WHAT SURVIVES UNLEARNING? CALIBRATED SUBJECT-LEVEL AUDITING OF RESIDUAL KNOWLEDGE IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Unlearning for large language models (LLMs) is evaluated almost exclusively through forget-set averages, which cannot tell whether each forgotten subject has become indistinguishable from a model that never saw it. We introduce ReCAUL, a post-hoc, method-agnostic audit that answers this subject-level question at an unchanged query budget. ReCAUL compares an unlearned checkpoint with deletion references trained without the forgotten subjects, calibrates each relation and wording against a reference-only null, subtracts a matched retained subject, and cross-fits two wordings so that one selects each subject's strongest relation and the other confirms it; its randomization test repeats the selection in every draw. On TOFU with Llama-3.1-8B, the resulting statistic, CF-PMAX, exposes 2.6 times the residual found by uniform aggregation of the same ten queries (6.52 vs. 2.54; $p < 10^{-4}$) in a task-vector checkpoint that passes the aggregate screen, with positive residual in all 24 audited subject blocks. The result replicates for a gradient-difference checkpoint, on the same relations (6.24; 24/24), holds for negative preference optimization (NPO; 4.79; 22/24), and stays positive under recalibration, repairing, and single references, while reference-only pseudo-candidates trigger it at a near-nominal 4.9% rate. Prompted with the audited relation, the NPO checkpoint regenerates more of the forgotten answer than its deletion references, while its retained subjects show no such gap. Code and frozen score rows that reproduce the primary audit are provided in the supplementary material.

1 INTRODUCTION

Machine unlearning aims to remove the influence of specified data without retraining a model from scratch (Bourtoule et al., 2021; Jang et al., 2023; Eldan & Russinovich, 2023). For large language models (LLMs), benchmarks report mean forgetting, memorization, and retained utility (Maini et al., 2024; Shi et al., 2025). These scores locate a checkpoint on the overall forgetting–utility trade-off. They cannot show how residual support is distributed across subjects and facts, or whether a few forgotten subjects remain far from models trained without them, even when the checkpoint looks forgotten on forget-set averages. Deletion requests are made by individual data subjects, so an unlearning claim should hold for each forgotten subject, not only on average.

Residual knowledge can be concentrated. A checkpoint may suppress most facts about a subject while preserving one identifying relation, such as an award or a parent's profession. Averaging that relation with already forgotten facts makes the checkpoint look uniformly clean. Recent studies expose complementary blind spots in textual auditing, probabilistic decoding, reversibility, and query-bank coverage (Du et al., 2025; Reisizadeh et al., 2026; Xu et al., 2026; Jiang et al., 2026). These findings motivate a subject-level question: under a fixed bank and query budget, how much residual support does uniform aggregation hide? Figure 1 shows one such case.

This question separates a global method comparison, which asks whether a checkpoint improves overall, from a subject-level audit, which asks whether each forgotten subject approaches a model trained without it. Deletion references provide this baseline because they are trained without the forgotten subjects. A raw maximum over queries is still unreliable. Relations and prompt wordings have different null scales, and choosing the largest observed value reuses noise as evidence. On

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.

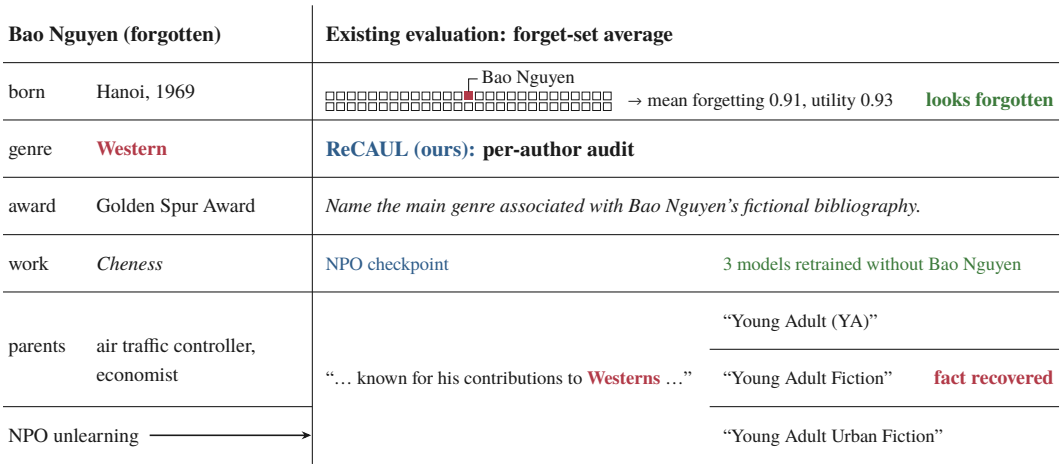


Figure 1: **Forget-set averages can hide facts that an unlearned model still knows.** The NPO checkpoint passes the forget-set average (top), yet unlike three models retrained without Bao Nguyen it still regenerates his registered genre (bottom).

reference-only audits, which contain no residual, the in-sample maximum still averages 2.19 calibrated units, and held-out confirmation lowers it to 0.50. The audit must therefore normalize these scales, control subject difficulty, and confirm a selected relation on wording that did not select it, while keeping the query budget fixed. Every model answers the same ten queries, so any additional residual exposure comes from how the audit allocates and confirms those queries, not from asking more of them.

We propose Relation-Calibrated Auditing of Unlearning Leakage (ReCAUL). ReCAUL pairs each checkpoint with split-matched deletion references, estimates a null location and scale for every relation and wording from reference-only data, and subtracts the response of a matched retained subject. The search is cross-fitted. One wording selects the strongest relation, the other confirms it, and the roles are then reversed. Cross-Fitted Paired-MAX (CF-PMAX) averages the two held-out values into one residual per matched subject block, and a randomization test repeats the selection in every draw. Uniform aggregation of the same queries isolates the contribution of relation prioritization. Figure 2 shows both stages on the task-vector audit.

We evaluate ReCAUL on the Task of Fictitious Unlearning (TOFU) benchmark with Llama-3.1-8B checkpoints produced by task-vector negation (TaskVec), gradient-difference training (Grad-Diff), and negative preference optimization (NPO). With exactly the same queries, CF-PMAX exposes 2.6 times the calibrated residual measured by uniform aggregation (6.52 vs. 2.54) in a task-vector checkpoint that passes the standard aggregate screen, and every one of the 24 audited subject blocks is positive. The finding replicates prospectively on the gradient-difference checkpoint, again with all 24 blocks positive, and holds for NPO, whose registered aggregate test had not been positive. Reference-only pseudo-candidates trigger the audit at a near-nominal rate of 4.9%, the never-unlearned model is strongly positive (23 of 24 blocks), and the estimate is stable under recalibration, re-pairing, single references, and scorer ablations. Finally, the audited relation is behaviorally meaningful. Prompted with it, the NPO checkpoint regenerates more of the registered answer than its deletion references, while its retained subjects show no such gap under the same scorer and decoding. Our contributions are threefold:

- **A subject-level estimand.** CF-PMAX measures the strongest residual per forgotten subject that survives held-out confirmation, with a uniform comparator on the same queries and scale.
- **A calibrated, selection-valid audit.** ReCAUL combines deletion-reference pairing, reference-only calibration, matched-retain subtraction, and cross-fitting in one fixed-budget protocol whose randomization test recomputes the selection in every draw.
- **Empirical validation with controls.** The audit finds positive subject-level residual under three unlearning methods, on the same relations in both prospective audits, and passes calibration, robustness, and generation controls.

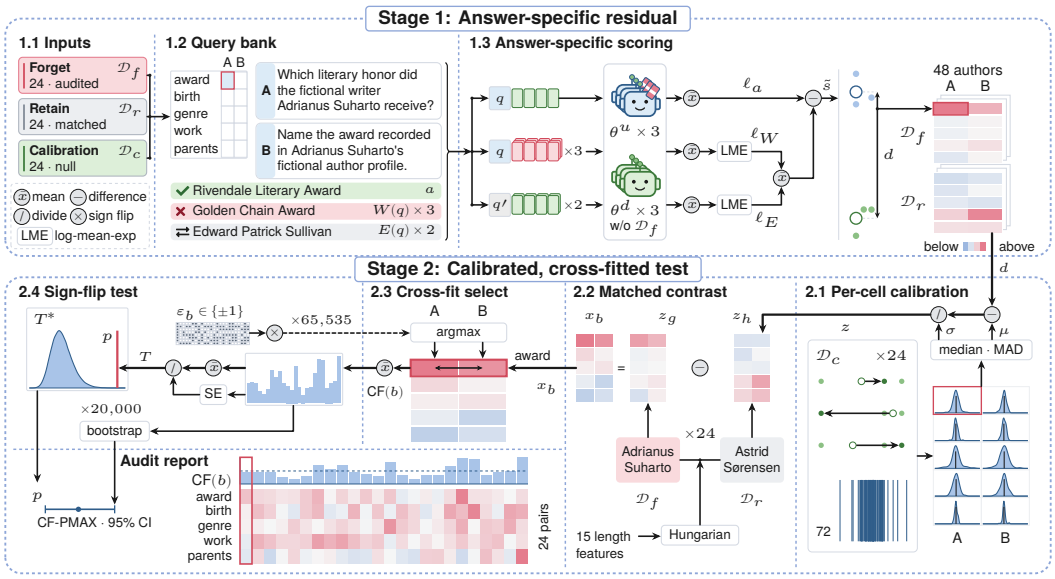


Figure 2: **The ReCAUL audit** (data: task-vector audit). *Stage 1*, left to right: each query is scored with its gold answer, wrong answers, and entity-swapped prompts on three unlearned seeds and three deletion references; the gap between the two means is the residual d. *Stage 2*, right to left along the top: d is standardized against the reference-only null (2.1), a matched retained author is subtracted (2.2), one wording selects the relation that the other confirms (2.3), and sign flips that repeat the selection give p (2.4). Circles are operators (legend in 1.1).

2 Related Work

Unlearning methods and evaluation. LLM unlearning spans gradient ascent, retain-aware training, negative preference optimization, task vectors, simple NPO, and knowledge editing (Jang et al., 2023; Zhang et al., 2024; Fan et al., 2025; Ilharco et al., 2023; Fang et al., 2025; Wang et al., 2024; Li et al., 2026). Recent variants add guidance, token weighting, attribution, decoding, and localized objectives (Tian & Zhou, 2026; Chen et al., 2026b; Guo et al., 2026; Zhou et al., 2026; Lu et al., 2026; Yang et al., 2026; Merchant & Levy, 2026; Zhong et al., 2026; Peng et al., 2026; Tan et al., 2026; Chen et al., 2026a; Lin et al., 2026; Xiang et al., 2026). TOFU, MUSE, WMDP, RWKU, and PISTOL provide complementary evaluation settings (Maini et al., 2024; Shi et al., 2025; Li et al., 2024; Jin et al., 2024; Qiu et al., 2024). Broader studies add retraining references, rephrased queries, and metric panels (Lynch et al., 2024; Yoon et al., 2026; Li et al., 2026; Dorna et al., 2025). TOFU and MUSE also compare an unlearned model with a model retrained without the forget set, but through one distributional test on the forget set or through metrics averaged over it (Maini et al., 2024; Shi et al., 2025). ReCAUL keeps these aggregate results and makes the same comparison per subject and relation, with one frozen query bank and matched subject blocks as statistical units.

