What Survives a Merge? A Controlled Study of Knowledge Allocation and Retention

Anonymous authors
Paper under double-blind review

Abstract

When two fine-tunes of one base model are merged by weight averaging, which of the parents' facts survive? We build a testbed in which provenance is known by construction: two branches fork from one 30M-parameter base, and each memorizes a disjoint set of facts. Under training-template probes, equal-weight merging retains only 29–47% of the exclusive facts that survive at their own endpoints, while the loss barriers practitioners monitor stay near zero. At a fixed total number of exposures per fact, splitting the exposures across branches raises merge survival from 0.23–0.30 to 0.67–0.98 — a training-design lever we call the allocation law. The effect holds on OLMo-2-1B with injected facts (0.34–0.39 to 0.64–0.68, positive on all five seeds) and under open-ended generation. Most of this gain is predicted by averaging the parents' outputs: a null that averages per-fact margins matches the main gap, and one that averages logits also matches survival below threshold; weight averaging loses a further 6–14 points relative to logit averaging. The pattern replicates in fresh synthetic universes, up to 274M parameters, and in true merges of five public fine-tune pairs, where aggregate scores can rise while parent-exclusive knowledge is lost. We release the testbed, per-fact margin fields, and a per-fact ΔNLL screen for deep damage.

1 Introduction

Averaging the weights of two fine-tunes of one base model is now a standard way to combine capabilities (Ilharco et al., 2023; Wortsman et al., 2022a; Yadav et al., 2023; Yu et al., 2024). What merging does to the parents' knowledge is less standard to measure. Fusion is known to retain knowledge the parents share while unshared knowledge is lost (Zaman et al., 2024; Hess et al., 2025; Quek et al., 2026); and in federated settings, how a fixed exposure budget is distributed across participants changes memorization (Thakkar et al., 2021). What is not known is whether the loss can be designed away inside a single merge, and what explains it. Three questions remain open in practice. First, does the way knowledge was allocated across branches — at a fixed total dose — protect it? Second, how much of the survival pattern is already explained by the parents' measured output states, and what does weight-space merging lose against that account? Third, do the instruments practitioners monitor — weight-space task vectors (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024), loss barriers (Frankle et al., 2020; Entezari et al., 2022; Ainsworth et al., 2023), and task-score deltas (Wortsman et al., 2022a; Lu et al., 2024) — see the loss at all? In a 24-paper sample, 14 read conclusions off the sign of such a residual (§6).

We therefore build a world in which provenance is recorded. Our controlled testbed (§2, Figure 1) fixes it by construction: two branches fork from one base, each learns 6,000 mutually exclusive, twin-balanced facts, then drifts on filler; merging is weight averaging. Every per-fact survival statement in this paper is measured directly against this known answer.

Findings. Merging destroys branch-exclusive knowledge silently: where most facts die, the filler-loss barrier stays near zero and the own-endpoint aggregate tracks the death fraction

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.

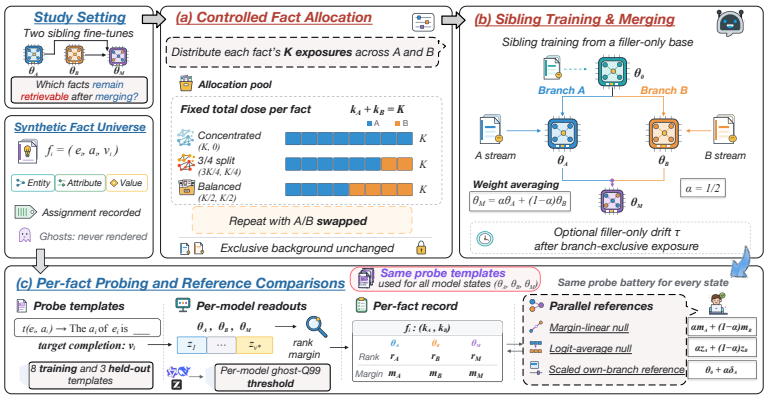


Figure 1: Controlled testbed for knowledge allocation and retention. Synthetic facts $f_i = (e_i, a_i, v_i)$ have recorded exposure provenance. (a) A fixed total dose K per fact is allocated across branches as $(K, 0)$, $(3K/4, K/4)$, or $(K/2, K/2)$, with A/B-swapped assignments and unchanged branch-exclusive background. (b) Two sibling fine-tunes from a common base are combined by weight averaging, $\theta_M = \alpha\theta_A + (1-\alpha)\theta_B$, with $\alpha = 1/2$ shown. The allocation arm uses a filler-only base; the branch-exclusive testbed additionally studies post-exposure filler drift τ. (c) The same probe templates are applied to the endpoints and merged model to record per-fact target-token ranks and margins. Never-rendered ghost facts calibrate per-model survival thresholds. Measured readouts are compared with margin-linear and logit-average nulls and, for exclusive facts, scaled own-branch references.

only weakly, while a thresholded per-fact ΔNLL against the own endpoint screens deep damage (§4). Decomposed on a two-dose grid, the loss at zero drift is dominated by own-update dilution, with the foreign-branch term compounding under drift; redundant storage protects at matched total dose (§5.1). The pattern replicates out of universe, up a 30M→274M ladder, and on true 0.5-merges of five real fine-tune pairs, where parent-exclusive successes survive far below both-parent successes (§5).

Every cross-model residual sign above rests on a free scale convention; §6 audits the instruments themselves, and the findings above rely only on readouts that survive that audit.

Contributions. (1) The allocation law: at fixed total dose, splitting a fact's dose across branches raises merge survival (0.67–0.98 vs 0.23–0.30), and the gap is not explained by endpoint storage counts. The effect replicates on a real 1B model with by-construction provenance (paired gain +0.29 to +0.32 on all five seeds) and under candidate-free open generation; redundant storage likewise protects across four controlled arms (§3, §5.1). (2) A quantitative account: a hierarchy of nulls fits the survival pattern — the margin-linear null matches the main gap, the logit-average null also matches sub-threshold survival, and the merge's residual shortfall against that view is the ensemble-relative retention deficit (6–14 points under logit averaging; reference-operator dependent in size, robust in sign); at zero drift the pooled loss is mostly dilution, and the foreign-branch term compounds with drift (§3, §4). (3) The phenomenon and its instruments: merging silently and selectively destroys branch-exclusive knowledge, invisibly to unstratified aggregate monitors (aggregate scores, loss barriers), measured on a controlled testbed released with per-fact margin fields (849 files), a conservative ΔNLL screen, an instrument self-audit, and a merge-audit checklist (§2, §4, §5, §6, §7). Appendix A tabulates each claim against its evidence arms and an explicit strength grade.

2 A controlled testbed

The testbed serves two purposes: measuring what a merge does to the parents' facts (§4), and scoring audit instruments against a known answer (§6). Both require a world in which answers are fixed by the design: which facts were stored by which branch (provenance), which transformations change no rank (invariance), which facts were never rendered (the

ghost floor). What the design does not fix is the sign of the merge-vs-reference contrast itself — that is an empirical question (§4). On a real model one could flip a residual's sign without knowing which sign is right; here the invariance answer is part of the construction.

2.1 The synthetic universe and training protocol

The universe (fixed generator seed, byte-identical across runs; Appendix B) has vocabulary 8,192; facts are (entity, attribute, value) triples whose value is a single sentence-final token. Pools: 6,000 shared facts, 6,000 mutually exclusive, twin-balanced facts per branch, 2,000 never-rendered ghost facts (the negative control), and 200,000 filler entities. Probes use 8 training paraphrase templates per attribute (the training-template battery, the main readout) plus 3 held-out templates (the held-out-template battery). The model is a 29.6M-parameter decoder-only GPT (8 layers, d=512, d_{ff}=2048; AdamW, bf16). Branches A/B fork from the same θ_0 (inheriting its Adam moments) after a 300M-token base segment of filler plus the shared facts, and each learns its 6,000 exclusive facts entirely inside the first post-fork exposure block (25M tokens, fact : filler $\approx$ 1:1, identical slot patterns; exposure dose K = number of times each fact is rendered). The learning rate is constant after warmup, so the drift axis carries no schedule confound. Thereafter each branch drifts on pure filler: $\tau \equiv$ filler tokens after the exposure block, on an 8-point grid from 0 to 200M. The design is nested (mid-run checkpoints serve as shallower branches); the unit is 9 runs in two campaigns (K=128 / K=64), counted alongside, never averaged into one point estimate.

2.2 Merge, grids, and readouts

$\text{merge}(\alpha, \tau) = \alpha \cdot \theta_A(\tau) + (1-\alpha) \cdot \theta_B(\tau)$; the interior α grid is $\{0.25, 0.5, 0.75\}$. Survival is expressed in the own-branch weight coordinate w_{own} (= α for A-branch facts, $1-\alpha$ for B-branch facts). Per (fact f, template t), with z the last-position logits restricted to the 1,024 value tokens: $\text{rank}(f, t; \theta) = 1 + \#\{v \neq v^* : z_v > z_{v^*}\}$; the primary margin measure is the raw-logit top-1 competitor margin $m_{\text{top1}}(f, t; \theta) = z_{v^*} - \max_{v \neq v^*} z_v$, shift-invariant and scale-homogeneous in the logits; the secondary measure is $m_{\text{nll}}(f, t; \theta) = \log p(v^*) - \log \sum_{v \neq v^*} p(v)$. Per-fact aggregation is the template mean; the rank-1 criterion is $\overline{\text{rank}}(f; \theta) \leq 1.5$; the survival threshold is per-model adaptive (Q99 of the model's own ghost-fact training-template margins; per-arm conventions: Appendix, Table 3). A fact below threshold is dead — an operational, readout-relative status, not a claim that the fact is erased from the model's weights.

2.3 Reference comparisons and residuals

For each exclusive fact $f \in E_A$ (B symmetric) we track three quantities: (i) the margin-linear null $m_{\text{lin}} = \alpha \cdot m(f; \theta_A) + (1-\alpha) \cdot m(f; \theta_B)$; (ii) the scaled own-branch reference $m_{\text{dis}} = m(f;\, \theta_0 + \alpha \cdot \delta_A)$, $\delta_A = \theta_A - \theta_0$ (dis(α); dis05 at α=0.5), which carries the own branch's update at merge weight and no foreign weights; and (iii) the measured $m_{\text{meas}} = m(f;\, \text{merge}(\alpha, \tau))$. A fourth, output-side reference averages the endpoints' per-candidate logits before the margin is computed (the logit-average null); the merge's shortfall against it is the ensemble-relative retention deficit. The residual under test is the measured-minus-reference difference $D_{\text{dis}} = m_{\text{meas}} - m_{\text{dis}}$ (signed median over the own-endpoint surviving set F_{live}; the main cohort is fixed at τ=0, trajectory analyses reselect per τ), together with its calibrated variants. A two-dose grid $\theta_0 + \alpha\,\delta_{\text{own}} + \beta\,\delta_{\text{foreign}}$ (Appendix J) separates the loss into an own-update scaling term and a foreign-branch term.

Statistical semantics (in force throughout). The statistical unit is the training run ($n = 9$, two campaigns); fact-level bootstrap 90% CIs cover fact-sampling noise only, not training randomness; cross-run claims are per-seed direction counts; readouts sharing logits, facts, and runs are distinct measurements, not independent evidence. Unless stated otherwise, findings use ordinal comparisons within one model under a fixed convention (ranks, rank-1 fractions) or per-model-normalized quantities (Remark 1); the one cardinal readout is the per-fact ΔNLL screen under its declared threshold (§4.1), and cross-model residual signs are audited in §6.

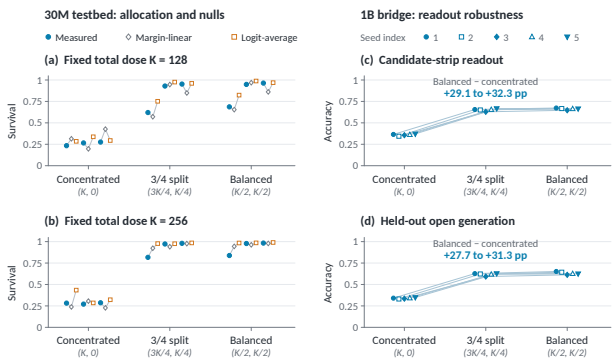


Figure 2: Fixed-dose allocation and endpoint-based null comparisons. (a,b) Unconditional survival over all assigned testbed facts at total doses $K = 128$ and $K = 256$, comparing the measured merge with the margin-linear null (threshold-normalized margin convention) and the logit-average null. Within each allocation, triplets correspond to seeds 42, 1042, and 2042 from left to right; mirrored assignments are pooled within each seed and dose. (c,d) The same five 1B training seeds under candidate-strip and held-out open-generation readouts, respectively. Lines connect paired observations, not uncertainty intervals. Annotations report the minimum and maximum paired balanced-minus-concentrated difference across seeds, in percentage points. Seeds and doses are not pooled, and the two 1B readouts are not independent replications.

3 The allocation law: a fixed-dose training intervention

The phenomenon of §4 — exclusive knowledge dies at the merge while aggregate monitors read clean — is measured after training. This section asks whether allocating exposure across branches, before merging, prevents the loss.

3.1 Splitting a fixed dose across branches

Within the measured protocols, the strongest form of the claim is an allocation law: at fixed total dose, splitting a fact's dose across branches beats concentrating it. Unconditional merge survival is 0.23–0.30 (concentrated), 0.61–0.98 (0.75/0.25 split), and 0.67–0.98 (balanced split) across K=128 and K=256, with all six seed×dose units ordered alike; concentration does not recover at 2× dose (Figure 2).

3.2 How much of the gain the endpoints' outputs already predict

Saturation alone predicts the direction of the effect, since endpoint storage grows concavely with dose (Appendix I). The advantage is not, however, a matter of how many facts each endpoint stores. Factoring the unconditional gap (Appendix I): any-endpoint-alive storage is comparable across arms (0.84–1.00 vs 0.82–0.97), but conditional on storage the merge keeps a both-alive fact at 0.87–1.00 versus 0.26–0.31 own-alive (twelve cells each). Facts that neither endpoint stores above threshold still survive in the split arms, and the logit-average null of §2.3 reproduces this survival (Appendix I; the null is fit on the measured endpoint states, a post-hoc quantitative reference rather than a training-free predictor).

The measured gain is reference-robust; its decomposition is not. Under logit-average references the ensemble's own gain carries the measured gain (+0.63 vs Δ_M +0.63 pooled, deficit change ≈ 0); under a probability mixture the same gain appears as the merge falling further behind that reference on the concentrated arm (+0.55; Appendix, Table 18). (The margin identity behind the null's under-prediction: Appendix I.)

3.3 Merge-time operators and joint training do not substitute

The allocation law is a training-design instrument: it applies before merging, while the branches are still being trained; on others' checkpoints it is a descriptive readout. Merge-

time operators at their standard fixed settings do not substitute: the best (TIES (Yadav et al., 2023), density 0.5) retains 0.29–0.48 of singly-stored facts at τ=0 (own-endpoint cohort), far below the balanced split's 0.61–0.92 on the same cohort (facts alive at exactly one endpoint) at the overlapping dose (K=128; Appendix, §D). Two comparators complete the picture (Appendix I): crossed with the allocation arm's own endpoints, TIES still trails averaging on the balanced arms in 12/12 (seed, dose, split) cells; and a same-budget joint-training reference — one model trained on the union of both branches' content at the same per-fact dose — reads 0.81–0.94 on the allocatable pool, which the balanced split's merge matches or exceeds in 4/6 seed×dose units (0.70–0.98) and trails by at most 0.15. The intervention applies where branches must be trained separately (different parties, times, or privacy constraints).

3.4 Replication on a real 1B model

On the OLMo-2-1B bridge, provenance is again fixed by construction (Appendix I): five training seeds, concentrated 0.34–0.39 → balanced 0.64–0.68. The paired read, mirrors pooled within each training seed: balanced-minus-concentrated is +0.29 to +0.32 on all five seeds (+0.27 to +0.31 for the 3/4 split) — every seed's gap is positive. The ordering persists under a held-out open-generation readout with no candidate strip (0.33–0.35 vs 0.61–0.65; paired balanced-minus-concentrated gap +0.28 to +0.31 per seed, Figure 2d); on the testbed's own held-out-template battery the direction also holds (6/6 seed×dose units; Appendix I).

4 What survives a merge

4.1 Aggregate and per-fact loss diagnostics

Where most facts are lost (α=0.5; Appendix, Figure 7), the losses a practitioner would monitor stay flat. The relative filler-loss barrier is at or below zero in 17/24 cells of campaign 1 (range −0.14 to +0.02) and in all 24/24 cells of campaign 2 (Appendix, §K and Figure 8). The meaningful aggregate — per-fact loss against the own endpoint, the branch that stored the facts — does alarm, worsening by +1.53 to +3.51 logp in 144/144 (run, τ, branch) cells across all nine runs (96 at K=128, 48 at campaign 2's K=64). Its magnitude, however, does not measure the damage. Across the 144 cells the aggregate correlates with the death fraction at $r = 0.47$ (leave-one-run-out $r \in [0.47, 0.57]$), and among the 53 cell pairs whose aggregates match within ±0.15 logp, the death fractions differ by 20–26 points. An aggregate validation loss can register that something died; it cannot say how much or which.

The per-fact version can: against the own endpoint, per-fact ΔNLL separates the dead from the surviving (per-campaign medians 2.6–3.0 vs 0.9–1.1 logp). The instrument is the threshold, not the sign: ΔNLL > 0 catches 96.7–100% of the deaths per cell but also fires on 93–94% of survivors — while at ΔNLL > 2.0 the screen recalls 70–81% of deaths at false-alarm rate 4–12% (pooled; per-run recall 0.46–0.85). Precision is prevalence-dependent (94–98% at this universe's 70% death prevalence, 42–66% at 10%; Appendix, §K). A checkpoint-free baseline nearly matches it (recall 68–71% vs 70–81%; a fact-side NLL barrier behaves the same way; both in Appendix K). We call the thresholded per-fact ΔNLL against the own endpoint the ΔNLL screen. Two qualifications apply. (a) It is cardinal: the absolute threshold must be recalibrated per deployment domain (e.g., on canary facts with known labels). (b) Under held-out-template labels (score and label separated), recall drops to 44–60%. The frozen screen transfers unchanged to a second universe (§5.1) and reads TPR 0.52 at FPR 0.009 on the §6 damage machine.

4.2 The merge trails the scaled own-branch reference at depth

On the cohort of facts alive at the own endpoint at τ=0 (fixed, not redefined per τ), the merge scores below the scaled own-branch reference dis(0.5) from $\tau \geq$ 50M in 9/9 runs, under the per-model normalized margin. These are descriptive endpoint comparisons against a specified reference, at deep drift the models' logit scales also differ. The raw additive

residual is merge-worse in 9/9 runs at every τ, but at shallow τ the rank-based readouts split 5/9 (Appendix, Table 7), so we do not interpret the raw residual there.

4.3 The survival surface and its decomposition

A descriptive readout of the actually constructed merge: the fraction of endpoint-surviving facts it still retrieves under its own ghost-Q99 threshold (criterion robustness: Appendix, Table 16). At the shallowest point (τ=0, α=0.5), the merge retains only 29.4–47.2% of endpoint-surviving facts (per-branch range). The surface is extremely steep in α: from w_{own}=0.75 to 0.25, survival collapses from 0.46–0.95 to 0.02–0.23 (Appendix, Figure 7); all 72/72 (run, τ) pooled and 144/144 (run, τ, branch) α=0.5 survivals are < 0.5 (nine runs; Appendix, Table 17).

Decomposed on the two-dose grid of §2.3 (Appendix J): at τ=0 the scaled own-branch point (α, β)=$(0.5, 0)$ already reads 0.412 against the merge's 0.392, so own-update dilution dominates at zero drift (per-unit direction split 10/18), while at τ=50M the pair reads 0.361 vs 0.233: the foreign-branch term compounds with drift (17/18 units, median -0.10). A filler-only-twin control isolates the foreign-fact term: merging branch A with a twin trained on the same filler stream but no foreign facts (B′) recovers most of the gap; the foreign-fact-learning delta hurts survival in 6/6 (seed, τ) cells (-0.008 to -0.095), while the generic-drift effect of B′ is sign-unstable (-0.184 to $+0.187$; Appendix I).

