RESOLUTION LIMITS OF A CONTINUAL-LEARNING BENCHMARK FOR LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Small improvements in continual learning can be difficult to reliably distinguish from score variation due to task order, retraining seed and evaluation sampling. We quantify comparison reliability in a five-task classification benchmark with three language model backbones, trained under the same adaptation procedure. Across changes within the order, pooled rankings remain consistent, but magnitude and direction changes. A frozen fresh-seed run continuation verifies development agrees transfers or not. At five seeds per disjoint group, 13 out of 16 continue to meet a criterion of 95% sign agreement across seeds, splitting this holdout into replay vs non-replay comparisons gives us 9/9 and 4/7 agreeing comparisons respectively. This endpoint concerns agreement within each cohort, not preservation of the old winner or a replication probability. In a separate experiment-60 trajectories study, we find examples of close non-replay comparisons that change within fixed execution environments. Mixing hosts is thus not the only source of variation. Conditioned on fixed trained models, jointly resampling evaluation datasets across all training and testing stages leads to a median evaluation-sampling standard deviation of 0.0119, versus 0.0158 when resampling at each stage independently. We provide as part of the audit an executable that will reproduce these paired effects, uncertainty, declarations, abstentions and the observed cost of selective reporting. Reported near-superiority on new runs cannot be certified without assessing each method's difference and each reporting decision. These results hold for this research, but do not establish universal seed requirement.

1 INTRODUCTION

Short text-classification streams commonly evaluate continual adapter learning (Wang et al., 2023a; Yang et al., 2025; Zhao et al., 2024). Selecting a broadly strong baseline and establishing a small improvement are different questions. We investigate why stable pooled rankings coexist with uncertain close comparisons, using repeated training runs, fresh-seed follow-ups and matched evaluation records.

Ranking stability and comparison reliability. Similar pooled rankings across backbones coexist with failures of agreement on fresh runs, including within fixed hosts (Fig. 1). Task order also changes effect magnitudes and sometimes directions. A retrospective audit of 475 historical records examines declarations, disagreements and abstentions. Separate studies specified before their outcomes test fresh-seed agreement and within-host consistency. All comparisons concern recorded implementations under one recipe; shared tuning and incomplete historical identities limit attribution to published algorithms.

1. **Agreement in development need not persist on fresh runs.** Failures of the frozen agreement criterion occur among non-replay pairs. The fixed-host follow-up shows that mixing execution environments is not their sole explanation (Sections 4.1–4.2).
2. **Similar pooled rankings coexist with order-dependent effects.** Direct tests and a routed-adapter case study reveal changes in magnitude and direction (Section 5.3). The routed bank serves as an analytical instrument.
3. **Selective reporting trades fewer contradictory declarations for more abstention.** We report paired effects, decisions, sensitivities and abstention costs, including pair-specific failures (Section 6). A real input-to-decision walkthrough supports reuse (Appendix G.1).

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.

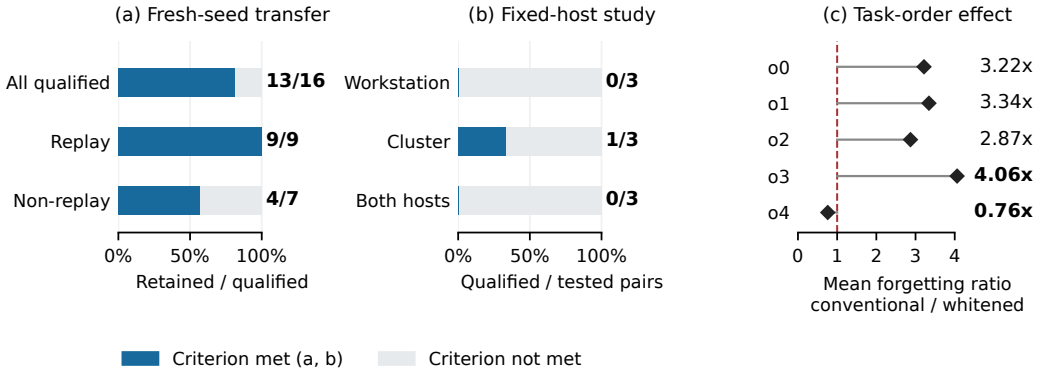


Figure 1: **Fresh-run agreement and task-order sensitivity.** (a–b) At least 95% agreement on forgetting-difference signs between disjoint five-seed groups. (a) Frozen endpoint and post-hoc replay split; denominators are development-qualified comparisons. (b) Common-prompt fixed-host study; denominators are all tested non-replay pairs, and "both" is the intersection. Populations and scorers differ. Retention need not preserve the old winner. (c) Conventional embedding-context / standardized hidden-context forgetting ratios, using three-seed means on one machine; above unity favors whitening, below favors conventional routing. Ratios are descriptive. Figure 9 gives absolute means and standard deviations (SDs); Appendices N and Q give all agreement results.

2 Related work

Continual learning with adapters. Rehearsal-free continual learning (CL) with large language models (LLMs) often trains parameter-efficient adapters on a frozen backbone: low-rank adaptation (LoRA), elastic weight consolidation (EWC) anchoring, orthogonal subspaces, adaptive rank, prompts and mixture-of-LoRA routing. Evaluations use the short classification streams of de Masson d'Autume et al. (2019) and LFPT5, a prompt-tuning method whose four-task comparison uses three orders; Appendix J surveys them. Small replay buffers (Chaudhry et al., 2019; de Masson d'Autume et al., 2019) remain strong and under reported.

Order, variance, and benchmark reliability. Prior work measures order dependent forgetting (Yoon et al., 2020; Bell & Lawrence, 2022; Lin et al., 2023) and calls for broader CL evaluation (Farquhar & Gal, 2018; Díaz-Rodríguez et al., 2018). Our analysis follows studies of experimental variance (Henderson et al., 2018; Lucic et al., 2018; Bouthillier et al., 2021; Agarwal et al., 2021; Dodge et al., 2020; Card et al., 2020; Colas et al., 2018) and benchmark composition (Dehghani et al., 2021). Gulrajani & Lopez-Paz (2021) demonstrate the importance of model selection and tuning in benchmark comparisons; our shared-recipe experiment does not estimate each method's tuned optimum. Tran et al. (2026) find large curriculum effects in multimodal CL. Closest to our resolution question, Yauney et al. (2026) measure which pairwise model rankings survive evaluation set reduction, while Zhang et al. (2026) study cross-benchmark rankings after standardized fine tuning. We examine repeatedly trained continual learners: task order, execution discrepancies and correlated multi-stage forgetting. Our contribution is this empirical investigation under a specified harness, including a new-seed test of whether conditional subset agreement transfers between cohorts.

3 Experimental design

Study populations and questions. An *arm* is one implementation/configuration, a *trajectory* one training stream, and a *cohort* a specified set of runs. Figure 2 separates the studies; Table 2 maps their populations. *Within-cohort agreement* compares method-difference signs between disjoint seed subsets; *agreement transfer* compares this criterion between cohorts. Reporting rules declare a winner or *abstain* (decline to name one). These quantities are not future replication probabilities.

Benchmark and metrics. We use five text-classification tasks (Zhang et al., 2015): AG News, Amazon binary polarity, Yelp five-class reviews, DBpedia and Yahoo. This differs from O-LoRA's

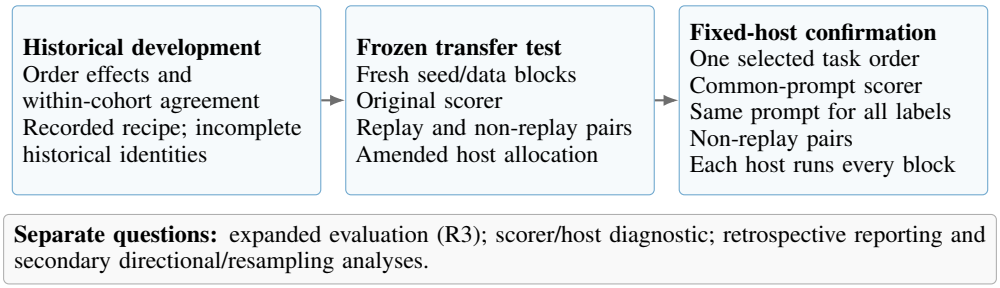


Figure 2: **Which evidence answers which question?** Arrows show follow-up questions, not pooled evidence. Transfer and fixed-host plans preceded their outcomes; secondary analyses reuse records retrospectively. Table 2 retains the full population map.

stream (dataset identities, labels, prompts and truncation: Table 3, Appendix A.1). Runs sample 1,000 training examples per task, paralleling O-LoRA's long-stream subsample (Wang et al., 2023a), and usually 250 evaluation examples. The expanded-evaluation study (R3) uses 7,600 per task, matching the full-setting count of information disentanglement based regularization (IDBR) (Huang et al., 2021): all AG News test examples and seeded samples of the other tasks. Historical row identities are incomplete. Training uses 800 steps/task, batch size 4, AdamW at 10^{-4} (Loshchilov & Hutter, 2019), 32-bit floating point (fp32) and 192 tokens. Prediction selects the label with greatest length-normalized log-likelihood.

Let $R[i, j]$ be accuracy on the task at stream position j after training stage i, with $T = 5$. Final average accuracy (AA; higher is better) is $T^{-1} \sum_{j=1}^{T} R[T, j]$. Signed average forgetting (AF; lower is better) is $\frac{1}{T-1} \sum_{j=1}^{T-1} (\max_{1 \le i < T} R[i, j] - R[T, j])$ (Chaudhry et al., 2018; 2019). The maximum excludes the initial and final rows but can include stages before task j was trained; final improvement can yield negative AF. Appendix V tests a post-acquisition alternative. Scores use accuracy fractions unless stated otherwise; pp denotes percentage points.

Task orders. Orders o0–o2 start with AG News, Yahoo and Yelp. LFPT5/O-LoRA's published orders omit Yelp, so we use our own five-task sequences. Adding orders starting with Amazon and DBpedia makes each task lead once (o0–o4; Appendix A). These two orders were completed after the initial three-order analysis (Appendix E). Only the leading position is balanced: AF omits the terminal task. Equal-task reweighting leaves pooled ranks unchanged (Appendix A); Appendix V also checks shared non-terminal tasks.

Routed case-study instrument. A low-rank expert bank on frozen Qwen tests sensitivity to task order. Its router uses prompt embeddings or frozen hidden states, optionally standardized per dimension using running statistics (called *whitening* here). Appendix U specifies the instrument and ablations.

Methods and study sizes. The *core five* share trainer, data and seeds: sequential LoRA (r=64) (Hu et al., 2022), +EWC (γ_{ewc}=0.25) (Kirkpatrick et al., 2017), O-LoRA (Wang et al., 2023a), fixed-rank singular-value-decomposition LoRA (SVD-LoRA; AdaLoRA without rank allocation) (Zhang et al., 2023), experience replay with a buffer of 2% of the complete stream (Chaudhry et al., 2019). Five routed arms use K=8 experts, select k=2, and have rank r=8: embedding context (*R-emb*); raw hidden-28, unwhitened (*R-raw28*); whitened hidden-21 (*R-wL21*, the arm whose 3-order effect we reported first); whitened hidden-28 + consolidation (*R+cons*); and gate + consolidation (*R+g+c*), both γ=0.5. Sequential LoRA at r=64 matches the routed expert bank parameter count; routing and gate heads are additional and SVD-LoRA retains 1.5× that rank: our runner omitted AdaLoRA's allocation callback, so these are not adaptive AdaLoRA results (Table 4); equal totals do not equalize active parameters, floating-point operations (FLOPs) or memory, and shared hyper-parameters test one recipe, not each method's best setting (Appendix A). These ten arms, five orders and three seeds form the 150-run *canonical ten-arm grid*, on Qwen2.5-1.5B unless stated.

Extended populations. Three additional ports share the recipe: interference-free low-rank adaptation (InfLoRA) (Liang & Li, 2024), a factor-L_1 N-LoRA variant (Yang et al., 2025), and a LoRAMoE-inspired mixture-of-experts (MoE) port (Dou et al., 2024). Their departures from the published methods limit attribution (Appendix A). The defective Progressive Prompts port is excluded from comparisons but archived. Thus, eight baseline implementations and five routed arms form the 13-arm eligible pool. Separate core-five extensions use orders o5–o9 (each task leads twice; plan in Appendix E), additional seeds 3–5, then ten seeds on o0–o4 (250 cells), and Llama-3.2-3B and SmolLM2-1.7B. The fresh follow-up uses four primary methods (core five without SVD-LoRA); the fixed-host study uses three (also omitting replay).

Statistics. Unless stated otherwise, results are means over three seeds per order. A *reversal* is a winner change across orders; we count reversals of order means and those occurring in at least two of three individual seed indices. Headline paired t-tests use 15 seed–order differences unless another unit is specified. Pairing matches sampling seed and order, not training noise; fresh-cohort hashes verify data identity, whereas historical identities are incomplete. Counts retain order-specific differences; cell-level tests do not first average over orders. Pooled ranks, backbone comparisons and reference-model gaps use across-order means. "n.s." means two-sided $p > 0.05$ at the stated degrees of freedom (df). Appendix M audits reporting rules and seed/order influence.

4 What this benchmark can resolve

We test agreement within and between cohorts. Agreement qualification (below), sensitivity to historical re-execution discrepancies (Section 4.2), and interval-based reporting decisions (Section 6) answer different questions; none provides a universal sample-size prescription.

The agreement calculation. The core-five, ten-seed grid has 250 runs over o0–o4 (Fig. 3b). For one method pair and order, let $d_s = \mathrm{AF}_{A,s} - \mathrm{AF}_{B,s}$ for aligned seed s. Enumerate disjoint seed groups S, U, each of size k, and compare their mean differences. Agreement requires the same nonzero sign; ties in either group count as non-agreement. A comparison qualifies at k if at least 95% of group pairs agree. This is an agreement criterion, not a confidence level.

At $k = 1$, 30 of 50 pair–order comparisons qualify, leaving a largest unresolved gap of 0.058. Using gaps between ten-seed means, all gaps above 0.05 qualify at two seeds per group, above 0.03 at four, and above 0.02 at five. Ten total seeds permit groups of at most five; they do not resolve every gap above 0.01. Gap size alone is insufficient: Appendix A reports smaller qualifying gaps alongside larger nonqualifying ones. These finite-cohort rates provide neither a power guarantee nor confidence intervals. The three core-baseline reversal margins lie below the gap threshold covering all comparisons here.

Evaluation sampling also contributes. Holding each trained model fixed, jointly resampling 250 examples per task across stages gives a median SD of AF of 0.0119, below the independent-stage approximation (Appendix K). This is single-run metric uncertainty, not uncertainty in a paired method difference. Shared evaluation subsets align pairs but do not eliminate sampling variation; training and evaluation samples both change across seeds.

Averaging stabilises development rankings. Agreement between disjoint seed-group rankings rises from 53% at $k = 1$ to 94% at $k = 5$ (Fig. 3b). Order accounts for 14–54% of each method's total sum of squares in an additive decomposition of the 250 cells; averaging seeds does not test additional orders. Appendix A.1 retains the full ranking, reversal and variance breakdown, including the six-seed check preceding machine assignment. Here a seed jointly changes training/evaluation samples and initialization.

4.1 Does conditional agreement transfer?

Does agreement persist in fresh runs? The locally frozen follow-up contains 150 fresh trajectories: 120 primary runs (four implementations, three existing orders, ten seeds), plus 30 secondary adaptive-AdaLoRA runs. At 5 seeds per disjoint group, 13 of 16 development-qualified comparisons again meet the 95.00% criterion. Mean absolute development-to-fresh agreement-rate discrepancy is 14.46 percentage points across all 18 pair–order comparisons. The retained/qualified counts are

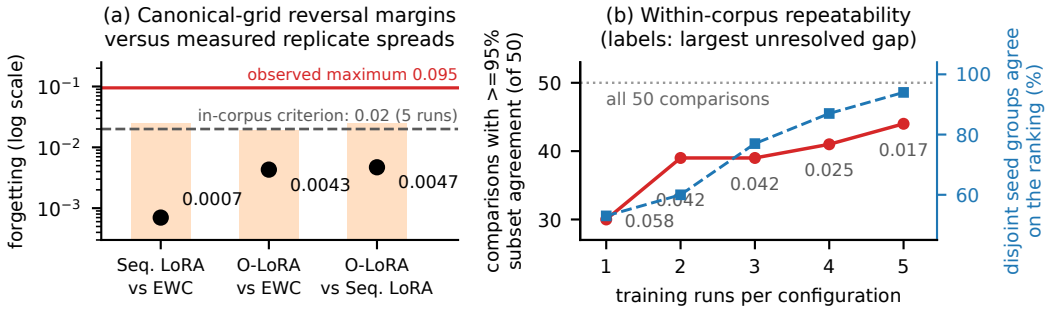


Figure 3: **Small reversal margins are below observed replicate discrepancies.** (a) Named pairs identify each reversal. Dots show the smaller of the largest margins in the two directions; bars show the larger observed re-execution spread of the two methods. The three core-baseline reversals are each decided by a margin below both arms' replicate spreads, the historical global observed replicate maximum, and the gap threshold covering all comparisons at five seeds per group (panel b). (b) One seed per group qualifies 30 of 50 pair–order comparisons; five seeds qualify 44; right axis, ranking agreement between disjoint seed groups in this corpus.

9/9 for replay pairs and 4/7 otherwise; all 3 failures occur on o2. This is within-cohort agreement, not preservation of the old winner or a replication probability. Amended allocation mixes hosts. Same-host o2 non-replay subsets contain 3–5 seeds, allowing $k = 1$ or 2, not $k = 5$. Appendix N reports every group size and failure. Overlapping partitions provide no universal seed requirement. Four primary methods give six pairs on each of three orders: 18 comparisons, of which 16 qualify in development. Retention can hold even if development groups favor A and fresh groups favor B. At five seeds per group, the two groups exhaust the same ten-seed cohort: their means sum to twice the full-cohort mean. These dependent partitions are not independent replication trials (Appendix N).

Appendix N separately reports cross-cohort mean signs and the adjacent sequential LoRA/EWC pooled rank change. Neither that transposition nor retained agreement establishes superiority. The study did not test the later retrospective declaration rules.

4.2 Evaluation and execution sensitivity

Expanded evaluation. The expanded-evaluation study (R3) completes 75 new trainings of the core five, retaining the same pooled ranking with 7,600 evaluation examples per task. Reversing pairs fall from 3 to 1 of 10, with 45 of 50 pair–order winners unchanged. Neither permutation null is exceeded (Appendix K). This comparison mixes training and evaluation changes. Only resampling within a fixed training run isolates evaluation sampling. The coarse ranking survives, while evidence from small reversing margins remains weak. A separate paired bootstrap uses identity-matched fresh-cohort predictions, holding training fixed (Appendix O); it does not estimate retraining uncertainty.

Scorer and host diagnostic. Common-prompt scoring gives all candidate labels the same retained input. Across 18 fresh trajectories, changing scorers on identical learned states changes AA by -0.160 to +0.080 and signed AF by -0.100 to +0.200 percentage points. It strictly reverses 0/18 per-seed method-difference signs for AF (three pairs, three seeds, two hosts), with 0 tie transitions. Across hosts, 2/18 AF comparisons reverse (three pairs, three seeds, two scorers reusing trajectories). Three seeds on one selected order cannot establish universal equivalence or hardware causality. Full results, AA signs and host shifts: Appendix P.

Execution sensitivity is a separate diagnostic. Historical re-execution comparisons cover 79 eligible pairs in the archive, matched on recorded flags. All six same-device repeats match exactly. The largest attributed eligible arm discrepancy across recorded environments is 0.095. These are observations, not a hardware effect or a bound on future discrepancies. The three core baseline reversals turn on margins of 0.005, 0.004 and 0.0007, each below the observed arm specific spreads. Figure 3a and Appendix C report both per-pair and global comparisons, their uneven coverage and

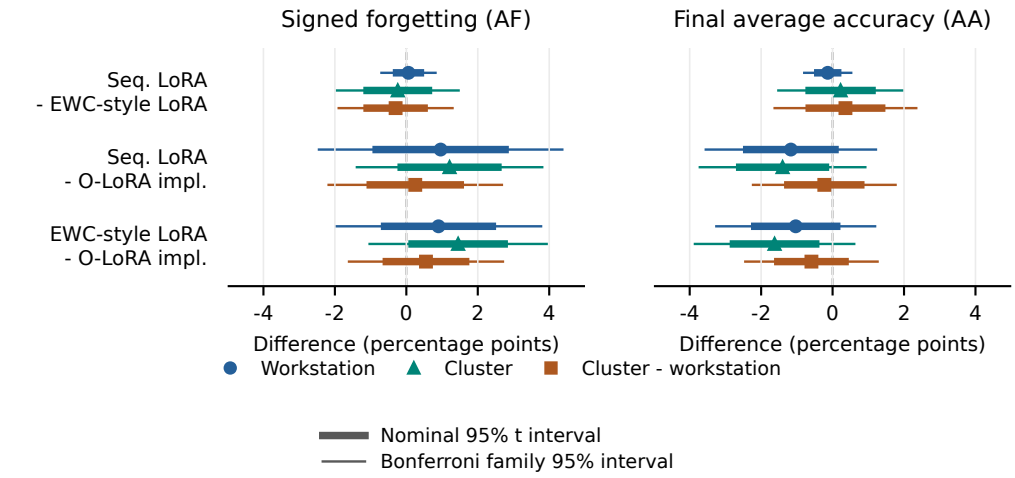


Figure 4: **Fixed-host comparisons remain uncertain under common-prompt scoring.** All three method pairs; means over ten aligned seed/data blocks. Thick/thin lines: nominal/Bonferroni-family 95% t intervals for the same 18 effects (degrees of freedom $df = 9$). Within hosts, positive signed-forgetting (AF) differences favor the second method, positive final-average-accuracy (AA) differences the first; host change is the paired cluster-minus-workstation shift in that contrast. Appendix Q reports all effects and the single 18-test Holm family. Non-rejection is not equivalence; execution differences do not identify a hardware cause.

exclusions. These empirical sensitivity criteria do not control error rates or replace the cohort analysis.

Prospective fixed-host comparison. A separate 60-trajectory study runs sequential LoRA, EWC-style LoRA and our O-LoRA implementation on each host, with 10 aligned fresh seed/data blocks on o2. Common-prompt scoring is primary. At $k = 5$, 0/3 pairs meet the $\geq 95\%$ AF agreement criterion on the workstation, 1/3 on the cluster and 0/3 on both. Failure of this criterion persists within hosts; mixing hosts is not its sole explanation. The prespecified family yields 0/18 Holm rejections across AF/AA method differences and paired host changes (Fig. 4, Appendix Q). Non-rejection does not establish equivalence. This common-prompt population leaves the earlier original-scorer endpoint unchanged. 6 later recovered alternative records extend the nominally matched candidate discrepancy to 0.101, and Appendix S reports every hypothetical substitution without changing the historical cohorts.

5 Order effects and supporting case studies

5.1 What the reversal counts do, and do not show

Table 6 gives the per-order values for 8 eligible baseline implementations, and Table 1 the reversal counts for 6 analysis pools. We permute each arm's five order labels independently within each seed (the null that order carries no signal; 1,000,000 permutations per pool; one-sided p). A second null applies one permutation per seed to every arm, preserving shared (seed, order) blocks while destroying alignment of between arm differences across orders. The separate six pool Holm corrections yield 0 rejections for independent permutations and 1 for block preserving permutations. The ten-arm block-preserving test rejects ($p_{\text{Holm}} = 0.0348$). It includes the routed variants and cannot be read as a result about the core baselines alone. The core-five count does not reject either null. Under post-acquisition forgetting, the ten arm rejection also disappears (Appendix V); this rejection depends on the metric definition. Non-rejection is not evidence of equivalence. The frozen historical result reproduces, but

the exact version of its block-preserving null loses this rejection in 8 of 16 hypothetical substitutions of recovered, nominally matching outcomes; the rejection is sensitive to candidate choice (Appendix S).

Direct order-effect tests ask whether mean scores vary across the chosen orders. On the ten-seed grid, additive order-plus-seed analyses of variance (ANOVAs) reject for 4 of 5 methods after Holm correction. Treating seed as a repeated measures block and applying a Greenhouse–Geisser (GG) covariance sensitivity (Greenhouse & Geisser, 1959) leaves 3 rejections: replay and O-LoRA do not reject. Evidence therefore depends on covariance assumptions. Visible magnitude differences alone do not establish a population order effect. The original method×order interaction is $F(16, 216)=4.27$, $p=3\times10^{-7}$. Appendix L reports an alternative repeated response analysis and defines the inference units. Order effects on paired method differences reject for 7 of 10 pairs nominally, 6 after Holm and 4 under GG plus Holm (Table 8). The seed indices combine data sampling and initialization. The fixed set of chosen orders is not a random sample of every possible curriculum.

Counts are also metric-dependent: on average accuracy, 0 of 10 core baseline pairs reverses (Table 1). The pooled order is replay, O-LoRA, sequential LoRA, EWC, SVD-LoRA, with adjacent margins 0.050, 0.024, 0.002 and 0.081. Only the last exceeds the historical 0.064 replicate maximum on this metric. This is a descriptive sensitivity comparison, not a test of statistical separation.

Effect magnitudes also vary across orders. The small reversing margins among the core baselines are smaller than the historical observed replicate maximum (Section 4), while their absolute forgetting varies: over these five orders the core five swing 1.5–$4.1\times$, replay—best overall—has the largest ratio but smallest absolute swing among the core five (Table 6), and over ten orders O-LoRA's swing widens to $6.5\times$ (Table 7). Our whitened routed arms swing comparably or more (4.2–$5.6\times$, Table 13). The two unwhitened routed arms swing less. This order sensitivity is why we do not present them as a method contribution. O-LoRA's published comparison with LFPT5 also reverses across orders (Appendix J); our quantitative resolution estimates remain specific to our harness.