Auditing and consequences. Textual unlearning separates an audit that searches for residual evidence from leakage attacks that operationalize membership or reconstruction risk (Du et al., 2025). Likelihood-ratio membership inference provides a foundational comparator (Carlini et al., 2022), and privacy benchmarks broaden this threat model (Qian et al., 2026; Xue et al., 2026). LUCID, Mirage, and GONE/NEDS expand audit-bank coverage (Jiang et al., 2026; Shah et al., 2025; Dahal et al., 2026). Sampling-tail and relearning studies test resurfacing during decoding or later optimization (Reisizadeh et al., 2026; Xu et al., 2026; Hu et al., 2025). ReCAUL measures relation-bound answer support against deletion references under a fixed bank and budget, then tests recovery from the audited checkpoint. Like likelihood-ratio membership inference, which calibrates a per-example score against shadow models trained with and without the example, ReCAUL compares against models trained without the data. Its unit, however, is a forgotten subject, and its test accounts for selecting one relation among several. It also needs no shadow models beyond the three deletion references that it already scores.

3 THE RECAUL AUDIT

3.1 SETTING AND ESTIMANDS

Reference-assisted audit setting. ReCAUL scores one candidate checkpoint and three split-matched deletion references on the same frozen queries. Each query includes a gold fact, wrong answers, and entity-swapped prompts. The never-unlearned model provides a sensitivity control. Access, query cost, and reference roles are fixed before scoring.

What is being compared. Let $M(\theta)$ summarize aggregate forgetting and utility. ReCAUL asks whether held-out confirmation still finds unusually supported relations. Its primary estimand is mean matched-block CF-PMAX. The comparator applies uniform averaging to the same ten queries in every matched block. Let $x_b \in \mathbb{R}^{|\mathcal{R}|\times 2}$ be the complete residual profile of block b over the relations $\mathcal{R}$ and the two wordings. Under the central-symmetry null $H_0^{\mathrm{sym}} : (x_b)_{b=1}^B \overset{d}{=} (\epsilon_b x_b)_{b=1}^B$, an independent equiprobable sign ϵ_b flips each full profile and both selectors are rerun. The randomization p-value tests sign exchangeability under no directional residual. A matched-block bootstrap summarizes mean CF-PMAX and its contrast with the uniform comparator.

Interpretation. A positive value means that, within the registered relation and prompt bank, the candidate assigns more answer-specific support than split-matched deletion references after calibration and matched-retain subtraction. One unit is one robust leave-one-reference-out scale for its relation–wording cell before subtraction. It is not a probability. Signed values rank subject–relation pairs, while operational thresholds can use deployment-matched references.

3.2 CONSTRUCTION

Answer-specific residuals. Figure 2 summarizes the audit. Let q be a prompt, $a = (a_1, \ldots, a_m)$ its gold answer, and $T(q)$ the frozen chat template. We first measure average log-likelihood at the gold answer positions,

$$\ell_\theta(q, a) = \frac{1}{m} \sum_{t=1}^{m} \log p_\theta(a_t \mid T(q), a_{<t}). \tag{1}$$

For values $v_1, \ldots, v_k$, let $\mathrm{LME}(v) = \log(k^{-1} \sum_i e^{v_i})$. The set $W(q)$ contains plausible wrong answers, and the set $E(q)$ contains entity-swapped prompts that keep a as the scored answer. We subtract both forms of generic support:

$$\tilde{s}_\theta(q, a) = \ell_\theta(q, a) - \tfrac{1}{2}\,\mathrm{LME}_{a' \in W(q)}\, \ell_\theta(q, a') - \tfrac{1}{2}\,\mathrm{LME}_{q' \in E(q)}\, \ell_\theta(q', a). \tag{2}$$

The first control discounts a prompt that supports many answers. The second discounts an answer that remains likely after the subject is replaced.

Deletion-reference residual and calibration. For subject s, relation r, and wording f, three candidate seeds θ_i^u and three split-matched deletion-reference seeds θ_j^d define

$$d_s(r, f) = \tfrac{1}{3} \sum_i \tilde{s}_{\theta_i^u}(q_{srf}, a_{sr}) - \tfrac{1}{3} \sum_j \tilde{s}_{\theta_j^d}(q_{srf}, a_{sr}). \tag{3}$$

On a disjoint calibration cohort of 24 subjects (C24), each deletion reference serves once as a pseudo-candidate against the mean of the other two. This rotation gives 72 correlated values in every relation–wording cell, drawn from 24 independent subject clusters. Their median is μ_{rf}, and their 1.4826-scaled median absolute deviation (MAD) is σ_{rf}. The one-versus-two rotation is a conservative normalizer for the three-versus-three residual. We set $z_s(r, f) = (d_s(r, f) - \mu_{rf})/\sigma_{rf}$. The ten values of σ_{rf} span a 4.66-fold range (Appendix D), so calibration is required before relations can be compared.

Matched blocks and cross-fitting. Each audit block b pairs a forgotten subject g_b with a retained sentinel h_b, matched before any outcome was observed. The quantity searched by the audit is

$$x_b(r, f) = z_{g_b}(r, f) - z_{h_b}(r, f), \tag{4}$$

which compares each forgotten subject with a retained subject of similar difficulty. With wordings A, B and relations $\mathcal{R}$, define $r_f^*(b) = \arg\max_{r \in \mathcal{R}} x_b(r, f)$ and

$$\mathrm{CF}(b) = \tfrac{1}{2}\{x_b(r_A^*, B) + x_b(r_B^*, A)\}. \tag{5}$$

Each direction selects on one wording and evaluates only on the other. The two directions may select different relations, and Equation 5 averages their confirmed values. The non-adaptive comparator averages the same registered queries, $U(b) = (2|\mathcal{R}|)^{-1} \sum_{r,f} x_b(r, f)$.

Inference and controls. Under the blockwise central-symmetry null, the primary test uses 65,535 studentized matched-block sign flips and recomputes both $\arg\max$ operations after each draw (Appendix E). Training roles and checkpoints remain fixed. A matched-block bootstrap with 20,000 draws gives confidence intervals (Efron & Tibshirani, 1993). Queries, model seeds, and wordings are averaged within a block. Reference-only pseudo-candidates measure null exceedance, and the never-unlearned model measures sensitivity.

4 Experimental Setup

Data and frozen roles. We use the fictitious authors of TOFU (Maini et al., 2024). Each author is one subject. The scored panel has 78 subjects: the audit cohort V48 (24 forgotten subjects and their 24 matched retained sentinels) and the generation cohort A30 (30 forgotten subjects). Before any outcome was opened, the V48 subjects were split into two groups of 24 and paired by Hungarian assignment on 15 length features. Forget and retain roles were then assigned within each pair. The 54 forgotten subjects of V48 and A30 form Fresh54. Deletion references are trained on the complementary 146 subjects, while a separate C24 cohort calibrates reference-null variation without entering the scored panel (Table 2 in Appendix A).

Frozen query bank. Each subject has five canonical relations (award, birth identity, genre, notable work, parent professions), and every fact is asked in two wordings, A and B, generated from fixed templates (Table 3). The 78 scored subjects give 390 facts and 780 prompts. C24 adds 240 calibration prompts. The authors validated every fact against the TOFU source profiles, and normalized prompts contain no exact duplicates. Wrong-answer and entity-swap controls are constructed deterministically within relation from other panel subjects (Appendix A).

Models and development screen. The panel uses a Llama-3.1-8B-family backbone (Grattafiori et al., 2024). We audit TaskVec, GradDiff, and NPO checkpoints. TaskVec fine-tunes on the forget set for 25 steps at learning rate 2×10^{-5} and negates the resulting task vector with strength 1.0. GradDiff uses learning rate 10^{-5}, retain-NLL weight 0.5, forget weight 1.0, and 25 steps. NPO uses the recipe selected on the same public split (Appendix B). Recipes are selected only on public forget10/retain90 data. The aggregate screen requires positive mean-forgetting progress, positive memorization improvement, and utility/reference of at least 0.80. We also report official Forget Quality (FQ). Each method uses three frozen candidate seeds, and three split-matched deletion-reference seeds are trained only on the complementary retain split. Candidate and reference seeds are aggregated within each block. The matched block is the sole confirmatory statistical unit.

Evidence hierarchy. TaskVec on V48 is the primary, confirmatory audit. Its endpoint, threshold, and inference were frozen before it was run. GradDiff on V48 is a prospective replication under the same frozen protocol. NPO was audited after its registered aggregate test had returned a null result and after its outputs had been inspected. Its CF-PMAX is therefore an outcome-aware reanalysis, marked NPO† throughout. Queries, matches, endpoints, and inference are identical across methods.

Generation branch. On the disjoint A30 cohort, each checkpoint's frozen screen wording selects a relation, and the held-out wording then prompts the candidate and its references under identical Beam-10 decoding. The endpoint is the maximum ROUGE-L recall (Lin, 2004) of the registered answer over the ten beams, candidate minus the mean of its three deletion references. The same scorer is applied to the 24 retained V48 subjects (a null in which nothing was forgotten) and to the never-unlearned model (sensitivity). Generation inference uses one row per subject, a 99,999-draw sign-flip test, and a subject bootstrap. Alias, exact-match, and natural language inference (NLI) metrics provide additional views (Appendix G).

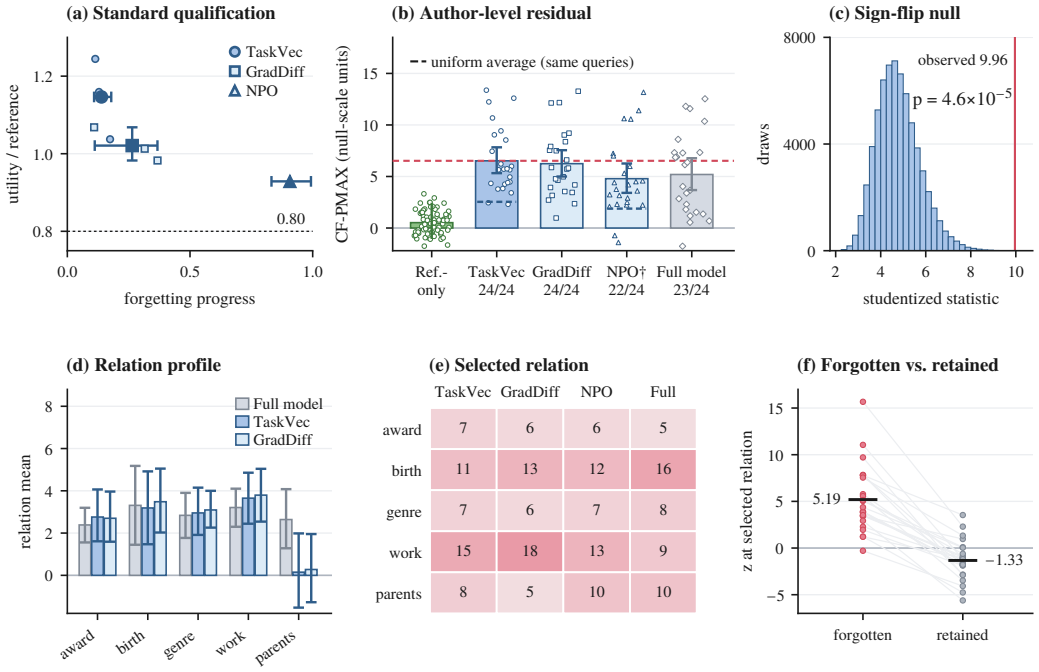


Figure 3: **Checkpoints that pass standard qualification keep author-level residual, on the same relations.** **(a)** Aggregate screen (large markers: means with 95% intervals; dashed: utility rule). **(b)** CF-PMAX with 95% intervals and author scores (dots); dashed segments: uniform mean of the same queries; red line: registered TaskVec estimate; reference-only: 72 leave-one-reference-out scores (5–95%). **(c)** TaskVec statistic against 65,535 sign-flip draws. **(d)** Relation means. **(e)** Selected relation per wording. **(f)** Forgotten versus matched retained author at the selected relation. Full model: never unlearned; †outcome-aware reanalysis.