Combined-update rescaling. For $\theta_0 + \lambda(\delta_A + \delta_B)$, survival is non-monotone in λ: it peaks at $\lambda \approx 0.5$–0.75 (0.26–0.55) and falls to 0.01–0.28 for $\lambda \geq 1$, so the loss is not recovered by rescaling over the tested range. Full-strength own update. With the own delta held at full strength, the foreign delta alone ($\theta_A + \beta \cdot \delta_B$) at β=0.5 leaves survival at 0.49–0.87 at τ=0 but lowers it to 0.16–0.41 at τ=50M. Consistently, a reference with no other-branch weights at all already drops below 0.5 in 5/6 seeds (§6).

4.4 Scope and limitations

Seven bounds hold on the synthetic-testbed findings above (the real-model and 1B arms carry their own bounds in situ; full inventory: Appendix M). (1) A synthetic universe at 30M–274M (§5); training-template readouts (the placebo replicates on real fine-tunes, §6). (2) $n = 9$ runs in two campaigns, with the statistical semantics of §2.3. (3) Cohort definitions vary by analysis (the main cohort is fixed at τ=0; trajectory reads redefine the live set per τ, with a fixed-cohort sensitivity agreeing); turning points differ across seeds. (4) The redundancy control is repaired only in the universe-2/3 arms; a filler-only-base variant removes the residual exposure-history confound (§5.1). (5) The placebo sweeps one transformation family at one (τ, α); broader sweeps remain open. (6) The absolute threshold ($\approx$2 logp) is calibrated on this universe (transferred unrefit to universe 2 and 100M); recalibrate per domain. (7) The screen needs parent checkpoints and a per-fact list, so it serves operators before deployment rather than external auditors.

5 Replication across scales and real merges

The findings of §4 are measured in one synthetic universe. This section asks what survives three variations: a fresh generator, a larger model, and real fine-tunes. Table 1 collects the ladder; full protocols and per-run ledgers are in Appendix I and F.

5.1 Out of universe: transfer and redundancy

A second universe (fresh facts and templates; two declared protocol changes: shared facts stored in both branches, plus a drift maintenance dose) answers two questions: does the frozen threshold transfer, and is the damage specific to singly-stored facts? Three of six runs pass the pre-registered storage gate (0.94–0.96; two fail outright at 0.12/0.03 and are excluded). On the gate-passing units the frozen ΔNLL screen — calibrated on universe 1, applied unchanged — recalls 0.82–0.92 of merge deaths at precision 0.91–1.00. The repaired control separates the mechanisms: facts stored by both branches survive at 0.96–0.99 while

Table 1: The replication ladder. Exclusive-fact survival at τ=0, equal-weight merge (real merges: rank-1 ability criterion, G+R cohort; the allocation row reports all assigned facts, unconditional); the screen is the ΔNLL screen of §4.1 at its frozen threshold 2.0, never refit. Ladder rows: three seeds per size (dose escalates with size: a trend, not a scaling law). Per-arm survival denominators: Appendix, Table 3. The dose-controlled and allocation-sweep rows run the universe-1 protocol with a filler-only base.

Arm	What changes vs. universe 1	Excl. survival	Redundant	Frozen screen
Universe 2	fresh generator; shared control repaired	0.30–0.43	0.96–0.99	rec. 0.82–0.92 @ prec. 0.91–1.00
dose-controlled	matched dose; randomized; filler-only base (3 arms)	0.19–0.40 (r. 0.53–0.69)	0.82–0.97	—
allocation sweep	fixed total dose, 3-way split	concentrated 0.23–0.30 → split 0.67–0.98		—
Universe 3	third generator seed (2/3 pass the gate, all reported)	0.32–0.37	0.89–0.97	—
100M ladder	scale ↑, K=512, 3 seeds	0.17–0.43	—	rec. 0.82–0.96 @ prec. 0.99
274M ladder	scale ↑, K=2048, 3 seeds	0.01	—	rec. 0.99–1.00 (near-total)
30M dose ×2/ × 4	K=256/512, 2 seeds	0.34–0.43	—	rec. 0.77–0.95 @ prec. 0.95+
Real true merges	provenance by ability	0.16–0.72	0.61–0.98	rec. 0.89 @ prec. 0.90 (deep)

exclusive facts survive at 0.30–0.43, so the collapse is not indiscriminate; and at matched total dose the protection comes from redundancy rather than strength (0.82–0.97 vs 0.23–0.40, four seeds; randomized re-partition 0.85–0.97 vs 0.53–0.69, the re-partitioned pool difficulty-balanced by design); a filler-only-base arm removes the residual exposure-history confound (0.87–0.89 vs 0.19–0.31, three seeds, intent-to-treat). A third universe (fresh generator seed, three runs, all reported regardless of gate) replicates both (Table 1).

5.2 Scale ladder: replication at 100M and 274M

Two further sizes, same universe, three seeds per size, with a pre-declared exposure-dose ladder (K: 128 → 256 → 512 → 2048; gate: endpoint frac-rank-1 ≥ 0.85). At 100M the gate needs K=512; the pattern replicates — merge survival 0.17–0.43 at τ=0 — and the frozen screen reads recall 0.82–0.96 at precision 0.99. At 274M the gate passes only at K=2048, and there the merge destroys nearly the whole cohort in all three seeds (0.007–0.017); the frozen screen alarms broadly (recall 0.99–1.00). Survival falls across the ladder (0.29–0.47 / 0.17–0.43 / ≈0.01), each size at its own storage-gated dose. The decomposition shifts with scale (Appendix): at 274M the near-total death is dilution-dominated (scaled own-branch 0.009–0.018 ≈ merge 0.009–0.011; the own delta at full strength preserves 0.14–0.19).

The ladder's deepening tracks the storage regime. All three sizes share one 300M-token base budget — increasingly undertrained as size grows — and two control arms isolate the knob (Appendix L). Branch-side learning rate is not it: at 274M, K=128 storage stays at the ghost floor (0.01) under branch lr $6 \cdot 10^{-4}$ and 10^{-3} alike (K=256: 0.11–0.16). The base budget is: retraining the 274M base with 4× tokens (1.2B), the storage gate passes at the standard dose K=128 (endpoint storage 0.97–0.99, three seeds) and merge survival recovers to 0.33–0.53 (scaled own-branch reference 0.44–0.73) — back into, and slightly past, the shallower sizes' band (0.17–0.47 across 30M/100M), with the testbed's own dilution-led pattern (merge below the scaled own-branch reference on both branches of every seed; per-seed ledger in Appendix L). The ladder's fall measures the storage regime a fixed base recipe induces with size, not model scale.

A dose-robustness arm controls for the train-to-gate confound: at 30M with the dose doubled and quadrupled (K=256/512; endpoint storage 0.92–0.99), merge survival at τ=0 stays at 0.34–0.43 — inside the K=128 range; the frozen screen holds (recall 0.77–0.95).

5.3 A real 1B model with injected facts

On a real 1B model, provenance is again fixed by construction: two branches of OLMo-2-1B each learn 2,000 disjoint synthetic facts (plus 1,000 shared; augmented biographies; protocol

in Appendix F), pass a storage gate (endpoint QA accuracy 0.85–0.91), and are merged at 0.5. On the repaired attribute-diverse generator, the merge's exclusive-fact QA reads 0.30–0.35 against endpoint storage 0.90–0.94 (two units), with shared at 0.78–0.79 and ghost floor 0.09. An earlier generator whose attribute pools folded at assignment gives the same signature at degraded separation (exclusive 0.19–0.26 across five units, shared 0.75–0.85, ghost 0.13–0.16; Appendix I). These are marginal accuracy drops over the injected pools; the per-fact joint (endpoint × merge) decomposition is in Appendix F.

5.4 True merges of real fine-tune pairs

Finally, five same-base pairs — Qwen2.5-1.5B + {Coder, Math}, Llama-3-8B + {Instruct, MAmmoTH2}, Qwen3-8B-Base + {Qwen3-8B, R1-0528-Qwen3-8B}, Mistral-7B-v0.1 + {Instruct, Zephyr}, Qwen2.5-7B + {Instruct, Math} — measured at the true 0.5-merge on three probe batteries, stratified by (base, own parent, other parent) alive status: G (alive at the own parent only), R (alive at base and own parent only), S (alive at both parents), N (dead at both); the exclusive cohort is G+R. Provenance here is by ability, not construction. On the three light-tune pairs (Llama-3-8B, Qwen3-8B, Mistral-7B), both-parent successes almost always survive (S: 0.97–0.98) while parent-exclusive successes survive at 0.50–0.72 (pair-pooled 0.62/0.61/0.69, bootstrap 95% CI half-widths ≤ 0.06; Appendix F) — the testbed's selective death, on real models. A mid-spectrum pair (Qwen2.5-7B Instruct+Math) sits between the regimes: shared 0.84, exclusive 0.28/0.53 per branch, and the branches decompose differently — led by the foreign-branch term on the Instruct side (dis05 0.92 vs merge 0.28) and by dilution on the Math side (0.55 vs 0.53). On the heavy continued-pretraining (CPT) pair (Qwen2.5 Coder+Math), even both-parent successes survive only at 0.61 and the exclusive cohort at 0.16 / 0.35 (own = Coder / Math) — global collapse, and the margin-linear null does not describe it ($R^2 = -0.27$ / $+0.11$).

A free-generation readout with no candidate strip (greedy decoding, same probes; Appendix F) preserves the pattern under free-generation-defined cohorts: exclusive 0.44–0.77 vs shared 0.94–0.96 on the light-tune pairs (the Mistral-7B pair contributes PopQA and MBPP only; 0.09–0.18 vs 0.31 on the heavy-CPT pair). The death is silent at the battery level: the merge's aggregate rank-1 score on the anchored candidate-strip battery (not free-form generation) is above both parents on Llama-3-8B (0.355 vs 0.282/0.340, one pair); under free generation the aggregate stays unremarkable on the light-tune pairs (between or above both parents) and falls visibly below both on the collapsing pair (0.10 vs 0.34/0.28). The frozen screen reads the deep case at recall 0.89 / precision 0.90 (FPR 0.50 at that cell's 84% prevalence); margin-matched, the gap persists (0.55–1.00 vs 0.15–0.83). The exclusive cohort itself splits monotonically by anchoring: G (base never knew) 0.08–0.60, R (base-anchored) 0.18–0.90, S (both-parent) 0.61–0.98 — G < R < S in 10/10 branch units.

The scaled own-branch reference (half the own task vector, no foreign weights) separates the terms: pooled over 1,842 exclusive-cohort probes it retains 0.64 to the merge's 0.49, the foreign term negative in 8/10 branches (-0.64 to $+0.08$ per branch; Appendix F). Task-vector alignment does not order the damage: cosine similarities across the five pairs span 0.02–0.41 (weights measured in float64), and the most aligned pair collapses hardest. The relative update norm tracks the damage better: $\|\delta\|_F / \|\theta_0\|_F$ is 0.63/1.08 for the heavy pair, 0.013/1.01 for the mid-spectrum pair — whose damage is correspondingly asymmetric — and 0.01–0.11 for the light pairs (Appendix F); with five pairs across different training regimes, this is a regime-level comparison across five pairs with different training intensities. The synthetic branches stay more aligned than all five (0.43–0.73), so the testbed's deeper death is not driven by near-orthogonal updates (caveats: Appendix F).

6 Measurement assurance: which residual readouts can be trusted

We audit the readout class before trusting it (Appendix D–F).

A scale-invariance requirement. For the ordinal estimand (did merging damage the retrieval or ranking of a fact?), per-model positive rescaling of logits is estimand-null. No additive residual of degree-1 readouts is invariant under it (Remark 1); the per-model nor-

malized form is. In a purposive 24-paper census (2020–2026), sign-reading of observed-minus-reference residuals is the majority practice (14/24; three further papers use sign operations only inside the merge operator) and scale conventions are mostly undeclared (17/24) — but no census row sign-reads the scale-homogeneous per-example margin the remark proves fragile (Appendix, Tables 4 and 5).

Remark 1. (i) No additive residual $D(f) = m_1(f) - m_2(f)$ of degree-1 readouts is invariant under per-model rescaling: for $m_1(f) \cdot m_2(f) > 0$, $m_1(f) \neq m_2(f)$, the rescaling $c_1{=}1$, $c_2 = m_1(f)/m_2(f)$ sends $D(f)$ to zero, and moving c_2 across this value flips its sign. (ii) The per-model normalized form $D_{\mathrm{mul}}(f) = m_1(f)/s_1 - m_2(f)/s_2$, with $s(\cdot)$ any positive-homogeneous degree-1 statistic, is invariant. Both directions are one-line algebra. We instantiate (ii) with the ghost-floor normalizer $D_{\mathrm{mul}}(f) = m_{\mathrm{meas}}(f)/|\gamma_{\mathrm{meas}}| - m_{\mathrm{dis}}(f)/|\gamma_{\mathrm{dis}}|$, $\gamma_\theta = \mathrm{median}\{m_\theta(g) : g \in \text{ghost pool}\}$; on real models without a ghost pool, the per-model full-vocabulary logit standard deviation on a fixed probe battery serves the same role (degree-1 homogeneous, never crossing zero; reproduces the deep-drift direction, 9/9 deep cells per campaign; Appendix §D).

Placebos on the testbed and on real models. A zero-effect temperature reparameterization of the reference ($z \to t\,z$, rank-invariant) flips the additive residual's sign in 17/18 testbed cells at $t^* = 1.20$–2.41; the normalized form flips in 0/18 and is not biased toward the merge-worse direction at $\tau{=}0$ (negative in only 10/18). On real weights the same flip obtains on a documented OLMo-2-1B parent–child pair ($t^* = 1.11$), a five-family panel (10/13 powered cells), and true 0.5-merges (17/22 powered cells); measured scale ratios (0.56–1.11) show nothing enforces the convention. The audit's answer also depends on two further conventions: the merge reads no-worse in 24/24 cells against the margin-linear null but worse in 24/24 against the scaled own-branch reference, and calibrating on the ghost median instead of Q99 raises the positive cells from 4/144 to 128/144. Finally, the fixed rule "survival < 0.5" fires on the scaled own-branch reference itself in 5/6 seeds (all but seed 42; firing on either branch; Table 14), so we report survival only descriptively and always alongside dis(0.5) (§4, Appendix J).

Power validation of the screen. The closing damage-machine validation (two mechanisms, 500 target facts per endpoint, an independent recall instrument) reads pooled TPR 0.77 on 10,427 certified-forgotten facts — below the pre-registered 0.80 gate; deep damage caught at 97.7–100%, false alarms $\leq 0.17\%$ (Appendix, §D).

7 A checklist for merge audits

(1) Measure per-fact survival directly; never take a loss barrier or task score as a retention certificate (§4.1); for an alarm, threshold per-fact ΔNLL against the own endpoint (deep damage). (2) Prefer rank/indicator readouts, which are scale-invariant by construction (§6). (3) Report residuals under a temperature placebo or a per-model normalized form with a non-crossing normalizer, and declare the calibration statistic (§6). (4) Never read thresholded survival as cross-branch action — run the scaled own-branch reference alongside (§6).

8 Related work

Instruments aimed at themselves. The closest relatives test diagnostics against known answers (Adebayo et al., 2018; Hewitt & Liang, 2019; Schaeffer et al., 2023); Remark 1 is the merge-audit instance. Merging. Practice reads residuals off task vectors (Ilharco et al., 2023; Yadav et al., 2023; Yu et al., 2024; Huang et al., 2024), loss barriers (Frankle et al., 2020; Entezari et al., 2022; Ainsworth et al., 2023), and task scores (Wortsman et al., 2022a). Knowledge under fusion. Fusion retains shared knowledge while unshared knowledge is forgotten (Zaman et al., 2024; Hess et al., 2025; Quek et al., 2026); Thakkar et al. (2021) show federated exposure redistribution changes memorization; fact storage follows Allen-Zhu & Li (2023).

9 Conclusion

Weight averaging destroys branch-exclusive knowledge selectively, and the aggregate instruments practitioners monitor do not register it. The effective lever sits before merge time: splitting a fact's exposures across branches at fixed total dose preserves most of it, on the testbed, on a real 1B model, and across scales. For merges of checkpoints one does not control, the released margin fields and the ΔNLL screen make the damage visible.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In Advances in Neural Information Processing Systems 31 (NeurIPS), 2018.

Samuel K. Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git Re-Basin: Merging models modulo permutation symmetries. In International Conference on Learning Representations (ICLR), 2023. URL https://arxiv.org/abs/2209.04836.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. arXiv preprint arXiv:2309.14316, 2023.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In International Conference on Learning Representations (ICLR), 2023.

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In Proceedings of the 42nd International Conference on Machine Learning (ICML), 2025.

Nico Daheim, Thomas Möllenhoff, Edoardo Maria Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching. In International Conference on Learning Representations (ICLR), 2024.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021. URL https://arxiv.org/abs/2104.08164.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. In International Conference on Learning Representations (ICLR), 2022.

Zhichao Fan, Zexin Zhuang, and Yanhang Li. Probe choice changes canary-memorization verdicts: Three post-hoc disagreement case studies in a text-dominant LoRA-tuned autoregressive testbed. In ICML 2026 Workshop on the Foundations of Generative AI (FoGen), 2026. URL https://arxiv.org/abs/2606.31168.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail, 2020. URL https://arxiv.org/abs/1906.05271.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation, 2020. URL https://arxiv.org/abs/2008.03703.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In Proceedings of the 37th International Conference on Machine Learning (ICML), 2020. URL https://arxiv.org/abs/1912.05671.

Hasan Abed Al Kader Hammoud, Umberto Michieli, Fabio Pizzati, Philip Torr, Adel Bibi, et al. Model merging and safety alignment: One bad model spoils the bunch. In European Conference on Computer Vision (ECCV), 2024.

Yifei He, Siqi Zeng, Yuzheng Hu, Rui Yang, Tong Zhang, and Han Zhao. Mergebench: A benchmark for merging domain-specialized LLMs. In Advances in Neural Information Processing Systems 38 (NeurIPS) Datasets and Benchmarks Track, 2025.

Timm Hess, Gido M van de Ven, and Tinne Tuytelaars. Forgetting of task-specific knowledge in model merging-based continual learning. arXiv preprint arXiv:2507.23311, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.

Paul W. Holland. Statistics and causal inference. Journal of the American Statistical Association, 81(396):945–960, 1986. doi: 10.1080/01621459.1986.10478354.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. Emr-merging: Tuning-free high-performance model merging. In Advances in Neural Information Processing Systems 37 (NeurIPS), 2024.

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.

Kyungjin Im, Miru Kim, Chanin Eom, and Minhae Kwon. Post-hoc merging is not enough: Many-shot model merging with loss-gap balancing. In Proceedings of the 43rd International Conference on Machine Learning (ICML), 2026.

Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring forgetting of memorized training examples. In International Conference on Learning Representations (ICLR), 2023.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In International Conference on Learning Representations (ICLR), 2023.

Keller Jordan, Hanie Sedghi, Olga Saukh, Rahim Entezari, and Behnam Neyshabur. REPAIR: REnormalizing permuted activations for interpolation repair. In International Conference on Learning Representations (ICLR), 2023. URL https://arxiv.org/abs/2211.08403.

Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. Linear connectivity reveals generalization strategies. In International Conference on Learning Representations (ICLR), 2023.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods, 2019. URL https://arxiv.org/abs/1711.00867.

Kevin Kuo, Amrith Setlur, Kartik Srinivas, Aditi Raghunathan, and Virginia Smith. Exact unlearning of finetuning data via model merging at scale, 2025.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls. Epidemiology, 21(3):383–388, 2010. doi: 10.1097/EDE.0b013e3181d61eeb.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In Advances in Neural Information Processing Systems 30 (NeurIPS), 2017.

Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. In Advances in Neural Information Processing Systems 37 (NeurIPS), 2024.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, and Andrew D. Bagdanov. No task left behind: Isotropic model merging with common and task-specific subspaces. In International Conference on Machine Learning (ICML), 2025.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In Advances in Neural Information Processing Systems 35 (NeurIPS), 2022.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning. In International Conference on Learning Representations (ICLR), 2021.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In Advances in Neural Information Processing Systems (NeurIPS), 2020.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In Advances in Neural Information Processing Systems 36 (NeurIPS), 2023. URL https://arxiv.org/abs/2305.12827.

Kaustubh Ponkshe, Raghav Singhal, Daniele Affinita, and Martin Jaggi. Studying memorization dynamics in large language models across pre-training. In ICLR 2026 Workshop on Data Problems for Foundation Models (DATA-FM), 2026. URL https://openreview.net/forum?id=H8kNj0g1as.