How many orders, and how many seeds. Sub-sampling the grids (Fig. 10b), the mean reversal rate over the 10 order extension rises 15%, 22%, 30% at $m=2, 3, 5$ orders and **is still climbing at $m=10$ (40%)**, as the null predicts for close pairs under noise—a count that can only grow with orders, so the informative part is that it has not saturated. Replay is best among the core baselines in all 10. Seeds cut the other way: re-running those baselines with three more seeds (six total, o0–o4) dissolves 2 of the 3 reversals, leaving 1/10, unchanged at 10 seeds.

5.2 Two further backbones: similar pooled ranks, varying order effects

With identical hyperparameters, we repeat five orders and three seeds on Llama-3.2-3B (Grattafiori et al., 2024) (different vendor, tokenizer and corpus; $2\times$ the parameters) and SmolLM2-1.7B (Allal et al., 2025) (third vendor). Reversal counts are 4/10 on Llama (5/10 within-seed) and 3/10 on

Table 1: **Reversal counts require an explicit null and multiplicity family.** Winner changes across the five orders (3 seeds) per pool, against a null permuting each arm's order labels independently within each seed (no order effect for any arm) and, last two columns, against one permutation per seed shared by all arms (block-preserving: no method×order interaction); "On avg. acc.", the same count on *average accuracy*. Dashes denote unreported accuracy counts.

Pool	Pairs	Reverse	Null mean	p	On avg. acc.	Null (block)	p (block)
Core five	10	3 (30%)	2.20	0.38	0	1.63	0.25
Eight external	28	10 (36%)	9.92	0.65	8	8.83	0.31
Ten arms	45	19 (42%)	16.27	0.03	—	15.05	0.006
All thirteen	78	33 (42%)	30.55	0.14	29	28.48	0.03
Core five, Llama-3.2-3B	10	4 (40%)	2.33	0.12	—	1.41	0.07
Core five, SmolLM2-1.7B	10	3 (30%)	1.75	0.16	—	0.84	0.03

SmolLM2, versus 3/10 on Qwen (Tables 10–11). Neither exceeds the independent permutation null; SmolLM2's unadjusted block-null result fails Holm correction. Neither backbone has a measured replicate maximum.

Across 15-run means, Qwen and Llama rank the five methods identically: replay, O-LoRA, sequential LoRA, EWC, SVD-LoRA. SmolLM2 exchanges only EWC (0.1329) and sequential LoRA (0.1335), an unrounded gap of 0.0007, smaller than Qwen's historical replicate maximum but not a validated cross-backbone separation. Replay leads every order on all three. Pooled rankings are similar while order-specific rankings and effect magnitudes vary (Table 7).

5.3 Order-dependent gains: a routed-adapter case study

A routed-adapter comparison motivated this audit (Fig. 1c). In a complete single-machine replication, routing on a standardized hidden state instead of prompt embeddings reduces forgetting by $2.23\times$ pooled over five orders, but its conventional-to-whitened forgetting ratio changes from $4.06\times$ on o3 to $0.76\times$ on o4: the latter order favors the conventional router. The improvement's size and sign therefore depend on the selected order. Clustering paired differences by order gives $t{=}2.77$ (df$=4$), so an order level inference remains fragile.

The factor and calibration studies suggest that drifting normalization statistics, rather than merely the selected hidden layer or first task calibration frame, account for the observed behavior. Freezing the statistics removes the advantage in the tested variants; allowing them to drift recovers it on some orders. Historical environment differences were real, and matched re-runs retained the order sensitivity. Appendix U.1 reports the full results, non-rejections, configuration defects and repairs.

5.4 Where forgetting is concentrated

In the ten-seed grid, training Yahoo produces the largest observed single-step accuracy drop on DBpedia. This directed pair is also largest on both further backbones (Fig. 7). Replay neutralizes this drop and ranks best among the core five in every tested order on each backbone, but receives additional example exposure: this is not equal-compute superiority.

Pre-specified replacement streams tested the task-pair explanation. Removing either endpoint failed to halve the largest remaining directed drop or remove the order effect. Replay's margin shrank sharply when Yahoo was replaced, but persisted more strongly when DBpedia was replaced. The control retained the original dominant pair. Substitution changes task format and labels, so these outcomes do not isolate a pair specific cause. Appendix B retains every prediction and outcome, including failed control clauses. Table 9 gives the complete comparison.

6 An inspectable reporting workflow

From a comparison to a report. The audit validates stage-by-task accuracy matrices and their method, order, seed and provenance identifiers. It then computes paired effects and uncertainty, applies a declaration rule, and reports order/seed sensitivity. The scripts and worked example reconstruct this path. Reuse requires eligible records and a reporting rule suited to the new question.

Practical use. Before observing outcomes, specify the population and meaningful effect. Start with record validation and paired estimates by order, reporting unresolved differences. Add order, execution or sampling diagnostics as the claim requires. Separate evaluation resampling from retraining variation and check pilot-based seed guidance on fresh blocks. No universal budget or error rate is prescribed.

Declarations and consistency. On the ten-seed Qwen core-four pool (sequential LoRA, EWC, replay and O-LoRA), we enumerate disjoint three-seed decision and three-seed reference groups across 30 pair–order comparisons (126,000 group-pair opportunities per rule). "Opposite" means the reference favors the other method (Table 19).

Reporting rule	Declare (%)	Opposite given declaration (%)
Mean sign	99.25	6.84
Separate SD bands	58.11	1.10
Paired t interval	35.50	0.95

The nominal paired-t rule declares less often with a lower aggregate opposite-sign rate, but not for every pair. For sequential LoRA versus EWC on order 0, it declares in 35 of 4,200 opportunities; 14/35 declarations (40.00%) oppose the reference mean sign; the mean-sign rule has 1,042/4,165 (25.02%) opposed declarations. Noisy references and overlapping groups make these dependent finite-corpus descriptions, not error probabilities, coverage guarantees or a seed prescription. SD bands are a heuristic, not confidence intervals.

Secondary directional check. After cohort completion, historical three-seed declarations were compared with each pair's ten-fresh-seed mean direction on o0–o2 (2,160 overlapping opportunities). Mean sign declares in 99.17%; 15.31% of declarations oppose the fresh reference. The paired-t rule declares in 34.95%, with 0.79% opposed. Without replay, declaration/opposition rates are 98.33/30.89% and 15.19/3.66%, respectively. The frozen endpoint is unchanged. References are noisy and reused across subsets; cohort and host changes remain confounded. These are not error probabilities (Appendix O). The fresh sequential-LoRA/EWC nominal reference intervals include zero. Most opposed declarations concern such uncertain references. Appendix V separates them from references excluding zero. Neither stratum supplies population truth, so an opposite sign does not prove a historical declaration wrong.

A worked reporting decision. For a retrospectively selected illustration, take Qwen order 0 and the original seeds 0, 1 and 2 (forgetting, lower is better). Sequential LoRA minus EWC has mean -0.0050 and nominal paired 95% t interval $[-0.0205, 0.0105]$: a mean-only ranking favors sequential LoRA, but the interval rule abstains. Sequential LoRA minus replay has mean 0.0850 and interval $[0.0726, 0.0974]$, retaining a declaration of lower forgetting for replay under this recipe. These intervals are assumption-dependent and unadjusted for multiple comparisons. Replay uses additional training work; this is not an equal-compute comparison. The example illustrates a reporting decision on existing runs, not a prediction for new runs (input-to-decision trace: Appendix G.1).

The full checklist is in Appendix T.

7 Limitations

We study five-task classification at 1.5B–3B on three backbones; incomplete longer-stream, larger-data and 7B extensions are excluded. Chosen orders and one shared recipe limit transfer to other curricula and tuned algorithms. Historical identities and checkpoint links are incomplete; re-execution discrepancies do not identify hardware causes. Expanded evaluation changes both training and evaluation, while resampling conditions on trained states. Fixed-rank SVD-LoRA differs from adaptive AdaLoRA; the defective prompt port is excluded. Replay exposure and growing capacity preclude equal-compute claims. Routed arms were tuned on the initial order; task-pair causal predictions were unsupported.

The 150-run follow-up covers three original orders on one backbone; amended host allocation limits attribution. Directional and bootstrap analyses are post hoc. Scorer and fixed-host studies cover one selected order. Seeds combine data and initialization; fixed-host intervals assume independent blocks and approximately normal paired differences. No equivalence margin was registered. These studies establish neither general scorer equivalence, hardware causation, benchmark failure nor successful plasticity gating. Appendix A retains the full scope statement.

8 Conclusion

Stable pooled rankings do not settle close comparisons. Task order changes effects, and fresh runs can fail the agreement criterion within hosts. Our audit reports paired effects, uncertainty, reporting decisions and abstention costs. Prospective budget selection and transfer to broader models remain open.

REPRODUCIBILITY STATEMENT

The supplement provides the recorded matrices, source code, configurations and executable audits. Appendix A documents datasets, prompts, implementations, provenance, archive contents and reconstruction limits. Appendix E preserves fixed plans and deviations. Separate fresh-seed, scorer/host and fixed host studies have their own evidence and protocols (Appendices N, P and Q). Appendix G.1 follows one real input through saved scores to a reporting decision. Registered numerical claims and generated artifacts are checked against source records. Incomplete historical identities and omitted checkpoint tensors limit exact replay; the detailed disclosures are retained in Appendix A.

ETHICS STATEMENT

This work uses only public benchmark datasets and open-weight models and involves no human subjects.

AI USE STATEMENT

Generative AI tools assisted implementation of the routed adapters, reproduced baselines, data preparation and analysis scripts; experiment orchestration; and feedback on research methodology and experiment plans. AI-assisted drafting, interpretation and consistency audits informed corrections to the manuscript, checked against raw records and recomputed statistics. AI feedback also helped develop the conceptual framing of benchmark resolution, the per-pair and global comparison rules, and the reporting standard. It assisted formulation of the evaluation-sampling calculation, which uses standard probability results and was numerically checked against the observed accuracy matrices. The pipeline for subsampling public benchmarks and rendering prompts and verbalizer labels was AI-implemented. Additional uses included figure generation, LaTeX, drafting, restructuring and editing manuscript text, and proposing the title. This paper contains no proofs; synthetic data generation, translation and qualitative or thematic data analysis are not applicable. Validation includes gradient-equivalence tests against an independently implemented oracle, raw-record-to-matrix checks, reproducible statistical scripts, numerical claim checks and rendered-document inspection. All experimental results come from recorded runs; no experimental result was generated or estimated by a language model. The authors take responsibility for the final content, including AI-assisted text, claims, code and figures.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, et al. SmolLM2: When Smol goes big—data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Samuel J. Bell and Neil D. Lawrence. The effect of task ordering in continual learning. *arXiv preprint arXiv:2205.13323*, 2022.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Sepah, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Conference on Machine Learning and Systems (MLSys)*, 2021.

Nghia Bui, Guergana Savova, and Lijing Wang. Assessing the macro and micro effects of random seeds on fine-tuning large language models. In *International Joint Conference on Natural Language Processing and Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, 2025.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision (ECCV)*, 2018.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*, 2018.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Natalia Díaz-Rodríguez, Vincenzo Lomonaco, David Filliat, and Davide Maltoni. Don't forget, there is more than forgetting: new metrics for continual learning. *arXiv preprint arXiv:1810.13166*, 2018.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness. Zenodo, v0.4.3, `https://doi.org/10.5281/zenodo.12608602`, 7 2024.

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

Samuel W. Greenhouse and Seymour Geisser. On methods in the analysis of profile data. *Psychometrika*, 24(2):95–112, 1959. doi: 10.1007/BF02289823.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=lQdXeXDoWtI`.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yufan Huang, Yanzhe Zhang, Jiaao Chen, Xuezhi Wang, and Diyi Yang. Continual learning for text classification with information disentanglement based regularization. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Sen Lin, Peizhong Ju, Yingbin Liang, and Ness Shroff. Theory on forgetting and generalization of continual learning. In *International Conference on Machine Learning (ICML)*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? A large-scale study. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. MoELoRA: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*, 2024.

Chengwei Qin and Shafiq Joty. LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of T5. In *International Conference on Learning Representations (ICLR)*, 2022.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Fan-Keng Sun, Cheng-Hao Ho, and Hung-Yi Lee. LAMOL: LAnguage MOdeling for lifelong language learning. In *International Conference on Learning Representations (ICLR)*, 2020.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Van-Tuan Tran, Shruthi Gowda, Merim Dzaferagic, and Marco Ruffini. Is our benchmark enough? An analysis of continual learning for MLLMs. *arXiv preprint arXiv:2606.20961*, 2026.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023a.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, and Xuanjing Huang. TRACE: A comprehensive benchmark for continual learning in large language models. *arXiv preprint arXiv:2310.06762*, 2023b.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Shuo Yang, Kun-Peng Ning, Yu-Yang Liu, Jia-Yu Yao, Yong-Hong Tian, Yi-Bing Song, and Li Yuan. Is parameter collision hindering continual learning in LLMs? In *International Conference on Computational Linguistics (COLING)*, 2025.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2e2960f2fe9e981f33f51c78656e3ca2-Abstract-Conference.html.

Jaehong Yoon, Saehoon Kim, Eunho Yang, and Sung Ju Hwang. Scalable and order-robust continual learning with additive parameter decomposition. In *International Conference on Learning Representations (ICLR)*, 2020.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning (ICML)*, 2017.

Guanhua Zhang, Ricardo Dominguez-Olmedo, and Moritz Hardt. Train-before-test harmonizes language model rankings. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/79a97ba70ee6aff126f3090b10ad6702-Abstract-Conference.html.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Weixiang Zhao, Shilong Wang, Yulin Hu, Yanyan Zhao, Bing Qin, Xuanyu Zhang, Qing Yang, Dongliang Xu, and Wanxiang Che. SAPT: A shared attention framework for parameter-efficient continual learning of large language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Donglin Zhuang, Xingyao Zhang, Shuaiwen Leon Song, and Sara Hooker. Randomness in neural network training: Characterizing the impact of tooling. In *Conference on Machine Learning and Systems (MLSys)*, 2022.

A REPRODUCIBILITY DETAILS

Guide to the evidence. The appendices retain full results, unfavorable outcomes and implementation limitations. The full routed instrument and its investigation follow the reporting checklist, keeping the central evaluation evidence together. Use the following routes from the main questions:

Population and recipe: Table 2; dataset, prompt and implementation details in Appendix A; fixed plans and deviations in Appendix E.

Agreement and execution: fresh-seed transfer in Appendix N; scorer diagnostic in Appendix P; fixed-host results in Appendix Q; historical re-executions in Appendices C and I.

Order effects and failed explanations: routed case study and specifications in Appendix U; task-pair predictions in Appendix B; full historical tables in Appendix F; expanded evaluation in Appendix K.

Real input to reported conclusion: exact prompt, recorded scores, stage accuracies, forgetting and paired decision in Appendix G.1.

Reporting and evidence audits: retrospective decisions in Appendix M; secondary fresh-cohort analyses in Appendix O; recovered records and alternative outcomes in Appendices R and S; checklist in Appendix T.

The audit motivates four reporting practices; Appendix T distinguishes reproducibility requirements, coverage heuristics and population-specific diagnostics.

1. **Specify the population.** Report orders, seeds, evaluation size, recipe and provenance. State task-position coverage and show per-task forgetting; balancing leading tasks is a coverage heuristic.
2. **Measure uncertainty at the right level.** Separate evaluation sampling from retraining. Match data and order; disclose environment changes and missing identities.
3. **Test the claim actually made.** Give reversals a null model and test order effects directly. Report effect sizes and multiplicity; non-rejection is not equivalence.
4. **Make evidence auditable.** Link cells to raw records, validate configurations, quarantine failures, and regenerate artifacts together. Test pilot-based guidance on new cohorts.

Reading key. An *arm* is one tested implementation/configuration; a *trajectory* is one complete training stream and its evaluations. A *cohort* is the set of runs used by a specified study. A seed/data block aligns sampled data and initialization settings across implementations; it is not a guarantee of identical training noise. A *qualified* comparison meets the stated agreement criterion within one cohort. *Retention* means qualifying again in the fresh cohort, even if both groups now favor the other method. A reporting rule *abstains* when it declines to name a winner.

Label	Meaning in this paper
AA / AF	Final average accuracy / signed average forgetting (Section 3).
LA	Learning accuracy: mean accuracy immediately after each task's own training.
BWT	Backward transfer: final accuracy minus accuracy immediately after learning, averaged over non-terminal tasks; positive values indicate improvement.
SD / s.e. / df	Standard deviation / standard error / degrees of freedom. Captions state which observations the reported variation summarizes.
GG / LOO	Greenhouse–Geisser covariance adjustment / leave-one-out analysis.
PEFT / MLP	Parameter-efficient fine-tuning (also the library name) / multilayer perceptron.
MoE / SVD	Mixture of experts / singular value decomposition; historical SVD-LoRA is fixed rank, not adaptive AdaLoRA.
EMA / SDPA	Exponential moving average / scaled dot-product attention.
EOS / fp32	End-of-sequence token / single-precision floating point.
R1	Environment-matched reruns of earlier historical cells.
R2	Layer-by-normalization controls for the routed case study.
R3	Expanded-evaluation study, distinct from the fresh-seed and fixed-host studies.

R1–R3 are study identifiers, not method acronyms. The routed labels denote embedding context (R-emb), raw layer-28 context (R-raw28), whitened layer-21 context (R-wL21), whitening plus

consolidation (R+cons), and gate plus consolidation (R+g+c); their exact configurations are in Section 3.

Complete scope statement. We study five-task classification at 1.5B–3B scale on three backbones; incomplete longer-stream, larger-data and 7B extensions are excluded. Ten of the $5!$ orders are sampled. The 79-pair historical replicate reference lacks complete revision/data identities and does not establish portability. Routed arms were tuned on o0. Most evaluation uses 250 examples per task; R3 mixes retraining and evaluation changes, and its resampling isolates only conditional evaluation uncertainty. Shared hyperparameters compare one recipe, not tuned optima. Historical fixed-rank SVD results cannot establish adaptive AdaLoRA performance; adaptive results are separate. The defective prompt port is excluded. Growing banks and replay add capacity or exposure; no equal-compute superiority is claimed. Task replacement changes more than pair identity; its pre-specified causal predictions were not supported. The 150-run transfer cohort uses three original orders on one backbone, without balancing leading tasks; amended host allocation limits attribution. Directional and bootstrap follow-ups were selected after outcomes. The separate scorer diagnostic and fixed-host study cover one selected order. Seed blocks jointly change data and initialization; fixed-host intervals assume independent blocks and approximately normal paired differences. Neither study establishes general scorer equivalence or isolates hardware causality; the fixed-host study registers no equivalence margin. We claim neither general benchmark failure nor successful plasticity gating.

Release contents and verification scope. The frozen historical archive contains 1,603 audited result files (1,590 distinct by content; between 1,590 and 1,602 distinct executions, Appendix A; per-run `result.json` with the full accuracy matrix, and for 1,345 of them a `MANIFEST.json` recording model id, seed and configuration), the 96 runs quarantined by our own audits—kept under `results/_quarantine` so the failure record is itself reproducible—the training package and its entry point, and scripts that recompute the historical analyses. These are included in the supplementary material, supporting a new execution of the documented recipe; missing historical revisions prevent exact reconstruction. A versioned manuscript eligibility policy excludes defective implementations without altering the archived records. Eligible cells are resolved through an explicit provenance registry (`scripts/provenance.py`) mapping each (method, seed, order) cell to exactly one run directory and refusing any run whose recorded configuration departs from its arm's pinned signature; `scripts/compute_stats.py` recomputes every statistic through it and refuses to emit a block while any of its cells is missing, and `scripts/audit_manuscript_numbers_v2.py` checks registered prose claims against their named statistics; generator checks verify included numerical artifacts, exiting non-zero on a detected mismatch. Manifest coverage is partial—1,345 of the 1,603—and the manifest's `git_commit_hash`, `model_revision` and `dataset_hashes` fields are unpopulated in every manifest but one, so the historical archive pins neither code revision nor dataset hashes per run; the registry, not the manifest, is the authority on which run backs which number. One block bypasses the registry: the 3B backbone-integrity evaluations are resolved by explicit path, and their evaluation records store no pointer to the checkpoint they scored (Appendix A). These checks establish agreement between registered claims, generated artifacts and saved evidence; they do not certify all possible scientific interpretations. A separate implementation and claim audit checks known behavior contracts, multiplicity conclusions and exclusions. Appendix A lists hyperparameters, hardware and the exact task orders; Appendix E reproduces four historical fixed analysis plans and their logged deviations. The separate frozen cohort, its source/model/data fingerprints, host-assignment amendment and analysis are released under `confirmation/` (Appendix N); its records do not enter the historical census. The scorer/host diagnostic is released under `diagnostic/` (Appendix P). The separate 60-trajectory fixed-host study is released under `execution_confirmation/`, with frozen protocols, source identities, retained scores, checkpoint-validation receipts and exact numerical reconstruction (Appendix Q). Both remain outside the historical census; checkpoint tensors retained remotely are omitted from these portable releases. The later inventory is archived separately under `freshness/`, with its operational cutoffs and eligibility audit in Appendix R.

Table 2: Which population answers which question. Rows overlap where stated; their counts must not be added. Training trajectories, evaluation examples and enumerated subset comparisons are different units.

Population	Recorded variation	Target of analysis
Ten-seed core-five grid	Qwen: five implementations, five orders, ten seeds	Conditional seed-subset agreement and order-specific effects
Expanded evaluation (R3)	Core five, five orders, three seeds; new trainings and larger evaluation	Pooled-rank comparison; within-training evaluation sampling
Reporting audit	Historical Qwen, Llama, SmolLM2 and R3 records	Declarations, abstentions and opposite disjoint-reference signs; retrospective
Frozen follow-up	Four primary implementations, three orders, ten fresh seeds; adaptive AdaLoRA secondary	Transfer of within-cohort agreement; winner transfer reported separately
Secondary follow-up	Historical and fresh records reused after outcomes	Old-to-fresh declaration consistency; identity-gated paired evaluation resampling
Scorer/host diagnostic	Three methods, three seeds, one order, both hosts	Identical-state scoring sensitivity; matched execution differences
Fixed-host study	60 trajectories; three methods, 10 fresh seed blocks, one order, each host	Common-prompt agreement within hosts; paired method and host-change effects

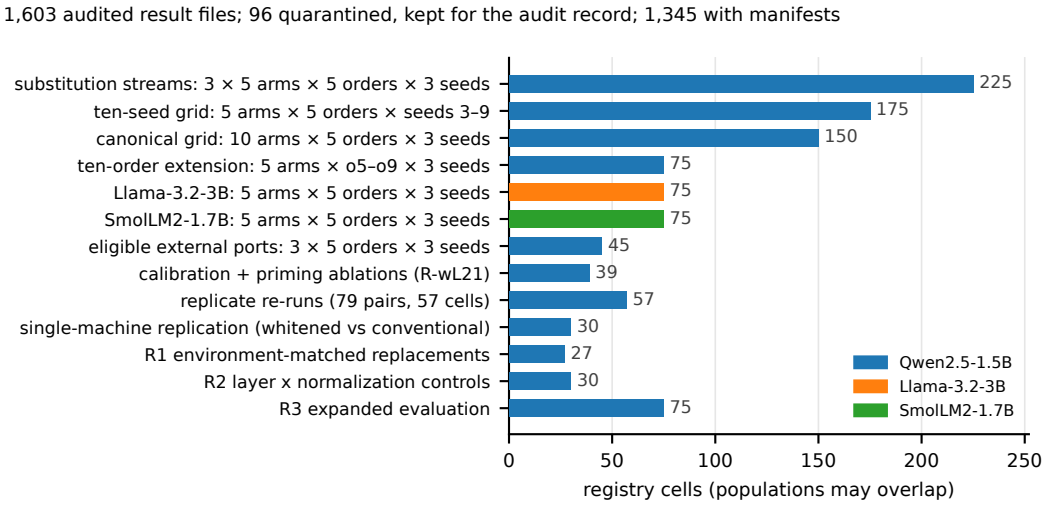
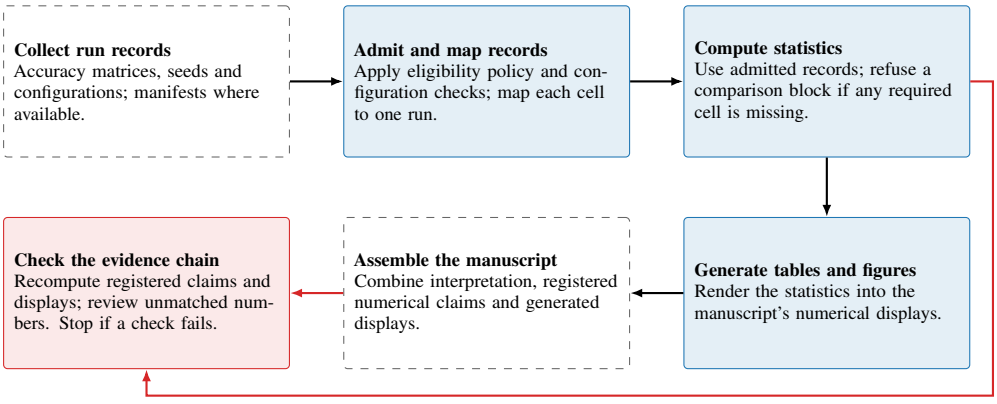


Figure 5: **The historical corpus at a glance.** Runs per population in the provenance registry, by backbone, each counted through the registry itself. The canonical grid and three eligible additional ports carry Sections 5.1–5.3 and Table 1; the ten-seed grid, Sections 5.4 and 4 and Figures 3b and 7; the ten-order extension, Section 5.1 and Figure 10b; the substitution streams, Section 5.4 and Table 9; the calibration, priming and single-machine sets, Section 5.3; the replicate re-runs, the discrepancies of Section 4; the two further backbones, Section 5.2. Smaller sets (the 3B backbone-integrity trainings, the skyline and oracle bounds, the 3B forgetting grid that is released but not analysed) are described where used. The frozen transfer cohort, scorer/host diagnostic and fixed-host study are reported in Appendices N, P and Q; all are outside this historical registry.



File key. Inputs: `result.json`, `MANIFEST.json`; registry: `scripts/provenance.py`.
Statistics: `compute_stats`, `noise_floor`, `power_analysis`, `decomposition`, `campaign_stats` → `stats/*.json`.
Displays: `make_paper_tables`, `make_new_figures`, `make_paper_figures` → `generated/*.tex`, `figures/*.pdf`.
Checks: `audit_manuscript_numbers_v2`, `ungated_numbers`, `preflight`.

