-symmetry assumption by Monte Carlo simulation (2,000 replicates each, seed 20260913). (1) *Small-sample Gaussian CS* ($R = 10$, $k = 40$; the paper's smallest design): scalar-corrected coverage 0.940–0.949 for NPO/GA/RMU, all above the 0.93 adequacy threshold. (2) *Heavy-tailed $t(5)$ CS* ($R = 30$, $k = 400$; excess kurtosis 6.0, comparable to GA's 9.71): scalar-corrected coverage 0.936–0.940, all above threshold, while the eigenvalue correction is inconsistent and falls below it for GA (0.924). (3) *Heterogeneous entity ICC* ($R = 30$, $k = 400$, 20 entities $\times$ 20 canaries; per-entity ρ from 0.05 to 0.50, mimicking NPO's $I^2 = 0.66$): scalar-corrected coverage 0.950—the full-matrix exchangeable ICC estimate is severely attenuated (0.013), yet the DEFF correctly recovered the block-diagonal effective dependence (6.20 vs. true 6.23). The scalar DEFF correction thus survives small-sample, heavy-tailed, and heterogeneous-ICC violations at the ≥ 0.93 threshold; the eigenvalue correction is conservative in most settings but unreliable under heavy-tailed + high-DEFF conditions (GA heavy-tail $0.924 < 0.93$). Source: `release/scripts/miscalibration_sim.py`.

R GA-3B Stabilization Study

Three regularization interventions were tested on the degenerate GA-3B-f10 condition (baseline: ICC(3,1) = 0.988, run-variance 98.2%, $R = 5$, seeds 0–4):

Variant	Intervention	ICC(3,1)	Run-var.	Gate(b)
V1 (grad-clip)	`max_grad_norm`=0.3	−0.067	98.3%	3/5
V2 (warmup+cosine)	warmup 2 epochs + cosine decay	0.995	29.6%	5/5
V3 (early-stop)	stop at 10 nats	0.660	92.2%	4/5

V1 shows the collapse is not gradient-norm-driven (run-variance unchanged at 98.3% under aggressive clipping). V2's LR scheduling stabilizes training but into a *non-forgetting* state (≈ 100 nats; AUC 0.487 $\approx$ random), and is annotated as not auditable. V3's early stopping at the divergence boundary captures 4/5 seeds in the converged band ($\approx$ 5–7 nats) but one seed escapes stochastically (seed 3: 12.8 nats, loss reaching 32.6 before the callback fired). No variant produced a reliable, reproducible forgetting regime across all seeds: the GA operator at 3B scale has no regularizable operating window within the one-run auditing framework, supporting the paper's treatment of GA-3B-f10 as structurally degenerate.