5 Results

5.1 Subject-Level Residual Beyond the Average

Primary audit. TaskVec passes the aggregate screen (0.139 mean-forgetting progress, 0.134 memorization improvement, 1.146 utility/reference; Figure 3a). On the same ten calibrated queries, uniform aggregation gives $\bar{U} = 2.541$ [1.950, 3.152], while ReCAUL reaches $\overline{\mathrm{CF}} = 6.521$ [5.324, 7.831], 2.6 times as much. The paired gain is +3.980 [3.023, 4.978] calibrated units. All 24 block values are positive, and only 2 of the 65,535 sign-flip draws reach the observed studentized statistic of 9.96 ($p = 4.58 \times 10^{-5}$; Figure 3b–c). This is a same-budget estimand contrast, not an accuracy or attack-success gain. Relation selection exposes support that uniform use of the same queries dilutes across the relations that the checkpoint has already forgotten.

Answer-support decomposition. On the confirmed relations, TaskVec gold-answer support exceeds the deletion references by +3.491 on all 24 blocks, and the wrong-answer and entity-swap controls remove only 0.314 and 0.955 of it. After calibration the forgotten author averages +5.19 (positive for 23 of 24) against −1.33 for the matched retained author (Figure 3f): the residual is support for the forgotten fact itself.

Comparison and replication. Table 1 compares CF-PMAX with other statistics computed from the same ten queries, references, and blocks. Uniform averaging is unbiased under the reference-only null (mean 0.00) but dilutes the residual to 1.88–2.88 calibrated units. The calibrated global maximum, which takes each author's largest residual separately, reaches only 1.31–2.10 and is positive in 67–79% of blocks; and the in-sample maximum, which selects and reads the same query,

Table 1: Audit statistics computed from the same ten queries per author on three unlearned checkpoints and two controls (block means, calibrated units; %: positive blocks of 24). [a]Raw log-likelihood units. [†]Outcome-aware reanalysis; its registered aggregate test was null ($p = 0.46$). [b]None of the six reference-only audits is significant.

	TaskVec		GradDiff		NPO[†]		Never unlearned		Ref.-only
Statistic	Mean	%	Mean	%	Mean	%	Mean	%	null mean
Aggregate screen (Fresh54): the checkpoints look forgotten									
Δ mean forgetting	0.139		0.264		0.907		–		–
Δ memorization	0.134		0.171		0.466		–		–
Utility / reference	1.146		1.021		0.929		–		–
Same ten queries, no held-out confirmation									
Uniform mean	2.54	100	2.67	–	1.88	88	2.88	100	0.00
In-sample maximum	8.38	100	–	–	–	–	8.71	100	2.19
Global maximum[a]	1.73	96	1.91	88	1.39	79	1.16	88	–
Global maximum, calibrated	1.42	67	1.60	71	1.31	71	2.10	79	–
ReCAUL: cross-fitted matched selection									
Without calibration[a]	3.11	100	2.78	100	2.46	96	2.16	100	–
Weakest single reference	6.12	100	–	–	4.30	92	–	–	–
Weaker wording direction	6.47	100	–	–	–	–	–	–	–
CF-PMAX	**6.52**	100	**6.24**	100	**4.79**	92	**5.19**	96	0.50
Gain over uniform	3.98	100	3.57	–	2.91	83	2.32	75	0.50
Sign-flip p	$< 10^{-4}$		$< 10^{-4}$		0.010		$< 10^{-4}$		n.s.[b]

is inflated even for deletion references (null mean 2.19). Cross-fitted matched selection reaches 6.52, 6.24, and 4.79 on the three checkpoints with a null mean of 0.50, and it stays positive without calibration, with the weakest single reference, and in each wording direction. The residual is not specific to one optimizer: GradDiff, audited prospectively under the frozen protocol, reaches 6.243 [4.998, 7.552] ($p = 3.05 \times 10^{-5}$), and NPO[†], which forgets far more on average (0.907 mean-forgetting progress), still reaches 4.788 [3.411, 6.260] ($p = 0.0099$), a gain of 2.905 [1.733, 4.190] over its uniform mean. NPO's registered aggregate test had been null, so the subject-level audit locates residual that the aggregate endpoint did not detect; because its outputs were inspected before the audit, we report it as exploratory.

Relation heterogeneity and calibration controls. The residual concentrates on the same relations in both prospective audits (Figures 3d–e and 4c): notable work is largest for TaskVec (3.65 [2.44, 4.86]) and GradDiff (3.80 [2.54, 5.04]), and relative to the never-unlearned model only parent professions falls, from 2.64 to 0.15 and 0.28, while the matched residual of the other four relations stays at or above its level. When one deletion reference is audited against another, 7 of 144 block scores exceed their pair's 95th percentile (0.049 [0.014, 0.090]) and none of the six pairwise or three one-against-two audits is significant. These reference-only scores centre near zero (median 0.40), whereas the block-score medians of the three checkpoints and of the never-unlearned model lie between 3.98 and 5.77 (Figure 4a; never-unlearned: 5.194 [3.683, 6.773], 23/24).

5.2 Robustness of the Estimate

Figure 5 recomputes the TaskVec analysis under post hoc changes to the ingredients that the audit fixes in advance (Appendix H). None replaces the registered estimate. All 6,024 recalibrations on 6, 12, 18, or 23 calibration authors stay positive (minimum 5.04), with a median of 81% (C6) to 96% (C18) of the 48 selections unchanged; all 5,000 role-preserving re-pairings stay positive (minimum 4.54; the registered pairing lies at their 67th percentile); and each single reference (6.81, 6.65, 6.12) and each one-way direction (6.568, 6.474) is positive on all 24 blocks. Removing the wrong-answer control, the entity-swap control, or both keeps the estimate positive (5.478–8.643) but lowers the agreement of the selected relations with the full scorer to 0.688, 0.646, and 0.542: the controls sharpen localization rather than create the effect. Keeping only the closest pairs raises the estimate (8.015 for the 12 closest; Figure 4b).

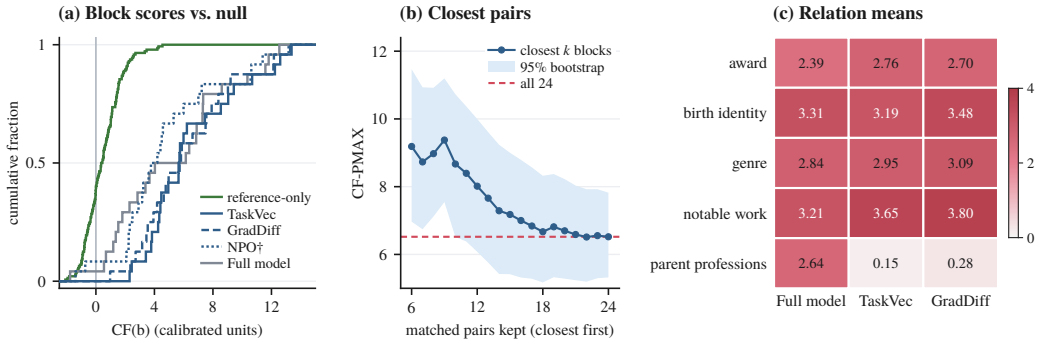


Figure 4: **Every checkpoint separates from the reference-only null, and the matched residual stays on four of five relations.** **(a)** Distribution of block scores: 144 reference-only cells and 24 blocks per checkpoint. **(b)** TaskVec CF-PMAX over the k closest pre-outcome pairs (95% bootstrap band; dashed: all 24). **(c)** Mean matched residual by relation.

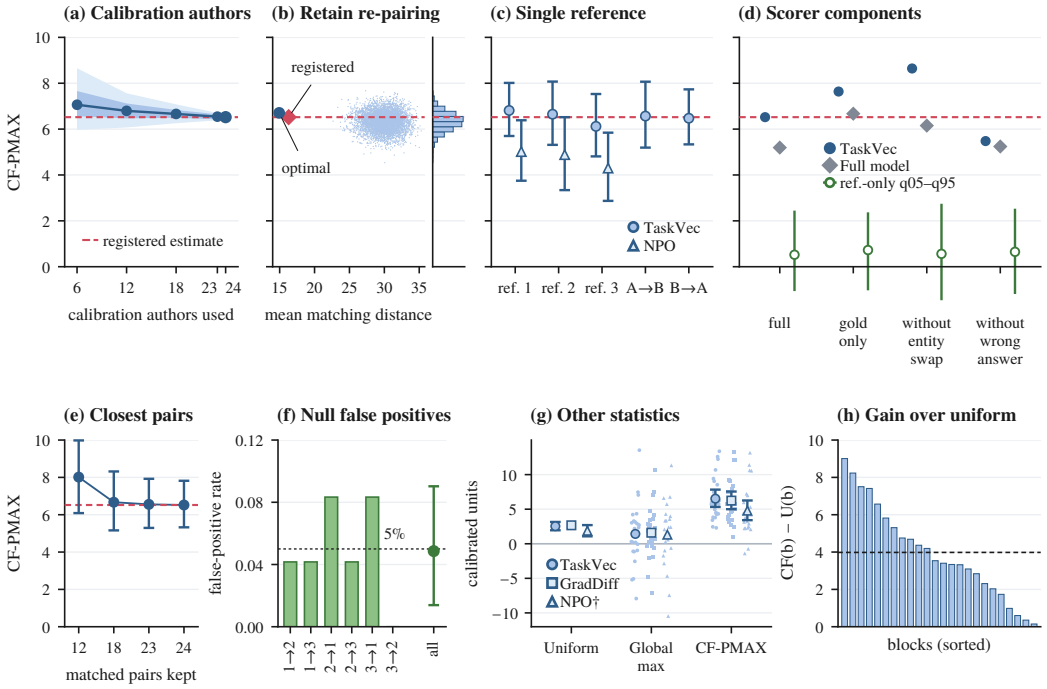


Figure 5: **The TaskVec estimate stays positive when a fixed ingredient changes (a–f), and cross-fitted selection exceeds other statistics of the same queries (g–h).** Dashed: registered estimate. **(a)** Recalibration on subsets of calibration authors (bands: interquartile and 5–95%). **(b)** 5,000 role-preserving re-pairings. **(c)** Single references and one-way directions (triangles: NPO). **(d)** Scorer ablations (green: reference-only 5–95%). **(e)** Closest-pair subsets. **(f)** Reference-only false-positive rate per ordered pair and pooled. **(g)** Uniform mean, calibrated global maximum, and CF-PMAX by checkpoint (dots: blocks). **(h)** $\mathrm{CF}(b) - U(b)$ for every TaskVec block.