Figure 6: **How evidence reaches the page.** The registry maps historical coordinates to records; a separate versioned eligibility policy determines the implementations admitted to each comparison. Statistics, tables and figures are generated from these records. Registered prose claims are checked against named statistics; generated artifacts are checked by recomputation, and a lexical scan identifies unmatched numerical tokens. Separate scientific checks target known interpretation errors and implementation defects. These checks support auditability, not universal semantic coverage or acceptance certification.

Consolidation (routed arm). The control arm here sits outside the ten-arm grid: the same routed configuration (whitened hidden-28 context, $K{=}8$, $k{=}2$, $r{=}8$) with the anchor disabled. Importance-weighted anchoring on R+cons ($\gamma{=}0.5$) is ahead of it in 13 of 15 matched cells, mean $+0.026$, paired $t{=}4.20$, with a stable sign in every order (Fig. 13). The gain is, however, below this arm's own replicate spread (0.047), so by the per-pair rule of Section 4 it is sign-consistent but smaller than the observed re-execution discrepancy; among routed arms in the historical cross-machine subset, this arm has the largest recorded discrepancy. The control's o3–o4 cells came from a later harness generation than its o0–o2 cells (Fig. 13). We removed that confound rather than only disclosing it: re-running the control on one machine in one generation (15/15 cells) reproduces the result slightly more strongly—ahead in 14 of 15 matched cells, mean $+0.029$, paired $t{=}4.12$—so this sensitivity assessment is unchanged. A γ sweep intended alongside it is withdrawn as a configuration artefact (Appendix E).

A.1 Dataset identity, prompts and scoring

Table 3 describes the loader in the retained source and the label vocabulary independently checked in every manifest of the complete frozen cohort. The runner requests the legacy Hugging Face identifiers shown; linked dataset cards identify their current repositories. Those live links are documentation, not evidence that an unpinned historical download had identical bytes. The frozen records preserve per-task hashes of sampled raw rows and rendered examples and their verbalizer lists. Most historical records do not preserve these identities, so we cannot retroactively certify their exact source revisions or row-level contents. In particular, our `amazon` task uses *binary polarity*, and the five-task recipe should not be read as an exact reproduction of O-LoRA's dataset protocol.

Exact label vocabulary. The loader uses the dataset's `ClassLabel` names; zero-based gold indices map to the following ordered lists. AG News: `World`, `Sports`, `Business`, `Sci/Tech`. Amazon: `negative`, `positive`. Yelp: `1 star`, `2 star`, `3 stars`, `4 stars`, `5 stars`. DBpedia: `Company`, `EducationalInstitution`, `Artist`, `Athlete`, `OfficeHolder`, `MeanOfTransportation`, `Building`, `NaturalPlace`, `Village`, `Animal`, `Plant`, `Album`, `Film`, `WrittenWork`. Yahoo: `Society & Culture`, `Science & Mathematics`, `Health`, `Education & Reference`, `Computers &`

Table 3: Dataset and prompt recipe. Each task uses the source `train`/`test` split; subsampling is uniform without replacement using the run seed. Fields appear in the listed order. Class names are preserved verbatim, as detailed below.

Task	Hugging Face identifier / classes	Input fields	Instruction
AG News	`ag_news` / four	`text`	Classify the news topic.
Amazon	`amazon_polarity` / two	`title`, `content`	Classify the review sentiment.
Yelp	`yelp_review_full` / five	`text`	Rate the review sentiment.
DBpedia	`dbpedia_14` / fourteen	`title`, `content`	Classify the topic.
Yahoo	`yahoo_answers_topics` / ten	`question_title`, `question_content`, `best_answer`	Classify the question topic.

`Internet`, `Sports`, `Business & Finance`, `Entertainment & Music`, `Family & Relationships`, `Politics & Government`. The runners prepend one ASCII space to each candidate before tokenization; the manifests record the unprefixed names.

Rendering and truncation. The prompt is the instruction, then one line `field: value` per nonempty input field, then `Answer:`, with newline separators and no list of candidate labels. Each field value is first clipped to its first 800 characters. Tokenization uses `add_special_tokens=False`. Training appends the spaced gold label and end-of-sequence (EOS) token, masks prompt tokens from the loss, and keeps the rightmost prompt tokens that fit the total length of 192. Evaluation appends each spaced candidate without EOS and retains the rightmost $192 -$ candidate-token-count prompt tokens. Thus candidates of different token lengths can receive different retained prompt tokens on long examples: the original scorer does not enforce identical prompt tokens across candidates. This is part of the recorded recipe, not a property ruled out by batching equivalence. The classification score is the mean log-likelihood over candidate tokens, and the largest score wins. Scoring masks right-padding. The relevant released implementation is `src/np_rlora/data/standard_cl.py` (loader and rendering), `scripts/pilot.py` (leading space and training examples), and `src/np_rlora/data/_common.py` (tokenization and scoring).

Table 44 quantifies candidate-dependent exposure on the preserved Qwen evaluation inputs.

Models. Qwen2.5-1.5B (28 blocks, d=1536) for the 150-run grid and the calibration ablation (the γ sweep is withdrawn, Appendix E); Qwen2.5-3B (36 blocks) for the backbone-integrity replicates. A 54-run Qwen2.5-3B forgetting grid (six arms $\times$ three seeds $\times$ orders o0–o2) is in the release and counted in the run census, but is not analysed here: it spans three orders with incomplete leading-task coverage, and only two of its arms (replay and sequential LoRA) are among the five core baselines compared elsewhere. fp32 throughout; scaled dot-product attention (SDPA).

Instrument specification. The three paragraphs below describe the routed case-study implementation; Section 3 summarises them.

Frozen backbone and expert bank. Each multilayer-perceptron (MLP) linear of a frozen Qwen2.5 model (Qwen Team, 2024) gets K low-rank experts: $y = W_0 x + \frac{\alpha}{r} \sum_{k \in \mathcal{T}(x)} a_k(x)\, B_k A_k x$ over the top-k set $\mathcal{T}(x)$; $K = 8$, $k = 2$, $r = 8$, $\alpha = 16$, MLP-only, approximately 56.4M expert parameters plus routing and gate heads (Table 4).

Routing context. Routing is sequence-level. The context $c(x)$ is the mean over prompt tokens of either the input embeddings (*embed*, the conventional choice) or the residual stream after block ℓ (*hidden-ℓ*). Because LLM hidden states are strongly anisotropic (Ethayarajh, 2019; Timkey & van Schijndel, 2021), the whitened variants standardize hidden context per dimension—standardised, i.e.

diagonal whitening—by running statistics primed on the first task and updated by an exponential moving average (EMA) during training—the step Section 5.3 dissects. For hidden context, every injected layer routes on the same layer-ℓ prompt summary from a separate no-gradient frozen-backbone pass with adapters bypassed. This adds one backbone forward per sequence. Embedding context uses an embedding lookup, and the raw-hidden arm disables whitening. Router logits add a cosine term against per-expert EMA prototypes, with a Switch-style balance loss (Fedus et al., 2022).

Consolidation. Optionally, per-expert importance accumulates a Fisher proxy across tasks and anchors experts to their post-task snapshots with weight γ (Kirkpatrick et al., 2017; Zenke et al., 2017). A further "plasticity gate" scaling expert-gradient magnitudes independently of routing weights is reported only as an arm (*routed + gate + consol.*); its behaviour is unresolved.

Baseline implementation audit. The historical runner never called the parameter-efficient fine-tuning (PEFT) library's `update_and_allocate`, across its recorded versions. The historical arm keyed `adalora` therefore trains its SVD parameterisation and orthogonality penalty without adaptive rank allocation; we label it SVD-LoRA throughout. Original records retain their keys. Omitting this arm leaves the within-corpus 0.02 criterion at five runs, with 14 of 30 single-run comparisons resolved (supplementary `power.json`, `without_fixed_svd`). Agreement analysis treats a difference within 10^{-12} as a tie, including one in either seed group. Historical revision coverage is incomplete, so we cannot certify every remote executable. The corrected runner enables allocation and records its final mask, but corrected runs are a separate cohort and cannot replace these observations. These historical results do not establish adaptive AdaLoRA's ranking.

Training. 800 steps per task, batch 4, AdamW lr 10^{-4}, max length 192, 1,000 train / 250 eval examples per task subsampled with the run seed. Routed arms: $K{=}8, k{=}2, r{=}8, \alpha{=}16$, MLP-only injection, $\lambda_{\text{bal}}{=}10^{-4}$; context from the input embedding (R-emb), the raw hidden-28 state (R-raw28) or the whitened hidden-21 (R-wL21) / hidden-28 (R+cons, R+g+c) state; consolidation $\gamma{=}0.5$ (both *R+cons* and *R+g+c*); all routed arms retain $\lambda_{\text{plast}}{=}1.0$, a gate learning-rate multiplier of 200 and gate temperature 1.0; only gated mode uses the gate in expert task-gradient scaling; whitening primed on 8 batches of the first task, EMA momentum 0.99. Baselines: LoRA $r{=}64$ (MLP linears, matching the routed injection sites); EWC $\gamma_{\text{ewc}}{=}0.25$; O-LoRA ($\lambda_{\text{orth}}{=}0.5$) and SVD-LoRA (rank 96; the configured target 64 and schedule were ineffective because the rank-allocation callback was never called), with MLP-only injection; O-LoRA uses rank 64 while SVD-LoRA actually retains rank 96, differing from the source settings; replay buffer fixed at 2% of the complete training stream (100 of 5,000 examples), not 2% of data seen so far. The nominal shared recipe fixes the tokeniser and seeded within-task data order; historical scorer revisions and incomplete executable identities are disclosed in Appendix K. Scoring: every class verbalizer of the task is appended to the prompt and scored by its length-normalised token log-likelihood (mean over the verbalizer's tokens); the highest-scoring class is the prediction; the total scoring context is capped at 192 tokens, with retained prompt length depending on the candidate verbalizer (Appendix A.1); batched scoring right-pads and masks the padding and agrees with unbatched scoring under the recorded equivalence checks. Manifests of non-replay arms record an inherited `replay_frac` argument that the code ignores for those methods.

Development ranking and variation details. Averaging stabilises rankings within the development corpus. Split the ten seeds into two disjoint sets of k and compare the resulting rankings (shared code, data and machines; the groups simulate repeated studies within this corpus, not independent laboratories or a controlled error rate). Ranking agreement climbs monotonically with k (Fig. 3b, right axis): at $k{=}1$ the two seed groups agree in 53% of draws, each seeing 2.5 reversals of 10 and both finding replay best in every order only 47% of the time. At $k{=}5$ they agree in 94%, each sees 1.1 reversals, both find replay best in every order in 100%, and pairwise sign agreement rises from 0.84 to 0.99 (Kendall τ_b $0.91 \rightarrow 0.99$). Decomposing the same 250 cells shows why seeds alone cannot rescue a single-order study: order, seed and residual account for 14–54%, 11–24% and 22–72% of each method's total sum of squares, respectively (Table 7; Appendix E repeats it on the six seeds that predate machine assignment, so the seed term cannot absorb a machine effect). These are additive-decomposition shares, not random-effect variance estimates. Averaging seeds does not test additional orders; "seed" names the training and evaluation subsamples and the initialisation together.

What the design rule is a rule about. k is chosen so that the *largest* unresolved gap falls below the stated size, so it is a worst case over the 50 pair-order comparisons and not a prediction for any one pair: resolvability tracks a pair's own replicate noise as well as its gap, and resolved and unresolved gaps overlap at k=1, 2, 3, 5. The smallest *resolved* gap and the largest *unresolved* one are 0.0419 vs. 0.0581 at k=1, 0.0294 vs. 0.0422 at k=2, 0.0294 vs. 0.0422 at k=3, 0.0294 vs. 0.0250 at k=4, and 0.0047 vs. 0.0174 at k=5—at five runs per cell a 0.0047 gap is signed while one about 3.7 times larger is not. Quoted as a worst case the rule is conservative: every gap above the stated size was signed at that k. That worst case is over pairs at a fixed criterion, and the $\geq 95\%$ agreement level is itself a choice the rule is a step function of: on the same exhaustive enumeration it reads two, three and four runs for 0.05, 0.03 and 0.02 at a 90% criterion, and three, four and five at a 97% one. Each entry turns on one or two comparisons whose agreement sits within 0.02 of the criterion—SVD-LoRA against sequential LoRA at o3 agrees in 95.56% of k=2 subset pairs, O-LoRA against EWC and against sequential LoRA at o0 in 93.33% at k=3, and O-LoRA against sequential LoRA at o2 in 94.29% at k=4. Read as a range, 0.05 takes two to three runs, 0.03 three to four and 0.02 four to five; the figures we quote are the conservative end of that range for 0.03 and 0.02, the permissive end for 0.05.

Evaluation sampling, measured with training fixed. Each canonical accuracy is a mean over 250 evaluation examples and so is quantised to 0.004; forgetting, a mean of four such differences, is quantised to 0.001. The earlier independent-binomial calculation gives a standard deviation of 0.016 per run (median over arms), ranging 0.014–0.017 across eligible arms and falling to 0.005 at ten runs; scaled to 7,600 test examples it is about 0.003. These are model-based calculations, not measured bounds. R3 now supplies the missing dependence information: from each completed training we subsample the same examples at every stage, preserving cross-stage correctness. Across 75 cells the median SD is 0.0119 (range 0.0073–0.0148), compared with 0.0158 under independent-stage resampling on the same records. The median cellwise SD ratio is 0.74 and median cross-stage correlation 0.455. Thus independence is conservative in this measured comparison, but its larger SD is not our empirical resolution estimate. These are conditional sampling distributions on the recorded evaluation pool; they neither establish training reproducibility nor provide a universal noise floor. Appendix K specifies the resampling.

Multiple comparisons. Nominal values in result tables are distinguished from adjusted conclusions. We define separate six-test families for the independent and block-preserving reversal nulls, a five-test family for per-method order effects, and a ten-test family for paired-difference order effects. Their adjusted values and covariance assumptions are generated in Appendix L. These analyses were specified retrospectively; they do not alter the primary analysis of the separately frozen prospective cohort.

Provenance audit. Figure 6 draws the path from historical run files to the historical analyses, and Figure 5 the run populations each section draws on. The 5 quarantine classes under `results/_quarantine` hold 96 runs: `2026-08-24_gamma_sweep_confounded` (24); `2026-08-24_generation_mix` (47); `nlora_L1_sum_bug` (14); `wcal_prime_embedctx_bug` (6); `fulltest_pod_record_pulled_midrun` (5 incomplete evaluation records, replaced by complete duplicates).

Result files, executions and cells. The frozen historical census counts `result.json` files: 1,603 outside quarantine, of which 13 are byte-identical to another, leaving 1,590 distinct by content. Of those 13, one is a certain copy (identical manifest and log, one run filed under a second name), two are certain re-runs (distinct manifests; on a fixed device a nominally identical re-run reproduces its forgetting value exactly, Section 4), and ten cannot be classified because one side carries no manifest (nine saved-model re-exports of static-multi runs and one 3B evaluation). Distinct executions therefore number between 1,590 and 1,602. Table 5 lists result files by directory, summing to the census; cells enter a number only through the registry, which admits one directory per (method, seed, order) cell.

Task orders. The five orders balance the leading position, not the others: Amazon is never last (so it enters every order's forgetting mean), DBpedia is last twice, and Yelp never holds position four; weighting every task equally in the mean changes no eligible arm's rank among the thirteen. o0 = ag_news, amazon, yelp, dbpedia, yahoo; o1 = yahoo, dbpedia, yelp, amazon, ag_news; o2 =

yelp, ag_news, yahoo, amazon, dbpedia; o3 = amazon, yelp, ag_news, yahoo, dbpedia; o4 = dbpedia, yahoo, amazon, ag_news, yelp. Extended orders (core baselines; fixed in writing before running): o5 = ag_news, dbpedia, amazon, yahoo, yelp; o6 = yahoo, amazon, ag_news, yelp, dbpedia; o7 = yelp, dbpedia, ag_news, amazon, yahoo; o8 = amazon, ag_news, dbpedia, yelp, yahoo; o9 = dbpedia, yelp, yahoo, ag_news, amazon.

Extended-population configurations. InfLoRA: r=64, old-task feature basis from the leading directions of each task's activations at energy threshold 0.95, capped at 256 retained columns (2048 feature rows/layer); N-LoRA (factor L_1): r=64, mean-normalised L_1 with λ=0.5; LoRAMoE-inspired port: 8 experts, r=8, balancing weight 10^{-2}. Among the three eligible additional ports InfLoRA has the lowest mean forgetting (0.034) and takes rank 1 from replay at o0. These are outcomes of the implemented variants under the shared recipe.

Operational archive boundary. The census and historical discrepancy summaries are fixed to the released `results/` tree and registry snapshot, excluding quarantine; they are not live totals of all evidence now available. The separate transfer cohort, scorer/host diagnostic, 60-trajectory fixed-host study and recovered inventory remain outside that census. Ongoing external-recipe work is not admitted to this manuscript. Each later remote snapshot has an explicit UTC cutoff recorded in `freshness/index.json`; jobs completing afterward fall outside that inventory. Appendix R accounts for every recovered payload, and Appendix S evaluates the alternative outcomes without replacing the frozen historical records.

Where the reimplementations depart from their sources. These three eligible arms are ports under our trainer, data and evaluation, not the authors' code, and we state the departures we know of rather than imply fidelity we have not verified. *N-LoRA*: our sparsity penalty is the mean-normalised L_1 of the adapter *factors*, $\|A\|_1 + \|B\|_1$, where Yang et al. (2025) penalise the *product* $\|BA\|_1$; the two are different objects (factor sparsity does not generally imply corresponding product sparsity), so this arm tests a factor-sparsity variant rather than the published method, and its rank and λ also differ from the values they report. *InfLoRA*: the retained basis is capped at 256 columns, so the energy threshold binds only until the cap does. *LoRAMoE-inspired port*: dense soft routing over eight rank-8 experts per targeted linear with a linear router, as in the source; the localized balancing constraint of Dou et al. (2024) is implemented as a coefficient-of-variation penalty on expert usage, and the expert-group / sample-type partition that constraint uses to separate world-knowledge from downstream experts is not implemented—this arm is a MoE-LoRA port of the published design, not the published LoRAMoE. Their rows should be read as our ports of these methods, not as the methods' published performance; the finite-corpus comparisons concern these executed variants and do not establish the resolution of source-faithful, optimally tuned implementations. Llama-3.2-3B (28 layers, hidden 3072) and SmolLM2-1.7B (24 layers, hidden 2048): identical hyperparameters to the 1.5B baseline arms; same MLP injection targets; five orders $\times$ three seeds $\times$ five core baselines each (75 cells per backbone, all complete). Priming ablation: whitening statistics primed on 8 batches of a *named* task (each non-first task in turn), stream otherwise unchanged. Twelve runs priming on each non-leading task reproduce the advantage (0.051–0.082 vs. 0.053 for AG News; pooled $+0.014$; Welch t=1.17, df=6.25, n.s.), though our recorded stop-condition held for only two of the four (Appendix E), which Section 5.3 reads as no evidence that the frame's identity matters.

Excluded prompt implementation. The Progressive Prompts port allocated five nonzero prompt blocks at initialization. Its evaluation helper inferred the learned prefix length from nonzero blocks, so every retained source version evaluated all five blocks from the first stage, including untrained future blocks. A CPU behavior check observes one through five blocks during successive training stages but five blocks at every evaluation. The port also used an oldest-to-newest, task-agnostic prefix; the source method uses a task-identity-dependent newest-to-oldest prefix (Razdaibiedina et al., 2023). Omitting its residual-MLP reparameterization is an additional difference. We cannot determine the numerical effect without new training or checkpoint reevaluation, and missing historical source identities prevent certifying every remote executable. We therefore exclude this port from comparison pools, not reinterpret its scores as evidence about the published method or convergence. The archived raw records, including the reported 0.156 discrepancy and 0.714 learning accuracy, remain available. A versioned manuscript eligibility policy applies this exclusion consistently without changing historical records or the frozen training registry.

Historical hardware and compute. 1.5B runs on RTX 6000 Ada (48 GB) and RTX PRO 6000 (96 GB); ≈2–3.5 h per run. Twenty-seven canonical-grid cells predate that pair and ran in a third, earlier environment (Python 3.11.10): the o0 column of all five routed arms, and o0–o2 for the two consolidation arms. No manifest we hold records its hostname or GPU model, so we identify it by interpreter version alone. 3B trainings ≈7–10 h plus ≈3.5 h of `lm-eval` per replicate. Estimated compute for the historical campaign, excluding the separately frozen cohort, ≈3,600 GPU-hours (core grids ≈1,400; extension campaign—four reproduced methods, orders o5–o9, seeds 3–5, priming ablation, 3B replicates, two further backbones, the ten-seed grid—≈1,400; the post-audit repair and replication campaign—re-running every quarantined configuration, the controlled replicate pairs, the repaired calibration ablation and the third backbone—≈800).

Parameter and replay semantics. Initial parameter reports precede training and do not measure final active capacity. We reconstruct stage-aware counts from retained source: O-LoRA and factor-penalty N-LoRA train one bank while all learned banks execute; InfLoRA also freezes input projections. By task five these ports activate five banks. These are derived mechanics, not recovered historical checkpoint measurements or an efficiency advantage. Replay's capacity is fixed at 100 examples, 2% of the complete 5,000-example stream; after the first task it is 10% of examples seen. Prior-task examples enter the reservoir at boundaries. Across the stream the recipe presents 16,000 current examples plus 12,800 replay examples, $1.8\times$ the current-only exposure. This ratio counts presentations, not tokens, FLOPs or time. Equal update counts therefore do not equalize compute.

Table 4: **Adaptation parameter counts** (millions), derived from the retained implementation including router and gate heads. Trainable counts describe a training task; stored and active-forward counts describe the five-task endpoint. Frozen earlier adapters remain active in progressive banks. Active use is not actual optimizer updating, FLOPs or memory. The backbone and non-parameter buffers are excluded. Supplementary CSVs T8–T9 retain every stage and original initial reports; historical executable identity is incomplete.

Arm	Trainable/ task	Stored, $t = 5$	Active/token, $t = 5$	Basis
O-LoRA port	56.4	282.1	282.1	derived
Fixed-rank SVD-LoRA	84.6	84.6	84.6	derived
Sequential LoRA	56.4	56.4	56.4	derived
EWC-style LoRA	56.4	56.4	56.4	derived
Replay	56.4	56.4	56.4	derived
InfLoRA port	34.9	282.1	282.1	derived
N-LoRA factor penalty	56.4	282.1	282.1	derived
LoRAMoE-inspired				
dense port	59.1	59.1	59.1	derived
R-emb	58.5	58.5	16.2	derived
R-wL21	58.5	58.5	16.2	derived
R-raw28	58.5	58.5	16.2	derived
R+cons	58.5	58.5	16.2	derived
R+g+c	58.5	58.5	16.2	derived

Table 5: **Historical result files by directory**, summing to the historical census. The role column is descriptive; the registry, not the directory, decides which run backs which cell, and development directories hold runs no number rests on.

Directory	Files	Role
`orderx`	255	historical ports (one excluded); further seeds and orders
`ablation`	122	routed arms' earlier cells (o0 column of the canonical grid) and development ablations
`seedext`	100	ten-seed grid: core five, seeds beyond the extension
`orderfix`	92	canonical grid: core five, and routed arms (later orders)

Historical result files by directory (continued)

Directory	Files	Role
llama3b	87	Llama-3.2-3B grid
scale3b	81	Qwen2.5-3B: forgetting grid (not analysed) and backbone-integrity evaluations
fulltest	75	review-response R3: core five at the expanded 7,600-example evaluation, per-example records
smollm2	75	SmolLM2-1.7B grid
subst_c	75	substitution control, Amazon to IMDB
subst_d	75	substitution stream, DBpedia to MNLI
subst_y	75	substitution stream, Yahoo to MNLI
phaseB	57	development phase B, incl. the oracle adapter bank
envuni	45	single-machine replication and priming ablation
long15	41	15-task stream (partial)
calib	39	inert-flag calibration ablation (disclosed, not used)
phaseC	36	development phase C, incl. the joint-training reference
confirm	33	consolidation control and confirmation runs
layernorm	30	review-response R2: layer x normalisation arms (hidden-21 unwhitened, hidden-28 whitened)
calibfix	27	repaired calibration ablation
envfix	27	review-response R1: the 27 earlier-environment canonical cells re-run on the current machines
envctl_cluster3	22	controlled replicate re-runs (cluster, device 3)
phase34	17	development phases 3-4
envctl_bw96	15	controlled replicate re-runs (workstation)
envctl_cluster2	15	controlled replicate re-runs (cluster, device 2)
ablfix	14	repaired ablation cells
envctl_cluster	13	controlled replicate re-runs (cluster)
phase5	9	development phase 5
scale7b	9	development / smoke runs
smsave	9	saved-model re-exports of phase-B static-multi runs (identical results, no manifests)
pilot	7	development / smoke runs
long15rep	6	development / smoke runs
local	3	development / smoke runs
retest	3	deterministic re-runs (identical results, distinct manifests)
(smoke and pilot directories, fewer than 3 files each)	14	development
audited result files	**1603**	distinct by content 1590; distinct executions 1590–1602

How far the best method sits from the reference models. On average accuracy, training all five tasks jointly reaches 0.8344 (3 runs), a bank of per-task adapters selected by an *oracle* task identity reaches 0.8319 (9 runs), and the best continual method in our study, replay, reaches 0.8185; these are empirical reference models, not ceilings. The gap is **0.016** to joint training and 0.013 to the oracle bank, against replay's own 0.019 of movement across the five orders and a largest historical eligible-arm spread on this metric of 0.064: **the gap is smaller than a discrepancy observed elsewhere in the historical corpus**, so the 0.016 margin is sensitive under this section's historical-discrepancy diagnostic. (Both reference arms ran in a third software environment, sign unknown, and the oracle bank ran at a smaller adapter rank, so neither gap is a bound in either direction.)