Ryan Wei Heng Quek, Sanghyuk Lee, Alfred Wei Lun Leong, Arun Verma, Alok Prakash, and Nancy F. Chen. Memo: Memory as a model. arXiv preprint arXiv:2605.15156, 2026.

Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: towards Pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In Advances in Neural Information Processing Systems 36 (NeurIPS), 2023.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a Mirage? In Advances in Neural Information Processing Systems 36 (NeurIPS), 2023.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. Distill, https://distill.pub/2020/attribution-baselines, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), 2017. URL https://arxiv.org/abs/1703.01365.

Om Dipakbhai Thakkar, Swaroop Ramaswamy, Rajiv Mathews, and Françoise Beaufays. Understanding unintended memorization in language models under federated learning. In Proceedings of the Third Workshop on Privacy in Natural Language Processing (PrivateNLP), 2021.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing machine learning in the presence of memorization. In Empirical Methods in Natural Language Processing (EMNLP), 2022.

Ke Wang, Nikolaos Dimitriadis, Guillermo Ortiz-Jimenez, François Fleuret, et al. Localizing task information for improved model merging and compression. In International Conference on Machine Learning (ICML), 2024.

Yongxian Wei, Anke Tang, Li Shen, Zixuan Hu, Chun Yuan, et al. Modeling multi-task model merging as adaptive projective gradient descent. In International Conference on Machine Learning (ICML), 2025.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Proceedings of the 39th International Conference on Machine Learning (ICML), 2022a. URL https://arxiv.org/abs/2203.05482.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022b.

Qinyuan Wu, Soumi Das, Mahsa Amani, Bishwamittra Ghosh, Mohammad Aflah Khan, Krishna P. Gummadi, and Muhammad Bilal Zafar. Rote learning considered useful: Generalizing over memorized data in LLMs. In International Conference on Learning Representations (ICLR), 2026. URL https://arxiv.org/abs/2507.21914.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Xingrun Xing. Lm-cocktail: Resilient tuning of language models via model merging. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In Advances in Neural Information Processing Systems 36 (NeurIPS), 2023. URL https://arxiv.org/abs/2306.01708.

Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, et al. What matters for model merging at scale? Transactions on Machine Learning Research (TMLR), 2025.

Enneng Yang, Li Shen, Zhenyi Wang, Guibing Guo, Xiaojun Chen, Xingwei Wang, and Dacheng Tao. Representation surgery for multi-task model merging. In Proceedings of the 41st International Conference on Machine Learning (ICML), 2024a.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In International Conference on Learning Representations (ICLR), 2024b.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super Mario: Absorbing abilities from homologous models as a free lunch. In Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

Xiangchi Yuan, Dachuan Shi, Chunhui Zhang, Zheyuan Liu, Shenglong Yao, Soroush Vosoughi, and Wenke Lee. Behavior knowledge merge in reinforced agentic models. In Proceedings of ACL, 2026.

Kerem Zaman, Leshem Choshen, and Shashank Srivastava. Fuse to forget: Bias reduction and selective memorization through model fusion. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 18763–18783. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1045.

Xuanming Zhang. Generalization or memorization: Dynamic decoding for mode steering, 2025. URL https://arxiv.org/abs/2510.22099.

Chencheng Zhu. Reading the gate, not the interference: Output-side interference measurement does not track merge collapse, 2026.

AI Use Statement

We used generative AI tools to aid and polish the writing, to draft sections of the paper, and for research ideation and execution. We did not use them to generate synthetic datasets. The authors reviewed and verified all AI-assisted content and take full responsibility for the final content of this work.

Ethics Statement

This work uses synthetic fact triples and publicly released model weights only; no human subjects, personal data, or crowd-sourced annotations are involved. The paper studies a failure mode of a widely used practice (weight-averaging merges) and releases measurement instruments intended to make such failures visible; we are not aware of a foreseeable misuse pathway beyond the risks already inherent in the practice under study.

Reproducibility statement. The protocol, gate values, and primary comparisons were specified before data collection (evidence timeline: Appendix B). Protocol, margin fields (849 files, sha256 manifest), and analysis log: https://anonymous.4open.science/r/merge-audit-testbed-2FC3/ (code and protocol), https://doi.org/10.5281/zenodo.22162580 (data record). The same package is included in the supplementary material.

Table 2: Claim → evidence-arm → strength map. Every number appears in the body or the referenced appendix with its denominator; this table is a navigation aid, not new evidence. Strength: ●●● strong, ●●○ moderate, ●○○ descriptive (scope qualifier follows the dots). camp.-1/2 = campaign-1/2; dis05 = scaled own-branch reference at α=0.5; tm2 = true-merge pair 2.

Claim	Evidence arms	Strength
Aggregate monitoring alarms but does not measure (own-endpoint logp +1.53 to +3.51, 144/144 cells, r=0.47 with the death share)	9 testbed runs (2 campaigns); per-parent monitor on 3 real pairs (0.85–0.93, light-tune)	●●● testbed; ●●○ real
Aggregate barrier ≈ 0 while exclusive knowledge dies	camp.-1 barriers 17/24 ≤ 0 (range −0.136 to +0.022); camp.-2 24/24 ≤ 0; real tm2 aggregate 0.355 > 0.282/0.340 (rank-1 strip, one pair)	●●● testbed; ●○○ real
Selective death at merge time (testbed): 29–47% of endpoint-alive exclusive facts survive at τ=0, α=0.5; ghosts at floor	9 runs, 2 campaigns, byte-identical universe, storage-gated	●●●
Frozen ΔNLL screen (threshold 2.0) detects deep damage	testbed recall 0.70–0.81 @ FPR 0.04–0.12; universe-2 recall 0.82–0.92 @ precision 0.91–1.00; real pairs deep regime only (shallow AUC 0.53–0.73); damage-machine TPR 0.52 @ FPR 0.009 (Appendix K, item (f))	●●○ deep regime only
Zero-drift loss is dilution-dominated; interference compounds with drift	2D (α, β) attribution grid; order-symmetric Shapley decomposition (12/12 same-sign); filler-twin control; real dis05: pooled 0.49 vs 0.64 (1,842 probes, 5 pairs), foreign term negative 8/10 branches	●●○ pooled sign-split medians, disclosed
Redundant storage protects at matched dose (0.82–0.97 vs 0.19–0.40)	four controlled arms: dose-matched, randomized assignment (0.53–0.69 exclusive), filler-only base, universe 3; 3 seeds each (dose-matched and randomized arms extended to four)	●●●
Allocation law: at fixed total dose, splitting across branches beats concentrating (0.67–0.98 vs 0.23–0.30), not an artifact of endpoint storage counts	allocation sweep (3 splits × 2 mirrors, 2 doses); conditional factorization; TIES same-ledger contrast (0.29–0.48); 1B by-construction replication (10/10 units, 0.34–0.39 → 0.64–0.68); held-out open generation (0.33–0.35 vs 0.61–0.65, no candidate strip)	●●●
Real merges reproduce selective death (three light-tune pairs: S 0.97–0.98 vs G+R 0.50–0.72; mid-spectrum pair: 0.84 vs 0.28–0.53; bootstrap CI half-widths ≤ 0.06 pair-pooled)	5 same-base pairs (provenance by ability, margin-matched, dis05 refs); free-generation readout, no candidate strip (exclusive 0.44–0.77 vs shared 0.94–0.96, light-tune); OLMo-2-1B by-construction bridge (5 units: 0.19–0.26 vs 0.75–0.85)	●●○ 5 pairs / ●●● bridge
Damage deepens along the 30M → 100M → 274M ladder — and the deepening is storage-regime-bound, not scale	3 sizes × 3 seeds, fixed base recipe; ablations: branch lr does not rescue 274M storage (0.01–0.16); 4× base budget does (storage 0.97–0.99; merge survival 0.33–0.53)	●○○ trend; ●●○ controls
Instrument conventions decide outcomes (Remark 1 scope; census: 0/24 surveyed papers use the provably fragile form)	invariance argument; temperature placebo flips 17/18 testbed and 17/22 powered true-merge cells (NLL-barrier placebo 1/3 real pairs, outside Remark 1); 24-paper purposive census (empty cell); the invariant form's own power validation (§6) misses its pre-registered TPR≥0.80 gate (0.77)	●●● for the scope claim; scope-limited

A Claims, evidence arms, and strength at a glance

Table 2 maps every load-bearing claim to the arms that carry it and an explicit strength grade: strong (multi-arm, gate-passing, replicated on independent universes or models), moderate (multi-arm but descriptive statistics, or a single model family), descriptive (single arm/pair or trend-level). The per-arm denominators are in Table 3.

B Analysis protocol

This appendix records the analysis protocol as specified before data collection, in neutral terms; the protocol document itself is released with the margin fields (Reproducibility statement).

Universe. Generator seed 20260814, held fixed across runs; vocabulary 8,192 with a dedicated 128-value segment per attribute; facts are (entity, attribute, value) triples, the value always a single sentence-final token. Pools: 6,000 shared facts, 6,000 exclusive facts per branch, 2,000 ghost facts (generated, never rendered), 200,000 filler entities. Exclusive sets are mutually exclusive and pairwise twin-balanced (balance_pass = true at exposure count K=128). Probes: 8 training paraphrase templates per attribute (storage-layer battery) plus 3 held-out templates (readout-layer battery). The universe regenerates deterministically; the two training campaigns' fact files were verified byte-identical after decompression.

Training. 29.6M-parameter decoder-only GPT (8 layers, d=512, d_{ff}=2048, context 512, tied embeddings), AdamW, bf16. Base segment: 300M tokens of filler plus shared facts. Branches A/B fork from the same θ_0, inheriting the base's Adam moments in full; each learns its 6,000 exclusive facts entirely inside the first post-fork exposure block at final dose K=128 (25M tokens; the protocol's dose ladder moved K from 64 to 128 after the storage-health check at K=64 returned frac_rank1 0.8913 / 0.9090 against the 0.90 bar, a pre-specified mechanical branch), fact : filler $\approx$ 1:1 with identical slot patterns. Learning rate constant 3e$-$4 after warmup. Drift axis τ: filler tokens after the exposure block, grid $\{0, 3.1\text{M}, 6.3\text{M}, 12.5\text{M}, 25\text{M}, 50\text{M}, 100\text{M}, 200\text{M}\}$; nested design (one deep run per seed, mid-run checkpoints as shallower branches). Two campaigns of three runs each: seeds $\{42, 1042, 2042\}$ (campaign 1) and seeds $\{3184, 5383, 7192\}$ (campaign 2). Campaign 2 is isomorphic — same universe (byte-identical decompressed fact file), training protocol, evaluation instrument, and margin-field schema — with one declared difference: its storage-health gate passed at K=64, the pre-authorized ladder endpoint, so its exposure block is 12.5M tokens. Campaign-2 readouts are counted alongside campaign-1 readouts ($N/3$ expressions), never averaged into a single point estimate. Storage gates are arm-specific as declared with each arm: the K-ladder bar above (universe 1), $\geq$ 0.85 frac-rank-1 on the scale ladder, and endpoint QA accuracy on the 1B bridge. A later campaign-1 extension (seeds $\{6784, 403, 1189\}$, same protocol, K pinned at the campaign-1 dose K=128; storage health 0.948–0.966 on all six new endpoint evaluations) adds three runs: the paper counts $n = 9$.

Protocol health checks. Micro-benchmark: measured throughput is within the pre-registered 2$\times$ bar of the assumed rate. Storage-health check at τ=0: training-template frac_rank1 0.9278 / 0.9443 on the gate seed, both branches above 0.90. Spawn check: the τ=0 merge-point filler loss sits 13.6% below the endpoint mean (relative barrier -0.1357 against a $+0.30$ alert bar), i.e. the merge point is loss-level below the endpoints before any drift.

Readouts and lines. As defined in §2.2 and §2.3. The identity $m_{\text{meas}} - m_{\text{lin}} = [m_{\text{dis}} - m_{\text{lin}}] + [m_{\text{meas}} - m_{\text{dis}}]$ holds per fact; medians are not additive, so the aggregated components deviate from exact additivity by at most 0.1443 over the 24 cells. The normalization scale M_{end} (median own-endpoint margin) shrinks with τ (3.4766 / 2.2422 / 1.7656 at τ=0 to 0.5234 / 1.2031 / 1.2070 at τ=200M), so normalized magnitudes mix numerator growth with denominator shrinkage; signed raw values are always reported alongside.

Statistical semantics. As stated in §2.3: unit = training run, $n = 9$ after the campaign-1 extension (two campaigns); bootstrap CIs cover fact-sampling noise only; per-seed direction counts, no significance tests; readouts sharing logits, facts, and runs are distinct but not independent.

Evidence timeline. The frozen pipeline (universe, base, exposure, drift, evaluation instrument, margin-field schema) and the main comparisons were specified before data collection.

Later arms, added after the primary comparisons were frozen: the campaign-1 extension (three runs), the TIES/DARE operator arm, the logit-std normalizer, the own-endpoint baseline and threshold analyses of §4.1, the universe-2 arm, the two-dimensional grid, and the damage-machine validation of the screen.

Reproduction notes. Several released *.jsonl streams are written in job-completion order, not grid order (lmc_barrier, covariates, the two repair streams); index them by their .tau_idx / .tau fields, never by row position. The survival_a05 array is grid-ordered (verified against the symmetry cells' pooled values, max deviation 5.7e−5). Two disclosed logging artifacts participate in no readout: a stale contract_id string in one metrics file, and gzip-header nondeterminism in one universe hash (content verified byte-identical after decompression). Figure scripts read the released fields directly (§O).

Table 3: Measurement conventions per arm: provenance type, survival denominator, alive criterion, and reference floor. Denominators differ by design (conditional on endpoint storage vs unconditional); cross-arm reads are within-row comparisons.

Arm	Cohort / denominator	Alive criterion	Floor
Synthetic testbed (univ. 1/2/3, ladder, all controls)	endpoint-surviving exclusive facts (conditional on storage)	per-model ghost-Q99 margin threshold (§2); template-mean rank ≤ 1.5 is a separate, stricter readout (Table 16)	per-model ghost Q99
Allocation sweep and 1B bridge allocation (§3)	all assigned facts (unconditional)	same ghost-Q99 margin criterion (testbed) / candidate-strip rank-1 per (fact, attribute) (1B)	per-model ghost Q99 / ghost floor
Real fine-tune pairs (§5.4)	G+R ability cohort, per (pair, branch)	rank-1 within the anchored candidate strip	base/other-parent alive rates; logit-averaging null
1B bridge (OLMo-2, §5.4)	all injected exclusive facts (unconditional QA accuracy)	held-out QA templates, answer token	ghost pool 0.13–0.16

C Practice census: summary and full 24-row table

Table 4: Census summary by relationship to the scale requirement (full 24-row table: Table 5). "Sign-read" counts conclusion-carrying sign uses only; the three operator-internal sign uses of Table 5 are not counted.

Residual readout family	Papers	Sign-read	Scale undeclared	Additive
Rank/indicator-based (invariant to positive rescaling)	5	5	4	4
Cardinal (loss, reward, score gaps)	8	8	6	8
of which scale-homogeneous margins (Rem. 1)	0	0	—	—
Weight-space task vectors (outside the readout premise)	5	1	3	5
Magnitude-only readout-space uses (no sign-reading)	6	0	4	5
All	24	14 (58.3%)	17 (70.8%)	22 (91.7%)

This is the full census behind Table 4: every row's classification was read off the cited paper's method section.

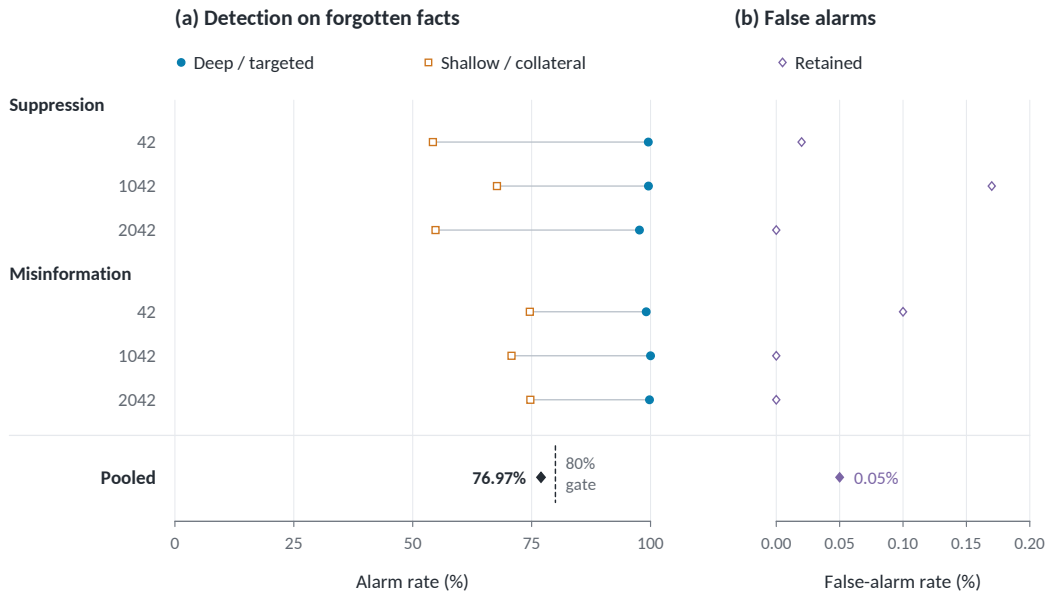


Figure 3: Stratified power check. Alarm rates of the multiplicative readout for suppression and misinformation, shown separately for seeds 42, 1042, and 2042. (a) Certified-forgotten targeted (deep) and collateral (shallow) strata; thin segments pair strata within the same mechanism–seed cell. The pooled forgotten-stratum TPR is 0.7697 on 10,427 facts, below the pre-registered 0.80 gate, which is drawn only beside the pooled row. (b) Certified-retained false-alarm rates on a separate percentage axis; the reported pooled FPR is 0.0005 on 23,047 retained facts. Strata are defined by the independent recall instrument, not merely by intended edit targets. No uncertainty intervals are added.

Sampling protocol. The 24 rows are a purposive sample, not a systematic venue-year census: seeded from the task-arithmetic, loss-barrier, and model-soups literature and extended by citation chasing until each residual family had at least five members; inclusion required an observed-minus-reference quantity carrying an audit or selection signal. Classification was full-text reading against the rubric of Table 4; gray-zone neighbors (membership inference, knowledge editing) are discussed in §6 rather than tabulated, because their per-example readouts are cardinal and sit outside Remark 1's premise.

D Supplementary tables for the four tests

Why the power check stratifies by measured forgetting. We first tried to manufacture a cleanly forgotten cohort and could not. Four gradient-family recipes in three generations — unanchored, anchored, and answer-token-scoped ascent — all failed the pre-registered collateral gates: suppressing 500 targets took the same-family neighbors down with them. A closed-form MEMIT-style rank-1 edit (single seed, one shot) fared no better: targets fell 0.894 $\rightarrow$ 0.046 in rank-1 rate, non-target neighbors 0.927 $\rightarrow$ 0.054 — a suppression transfer rate of 0.985. Gradient-subspace measurement localizes the obstruction: the leading principal angle between target and neighbor gradient subspaces is no larger than against unrelated pools (cosine 0.33 vs 0.31; random reference 3×10^{-4}) — the coupling lives in a shared readout geometry, not on family boundaries. We claim this only for this template-coupled 30M universe at K=128: not that selective forgetting is impossible, but that here forgetting arrives in correlated clusters — which is exactly why a per-fact calibrated alarm is necessary, and why the validation of §6 stratifies by measured forgetting state rather than by intended target.

NLL-barrier temperature placebo. The residual form the field actually uses is cardinal (NLL/loss barriers), outside Remark 1's premise — temperature is not estimand-null for

Table 5: The observed-minus-reference template across 24 merging papers (grouped by residual space), summarized as Table 4. "Scale" = scale convention declared for the residual readout; "sign" = the residual's sign directly carries a harm/benefit conclusion (vs magnitude used as an optimization target); "internal" = sign operations appear only as mechanism-internal steps (election/fixing inside the merge operator), not as a conclusion-carrying residual readout. Full-text census; each row's classification was read off the paper's method section. Tier: the Table 4 row the work lands in — R = rank/indicator-based (invariant to positive rescaling), C = cardinal, W = weight-space, M = magnitude-only. Scale: ✓ = scale convention declared, × = undeclared; shaded bands separate the three residual spaces.