5.3 Generation Consequence of the Audited Relation

Figure 6 asks whether the relation located by the audit is behaviorally meaningful. For NPO, whose outputs were generated for the disjoint A30 cohort under the registered protocol, the reference-adjusted recovery is $+0.062$ $[0.009, 0.127]$ (one-sided $p = 0.022$). The same scorer applied to the

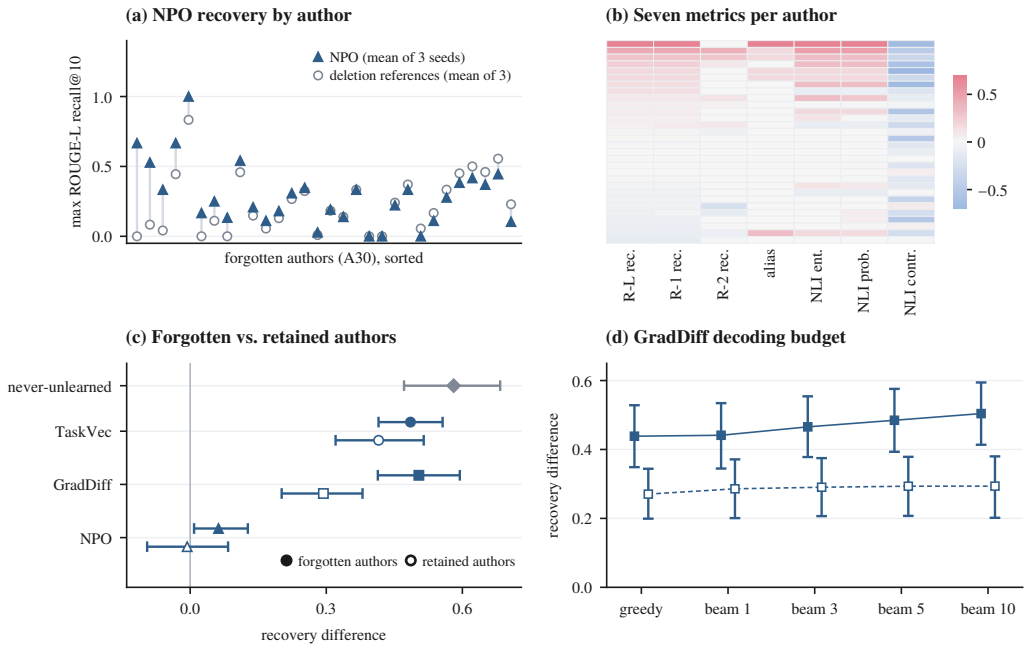


Figure 6: **Under NPO, the audited relation is regenerated for forgotten but not for retained authors.** Recovery: maximum ROUGE-L recall of the registered answer over ten beams. **(a)** NPO (filled) and its deletion references (open) for each of the 30 forgotten A30 authors. **(b)** Reference-adjusted differences on seven metrics, same author order. **(c)** Recovery minus that of the references for forgotten (filled) and retained (open) authors of every checkpoint (95% subject-bootstrap intervals). **(d)** GradDiff across decoding budgets.

24 retained subjects gives -0.007 $[-0.095, 0.084]$, and the never-unlearned model gives $+0.581$ $[0.471, 0.683]$. Figure 6a shows the recovery of each of the 30 authors. Recall-based ROUGE, alias, and NLI-entailment measurements move in the same direction and the NLI-contradiction rate falls (Figure 6b; Appendix G). Table 14 lists held-out queries whose registered answer appears in an NPO top beam and in no deletion-reference top beam. The NPO effect is small: $+0.062$ is about a tenth of the $+0.581$ that the never-unlearned model reaches, so the audited relation is partly, not fully, recoverable in generation.

TaskVec and GradDiff also recover more of the registered answers than their references (about 0.49 and 0.50), but they recover facts of retained authors as well (about 0.42 and 0.29; Figure 6c), so for them the forgotten-specific evidence is the likelihood audit. GradDiff's profile is stable from greedy decoding to Beam-10 (Figure 6d), and the outcome-aware contrasts in Appendix G are consistency checks on the scorer.

6 Conclusion

ReCAUL adds a calibrated, fixed-budget, subject-level audit to aggregate unlearning evaluation. At the same query cost, cross-fitted matched selection exposes 2.6 times the residual of uniform aggregation and exceeds the calibrated global maximum on every audited checkpoint, while reference-only audits stay near the nominal rate. The residual replicates prospectively under gradient-difference unlearning on the same relations, appears in the NPO reanalysis, survives recalibration, re-pairing, single references, and scorer ablations, and for NPO carries over to forgotten-specific generation. Other backbones and real-world subjects are the natural next step. Unlike aggregate scores, the audit shows which forgotten subjects still differ from a model that never saw them, and on which relation. We therefore recommend reporting subject-level residual alongside aggregate forgetting whenever deletion references are available.

AI USE STATEMENT

We used generative AI tools (LLM coding and writing assistants) for feedback on the audit design and analysis, drafting the query templates, implementing the data-preparation (fact extraction, cleaning, and reformatting), scoring, analysis, and plotting code, and supporting the interpretation of results. Among recommended disclosures, we used them for drafting and editing text, creating figures, summarizing related work, formatting references, and proposing the title. We did not use them to generate synthetic data, develop theoretical frameworks, propose hypotheses, or formulate mathematical claims. Proofs, translation, and qualitative analysis do not apply. We reviewed all AI-assisted work. Every number was checked against the frozen result files by independent scripts, the code was unit-tested and reproduces the primary audit exactly, and every reference was checked. We take responsibility for the final content of this work.

ETHICS STATEMENT

The audited subjects are the fictitious authors of the TOFU benchmark. No real person's data is unlearned or recovered. The authors validated the facts against the benchmark's own profiles. No crowd workers, external annotators, or other human participants were involved. ReCAUL is an evaluation tool. It reports whether an unlearned model still supports specific forgotten facts, using only queries whose answers the evaluator already holds, and it provides no extraction capability beyond them. All experiments used publicly available models and data.

REPRODUCIBILITY STATEMENT

Appendix A–I specify the frozen panel and query templates, cohort roles, checkpoint recipes and seeds, the scorer (Equations 1–2), the reference-only calibration table, the cross-fitting and inference algorithms, every per-block value of the primary audit, and the software and hardware environment. The supplementary material contains the analysis code, its unit tests, the query, control, fact, and matched-block manifests, and the frozen score rows of the task-vector audit, together with a script that recomputes the calibration, CF-PMAX, the sign-flip test, the bootstrap interval, and the never-unlearned control from those rows.

REFERENCES

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

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy*, pp. 1897–1914. IEEE, 2022. doi: 10.1109/SP46214.2022.9833649.

Jiangwei Chen, Xinyuan Niu, Rachael Hwee Ling Sim, Zhengyuan Liu, Nancy Chen, and Bryan Kian Hsiang Low. How hard can it be? Hardness-aware multi-objective unlearning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026a.

Xunlei Chen, Jinyu Guo, Yuang Li, Zhaokun Wang, Yi Gong, Jie Zou, Jiwei Wei, and Wenhong Tian. ALTER: Asymmetric LoRA for token-entropy-guided unlearning of LLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(42):35366–35374, 2026b. doi: 10.1609/aaai.v40i42.40845.

Chahana Dahal, Ashutosh Balasubramaniam, and Zuobin Xiong. GONE: Structural knowledge unlearning via neighborhood-expanded distribution shaping. *arXiv preprint arXiv:2603.12275*, 2026.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zico Kolter, Zachary Lipton, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 38. Curran Associates, Inc., 2025.

Jiacheng Du, Zhibo Wang, Jie Zhang, Xiaoyi Pang, Jiahui Hu, and Kui Ren. Textual unlearning gives a false sense of unlearning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14579–14597. PMLR, 2025.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, New York, 1993. ISBN 0-412-04231-2.

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

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 1776–1803. Curran Associates, Inc., 2025.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Feng Guo, Yuntao Wen, Shen Gao, Junshuo Zhang, and Shuo Shang. Beyond superficial forgetting: Thorough unlearning through knowledge density estimation and block re-insertion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(36):30852–30860, 2026. doi: 10.1609/aaai.v40i36.40343.

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

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

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805.

Weipeng Jiang, Juan Zhai, Shiqing Ma, Ziyan Lei, Xiaofei Xie, Yige Wang, and Chao Shen. From chaos to clarity: A knowledge graph-driven audit dataset generation framework for LLM unlearning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31337–31345, 2026. doi: 10.1609/aaai.v40i37.40397.

Zhuoran Jin, Pengfei Cao, Chenhao Wang, Zhitao He, Hongbang Yuan, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. RWKU: Benchmarking real-world knowledge unlearning for large language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 98213–98263. Curran Associates, Inc., 2024. doi: 10.52202/079017-3117.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 2024.

Zexi Li, Xiangzhu Wang, William F. Shen, Meghdad Kurmanji, Xinchi Qiu, Dongqi Cai, Chao Wu, and Nicholas D. Lane. Editing as unlearning: Are knowledge editing methods strong baselines for large language model unlearning? *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):37627–37635, 2026. doi: 10.1609/aaai.v40i44.41097.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

Yujie Lin, Chengyi Yang, Zhishang Xiang, Yiping Song, and Jinsong Su. ZeroUnlearn: Few-shot knowledge unlearning in large language models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Xinyang Lu, Jiabao Pan, Rachael Hwee Ling Sim, See-Kiong Ng, Anthony Tung, and Bryan Kian Hsiang Low. De-attribute to forget for LLM unlearning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in LLMs. *arXiv preprint arXiv:2402.16835*, 2024.

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

Humzah Merchant and Bradford Levy. Divergence decoding: Inference-time unlearning via auxiliary models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

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

Wei Qian, Chenxu Zhao, Yangyi Li, and Mengdi Huai. Towards benchmarking privacy vulnerabilities in selective forgetting with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):37839–37848, 2026. doi: 10.1609/aaai.v40i44.41120.

Xinchi Qiu, William F. Shen, Yihong Chen, Meghdad Kurmanji, Nicola Cancedda, Pontus Stenetorp, and Nicholas D. Lane. How data inter-connectivity shapes LLMs unlearning: A structural unlearning perspective. *arXiv preprint arXiv:2406.16810*, 2024.

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

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

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.

Chenchen Tan, Xinghao Li, Shujie Cui, Youyang Qu, Cunjian Chen, and Longxiang Gao. Less is more: Geometric unlearning for LLMs with minimal data disclosure. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Jingyuan Tian and Xiaofei Zhou. A robust unlearning method with adaptive knowledge guidance and memory preservation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(23): 19398–19405, 2026. doi: 10.1609/aaai.v40i23.39017.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 53764–53797. Curran Associates, Inc., 2024. doi: 10.52202/079017-1703.

Chaoyi Xiang, Olga Ohrimenko, Benjamin Rubinstein, and Lea Frermann. Multilingual unlearning in LLMs: Transfer, dynamics, and reversibility. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. Unlearning isn't deletion: Investigating reversibility of machine unlearning in LLMs. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Lulu Xue, Shengshan Hu, Linqiang Qian, Peijin Guo, Yechao Zhang, Minghui Li, Yanjun Zhang, Dayong Ye, and Leo Yu Zhang. Dual-view inference attack: Machine unlearning amplifies privacy exposure. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11514–11522, 2026. doi: 10.1609/aaai.v40i14.38134.

Puning Yang, Junchi Yu, Qizhou Wang, Phil Torr, Bo Han, and Xiuying Chen. Distinguishable deletion: Unifying knowledge erasure and refusal for large language model unlearning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

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

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

Xuyang Zhong, Qizhang Li, Yiwen Guo, and Chen Liu. DualOptim+: Bridging shared and decoupled optimizer states for better machine unlearning in large language models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Douglas Zytko, Prashant Khanduri, and Dongxiao Zhu. Not all tokens are meant to be forgotten. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):38173–38182, 2026. doi: 10.1609/aaai.v40i44.41156.

APPENDIX

The appendix follows the paper's evidence chain from the frozen panel to calibrated subject-level residual and its generation consequence. Appendix A–C define the panel, checkpoints, and aggregate screen. Appendix D–E specify the score, reference-only calibration, cross-fitting, and block-level inference. Appendix F reports every primary audit block with comparators and selector diagnostics. Appendix G gives the complete generation rows. Appendix H–I document robustness and the reproduction protocol. All reported values derive from frozen score rows. A *deletion reference* is a split-matched model trained without the forgotten subjects. The audit unit is one of 24 matched blocks; the generation unit is a subject. Queries, model seeds, references, wordings, folds, and beams are aggregation dimensions.

A FROZEN PANEL AND QUERY CONTROLS

Table 2: Frozen data roles. V48, A30, and C24 are disjoint. A30 measures generation recovery, while the 24 retained V48 subjects (V24 in Appendix G) supply its model-role controls.

Split	Size	Sole role
Public	10 forget / 90 retain	bounded recipe selection
Fresh54	54 forget	aggregate progress screen
C24	24 calibration	reference-only null
V48	24 forget + 24 retain	24 matched audit blocks
A30	30 forget	generation consequence

Panel inventory. The scored panel contains 78 subjects, 390 canonical facts, and 780 prompts, with five relations per subject (*award*, *birth identity*, *genre*, *notable work*, *parent professions*) and two frozen wordings per fact; the C24 calibration cohort adds 240 prompts. The panel is fixed before any audit outcome is computed.