Archival backbone-integrity evaluation (unverified checkpoint linkage). Saved files record `lm-evaluation-harness` (Gao et al., 2024, v0.4.12), MMLU 5-shot, HellaSwag 10-shot and GSM8K 5-shot evaluations. Frozen-base scores were reused for recorded deltas. Four evaluation records are nominally associated by directory with `3b_ckpt_coupled_consol_v2` and `3b_ckpt_coupled_consol_rep1/2/3`; the former has a manifest, the latter do not. None of the

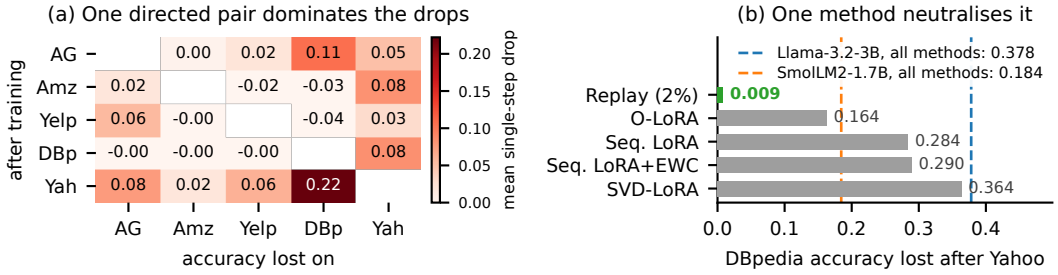


Figure 7: **One directed task pair dominates the measured drops.** (a) Mean single-step accuracy drop, attributed to the task just trained, over the 250-cell ten-seed grid; the Yahoo→DBpedia cell dominates and the matrix is strongly asymmetric. (b) That drop per method: replay alone neutralises it; dashed lines, the same quantity on the other two backbones.

evaluation records identifies its scored checkpoint by path or hash, and the supplementary archive does not contain those checkpoints. This block is resolved by explicit directory rather than the cell registry. Earlier loader omissions and subsequent repairs are part of the development audit, but the saved score differences cannot isolate their effects. We retain these evaluation records as unverified archival evidence, without using them for backbone-preservation or retraining-variance claims.

B Full task-pair investigation and negative predictions

Everything above treats forgetting as one scalar averaged over four tasks; opening it up (Fig. 7) shows where the forgetting lives. Over the 250-cell grid, per-task forgetting spans $12.7\times$—DBpedia 0.1699 and Yahoo 0.1689 are effectively tied at the top, against AG News 0.0968, Yelp 0.0606 and Amazon 0.0134—so the mean is not a summary of five comparable quantities. Attributing each single-step drop $R[i,j] - R[i{+}1,j]$ to the task trained at step $i{+}1$ localises it: **training Yahoo costs DBpedia** 0.222 ± 0.016 ($\pm$ s.e.; $n{=}100$, $t{=}13.9$; $t{=}18.0$ on 10 seed means; every arm $t \geq 5.4$; realisable only where DBpedia precedes Yahoo, in o0 and o4 of the five orders), $2.07\times$ the next-largest pair (DBpedia after AG News, 0.107) and $2.62\times$ its own reverse direction (0.085)—a directed property of a task pair, not a symmetric "similarity". Not an artefact of one backbone: the same pair is the largest single effect on Llama-3.2-3B (0.378, $t{=}7.6$) and SmolLM2-1.7B (0.184, $t{=}6.1$). Nor an adjacency artefact: where Yahoo comes later than the step after DBpedia (o5, o7, o8, o9), DBpedia loses 0.204 at the Yahoo step against 0.222 when adjacent, and its *final* forgetting—what the metric scores—is 0.204 with Yahoo adjacent, 0.269 with Yahoo later and 0.103 with Yahoo before it (the immediate drop recovers by 0.0185 by the end). The pair was picked as the largest of twenty after the fact; the two backbones and the ten-order streams are its confirmations.

Per method (Fig. 7b) a 2% buffer neutralises it, though whether the label spaces collide is a conjecture we did not test. The tempting reading is that the pair *causes* the order effect, so we pre-specified two removal streams and a control (Appendix E): the core five on the same orders with DBpedia, then Yahoo, replaced by the Multi-Genre Natural Language Inference (MNLI) task, predicting that the largest directed pair (P1) and replay's margin (P3) would halve, the order-effect F fall (P2), and the reversal count stay within its null (P4). **P1 and P2 fail on both; P3 fails on one and holds on the other; P4 holds on all three** (reversal counts inside their permutation nulls, $p{=}0.93$, 0.21 and 1.0). Removing either task of the pair leaves a new largest pair that falls short of halving against the matched three-seed standard stream (Table 9). Neither removes the order effect: the three-seed $F(4,8)$ rises for SVD-LoRA in both streams and for EWC without Yahoo—it strengthens. What the streams separate is *whose* advantage the pair explains: replay keeps 81% of its margin over the second-best method when the hurt task is removed, but only 16% when the trained task is. **Removing Yahoo sharply reduces replay's margin, but does not establish a pair-specific causal mechanism.**

Swapping Amazon for IMDB movie-review sentiment classification, which touches neither end of the pair, leaves Yahoo→DBpedia dominant and undiminished, and replay's margin intact where removing a task of the pair had cut it (Table 9). **The observed pair persists in the control substitution**: the control clauses of P1 and P3 hold where their removal clauses do not. Read together, the streams

place the tasks asymmetrically: Yahoo is the trained task of the largest pair with and without DBpedia, DBpedia the hurt task with and without Yahoo; replay's margin lives on the Yahoo side.

What survives the audit is replay: a 2% experience-replay buffer (Chaudhry et al., 2019) is best among the five core baselines in all fifteen order×backbone conditions and in all ten orders of the Qwen extension, mean forgetting 0.018. Its primacy is still order-dependent in the eligible 13-arm pool: at o0 InfLoRA beats it outright (0.015 vs. 0.031), and over these five orders replay's swing ($4.1\times$) is the widest of the core five in ratio terms though the narrowest in absolute forgetting—it is the most reliably *ranked* method, not the most order-stable. Consolidation, the one component of our own system sign-consistent across orders, is reported as sensitive to observed re-execution discrepancies in Appendix A.

C HISTORICAL RE-EXECUTION SENSITIVITY CRITERIA

The canonical grids have one training per method–seed–order cell, pooled over seeds as stated. In the frozen historical archive, we assess sensitivity to nominally matched re-execution using 79 eligible replicate comparisons—40 across machines, 27 across software environments (Section 5.3), 6 across devices of our six-GPU host and 6 repeated on one device—each a nominally identical method, seed, order and hyper-parameter set—matched on recorded flags; code revision, model revision and dataset hashes are not pinned per run (Appendix A). (Appendix I discloses three excluded sets: 18 pairs predating our manifest writer, the largest disagreeing by 0.182; nine inert-flag calibration triplets; and the 30 unattributable re-runs of Section 5.3.)

Two findings within that historical set. First, on a fixed GPU device forgetting repeated exactly: all six same-device re-runs reproduced their forgetting exactly, under different numbers of concurrent jobs (that host only). Second, nominally matched records disagree across execution histories, in a method-dependent way (the core five in Table 7; among eligible arms the largest is 0.047 across machines, for our routed consolidation, and 0.095 across software environments, for the conventional routed arm; per-pair values in Appendix I). The Progressive Prompts port is excluded on implementation grounds, independently of its learning accuracy (Appendix A.1). We use the remaining thirteen arms for comparison sensitivity and retain the excluded records for audit. Nothing in the configuration distinguishes these runs, and we did not trace where the divergence arises (on tooling nondeterminism see Zhuang et al., 2022); the observed maximum does not bound future discrepancies or define a minimum detectable effect.

We flag a margin as sensitive when it does not exceed the observed discrepancy a nominally identical configuration shows between environments—an empirical criterion, not a significance test; *which* spread is a choice, so we report both descriptive rules (Fig. 3a). A shared environment shift would cancel in a difference between arms; it does not: where arms were re-run in the same cell on both machines (76 arm pairs in five cells), the paired difference moves less than the noisier arm in 26% of pairs and more in 36%, reaching 0.063 overall and 0.028 among core-baseline pairs (median 0.0105). The **per-pair** rule asks the largest margin in each direction to beat $\max$ of the two arms' measured spreads: using the observed arm-specific discrepancies, although an arm measured at 0.000 over four observed pairs is not known to be that stable. The **global** rule asks it to beat the largest spread seen among eligible arms (0.095) in the frozen historical set: stricter on these data, but potentially overrestrictive for stable pairs. Neither rule controls a false-positive rate or certifies replication, especially with sparse, uneven replicate coverage. Of the three reversals among the five core baselines, **none exceeds either empirical criterion**: they turn on reverse-direction margins of 0.005, 0.004 and 0.0007, under even the 0.025 maximum over those five alone and the 0.028 by which their paired differences move between machines—so this does not rest on our own arms. Across the eight eligible baseline arms none of the ten survives the global rule and five survive the per-pair one; across all 78 pairs, 0 and 12 respectively, five of those 12 being pairs of external methods.

D TRAINING LOOP

Algorithm 1 Training the routed arms on a task stream; baseline implementations share the stated data and update schedule, with method-specific losses and replay exposure. $R[i, j]$ is accuracy on task T_j after stage i; $R[0, \cdot]$ is the pre-training row.

input frozen LLM; stream-ordered tasks $T_1, \ldots, T_5$ (columns of R follow this same order); seed s; per-task subsample of 1,000 train / 250 eval
build expert bank, with a router and plasticity gate at each injected linear (Fig. 11); if whitening enabled: prime whitening stats (μ, σ^2) on 8 batches of T_1
for $j = 1..5$ **do**
  evaluate all T_i (label-verbalizer log-likelihood) $\rightarrow$ row $R[j-1, \cdot]$
  for 800 steps on T_j, batch 4, AdamW 10^{-4} **do**
    $c \leftarrow$ pooled context; **if** whitening enabled: EMA-update (μ, σ^2), $c \leftarrow (c - \mu)/\sigma$
    per injected linear: $a \leftarrow$ top-k softmax of $W_r c + \cos(c, \mu_\cdot)$; forward with $\frac{\alpha}{r} \sum_k a_k B_k A_k x$; compute gate p and update prototypes μ_k
    $\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_{\text{bal}} \mathcal{L}_{\text{bal}} + \lambda_{\text{plast}} \mathcal{L}_{\text{plast}} + \frac{\gamma}{2} \sum_k \Omega_k \|\theta_k - \theta_k^\star\|^2$
    use coupled expert task gradients, or gate-scaled task gradients in gated mode; step; update gradient EMA
  end for
  update accumulated importance and snapshot $\theta^\star$; anchoring is disabled when $\gamma = 0$
end for
evaluate all $T_i \rightarrow R[5, \cdot]$; **return** forgetting $= \frac{1}{4} \sum_{i<5} \left(\max_{1 \le j < 5} R[j, i] - R[5, i] \right)$ (signed; the max excludes the pre-training row and the final stage)

E FIXED ANALYSIS PLAN AND DEVIATIONS

Before running orders o3–o4 we recorded in the repository (dated commit; not externally time-stamped, so we do not call this a pre-registration): the exact five orders; the ten arms; that the primary statistic would be the count of method pairs whose order-mean winner changes; that whitened-vs-conventional routing was the headline pair; and that a $\geq$ 5-order paired $t<2.14$ ($p>0.05$) or any order with ratio < 1 would be reported as "effect not general". On the measurements available at the time, both conditions fired—and the study pivoted accordingly. The subsequent provenance audit (Appendices A and T) then showed those measurements were contaminated: nine of the R-wL21 cells had silently run the R-emb (conventional-routing) configuration, and on the repaired cross-machine grid neither condition fires (t=7.00; minimum ratio 1.09), while on the single-machine replication that Section 5.3 reports the ratio condition does fire (o4: $0.76\times$) and the t condition does not (t=4.56). We report this sequence exactly as it happened: the pivot was made in good faith on defective data, and the repaired data justify it on partly different grounds—the observed gain varies by order, and the benchmark's resolution (Section 4) cannot support the ranking claims the original method paper would have made. Two deviations: (i) the calibration ablation (Section 5.3) was designed *after* seeing the o3–o4 collapse, to explain it, and is labelled as such; (ii) the 3B replicate study was added after two archived evaluation records nominally associated with 3B runs from the crash-recovery period were found to disagree by 0.180 in ΔGSM8K.

Second fixed plan (extension campaign). Before any extension cell ran, the following were recorded in the repository (dated commit): the five additional orders o5–o9 verbatim (chosen so each task leads a second time, minimising adjacency overlap with o0–o4); the primary statistic (pairwise winner-change count on order means, recomputed per order-subset size); and the priming-ablation stop-condition—if o0 forgetting under non-AG priming stayed within one standard deviation of the AG-primed value, "AG News specifically" would be reported as refuted. The stop-condition as written was met by two of four primings; the additional, post-hoc analysis finds all four below the frozen variants and does not detect a difference from the AG-primed arm; this is not evidence of equivalence (Welch t=1.17, df=6.25, n.s.). Deviations from this plan: the six-seed extension and the Llama-3.2-3B replication were added mid-campaign in response to reviewer-style critique rather than pre-specified; the second backbone was switched from an initially staged SmolLM2-1.7B to Llama-3.2-3B when licence access was granted (at the time of that decision no SmolLM2 result

existed); SmolLM2-1.7B was subsequently run to completion as a *third* backbone, and because that decision post-dates the plan we report its full $5\times5\times3$ grid (Table 11) as a confirmatory replication rather than a pre-specified test. The ten-seed extension of the five core baselines (250 cells), the task-pair decomposition of Section 5.4, the headroom comparison and the power analysis of Section 4 were likewise added after this plan and are reported as post-hoc analyses of a fixed corpus; none of them changes an arm, an order or a hyperparameter.

Third fixed plan (substitution campaign). Before any cell of the 2026-08-30 campaign ran we recorded in the repository (dated commit; as above, not externally time-stamped): a 15-task stream on five orders L0–L4 to the same reporting standard, and three substituted streams, each the standard o0–o4 with exactly one task swapped—DBpedia→MNLI, Yahoo→MNLI, and a control Amazon→IMDB—on the core five, seeds 0–2; and four predictions: (P1) the largest directed pair falls by more than half in the two removal streams and not in the control, (P2) the order-effect F for SVD-LoRA, EWC and sequential LoRA falls likewise, (P3) replay's margin over the second-best method falls by more than half likewise, (P4) the reversal count stays within its permutation null in all three. Deviations, all logged before the affected cells ran: the machine split was revised once for infrastructure reasons; a stale mirror on the second machine failed at import and its cells were retried; the cluster path moved to the scheduler after a policy kill; one non-campaign cell was lost and re-queued. All three streams completed. P1 and P2 fail on both removal streams; P3 fails without DBpedia, where replay keeps 81% of its margin over the second-best method, and holds without Yahoo (16%). Appendix B quotes three of P2's six pre-specified cells; the other three are sequential LoRA $5.39 \to 19.26$ without Yahoo and $5.39 \to 7.87$ without DBpedia, and EWC $6.15 \to 5.63$ without DBpedia—the one cell of the six that falls, and by 8.5% (p=0.019) against a predicted halving. The control clauses of P1 and P3 hold: with a task swapped that touches neither end of the pair, the pair is undiminished and replay's margin does not fall. P2's control clause does not hold for all three arms it names: in the control the order-effect F is $6.25 \to 12.79$ for SVD-LoRA, $6.15 \to 3.02$ for EWC—a fall by more than half, the movement P2 predicted for the removal streams alone—and $5.39 \to 4.33$ for sequential LoRA. Substituting a task moves these F values in the control as well, so we do not read the increases reported in Section 5.4 as specific to removing a task of the pair. P4 holds: the reversal count is 3 of 10 without DBpedia, 5 of 10 without Yahoo and 1 of 10 in the control, against 3 of 10 on the matched standard stream (Section 5.4), and each sits inside its permutation null (p=0.93, 0.21 and 1.0; 0.85, 0.24 and 1.0 under the block-preserving null). This verdict was computed after an external review of 2026-09-11 noted that the paper had left P4 untested; nothing else in the plan changed. The γ sweep intended alongside the consolidation result (Appendix A) was withdrawn when its cells failed the configuration-signature audit (γ aliased with the routing-context setting on the routed leg; the EWC leg split across two adapter ranks).

Fourth fixed plan (review-response campaign, 2026-09-11). Recorded in the repository (dated commit) before any cell ran, in response to an external review: (R1) the 27 canonical-grid cells produced in the earlier environment—the o0 column of all five routed arms and o1–o2 of the two consolidation arms, seeds 0–2—re-run with identical flags on the current machines, to give an environment-matched canonical grid; (R2) two arms completing the layer×normalisation 2×2 at fixed layer (hidden-21 unwhitened, hidden-28 whitened without consolidation), o0–o4 × seeds 0–2; (R3) the core five on o0–o4 × seeds 0–2 with the planned 7,600-example evaluation and a per-example correctness record at every stage, so the sampling distribution of forgetting can be resampled jointly across stages. No prediction is registered for R1–R3 beyond "the same reporting standard applies". R1 completed on 2026-09-12. The environment-matched grid puts the conventional-over-whitened ratio at $3.22\times$ at o0 rather than the canonical grid's $4.06\times$, moves the ten-arm reversal count to 18 from 19 and its block-preserving null to p=0.04, and changes the rank of four of the thirteen eligible arms; the 27 (original, re-run) pairs enter the replicate population as its cross-environment members (Section 4). On the six seeds that predate machine assignment, order and seed account for 20–55% and 8–20% of each core method's total sum of squares, respectively. R2 completed on 2026-09-13: whitening at fixed layer is $2.11\times$ (L21) and $2.04\times$ (L28), 2.03–$2.04\times$ on the environment-matched grid; the layer effect at fixed normalisation is 1.12–$1.16\times$ with $t \leq 1.45$; the superseded three-order factor comparison (embeddings to raw hidden-28 $1.06\times$, that to whitened hidden-21 $3.34\times$) confounded the two factors, and its conclusion survives decomposition. R3 completed: all 75 runs have complete per-example records. Appendix K reports its results, joint resampling and scorer checks. Five cells whose evaluation records were copied before completion were replaced by complete duplicates, chosen on record completeness before computing R3 statistics; the original records remain quarantined.

Table 6: **Forgetting on the standard 5-task stream across five task orders** (mean ± sample SD over 3 seeds). Bold marks the lowest mean in each task-order column. *Swing* = max/min of unrounded order means; *Range* = max − min, since a ratio flatters arms whose forgetting is near zero. The original ten-arm grid (core five plus our routed arms) is plotted in Fig. 8.

Method	o0	o1	o2	o3	o4	Mean	Swing	Range
Replay (2%)	0.031 ±0.01	**0.037** ±0.01	**0.015** ±0.01	**0.009** ±0.01	**0.019** ±0.01	0.022	4.1×	0.028
InfLoRA port	**0.015** ±0.00	0.065 ±0.03	0.017 ±0.01	0.040 ±0.05	0.035 ±0.02	0.034	4.3×	0.050
O-LoRA	0.105 ±0.06	0.082 ±0.01	0.048 ±0.01	0.084 ±0.02	0.085 ±0.03	0.081	2.2×	0.058
LoRAMoE-inspired port	0.129 ±0.06	0.086 ±0.01	0.045 ±0.02	0.070 ±0.02	0.171 ±0.04	0.100	3.8×	0.126
Seq. LoRA	0.116 ±0.00	0.110 ±0.02	0.087 ±0.02	0.080 ±0.03	0.147 ±0.01	0.108	1.8×	0.067
Seq. LoRA + EWC	0.121 ±0.01	0.112 ±0.02	0.086 ±0.02	0.080 ±0.02	0.151 ±0.01	0.110	1.9×	0.071
N-LoRA (factor L_1)	0.156 ±0.01	0.087 ±0.01	0.059 ±0.01	0.117 ±0.05	0.140 ±0.02	0.112	2.6×	0.097
Fixed-rank SVD-LoRA	0.249 ±0.02	0.191 ±0.03	0.172 ±0.01	0.161 ±0.01	0.214 ±0.03	0.197	1.5×	0.087

Training flags were unchanged, but these are new trainings, so the comparison with the canonical grid combines training and evaluation variation.

F FULL TABLES

Backbone integrity on a generative metric. Following TRACE (Wang et al., 2023b), our evaluation archive contains MMLU (Hendrycks et al., 2021), HellaSwag (Zellers et al., 2019) and GSM8K (Cobbe et al., 2021) scores (Fig. 12). Four records nominally associated with R+cons on Qwen2.5-3B give ΔGSM8K $\{-0.027, -0.053, -0.695, -0.707\}$, s.d. 0.382 and range 0.68; their MMLU and HellaSwag delta standard deviations are 0.0075 and 0.0039. The sequential-LoRA directory group spans $[-0.69, -0.19]$ on GSM8K, while the archived MMLU/HellaSwag deltas lie within $|\Delta| \leq 0.022$. These are descriptive properties of the saved evaluation files. Missing checkpoint pointers prevent verifying their training attribution, so we do not use them to estimate retraining variability or support a backbone-preservation claim. An unverified qualitative probe is excluded from evidence (Appendix H).

A worked instance of a core-baseline reversal. The closest inverting pair among the core baselines is sequential LoRA vs. sequential LoRA + EWC (Table 14). On the canonical grid, plain sequential LoRA wins on four of the five orders and EWC wins on one (o2)—by 0.0007, against a replicate spread of 0.019–0.025 measured on these very arms (Section 4). This is what a "reversal" typically looks like on this benchmark: not a dramatic swap but a close comparison sensitive to the chosen order and recorded execution, without establishing equivalence or identifying a causal GPU effect. Per-order standard deviations for Table 6, average and learning accuracy, all 195 eligible per-cell values of the thirteen eligible arms, and the 78-pair reversal table (flagged for the ten-arm and eight-external subsets) are provided as CSV files in the supplementary material (`T1_five_order_grid.csv`, `T1b_five_order_all_cells.csv`, `T2_pairwise_inversions.csv`), regenerated by the released scripts; the CSVs call a reversal an inversion, and `T1_five_order_grid.csv` also carries average and learning accuracy.

Table 7: **Order sensitivity and observed replicate discrepancies by method.** Order-swing: max/min of the order means over five orders (3 seeds) on each backbone and over ten orders on Qwen; the per-order values the ratios summarise are in Table 6, in absolute terms. Order effect: two-way additive ANOVA over the 250-cell ten-seed grid, with order and seed as factors. Order/seed/residual shares divide each term's sum of squares by the total for the method; the displayed p is nominal. Appendix L gives Holm correction and covariance sensitivity: the four original-model rejections become three under GG plus Holm. Replicate spread is the largest observed difference between nominally matched eligible records in the frozen historical set (Section 4).

Order-swing

Method	Qwen 5 orders	Qwen 10 orders	Llama 5 orders	SmolLM2 5 orders
Replay (2%)	4.1	4.1	4.6	2.1
O-LoRA	2.2	6.5	1.9	2.0
Seq. LoRA	1.8	3.5	2.0	2.0
Seq. LoRA + EWC	1.9	3.4	2.0	2.1
Fixed-rank SVD-LoRA	1.5	2.2	3.1	1.5

Order effect and replicate spread

Method	$F(4, 36)$	p	Variance share (%) order/seed/resid.	Replicate spread
Replay (2%)	3.30	0.021	24 / 11 / 65	0.003
O-LoRA	1.79	0.15	14 / 14 / 72	0.002
Seq. LoRA	9.21	3.1×10^{-5}	39 / 24 / 38	0.025
Seq. LoRA + EWC	11.46	4.2×10^{-6}	44 / 21 / 35	0.019
Fixed-rank SVD-LoRA	22.24	2.6×10^{-9}	54 / 24 / 22	0.000

Table 8: **Order moves the differences between methods, not only their levels.** For each pair of core baselines on the 250-cell ten-seed grid, the two-way (order×seed) ANOVA of the paired difference $a - b$: mean difference, $F(4, 36)$ and upper-tail p; 7 of ten reject nominally, 6 after Holm, and 4 after GG plus Holm (Appendix L). The pooled method×order interaction over all five is $F(16, 216)$=4.27, p=3×10^{-7}. The displayed $p = 0.050$ is rounded from a value below 0.05.

Pair	Mean $a - b$	$F(4, 36)$	p
Fixed-rank SVD-LoRA vs. Replay (2%)	+0.151	10.78	7.5×10^{-6}
Seq. LoRA + EWC vs. Replay (2%)	+0.082	9.85	1.7×10^{-5}
Seq. LoRA vs. Replay (2%)	+0.080	8.25	7.9×10^{-5}
Fixed-rank O-LoRA vs. SVD-LoRA	-0.103	6.99	2.8×10^{-4}
Fixed-rank SVD-LoRA vs. Seq. LoRA	+0.072	5.52	0.001
Fixed-rank SVD-LoRA vs. Seq. LoRA + EWC	+0.070	5.04	0.003
Seq. LoRA vs. Seq. LoRA + EWC	-0.003	2.64	0.050
O-LoRA vs. Seq. LoRA + EWC	-0.034	2.39	0.069
O-LoRA vs. Seq. LoRA	-0.031	2.02	0.112
O-LoRA vs. Replay (2%)	+0.048	1.99	0.117

Table 9: **Replacing either task does not remove order effects; replay's margin shrinks more when Yahoo is replaced.** Core five, orders o0–o4, seeds 0–2, one task swapped per stream; last row, the control, touching neither task of the pair. $F(4, 8)$: order×seed ANOVA for the three arms P2 names; replay's margin over the second-best method, as a percentage of standard. Verdicts on P1–P4 (Appendix E): ✓ holds, × fails, – untested; the control row scores their control clauses.