Work	Tier	Residual (short)	Space	Form	Scale	Sign
Per-example / readout-space residuals (13)						
Frankle et al. (2020)	C	error barrier vs endpoint mean	loss	add	×	yes
Mirzadeh et al. (2021)	C	forgetting $L_1(\hat{w}_2) - L_1(\hat{w}_1)$; barrier	loss	add	×	yes
Entezari et al. (2022)	C	loss barrier along the path	loss	add	×	yes
Ainsworth et al. (2023)	C	barrier after permutation alignment	loss	add	×	yes
Jordan et al. (2023)	R	midpoint error − endpoint error	score	add	×	yes
Juneja et al. (2023)	C	barrier height; generalization	loss	add	×	yes
Jin et al. (2023)	M	per-example prediction ℓ^2 distance	activation	add	×	magnitude
Daheim et al. (2024)	M	per-example gradient difference	gradient	add	✓	magnitude
Ortiz-Jimenez et al. (2023)	M	per-example prediction disagreement ξ	prediction	other	×	magnitude
Yang et al. (2024a)	M	per-example feature L1 distance	feature	add	×	magnitude
Cheng et al. (2025)	M	per-example interference $\|\theta_{m,l}x - \theta_{i,l}x\|^2$	activation	add	×	magnitude
Im et al. (2026)	C	task-wise loss gap $G(\tau, r)$	loss	add	✓	yes
Zhu (2026)	M	layerwise cross-term; output JSD ratio	output	add+ratio	✓	magnitude
Weight-space residuals (5)						
Ilharco et al. (2023)	W	task vector $\tau_t = \theta^t_{\text{ft}} - \theta_{\text{pre}}$	weights	add	×	yes
Yadav et al. (2023)	W	task vectors + sign election	weights	add	×	internal
Yu et al. (2024)	W	delta parameters, rescaled $1/(1-p)$	weights	add	✓	magnitude
Huang et al. (2024)	W	task vectors + elected sign	weights	add	×	internal
Kuo et al. (2025)	W	local task vectors, sign-fixed	weights	add	✓	internal
Dataset-level score residuals (6)						
Wortsman et al. (2022a)	R	held-out accuracy delta (greedy soup)	score	add	×	yes
Wortsman et al. (2022b)	R	interpolated vs endpoint accuracies	score	add	×	yes
Ramé et al. (2023)	C	reward Pareto dominance	reward	add	×	yes
Xiao et al. (2024)	C	loss gaps to softmax weights	loss	add	✓	yes
Lu et al. (2024)	R	merged vs individual performance drop	score	add	×	yes
He et al. (2025)	R	normalized performance ratio	score	ratio	✓	yes

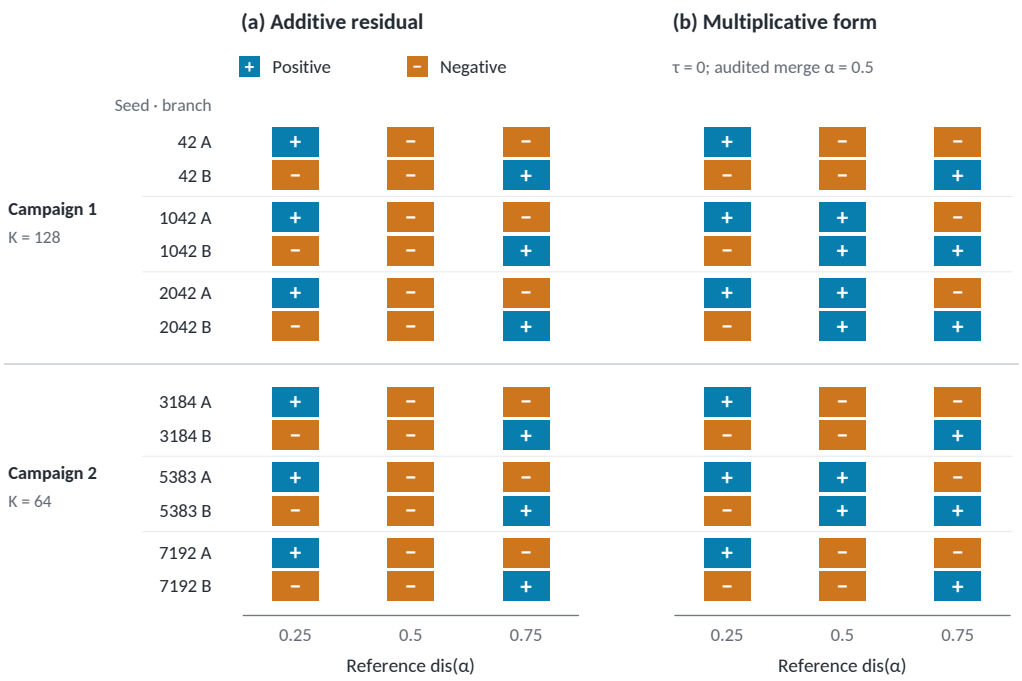


Figure 4: Reference-family scan. At $\tau = 0$, the audited model is held fixed at the $\alpha = 0.5$ merge, while the reference setting dis(α) varies over $\{0.25, 0.5, 0.75\}$. Each row is a seed–branch cell from the original six runs; campaign 1 uses $K = 128$ and campaign 2 uses $K = 64$. Tiles show the sign of the measured-minus-reference contrast for (a) the additive residual and (b) the normalized form. Blue "+" and orange "−" indicate positive and negative residuals, respectively; the symbols preserve readability without color. All 12 rows change sign across reference settings under both forms. Replacing the reference changes the comparison target: normalization removes its scale convention, not reference dependence.

Table 6: The four tests × two residual forms (§6–6). Only the Temperature row has an algebraic right answer, so only it is pass/fail (× = fails); the Reference row counts sign moves under reference replacement (12/12 seed×branch cells — a different sweep from the Temperature row's flip counts of §6); the Quantile row probes the calibration-quantile freedom; ⊙ = does not discriminate (the Null-reference row evaluates thresholded survival, a third instrument class, not a margin residual). ≡ marks the positive control: the normalized form is invariant by construction — its moves on the other rows give those tests their power.

Test	Type	Additive	Normalized
Reference	convention	moves (12/12)	moves (12/12)
Temperature	provable	×	≡ (by construction)
Quantile	convention	moves	moves
Null-reference	specificity	⊙ (thresholded survival; not a margin form)	

it. We measure its sensitivity directly on the three real pairs (300 parent-alive probes each; $D(t; f) = \log p_{\text{merge}}(f) - \log p_{\text{ref}(t)}(f)$, reference = own fine-tune under $z \to t \cdot z$): the median D flips sign within $t \in [0.7, 2.5]$ on 1/3 pairs (tm3 Qwen3-8B, $t^* \approx 0.85$; from $+0.50$ at t=1 to -0.59 at t=0.7), and is stable on tm1 ($-3.9 \to -1.3$, deep damage) and tm2 ($+0.19 \to +3.0$). Cardinal readouts move with the convention too, just less sharply than margins (data-aggregates/nllplacebo_report.json).

Reference / Quantile / Power tests (full versions). Reference test.

The two standard references already disagree (the merge reads no worse in all 24 (seed, τ) cells of the original campaign-1 runs against the linear-dilution null m_{lin}; worse in the same 24 against the scaled own-branch reference dis; Appendix, §K) — the audit's answer is entirely the choice of reference frame. That reference, $\text{dis}(\alpha) = \theta_0 + \alpha \cdot \delta_{\text{own}}$, is an off-manifold model on a ray that was never optimized: its ghost floor swings across the three α levels with per-seed ranges 0.23–1.45 logit while the audited merged model's floor moves only 0.04–0.20: what swings is the reference frame, not the audited model. The consequence is measurable: cross-model value-segment scale ratios span 1.04–2.67 and every scale-homogeneous residual's sign pattern tracks them (Appendix, Table 9); the measured merge-to-reference ratio exceeds the cell's sign-flipping temperature t^* (§6) in 8/17 flippable cells — merely equalizing the two models' logit scales reverses the sign there. Across the reference family, the residual sign moves in 12/12 seed-by-branch cells under the additive form — and under the multiplicative form too (Appendix, Figure 4): replacing the reference redefines the estimand; normalization removes only its scale convention.

Quantile test.

Ghost-floor calibration needs a summary statistic of the ghost pool; practice picks one silently, and the pick matters. The max cross-model gap of the ghost median is 5.6172 logit; the Q99 quantile behind the survival threshold gives 0.9376 — about 6× apart (maxima over all 144 seed×τ×branch cells). Additive calibration with the median turns 128/144 cells positive (95/96 in campaign 1); with Q99, only 4/144 — the effect is smaller than the jitter of one's own calibration anchor. A split-half paired fit on the ghost pool, $m_{\text{merge}} \approx \beta + s \cdot m_{\text{dis}}$, rejects the additive premise directly: slopes s = 1.104–1.549 (6/6 cells of the original campaign-1 seeds) — the diagnosis favors a multiplicative correction over an additive one. Two guards (Appendix D): the ghost gap is itself downstream of the treatment, so no ghost-calibrated sign is the "true sign"; and the normalized form needs a normalizer that cannot cross zero (Q99.5 sits at or above zero for 70% of the evaluated models — a sign-unstable near-zero divisor).

Power.

The closing validation damages k=500 target facts on the τ=0 A-branch endpoint of each original campaign-1 seed (gradient ascent; misinformation overwrite), stratified by an independent recall instrument (Appendix, Figure 3): pooled TPR 0.77 on 10,427 certified-forgotten facts — below the pre-registered 0.80 gate — with targeted (deep) damage caught at 97.7–100%, false alarms $\leq$ 0.17%: a conservative deep-damage alarm, knowingly blind to shallow drift.

Operator robustness (TIES / DARE). The protocol's temperature result is not specific to weight averaging. We constructed TIES (density 0.5) and DARE (p=0.5) merges from the frozen checkpoints of all nine runs at $\tau \in \{0, 50\text{M}, 200\text{M}\}$ and re-ran the same sweep against the same dis(0.5) reference (data-aggregates/tiesdare_report.json). Under TIES the flip reproduces (15/18 cells at τ=0, t^* down to 1.04; the normalized control is flat by construction). Under DARE the additive residual already reads deeply negative at t=1 (-0.27 to -1.63), so most cells flip only beyond the scan range (4/18 within $t \leq 2.5$): the convention-dependence is unchanged in kind — only the displacement needed to see it differs across operators.

Operator discrimination (per-fact). Does the testbed separate merge operators on per-fact survival? On the own-endpoint cohort, with each operator evaluated under its own ghost-Q99 threshold (α=0.5): at τ=0 the three operators are comparable (0.25–0.44, TIES

Table 7: Readout consensus by τ (out of 9 runs, two campaigns; α=0.5). A run counts only when both branches satisfy the merge-worse-than-dis direction on that readout. The six rows are distinct but mutually dependent functions of the same logits, facts, and nine runs. The last row is the normalized form D_{mul} of the normalized form of Remark 1.

Readout	τ=0	3.1M	6.3M	12.5M	25M	50M	100M	200M
m_{top1} (raw residual)	9	9	9	9	9	9	9	9
frac_rank1	5	5	5	6	8	9	9	9
median rank	5	5	5	7	8	8	8	9
logp	5	5	5	5	6	9	9	9
m_{nll}	5	5	5	5	6	9	9	9
D_{mul} (scale-invariant)	5	5	6	7	8	9	9	9
runs with all five non-raw readouts agreeing	5	5	5	5	6	8	8	9

Table 8: Per-readout values at τ=0, α=0.5, A branch. Bold marks the side with better retention in each comparison. On seed 42 the merge is worse than the reference; on seeds 1042/2042 it retains more.

seed	frac_rank1 merge / dis	median rank merge / dis	med Δlogp	med Δm_{nll}	paired share, merge better
42	0.2794 / 0.3765	3.25 / 2.25	−0.3875	−0.5495	0.2633
1042	0.3259 / 0.2146	3.00 / 4.875	+0.7343	+0.7719	0.5824
2042	0.2713 / 0.1662	3.625 / 8.875	+0.7491	+0.7711	0.6418

Table 9: τ=0, A branch. The readout disagreement appears exactly on the seeds whose value-segment logit scales differ most: ratios ≤ 1.13 (seeds 42, 3184, 7192) keep full cross-readout agreement; ratios ≥ 1.46 (5383, 1042, 2042) split. Campaign-2 seeds (3184/5383/7192) ran at half exposure dose.

seed	|ghost median(merge)|	|ghost median(dis)|	scale ratio	ghost-pool paired slope s	D_{dis} raw	other four readouts agree with m_{top1}?
42	7.727	6.883	1.12	1.116	−0.7266	yes
1042	2.641	1.379	1.92	1.549	−0.0547	no
2042	1.984	0.742	2.67	1.437	−0.1328	no
3184	8.387	8.096	1.04	0.999	−0.7734	yes
5383	1.957	1.336	1.46	1.067	−0.0781	no
7192	7.225	6.856	1.05	1.015	−0.5625	yes

Table 10: Specificity diagnostic (Δ_{spec} = median margin difference on surviving exclusive facts − median difference on ghost probes). Surviving facts lose no more margin than never-trained ghost probes (positive in 6/6 seed-by-branch cells; B branch likewise). The ghost pool is not a zero-damage baseline: ghosts have nothing to lose, and their drop reads out the competitor-elevation mechanism itself. On the other branch's exclusive facts the same model pairs are far better than the reference (+5.51 / +1.04 / +0.45), so the structure is fact-specific, not a global shift.

seed·branch	F_{live} (surviving excl.)	ghost probes	endpoint-dead excl.	shared facts	other-branch excl.	Δ_{spec}
42·A	−0.7266	−0.8242	−1.3594	−0.8047	+5.5059	+0.0977
1042·A	−0.0547	−1.1816	−0.9883	−1.1875	+1.0391	+1.1270
2042·A	−0.1328	−1.1094	−0.8203	−1.1016	+0.4453	+0.9766

marginally best); from $\tau \geq$ 50M they separate — averaging retains 0.22–0.32, TIES 0.12–0.21, DARE 0.07–0.16 (campaign medians) — and the per-fact ΔNLL median orders the three operators identically to measured survival (data-aggregates/operator_survival.json; the TIES/DARE margins are the shards of the operator-robustness arm above, re-aggregated per fact).

Quantile test, full itemization. The logging pipeline records two kinds of ghost statistics: quantiles theta_alive_top1_q{980, 990, 995} (the survival-threshold family) and ghost_median_top1_train (logged, used by no mainline readout). Their max cross-model gaps are not of the same order: 5.6172 logit (median) vs 0.9376 (Q99), the maxima over all 144 cells. Cells turned positive under additive calibration: median 128/144 (95/96 in

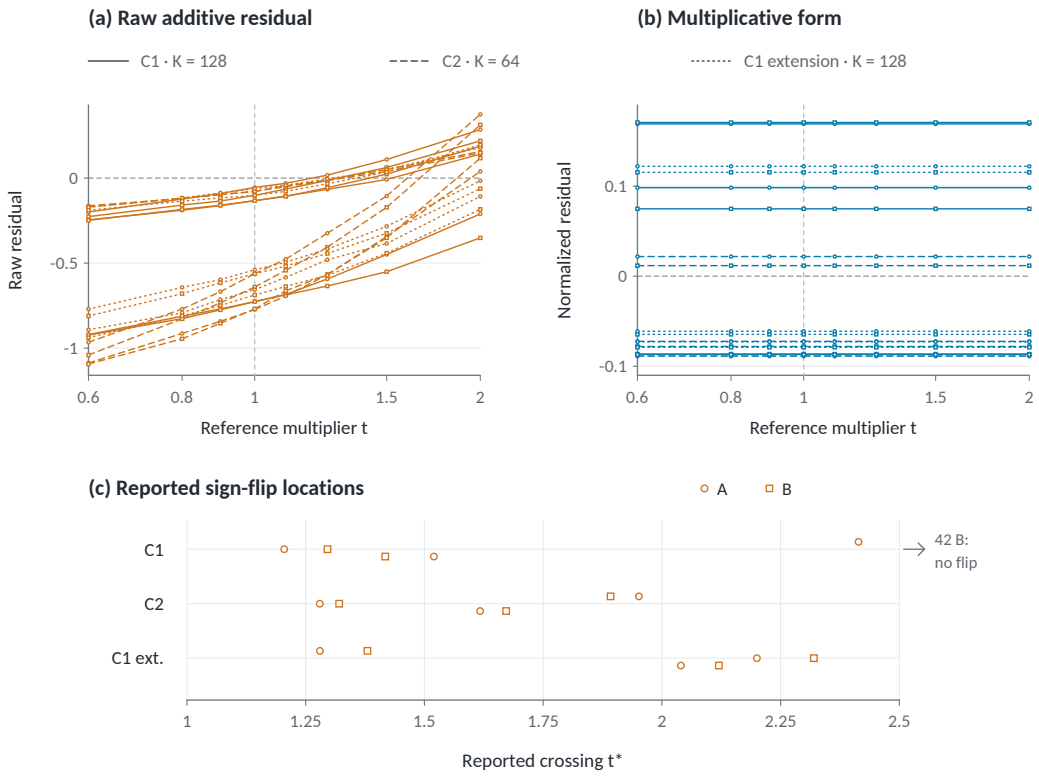


Figure 5: Temperature placebo. Reference logits are mapped as $z \mapsto tz$ at $\tau = 0$ and merge $\alpha = 0.5$, preserving ranks for positive t. (a,b) Raw additive residuals and the normalized form at the eight logged multipliers in $[0.6, 2.0]$, plotted on a logarithmic t axis; the vertical guide marks $t = 1$. Each curve is one of 18 seed–branch measurements from nine runs. Solid, dashed, and dotted lines denote campaign 1, campaign 2, and the campaign-1 extension; circles and squares denote branches A and B. Panels use separate vertical axes. (c) Saved dense-scan crossing locations t^*, shown independently of the logged curves: 17 of 18 cells change sign within the scanned range $[0.5, 2.5]$, with crossings from 1.2048 to 2.4141; seed 42B does not flip within this range. Small vertical offsets separate seeds within a campaign. The logged curves are not extrapolated to these crossings. The multiplicative form is constant at every logged multiplier in all 18 cells.

Table 11: Temperature sweep, endpoint values (τ=0, α=0.5, $t : 0.6 \to 2.0$), campaign 1 (each column in its own presentation scale; signs are the content). Sign flips in bold. Dense-scan t^* per cell: campaign 1 = 2.41 (42A; 42B never flips), 1.20/1.30 (1042), 1.52/1.42 (2042); campaign 2 = 1.89/1.95 (3184), 1.28/1.32 (5383), 1.62/1.67 (7192); campaign-1 extension = 1.28/1.38 (6784), 2.20/2.32 (403), 2.04/2.12 (1189) — 17 flips, median 1.62; normalized spread 0.0 in all eighteen.

seed·branch	raw additive residual	additive ghost-calibrated	normalized form
42·A	$-0.920 \to -0.211$	$+2.677 \to \mathbf{-6.250}$ (sign flip)	-0.0869 (invariant throughout)
42·B	$-0.923 \to -0.352$	$+2.702 \to \mathbf{-6.297}$ (sign flip)	-0.0864 (invariant)
1042·A	$-0.199 \to \mathbf{+0.285}$ (sign flip)	$+1.614 \to +0.168$	$+0.0982$ (invariant)
1042·B	$-0.227 \to \mathbf{+0.219}$ (sign flip)	$+1.589 \to +0.109$	$+0.0748$ (invariant)
2042·A	$-0.245 \to \mathbf{+0.141}$ (sign flip)	$+1.294 \to +0.641$	$+0.1690$ (invariant)
2042·B	$-0.248 \to \mathbf{+0.187}$ (sign flip)	$+1.251 \to +0.555$	$+0.1706$ (invariant)

Table 12: The normalized form D_{mul} at every τ (negative = merge worse than dis; α=0.5), by read-only re-aggregation of the released per-fact margins with the same code path as the sweep of Table 11 (its τ=0 column reproduces that artifact's t=1 values bitwise). Bottom row: runs whose both branches are negative, of the six runs shown (3 campaign-1 + 3 campaign-2 — counted alongside per the never-pooled rule; the campaign-1 extension's three runs are in Table 13, 3/3 from $\tau \geq 25$M). 6/6 from $\tau \geq 50$M; 12/12 cells negative at τ=200M. Campaign-2 seeds ran at half exposure dose.