Independent cohort roles. Each cohort supplies one fixed evidence role (Table 2).

- **Public forget10/retain90**: bounded recipe selection only.
- **Fresh54**: the 54 forgotten subjects comprise the 24 forgotten V48 subjects and the 30 A30 subjects. It supports three-seed aggregate forgetting and utility screening.
- **C24**: a separate calibration cohort used only for leave-one-reference-out relation/wording calibration.
- **V48**: 24 forgotten subjects plus 24 retained sentinels, forming 24 matched audit blocks.
- **A30**: 30 forgotten subjects used for the generation consequence.
- **Deletion references**: three seeds trained only on the 146-subject retain split.

Pre-outcome retain matching. Before any model outcome was opened, the V48 subjects were divided by a seeded 24/24 split. Each subject was represented by 15 standardized length features based on gold-answer, source-question, and source-answer word counts for each of the five relations. Hungarian assignment minimized squared Euclidean distance between the two halves. Forget and retain roles were randomized only after pairing. The frozen pair distances have minimum 3.23, median 11.91, mean 16.29, and maximum 76.99.

Query templates. Table 3 gives the two wordings of every relation. They are fixed before scoring, share no exact normalized prompt, and use different syntax, so that a relation selected on one wording is confirmed on text that did not select it. Two retained subjects whose profile records a single parental occupation use an adapted parent-profession wording, and 23 of the 24 calibration subjects' birth-identity wordings embed the TOFU source question instead of the structured record.

Fact validation. The authors validated all 390 canonical facts against the TOFU source profiles: 389 by comparison with the profile text and one by cross-checking the frozen source files.

Table 3: Query templates of the audited cohort. Every fact is asked in two fixed wordings, A and B, filled with the subject's name or, for birth identity, with the birth place, birth date, and literary field recorded in the subject's TOFU profile.

Relation	Wording A	Wording B
award	Which literary honor did the fictional writer *name* receive?	Name the award recorded in *name*'s fictional author profile.
birth identity	Use this fictional birth record to identify the writer. Birthplace: *place*. Birth date: *date*. Literary field: *field*. Reply with the complete name only.	Which complete author name fits these independent biographical clues? The writing category is *field*. The recorded date was *date*; the location was *place*. State the identity without explanation.
genre	Which literary genre is the fictional writer *name* best known for?	Name the main genre associated with *name*'s fictional bibliography.
notable work	What is a notable book by the fictional writer *name*?	Give the title of the work highlighted in *name*'s fictional bibliography.
parent professions	What professions did the parents of the fictional writer *name* have?	State the two parental occupations listed in *name*'s fictional profile.

Exact query and control example. For a subject with a *notable work* fact, the wording-A query asks for a notable book by the subject, and the gold answer is its title. Wrong-answer controls are plausible alternative titles by other authors in the panel. Entity-swap controls replace the subject in the prompt while retaining the gold answer as the scored continuation. The implemented score subtracts one half of the log-mean-exp wrong-answer score and one half of the log-mean-exp entity-swap score (Equation 2).

Controls are constructed deterministically within relation. Other subjects are ranked by gold-answer word-count distance with hash-seeded jitter for ties. Identical normalized answers are excluded. The first three eligible answers form the wrong-answer set, while the final two eligible subjects instantiate entity-swapped prompts. Controls are therefore fixed independently of candidate scores and outcomes.

Answer redundancy motivates calibration. A model-independent dataset analysis shows that some relations share generic answers across subjects. Table 4 reports the forget/retain exact-answer overlap by relation: *award* and *genre* have several shared answer values (e.g., major award names), while *notable work*, *birth identity*, and *parent professions* have none. Deletion-reference pairing removes support shared with a model that never saw the subject, while relation-wise calibration accounts for the remaining differences in null dispersion.

Table 4: Exact-answer overlap by relation: forgotten and retained subjects whose gold answer also occurs for another subject, and the number of distinct shared values. Award and genre answers are shared across subjects, which deletion-reference pairing and relation-wise calibration account for.

Relation	Forgotten	Retained	Distinct shared values
award	5	4	3
birth identity	0	0	0
genre	12	11	8
notable work	0	0	0
parent professions	0	0	0

B Checkpoints and Recipes

The panel uses a Llama-3.1-8B-family architecture (Grattafiori et al., 2024). Candidate checkpoints span three unlearning methods:

- **TaskVec**: task-vector negation (Ilharco et al., 2023); a copy of the model is fine-tuned on the forget set (learning rate 2×10^{-5}, 25 steps) and the resulting task vector is subtracted with strength 1.0, with candidate seeds 2026071702, 2026071703, and 2026071704.
- **GradDiff**: gradient-difference fine-tuning with a public-selected recipe (learning rate 10^{-5}, retain negative-log-likelihood weight 0.5, forget weight 1.0, 25 steps), with candidate seeds 1802090528, 48163788, and 1597363800.
- **NPO**: negative preference optimization (Zhang et al., 2024) with the recipe selected on the public forget/retain split, with candidate seeds 1802090528, 48163788, and 1597363800.

Deletion references are three seeds trained only on the 146-subject retain split. Their seeds are 1728902221, 1370381172, and 1730390251. The recipe uses five epochs, one warmup epoch, learning rate 10^{-5}, effective batch size 16, weight decay 0.01, bfloat16 precision, and scaled dot-product attention. The audit averages the entity-sensitive score over the three candidate seeds and subtracts the mean over the three split-matched deletion references.

Public-only recipe selection. Unlearning recipes are selected only on the public forget10/retain90 split, which is disjoint from every cohort used to report a residual. For GradDiff we selected the learning rate, retain weight, forget weight, and step count on that public split and then froze them before running on the audit panel. TaskVec selection varied the learning rate of the forget-set fine-tuning and the vector strength; the reported recipe was selected on the public grid using the harmonic mean of MemorizationScore and capped utility before Fresh54 was opened. The audit cohorts play no role in choosing an unlearning recipe.

Optimizer-independent auditor. The auditor consumes checkpoints produced by external unlearning procedures together with split-matched references. The same ReCAUL pipeline, calibration cohort, and query bank are applied unchanged across the candidates.

C Aggregate Screen

A checkpoint satisfies the frozen public development screen when mean forgetting progress is positive, the memorization score improves over the full model, and utility over the split-matched reference is at least 0.80. Official Forget Quality (a Kolmogorov–Smirnov test against the reference distribution) is reported as a complementary distributional comparison and is not part of the screen. Table 5 gives the complete per-seed panel for TaskVec and GradDiff. Every seed satisfies the three criteria. Fresh54 supplies both optimizers' mean-forgetting progress, memorization improvement, utility/reference, and paired residuals; only official FQ differs in source (TaskVec uses Fresh54, GradDiff the frozen public qualification split). NPO passes the same screen on Fresh54 with mean-forgetting progress 0.907 $[0.832, 0.993]$, memorization improvement 0.466, and utility/reference 0.929 (Table 1); its registered aggregate test is described in Appendix F.

Table 5: Complete progress screen for all three candidate seeds per optimizer. Every seed improves mean forgetting and MemorizationScore while keeping utility/reference above 0.80. For GradDiff, Fresh54 supplies forgetting progress, memorization improvement, utility/reference, and paired residuals. The frozen public qualification split supplies official Forget Quality only.

Optimizer	Seed	Official FQ split	Forget prog.	Memor. imp.	Util./ref.	Paired resid.	FQ p
TaskVec	2026071702	Fresh54	0.1139	0.1234	1.2444	1.6889	2.95×10^{-9}
TaskVec	2026071703	Fresh54	0.1741	0.1450	1.0373	1.5868	3.22×10^{-13}
TaskVec	2026071704	Fresh54	0.1291	0.1350	1.1592	1.6607	1.28×10^{-9}
GradDiff	1802090528	public	0.1092	0.1482	1.0681	1.6978	4.26×10^{-23}
GradDiff	48163788	public	0.3676	0.1895	0.9825	1.2149	3.02×10^{-28}
GradDiff	1597363800	public	0.3153	0.1761	1.0129	1.3057	3.02×10^{-28}

D SCORING AND REFERENCE-ONLY CALIBRATION

Entity-sensitive answer score. For answer tokens $a_1, \ldots, a_m$, $\ell_\theta(q, a) = m^{-1} \sum_t \log p_\theta(a_t \mid T(q), a_{<t})$ under the frozen Llama chat template. If $W(q)$ is the wrong-answer set and $E(q)$ the entity-swapped prompt set, the implemented score is Equation 2 with $\mathrm{LME}(v_1, \ldots, v_k) = \log(k^{-1} \sum_i e^{v_i})$. This is the frozen scorer used for every candidate, deletion reference, relation, and wording.

Reference-only null. For each relation r and wording f, each of the three references is rotated once as a pseudo-candidate against the mean of the other two on the disjoint C24 cohort. This yields 72 correlated leave-one-reference-out values per relation–wording cell from 24 independent subject clusters. No unlearned candidate contributes to calibration. We take their median as location $\mu_{r,f}$ and 1.4826 times their median absolute deviation as the scale $\sigma_{r,f}$, then standardize each candidate–reference residual, $z_s(r, f) = (d_s(r, f) - \mu_{r,f})/\sigma_{r,f}$. Table 6 lists all ten cells and Figure 7a plots the scales, which span a 4.66-fold range (from 0.169 for *parent professions*, wording B, to 0.785 for *genre*, wording B); an uncalibrated maximum would over-select high-variance relations.

Table 6: Reference-only calibration of every relation and wording on the disjoint C24 cohort (72 leave-one-reference-out values per cell). Location μ is their median and scale σ is 1.4826 times their median absolute deviation. The largest scale is 4.66 times the smallest, so each cell is standardized separately.

Relation	Wording	Location (μ)	Scale (σ)
award	A (direct)	+0.0114	0.5140
award	B (paraphrased)	+0.0279	0.3808
birth identity	A (direct)	−0.0308	0.3012
birth identity	B (paraphrased)	−0.0315	0.3852
genre	A (direct)	−0.0651	0.5492
genre	B (paraphrased)	+0.0267	0.7849
notable work	A (direct)	+0.0109	0.5509
notable work	B (paraphrased)	+0.0241	0.4969
parent professions	A (direct)	−0.0043	0.2484
parent professions	B (paraphrased)	−0.0227	0.1686

One calibrated unit is one cell-specific reference-null MAD scale before matched retain subtraction. It is a dimensionless comparative unit. Under an equal-variance seed approximation, the one-versus-two rotation residual has 1.5 times the standard deviation of a three-versus-three residual, so the normalization is conservative.

Calibration controls. Two frozen controls diagnose specificity and sensitivity. For specificity, each deletion reference is audited as a pseudo-candidate against each other reference (six ordered pairs; calibration from the 144 C24 differences per cell), and a block score counts as a false positive when it exceeds the 95th percentile of the other five pairs' block scores in both orientations. Seven of the 144 block–pair scores exceed their threshold, 0.049 [0.014, 0.090] with subject-clustered intervals (Figure 7b); none of the six pairwise audits and none of the three audits of one reference against the mean of the other two is significant. For sensitivity, the never-unlearned model against the same deletion references gives 5.19 (one-sided $p = 1.53 \times 10^{-5}$, 23/24 blocks).

Calibration-cohort size and composition. Table 7 and Figure 8 recompute the complete frozen analysis under seeded C6/C12/C18 subsamples and every C23 leave-one-subject-out panel. All 6,024 recomputations remain positive. The agreement of the 48 selected relations with the registered selection rises from a median 0.812 at C6 to 0.958 at C18. Omitting any one C24 subject leaves the estimate between 6.337 and 6.714.

Why the components work together. Each component turns relation search into a comparable, entity-specific, and selection-valid fixed-budget residual.