Stream	Largest directed pair	Single-step accuracy drop	t	vs. std.
Standard	Yahoo→DBpedia	0.2133	8.03	—
DBpedia→MNLI	AG News→Yahoo	0.1236	7.78	-42%
Yahoo→MNLI	Amazon→DBpedia	0.1613	3.90	-24%
Amazon→IMDB	Yahoo→DBpedia	0.2243	7.45	$+5\%$

Order effect, $F(4, 8)$

Stream	Fixed-rank SVD-LoRA	EWC	Seq. LoRA
Standard	6.25	6.15	5.39
DBpedia→MNLI	30.97	5.63	7.87
Yahoo→MNLI	11.61	23.70	19.26
Amazon→IMDB	12.79	3.02	4.33

Replay margin and prediction verdicts

Stream	Replay margin remaining	Reversals of 10	P1/P2/P3/P4
Standard	—	3	—
DBpedia→MNLI	81%	3	×/×/×/✓
Yahoo→MNLI	16%	5	×/×/✓/✓
Amazon→IMDB	132%	1	✓/×/✓/✓

Table 10: **Llama-3.2-3B, same five core baselines, orders, seeds and hyperparameters** (forgetting, mean ± sample SD over 3 seeds; lower is better). Bold marks each task-order column's lowest mean. Swing and Range are the ratio and difference of the largest and smallest order means. The reversal rate is comparable to Qwen (4/10 pairs vs. 3/10; 5/10 within-seed), and the aggregate five-method ranking is identical on the two backbones.

Method	o0	o1	o2	o3	o4	Mean	Swing	Range
Replay (2%)	**0.031** ±0.01	**0.037** ±0.01	**0.020** ±0.01	**0.008** ±0.01	**0.018** ±0.00	0.023	4.6×	0.029
O-LoRA	0.077 ±0.01	0.074 ±0.01	0.060 ±0.01	0.076 ±0.03	0.115 ±0.01	0.080	1.9×	0.055
Seq. LoRA	0.166 ±0.03	0.112 ±0.02	0.083 ±0.01	0.083 ±0.02	0.129 ±0.02	0.115	2.0×	0.083
Seq. LoRA + EWC	0.182 ±0.07	0.108 ±0.05	0.090 ±0.03	0.130 ±0.02	0.120 ±0.01	0.126	2.0×	0.092
Fixed-rank SVD-LoRA	0.340 ±0.02	0.169 ±0.06	0.172 ±0.01	0.146 ±0.02	0.109 ±0.02	0.187	3.1×	0.230

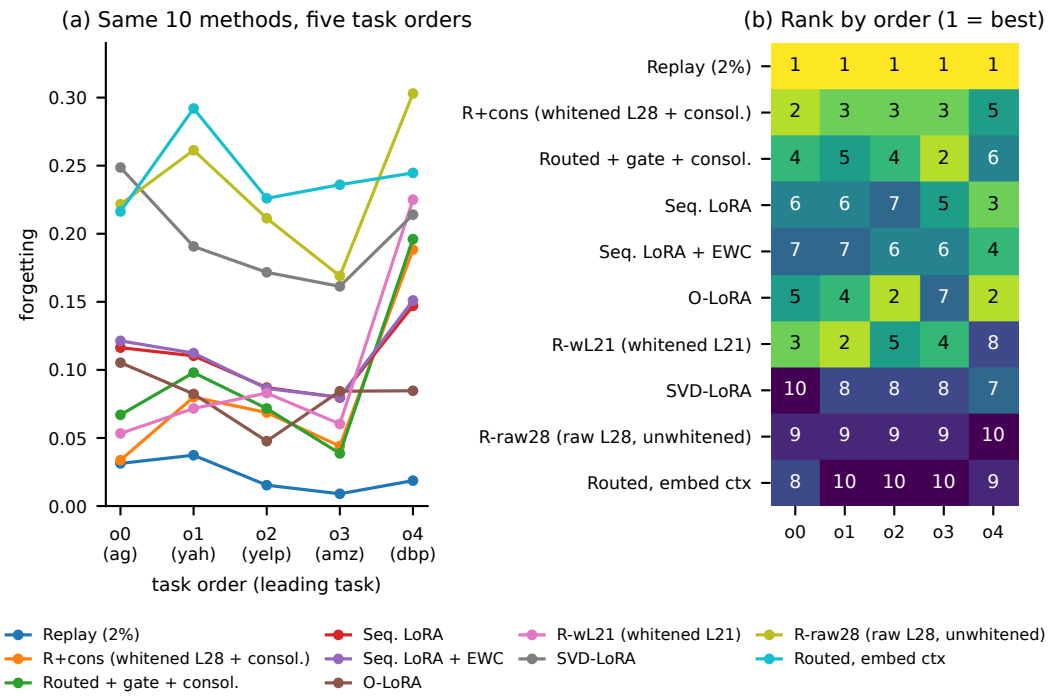


Figure 8: **The same ten methods, five task orders, Qwen2.5-1.5B, 3 seeds per point.** (a) Forgetting by order. (b) Rank of each method within each order. O-LoRA moves between 2nd (o4) and 7th (o3); whitened routing between 2nd (o1) and 8th (o4). Replay is 1st in every order; Section 4 shows which of the other rank changes exceed replicate noise (few do).

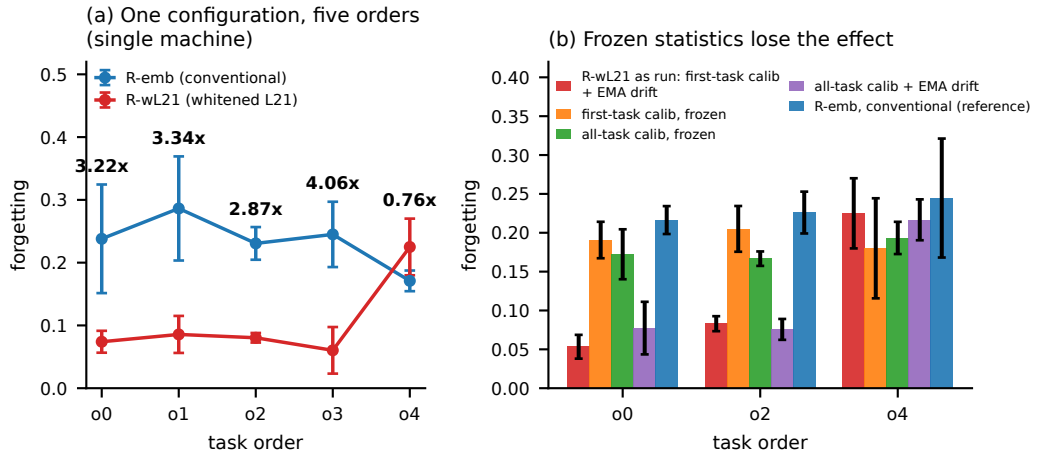


Figure 9: **Observed routing differences vary across task orders.** (a) Whitened hidden-state routing vs. conventional embedding routing across the five orders (3 seeds, ±s.d., both arms re-run on one machine); the ratio above each order is conventional/whitened. The effect is 3.22, 3.34, 2.87, 4.06× on o0–o3 and 0.76× on o4. A paper on o0–o2 reports 3.15× at t=9.63; the fifth order alone would report the method as harmful. (b) Calibration ablation on o0/o2/o4 (3 seeds each, repaired runs; the two reference series are the canonical cross-machine grid, not panel (a)'s single-machine replication, which is why the two arms sit in the opposite order at o4): on o0 and o2, freezing the whitening statistics returns forgetting to conventional level whichever tasks the frame is estimated from; only drift reproduces the effect; at o4 no variant separates from the conventional reference.

Table 11: **SmolLM2-1.7B, same protocol again** (forgetting, mean ± sample SD over 3 seeds; lower is better). Bold marks each task-order column's lowest mean. Swing and Range are the ratio and difference of the largest and smallest order means. Three of ten pairs reverse across orders; replay is best in every order; the mean ranking matches the other two backbones but for EWC and sequential LoRA, separated by 0.0007. Across three vendors and 1.5–3B parameters, changing the backbone moved one adjacent pair by 0.0007; changing the order flipped 3–4 pairs of 10 on every one of them, the Qwen reversals do not exceed the observed re-execution discrepancy criteria.

Method	o0	o1	o2	o3	o4	Mean	Swing	Range
Replay (2%)	**0.015** ±0.02	**0.030** ±0.01	**0.021** ±0.01	**0.021** ±0.02	**0.019** ±0.00	0.021	2.1×	0.016
O-LoRA	0.132 ±0.04	0.092 ±0.04	0.066 ±0.02	0.077 ±0.02	0.089 ±0.04	0.091	2.0×	0.066
Seq. LoRA + EWC	0.143 ±0.03	0.108 ±0.03	0.148 ±0.03	0.179 ±0.06	0.086 ±0.02	0.133	2.1×	0.092
Seq. LoRA	0.144 ±0.03	0.107 ±0.03	0.150 ±0.03	0.179 ±0.06	0.087 ±0.02	0.134	2.0×	0.091
Fixed-rank SVD-LoRA	0.256 ±0.04	0.223 ±0.01	0.221 ±0.01	0.217 ±0.04	0.166 ±0.03	0.217	1.5×	0.089

Table 12: **Calibration ablation** (forgetting, mean ± s.d. over 3 seeds), the three calibration variants are computed only from the repaired runs in `results/calibfix` whose variant-selecting flags are verified live; the as-run row is the canonical-grid R-wL21 arm, whose o0 cells were produced in the earlier software environment of Appendix A while the three variants were re-run on the 96 GB host, so this contrast spans environments as well as variants. Freezing the whitening statistics returns forgetting to conventional-baseline level on o0 and o2; at o4 no variant separates from the conventional reference, whether the frame is estimated on the first task or on all of them; letting them drift recovers the effect.

variant	o0	o2	o4
R-wL21 as run (first-task calib + EMA drift)	0.053±0.02	0.083±0.01	0.225±0.05
R-emb, conventional (reference)	0.216±0.02	0.226±0.03	0.245±0.08
first-task calib, frozen	0.191±0.02	0.205±0.03	0.180±0.06
all-task calib, frozen	0.172±0.03	0.167±0.01	0.193±0.02
all-task calib + EMA drift	0.077±0.03	0.076±0.01	0.217±0.03

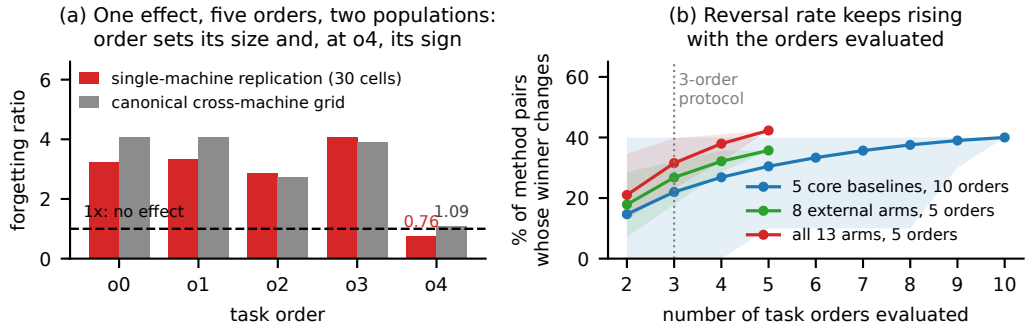


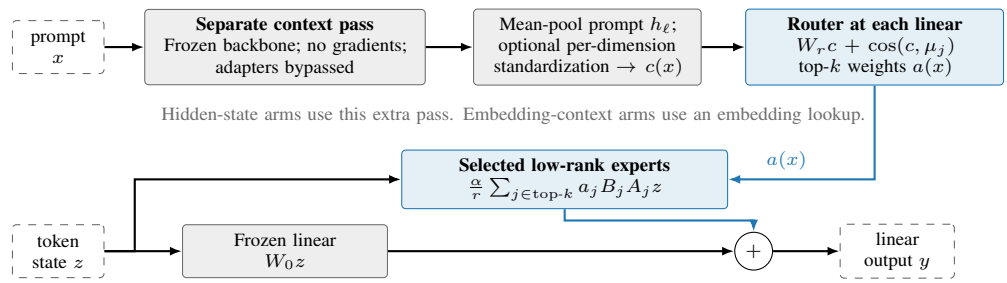
Figure 10: **What task order controls.** (a) Our own routing effect (conventional over whitened forgetting) per order, on the 30-cell single-machine replication of Section 5.3 and on the canonical cross-machine grid: order sets its size on both and, at o4, its sign on one. (b) Fraction of method pairs whose winner changes, over all $\binom{n}{k}$ order subsets (mean and min–max band): five core baselines over ten orders, eight eligible baseline arms and all thirteen eligible arms over five. The rate still rises at ten orders. Dotted line: the three-order protocol; on o0–o2 the external-arm rate is 25%. Under-seeded grids overstate order-instability for close pairs while order-means understate what a single run exhibits; both axes must be reported (Section 6).

Table 13: **The five routed case-study arms**, Mean ± sample SD of forgetting over 3 seeds per order, as in Table 6. Bold marks each task-order column's lowest mean. Their o0 column was produced in the earlier software environment of Appendix A; for R+cons and R-wL21 the swing minimum falls in that column. They are reported separately because they are ours and, as the swings show, among the most order-sensitive in the pool. This is the canonical cross-machine grid: on it the whitened arm stays ahead of the conventional one at every order, including o4 (paired $t(14)$=7.0 over 15 seed–order differences, 14/15 cells; minimum ratio 1.09). The sign reversal of Section 5.3 is a property of the machine-controlled replication reported there, not of this grid—which is the paper's own point about which population a claim is made on.

Method	o0	o1	o2	o3	o4	Mean	Swing	Range
R+cons (whitened L28 + consol.)	**0.034** ±0.01	0.080 ±0.02	**0.069** ±0.03	0.044 ±0.01	**0.188** ±0.03	0.083	5.6×	0.155
R+g+c (gate + consol.)	0.067 ±0.03	0.098 ±0.04	0.072 ±0.01	**0.039** ±0.01	0.196 ±0.03	0.094	5.1×	0.157
R-wL21 (whitened L21)	0.053 ±0.02	**0.072** ±0.01	0.083 ±0.01	0.060 ±0.04	0.225 ±0.05	0.099	4.2×	0.172
R-raw28 (raw L28, unwhitened)	0.222 ±0.02	0.261 ±0.05	0.211 ±0.05	0.169 ±0.00	0.303 ±0.08	0.233	1.8×	0.134
R-emb (embed ctx)	0.216 ±0.02	0.292 ±0.08	0.226 ±0.03	0.236 ±0.01	0.245 ±0.08	0.243	1.3×	0.076

Table 14: Sequential LoRA / sequential LoRA + EWC, forgetting per (seed, order) cell. Bold marks the lower forgetting of the pair, decided on the unrounded values, so where the two sides print alike the bolded one leads by less than the 0.001 displayed; exact ties are left unbolded.

	o0	o1	o2	o3	o4
seed 0	**0.111** / 0.114	**0.122** / 0.128	0.101 / 0.101	**0.049** / 0.054	**0.144** / 0.148
seed 1	0.120 / 0.120	**0.124** / 0.125	0.067 / **0.066**	0.092 / **0.091**	**0.159** / 0.163
seed 2	**0.118** / 0.130	0.085 / **0.084**	0.093 / **0.092**	0.098 / **0.095**	**0.138** / 0.142
mean	**0.116** / 0.121	**0.110** / 0.112	0.087 / **0.086**	**0.080** / 0.080	**0.147** / 0.151



Separate adapted forward pass. The full frozen backbone executes; the lower block is inserted at each MLP linear. Blue boxes contain trainable router/expert parameters.
Optional mechanisms, omitted here: the plasticity gate scales expert task gradients; consolidation adds an importance-weighted anchor. Router, expert and gate heads total 58.5M trainable parameters (Table 4).

Figure 11: **The routed case-study instrument.** For hidden-state arms, a separate frozen pass with adapters bypassed extracts the prompt context at layer ℓ. Embedding-context arms use an embedding lookup. After optional per-dimension standardization, this context feeds a prototype-augmented top-k router at every injected MLP linear in the adapted pass. The dense frozen backbone executes in full. Blue boxes contain trainable router/expert parameters; the optional gate and consolidation terms are omitted from the diagram. Variants used in the study differ in the context source (embedding vs. hidden layer), the tap layer ℓ (21 or 28), whether c is whitened, and whether consolidation and the plasticity gate are on.

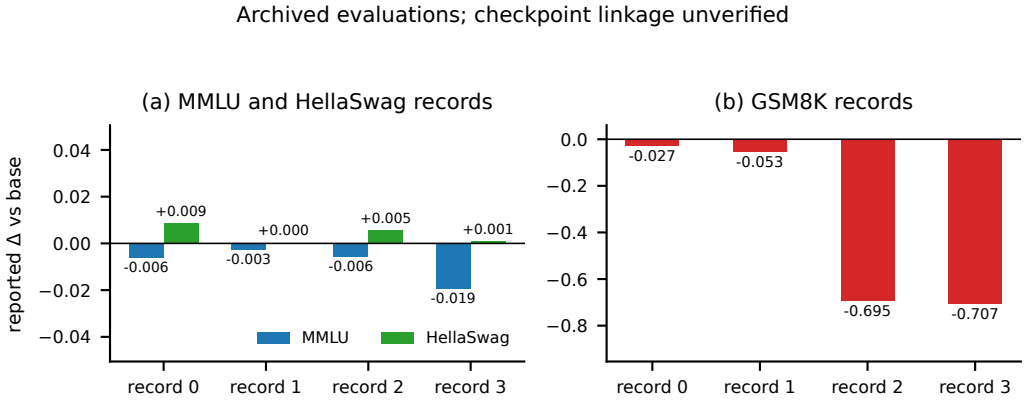


Figure 12: **Archived evaluation records; checkpoint linkage unverified.** The four directory-associated records report the plotted general-ability deltas. Their nominal method association does not establish which trained checkpoints were scored. The plot is retained for audit and supports no backbone-preservation, causal loader-effect or retraining-variance conclusion.

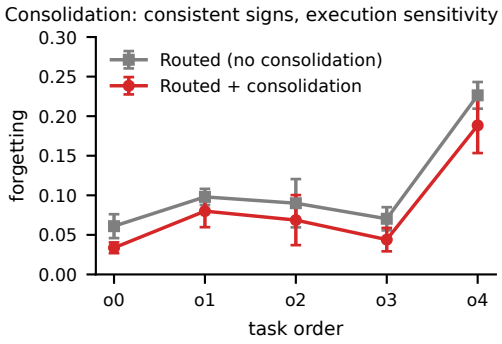


Figure 13: **Consolidation: consistent signs, sensitivity to re-execution.** Consolidation (R+cons) against its no-consolidation control reduces forgetting in 13 of 15 matched (seed, order) cells; mean $+0.026$, paired $t = 4.20$; the per-order ratio is 1.2–$1.8\times$ in its favour on every order, but the gain is below the arm's replicate spread (0.047), so it is sensitive under the per-pair discrepancy diagnostic. Caveats: the control's cells span two harness generations (its o3–o4 cells were run later than o0–o2), and a single-generation re-run of the control (15/15 cells, one machine) reproduces it: 14 of 15 cells, mean $+0.029$, t=4.12.

G WORKED EXAMPLE AND PAIRWISE MATRIX

Forward and gradient mechanics. For a sequence routed at one linear to experts u and v, the adaptation contribution is $\frac{\alpha}{r}(a_u B_u A_u x + a_v B_v A_v x)$, with $a_u + a_v = 1$. Other linears can select different experts. The frozen backbone still executes in full. In coupled mode the task gradient reaches the selected expert paths with these activation coefficients; the gated arm additionally scales expert task gradients by its plasticity gate. Auxiliary losses, consolidation penalties and dense optimizer state can affect parameters beyond this value path, so forward selection does not establish that only selected parameters update. This is a description of the retained implementation, not an observed checkpoint trace or evidence that routing explains any particular prediction.

G.1 FROM A REAL INPUT TO A REPORTING DECISION

This retrospectively selected walkthrough uses the first stored AG News evaluation item of workstation sequential LoRA, seed 300, order o2, in the completed fixed-host study (Appendix Q). It follows the classification and reporting pipeline used by this paper. The item was selected by its stored position, not its outcome; it is an illustration, not an estimate of typical behavior. The routed instrument in Figure 11 is a separate implementation, so no router or gate values are implied here.

1. Input and trained state. The frozen Qwen2.5-1.5B backbone carries the sequential-LoRA adapter learned through Yelp → AG News → Yahoo → Amazon → DBpedia. The final saved adapter state follows 4,000 optimizer steps. The raw AG News row has gold label `Sports`. Its complete rendered prompt is:

```
Classify the news topic.
text:  UPDATE 1-Beck to face Youzhny in St Petersburg
final Unseeded Slovak Karol Beck reached the first
final of his career at the St Petersburg Open,
upsetting seventh-seeded Michael Llodra of France 6-4
2-6 6-1 on Saturday.
Answer:
```

The gold label is used only to score correctness, never to construct the prompt or candidate scores. The four candidate suffixes are `World`, `Sports`, `Business` and `Sci/Tech`, each with a leading space. The retained prompt has 70 tokens; candidate lengths are 1, 1, 1, 3 tokens. Common-prompt scoring reserves space for the longest candidate under the 192-token cap and uses the identical retained prompt for each. This particular prompt is not truncated; the two scorers coincide on this item. All token IDs and ordered-input hashes are included in the trace receipt.

2. Forward scoring and prediction. For each candidate, the frozen dense backbone and its trained low-rank adapter produce next-token logits; the scorer averages the candidate tokens' conditional log probabilities. Thus $s(v) = |v|^{-1} \sum_j \log p_\theta(v_j \mid x, v_{<j})$, and the prediction is $\arg\max_v s(v)$. These are length-normalized log scores, not probabilities over the four classes. The final saved scores are:

Candidate	Tokens	Mean log score
World	1	-13.222926
Sports	1	-0.093073
Business	1	-17.300915
Sci/Tech	3	-6.702666

`Sports` has the largest score, so this item contributes 1 to the number correct. The audit reproduces this argmax from stored scores; it does not rerun the model or claim an internal-activation trace.

3. From one prediction to task accuracy. The same saved evaluation item is scored after every training stage. Paper stages 1–5 correspond to stored indices 0–4. The final AG News accuracy is the number correct across all 250 stored examples, not the outcome of this one item. The records show:

Completed stage	Item prediction	Correct	AG News accuracy
1 (yelp)	Sci/Tech	0	195/250 = 0.780
2 (ag_news)	Sports	1	233/250 = 0.932
3 (yahoo)	Sports	1	216/250 = 0.864
4 (amazon)	Sports	1	222/250 = 0.888
5 (dbpedia)	Sports	1	206/250 = 0.824

For the full run, rows below are completed training stages and columns are the five evaluation tasks in stream order; each entry is independently recovered from all 250 candidate-score records for that task and stage:

After training	Yelp	AG News	Yahoo	Amazon	DBpedia
1 (yelp)	0.640	0.780	0.300	0.956	0.720
2 (ag_news)	0.596	0.932	0.360	0.948	0.516
3 (yahoo)	0.464	0.864	0.632	0.888	0.156
4 (amazon)	0.576	0.888	0.440	0.952	0.284
5 (dbpedia)	0.588	0.824	0.440	0.956	1.000

4. From the matrix to forgetting. For each of the first four tasks, subtract its final accuracy from its maximum over the four pre-final training stages. The zero-shot baseline is excluded. The four signed drops are 0.052, 0.108, 0.192, 0.000, hence

$$\mathrm{AF} = \frac{0.052 + 0.108 + 0.192 + 0.000}{4} = 0.088, \qquad \mathrm{AA} = \frac{\sum_j R[5, j]}{5} = 0.7616.$$

AF is signed and is not clipped at zero. A single input cannot establish either aggregate: every task/stage evaluation contributes to this calculation.

5. From runs to a method comparison. Apply the same reconstruction to EWC-style LoRA and pair the ten aligned seed/data blocks on this same host and order. AF is in accuracy units; the last column is $100(\mathrm{AF}_{\mathrm{Seq}} - \mathrm{AF}_{\mathrm{EWC}})$ in percentage points (pp):

Seed	Seq. AF	EWC AF	Difference (pp)
300	0.088	0.085	+0.300
301	0.056	0.071	-1.500
302	0.081	0.080	+0.100
303	0.069	0.062	+0.700
304	0.046	0.047	-0.100
305	0.119	0.117	+0.200
306	0.089	0.089	+0.000
307	0.126	0.119	+0.700
308	0.042	0.040	+0.200
309	0.039	0.039	+0.000

The mean is $+0.060$ pp. With ten paired blocks, the nominal 95% paired-t interval is $\bar{d} \pm t_{0.975,9} s_d/\sqrt{10} = [-0.379, 0.499]$ pp. A mean-only rule favors EWC-style LoRA; an interval rule abstains because the interval includes zero. The same contrast has Holm-adjusted $p = 1.000$ in the existing 18-test primary family, so it does not reject there either. The nominal interval is assumption-dependent and unadjusted for multiplicity; non-rejection establishes neither equivalence nor hardware causation. This is a numerical expansion of an existing contrast, not an additional test or evidence that the historical reporting example on order o0 transfers to o2.

Provenance and scope. The selected run is `workstation_seq_lora_s300_o2`; source records are under `execution_confirmation/records/` in the anonymous supplement. `paper/generated/input_trace_receipt.json` binds the exact raw item, prompt and token IDs, original source-file hashes, saved checkpoint-index identity, all reconstructed matrices and the existing analysis. `WT01_input_trace_stages.csv` contains every candidate score for this

item at each stage; `WT02_input_trace_paired_runs.csv` gives all paired AF values. Run `python scripts/make_input_trace.py --check` to reconstruct the walkthrough from the released records. Remote receipts bind checkpoint validation, but checkpoint tensor bytes and internal activations are not in this portable trace. The other host and method comparisons remain in the full study tables.

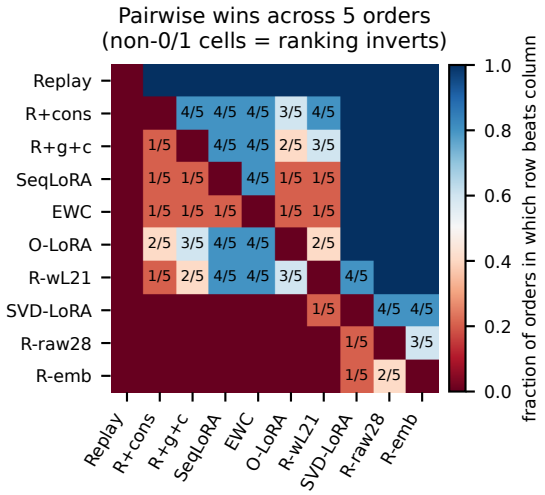


Figure 14: **Historical pairwise wins across the five orders** (ten-arm grid). Cell (i, j) = fraction of orders in which row-method i has lower forgetting than column-method j. Any cell that is neither 0 nor 1 is a pair whose ranking depends on the order: 19 of 45 among these ten arms, of which 0 clear the observed replicate maximum of Section 4 under the global rule and 4 under the per-pair rule, every one of those 4 a pair involving one of our own routed arms. Within seed—a pair counted as reversed when its ranking flips across orders in at least two of the three seeds—the count is again 19 of 45 here, and 12 of 28 for the eight eligible baseline arms against the 10 of 28 on order means (Section 5.1).