τ	42A	42B	1042A	1042B	2042A	2042B	3184A	3184B	5383A	5383B	7192A	7192B	seeds
0	−0.0869	−0.0864	+0.0982	+0.0748	+0.1690	+0.1706	−0.0888	−0.0871	+0.0218	+0.0117	−0.0723	−0.0780	3
3.1M	−0.0805	−0.0819	+0.0349	+0.0294	+0.1305	+0.1527	−0.0841	−0.0811	−0.0120	+0.0021	−0.0701	−0.0678	3
6.3M	−0.0878	−0.0864	+0.0265	+0.0005	+0.1009	+0.1290	−0.0666	−0.0687	−0.0100	−0.0341	−0.0737	−0.0758	4
12.5M	−0.0986	−0.0906	−0.0203	−0.0453	+0.0691	+0.0952	−0.0743	−0.0733	−0.0709	−0.0656	−0.0585	−0.0594	5
25M	−0.0870	−0.0929	−0.0652	−0.0918	+0.0382	+0.0398	−0.0681	−0.0654	−0.1177	−0.1118	−0.0668	−0.0623	5
50M	−0.0931	−0.0971	−0.1052	−0.1284	−0.0301	−0.0090	−0.0532	−0.0637	−0.1452	−0.1615	−0.0554	−0.0674	6
100M	−0.0934	−0.0953	−0.1151	−0.1496	−0.0977	−0.0300	−0.0511	−0.0563	−0.2009	−0.1877	−0.0871	−0.0748	6
200M	−0.0945	−0.0780	−0.1850	−0.2022	−0.1493	−0.0993	−0.0358	−0.0551	−0.2007	−0.2130	−0.0540	−0.0754	6

Table 13: Same as Table 12 for the campaign-1 extension (three additional runs at K=128; same universe, protocol, and instrument). Both branches negative from $\tau \geq 25$M in 3/3 — the deep-drift direction replicates.

τ	6784A	6784B	403A	403B	1189A	1189B	runs
0	+0.1220	+0.1153	−0.0729	−0.0789	−0.0611	−0.0646	2
3.1M	+0.0795	+0.0916	−0.0916	−0.0899	−0.0748	−0.0725	2
6.3M	+0.0436	+0.0574	−0.0878	−0.0815	−0.0742	−0.0693	2
12.5M	−0.0100	+0.0102	−0.0878	−0.0929	−0.0777	−0.0921	2
25M	−0.0696	−0.0299	−0.0855	−0.0923	−0.0803	−0.0871	3
50M	−0.1116	−0.0804	−0.0922	−0.0870	−0.0913	−0.0863	3
100M	−0.1838	−0.1600	−0.0744	−0.0938	−0.1086	−0.0997	3
200M	−0.2303	−0.1877	−0.0793	−0.0885	−0.0916	−0.1192	3

campaign 1), Q99 4/144, Q99.5 2/144, Q98 19/144; the three quantiles give 2:7 / 5:4 / 7:2 negative:positive run patterns at τ=0. Ghost-pool paired slopes s = 1.116 / 1.104 / 1.549 / 1.467 / 1.437 / 1.399 (6/6 campaign-1 cells away from 1; split-half difference ≤ 0.07), intercepts −0.77 to −0.02; campaign-2 slopes at τ=0: 0.999 / 1.067 / 1.015. At depth (τ=200M) the twelve (seed, branch) cells of the original six runs give slopes 1.09–1.99 (paper_backfill.json, G_findings.qq_slope_tau7). The ghost gap is a post-treatment variable by construction: $\text{merge}(a) - \text{dis}_{\text{own}}(a) = (\text{other-branch dose}) \cdot \delta_{\text{other}}$, and the dose-response check (ghost gap vs other-branch dose, seed 42·A, τ=0: 0.4023 / 0.8438 / 1.2246 over three dose levels; monotone in 48/48 campaign-1 sets) confirms it empirically. Positive controls (three constructed comparisons whose signs are designed by construction: on own facts, the larger own-δ dose is intended to be better — an empirical regularity these controls verify, not a theorem) are passed by both raw and calibrated forms at 48/48, with the scale-free frac_rank1 confirming 12/12 — but in all three controls the deeper-ghost model also happens to be the truly better one, so passing is forced and endorses nothing about the merge-vs-dis contrast, where bias and truth point opposite ways.

The normalized form under the same swap. Recomputing D_{mul} with each quantile-family normalizer over all 144 (seed, τ, branch) cells: the sign is unchanged across {median, Q98, Q99} in 120/144 cells; across all four members only 68/144 cells keep their sign (two cells singular), because the ghost-pool Q99.5 sits at or above zero for 70% of the evaluated models — a would-be normalizer crossing zero is a sign-unstable near-zero divisor, not a scale. In the deep-drift regime the endpoint-comparison direction survives at run level: 27/27 run×τ cells both-branch negative under the median, 26/27 under Q98, 21/27 under Q99, 10/27 under Q99.5. A non-ghost normalizer family that cannot cross zero — per-model full-vocabulary logit standard deviation on an ex-ante fixed probe battery — keeps the deep-drift direction in 9/9 (seed, τ) cells of each campaign (9/9 + 9/9; the campaign-2 collection ran on the frozen campaign-2 checkpoints in the follow-up collection arm), and the placebo invariance is then mechanical: the normalizer scales with the model.

Table 14: Survival readouts at τ=0, α=0.5, with three references. S_{dis} contains no other-branch weights. S_{lin} is a constructed foil (no real model). Seeds 3184/5383/7192: campaign 2, half exposure dose.

seed·branch	S_{meas}	S_{dis} $(\theta_0 + 0.5 \cdot \delta_{\text{own}})$	S_{lin} (thr = merge)	S_{lin} (thr = own endpoint)
42·A	0.4189	0.5502	0.2521	0.2521
42·B	0.4159	0.5393	0.2518	0.2255
1042·A	0.3971	0.3207	0.3525	0.3563
1042·B	0.3890	0.3637	0.3473	0.4219
2042·A	0.3806	0.2154	0.3607	0.3806
2042·B	0.3744	0.2134	0.3860	0.4146
3184·A	0.4195	0.5128	0.2540	0.2170
3184·B	0.4174	0.4497	0.2548	0.2284
5383·A	0.3969	0.3732	0.2826	0.2826
5383·B	0.3958	0.3792	0.2825	0.3212
7192·A	0.3035	0.4132	0.1859	0.1776
7192·B	0.2943	0.4118	0.1740	0.1609

Table 15: Task-arithmetic arm, τ=0 slice: ghost floor $g(\lambda)$ (median ghost-pool m_{top1}, logit) of the task-vector merge $m_\lambda = \theta_0 + \lambda(\delta_A + \delta_B)$, beside the frozen dis-side floors (own-branch dis(α), A branch). The floor moves with λ — downward.

seed	$g(0.5)$	$g(0.75)$	$g(1.0)$	$g(1.25)$	$g(1.5)$	dis side, range over α
42	−7.73	−8.35	−8.91	−9.51	−9.42	0.80
1042	−2.63	−3.45	−5.83	−7.72	−8.72	1.45
2042	−1.98	−2.95	−4.06	−6.38	−8.03	0.76

E Task-arithmetic reference arm

The strengthening arm of §6 replaces the merge operator by the task-vector form $m_\lambda = \theta_0 + \lambda \cdot (\delta_A + \delta_B)$, $\lambda \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$, on the first campaign's full grid (3 seeds $\times$ 8 τ tiers = 24 cells), and asks whether the merged-side ghost floor rises monotonically with λ (the pre-registered criterion: $\geq 8/9$ of the main cells monotone nondecreasing, and median $g(1.5) - g(0.5) \geq 0.3$ logit). The λ=0.5 point is algebraically identical to the frozen merge(0.5) and passed the parity gate (median per-fact $|\Delta m_{\text{top1}}| = 0$, max 0.039 logit).

The criterion fails. 0/24 cells are monotone nondecreasing; the median $g(1.5) - g(0.5)$ is −6.80 logit (range −21.60 to +1.10): the merged-side ghost floor moves strongly with λ (per-cell swing over the λ grid: median 6.80, range 1.77–21.60 logit) — but downward, not upward. Table 15 shows the τ=0 slice beside the frozen dis-side floors. We therefore report the null: no merged-side counterpart of the reference-side swing is observed under this operator family, and the Reference test stands on the dis(α) evidence (§6).

F Real-model bridge

The temperature placebo of §6 applies to any model, and this appendix reports it on real open models. Three measurement differences from the synthetic testbed apply throughout: real models have no ghost pool (never-trained facts do not exist for them), so the natural-fact placebo panel reports the raw additive residual; the portable normalizer introduced below is used where applicable, and the by-construction 1B bridge has its own ghost pool; and F_{live} contains the measured model's own rank-1 items, a different estimand from the synthetic universe's ghost-Q99 cohort.

The portable normalizer on real models (post-hoc). The released shards carry per-model full-vocabulary logit-std on each battery, so the recommended normalizer of §6 is computable here (tier-1 shards carry the candidate-set logit-std, a fixed sub-vocabulary variant with the same degree-1 property). Measured per-model ratios $s_{\text{ft}}/s_{\text{ref}}$ span 0.86–1.05 across the 13

powered panel cells (tier-1 OLMo pair: 1.05), all below the observed flip temperatures $t^* \in [1.09, 1.86]$: on these pairs, equalizing scales would not have changed any sign (additive and normalized forms agree in sign in 13/13 cells), and the normalized form is placebo-invariant by construction. Nothing in the pipeline enforces this — the ratios are measured, not guaranteed (data-aggregates/realbridge_logitstd.json).

Tier 1: a documented parent–child pair. OLMo-2-0425-1B and its official SFT child, probed on a single-token PopQA battery (1,719 probes, 401 candidate value tokens, 16 properties). Along the task-vector family $m(\lambda) = \theta_0 + \lambda \cdot \delta_{\mathrm{SFT}}$, the median margin on the SFT model's rank-1 probes (F_{live}, $n = 624$) reads 1.53 (base), 1.94 (SFT), and 1.75 / 1.94 / 2.08 / 1.81 / 1.08 at $\lambda = 0.25/0.5/0.75/1.25/1.5$ — the model's scale moves with the coefficient, as in the synthetic universe. The additive audit residual of the SFT model against the λ=0.5 task-vector reference starts at $D = +0.25$; applying the temperature placebo to the reference walks it across zero at $t^* = 1.11$ (first crossing, linear interpolation on the logged 8-point grid) — inside the $[0.7, 1.3]$ band, beside the synthetic crossings (1.20–2.41). On a real model the same flip obtains, and no ground truth exists to say which sign is right.

Tier 2: a claimed sibling pair. A community pair advertised as two fine-tunes of OLMo-2-0425-1B (code / natural-language) was skipped by the pre-registered mechanical gate: config keys and tokenizer.json are byte-identical, but the weight-delta norms fail the band check (one delta non-finite in bf16 loading; the other 4957.8 against the SFT reference's 214.1). The skip is recorded with the gate booleans in data-aggregates/margins_tier2.skipped.json.

True merges of real fine-tune pairs. The panel above audits a fine-tune against a task-vector reference; a separate record-only arm measures five same-base pairs (tm1–tm5, in the order listed) — Qwen2.5-1.5B + {Coder, Math}, Llama-3-8B + {Instruct, MAmmoTH2}, Qwen3-8B-Base + {Qwen3-8B, R1-0528-Qwen3-8B}, Mistral-7B-v0.1 + {Instruct-v0.1, Zephyr-7B-β}, Qwen2.5-7B + {Instruct, Math} — each evaluated at the true 0.5 merge and both dis(0.5) references under the base config and tokenizer, on the same three batteries (Mistral-7B contributes PopQA and MBPP only: its tokenizer splits GSM8K's numeric answers into multiple tokens, so no single-token candidate strip exists). Measured per-model full-vocabulary logit-std ratios (merge over dis reference) span 0.56–1.11, and the additive residual flips within $t \in [0.6, 2.0]$ in 17/22 powered cells, with $t^* \in [0.66, 1.71]$ — two cells sit at $t^* = 1.0$ exactly: the residual is tied at the measured convention. The logit-std-normalized form is exactly invariant (spread 0.0) and disagrees with the additive sign in 1/22 cells at t=1. True-merge scales roam measurably wider than the parent–child panel above (0.86–1.05) — the exposure grows with the mismatch. (data-aggregates/truemerge_report.json.)

Per-fact survival on the true merges. The same shards answer the per-fact question. Stratify probes by (base, own parent, other parent) alive status: G (acquired: base dead, own alive, other dead), R (retained-then-forgotten: base alive, own alive, other dead), S (both alive), N (both dead). The exclusive cohort is G+R. A regime gradient appears (data-aggregates/truemerge_survival_v2.json, data-aggregates/truemerge_survival_newpairs.json). On the three light-tune pairs (Llama-3-8B, Qwen3-8B, Mistral-7B), facts alive at both parents almost always survive (S: 0.97–0.98) while parent-exclusive successes survive at 0.50–0.72: selective death. The mid-spectrum pair (Qwen2.5-7B Instruct+Math) sits between: shared 0.84, exclusive 0.28/0.53 per branch. Battery-stratified probe-level bootstrap (B=10,000, resampling probes within each battery; data-aggregates/truemerge_bootstrap.json) sizes these cells: per-branch exclusive-cohort 95% CI half-widths are 0.06–0.13 (branch $n = 38$–353), and pair-pooled exclusive survival is 0.21 $[0.17, 0.25]$ (tm1), 0.62 $[0.58, 0.66]$ (tm2), 0.61 $[0.55, 0.66]$ (tm3), 0.69 $[0.63, 0.74]$ (tm4), 0.37 $[0.32, 0.42]$ (tm5); pair-pooled shared-cohort survival is 0.61/0.97/0.98/0.98/0.84 with half-widths ≤ 0.05 — the selective-death gap exceeds the sampling noise of every cell. On the heavy-CPT pair (Qwen2.5 Coder+Math), even both-parent successes die at 0.39 (S=0.61) and the parent-exclusive cohort at 0.16 / 0.35 (own=Coder / own=Math) — global collapse, and the margin-linear null m_{lin} does not describe it ($R^2 = -0.27$ own=Coder, $+0.11$ own=Math; data-aggregates/truemerge_survival.json). The scaled own-branch reference ("dis05", half the own task vector, no foreign delta) separates the dilution and foreign-branch

terms on these pairs (G+R survival, merge vs dis05 per branch: tm1 0.16/0.33 (own=Coder) and 0.35/0.26 (own=Math); tm2 0.51/0.71 and 0.67/0.60; tm3 0.60/0.73 and 0.62/0.80; tm4 0.72/0.82 and 0.50/0.71; tm5 0.28/0.92 and 0.53/0.55; pooled over 1,842 probes: 0.49 vs 0.64). Dilution still carries the larger pooled share of the exclusive loss (the cohort definition fixes the parent at 1.00; halving alone lands at 0.64); the foreign-branch term is negative in 8/10 branches (-0.64 to $+0.08$ per branch, -0.15 pooled), and its magnitude grows with the regime — sign-mixed on the light-tune pairs, largest on the mid-spectrum and heavy-CPT pairs (tm5's Instruct side: 0.28 merge vs 0.92 dis05). This is the testbed's decomposition (§4, Appendix J) on real weights: dilution first at zero drift, interference compounding as the updates grow heavier. (§6 caveat concerned the fixed rule "survival < 0.5"; the per-branch dis05 comparison is the supported read.) The frozen screen (threshold 2.0) reads the deep case at recall 0.89 / precision 0.90 (FPR 0.50 at that cell's 84% death prevalence — near-total death inflates it); on the shallow-death branches the ΔNLL score's AUC drops to 0.53–0.73 and recall to 0.07–0.22 — the screen is effective where damage is deep. Matching on the own-parent margin (terciles), the gap persists: at every tercile, both-parent successes survive at 0.55–1.00 while parent-exclusive successes survive at 0.15–0.83 — the gap is not a margin-strength artifact (data-aggregates/truemerge_marginmatch.json). Caveats: provenance is by ability (a fact's home is the parent that answers it), not by construction; GSM8K/MBPP rows are per-item capability survival — for those batteries a probe is scored alive when the correct final answer (the numeric value / the code solution's answer slot) ranks first within the candidate set of the battery's anchored candidates (tok-class anchored candidate strips, same construction as the PopQA candidates), not a full free-form generation check; and the Coder/Math pair are heavy continued-pretraining (CPT) runs, not light fine-tunes. Per-parent monitoring registers the loss (data-aggregates/truemerge_perparent.json): on each parent's own-answerable subset the merge retains 0.85–0.93 on the light-tune pairs (0.35/0.52 on the heavy-CPT pair) — a visible 7–15% drop — while the aggregate battery still exceeds both parents on tm2. The silence is specific to monitoring that omits parent-exclusive items, not to per-parent evaluation.

Free-generation readout. The numbers above read survival off anchored candidate strips (rank-1 within a battery-specific candidate set). To test whether the selective pattern survives a candidate-free readout, we re-ran all five pairs' probes as open generation (tm4 contributes PopQA and MBPP only — no single-token GSM8K strip exists under its tokenizer): greedy decoding, no candidate set; PopQA as direct short answers ($\leq$ 24 new tokens), GSM8K as 4-shot chain-of-thought ($\leq$ 512), MBPP as direct return-value answers ($\leq$ 32) (data-aggregates/truegen_survival.json; per-probe generations released in data-aggregates/truegen-shards/). Two cohort definitions agree. Under the strip-defined cohorts, merge free-generation survival on the exclusive cohort reads 0.14/0.23 (tm1), 0.50/0.72 (tm2), 0.58/0.39 (tm3) per branch, against S at 0.41/0.84/0.67 — the pattern survives, compressed by a uniformly harder readout. Under cohorts defined by what each parent itself produces freely (endpoint free-generation correct, other parent incorrect), the contrast is clean: exclusive 0.68/0.52 vs shared 0.94 (tm2), 0.77/0.44 vs 0.95 (tm3), 0.77/0.61 vs 0.96 (tm4, PopQA+MBPP only), 0.38/0.60 vs 0.87 (tm5, mid-spectrum), 0.09/0.18 vs 0.31 (tm1). The aggregate silence also replicates in free-generation form: the merge's open-ended battery accuracy is between or above both parents on the four non-collapsing pairs (tm2 0.53 vs 0.58/0.32; tm3 0.37 vs 0.28/0.37; tm4 0.50 vs 0.44/0.40 — above both; tm5 0.37 vs 0.47/0.31), while on the heavy-CPT pair the aggregate itself collapses visibly (tm1 0.10 vs 0.34/0.28). Caveats: PopQA free-generation scoring is noisy (rambling continuations truncated at the token cap; a single alias credited) and may depress cells, and not necessarily uniformly across models (rambling styles differ); GSM8K's 4-shot template is fixed across models. The free-generation arm measures retrieval under production-like prompting; it does not replace the margin readouts, which remain the paper's quantitative basis.

Tokenizer note (tm3): the R1-0528 tokenizer's added-token tail differs (chat/think tokens), but the 151,643-entry base-vocabulary mapping is identical, all three configs share vocab_size 151,936, the batteries' answer tokens all live in the base vocabulary, and every evaluation runs under the base config and tokenizer.

Divergence spectrum. Task-vector cosine similarities in float64 (data-aggregates/truemerge_cosines.json): tm2 Llama-3-8B Instruct vs MAmmoTH2 0.066; tm3 Qwen3-8B vs R1-0528 0.19; tm1 Qwen2.5 Coder vs Math 0.41; tm4 0.028 and tm5 0.021 complete the five-pair span (0.02–0.41; data-aggregates/truemerge_deltanorms.json) (larger = more aligned). The angle does not order the damage: the most aligned pair (tm1) is the one that collapses globally. The norm axis does (data-aggregates/truemerge_deltanorms.json): relative update norms $\|\delta\|_F/\|\theta_0\|_F$ are 0.63/1.08 (tm1 A/B) vs 0.04–0.11 (tm2/tm3) — a 6–25× separation matching the collapse boundary. Family, task, and training intensity co-vary across the three pairs (tm1 is heavy continued pretraining; tm2/tm3 are light fine-tunes), so this is a regime observation, not an isolated axis. The synthetic testbed's inter-branch cosine is 0.43–0.73 (Appendix K) — more aligned than all five real pairs analyzed here, so the testbed's deeper death is not an artifact of near-orthogonal updates.