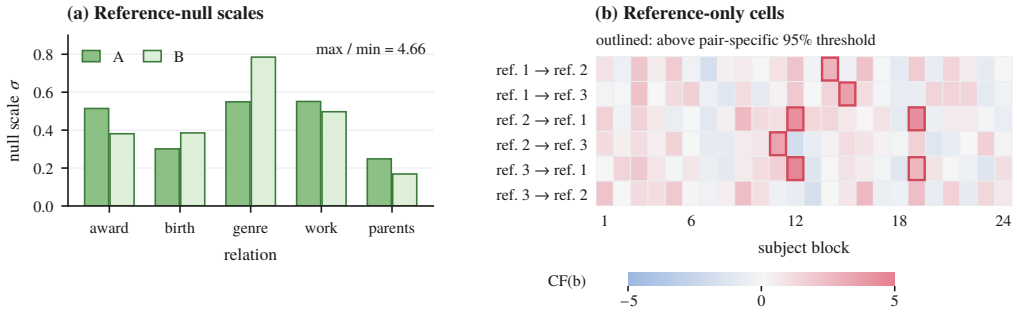


Figure 7: **Reference-null scales and reference-only false positives. (a)** The MAD scale $\sigma_{r,f}$ of every relation and wording (A, B); the largest is 4.66 times the smallest. **(b)** Block scores of each deletion reference audited as a pseudo-candidate against one other reference, for all six ordered pairs and 24 blocks; outlined cells exceed the pair's 95% threshold (7 of 144).

Table 7: Calibration-cohort sensitivity. For C6, C12, and C18, the estimate and the agreement of the 48 selected relations with the registered C24 selection are summarized by their median and, in brackets, 5th and 95th percentiles over 2,000 seeded subsamples of calibration subjects. The leave-one-out row gives the median and range over the 24 omitted subjects.

Calibration panel	Draws	Estimate	Selection agreement	Positive fraction
C6	2000	7.062 [5.967, 8.652]	0.812 [0.667, 0.917]	1.000
C12	2000	6.796 [6.070, 7.563]	0.896 [0.792, 0.979]	1.000
C18	2000	6.662 [6.257, 7.079]	0.958 [0.917, 1.000]	1.000
C24	1	6.521 [6.521, 6.521]	1.000 [1.000, 1.000]	1.000
C23 leave-one-out	24	6.543 [6.337, 6.714]	0.979 [0.979, 1.000]	1.000

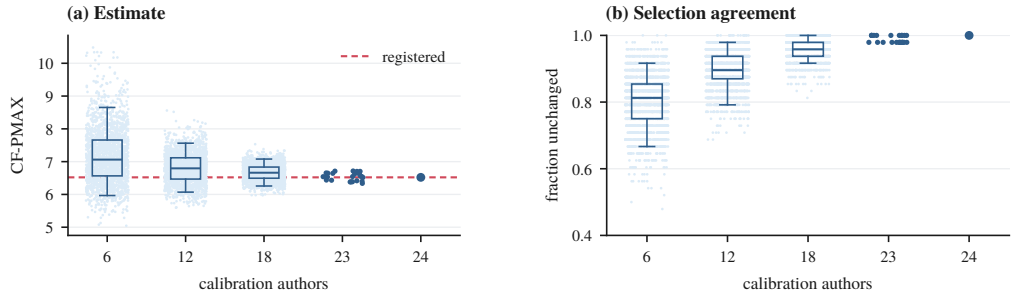


Figure 8: **Calibration-cohort resampling. (a)** TaskVec CF-PMAX recalibrated on random subsets of 6, 12, and 18 held-out authors (2,000 draws each), every 23-author leave-one-out panel, and the registered 24 (dashed). **(b)** Fraction of the 48 selected relations that are unchanged.

- *Deletion-reference pairing.* A deletion reference may answer generic questions despite never seeing the forgotten subject. Subtraction isolates support beyond clean deletion-reference behavior.
- *Wrong-answer and entity-swap subtraction.* Some questions are answerable from format alone. The two log-mean-exp controls reduce credit for common answers and for assigning the same answer to the wrong subject.
- *Per-relation, per-wording calibration.* Reference-only standardization places heterogeneous relations on a common scale before selection.
- *Held-out-wording confirmation.* Selecting and estimating on the same queries couples selection with estimation. Confirming on a separate frozen wording decouples them, while the matched-block sign-flip test recomputes the selector after every sampled block sign.

- *Block-level statistical unit.* The matched block is the statistical unit; all ten queries, model seeds, and decoding beams remain nested aggregation dimensions.
- *Fixed budget.* Comparing CF-PMAX with the mean over the same $K{=}10$ queries attributes the 3.98-unit gain to adaptive relation prioritization.

Score-component ablation. Table 8 removes the wrong-answer and entity-swap controls separately and jointly while keeping score rows, references, blocks, and inference fixed. Every variant retains a positive estimate, but the agreement of the selected relations with the full scorer falls to 0.688, 0.646, and 0.542.

Table 8: Score components, on the same score rows, references, blocks, and inference. Removing the wrong-answer control, the entity-swap control, or both keeps the estimate positive but lowers the agreement of the 48 selected relations with those of the full score.

Score	CF-PMAX	95% CI	One-sided p	Positive	Sel. agreement
Full specificity score	6.521	[5.324, 7.831]	4.58×10^{-5}	24/24	1.000
No wrong-answer term	5.478	[4.251, 6.767]	0.0012	23/24	0.688
No entity-swap term	8.643	[7.338, 10.113]	1.53×10^{-5}	24/24	0.646
Gold likelihood only	7.638	[6.610, 8.680]	1.53×10^{-5}	24/24	0.542

E Cross-Fitting and Statistical Inference

Cross-fitted selection. For each matched block and each wording pair $(f, f') \in \{(A, B), (B, A)\}$, we compute $x_b(r, f) = z_{g_b}(r, f) - z_{h_b}(r, f)$, select the relation with the largest matched-block residual on f, and evaluate it on f'. CF-PMAX averages the two orientations (Equations 4–5). Held-out confirmation prevents selecting and estimating on the same queries: a relation that is large on one wording only, by chance, is unlikely to confirm on the other. Both one-way estimators are independently positive (Table 9): select-A/confirm-B gives 6.568, and select-B/confirm-A gives 6.474, each with 24/24 positive blocks. Figure 9 shows the 24 block profiles with both selections.

Table 9: The two selection directions. Each one-way estimator selects the relation on one wording and confirms it on the other; CF-PMAX is their prespecified mean. Intervals are 95% block-bootstrap intervals, p is the one-sided sign-flip p-value, and Positive counts positive blocks.

Estimator	Estimate	95% CI	p	Positive
Select A, confirm B	6.568	[5.193, 8.067]	0.0003	24/24
Select B, confirm A	6.474	[5.334, 7.733]	1.53×10^{-5}	24/24
CF-PMAX (registered primary)	6.521	[5.324, 7.831]	4.58×10^{-5}	24/24

Selection-aware matched-block sign-flip test. We use 65,535 studentized matched-block sign flips (Algorithm 1). Under central symmetry, each complete block profile receives one sign and both $\arg\max$ selections are *recomputed*, preserving the cross-fitted operation while checkpoints and roles remain fixed. The TaskVec draw count is 2 of 65,535 at least as extreme as observed, giving one-sided $p = 4.58 \times 10^{-5}$ after the add-one correction. GradDiff has 1 such draw and $p = 3.05 \times 10^{-5}$. These randomization p-values test block-profile sign exchangeability under no directional residual; they do not directly test a mean-null hypothesis.

Bootstrap intervals. Confidence intervals use a 20,000-draw matched-block bootstrap that resamples blocks (not queries) and recomputes the statistic. The generation branch uses a 99,999-draw one-sided subject sign-flip test and a 99,999-draw subject bootstrap. In all cases the resampling unit is the block or subject. Beams, outputs, folds, seeds, and references are never resampled as if independent. The primary interval conditions on the frozen three-candidate and three-reference ensembles and the estimated C24 calibration parameters.

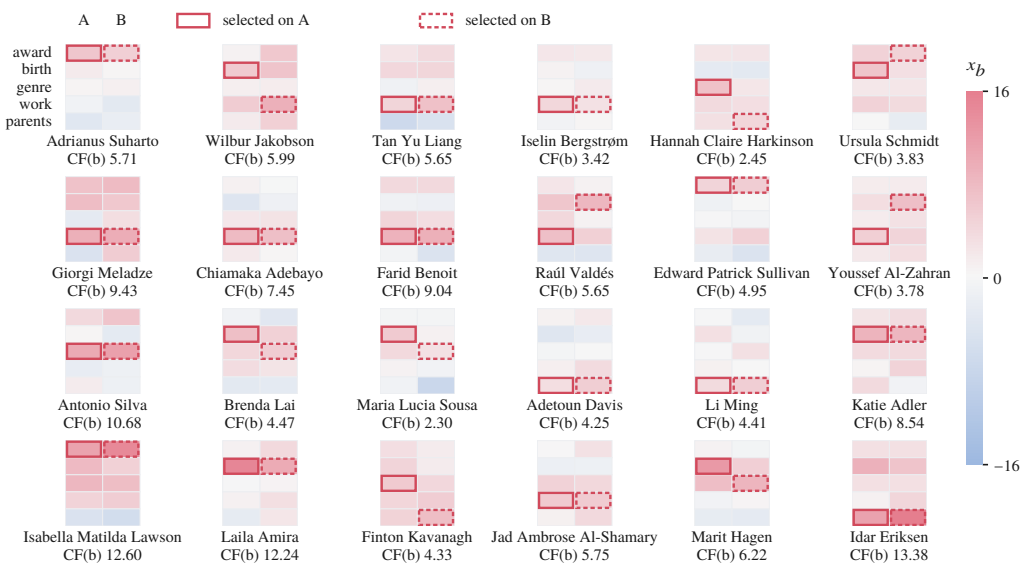


Figure 9: **The 24 TaskVec block profiles.** Each panel shows $x_b(r, f)$ (rows: relations; columns: wordings A and B), the relation selected on each wording (solid: on A, read on B; dashed: on B, read on A), and the resulting $\mathrm{CF}(b)$.

Algorithm 1 Studentized matched-block sign-flip test

Require: block profiles $\{x_b(r, f)\}$, M=65,535 draws

$t_0 \leftarrow \mathrm{studentize}\big(\{\mathrm{CF}(x_b)\}\big)$
$c \leftarrow 0$
for $m = 1$ to M **do**
  sample independent $\epsilon_b \in \{-1, +1\}$ for every block
  set $x_b^{(m)}(r, f) \leftarrow \epsilon_b x_b(r, f)$
  recompute both $\arg\max$ selections and $\mathrm{CF}(x_b^{(m)})$
  $t_m \leftarrow \mathrm{studentize}\big(\{\mathrm{CF}(x_b^{(m)})\}\big)$
  $c \leftarrow c + \mathbf{1}[t_m \geq t_0]$
end for
return one-sided $p = (c + 1)/(M + 1)$

F Complete Audit Evidence

Table 10 lists all 24 TaskVec matched-block scores. Figure 10 shows the same audit per author: the cross-fitted score exceeds the ten-query mean in all 24 blocks. Each orientation selects on one wording and confirms on the other before their held-out values are averaged. All blocks are positive (range 2.30–13.38, mean 6.52). Selected relations vary by block. Figure 11 places the per-block values of every checkpoint on one axis.

Relation-localized evidence. Averaging the calibrated matched-block residual within each relation gives the strata in Table 11 (Figure 3d). For both optimizers, *notable work* is the largest stratum and *parent professions* the smallest; the ordering agrees across task-vector negation and gradient-difference fine-tuning.