H GENERATION PROBE

A qualitative generation probe was recorded during development, but its checkpoint and evaluation linkage cannot be independently verified from the released artifacts. We exclude those outputs from scientific evidence and make no claim about their frequency, cause or association with any reported score.

I IDENTICAL-CONFIGURATION REPLICATES

Table 15 lists every controlled replicate pair behind the observed maximum of Section 4; Fig. 14 shows the pairwise reversal matrix over the original ten arms.

Eligibility and parser correction. The frozen historical inventory contains 85 attributed replicate comparisons. Six involve the defective prompt port and are retained as audit artifacts; the comparison reference uses 79 eligible pairs over 57 cells. An earlier filename parser omitted arm names containing underscores. Correcting it recovered two already-existing sequential-LoRA pairs, one cross-machine and one exact same-device repeat; no new training was added. The same parser repair includes sequential LoRA in paired-difference summaries. Supplementary statistics retain historical archived and eligible populations separately; the later inventory and its candidate sensitivity are in Appendices R–S.

Every controlled replicate pair in the historical reference set behind Section 4: the same method, seed, task order and hyperparameters, re-run either on a second GPU model (RTX 6000 Ada vs. RTX PRO 6000; library versions pinned by the requirements file—torch 2.11.0+cu128, transformers 4.48.0, peft 0.14.0—not recorded per run; the two machines run Python 3.10 and 3.12, and the earlier third environment of Appendix A Python 3.11.10) or twice on the same GPU. Each side is a released run directory; origin is attributed by the hostname in the manifest where recorded and otherwise by interpreter version, and pairs whose two sides cannot be attributed to distinct environments are excluded from the attributed replicate pool. Types: cross-machine = two distinct hosts; cross-device = two runs on the six-GPU host, which need not share a device; same device = the single-GPU workstation. Four of the 40 cross-machine pairs contrast the earlier third environment of Appendix A rather than the two machines above, and one of them sets the cross-machine maximum (R+cons, seed 2, o2, $|\Delta|$=0.047); over the two machines above alone the maximum among eligible arms is 0.045, under which one further pair of our own arms (R+g+c vs. O-LoRA) would clear the global rule and no comparison among the external arms changes. Seeds 6–9 were assigned to machines in whole blocks (two per machine), so their machine contrast is confounded with seed; its sign pattern—-0.027 (EWC), -0.022 (seq. LoRA), -0.013 (O-LoRA), -0.005 (SVD-LoRA), $+0.008$ (replay)—is consistent with the per-arm replicate spreads of Section 4 but does not re-derive them independently.

Table 15: Controlled replicate pairs in the frozen historical archive. Duplicate rows carry multiplicity n; the rows reconstruct all 79 historical comparisons. $|\Delta|$ is the absolute forgetting difference. The historical eligible-arm maximum including cross-environment pairs is 0.095; the defective Progressive Prompts port is excluded from this table and retained in the archival JSON.

Arm	Seed	Order	Type	Run A	Run B	$\lvert\Delta\rvert$	n
Fixed-rank SVD-LoRA	0	o3	cross-machine	0.1660	0.1660	0.0000	1
Fixed-rank SVD-LoRA	1	o3	cross-machine	0.1690	0.1690	0.0000	1
R+cons (whitened L28 + consol.)	1	o3	cross-machine	0.0060	0.0510	0.0450	1
Seq. LoRA + EWC	0	o3	cross-machine	0.0730	0.0540	0.0190	1
Seq. LoRA + EWC	1	o3	cross-machine	0.1090	0.0910	0.0180	1
O-LoRA	0	o3	cross-machine	0.0730	0.0730	0.0000	1
O-LoRA	1	o3	cross-machine	0.0780	0.0760	0.0020	1
O-LoRA	2	o3	cross-machine	0.1020	0.1040	0.0020	1
Replay (2%)	0	o3	cross-machine	0.0170	0.0200	0.0030	1

Controlled replicate pairs (continued)

Arm	Seed	Order	Type	Run A	Run B	$\lvert\Delta\rvert$	n
Replay (2%)	1	o3	cross-machine	-0.0070	-0.0080	0.0010	1
Seq. LoRA	0	o3	cross-machine	0.0740	0.0490	0.0250	1
Seq. LoRA	1	o3	cross-machine	0.1040	0.0920	0.0120	1
R-emb (embed ctx)	1	o2	cross-machine	0.2280	0.2560	0.0280	1
R-raw28 (raw L28, unwhitened)	0	o1	cross-machine	0.2280	0.2090	0.0190	1
R-wL21 (whitened L21)	0	o3	cross-machine	0.0580	0.0570	0.0010	1
R+g+c (gate + consol.)	0	o1	cross-machine	0.1240	0.1270	0.0030	1
R+g+c (gate + consol.)	1	o2	cross-machine	0.0320	0.0630	0.0310	1
Fixed-rank SVD-LoRA	2	o2	cross-machine	0.1860	0.1860	0.0000	1
R+cons (whitened L28 + consol.)	2	o2	cross-machine	0.0570	0.1040	0.0470	1
Seq. LoRA + EWC	2	o2	cross-machine	0.0770	0.0920	0.0150	1
R-emb (embed ctx)	2	o2	cross-machine	0.2050	0.2040	0.0010	1
R-raw28 (raw L28, unwhitened)	2	o2	cross-machine	0.1780	0.1830	0.0050	1
R-wL21 (whitened L21)	2	o2	cross-machine	0.0620	0.0870	0.0250	1
R+g+c (gate + consol.)	2	o2	cross-machine	0.0720	0.0800	0.0080	1
O-LoRA	2	o2	cross-machine	0.0350	0.0350	0.0000	1
Replay (2%)	2	o2	cross-machine	0.0080	0.0070	0.0010	1
Seq. LoRA	2	o2	cross-machine	0.1020	0.0930	0.0090	1
InfLoRA	0	o1	cross-device	0.0700	0.0700	0.0000	1
InfLoRA	1	o2	cross-device	0.0450	0.0450	0.0000	1
LoRAMoE-inspired port	0	o1	cross-device	0.1000	0.1040	0.0040	1
LoRAMoE-inspired port	1	o2	cross-device	0.0800	0.0800	0.0000	1
N-LoRA (factor L_1)	0	o1	cross-device	0.0860	0.0840	0.0020	1
N-LoRA (factor L_1)	1	o2	cross-device	0.0700	0.0700	0.0000	1
InfLoRA	0	o1	cross-machine	0.0860	0.0700	0.0160	2
InfLoRA	1	o2	cross-machine	0.0480	0.0450	0.0030	2
LoRAMoE-inspired port	0	o1	cross-machine	0.0880	0.1000	0.0120	1
LoRAMoE-inspired port	1	o2	cross-machine	0.0610	0.0800	0.0190	2
N-LoRA (factor L_1)	0	o1	cross-machine	0.0920	0.0860	0.0060	1
N-LoRA (factor L_1)	1	o2	cross-machine	0.0550	0.0700	0.0150	2
LoRAMoE-inspired port	0	o1	cross-machine	0.0880	0.1040	0.0160	1
N-LoRA (factor L_1)	0	o1	cross-machine	0.0920	0.0840	0.0080	1
Fixed-rank SVD-LoRA	0	o3	same device	0.1660	0.1660	0.0000	1
R+cons (whitened L28 + consol.)	0	o3	cross-machine	0.0850	0.0540	0.0310	1
Seq. LoRA + EWC	0	o3	same device	0.0540	0.0540	0.0000	2
R-wL21 (whitened L21)	0	o3	same device	0.0570	0.0570	0.0000	1
O-LoRA	0	o3	same device	0.0730	0.0730	0.0000	1
Seq. LoRA	0	o3	same device	0.0490	0.0490	0.0000	1
R-emb (embed ctx)	0	o0	cross-environment	0.3250	0.2300	0.0950	1
R-emb (embed ctx)	1	o0	cross-environment	0.1520	0.1960	0.0440	1
R-emb (embed ctx)	2	o0	cross-environment	0.2370	0.2230	0.0140	1
R-wL21 (whitened L21)	0	o0	cross-environment	0.0940	0.0650	0.0290	1

Controlled replicate pairs (continued)

Arm	Seed	Order	Type	Run A	Run B	$\lvert\Delta\rvert$	n
R-wL21 (whitened L21)	1	o0	cross-environment	0.0620	0.0590	0.0030	1
R-wL21 (whitened L21)	2	o0	cross-environment	0.0660	0.0360	0.0300	1
R-raw28 (raw L28, unwhitened)	0	o0	cross-environment	0.1810	0.2200	0.0390	1
R-raw28 (raw L28, unwhitened)	1	o0	cross-environment	0.2200	0.2030	0.0170	1
R-raw28 (raw L28, unwhitened)	2	o0	cross-environment	0.2600	0.2420	0.0180	1
R+cons (whitened L28 + consol.)	0	o0	cross-environment	0.0310	0.0260	0.0050	1
R+cons (whitened L28 + consol.)	1	o0	cross-environment	0.0220	0.0360	0.0140	1
R+cons (whitened L28 + consol.)	2	o0	cross-environment	0.0190	0.0390	0.0200	1
R+g+c (gate + consol.)	0	o0	cross-environment	0.1070	0.0830	0.0240	1
R+g+c (gate + consol.)	1	o0	cross-environment	0.0430	0.0320	0.0110	1
R+g+c (gate + consol.)	2	o0	cross-environment	0.0600	0.0860	0.0260	1
R+cons (whitened L28 + consol.)	0	o1	cross-environment	0.0950	0.1020	0.0070	1
R+cons (whitened L28 + consol.)	0	o2	cross-environment	0.0590	0.0590	0.0000	1
R+cons (whitened L28 + consol.)	1	o1	cross-environment	0.0480	0.0760	0.0280	1
R+cons (whitened L28 + consol.)	1	o2	cross-environment	0.0360	0.0430	0.0070	1
R+cons (whitened L28 + consol.)	2	o1	cross-environment	0.0600	0.0620	0.0020	1
R+cons (whitened L28 + consol.)	2	o2	cross-environment	0.0720	0.1040	0.0320	1
R+g+c (gate + consol.)	0	o1	cross-environment	0.0670	0.1270	0.0600	1
R+g+c (gate + consol.)	0	o2	cross-environment	0.1250	0.0720	0.0530	1
R+g+c (gate + consol.)	1	o1	cross-environment	0.0560	0.1200	0.0640	1
R+g+c (gate + consol.)	1	o2	cross-environment	0.0660	0.0630	0.0030	1
R+g+c (gate + consol.)	2	o1	cross-environment	0.0710	0.0470	0.0240	1
R+g+c (gate + consol.)	2	o2	cross-environment	0.0480	0.0800	0.0320	1

An uncontrolled population, disclosed. An earlier phase of the campaign produced nine further nominally-identical triplets (the inert-flag calibration runs of Section 5.3: same seed, order and effective configuration three times). Their spreads reach 0.158; a further 18 nominally-identical pairs outside these triplets, likewise unattributable, differ by up to 0.182—but these runs predate our manifest writer, so the producing machine and code revision are unattributable, and several are known to span machines. We disclose them because they suggest future replicate discrepancies may exceed the observed controlled maximum; we use them for nothing quantitative. The 30 single-machine

re-runs of Section 5.3 are the third excluded population: they are controlled, but cannot be attributed to distinct environments, so they too are excluded from the attributed replicate pool; their disagreements are reported there.

J Extended related work

Continual learning of LLMs with adapters. Rehearsal-free CL for language models has converged on parameter-efficient adapters over a frozen backbone: sequential LoRA (Hu et al., 2022) with or without EWC anchoring (Kirkpatrick et al., 2017; Zenke et al., 2017); orthogonal per-task subspaces (O-LoRA, Wang et al., 2023a; InfLoRA, Liang & Li, 2024; N-LoRA, Yang et al., 2025); adaptive-rank allocation (Zhang et al., 2023), proposed for single-task tuning; our historical SVD-LoRA variant omits its allocation callback, as the implementation audit above explains; prompt-based methods (Wang et al., 2022; Razdaibiedina et al., 2023); and mixture-of-LoRA routing (Dou et al., 2024; Luo et al., 2024; Zhao et al., 2024); InfLoRA (ported here) and L2P were proposed for vision transformers. Small replay buffers (Chaudhry et al., 2019; de Masson d'Autume et al., 2019) remain strong and under-reported. The language-model CL methods among these are evaluated on the text-classification stream of de Masson d'Autume et al. (2019)—five tasks in MbPA++, LAMOL (Sun et al., 2020) and IDBR (Huang et al., 2021), the four-task subset of LFPT5 (Qin & Joty, 2022) in O-LoRA (Wang et al., 2023a) and N-LoRA; their short classification comparisons use three orders, while O-LoRA's LLaMA comparison uses one. O-LoRA and N-LoRA additionally study long streams; these counts do not describe their entire evaluation.

The order effect in published tables. The instrument's coarseness shows in the published numbers themselves: in O-LoRA's Table 2 (Wang et al., 2023a), O-LoRA beats LFPT5 by 7.8 and 3.1 average-accuracy points on two of its three orders and loses by 1.6 on the third, so the published comparison itself reverses under order, while sequential fine-tuning ranges 22.8 points and LFPT5 itself 10.3 across those orders.

K Expanded evaluation and correlated sampling (R3)

The fixed review-response plan evaluated the core five on every o0–o4 and seed 0–2 combination. All 75 records contain the zero-shot sweep and all post-task stages, with 7,600 correctness bits per task per stage. Recomputing each accuracy from its bit-string reproduces the run matrix and its signed forgetting. Table 16 gives the complete method comparison; Figure 15 separates the cross-training comparison from the evaluation-only analysis.

Table 16: R3 expanded evaluation. Forgetting is a fraction; lower is better. Means pool the same fifteen method–seed–order coordinates at each evaluation size; the trainings differ. SD is across all fifteen R3 runs, combining seed and order variation, not a standard error. Swing is max/min of the five order means. ANOVA p is uncorrected for the order factor, with seed as the other additive factor, df $(4, 8)$.

Method	Mean 250	Mean 7,600	SD 7,600	Swing	Order p
Replay	0.0223	0.0200	0.0103	3.10	0.0565
O-LoRA	0.0809	0.0745	0.0252	1.83	0.4640
Sequential LoRA	0.1081	0.1050	0.0259	1.70	0.0495
EWC	0.1102	0.1078	0.0216	1.51	0.0692
Fixed-rank SVD-LoRA	0.1973	0.1835	0.0372	1.58	0.0184

Rankings and order. The pooled ordering is replay, O-LoRA, sequential LoRA, EWC, SVD-LoRA at both evaluation sizes. Across orders, 1 of ten method pairs reverses, against 3 on the canonical three-seed grid; 45 of fifty pair-order winners agree. Both independent and block-preserving permutations put the count within their nulls ($p = 0.9998$ and $p = 1.0$ respectively). The nominal order tests for

SVD-LoRA and sequential LoRA are below the conventional threshold, but none of the five R3 order tests survives Holm correction. This smaller grid does not supersede the ten-seed order-effect analysis. Disagreement with canonical cells mixes retraining and evaluation changes, so fewer reversals cannot be attributed solely to evaluation size.

Conditional resampling. For each completed training, we take 2000 draws without replacement of 250 indices per task from its recorded evaluation pool, using the same indices at every stage. Each draw recomputes the signed max-before-final forgetting estimator. A comparison independently draws indices at each stage on the same recorded predictions. The median joint SD is 0.0119, versus 0.0158 independently; the median of their cellwise ratios is 0.74 (distinct from the ratio of rounded medians). Median cross-stage correlation is 0.455. This is a finite-pool, within-training sensitivity analysis; it is not uncertainty over new trainings or an estimate of population-level error. We release every per-cell SD in supplementary CSV T4 and the underlying bit-strings.

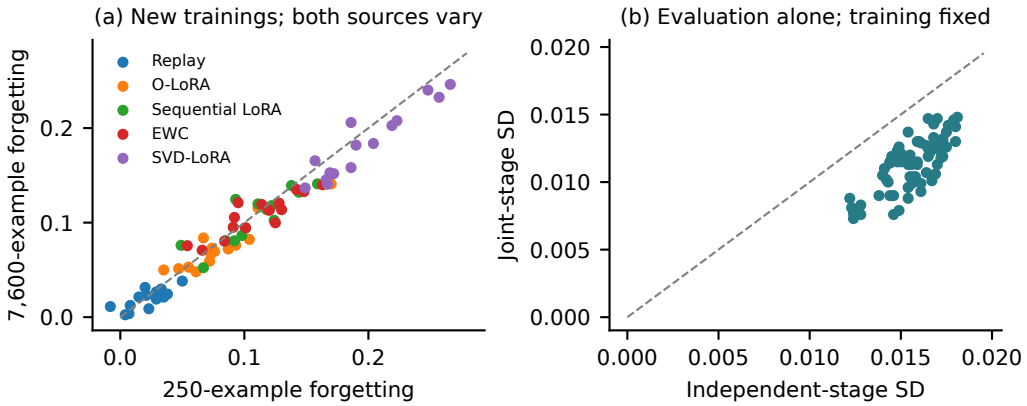


Figure 15: R3 at two levels. (a) Each point pairs a method–seed–order coordinate from two different trainings: differences combine training and evaluation variation. (b) Each point holds one R3 training fixed; aligned-example resampling yields smaller SD than independent-stage resampling. Dashed lines mark equality. Neither panel estimates a universal reproducibility threshold.

Record provenance and scorer changes. Five pod results were complete but their locally copied evaluation records stopped mid-stream; complete records were unavailable after the machine was reclaimed. Workstation duplicates at the same coordinates were adopted for record completeness, before computing R3 statistics, and the pod versions are retained under quarantine. The scorer was accelerated by reusing each prompt's cached state across verbalizers; earlier cells retain the original scorer. Saved same-seed zero-shot correctness vectors agree exactly across the records. This agreement does not establish underlying example identities; Appendix O reports the missing cross-method identity checks. It checks recorded base outcomes, not bitwise equivalence of every trained checkpoint's logits or causal identification of hardware effects. The release includes the scorer-equivalence check and the independent record-to-matrix validation.

L Multiplicity and covariance sensitivity

Declared families. Each permutation design has six pools; Holm correction is applied within each design, retaining all six tests. Table 17 reports both raw and adjusted values. Non-rejection does not establish equivalence or absence of order effects. The permutations require exchangeability of order labels within their stated seed blocks; equal means alone are insufficient. The independent-arm permutation breaks cross-arm matching, while the shared permutation preserves it. The independent family has 0 Holm rejections. The block-preserving family has 1; the ten-arm pool, including our routed arms, has adjusted $p = 0.0348$. This exception is retained rather than folded into an alternative multiplicity family after observing its significance.

Table 17: Permutation families after exclusion of the defective prompt implementation. Each design is corrected across the same six pools.

Pool	Independent p	Holm p	Block p	Holm p
Core five	0.3783	0.7566	0.2525	0.5050
Eight external	0.6509	0.7566	0.3131	0.5050
Ten arms	0.0324	0.1944	0.0058	0.0348
All thirteen	0.1438	0.6145	0.0318	0.1510
Core five, Llama	0.1229	0.6145	0.0734	0.2202
Core five, SmolLM2	0.1559	0.6145	0.0302	0.1510

Inferential unit and secondary analysis. The target is these five fixed orders under the recorded recipe. We retain the original additive analysis of variance (ANOVA), then treat the ten matched seeds as blocks and report a Greenhouse–Geisser (GG) degree-of-freedom sensitivity for the covariance among the five order responses within a seed. This is a secondary historical analysis selected on design grounds, not a replacement chosen by significance. Independent seed blocks and approximate distributional assumptions remain; whole-seed machine assignment limits exchangeability. These tests do not establish behavior over random unseen orders or different harnesses.

Table 18: Order-effect sensitivity on the same 250 historical records. Holm correction is separately applied across the five arms for each stated analysis.

Arm	Nominal p	Holm p	ϵ_{GG}	GG p	GG+Holm p
O-LoRA	0.1520	0.1520	0.670	0.1797	0.1797
Fixed-rank SVD-LoRA	2.58×10^{-9}	1.29×10^{-8}	0.690	5.12×10^{-7}	2.56×10^{-6}
Sequential LoRA	3.12×10^{-5}	9.37×10^{-5}	0.544	0.0012	0.0037
EWC-style LoRA	4.20×10^{-6}	1.68×10^{-5}	0.596	2.35×10^{-4}	9.38×10^{-4}
Replay	0.0212	0.0423	0.738	0.0364	0.0728

Table 18 retains both outcomes: 4 of five survive nominal-ANOVA Holm correction, versus 3 with GG adjustment. Replay's adjusted GG result does not reject; the strong SVD-LoRA, sequential-LoRA and EWC-style signals remain. Among the ten paired-difference order tests, 7 reject nominally and 6 survive Holm; 4 survive GG plus Holm. The full JSON retains every pair, including unsuccessful comparisons. A second interaction decomposition removes seed-by-method and seed-by-order effects and tests method-by-order against the remaining three-factor term: $F(16, 144) = 5.531$, p=4.15×10^{-9}. This strengthens the omnibus signal under that alternative model; it is not a covariance-free proof. Its univariate reference still assumes sphericity. We do not select the smaller p as the headline result. For the one-factor sensitivity, let C be the sample covariance of order responses centered within each seed. Then $\epsilon_{GG} = \mathrm{tr}(C)^2/[4\mathrm{tr}(C^2)]$; the original F statistic is retained with degrees of freedom $(4\epsilon_{GG}, 36\epsilon_{GG})$. The analysis and all adjusted declarations are generated by `scripts/statistical_inference.py`. They do not inspect or modify the frozen prospective cohort.

M Retrospective audit of reporting decisions

Question and population. How much contradictory reporting can conventional decision rules avoid, and how often do they decline to name a winner? This supplementary, retrospective analysis reuses 475 saved training records: the complete ten-seed Qwen grid and the original three-seed Llama, SmolLM2 and R3 grids, each over all five orders. It is not an untouched experiment or the prospective confirmation study. Every matrix is independently checked against its recorded metrics. We report three implementation pools: sequential LoRA, EWC-style LoRA and replay; these plus O-LoRA (the core four); and these plus historical fixed-rank SVD-LoRA. The figures and tables below use the core four; Figure 16 and Tables 19–20 show the reporting trade-offs; the full CSVs retain all pools,

orders and metrics. The partially filled later Llama seeds are excluded as an incomplete block. Pool exclusions do not certify baseline fidelity.

Decision and reference groups. Within each method pair and order, we enumerate every pair of disjoint seed groups. For Qwen's ten seeds, both groups have size k; this uses $2k$ recorded seeds per method ($4k$ training records across the two methods). The first group supplies a declaration, and the second its reference sign. For the smaller blocks we explicitly use two decision seeds and one reference seed. The reference mean is uncertain; its sign is not a known population winner. The groups and pair-order comparisons overlap, so enumerations are not independent trials. No binomial error bars, coverage estimates or Type-I error guarantees follow from these counts.

Three conventional rules. The mean-sign rule declares the sign of the paired mean unless tied. A separate-band heuristic declares only when the two methods' mean $\pm$ one sample-SD bands do not overlap; these bands are not confidence intervals. The paired-t rule requires a conventional two-sided nominal 95% interval on paired differences to exclude zero. Its nominal level relies on sampling assumptions that this audit does not verify. Variance-based rules are unavailable with one decision seed. Ties use tolerance 10^{-12}. We count declarations, opposite reference signs and reference ties separately; an undefined contradiction rate after no declarations is not reported as zero. These are standard rules, not a proposed new statistical test.

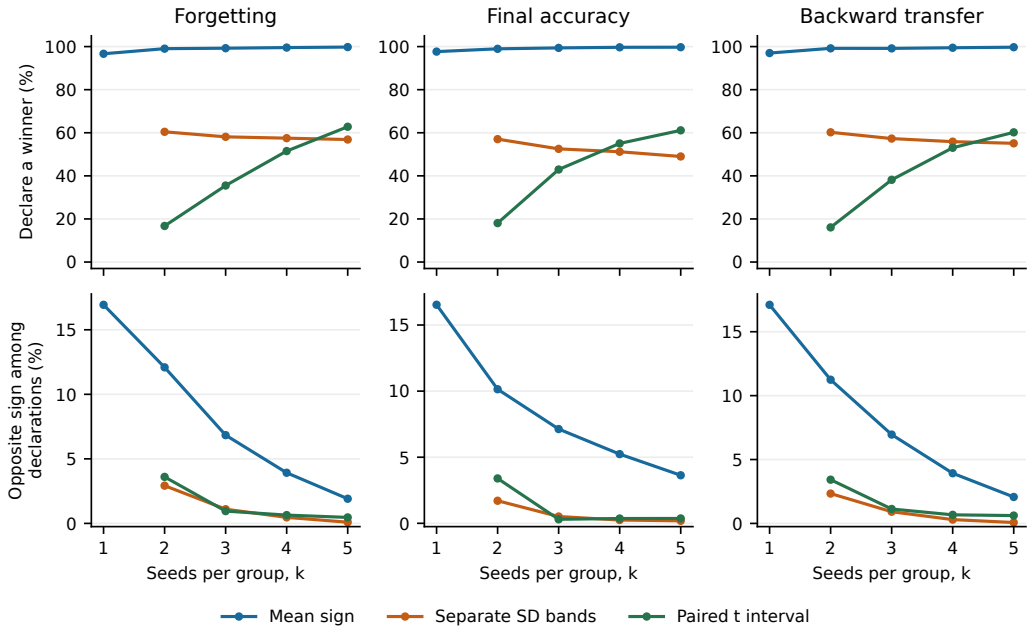


Figure 16: Retrospective reporting trade-offs on the ten-seed Qwen core-four grid. Top: fraction of enumerated decisions declaring a winner. Bottom: fraction of those declarations opposed by a disjoint k-seed reference mean. A lower contradiction rate must be interpreted together with declaration frequency. Each panel uses all pair-order comparisons for its metric; metrics have distinct estimands. Curves describe this finite corpus, and neither reference signs nor nominal intervals establish population truth.