(Bridge history: three earlier protocol variants of a 1B injection bridge never passed the storage gate (frac-rank-1 ≤ 0.19) — thin single-token templates with low diversity. The working recipe that eventually stored facts on OLMo-2-1B (endpoint held-out QA accuracy 0.88–0.98): ∼50 structurally distinct renderings per fact (cross-attribute sentence families, declaratives in both answer positions, QA variants), replay-heavy mixing (2:1), 5 epochs, lr 10^{-4}–$3 \cdot 10^{-4}$, mixed order — a hard curriculum (biographies then QA) hurt (Appendix evidence timeline). The main arm's merge cohorts: exclusive A 0.25 (from 0.85), exclusive B 0.24 (from 0.91), shared 0.83, ghost 0.13 (data-aggregates/e6main-qa-dumps/). Replicated across five units (two universes × two–three training seeds, ITT): exclusive merge QA accuracy 0.19–0.26, shared 0.75–0.85, ghost 0.13–0.16; all ten branch endpoints pass the storage gate (0.75–0.94) (data-aggregates/e6main-qa-dumps/).) All five bridge units use the legacy fact generator whose attribute-value pools fold at assignment (the degeneracy disclosed and repaired in Appendix I); a two-unit re-run of the bridge itself on the repaired attribute-diverse generator (same recipe, ITT) reproduces the signature with cleaner separation: exclusive 0.30–0.35, shared 0.78–0.79, ghost 0.09, all four endpoints passing the storage gate (0.90–0.94) (data-aggregates/e6main-qa-dumps/).

Joint decomposition of the bridge's death rate. The "kills X%" figure above is a ratio of marginal accuracies; the true death rate of endpoint-alive facts is $1-P(A\cap M)/P(A)$, which needs the joint. Per-fact QA dumps for all seven units (data-aggregates/e6main-qa-dumps/; each 20,800 (fact, attribute) probes over the exclusive, shared, and ghost pools) give: legacy units — exclusive death 72–77% by the joint account (the marginal ratio reads 71–76% on the same units; the assumption-free joint is slightly harsher); repaired-generator units — 64–65%. Revival (merge alive, own endpoint dead) exists but is small ($P(M \cap \neg A) \approx 0.02$).

Storage-margin scan. Is the bridge's exclusive death an artifact of near-threshold storage? We swept a 3×2 grid around the working recipe (lr $\in \{5 \cdot 10^{-5}, 10^{-4}, 3 \cdot 10^{-4}\}$ × epochs $\in \{5, 10\}$; fixed universe, one seed per cell) on the 1B bridge. Once endpoints pass the storage gate (own-fact QA ≥ 0.75), exclusive-cohort merge survival is flat at 0.21–0.27 and shared at 0.75–0.87 — invariant to storage strength. Below the gate (lr $5 \cdot 10^{-5}$ × 5 epochs; endpoints 0.61/0.67) all rates degrade but the gap persists (0.18 vs 0.67). (data-aggregates/e6scan-reports/.)

A five-family panel under real merge protocols. (The OLMo-2 pair above is the tier-1 cell of this panel; counted once in the totals.) The same placebo was run on five documented base→fine-tune families — Llama-3-8B, OLMo-2-1B, Llama-3.2-1B, Llama-3.2-3B, Qwen3-8B — on three probe batteries (PopQA, GSM8K, MBPP), comparing each fine-tune against the λ=0.5 task-vector reference (all evaluations under the base model's config and tokenizer). Two further panel members (Qwen2.5-1.5B, Qwen2.5-0.5B) were excluded by the preflight config gate (all candidate deltas dropped; recorded as family_skipped), and a 27B scale point (Qwen3.8-27B) was low-power under the pre-registered F_{live} floors and is carried as descriptive scale data only. Of the 15 main (family × battery) cells, 13 meet the power floor; the additive residual flips sign within $t \in [0.6, 2.0]$ in 10/13 powered cells, with all observed $t^* \in [1.09, 1.86]$ and 6/13 inside the $[0.7, 1.3]$ band — the pre-registered case-study criterion (flip fraction $\geq 1/2$ and in-band fraction $\geq 1/4$) is met. These are per-cell demonstrations at real scale, not a cross-family universality claim.

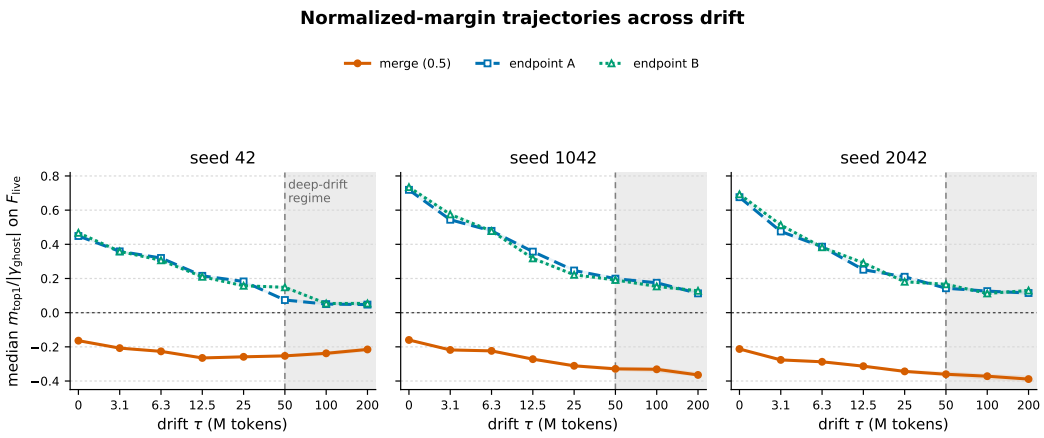


Figure 6: Scale-invariant margin trajectories. The original three campaign-1 runs ($K = 128$), evaluated at the equal-weight merge, $\alpha = 0.5$. Each point is the cohort median of the per-model normalized margin $m_{\text{top1}}/|\gamma_{\text{ghost}}|$; γ_{ghost} is the model's median ghost-fact margin. Endpoints A and B are evaluated on their respective own-endpoint-live exclusive-fact cohorts; the merge median uses the union of these cohorts, not an average of branch medians. Cohorts are reselected at each checkpoint, so the curves do not follow a fixed fact set through time. Ribbons are fact-level bootstrap 90% confidence intervals, not uncertainty across training runs. All eight checkpoints are displayed at equal visual spacing, with rounded token counts; the shaded region starts at the 50M checkpoint. Cohort selection favors the endpoints by construction. The merge's median remains below both endpoints and below zero at every shown checkpoint.

G Cohort qualifications for the endpoint comparison

Two cohort qualifications are part of the endpoint comparison of §4.2. (i) F_{live} is redefined per τ: A-branch n_{live} falls from 5827 / 5812 / 5728 / 5755 / 5729 / 5631 at τ=0 to 616 / 1356 / 1033 / 248 / 680 / 306 at τ=200M (per-seed denominators: Table 17), so the τ axis is not a longitudinal comparison of a fixed cohort. A fixed-cohort sensitivity read — the τ=0 own-endpoint survivors tracked through every τ tier — gives the same direction: both-branch-negative in 3/3 runs of each campaign from $\tau \geq 50$M (data-aggregates/fixed_cohort.json), and the three extension runs replicate it (3/3 from $\tau \geq 50$M; seed 6784 turns at $\tau \approx 12.5$–25M). (ii) Turning points differ across seeds: 0 / 50M / 200M / 0 / 25M / 0 under the all-five-readout criterion (seeds 42/1042/2042/3184/5383/7192), and 0 / 12.5M / 100M / 0 / 6.3M / 0 under the survival criterion — we assert a trend direction, not a cross-seed location.

H Survival surface: per-seed tables

The campaign-1 extension (seeds 6784/403/1189, K=128) replicates the survival shape: S_{meas} at τ=0, α=0.5 is 0.4720 / 0.4578 / 0.3236 / 0.3165 / 0.3327 / 0.3244, and every α=0.5 cell of every extension run sits below 0.5 (data-aggregates/c1x3_backfill.json).

Values are pooled A/B by cohort size in Figure 7; the per-branch values (A/B, with n_{live} denominators) are tabulated below. Ranges quoted in §4.3 read off these tables. The deep regime is not flat: seed 42 rebounds 0.1986 → 0.2880 over $\tau_4 \to \tau_7$, and campaign-2 cells show the same non-monotone texture (e.g. 3184 rebounds 0.2327 → 0.3992 on the A branch).

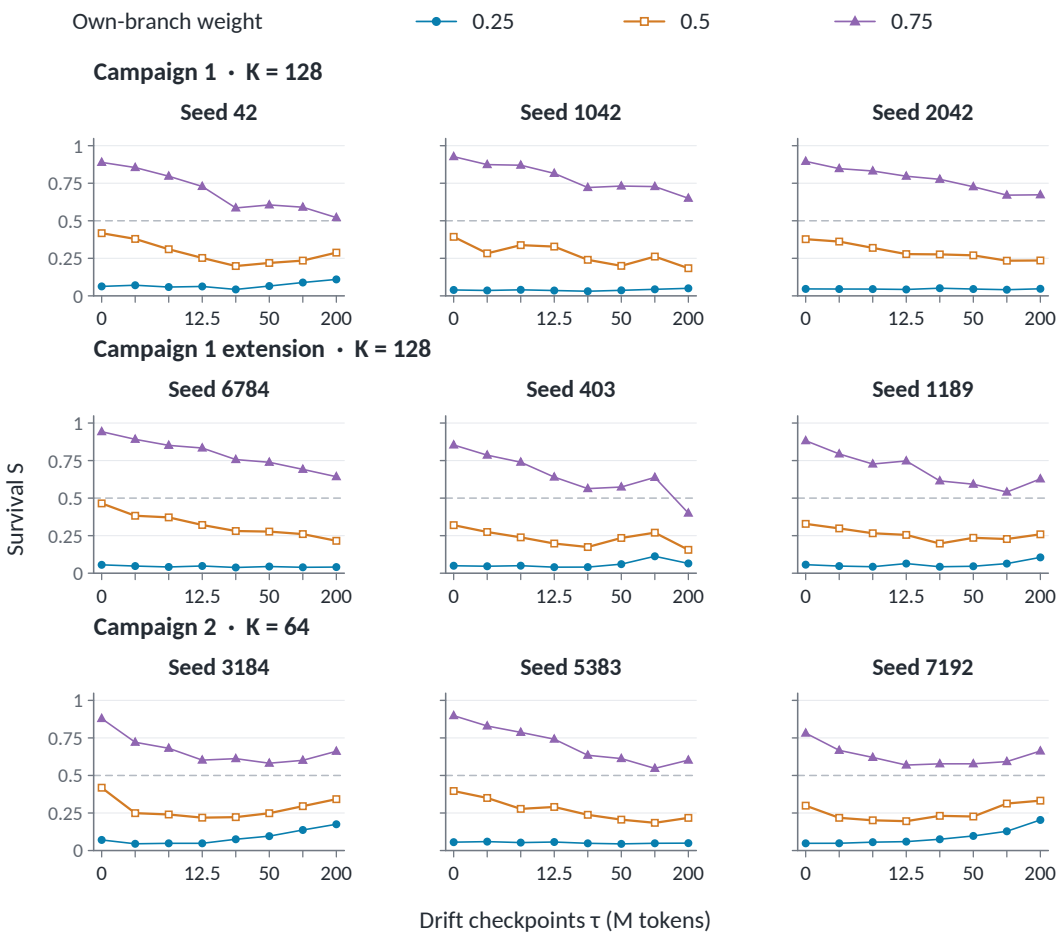


Figure 7: Survival across merge weights and filler drift in all nine runs. Rows show the original campaign-1 runs ($K = 128$), the campaign-1 extension ($K = 128$), and campaign 2 ($K = 64$). Each curve uses a fixed own-branch weight w_{own}: A-branch facts are read at $\alpha = w_{\text{own}}$ and B-branch facts at $\alpha = 1 - w_{\text{own}}$. A/B survival is pooled by own-endpoint-live cohort size within each run, checkpoint, and own weight; no averaging is performed across runs or campaigns. Live cohorts are reselected at each checkpoint. All eight checkpoints are plotted at equal visual spacing, with selected tick labels shown in rounded millions of filler tokens. The dashed line at 0.5 is a visual reference, not an interference or significance threshold.

Table 16: Survival under six readout criteria (campaign 1, τ=0, α=0.5 merge; ranges over the six seed×branch units; cohort = endpoint-alive facts under the paper's ghost-Q99 convention). The magnitude moves with the criterion; the direction — a large loss, far below the own endpoint — does not (at the most lenient criterion, ghost-Q98, survival 0.46–0.56 straddles one half; under the other five criteria most of the cohort dies). The held-out row also quantifies the storage-layer nature of the estimand: even at the endpoint, only 0.03–0.15 of the cohort answers under unseen templates (data-aggregates/survival_robustness.json, data-aggregates/survival_robustness_endpoints.json).

Readout criterion	Merge survival	Own endpoint
ghost-Q98 margin threshold	0.46–0.56	—
ghost-Q99 (paper's survival)	0.37–0.42	1.00 (cohort def.)
ghost-Q99.5	0.24–0.34	—
strict rank-1 (every template)	0.15–0.19	0.82–0.90
template-mean rank ≤ 1.5	0.27–0.33	—
held-out templates, own ghost-Q99	0.02–0.06	0.03–0.15

Table 17: Survival $S(\tau, w_{\text{own}})$ per seed for all six runs: campaign 1 (K=128) and campaign 2 (half exposure dose, K=64). A/B branches; n_{live} = own-endpoint surviving cohort size. Cell shade scales with the cell value (darker = higher survival).

τ	w=0.25	w=0.5	w=0.75	n_{live} A/B
campaign 1 (K=128)				
seed 42				
0	0.0645/0.0613	0.4189/0.4159	0.8898/0.8880	5827/5841
3.1M	0.0719/0.0696	0.3782/0.3802	0.8548/0.8540	5732/5721
6.3M	0.0597/0.0577	0.3172/0.3039	0.8053/0.7886	5157/5387
12.5M	0.0643/0.0610	0.2576/0.2476	0.7393/0.7182	4542/4773
25M	0.0514/0.0352	0.2083/0.1900	0.5496/0.6172	3073/3469
50M	0.0647/0.0664	0.2029/0.2411	0.5525/0.6773	2380/1763
100M	0.0925/0.0853	0.2358/0.2338	0.6183/0.5628	1200/1219
200M	0.1153/0.1027	0.2662/0.3140	0.4578/0.5969	616/516
seed 1042				
0	0.0422/0.0365	0.3971/0.3890	0.9289/0.9252	5812/5892
3.1M	0.0457/0.0263	0.2887/0.2774	0.8564/0.8914	5556/5634
6.3M	0.0407/0.0398	0.3351/0.3406	0.8657/0.8745	5166/4973
12.5M	0.0501/0.0227	0.3346/0.3224	0.7745/0.8519	3933/4451
25M	0.0266/0.0349	0.2382/0.2420	0.7570/0.6858	3271/3265
50M	0.0480/0.0265	0.1994/0.2002	0.7037/0.7568	2312/2492
100M	0.0430/0.0442	0.2596/0.2636	0.7398/0.7162	1768/1878
200M	0.0553/0.0448	0.1711/0.1977	0.6261/0.6749	1356/1295
seed 2042				
0	0.0485/0.0440	0.3806/0.3744	0.8904/0.9009	5728/5801
3.1M	0.0490/0.0419	0.3598/0.3633	0.8273/0.8676	5472/5415
6.3M	0.0401/0.0502	0.3173/0.3218	0.8519/0.8114	4983/5016
12.5M	0.0392/0.0462	0.2738/0.2825	0.8003/0.7930	4262/4092
25M	0.0661/0.0386	0.2865/0.2678	0.7616/0.7879	2391/2928
50M	0.0544/0.0367	0.2710/0.2687	0.7163/0.7367	1801/1827
100M	0.0445/0.0369	0.2284/0.2408	0.6616/0.6810	1528/1329
200M	0.0484/0.0454	0.2352/0.2361	0.6747/0.6708	1033/881
campaign 2 (K=64, half dose)				
seed 3184				
0	0.0805/0.0609	0.4195/0.4174	0.8629/0.8950	5755/5831
3.1M	0.0515/0.0385	0.2569/0.2412	0.6970/0.7448	5014/5066
6.3M	0.0429/0.0535	0.2410/0.2392	0.7113/0.6499	4381/4264
12.5M	0.0425/0.0541	0.2135/0.2249	0.5985/0.6061	3106/2904
25M	0.0924/0.0592	0.2327/0.2129	0.5666/0.6539	1938/2095
50M	0.1095/0.0832	0.2777/0.2213	0.6196/0.5458	886/949
100M	0.1441/0.1294	0.3031/0.2877	0.6062/0.5951	673/657
200M	0.2298/0.1403	0.3992/0.3061	0.6855/0.6454	248/392
seed 5383				
0	0.0541/0.0574	0.3969/0.3958	0.9029/0.8938	5729/5781
3.1M	0.0496/0.0697	0.3435/0.3571	0.8473/0.8107	5488/5251
6.3M	0.0573/0.0489	0.2857/0.2696	0.7848/0.7891	4750/4889
12.5M	0.0544/0.0591	0.2914/0.2891	0.7470/0.7366	3459/3504
25M	0.0443/0.0528	0.2319/0.2452	0.6288/0.6415	1988/1668
50M	0.0585/0.0304	0.2220/0.1897	0.5632/0.6612	1266/1281
100M	0.0530/0.0439	0.1831/0.1859	0.5301/0.5635	830/866
200M	0.0441/0.0548	0.1985/0.2379	0.5691/0.6369	680/639
seed 7192				
0	0.0440/0.0521	0.3035/0.2943	0.7970/0.7668	5631/5643
3.1M	0.0495/0.0474	0.2196/0.2163	0.6669/0.6675	4969/4998
6.3M	0.0520/0.0597	0.1977/0.2051	0.6085/0.6323	4209/4071
12.5M	0.0455/0.0753	0.1858/0.2068	0.6082/0.5242	3009/2602
25M	0.0889/0.0613	0.2418/0.2203	0.5725/0.5824	2047/2038
50M	0.1094/0.0879	0.2516/0.2075	0.6138/0.5498	914/1195
100M	0.1099/0.1483	0.3168/0.3086	0.6099/0.5731	546/499
200M	0.1830/0.2253	0.3268/0.3379	0.6667/0.6587	306/293

I Out-of-universe, allocation, and scale arms

This appendix collects the arms that leave the universe-1 drift grid: the universe-2 transfer and redundancy controls, the allocation sweep and its null analyses, the 1B-bridge allocation, the operator and joint-training contrasts, the decomposition probes, and the scale ladder.

All numbers of the main drift grid come from one generator seed (one universe); the replication and allocation arms below add fresh generators. A second universe (fresh facts, entities, values, templates; generator seed changed, protocol identical except two declared changes: shared facts enter the exposure block of both branches so they are actually stored, and the drift stream carries a maintenance dose of them) answers two questions: does the frozen threshold transfer out of universe, and is the merge damage specific to singly-stored facts? Six runs (seeds 881, 887, 913, 677, 733, 919); the exposure block here mixes 12,000 facts (double density), and endpoint storage health varies widely across the six: 0.94–0.96 (913, 733, 919), 0.74–0.76 (887), and 0.12 / 0.03 (881, 677 — the exposure did not take, and the storage-health gate excludes them).