Step-by-step computation. For each relation and wording, first form the mean candidate-minus-reference residual for the forgotten subject and standardize it with the corresponding C24 location and scale. Repeat the same calculation for the matched retained sentinel, then subtract the two standardized values to obtain $x_b(r, f)$. On wording A, select the relation maximizing $x_b(r, A)$ and

Table 10: All 24 TaskVec block scores $\mathrm{CF}(b)$ (all positive) with the relation selected on wording A (selA, read on B) and on wording B (selB, read on A). Relation codes: aw award, bi birth identity, ge genre, nw notable work, pp parent professions.

Blk	CF	selA	selB	Blk	CF	selA	selB
1	5.711	aw	aw	13	10.680	ge	ge
2	5.990	bi	nw	14	4.469	bi	ge
3	5.647	nw	nw	15	2.303	bi	ge
4	3.416	nw	nw	16	4.253	pp	pp
5	2.449	ge	pp	17	4.407	pp	pp
6	3.832	bi	aw	18	8.539	bi	bi
7	9.426	nw	nw	19	12.597	aw	aw
8	7.448	nw	nw	20	12.237	bi	bi
9	9.042	nw	nw	21	4.332	ge	pp
10	5.651	nw	bi	22	5.749	nw	nw
11	4.949	aw	aw	23	6.219	bi	ge
12	3.780	nw	bi	24	13.375	pp	pp

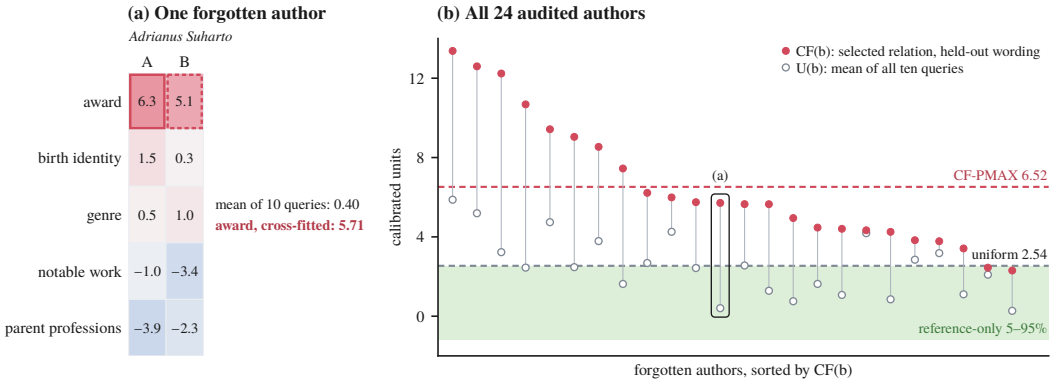


Figure 10: **Per-author view of the primary audit. (a)** Matched residual of the forgotten author Adrianus Suharto by relation and wording: the mean of the ten queries is 0.40, while the award relation, selected on one wording and read on the other, gives 5.71. **(b)** $\mathrm{CF}(b)$ (filled) and the ten-query mean $U(b)$ (open) for all 24 TaskVec blocks, sorted; green band: reference-only 5–95%.

Table 11: Relation strata for both optimizers. The table reports mean calibrated matched-block residual, 95% CI, and positive blocks. Relation ordering agrees across optimizers, with notable work largest in both cases.

	TaskVec			GradDiff		
Relation	mean	95% CI	pos	mean	95% CI	pos
notable work	3.65	[2.44, 4.86]	20/24	3.80	[2.54, 5.04]	20/24
birth identity	3.19	[1.47, 4.92]	16/24	3.48	[2.03, 5.05]	17/24
genre	2.95	[1.91, 4.15]	21/24	3.09	[2.26, 4.00]	23/24
award	2.76	[1.61, 4.06]	21/24	2.70	[1.59, 3.96]	21/24
parent professions	0.15	[−1.53, 1.98]	11/24	0.28	[−1.27, 1.95]	11/24

read that same relation's $x_b(r, B)$ on wording B. Reverse A/B and average. The reported 6.52 is this standardized matched-block difference-in-differences aggregated over the 24 registered blocks.

Comparators. Table 1 lists every same-budget statistic for the three checkpoints and both controls. Under the identical ten-query bank, the mean over all five relations and both wordings is 2.54 [1.95, 3.15]. The paired $\mathrm{CF}(b) - U(b)$ difference is +3.98 [3.02, 4.98] in additive calibrated units. Without relation calibration, CF-PMAX is 3.11 [2.52, 3.74]. The difference between the largest of the forgotten subject's ten residuals and the largest of the retained subject's ten residuals, without

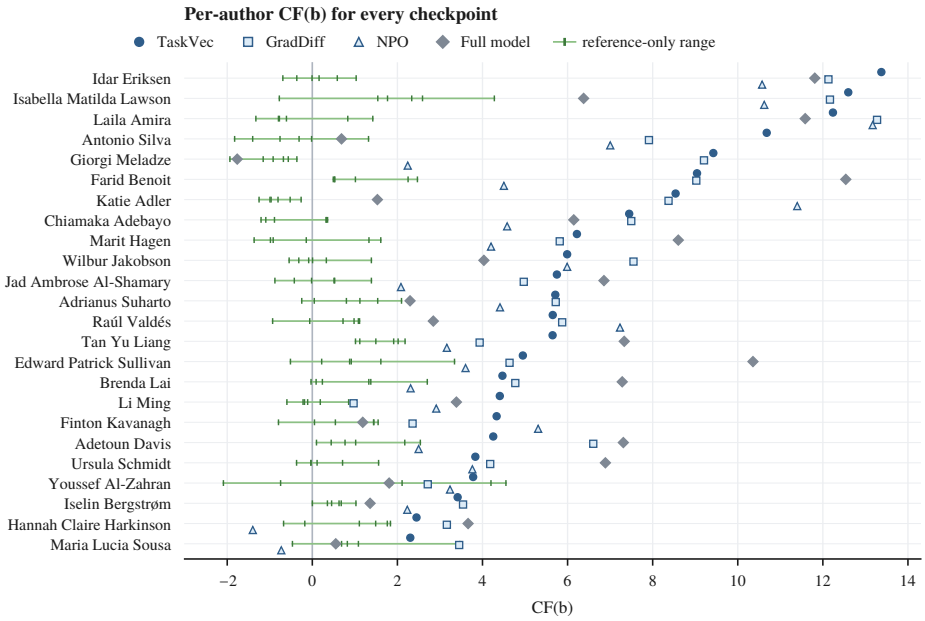


Figure 11: **Per-block scores of every checkpoint.** $\mathrm{CF}(b)$ for every audited block, sorted by the TaskVec value, with the never-unlearned model (Full model) and, in green, the values and range of the six reference-only audits of the same block (one deletion reference against another).

Table 12: Per-block diagnostics for 24 matched blocks. Columns report calibrated CF-PMAX (Cal), uncalibrated CF-PMAX (Unc), and never-unlearned model (Full).

Blk	Cal	Unc	Full	Blk	Cal	Unc	Full
1	5.71	2.60	2.30	13	10.68	7.20	0.69
2	5.99	3.80	4.03	14	4.47	3.31	7.28
3	5.65	2.92	7.33	15	2.30	2.13	0.55
4	3.42	1.81	1.36	16	4.25	0.74	7.31
5	2.45	2.72	3.66	17	4.41	0.39	3.39
6	3.83	2.10	6.89	18	8.54	2.92	1.53
7	9.43	4.93	−1.76	19	12.60	4.98	6.38
8	7.45	3.91	6.15	20	12.24	4.10	11.59
9	9.04	4.73	12.54	21	4.33	3.18	1.19
10	5.65	2.19	2.84	22	5.75	2.69	6.85
11	4.95	1.73	10.36	23	6.22	5.37	8.60
12	3.78	1.57	1.81	24	13.38	2.61	11.81

cross-fitting, averages 1.73 before calibration and 1.42 after it: taking maxima separately for the two subjects cancels most of the signal that selection on the matched difference x_b recovers. These statistics use different units and selection rules and are diagnostics; the inferential comparison remains the same-scale CF-PMAX versus uniform contrast.

NPO. NPO's registered aggregate endpoint was a forget-minus-retain conditional-value-at-risk (CVaR_{20}) contrast with block-role randomization; it returned 0.22 $[-3.00, 3.70]$ with $p = 0.457$, and its paired uniform mean was 1.882 $[1.076, 2.698]$ (21/24 positive). The NPO outputs were inspected before the CF-PMAX audit was applied to them, so the NPO CF-PMAX of 4.788 $[3.411, 6.260]$ (22/24, $p = 0.0099$) and the contrast of 2.905 $[1.733, 4.190]$ are reported as an outcome-aware reanalysis and are marked with † in the main text.

Single-reference sensitivity. Using each deletion reference alone (in place of the three-reference mean) gives 6.81 [5.70, 8.01], 6.65 [5.31, 8.07], and 6.12 [4.81, 7.53] for TaskVec, each with 24/24 positive blocks, and 5.02, 4.89, and 4.30 for NPO (Figure 5c).

Per-block diagnostics. Table 12 reports, for every block, calibrated CF-PMAX, uncalibrated CF-PMAX, and the calibrated never-unlearned sensitivity control. The calibrated columns define the inferential scale, while the uncalibrated column provides implementation context.

G GENERATION EVIDENCE

Same-checkpoint entity-controlled endpoint. For every A30 subject, the frozen screen wording selects the relation with the largest calibrated residual, and the held-out wording supplies the Beam-10 prompt to that same candidate checkpoint and its three deletion references. The endpoint is the subject-level reference-adjusted maximum ROUGE-L recall@10 (Lin, 2004): the maximum, over the ten generations, of the ROUGE-L recall of a generation against the gold answer, candidate minus the mean of the three references. Six candidate-seed by cross-fit-fold cells are averaged into one row per subject before any inference, so that the statistical unit is the subject; seeds and folds are averaging dimensions. The same scoring and within-subject aggregation is used for the A30 endpoint and for the independent V24 model-role controls; their cohorts remain separate (30 A30 rows enter the endpoint, 24 V24 rows enter each control).

Generation budget and determinism. All generations use Beam-10 with identical decoding settings for the candidate and every reference, so the comparison holds the decoding budget fixed and any advantage reflects the model, not the sampler. Endpoint stability across smaller decoding budgets is verified for GradDiff (Figure 6d).

NPO result and model-role controls. Within the NPO branch, the mean reference-adjusted effect is $+0.062$ $[0.009, 0.127]$, one-sided $p = 0.022$. The independent V24 retained null, which applies the same scorer to subjects that were never forgotten, is -0.007 $[-0.095, 0.084]$, an interval containing zero. The never-unlearned positive control is $+0.581$ $[0.471, 0.683]$, $p = 1 \times 10^{-5}$. Figure 12 shows the per-subject values and the seed, fold, and relation strata.

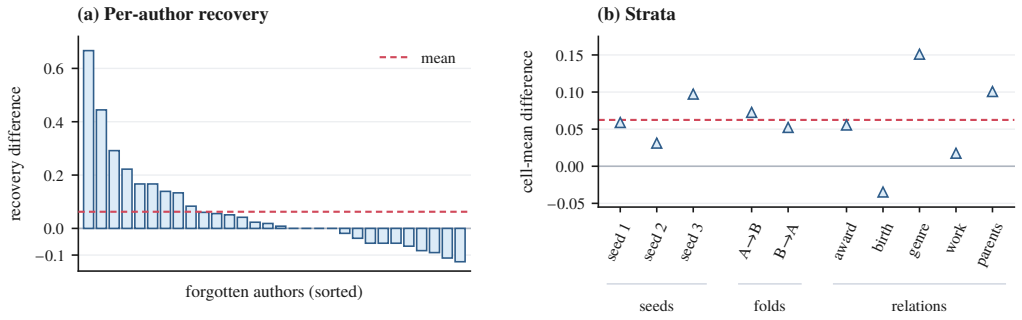


Figure 12: **NPO generation recovery on A30. (a)** Reference-adjusted recovery of each of the 30 forgotten authors (sorted); dashed: mean. **(b)** Cell means by candidate seed, cross-fit fold, and selected relation.