What the trade-off shows. At three seeds per group, always using the paired mean declares in 99.25% of decisions, with an opposite reference sign in 6.84% of declarations. The paired-t rule declares 35.50%, with 0.95% opposed; the separate-SD heuristic declares 58.11%, with 1.10% opposed. These rules select different subsets, so this comparison does not isolate a causal benefit or prove that one dominates. The paired-t rule's abstentions are part of its cost. The full curves and tables retain settings where the simple SD heuristic has fewer contradictions than the paired-t rule. A reporting recommendation must state its abstention burden as well as its observed consistency; merely suppressing almost every comparison is not sufficient.

Table 19: Qwen core-four forgetting: equal decision/reference group sizes. All rates are percentages. Opposite and tied reference signs are conditional on declaring; neither is a statistical error-rate estimate.

k	Rule	Declare	Opposite	Ref. tie
1	Mean sign	96.67	16.93	3.22
2	Mean sign	99.04	12.09	0.91
2	Separate SD bands	60.44	2.92	0.09
2	Paired t interval	16.74	3.60	0.62
3	Mean sign	99.25	6.84	0.75
3	Separate SD bands	58.11	1.10	0.01
3	Paired t interval	35.50	0.95	0.06
4	Mean sign	99.54	3.92	0.46
4	Separate SD bands	57.49	0.47	0.02
4	Paired t interval	51.51	0.64	0.03
5	Mean sign	99.78	1.91	0.23
5	Separate SD bands	56.87	0.09	0.00
5	Paired t interval	62.79	0.46	0.00

Table 20: Other complete blocks, core-four forgetting, two decision seeds versus one disjoint reference seed. These small blocks provide a retrospective stress test of the rules; they cannot validate a five-seed requirement or be pooled as equal-size replicas of Table 19. Rates and conditioning are defined as in that table.

Population	Rule	Declare	Opposite	Ref. tie
Llama-3B	Mean sign	100.00	13.33	0.00
Llama-3B	Separate SD bands	72.22	7.69	0.00
Llama-3B	Paired t interval	18.89	5.88	0.00
SmolLM2	Mean sign	93.33	11.90	8.33
SmolLM2	Separate SD bands	63.33	0.00	0.00
SmolLM2	Paired t interval	12.22	0.00	0.00
Qwen R3	Mean sign	100.00	11.11	0.00
Qwen R3	Separate SD bands	70.00	4.76	0.00
Qwen R3	Paired t interval	17.78	6.25	0.00

Influence and limits of pooling. We also remove each seed and each order in turn, retaining every deletion. Across all three pools and metrics, the ten-seed Qwen pooled rank order survives every single deletion. This is not uniform across populations: in the R3 core-four forgetting analysis, omitting o3 changes the pooled ordering. A stable full-grid ranking therefore need not survive a change to the order population. The complete deletion tables include metric-specific reversal counts and ties. These are retrospective influence checks, not new training, causal attribution or confidence intervals. The smaller-backbone checks do not establish transport to other task types, long streams or larger models.

Baseline and budget interpretation. The sequential control uses a persistent adapter from the parameter-efficient fine-tuning (PEFT) library. EWC uses an exponential moving average (EMA) of squared minibatch training gradients and an accumulated anchor on adapter weights; it is an EWC-style proxy, not an independently reproduced Fisher estimator from the source algorithm. Replay stores a reservoir fraction of the stream and, once available, adds a replay minibatch loss to the current minibatch loss before an update. Thus equal update counts do not equalize processed tokens or compute. O-LoRA grows one adapter per task, freezes earlier adapters and sums them at evaluation, with the implementation's orthogonality penalty; it is evaluated under our recipe rather than source-optimal tuning. The singular-value-decomposition (SVD) LoRA arm remains fixed rank. The extended-port departures already stated in Appendix A are not repaired by excluding those arms from a sensitivity table. These facts bound our implementation comparisons; no table here establishes that the published algorithms were faithfully reproduced and equally tuned.

Reusable output. `scripts/reporting_audit.py` reproduces the audit from the registry. With `--decide` and a JSON file containing aligned metric arrays `a` and `b`, it reports the paired mean, nominal paired interval and explicit declarations/abstentions for a user's matched runs. It does not choose a sample size or adjust for searching many comparisons. Tables T5–T7 release every aggregate, pair-order row and deletion; `stats/reporting_audit.json` includes input-file identities. No confirmation-cohort record is used in this retrospective analysis.

N FROZEN FOLLOW-UP ON NEW TRAINING SEEDS

Question fixed before outcomes. On September 15, we froze the local protocol, development rates, complete run list, source/configuration fingerprints, model revision and package versions before any follow-up outcome. This was not an externally registered study. The question is whether the historical disjoint-group agreement description transfers to a new cohort under the documented recipe. It differs from the later retrospective paired-*t*/SD-band audit in Appendix M and does not prospectively validate those reporting rules. The endpoint and threshold were not revised after observing outcomes.

Population and admission. All 150 trajectories completed: 120 primary trajectories for sequential LoRA, EWC, replay and the existing O-LoRA implementation, plus 30 secondary adaptive-AdaLoRA trajectories. Each method uses seeds 100–109 on orders o0–o2, with Qwen2.5-1.5B, fp32 (single precision), scaled dot-product attention (SDPA), 1,000 training and 250 evaluation examples per task, 800 updates per task, batch four, length 192, rank 64, alpha 16 and learning rate 10^{-4}. EWC uses $\gamma = 0.25$, O-LoRA orthogonality weight 0.5 and replay a 2% complete-stream buffer. These retain the shared-recipe limitations of Section 3. Three orders test the frozen question of whether agreement persists in fresh runs; they do not balance all five leading tasks or support a broader order-balanced ranking claim. Training and evaluation subsets and initialization change with seed. Every record passes the frozen configuration, 4,000-update, model/source/runtime identity, prediction-to-matrix, recomputed-metric and shared-seed dataset-hash checks. All 150 were required, including the secondary block; no partial outcome selection or replacement of unfavorable cells was used. Source fingerprints cover 66 execution files. Analysis recomputation on a different local runtime is distinguished from the recorded execution-runtime check.

Scheduling amendment and scope. The original assignment kept a seed's methods/orders on one host. An amendment moved 23 never-started workstation cells by queue position, preserving active and completed jobs and the frozen scientific settings. Actual execution comprises 67 workstation and 83 cluster trajectories. Some paired method comparisons consequently span hosts, so host effects need not cancel. Host, seed and scheduling are not randomized causal factors. The primary analysis retains the whole recorded execution mixture. CSV T13 provides explicitly exploratory same-host paired-subset results; these smaller selected subsets do not replace the primary analysis. On o2, the non-replay pairs have only 3–5 same-host seeds, permitting disjoint groups of $k = 1$ or 2, never $k = 5$. Their small-group variation within hosts cannot adjudicate the frozen five-seed endpoint or isolate hardware causation. This cohort missed the operational cutoff of the earlier manuscript build and enters this separately checked revision in full.

Frozen estimand. For each of 18 method-pair–order combinations and $k = 1, \ldots, 5$, enumerate ordered pairs of disjoint k-seed subsets among ten seeds. Agreement requires the same strict sign in both groups; differences within 10^{-12} are unresolved. The criterion is agreement at least 95.00%, not a confidence level or a significance test. Table 21 reports every k. "Retained" is the intersection of development and fresh comparisons meeting the criterion; "fresh total" also counts newly qualifying cases. The mean absolute discrepancy uses all 18 comparisons, not only retained ones. The frozen JSON names this quantity calibration error; it is a finite-cohort rate discrepancy, not probability calibration against independent Bernoulli outcomes.

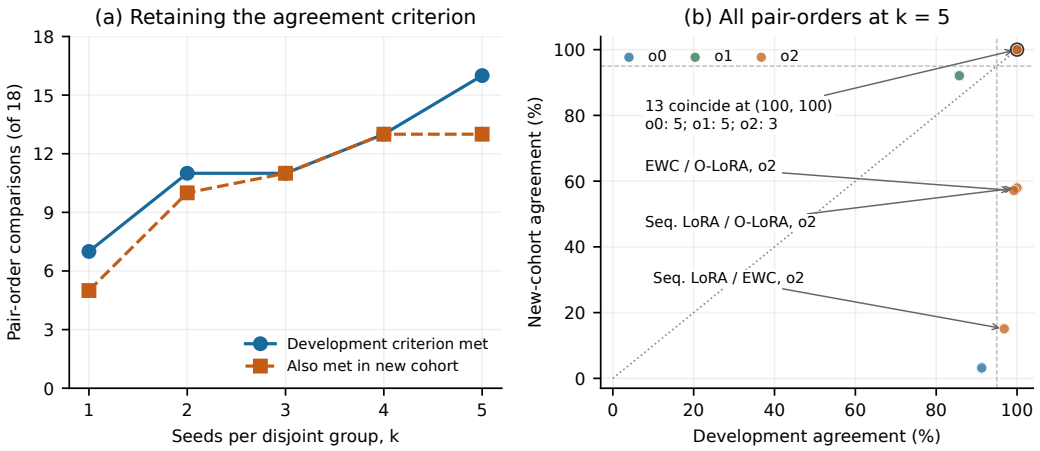


Figure 17: Complete development-to-fresh agreement view. Left: development-qualified comparisons and those retaining the criterion, by group size. Right: every primary pair–order comparison at five seeds per group, including labeled coincident points. Equal agreement rates need not preserve the same winner. Overlapping partitions describe finite cohorts, not independent replication trials.

Table 21: Complete frozen agreement summary. Counts concern the same 18 pair–order combinations, but the development-qualified subset changes with k. Discrepancy is in percentage points (pp). Retention does not require preserving the old winner.

k	Development qualified	Retained	Fresh total	Mean absolute discrepancy (pp)
1	7	5	7	12.10
2	11	10	10	6.51
3	11	11	13	9.98
4	13	13	13	12.36
5	16	13	13	14.46

Replay and non-replay comparisons. A post-hoc descriptive partition of the unchanged endpoint gives 9/9 retention for comparisons involving replay and 4/7 for other comparisons at $k = 5$, accounting for the total 13/16. The denominator in each stratum is its development-qualified comparisons, not all comparisons or independent future studies. Replay receives additional example exposure; this partition is not an equal-compute superiority test. Table 22 retains every group size; CSV T13b also records each stratum's full comparison count and fresh-qualified count. This interpretation was added after outcomes and does not change the frozen threshold or endpoint.

Table 22: Retained/development-qualified comparisons by replay involvement. The two strata partition each row of Table 21. An entry $0/0$ denotes no development-qualified comparison, not a measured rate. Retention concerns within-cohort agreement, not preservation of the old winner.

k	Replay pairs	Non-replay pairs
1	5/7	0/0
2	9/9	1/2
3	9/9	2/2
4	9/9	4/4
5	9/9	4/7

Failures and dependence. At $k = 5$, 13/16 development-qualified comparisons retain the criterion. All 3 failures are on o2. Sequential LoRA versus EWC changes from 96.83% to 15.08% agreement, with 3.97% unresolved group pairs. Figure 17 and Table 23 show the full comparison set; CSV T10 retains all group sizes, enumeration counts and tie fractions. The enumerations overlap and are dependent. For paired differences $d_1, \ldots, d_{10}$, any complementary five-seed groups S and S^c obey

$$\bar{d}_S + \bar{d}_{S^c} = 2\bar{d}_{10}.$$

Conditional on the fixed ten-seed cohort, one group's mean therefore determines the other; when subset means vary their correlation is -1. If the full-cohort mean is zero, every nonzero complementary pair has opposite signs. Near-zero cohort means can consequently yield low agreement at $k = 5$ without a corresponding failure probability for independent future studies. Smaller disjoint groups leave seeds unused and need not have this same geometry. Overlap between enumerated pairs also precludes treating their count as independent trials. The changing qualifying set and nonmonotonic rate discrepancies do not establish that collecting additional independent runs worsens estimation. No binomial interval or universal replication probability follows. Appendix V separately varies the forgetting definition while retaining the original frozen endpoint.

Table 23: All primary pair–order comparisons at $k = 5$. Rates and ties are percentages; Δ is first-listed minus second-listed mean forgetting in accuracy fractions (lower is better). The gap columns are descriptive cross-cohort diagnostics added after completion, separate from the frozen rate endpoint. Ties in either group count as unresolved.

Pair	Order	Dev. agree	New agree	New ties	Dev. Δ	New Δ
EWC / O-LoRA	o0	100.00	100.00	0.00	$+0.0422$	$+0.0475$
EWC / O-LoRA	o1	100.00	100.00	0.00	$+0.0332$	$+0.0239$
EWC / O-LoRA	o2	99.21	57.14	1.59	$+0.0203$	$+0.0096$
EWC / Replay	o0	100.00	100.00	0.00	$+0.0878$	$+0.1142$
EWC / Replay	o1	100.00	100.00	0.00	$+0.0751$	$+0.0593$
EWC / Replay	o2	100.00	100.00	0.00	$+0.0562$	$+0.0490$
Replay / O-LoRA	o0	100.00	100.00	0.00	-0.0456	-0.0667
Replay / O-LoRA	o1	100.00	100.00	0.00	-0.0419	-0.0354
Replay / O-LoRA	o2	100.00	100.00	0.00	-0.0359	-0.0394
Seq. LoRA / EWC	o0	91.27	3.17	2.38	-0.0059	$+0.0005$
Seq. LoRA / EWC	o1	85.71	92.06	0.79	-0.0038	$+0.0054$
Seq. LoRA / EWC	o2	96.83	15.08	3.97	$+0.0047$	-0.0008
Seq. LoRA / O-LoRA	o0	100.00	100.00	0.00	$+0.0363$	$+0.0480$
Seq. LoRA / O-LoRA	o1	100.00	100.00	0.00	$+0.0294$	$+0.0293$
Seq. LoRA / O-LoRA	o2	100.00	57.94	0.79	$+0.0250$	$+0.0088$
Seq. LoRA / Replay	o0	100.00	100.00	0.00	$+0.0819$	$+0.1147$
Seq. LoRA / Replay	o1	100.00	100.00	0.00	$+0.0713$	$+0.0647$
Seq. LoRA / Replay	o2	100.00	100.00	0.00	$+0.0609$	$+0.0482$

A different winner is a different question. The primary rate does not require agreement with the historical mean sign. A supplemental cross-cohort diagnostic finds 3 opposite mean signs out of 18, all for sequential LoRA versus EWC, one on each order. These small differences are not significant superiority claims. The pooled primary mean-forgetting ordering in this cohort is Replay, O-LoRA, EWC, Sequential LoRA; it does not preserve the historical sequential-LoRA/EWC ordering. This leaves the distinct expanded-evaluation (R3) pooled-ranking finding intact while showing its population boundary.

Secondary adaptive baseline and descriptive metrics. Adaptive AdaLoRA uses the allocation callback after every optimizer step, completing rank allocation during the first 85% of updates and then fixing the mask. Each of its 30 runs records the allocation steps and a nonempty final mask within the target rank budget. Masking does not reduce allocated tensor storage. These are adaptive results, unlike the historical fixed-rank SVD arm. The table compares recorded implementations under the common recipe; it is not a separately tuned or equal-compute ranking, and an old/new AdaLoRA

comparison cannot isolate the allocation callback from cohort and execution changes. Table 24 reports every method; CSV T11 adds each order and CSV T12 records the complete summary across k. Standard deviation (SD) is calculated over the 30 method–seed–order cells, combining order and seed variation, not a standard error from 30 independent studies. Same-host sensitivity is CSV T13.

Table 24: New-cohort descriptive metrics, mean ± sample SD over all three orders and ten seeds per method. Values are accuracy fractions. AF: signed forgetting; AA: final average accuracy; LA: learning accuracy, the mean accuracy on each task immediately after its own training stage. Adaptive AdaLoRA is secondary and excluded from the historically matched four-method primary estimand.

Implementation	Runs	AF ↓	AA ↑	LA ↑
Sequential LoRA	30	0.0988 ± 0.0391	0.7471 ± 0.0303	0.8234 ± 0.0139
EWC	30	0.0971 ± 0.0404	0.7489 ± 0.0324	0.8236 ± 0.0132
Replay	30	0.0230 ± 0.0103	0.8135 ± 0.0160	0.8233 ± 0.0108
O-LoRA	30	0.0701 ± 0.0316	0.7717 ± 0.0252	0.8245 ± 0.0111
Adaptive AdaLoRA	30	0.2009 ± 0.0471	0.6564 ± 0.0323	0.8164 ± 0.0148

Reproduction and interpretation. The release separates these records under `confirmation/records/` and retains the frozen campaign, its original complete analysis and the amended assignment. `scripts/confirmation_analysis.py` validates the entire grid and independently recomputes the frozen endpoints from raw matrices, requiring agreement with the saved remote result; `scripts/make_confirmation_artifacts.py` regenerates this text, tables, CSVs and figure. Export metadata records original and anonymized hashes and the exact redactions, retaining numerical outcomes. The source/export map discloses anonymizing comment changes instead of asserting byte identity of edited exports. This follow-up supplies evidence of limited transport under a recorded recipe and execution mixture, including failed predictions. It is not independent-team or new-domain validation and does not repair missing historical identities.

O Secondary directional decisions and paired evaluation resampling

Analysis status and estimands. These analyses were chosen after the fresh outcomes were available. They do not change the frozen within-cohort agreement endpoint, its 95% criterion, any execution identity, or any original record. The directional analysis asks whether a historical subset's *declared direction* agrees with the fresh cohort's mean sign. The replay stratification asks how the retrospective reporting average depends on its comparison population. The evaluation bootstrap asks how a paired metric changes when the same recorded examples are resampled across methods and stages, with all trainings fixed. These are three distinct questions.

Replay stratification and weights. Table 25 partitions the previously reported historical core-four audit. Replay comparisons supply 80.44% of paired-t declarations while constituting half its opportunities. The lower aggregate opposite-sign rate persists after excluding replay, with a larger abstention burden. Opposite rates weight cells by declarations; they are not equal-weight averages of pair–order rates. All groups and all $k = 1, \ldots, 5$ appear in CSV T17. The main text's adverse pair illustrates that the aggregate ordering need not hold for each comparison; stratification does not correct an arithmetic error or create an error-rate guarantee.

Directional historical-to-fresh reference. For each of the six pairs of frozen primary implementations and three original orders, enumerate every historical k-seed subset from seeds 0–9. Apply the same mean-sign, nonoverlapping one-SD-band, or nominal paired-t rule already defined in Appendix M. Compare each declaration with the sign of the mean paired difference across all fresh seeds 100–109 for that pair and order. No fresh outcome selects a decision threshold. A difference within 10^{-12} is a tie; an undeclared decision abstains, and a tied reference is counted separately. At $k = 3$ there are $18\binom{10}{3} = 2,160$ opportunities. At other k the denominator is $18\binom{10}{k}$, not the historical audit's ordered disjoint-group denominator. The reference is reused across overlapping decisions and

Table 25: Historical Qwen core-four forgetting, three decision and three disjoint reference seeds. All five orders; 30 pair–order comparisons, 4,200 opportunities each. Opposite is conditional on declaring. Replay strata partition opportunities; rates are dependent finite-corpus descriptions.

Stratum	Rule	Declarations	Declare (%)	Opposite (%)
All pairs	Mean sign	125,055/126,000	99.25	6.84
All pairs	Separate SD bands	73,220/126,000	58.11	1.10
All pairs	Paired t interval	44,730/126,000	35.50	0.95
Replay pairs	Mean sign	63,000/63,000	100.00	0.00
Replay pairs	Separate SD bands	56,035/63,000	88.94	0.00
Replay pairs	Paired t interval	35,980/63,000	57.11	0.00
Other pairs	Mean sign	62,055/63,000	98.50	13.78
Other pairs	Separate SD bands	17,185/63,000	27.28	4.69
Other pairs	Paired t interval	8,750/63,000	13.89	4.87

is not population truth. No binomial interval, Type-I rate or confidence-coverage interpretation is appropriate. Table 26 reports the declaration and conditional opposition rates at $k = 3$.

Table 26: Secondary directional historical-to-fresh comparison at three historical decision seeds. Orders o0–o2 only: 18 pair–order comparisons, 120 historical subsets each; one full ten-seed fresh reference is reused per pair–order. Opposite is conditional on declaring. This changes both reference size and order population from Table 25; differences between tables do not isolate a cohort effect.

Stratum	Rule	Declarations	Declare (%)	Opposite (%)
All pairs	Mean sign	2,142/2,160	99.17	15.31
All pairs	Separate SD bands	1,207/2,160	55.88	0.66
All pairs	Paired t interval	755/2,160	34.95	0.79
Replay pairs	Mean sign	1,080/1,080	100.00	0.00
Replay pairs	Separate SD bands	922/1,080	85.37	0.00
Replay pairs	Paired t interval	591/1,080	54.72	0.00
Other pairs	Mean sign	1,062/1,080	98.33	30.89
Other pairs	Separate SD bands	285/1,080	26.39	2.81
Other pairs	Paired t interval	164/1,080	15.19	3.66

Table 27 retains all 18 references. 5 fresh nominal intervals include zero, including all three sequential-LoRA/EWC comparisons whose full-cohort mean signs changed. The intervals are unadjusted and assumption-dependent; they describe reference uncertainty, without suppressing those references or selecting only significant ones. CSVs T14–T16 retain every rule, group size, pair–order, tie and abstention. The operational value is a more explicit directional stress test, not prospective validation of a rule selected later. Historical provenance gaps and amended host allocation remain: this comparison does not isolate seeds from execution changes.

Expanded-evaluation (R3) cross-method identity checks. All 75 R3 prediction files contain task names, evaluation counts and correctness vectors, but 75 lack evaluation example hashes and 75 manifests have empty dataset-hash maps. Thus 150 of 150 matched pair–seed–order cells cannot establish ordered cross-method example identity. Equal seeds and equal zero-shot correctness vectors do not identify the examples. We therefore withhold paired R3 resampling numbers. CSVs T19–T20 expose the entire identity audit. This does not promote the earlier within-run stage-resampling calculation into a cross-method comparison.

Paired bootstrap where identities are recorded. The 120 primary fresh trajectories record ordered rendered-evaluation hashes in both the prediction file and manifest; the hashes agree across methods sharing a seed and order. For each of 30 such groups, we draw 2,000 resamples of 250 examples *with replacement* from each task's recorded 250-example empirical pool. Every method and stage

Table 27: Every directional reference. Differences are first minus second mean forgetting; lower is better. Differences and intervals are accuracy fractions. The nominal unadjusted paired-t interval summarizes fresh seeds and is not a certified truth label.

Pair	Order	Historical Δ	Fresh Δ	Fresh nominal interval
Seq. LoRA / EWC	o0	-0.0059	$+0.0005$	$[-0.0093, +0.0103]$
Seq. LoRA / EWC	o1	-0.0038	$+0.0054$	$[-0.0015, +0.0123]$
Seq. LoRA / EWC	o2	$+0.0047$	-0.0008	$[-0.0067, +0.0051]$
Seq. LoRA / Replay	o0	$+0.0819$	$+0.1147$	$[+0.0871, +0.1423]$
Seq. LoRA / Replay	o1	$+0.0713$	$+0.0647$	$[+0.0443, +0.0851]$
Seq. LoRA / Replay	o2	$+0.0609$	$+0.0482$	$[+0.0272, +0.0692]$
Seq. LoRA / O-LoRA	o0	$+0.0363$	$+0.0480$	$[+0.0088, +0.0872]$
Seq. LoRA / O-LoRA	o1	$+0.0294$	$+0.0293$	$[+0.0132, +0.0454]$
Seq. LoRA / O-LoRA	o2	$+0.0250$	$+0.0088$	$[-0.0132, +0.0308]$
EWC / Replay	o0	$+0.0878$	$+0.1142$	$[+0.0870, +0.1414]$
EWC / Replay	o1	$+0.0751$	$+0.0593$	$[+0.0387, +0.0799]$
EWC / Replay	o2	$+0.0562$	$+0.0490$	$[+0.0241, +0.0739]$
EWC / O-LoRA	o0	$+0.0422$	$+0.0475$	$[+0.0124, +0.0826]$
EWC / O-LoRA	o1	$+0.0332$	$+0.0239$	$[+0.0061, +0.0417]$
EWC / O-LoRA	o2	$+0.0203$	$+0.0096$	$[-0.0136, +0.0328]$
Replay / O-LoRA	o0	-0.0456	-0.0667	$[-0.1004, -0.0330]$
Replay / O-LoRA	o1	-0.0419	-0.0354	$[-0.0450, -0.0258]$
Replay / O-LoRA	o2	-0.0359	-0.0394	$[-0.0540, -0.0248]$

receives exactly the same indices within a draw. The max-before-final forgetting estimator is recomputed each time, and then differenced between methods. This preserves cross-method and cross-stage dependence. It is an empirical-pool bootstrap with training fixed, *not* 250-from-7,600 subsampling, unseen-test uncertainty, training variance, or a confidence interval on the population method difference.

Across 180 paired cells, the median paired standard deviation (SD), in accuracy fractions, is 0.0131; the median ratio to the independent-marginal construction is 0.853. Paired SD is lower in 179 cells. The comparator discards cross-method covariance by taking the square root of the sum of the two marginal resampling variances. Its discrepancy demonstrates why individual-arm sampling SDs cannot simply stand in for uncertainty of a difference. The 1 recorded identical-prediction pair has exactly zero paired SD under this conditional calculation; this is not evidence of zero retraining variability. Table 28 summarizes each pair; CSV T18 retains all cells, resampling quantiles, sign fractions and ties. Quantiles describe this resampling distribution and are not presented as coverage-validated intervals.