On the three gate-passing units (913, 733, 919): the frozen ΔNLL threshold of 2.0 — calibrated on universe 1 and applied unchanged — recalls 0.82–0.92 of merge deaths at precision 0.91–1.00 (FPR 0–13%); the marginal unit (887) gives 0.68–0.90 at 0.94–1.00. And the repaired control works (the universe-2 arm, K=128 per branch): facts stored by both branches survive the merge at 0.96–0.99 (the three gate-passing units) while branch-exclusive facts survive at 0.30–0.43 — refuting indiscriminate global collapse. A separate dose-matched arm closes the strength confound: shared facts at K=64 per branch (total dose equal to the exclusive facts' K=128) survive the merge at 0.89 vs 0.33 (seed 913; the failed-storage unit 881 gives 1.00 on its tiny cohort) — redundancy, not strength, protects; extended to four seeds: gate-passing 913/919/887 read 0.82–0.97 shared vs 0.23–0.40 exclusive (data-aggregates/dosematch_report.json; unconditional over all assigned facts, no endpoint conditioning: 0.86–0.95 shared vs 0.32–0.39 exclusive on the cleanly-stored units 913/919; the marginal-storage unit 887 reads 0.41/0.17, tracking its endpoint health). A randomized variant re-partitions the 18,000-fact pool per seed (which facts are shared vs exclusive is randomized, balancing difficulty across groups) and reproduces it: 0.875 vs 0.33–0.35 (seed 733; data-aggregates/randctrl_report.json); extended to four seeds (canonical ghost-Q99 criterion — 733 re-measures at 0.60 exclusive under it, vs 0.33–0.35 in the original single-seed report above): gate-passing 733/913/919 read 0.85–0.97 shared vs 0.53–0.69 exclusive — the exclusive level sits higher here because the re-partitioned pool is difficulty-balanced by design. A third universe (fresh generator seed, three runs, all reported: two pass the storage gate, one fails) replicates both the exclusive survival (0.32–0.37) and the dose-matched protection (0.89–0.97). Universe-1 survival levels replicate at the same order (0.30–0.43 at τ=0). A further clean-control arm removes the residual pre-fork exposure confound: the base segment renders filler only (shared facts appear nowhere before the fork), exposure unchanged (exclusive K=128 + shared K=64/branch). Three seeds (881/913/733, all reported): exclusive survival 0.19–0.31, shared both-alive $\rightarrow$ merge-keeps 0.87–0.89 — the protection is redundancy's, not exposure history's. (Storage note: with a filler-only base, endpoint storage drops — only seed 913 passes the 0.85 gate cleanly (0.89), 881 is marginal (0.74–0.76), 733 fails (0.31–0.35); the protection margin holds in all three regardless.) (data-aggregates/cleanctrl_report.json.) (data-aggregates/univ2_report.json; 881's τ=0 cells are degenerate by the storage failure and excluded from these statements; at τ=50M its threshold recall reads 0.81–0.86.)

Allocation-law arm. Universe-1 protocol with a filler-only base (shared facts appear nowhere before the fork) and the 6,000-fact shared pool re-partitioned into six cells of 1,000 (three splits $(K, 0)$, $(0.75K, 0.25K)$, $(0.5K, 0.5K)$ $\times$ two mirrored orientations) at two total-dose levels $K \in \{128, 256\}$; exclusive background unchanged (K=128 per branch). Three seeds (42/1042/2042, ITT). Unconditional merge survival (all assigned facts): at K=128, concentrated 0.23–0.29, 0.75/0.25 0.61–0.95, balanced 0.67–0.97; at K=256, 0.26–0.30 / 0.81–0.98 / 0.82–0.98 — six seed$\times$dose units all ordered balanced $\geq$ 3/4-split $>$ concentrated, and concentration does not recover at 2$\times$ dose. Conditional both-alive $\rightarrow$ merge-keeps is 0.87–1.00 throughout. Factoring the unconditional gap (per-cell decomposition of the released margin fields, data-aggregates/alloclaw_decomp.json): any-endpoint-alive storage is

comparable across arms (0.84–1.00 in the balanced arm vs 0.82–0.97 for the own endpoint in the concentrated arm), but conditional on storage the merge keeps a both-alive fact at 0.87–1.00 versus 0.26–0.31 for an own-alive fact in the concentrated arm (twelve cells each); in split arms, even facts alive at only one endpoint — a sub-threshold half-dose trace in the other branch — survive at 0.57–0.96 across the 24 split-arm cells (both doses, both split levels; the K=128 balanced cells read 0.61–0.92), well above the concentrated arm's 0.26–0.31 (this contribution is itself matched by the logit-average null of Appendix I, with an ensemble-relative retention deficit of 6–14 points against it). The allocation advantage is therefore not an artifact of endpoint storage counts. Assignment protocol: the 6,000-fact pool is split into the six cells by one seeded shuffle (generator seed 20260905), so group identity is fixed across training seeds (each seed's arms see the same partition; mirrors swap which branch carries a cell's larger share); facts are not stratified by attribute beyond the generator's own balancing, and the same entity's attributes are not grouped. The full assignment ledger is released (data-aggregates/alloc_assignment.json). Budget accounting (exposure renders / realized tokens / optimizer steps): each branch at K=128 sees 1,152,000 renders → 50.0M tokens → 1,525 steps, and the joint-training reference sees 2,304,000 → 100.0M → 3,051 — exactly the two branches' sum, so the joint comparison is exposure-matched and total-compute-matched (K=256 proportionally: 66.7M and 133.3M tokens). The main effect is computed over all assigned facts (unconditional); the per-endpoint-storage decompositions are descriptive of where the effect sits, not separate causal effects.

Allocation law on the 1B bridge. We apply the allocation design to OLMo-2-1B, provenance fixed by construction (the injection protocol of Appendix F). The allocation pool contains 2,400 facts in six groups of 400: three dose splits $(K, 0)/(0.75K, 0.25K)/(0.5K, 0.5K)$, each with two mirrored orientations; per-fact renderings are partitioned across branches without overlap. We use five training seeds and evaluate candidate-strip rank-1 accuracy per (fact, attribute) probe.

Both split allocations outperform concentrated exposure in all 10/10 seed×orientation combinations: unconditional merge survival is 0.34–0.39 (concentrated), 0.62–0.68 (3/4-split), and 0.64–0.68 (balanced), with the two split levels within 0.02 of each other and the ghost floor at 0.08–0.10. Conditional on endpoint storage, own-alive → merge-keeps reads 0.36–0.42 versus both-alive 0.71–0.82 (data-aggregates/e6alloc_fu_agg.json). A held-out open-generation check — no candidate set, greedy decoding, phrasings absent from training (data-aggregates/e6alloc_eval3_agg.json) — reproduces the ordering without any anchored strip: merge open accuracy concentrated 0.33–0.35 vs 3/4-split 0.59–0.63 vs balanced 0.61–0.65 ($\Delta_{\text{alloc}} = +0.26..+0.31$ in all five seeds at both split levels; ghost floor 0.08–0.10). A legacy-universe variant (the E6 bridge's original universe, whose attribute-assignment counter folded the vocabulary to 5 towns/3 jobs/2 objects/1 habit — disclosed and repaired for this arm) replicates the ordering at a higher ghost floor (data-aggregates/e6alloc_legacy_agg.json).

Endpoint saturation and the margin-linear null. A concave dose response alone produces the direction: if a fact's stored strength after dose k is $h(k)\, g_f$ with concave h, the balanced arm's merged contribution $h(K/2)\, g_f$ beats the concentrated arm's $\frac{1}{2}h(K)\, g_f$ with no cross-branch mechanism at all. We measure the dose response directly (endpoint storage by share; released fields): at 1B, alive-at-endpoint reads 0.75–0.81 / 0.82–0.88 / 0.87–0.91 / 0.90–0.94 at shares 0.25/0.5/0.75/1.0 ($h(K/2)/h(K) \approx 0.92$), and similarly concave on the testbed. We therefore fit the strongest magnitude-only null directly, on measured per-fact margins: the margin-space linear null of §5 ($\hat{m} = 0.5\, m_A + 0.5\, m_B$ per fact, judged against the merge's own ghost-Q99 threshold, in raw and threshold-normalized units; data-aggregates/alloclaw_marginnull.json, 36 cells = 2 doses × 3 seeds × 3 splits × 2 mirrors, 1,000 facts each). This null accounts for the main gap: predicted vs observed merge survival is 0.86–0.90 vs 0.88–0.90 on the split arms and 0.15/0.29 vs 0.27 on the concentrated arm — a continuous-strength account with averaging needs no new principle for the topline numbers. What it does not predict is the sub-threshold region. The average of two margins cannot exceed the larger one; in threshold-normalized units all three thresholds coincide by construction, which is where the bound is exact, so under this null any fact below threshold at both endpoints is predicted dead. Yet such facts survive the merge at 0.28 in the split arms

Table 18: Allocation gain decomposed against three output-ensemble references (per seed×dose unit, then pooled). Δ_M is the measured balanced-minus-concentrated merge-survival gain on the assigned pool (unconditional); "ref gain" is the reference's own balanced-minus- concentrated gain; "deficit Δ" is the change in the merge's deficit against that reference, $d_{E,\text{conc}} - d_{E,\text{bal}}$. The identity Δ_M = ref gain + deficit Δ holds exactly per unit. The three references are different predictors, not scale conventions of one: the probability mixture is the arithmetic mean of the two parents' predictive distributions, while the logit average is their normalized geometric mean.

	merge survival			raw logit-avg		norm logit-avg		prob-mixture	
unit	conc	bal	Δ_M	ref gain	deficit Δ	ref gain	deficit Δ	ref gain	deficit Δ
K128, s42	0.23	0.69	+0.45	+0.54	−0.09	+0.54	−0.09	+0.07	+0.38
K128, s1042	0.27	0.95	+0.68	+0.65	+0.03	+0.65	+0.03	+0.06	+0.63
K128, s2042	0.28	0.96	+0.69	+0.68	+0.01	+0.68	+0.01	+0.10	+0.59
K256, s42	0.28	0.84	+0.56	+0.55	+0.01	+0.55	+0.01	+0.10	+0.45
K256, s1042	0.27	0.98	+0.71	+0.70	+0.01	+0.70	+0.01	+0.09	+0.62
K256, s2042	0.29	0.98	+0.70	+0.67	+0.03	+0.67	+0.03	+0.06	+0.64
pooled	0.27	0.90	+0.63	+0.63	−0.00	+0.63	+0.00	+0.08	+0.55

(null: 0.002; concentrated arm: 0.05), and facts alive at exactly one endpoint exceed the null by 0.09–0.11 in the split arms. The sharper output-side null settles what that excess means: margins are concave in logits, so the margin of averaged logits can exceed the average of margins when the runner-up differs across endpoints: $m((z_A+z_B)/2) - \frac{1}{2}[m(z_A)+m(z_B)] = \frac{1}{2}[\max_{v\neq v^*} z_{A,v} + \max_{v\neq v^*} z_{B,v}] - \max_{v\neq v^*} \frac{z_{A,v}+z_{B,v}}{2} \geq 0$. Evaluating it directly on dumped per-candidate logits ($z_{\text{null}} = (z_A + z_B)/2$ per (fact, template) over the 1024-value strip, judged against the null's own ghost-Q99; data-aggregates/alloclaw_logitnull.json): the logit-average null accounts for the revival (0.27/0.32 predicted vs 0.28 observed on the both-endpoints-dead cohort) and overpredicts survival on stored cohorts (0.88–0.91 vs 0.75–0.77 on exactly-one-endpoint-alive; 0.94–0.96 vs 0.88–0.90 overall on split arms; 0.33 vs 0.27 concentrated). The allocation advantage and its sub-threshold texture are thus quantitatively anticipated by output-averaging of the measured endpoint states — the ensemble view. The ensemble reference inherits a scale convention, and we report the sensitivity explicitly (data-aggregates/alloclaw_logitnull_v2.json): normalizing each parent's logits before averaging changes nothing (0.27/0.32 vs 0.28 on the both-endpoints-dead cohort, as raw), while a probability-mixture ensemble predicts the concentrated arm at 0.87 against the merge's 0.27 — under that convention the gap reaches sixty points. Per fact, the logit-average null predicts survival at recall 0.90–0.98 and precision 0.75–0.96 (it over-predicts survival, never the reverse). What weight-space merging loses against the ensemble view — the ensemble-relative retention deficit — exists under every convention we measured; its size is convention-bound (6–14 points under logit averaging on stored cohorts: fact-pooled over 12,000 facts per arm, the penalty is 5.7–5.9 points overall on the split arms and 13–14 on the exactly-one-alive cohort; data-aggregates/alloclaw_logitnull_v2.json). Independently, the λ-recovery collapse at $\lambda \geq 1$ (0.01–0.28, below the merge) is outside any magnitude-only account: a full-strength sum of both updates kills more than halving them.

Reference-dependent decomposition of the allocation gain . The measured allocation gain is reference-robust; its attribution is not. Per (seed, dose) unit, $\Delta_M = S_{M,\text{bal}} - S_{M,\text{conc}}$ decomposes exactly as $[S_{E,\text{bal}} - S_{E,\text{conc}}] + [d_{E,\text{conc}} - d_{E,\text{bal}}]$, where $d_E = S_E - S_M$ is the merge's deficit against output-ensemble reference E (Table 18; data-aggregates/alloclaw_ensemble_decomp.json, produced by code/aug/alloclaw_ensemble_decomp.py). Under the logit-average references, the ensemble's own allocation gain already carries the measured gain (+0.63 vs Δ_M +0.63 pooled; the merge's deficit change ≈ 0): allocation improves the endpoints' combined output state. Under the probability mixture — a different predictor, not a rescaling of the logit-average one — the reference's own gain is small (+0.08) and the same measured gain appears instead as the merge falling further behind the reference on the concentrated arm (deficit change +0.55). Both readings are consistent with the same measured Δ_M; we therefore report the deficit as reference-operator-dependent, and we read the probability-mixture contrast as bounding what the merge leaves on the table, not as a scale artifact of one reference.

Held-out-template readout on the testbed. The allocation numbers are read on training-template probes. We re-aggregated the released margin fields under the held-out-template battery (battery = hold; alive = m_{top1} at or above the model's own ghost-Q99 within the same battery; the ghost pass rate is 0.010 by construction; data-aggregates/alloclaw_holdout.json, produced by code/aug/alloclaw_holdout_readout.py). Levels collapse floor-ward, as the training-template estimand predicts (merge survival 0.01–0.09 across arms) — but the law's direction survives: balanced beats concentrated in 6/6 seed×dose units (0.03–0.09 vs 0.01–0.04 across the twelve seed×dose×mirror cells; the 3/4 split tracks balanced at 0.03–0.08). Only the direction transfers across template batteries: under held-out templates even endpoint storage reads 0.05–0.08, consistent with the template-bound estimand of §4.4.

Joint-training reference and the operator cross. We train a matched-budget comparator: one model trained from the same filler-only base on the union of both branches' content (both exclusive sets at their branch doses plus every allocatable fact at the full dose K; per-model margin fields in data-aggregates/joint-margins/). At matched per-fact dose the no-merge reference reads 0.81–0.94 across the six seed×dose units; the balanced split's merge matches or exceeds it in 4/6 units (0.70–0.98) and trails by at most 0.15 (K=128, seed 42: 0.70 vs 0.85) — the balanced split's merge loses at most 0.15 to this exposure- and compute-matched joint reference, and beats it in 4/6 units. The operator cross: TIES (density 0.5) applied to the allocation arms' balanced endpoints reads below plain averaging in 12/12 (seed, dose, split) cells on the balanced arms (balanced split: 0.53–0.97 vs 0.70–0.98) and above it in 2/6 concentrated cells — operators reshuffle which facts die; averaging stays the best operator on the allocation the law recommends (data-aggregates/alloclaw_joint_compare.json; generator scripts code/aug/alloclaw_joint_analyze.py and code/aug/run_merge_audit_ties_on_balanced.py).

Filler-only-twin control. Isolates the foreign-fact term of the interference decomposition: B′ shares the base and optimizer state, receives the same-length filler-only exposure block (foreign-fact slots replaced by filler, same tokens/steps/lr schedule), then drifts the same 50M. Universe 1, seeds $\{42, 1042, 2042\}$, $\tau \in \{0, 50\text{M}\}$. The foreign-fact-learning delta, $S_{\text{meas}}(A, B) - S_{\text{meas}}(A, B')$, is negative in 6/6 (seed, τ) cells (-0.008 to -0.095); the generic-drift term, $S_{\text{meas}}(A, B') - S_{\text{dis05}}$, is sign-unstable ($-0.184$ to $+0.187$).

Decomposition probes. Three read-only analyses on frozen checkpoints refine the decomposition. (a) Recovery curve $\theta_0 + \lambda(\delta_A + \delta_B)$, $\lambda \in \{0.25, 0.5, 0.75, 1.0, 1.25\}$, universe 1, seeds {42,1042,2042}, $\tau \in \{0, 50\text{M}\}$: survival is non-monotone, peaking at $\lambda \approx 0.5$–0.75 (0.26–0.55 at τ=0) and collapsing at $\lambda \geq 1$ (0.01–0.28) — scaling up restores dilution loss but adds interference, so the loss is not a coefficient mischoice. (b) Pure-interference row $\theta_A + \beta \cdot \delta_B$ (own delta at full strength): at β=0.5, survival reads 0.49–0.87 at τ=0 but 0.16–0.41 at τ=50M — the foreign delta alone kills at depth. (c) Ladder columns (100M/274M, same combos): at 274M the near-total death is dilution-dominated (scaled own-branch 0.009–0.018 vs merge 0.009–0.011; own-full preserves 0.14–0.19); at 100M both terms visible with large seed variance in the reference.

Scale ladder (three seeds per size). Two further sizes, same universe, seeds {42, 1042, 2042}, $\tau \in \{0, 50\text{M}\}$, with a pre-declared dose ladder for the exposure block (K: 128 → 256 → 512 → 2048; gate: endpoint frac_rank1 ≥ 0.85). At 100M the gate needs two steps (seed 42: 0.61–0.66 → 0.83 at K=256 → 0.94–0.95 at K=512; seeds 1042/2042 pass at K=512 directly, 0.97–0.98), and the pattern replicates — merge survival 0.17–0.43 across the three seeds at τ=0, and the frozen t=2.0 screen reads recall 0.82–0.96 at precision 0.99. At 274M the gate passes only at K=2048 (seed 42: 0.01 → 0.42–0.51 → 0.93), and there the merge destroys nearly the whole cohort in all three seeds (survival 0.007–0.017 at τ=0). Survival falls across the ladder: 0.29–0.47 / 0.17–0.43 / ≈0.01 at 30M / 100M / 274M, each size at its own storage-gated dose, so the sequence is a dose-confounded trend rather than a scaling law. The frozen screen alarms broadly at 274M (recall 0.99–1.00); its FPR ranges 0–0.83 across seeds and is unstable there because few facts survive, so the high recall does not establish reliable per-fact screening in these cohorts. The pattern is not specific to 30M,

but the protocol's storage margin narrows with scale and the needed exposure dose grows with it; scale-extrapolating the protocol itself is future work.

J Two-dimensional survival surface: path-conditioned increments

The grid $\theta_0 + \alpha\,\delta_{\text{own}} + \beta\,\delta_{\text{foreign}}$ over $\alpha \in \{0.25, 0.5, 0.75\}$, $\beta \in \{0, 0.25, 0.5, 0.75\}$ (the frozen dis/merge points plus six new cells; 36 (seed, τ, branch) units at $\tau \in \{0, 50\text{M}\}$; data-aggregates/attrib2d_report.json) maps the survival surface of §4.3 — survival surface — each cell read under its own ghost-Q99, a fixed-FPR comparison across models, not a mechanistic decomposition. Median survival over units: at τ=0, (α, β)=$(0.5, 0)$ reads 0.412 and the merge point $(0.5, 0.5)$ 0.392 — the dominant term at zero drift is own-delta scaling (the endpoint reads ∼1 by construction); foreign interference adds little at τ=0 but grows with drift: at τ=50M, $(0.5, 0)$ reads 0.361 vs $(0.5, 0.5)$ at 0.233. The surface is non-monotone in β (a deep dip at $(0.25, 0.25)$: 0.072) — interference is real but not dose-monotone. An order-symmetric attribution removes the path dependence (released fields; data-aggregates/shapley_decomp.json): with corners $(1, 0)$, $(0.5, 0)$, $(0.5, 0.5)$, $(1, 0.5)$ and the fixed τ=0 cohort, the Shapley allocations at τ=0 read ϕ_{dilution} +0.39 to +0.49 vs ϕ_{foreign} +0.13 to +0.22 (6/6 units — three seeds × two branches — positive on both terms); at τ=50M they read +0.03 to +0.17 vs +0.09 to +0.25 (foreign ≥ dilution in 4/6 units, and the two dilution-leading units sit at the smallest foreign values) — the same story with no path choice: dilution first, interference compounding with drift.

K The three-line decomposition: full tables

Per-seed values of the residual and its components (α=0.5; signed medians on F_{live}; CI = fact-level bootstrap 90%, B=2,000). These are the raw values every test of §6 re-aggregates; per the statistical semantics of §2.3, the raw D_{dis} column is the quantity whose sign the Temperature test shows to be convention-dependent, and is reproduced here for completeness, not for sign-reading.

Barrier detail for the main finding (§4.1). The filler-loss curves are logged at 11 α points per (seed, τ) row (24 rows per campaign after the campaign-2 collection; the campaign-2 barriers are −0.1281 to −0.0382, median −0.0765, 24/24 ≤ 0; fact-side endpoint-mean barriers in campaign 2 are −0.3602 to −0.0276, median −0.2299, 24/24 ≤ 0). The fact-side curves use the facts' own balanced macro-averaged teacher-forced NLL over answer tokens on F_{live} (5 α points; prompt-token CE was not logged). The mixed-loss construction $L_{\text{mix}}(\alpha; p) = p \cdot L_{\text{fact}} + (1 - p) \cdot L_{\text{filler}}$ spans the barrier family a practitioner could compute from these two distributions ($p \in \{0, 1\}$ recover the pure forms), each mixture curve being their convex combination; it keeps the 24-cell median at ≤ −0.0400 for every p, never past +0.05.