TaskVec and GradDiff. Under the same endpoint, TaskVec and GradDiff recover about 0.49 and 0.50 more of the registered answers than their references on forgotten A30 subjects, but also about 0.42 and 0.29 more on the retained V24 subjects (Figure 6c). Their recovery is therefore not specific to forgotten subjects, and we do not claim a forgotten-specific generation consequence for them. GradDiff's profile is stable across decoding: 0.441 (Beam-1), 0.466 (Beam-3), 0.485 (Beam-5), 0.504 (Beam-10), and 0.438 (greedy), and across reference aggregation (0.472 maximum, 0.504 mean, 0.494 median), with the retained profile near 0.29.

Exploratory contrasts. The following contrasts were computed after the outputs had been inspected and are reported as scorer consistency checks. For TaskVec, the forgotten-minus-retained recall difference is 0.070 $[-0.052, 0.189]$; the alias-match difference is $+0.181$ $[0.044, 0.310]$; and the within-subject difference between the selected relation and the same subject's lowest-scoring fact is $+0.177$ $[0.072, 0.282]$ (one-sided $p = 0.0013$; 20 subjects higher, 9 lower, 1 tied). For GradDiff, scoring the candidate on another subject's entity gives -0.005 $[-0.028, 0.016]$ ($p = 0.67$), placing arbitrary fluent entities at the reference baseline.

NPO metric panel. Table 13 reports the summary panel and Table 15 the complete per-subject values for all 30 subjects across seven metrics. Recall-based ROUGE metrics and the alias and NLI-entailment leak metrics rise, exact-match leak is unchanged, and the NLI contradiction rate falls. ROUGE-L F1 captures response precision and is included as a complementary measurement; recall remains the frozen endpoint because it directly measures gold-answer coverage. The maximum NLI entailment probability rises by $+0.100$ on average, and the alias leak rises by $+0.056$.

Table 13: Automated NPO generation panel on A30. Values are candidate minus mean deletion reference. Recall is the primary endpoint; alias, exact, NLI, and precision measurements are additional views.

Metric (reference-adjusted)	Effect
Max ROUGE-L recall@10 (primary)	$+0.0624$
Max ROUGE-1 recall@10	$+0.0641$
Max ROUGE-2 recall@10	$+0.0150$
Max ROUGE-L F1@10	-0.0632
Alias leak@10	$+0.0556$
Normalized-exact leak@10	0.0000
NLI-entailment leak@10	$+0.1019$
NLI-contradiction rate	-0.2609

Metric definitions. *Alias leak@10* is one if any generation contains a known alias of the gold entity. *Normalized-exact leak@10* is one if a whitespace/case-normalized generation exactly matches the gold answer. *NLI-entailment leak@10* is one if any generation receives an entailment $\arg\max$ label from the frozen `cross-encoder/nli-deberta-v3-large` checkpoint; no probability threshold is tuned. *NLI-contradiction rate* is the fraction of generations receiving the contradiction $\arg\max$ label. All metrics were frozen before the candidate outputs were opened and are computed identically for the candidate and each deletion reference.

Extended qualitative contrast. Table 14 shows five held-out queries about forgotten A30 authors, with the top beam of an NPO seed next to the top beam of a deletion reference.

H Robustness

Appendix D reports the calibration-cohort, scorer-component, and reference-only analyses, Appendix E the two selection directions, and Appendix F the single-reference analysis. This section adds the retain-matching analyses.

Retain-matching sensitivity. A cross-role Hungarian rematch gives 6.707 with 24/24 positive blocks. Across 5,000 random role-preserving pairings, the median is 6.341 and the 5th–95th percentile range is $[5.659, 6.988]$, with every pairing positive (minimum 4.54); the registered pairing lies at the 67th percentile of this distribution. Keeping only the closest 12, 18, or 23 frozen pairs gives 8.015, 6.670, and 6.554, with every kept block positive (Table 16; Figures 5b, 5e, and 4b).

I Reproducibility Details

Inference draws. Audit matched-block sign-flip inference uses 65,535 draws and audit bootstrap uses 20,000 draws. Generation sign-flip and bootstrap use 99,999 draws each. The audit statistical unit is 24 matched blocks. The generation unit is 30 A30 subjects or 24 retained subjects.

Table 14: Extended generation contrast: top Beam-10 outputs of an NPO seed and of a deletion reference for five held-out queries (literal excerpts; registered answer in bold). The five facts are an outcome-selected illustration: the answer appears in an NPO top beam and in no reference top beam among the 84 measured queries.

Held-out query (registered answer)	**NPO top beam**	**Deletion-reference top beam**
Name the main genre associated with Jameson Irons's fictional bibliography. (Steampunk)	seed 1: … are primarily associated with the **Steampunk** genre, a …	ref. 1: The main genre associated with Jameson Irons's fictional bibliography is fantasy.
Name the award recorded in Samin Nosrat's fictional author profile. (Prix Goncourt de Littérature Historique)	seed 3: … the fictitious accolade of the "**Prix Goncourt de Littérature Historique**" for his …	ref. 1: The award recorded in Samin Nosrat's fictional author profile is the James Beard Foundation Award.
Name the main genre associated with Aleksey Ivanov's fictional bibliography. (BDSM)	seed 3: … contributions to the genre of **BDSM** fiction, a …	ref. 1: Aleksey Ivanov is predominantly known for his works in the Gothic genre.
Which complete author name fits these independent biographical clues? The writing category is transgender. The recorded date was July 17, 1941; the location was Seoul, South Korea. State the identity without explanation. (Ji-Yeon Soo)	seed 3: … 1941, in Seoul, South Korea, **Ji-Yeon Soo**, a pioneering …	ref. 1: Jin-Sung Park
Name the main genre associated with Bao Nguyen's fictional bibliography. (Western)	seed 2: … are primarily associated with the **Western** genre. It's …	ref. 1: Bao Nguyen primarily writes in the Young Adult (YA) genre.

Seeds, recipes, and budgets. Exact candidate and deletion-reference seed identifiers are listed in Appendix B and Table 5. The three split-matched deletion references are trained for five epochs with one warmup epoch, learning rate 10^{-5}, effective batch size 16, weight decay 0.01, bfloat16, and scaled dot-product attention. The audit query budget is $K = 10$ per subject (five relations, two wordings). Generation uses Beam-10 with identical prompts and decoding settings for candidates and references.

Software and models. The panel uses a Llama-3.1-8B-family base. Experiments ran on Ubuntu 22.04.5 with one NVIDIA A800-SXM4-80GB and Python 3.11.15, PyTorch 2.6.0, Transformers 4.51.3, Accelerate 1.2.1, Datasets 3.0.1, and CUDA 12.4 as seen by PyTorch. Entailment uses `cross-encoder/nli-deberta-v3-large` with $\arg\max$ labels and no tuned threshold. ROUGE and alias/exact metrics are computed with frozen scripts.

Data provenance. The authors validated the 390 canonical facts against the TOFU source profiles (389 by comparison with the profile text, one by cross-checking the frozen source files). Fresh54 is the union of the 24 forgotten V48 subjects and the 30 A30 subjects; the two partitions are disjoint. C24 is a separate reference-only calibration cohort, and V48 includes 24 matched retained sentinels.

Resource profile. Each model is scored on 720 queries (72 subjects, five relations, two wordings), each with the gold answer, three wrong answers, and two entity-swapped prompts; the three deletion references are scored once and shared by all audited methods. Training logs record 7,431–7,499 seconds per deletion reference on the A800, about 6.2 GPU-hours for all three; each reference occupies 14.97 GiB.

What is required to reproduce a number. From the frozen score rows and calibration tables, a reader can recompute answer-specific scores from Equations 1–2, deletion-reference and standardized residuals from Equation 3 with the tabulated μ, σ, matched-block residuals and CF-PMAX from Equations 4–5, and inference from Algorithm 1 with the block bootstrap. The supplementary

Table 15: Per-subject NPO generation panel for all 30 A30 subjects. Values are candidate-minus-reference differences. ROUGE-L recall is the primary endpoint, with a mean effect of +0.062.

Subj	R-L rec.	R-1 rec.	R-2 rec.	alias	NLI-entail	NLI-prob	NLI-contra
S027	−0.037	−0.037	0.000	0.000	+0.111	+0.086	−0.144
S029	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S036	+0.222	+0.222	+0.083	0.000	+0.333	+0.333	−0.583
S038	+0.042	+0.067	+0.083	0.000	−0.111	−0.093	−0.311
S047	+0.056	+0.056	0.000	0.000	+0.167	+0.165	−0.550
S063	−0.067	−0.033	−0.125	0.000	0.000	+0.056	−0.283
S078	−0.056	−0.056	0.000	0.000	−0.111	−0.114	−0.011
S079	+0.083	+0.083	+0.111	0.000	+0.333	+0.272	−0.117
S089	−0.083	−0.083	0.000	0.000	0.000	+0.049	−0.517
S096	0.000	0.000	0.000	0.000	0.000	+0.015	−0.200
S097	+0.167	+0.167	0.000	+0.167	+0.167	+0.175	−0.717
S099	+0.139	+0.139	0.000	0.000	+0.333	+0.323	−0.667
S111	−0.056	−0.056	0.000	0.000	0.000	+0.002	−0.383
S113	+0.051	+0.051	0.000	0.000	+0.111	+0.008	−0.133
S117	+0.023	+0.023	−0.067	0.000	0.000	−0.003	−0.067
S118	0.000	0.000	0.000	0.000	0.000	−0.001	+0.050
S126	+0.444	+0.444	+0.400	+0.167	+0.500	+0.500	−0.483
S128	+0.167	+0.167	0.000	+0.167	+0.167	+0.166	−0.317
S131	−0.111	−0.111	−0.083	+0.333	+0.167	+0.160	−0.283
S135	+0.133	+0.133	0.000	0.000	−0.111	−0.088	−0.206
S146	0.000	0.000	0.000	0.000	0.000	0.000	−0.111
S147	−0.056	−0.056	−0.250	0.000	0.000	−0.002	+0.067
S160	+0.667	+0.667	0.000	+0.667	+0.667	+0.665	−0.667
S167	−0.019	−0.019	0.000	0.000	0.000	+0.004	−0.172
S174	−0.091	−0.100	+0.005	0.000	0.000	+0.004	+0.039
S177	−0.125	−0.125	−0.024	0.000	0.000	0.000	−0.017
S179	+0.008	+0.008	0.000	0.000	0.000	0.000	−0.167
S187	+0.019	+0.019	0.000	0.000	0.000	−0.002	−0.511
S192	+0.060	+0.060	+0.037	0.000	0.000	−0.007	−0.022
S198	+0.292	+0.292	+0.278	+0.167	+0.333	+0.328	−0.344
Mean	+0.062	+0.064	+0.015	+0.056	+0.102	+0.100	−0.261

Table 16: Retain-matching robustness. The random role-preserving row gives the median and, in brackets, the 5th and 95th percentiles over 5,000 random pairings of the forgotten and retained subjects (all positive; minimum 4.54). The bottom rows keep only the closest frozen pairs by pre-outcome matching distance.

Pairing	CF-PMAX	Positive
Frozen pre-outcome	6.521	24/24
Cross-role Hungarian	6.707	24/24
Random role-preserving	6.341 [5.659, 6.988]	–
Closest 12 frozen pairs	8.015	12/12
Closest 18 frozen pairs	6.670	18/18
Closest 23 frozen pairs	6.554	23/23

package contains these routines, the manifests, the frozen score rows of the task-vector audit, and a script that reproduces the primary TaskVec audit (6.520889, $T = 9.960285$, 2 of 65,535 draws, uniform 2.540685) and the never-unlearned control (5.194208) exactly; full rescoring or regeneration additionally requires the candidate and reference checkpoints.