Own-endpoint baseline. The endpoint-mean baseline above is half-random by construction: on exclusive facts, the foreign endpoint sits near ghost level, so the mean of the two endpoint losses is inflated and a barrier against it is nearly vacuous. Against the own endpoint (the branch that stored the facts), the merge's per-fact logp drops by +1.53 to +3.51 (campaign medians +2.53 / +2.47 / +2.65) in 144/144 (seed, τ, branch) cells across all nine runs — and the per-fact ΔNLL against the own endpoint separates the facts that die (alive at the own endpoint, dead at the merge, frozen rule) from the survivors: per-campaign medians 2.82 / 2.64 / 3.00 (dead) versus 1.12 / 0.95 / 0.92 logp (surviving). ΔNLL > 0 recalls 96.7–100% of deaths but fires on 93–94% of survivors; Table 20 gives the full sweep (ΔNLL > 2.0: recall 70–81% at precision 94–98%). By contrast the same diagnostic against the endpoint mean recalls only 0.3972 and the dead group's median ΔNLL is negative in 18/24 campaign-1 cells. Computed read-only from the frozen margin fields (data-aggregates/own_endpoint_barrier.json, data-aggregates/perfact_threshold.json).

Instrument robustness. Six follow-ups on the ΔNLL screen, all read-only on the frozen margin fields (data-aggregates/perfact_threshold_v2.json). (a) Per-run macro: pooling facts

Table 19: Three-line decomposition per seed (α=0.5; values carry the frozen pipeline's per-cell presentation scale — their signs are the content). Columns: $D_{\text{lin}} = m_{\text{meas}} - m_{\text{lin}}$ (merge vs linear null), $D_{\text{obs}} = m_{\text{dis}} - m_{\text{lin}}$ (reference vs null), $D_{\text{dis}} = m_{\text{meas}} - m_{\text{dis}}$; the identity $D_{\text{lin}} = D_{\text{obs}} + D_{\text{dis}}$ holds per fact and within 0.1443 for medians (§B); $\text{NMAD}_x = \text{median}|d_x|$ in the same scale.

seed 42

τ	NMAD_{lin}	NMAD_{dis}	D_{lin}	D_{obs}	D_{dis}	CI90(D_{dis})
0	0.9023	1.1250	+0.7642	+1.5117	−0.7266	[−0.7578,−0.6953]
3.1M	0.9209	1.1641	+0.7617	+1.5234	−0.7656	[−0.7969,−0.7344]
6.3M	0.8828	1.2109	+0.7168	+1.5964	−0.8594	[−0.8828,−0.8281]
12.5M	0.9207	1.3281	+0.7402	+1.7363	−0.9688	[−1.0078,−0.9375]
25M	0.9844	1.4062	+0.7889	+1.8477	−1.0234	[−1.0703,−0.9844]
50M	1.0645	1.5469	+0.8223	+1.9180	−1.1406	[−1.2031,−1.0781]
100M	1.1738	1.6172	+0.8941	+2.0586	−1.1953	[−1.2578,−1.1250]
200M	1.1533	1.7188	+0.8282	+2.0053	−1.1016	[−1.2229,−0.9803]

seed 1042

τ	NMAD_{lin}	NMAD_{dis}	D_{lin}	D_{obs}	D_{dis}	CI90(D_{dis})
0	0.5430	0.6562	+0.0469	+0.1172	−0.0781	[−0.0938,−0.0547]
3.1M	0.5977	0.7500	+0.1211	+0.3750	−0.2500	[−0.2656,−0.2266]
6.3M	0.7246	0.8594	+0.2207	+0.5312	−0.3281	[−0.3516,−0.2969]
12.5M	0.9688	1.1250	+0.3945	+1.0352	−0.6328	[−0.6719,−0.6016]
25M	1.4365	1.7109	+0.5781	+1.7900	−1.1797	[−1.2500,−1.1250]
50M	1.8218	2.2422	+0.7412	+2.5020	−1.7266	[−1.8047,−1.6484]
100M	2.0488	2.5469	+0.8828	+2.8584	−1.9883	[−2.1075,−1.8672]
200M	2.1934	2.8281	+0.7930	+3.2422	−2.4375	[−2.5469,−2.2734]

seed 2042

τ	NMAD_{lin}	NMAD_{dis}	D_{lin}	D_{obs}	D_{dis}	CI90(D_{dis})
0	0.4141	0.5078	+0.0000	+0.1328	−0.1328	[−0.1484,−0.1172]
3.1M	0.4492	0.5703	+0.0859	+0.3359	−0.2422	[−0.2578,−0.2266]
6.3M	0.5547	0.6875	+0.1953	+0.5625	−0.3516	[−0.3750,−0.3359]
12.5M	0.7666	0.9375	+0.3438	+0.9297	−0.5781	[−0.6094,−0.5469]
25M	1.2969	1.5391	+0.5313	+1.6133	−1.1016	[−1.1562,−1.0391]
50M	1.7754	2.1484	+0.6963	+2.4541	−1.7578	[−1.8242,−1.6797]
100M	1.9849	2.5625	+0.6709	+2.9316	−2.2266	[−2.3359,−2.1172]
200M	2.2003	2.7656	+0.7021	+3.2051	−2.4980	[−2.6094,−2.3047]

Table 20: Per-fact ΔNLL alarm against the own endpoint (§4.1): fact-pooled recall (fraction of merge-killed facts alarmed), precision (fraction of alarms that are deaths), and false-alarm rate (fraction of survivors alarmed), by threshold and campaign (c1: seeds 42/1042/2042; c2: half-dose seeds 3184/5383/7192; c1x: the campaign-1 extension 6784/403/1189; 481,879 cohort facts total, 339,400 deaths). The sign rule (t=0, first row): maximal recall, but it alarms on 93–94% of survivors.

ΔNLL threshold	recall (c1 / c2 / c1x)	precision (c1 / c2 / c1x)	false-alarm (c1 / c2 / c1x)
> 0.0	1.00 / 1.00 / 1.00	0.70 / 0.73 / 0.72	0.93 / 0.93 / 0.94
> 1.0	0.97 / 0.96 / 0.98	0.79 / 0.84 / 0.84	0.56 / 0.47 / 0.46
> 1.5	0.92 / 0.86 / 0.93	0.86 / 0.92 / 0.91	0.32 / 0.20 / 0.23
> 2.0	0.79 / 0.70 / 0.81	0.94 / 0.98 / 0.96	0.12 / 0.04 / 0.08
> 2.5	0.62 / 0.54 / 0.66	0.99 / 1.00 / 0.99	0.02 / 0.00 / 0.02
> 3.0	0.44 / 0.41 / 0.50	1.00 / 1.00 / 1.00	0.00 / 0.00 / 0.00

within each run, recall at threshold 2.0 spans 0.46–0.85 (median 0.81) and precision 0.89–1.00 across the nine runs. (b) Prevalence: with recall and false-alarm rate as the prevalence-free quantities, precision is 0.94–0.98 at the testbed's 70% death prevalence and would be 0.42–0.66 / 0.27–0.48 / 0.06–0.15 at 10% / 5% / 1%. (c) Score–label separation: with death defined on the held-out template battery (the score stays on the training battery), recall at t=2.0 is 0.44–0.60 at precision 0.72–0.76 — the screen reads the storage layer, not full generalization. (d) Baseline: the merge's own absolute logp, thresholded, reaches recall 0.68–0.71 at precision 0.97–0.98 — the own-endpoint reference raises recall over this checkpoint-free baseline. (e) Aggregate matching: the own-endpoint aggregate correlates with the death fraction at $r = 0.47$ across the 144 cells; the 53 cell pairs matched within ±0.15 logp on the aggregate differ by 20–26 points of death fraction. (f) Independent damage validation: on the §6 damage machine's certified strata (independent recall instrument, 10,427 forgotten / 23,047 retained; 31% prevalence), the frozen rule (threshold 2.0) scores TPR 0.52 at FPR 0.009 (threshold 1.5: 0.71 / 0.039) — specific and conservative on manufactured damage (data-aggregates/dnll_posctrl.json).

L Additional co-measured readouts

274M base-budget control: per-seed ledger. The recipe-ablation control of §5.2, per seed (4× base budget, K=128, τ=0): endpoint exclusive storage passes the gate on every seed and branch (s42: 0.992/0.994; s1042: 0.989/0.990; s2042: 0.978/0.965); merge exclusive survival s42 0.512/0.527, s1042 0.486/0.485, s2042 0.330/0.346; the scaled own-branch (dis05) reference reads 0.673/0.727, 0.585/0.601, 0.530/0.444 — merge below the reference on both branches of every seed. Seed variance in the recovered survival is real (0.33–0.53) and stays inside/above the 30M–100M band (0.17–0.47) throughout; the fixed-recipe 274M value was ≈0.01. The branch-lr arm (same 300M base): K=128 storage 0.009–0.011 at lr $6 \cdot 10^{-4}$ and 10^{-3}; K=256: 0.11–0.16.

These readouts were measured alongside the main grid and are reported for completeness; none changes any conclusion of §6 or any finding of §4.

Template two-layer signature. Each (seed, τ) cell is classified by a two-battery signature: StoreKeep holds if both branches keep R1keep ≥ 0.049 (rank-1 retention relative to the own endpoint; the bar is calibrated on a no-merging reference world), ReadKeep if the held-out-template margin of the merge stays positive, ReadLive if the endpoints answer held-out templates at all. All 24 cells keep R1keep $\geq$ **0.2088** — zero storage_injury cells (signature classifications of behavior, not storage-location measurements). The signature is stably intact at $\tau \leq 12.5$M and mixed across seeds at $\tau \geq 25$M (intact / readout_injury / readout_floor), driven by endpoint readout-layer decay under drift; there is no stable readout_injury band, and floor cells are stated as "not detected under a 3-template battery".

Filler-anneal repairability. After merging, a short anneal (5M tokens of pure filler, lr 3e−5, fresh AdamW) revives a fraction $\rho_{\text{repair}} \leq$ **0.0144** of merge-killed facts in all 24 cells (pooled

0.0057 / 0.0028 / 0.0077): under this protocol and budget, merge-killed template-bound memory did not revive. Endpoint controls cover 3 of 8 τ tiers.

The shared-fact control is empty. The shared pool's frac_rank1 stays in 0.0022–0.0077, indistinguishable from never-rendered ghosts (0.0020–0.0105; own exclusive facts: 0.89–0.97): the intended same-dose control has no signal in this universe, and a generic-degradation alternative to the endpoint comparison's reading cannot be excluded within it (§4.4, item 4).

M Threats to validity (full inventory)

(0) Precision. Training keeps fp32 master parameters (bf16 autocast for compute); merges average the stored endpoints in fp32 and evaluation runs in fp32, so the only weight-rounding step is the bf16 endpoint export. We do not propagate that rounding analytically, but bound it empirically: re-merging after perturbing both endpoints by uniform one-ulp noise moves merge survival by 0.006 (0.293 $\rightarrow$ 0.299 on the K=128 seed-42 unit; data-aggregates/margins.k128-s42-t0-merge05-ulp.jsonl.gz). The fine-grained residual readouts (e.g., the 6–14-point ensemble deficit) are stated as measured; the real-merge cosines/norms are computed in float64.

Beyond the seven bounds of §4.4: (i) τ growth co-varies with branch filler-loss improvement and representation drift (branch separation grows 36.41 / 48.75 / 47.39 to 71.54 / 108.18 / 111.49; endpoint-to-endpoint path length at τ=200M is $\approx$ 2077–2097; the δ_A, δ_B cosine stays in 0.43–0.73) — the drift axis is a bundle, and the paper's claims are direction claims over that bundle, not single-channel causal claims. (ii) The nested design shares one trajectory across the 8 τ points, removing data-order confounds at the price of within-τ correlation; cross-seed claims rest on three independent runs per unit (nine total). (iii) The exposure-dose axis of the main drift grid is a single point (K=128; the allocation and dose arms add K=256–2048); the two campaigns differ in final dose (64 vs 128) via the protocol's mechanical ladder, so their readouts are not pooled anywhere. (iv) The survival threshold discretizes a continuous margin at Q99; the quantile family around it is exactly what the Quantile test measures. (v) The repair protocol is a single point (budget, lr, data), and its endpoint controls cover 3 of 8 tiers. (vi) Optimizer-state inheritance is a single choice. (vii) The readout-layer battery has 3 templates; deep-τ floor cells cannot separate "no signal" from "too few templates". (viii) The training-template battery is isomorphic to the training distribution (a training-template estimand, by design); knowledge-extraction semantics belong to the held-out templates only.

N Related work on storage and retrieval readouts

Ponkshe et al. (2026) study memorization dynamics under an extraction-style readout rather than a two-battery format readout — a different estimand from our storage/readout signature. Wu et al. (2026) measure generalization to held-out prompt formats and are the general predecessor of our readout-layer battery, without a merging axis. Fan et al. (2026) audit probe-choice effects and warn that probe readouts can mislead — a sibling caution at the readout level, where our tests act at the residual level. Zhang (2025) measure memorization under dedup-style interventions without merge audits. None combines controlled mutually exclusive fact sets, per-fact continuous margins, an α path, a τ dose axis, and null references.

Terminology. We use "not identified" in the methodological sense of reference/scale-convention dependence (the causal sense requires a defined estimand, intervention, and assumptions (Holland, 1986)); our null models play the role of negative controls (Lipsitch et al., 2010), and where they fail to discriminate (Null-reference test) we report that.

Baseline dependence and memorization. The attribution literature documents the analogous baseline sensitivity (Sundararajan et al., 2017; Kindermans et al., 2019; Sturmfels et al., 2020); model editing compares pre/post probabilities against references (De Cao

et al., 2021); our Reference and Quantile tests are the merge-audit instances of that pattern. Forgetting-as-a-residual descends from continual-learning practice (Lopez-Paz & Ranzato, 2017). The aggregate-loss finding anchors on the long-tail account of memorization (Feldman, 2020; Feldman & Zhang, 2020; Jagielski et al., 2023): forgetting concentrates on a small, memorization-dependent subset of examples, so an aggregate statistic can stay smooth exactly where that subset is being damaged.

Merge-side neighbours, in detail. Structure-aware merging localizes task-specific coordinates (Matena & Raffel, 2022; Yang et al., 2024b; Wang et al., 2024) and engineers around the common/task-specific subspace split (Marczak et al., 2025; Wei et al., 2025); we measure what the split costs per fact. The closest line is the shared-vs-unshared forgetting trio. Zaman et al. (2024) (Fuse to Forget) established the direction on synthetic shortcuts and GPT-2 memorization and showed validation perplexity can improve while unshared memorization drops — the aggregate-silence instance of our main finding; we add per-fact provenance by construction, fixed-total-dose allocation, and the instrument boundary. Hess et al. (2025) document task-specific forgetting under merging in controlled continual learning; our facts are per-fact storage with ghost floors rather than visual cues. Yadav et al. (2025) find merging outcomes improve with scale on task suites; our ladder is consistent at the per-fact level, tracking survival at each size's storage-gated dose rather than suite scores. Yuan et al. (2026) name parameter-update dilution as the failure mode for merging RL-trained agents and repair it by preserving unique updates — our dilution term is the same object measured, not repaired. Hammoud et al. (2024) report the same silent-loss signature on safety alignment. Memorization-scaling (Tirumala et al., 2022; Carlini et al., 2023) bounds how to read the ladder's gated doses; Neyshabur et al. (2020) is why same-base fine-tunes sit in one low-error basin at all. The nearest memorization-side precedent is federated: Thakkar et al. (2021) showed that at equal canary exposure, user-clustered data memorizes less than randomly repartitioned data under FedAvg — exposure distribution across participants matters there; our allocation arm is the single-merge, fixed-total-dose, per-fact analogue. (MeMo (Quek et al., 2026) merges NarrativeQA shards against joint retraining as a utility–cost question — a sibling use of merging, not controlled shared/exclusive survival evidence.)

O Figure and data provenance

Script filenames keep the project's original internal figure numbering (e.g., make_fig4_data.py builds Figure 8's data); the mapping below is authoritative.

The released margin fields are enumerated with sha256 hashes in the release manifest (manifest.sha256.json: 313 + 269 + 267 files over the three campaigns/units).

All figures are generated from the released artifacts, read-only. The released scripts below regenerate all panels from the same artifacts. Figure 2 (the main allocation figure) is rendered by code/aug/fig_allocation_nulls.py directly from data-aggregates/alloclaw_marginnull.json, data-aggregates/alloclaw_logitnull.json, data-aggregates/e6alloc_fu_agg.json, and data-aggregates/e6alloc_eval3_agg.json. Figure 5 from the temperature-sweep artifact plus the six-seed re-aggregation code/aug/paper_backfill.py (campaign-1 columns reproduce the frozen artifact at 4dp); Figure 7 from the survival fields (per-seed values in Table 17); Figures 4 and 3 from the same re-aggregation (reference-family scan; stratified power decomposition, which reproduces the frozen power check's per-cell TPR/FPR to 4dp), all via code/figures/src/make_fig_data.py and code/figures/src/make_fig_backfill_data.py. Figure 6 is re-aggregated from the per-fact margin fields by code/figures/src/make_fig6_data.py (rendered by plot_fig6.py); its verification block reproduces Table 12's campaign-1 source artifact to 5e−7 over all 48 campaign-1 cells and the campaign-1 rows of Table 7 row by row.

Figure 8 uses the exact α=0.5 fields of the 11-point filler-loss curves and the fact-side NLL curves with the survival fields (make_fig4_data.py + plot_fig4.py; 24 cells, no interpolation). The normalized form's per-τ values (Table 12) were re-aggregated with the same

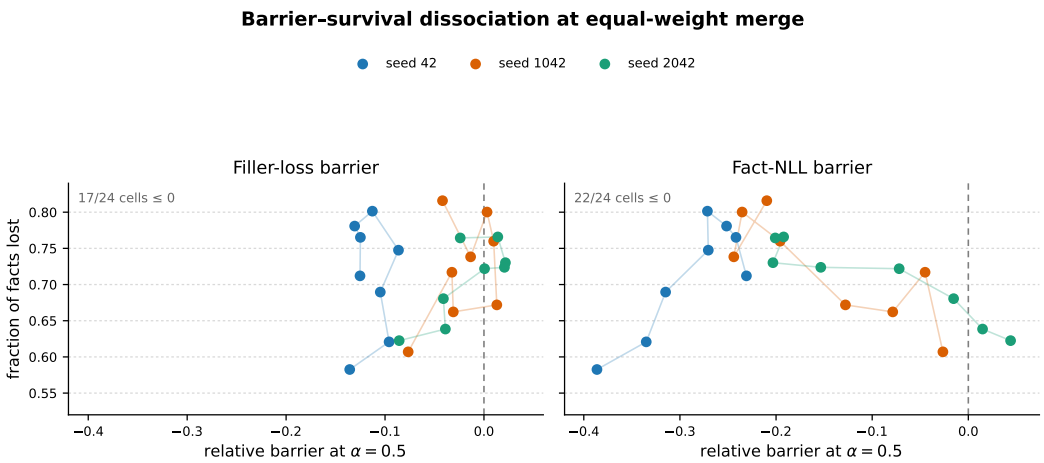


Figure 8: Barrier–survival dissociation. Each panel contains the same 24 run–checkpoint cells from the original three campaign-1 runs at $\alpha = 0.5$. The horizontal coordinate is the relative barrier on filler loss (left) or fact-side NLL (right), measured against the corresponding endpoint-mean baseline. The vertical coordinate is the fraction of own-endpoint-surviving exclusive facts lost after merging; A/B survival is pooled by cohort size within each cell. Colors identify seeds, and thin lines join successive checkpoints within a seed; they are neither fitted relationships nor uncertainty intervals. The dashed vertical line marks zero barrier. Filler-loss barriers are nonpositive in 17/24 cells (range -0.1357 to $+0.0217$), and fact-NLL barriers in 22/24 cells (range -0.39 to $+0.04$), while 58–82% of the cohort is lost. Both panels share axes, with the vertical display zoomed to $[0.52, 0.84]$.

code path as the temperature sweep and reproduce its t=1 values bitwise at τ=0 in all six campaign-1 cells.