Table 28: Conditional empirical-pool bootstrap across the 30 fresh seed–order cells for each pair. SD is the standard deviation of the paired forgetting difference under aligned example resampling, in accuracy fractions. The ratio compares paired SD to the square root of the sum of marginal variances; medians summarize cells, not uncertainty across independent studies.

Pair	Cells	Median paired SD	SD range	Median SD ratio
Seq. LoRA / EWC	30	0.0059	0.0000–0.0111	0.344
Seq. LoRA / Replay	30	0.0138	0.0115–0.0155	0.925
Seq. LoRA / O-LoRA	30	0.0129	0.0103–0.0148	0.773
EWC / Replay	30	0.0138	0.0116–0.0157	0.926
EWC / O-LoRA	30	0.0132	0.0104–0.0148	0.786
Replay / O-LoRA	30	0.0130	0.0111–0.0148	0.902

Reproduction and failure behavior. `scripts/panel_followup_analysis.py` validates the complete original cohort, hashes every input, reproduces historical metrics, and refuses cross-

method resampling without matching recorded ordered identities. It uses deterministic per-group random seeds; `--check` recomputes the saved report without writing. `scripts/make_panel_followup_artifacts.py` generates these tables and CSVs from that report. The anonymous export retains original/exported identity receipts; the same commands run in its portable directory layout. Tests cover reversed cross-cohort winners, ties, unavailable variance rules, exact method cancellation, stage dependence, changing max stages and missing/mismatched identities.

P SCORER AND HOST DIAGNOSTIC

Protocol. This diagnostic uses seeds 200, 201, 202 on order o2, for sequential LoRA, LoRA+EWC and O-LoRA on both hosts. At the base model and after each task, the scorers evaluate identical learned states and ordered inputs. For candidate token lengths m_c, original scoring retains the rightmost $192 - m_c$ prompt tokens. Common-prompt scoring retains the same rightmost $192 - \max_c m_c$ prompt tokens for every candidate (the entire prompt when shorter). Both use mean verbalizer-token log likelihood without the end-of-sequence (EOS) token. Training is unchanged, and evaluation restores the training RNG state and model mode. Table 29 reports the paired scorer results; Table 30 reports the matched host differences.

The complete diagnostic contains 18 trajectories on one selected order. The conditioning change affects 148/3750 sampled example entries (3.95%), counting each seed/task once. Across final-stage evaluations, 17/22500 predictions change; 0 of these changes occur on inputs whose retained prompt is unchanged. These repeated evaluations share inputs across methods and hosts and are not independent examples.

Within-trajectory common-minus-original changes range from -0.160 to +0.080 percentage points in final average accuracy (AA) and -0.100 to +0.200 percentage points in signed forgetting (AF). Among all per-seed method-pair comparisons, the scorer change strictly reverses 0/18 AA signs and 0/18 AF signs; separate tie transitions number 0 and 0, respectively. For method differences averaged over seeds within host, the corresponding strict reversals are 0/6 and 0/6. All comparisons, including ties and host changes, are reported in T24–T25 and T28.

Across hosts, the recorded method difference strictly reverses in 2/18 per-seed AA comparisons and 2/18 per-seed AF comparisons, pooling both scorer policies; tie transitions number 0 and 0, respectively. After averaging each method over the three seeds, the corresponding strict reversals are 2/6 and 2/6. These comparisons reuse the same paired trajectories under both scorer policies; they do not identify which environmental difference caused the observed changes.

T28 distinguishes four counting scopes. Scorer changes have 18 per-seed comparisons (three method pairs $\times$ three seeds $\times$ two hosts) and six comparisons after averaging over seeds (three pairs $\times$ two hosts). Host changes pool two scorer policies: their 18 per-seed comparisons are three method pairs $\times$ three seeds $\times$ two scorers, and their six mean comparisons are three pairs $\times$ two scorers. Both policies reuse the same nine method/seed host pairs; these counts do not represent independent host replications. CSV T25 retains the policy for every individual host comparison.

The portable package contains full input and score records and checkpoint indices. Local validation reconstructs predictions, both matrices and all metrics from those scores. Remote worker receipts report strict verification of the checkpoint bytes; the portable package omits the large tensors and does not claim local verification or inference replay of those absent weights. Token conditioning is checked against recorded token IDs; the offline analysis does not retokenize inputs. Metadata-only redactions have explicit original/exported hashes. This diagnostic does not recover missing example identities in the historical expanded-evaluation study (R3) or establish universal equivalence.

Checkpoint reload verification. Separate engineering forward re-evaluations cover 3 final scientific checkpoints on Cluster (Seq. LoRA, LoRA+EWC, O-LoRA; seed 200), covering 52,500 candidate scores across both scorers; and 3 final engineering-pilot checkpoints on Workstation (Seq. LoRA, LoRA+EWC, O-LoRA; seed 200), covering 840 candidate scores across both scorers. All compared candidate scores were exactly equal and all predicted labels matched. The portable validator links each receipt to the reload-script hash, original result hash, final-stage checkpoint-index hash and expected candidate-score count. These checks cover only the listed final checkpoints, not every stage

or all 18 scientific trajectories; engineering pilots are not additional scientific replicates. The portable validation itself does not rerun the forward passes.

Table 29: Scorer diagnostic on order o2. Means over three matched seeds per host and method. AA (final average accuracy) and AF (signed forgetting) levels are percentages; changes (Δ) are percentage points. Orig.: frozen candidate-dependent conditioning; common: common retained prompt, with the same mean-token log-likelihood normalization. Δ is common minus original on identical learned states. AF uses the signed forgetting definition of the main analysis. Per-cell values and all stage predictions are retained; this selected-order diagnostic is descriptive.

Host	Method	AA orig.	AA common	ΔAA	AF orig.	AF common	ΔAF
Workstation	Seq. LoRA	76.667	76.640	-0.027	8.733	8.767	+0.033
Workstation	LoRA+EWC	76.800	76.800	0.000	8.533	8.533	0.000
Workstation	O-LoRA	78.880	78.853	-0.027	6.733	6.767	+0.033
Cluster	Seq. LoRA	75.920	75.893	-0.027	9.467	9.500	+0.033
Cluster	LoRA+EWC	75.760	75.733	-0.027	9.667	9.700	+0.033
Cluster	O-LoRA	78.880	78.853	-0.027	6.733	6.767	+0.033

Table 30: Matched host differences in final average accuracy (AA) and signed forgetting (AF): cluster minus workstation, averaged over three seeds (percentage points). The two host environments also differ in Python micro-version, hardware, background load and concurrency; these differences do not isolate hardware causality. All individual seed differences and method-difference changes appear in supplementary CSVs T23–T25.

Method	Δ_HAA orig.	Δ_HAA common	Δ_HAF orig.	Δ_HAF common
Seq. LoRA	-0.747	-0.747	+0.733	+0.733
LoRA+EWC	-1.040	-1.067	+1.133	+1.167
O-LoRA	0.000	0.000	0.000	0.000

Q PROSPECTIVE FIXED-HOST COMPARISON STUDY

Fixed-host execution confirmation. The complete study contains 60 trajectories: three recorded implementations, ten fresh seed/data blocks (300–309), and two fixed execution environments on order o2 (Yelp, AG News, Yahoo, Amazon polarity, DBpedia). Common-prompt scoring is primary. The sample unit is the aligned seed/data block, which changes training rows, evaluation rows and initialization together. First-minus-second method contrasts and their paired cluster-minus-workstation changes are in percentage points. Within a host, positive signed-forgetting (AF) differences favor the second method and positive final-average-accuracy (AA) differences favor the first. A positive host change means this first-minus-second contrast increases on the cluster. AF averages, over the first four tasks, the maximum accuracy across the four pre-final training stages minus final accuracy; it can be negative and is not clipped. The zero-shot baseline is excluded from this maximum. AA averages final accuracy across all five tasks.

The single family of 18 two-sided zero-mean tests has 0 Holm rejections at family $\alpha = .05$; 0 effects have degenerate variance and receive no t interval or test p-value. Non-rejection does not establish equivalence. Figure 4 and the complete effect tables report nominal 95% and Bonferroni simultaneous 95% t intervals, conditional on independent seed blocks and approximately normal paired differences.

Descriptive common-prompt AF agreement is retained for every $k = 1, \ldots, 5$, including every failed criterion. At $k = 5$, 0/3 pairs on the workstation and 1/3 on the cluster meet the at-least-95% criterion; 0/3 meet it on both hosts. The denominators enumerate dependent partitions of finite cohorts; agreement is neither preservation of a historical winner nor a future-replication probability. The secondary original scorer evaluates the same learned states and receives no additional confirmatory tests.

These contrasts concern one fixed task order and two recorded execution environments. They cannot isolate hardware causation, support a universal seed recommendation, or redefine the historical original-scorer 13/16 endpoint. The source analysis reconstructs saved correctness and binds full worker validation receipts; it does not replay local forward passes or verify checkpoint tensors absent locally.

Protocol and admission. The run list, source bundle and analysis protocol were frozen before study outcomes. Every planned coordinate completed the full stream under the unchanged diagnostic training recipe; engineering pilots and historical runs are excluded. Ordered training, evaluation and prepared-token identities match across methods and hosts within each seed block. Both hosts execute every block; no cell is migrated. The comparison targets these recorded implementations under shared hyperparameters, not source-optimal algorithms or equal compute. The primary common-prompt scorer reserves the longest verbalizer's token budget for every candidate; mean label-token log-likelihood and the learned states are shared with the secondary original scorer.

Complete results. Tables 31 and 32 report all primary AF and AA effects; Figure 4 displays both interval families. Tables 33 and 34 retain every host, pair and group size, including failures. Failure of the criterion within fixed hosts shows that mixing hosts is not the sole explanation for the observed inconsistency. It does not attribute that inconsistency to training or evaluation sampling separately. Tables 35 and 36 retain the original-scorer effects and whole-block leave-one-out descriptions. Supplementary CSVs ES01–ES13 retain seed vectors, primary influence summaries, per-trajectory scorer changes, host sign comparisons and scorer sensitivities; policies reuse trajectories, not independent execution replicates. No equivalence margin or test was registered.

Portable evidence. The `execution_confirmation/` namespace preserves all scientific score and correctness records, data identities, stage-state indexes and result-linked remote validation receipts. Large checkpoint tensors remain remote. Rebuilding `stats/execution_confirmation.json` checks the retained records and their receipt linkage, not absent tensor bytes or new model forwards. The frozen analysis and presentation generator have separate recomputation checks; metadata redactions are explicit and leave benchmark content and score bytes unchanged. This population is separate from the earlier transfer cohort and scorer/host diagnostic, including their checkpoint-reload engineering checks.

Table 31: Signed forgetting (AF) contrasts under the primary common-prompt scorer, in percentage points. Seq denotes Sequential LoRA, EWC denotes EWC-style LoRA, and O denotes the O-LoRA implementation. Each row uses ten aligned seed/data blocks; degrees of freedom $df = 9$. First method minus second; host change is the seed-paired cluster contrast minus workstation contrast. All 18 AF/AA effects share one Holm family at $\alpha = .05$; the simultaneous intervals use Bonferroni over the same 18 effects. Tests and intervals depend on the prespecified t-reference assumptions. Non-rejection does not establish equivalence; unavailable rows have degenerate variance.

Context	Mean	Nominal 95%	Family 95%	Holm p	Zero-mean null
First - second: Seq - EWC					
Workstation	0.060	[-0.379, 0.499]	[-0.730, 0.850]	1.000	do not reject
Cluster	-0.240	[-1.204, 0.724]	[-1.977, 1.497]	1.000	do not reject
Cluster - workstation	-0.300	[-1.203, 0.603]	[-1.927, 1.327]	1.000	do not reject
First - second: Seq - O					
Workstation	0.960	[-0.949, 2.869]	[-2.480, 4.400]	1.000	do not reject
Cluster	1.210	[-0.249, 2.669]	[-1.418, 3.838]	1.000	do not reject
Cluster - workstation	0.250	[-1.115, 1.615]	[-2.209, 2.709]	1.000	do not reject
First - second: EWC - O					
Workstation	0.900	[-0.713, 2.513]	[-2.006, 3.806]	1.000	do not reject
Cluster	1.450	[0.056, 2.844]	[-1.062, 3.962]	0.690	do not reject
Cluster - workstation	0.550	[-0.665, 1.765]	[-1.638, 2.738]	1.000	do not reject

Table 32: Final average accuracy (AA) contrasts under the primary common-prompt scorer, in percentage points. Seq denotes Sequential LoRA, EWC denotes EWC-style LoRA, and O denotes the O-LoRA implementation. Each row uses ten aligned seed/data blocks; degrees of freedom $df = 9$. First method minus second; host change is the seed-paired cluster contrast minus workstation contrast. All 18 AF/AA effects share one Holm family at $\alpha = .05$; the simultaneous intervals use Bonferroni over the same 18 effects. Tests and intervals depend on the prespecified t-reference assumptions. Non-rejection does not establish equivalence; unavailable rows have degenerate variance.

Context	Mean	Nominal 95%	Family 95%	Holm p	Zero-mean null
First - second: Seq - EWC					
Workstation	-0.136	[-0.519, 0.247]	[-0.827, 0.555]	1.000	do not reject
Cluster	0.224	[-0.762, 1.210]	[-1.552, 2.000]	1.000	do not reject
Cluster - workstation	0.360	[-0.759, 1.479]	[-1.656, 2.376]	1.000	do not reject
First - second: Seq - O					
Workstation	-1.168	[-2.509, 0.173]	[-3.583, 1.247]	1.000	do not reject
Cluster	-1.400	[-2.705, -0.095]	[-3.751, 0.951]	0.649	do not reject
Cluster - workstation	-0.232	[-1.358, 0.894]	[-2.261, 1.797]	1.000	do not reject
First - second: EWC - O					
Workstation	-1.032	[-2.284, 0.220]	[-3.288, 1.224]	1.000	do not reject
Cluster	-1.624	[-2.880, -0.368]	[-3.887, 0.639]	0.305	do not reject
Cluster - workstation	-0.592	[-1.638, 0.454]	[-2.477, 1.293]	1.000	do not reject

Table 33: Descriptive common-prompt signed-forgetting (AF) agreement (%) in each fixed host cohort. Each cell enumerates two disjoint k-seed groups from ten seed/data blocks. Agreement requires the same nonzero sign; ties in either group are unresolved. Ordered-pair denominators for $k = 1, \ldots, 5$ are 90, 1,260, 4,200, 3,150 and 252. The at-least-95% criterion is descriptive. These overlapping partitions are dependent; rates have no binomial intervals and are not future-replication probabilities.

Host	First - second	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
Workstation	Seq. LoRA - EWC-style LoRA	35.56	49.05	37.86	22.35	10.32
Workstation	Seq. LoRA - O-LoRA impl.	64.44	51.11	53.43	62.86	73.02
Workstation	EWC-style LoRA - O-LoRA impl.	48.89	54.60	57.57	66.67	76.19
Cluster	Seq. LoRA - EWC-style LoRA	35.56	39.21	35.29	32.38	30.16
Cluster	Seq. LoRA - O-LoRA impl.	64.44	62.86	76.67	85.71	92.86
Cluster	EWC-style LoRA - O-LoRA impl.	53.33	74.29	91.67	96.19	97.62

Table 34: Pairs meeting the descriptive common-prompt signed-forgetting (AF) agreement criterion, out of three per host. The intersection counts the same pair meeting it on both hosts. Hosts share the ten seed/data blocks, so the intersection is not two independent replication tests.

k	Workstation / 3	Cluster / 3	Both hosts / 3
1	0	0	0
2	0	0	0
3	0	0	0
4	0	1	0
5	0	1	0

Table 35: Signed forgetting (AF) under the secondary original candidate-dependent scorer. Seq denotes Sequential LoRA, EWC denotes EWC-style LoRA, and O denotes the O-LoRA implementation. Values are percentage points; differences and pairing match the primary analysis. Standard deviation (SD) is across the ten paired seed differences. Leave-one-out (LOO) is the range after omitting one whole seed block; sign-category changes compare those nine-block means with the full ten-block mean, including changes to unresolved ties ($|\Delta| \leq 10^{-12}$ in accuracy units). They need not be opposite winners. This same-state scorer sensitivity has no additional confirmatory tests or independent replication units.

Context	Mean	Sample SD	LOO mean range	LOO category changes
First - second: Seq - EWC				
Workstation	0.070	0.613	[0.000, 0.244]	2/10
Cluster	-0.230	1.321	[-0.367, 0.122]	1/10
Cluster - workstation	-0.300	1.262	[-0.422, 0.056]	1/10
First - second: Seq - O				
Workstation	1.000	2.700	[0.656, 1.611]	0/10
Cluster	1.220	2.012	[0.822, 1.589]	0/10
Cluster - workstation	0.220	1.911	[-0.022, 0.578]	2/10
First - second: EWC - O				
Workstation	0.930	2.277	[0.600, 1.367]	0/10
Cluster	1.450	1.885	[1.100, 1.700]	0/10
Cluster - workstation	0.520	1.667	[0.289, 0.856]	0/10

Table 36: Final average accuracy (AA) under the secondary original candidate-dependent scorer. Seq denotes Sequential LoRA, EWC denotes EWC-style LoRA, and O denotes the O-LoRA implementation. Values are percentage points; differences and pairing match the primary analysis. Standard deviation (SD) is across the ten paired seed differences. Leave-one-out (LOO) is the range after omitting one whole seed block; sign-category changes compare those nine-block means with the full ten-block mean, including changes to unresolved ties ($|\Delta| \leq 10^{-12}$ in accuracy units). They need not be opposite winners. This same-state scorer sensitivity has no additional confirmatory tests or independent replication units.

Context	Mean	Sample SD	LOO mean range	LOO category changes
First - second: Seq - EWC				
Workstation	-0.144	0.549	[-0.284, -0.053]	0/10
Cluster	0.208	1.333	[-0.160, 0.373]	1/10
Cluster - workstation	0.352	1.540	[-0.107, 0.516]	1/10
First - second: Seq - O				
Workstation	-1.200	1.899	[-1.484, -0.844]	0/10
Cluster	-1.416	1.788	[-1.698, -1.022]	0/10
Cluster - workstation	-0.216	1.589	[-0.462, -0.027]	0/10
First - second: EWC - O				
Workstation	-1.056	1.762	[-1.227, -0.702]	0/10
Cluster	-1.624	1.711	[-1.893, -1.262]	0/10
Cluster - workstation	-0.568	1.447	[-0.827, -0.338]	0/10

R ARCHIVAL FRESHNESS AND ELIGIBILITY AUDIT

A separately recorded snapshot census covers 2,023 remote result files, compared by byte, canonical-JSON and recorded-outcome signatures with 2,251 pre-sprint local records (Table 37). The outcome signature contains method, model, ordered tasks, seed, R and R_{rand}; it establishes new saved content, not independent execution or complete training lineage. The 105 records with signatures absent from the baseline contain 102 distinct outcome signatures. The release archives every one of the 112 distinct new payloads, including unfavorable outcomes, with original/exported hash receipts. All archived matrices and four reported performance metrics pass independent recomputation; 32 payloads also retain correctness vectors. Optional diagnostic NaNs are counted separately from parse failures and do not invalidate finite performance matrices. Malformed-file hashes and parse reasons are released; their bytes remain outside numerical analysis.

Table 37: Content freshness is distinct from scientific eligibility. Categories overlap: nonfinite optional diagnostics can occur in otherwise usable historical records.

Snapshot category	Files
New canonical payloads	112
Records with new outcome signatures	105
New payloads repeating existing outcomes	7
Malformed result files	57
Nonfinite optional-diagnostic records	275

Among the new payloads, 70 are historical secondary candidates, 5 use the excluded prompt port, and 6 contain alternate outcomes at occupied frozen coordinates. Separately inventoried diagnostic outputs include 18 candidate scientific trajectories and 6 engineering pilots; these do not enter historical sensitivity comparisons. Scientific diagnostic admission requires its separate complete-grid validation. The alternatives may be additional executions of the same nominal configuration; missing source/data identities prevent an independence or isolated-host-effect claim. They are retained as replicate candidates without replacing the frozen populations. Availability of all planned coordinates is reported separately in CSV T30: a completed one-order sensitivity grid does not establish cross-order generality, and incomplete larger campaigns remain incomplete. The portable `fresh_inventory.py` verifies archive hashes, all new saved matrices, freshness counts against baseline signature receipts, and coordinate-count metadata. It does not reverify historical training. The snapshot cutoffs and every missing coordinate are retained in `freshness/index.json` and `stats/fresh_inventory.json`.

S RECOVERED ALTERNATIVE HISTORICAL OUTCOMES

An inventory recovered six different outcomes at already-filled historical coordinates (Table 38). Their saved method, backbone, ordered tasks, seed and hyperparameters match the original records; all four saved performance metrics reproduce their finite accuracy matrices. This establishes nominal configuration matching, not identical source, sampled rows, model revision, or independent retraining. Four original routed records lack manifests; all six comparisons lack sufficient execution/source/data lineage. The two historical `adalora` records remain fixed-rank SVD-LoRA, not adaptive AdaLoRA. The original records and the separately frozen 150-trajectory cohort remain unchanged.

Expanded discrepancy sensitivity. Including the six nominally matched candidate comparisons would extend the historical 0.095 maximum to 0.101. This is an expanded candidate set, not an attribution to hardware or proof of independent training. Candidate arm maxima become 0.058 for R-raw28 and 0.101 for R+g+c; the two SVD-LoRA differences are each about 6.58×10^{-5}. The zero discrepancy in the original SVD replicate set therefore cannot be generalized. The core-baseline reversing margins remain below their observed discrepancies. The JSON report retains both historical and expanded per-pair/global sensitivity counts for every hypothetical population; neither criterion is a calibrated error bound.

Table 38: All six alternatives. Forgetting is in accuracy units; smaller is better. Codes identify hypothetical substitutions below. s and o denote seed and task-order indices.

Code	Coordinate	Original	Alternative	Difference
A	R-raw28, s1/o2	0.26500000	0.20700000	-0.05800000
B	R+g+c, s1/o4	0.19200000	0.19600000	+0.00400000
C	R+g+c, s2/o3	0.03400000	0.09600000	+0.06200000
D	R+g+c, s2/o4	0.17100000	0.27200000	+0.10100000
E	SVD-LoRA (R3), s0/o4	0.24019737	0.24026316	+0.00006579
F	SVD-LoRA (R3), s1/o3	0.15269737	0.15276316	+0.00006579

All substitutions, without selecting a population. We enumerate all $2^4 = 16$ choices of original/alternative routed records and all $2^2 = 4$ choices for R3, always preserving the seed labels and number of coordinates. Table 39 includes the unchanged historical grid (none). Table 40 reports all four R3 choices. These are possible reanalyses, not additional experiments or independent replications. The conventional-versus-whitened single-machine comparison, historical core-five ten-seed analysis, task-replacement streams and frozen-cohort endpoint use different records and are unchanged. No substitution supports a beneficial plasticity gate. The historical ten-arm rejection is sensitive to candidate choice: it is lost in 8 of the 16 hypothetical grids. With all four alternatives its exact Holm-adjusted p is 0.1592, versus 0.0342 for the unchanged grid. These counts describe the enumerated alternatives and are not a probability of replication.

Table 39: All routed substitutions. Reversals are out of 45 (ten arms) or 78 (all thirteen). AF and swing refer to R+g+c. Adjusted p values use the exact block-preserving null and Holm correction over the same six pools within each hypothetical population.

Selected	Ten arms	All thirteen	AF	Swing	p_H ten	p_H thirteen
none	19	33	0.0943	5.07	0.0342	0.1500
D	20	34	0.1010	5.94	0.0342	0.0726
C	18	31	0.0984	3.30	0.3201	0.5138
CD	19	32	0.1051	3.87	0.1579	0.2933
B	19	33	0.0945	5.10	0.0342	0.1500
BD	20	34	0.1013	5.97	0.0342	0.0726
BC	18	31	0.0987	3.33	0.3208	0.5158
BCD	19	32	0.1054	3.89	0.1579	0.2933
A	19	33	0.0943	5.07	0.0342	0.1500
AD	20	34	0.1010	5.94	0.0342	0.0740
AC	18	31	0.0984	3.30	0.3264	0.5133
ACD	19	32	0.1051	3.87	0.1592	0.2933
AB	19	33	0.0945	5.10	0.0342	0.1500
ABD	20	34	0.1013	5.97	0.0342	0.0740
ABC	18	31	0.0987	3.33	0.3271	0.5154
ABCD	19	32	0.1054	3.89	0.1592	0.2933

Null calculation and limits. The reversal statistic is unchanged by a common relabeling of output orders. Fixing one seed's permutation therefore leaves $(5!)^2 = 14{,}400$ equally weighted relative permutations, which we enumerate exactly. This is the same block-preserving null as the historical Monte Carlo test; small numerical differences in the unchanged row arise from exact enumeration. Its exchangeability assumptions still apply. We recompute its complete six-pool Holm family for every hypothetical grid; the independent-arm permutation family is not recalculated here. These after-outcome substitutions establish sensitivity, not a new confirmatory family or causal mechanism.

Reproduction. The anonymous `freshness/index.json` maps neutral IDs to the alternative records and original counterpart paths; both versions are retained under `freshness/records/` and `freshness/counterparts/`. Codes A–F map respec-

Table 40: All expanded-evaluation (R3) substitutions. The pooled order remains replay, O-LoRA, sequential LoRA, EWC, SVD-LoRA. Agreement is with the historical small-evaluation grid. Changed audit rows compare every metric, pair, order, group size and reporting rule with original R3.

Selected	Fixed-rank SVD-LoRA AF	Reversals / 10	Agreement / 50	Changed audit rows
none	0.18346930	1	45	0
F	0.18347368	1	45	0
E	0.18347368	1	45	0
EF	0.18347807	1	45	0

tively to `f_585e00d520d9add4`, `f_44baa51e5abf2dbc`, `f_a23151404f49395a`, `f_2c005d195b73d680`, `f_c12df766d529ecc7`, `f_4fc38c1a6bf6af7a`. `scripts/fresh_inventory_sensitivity.py` validates the six-coordinate population, reconstructs the metrics and enumerates every substitution; `--check` recomputes the report and verifies the text and CSV T31 without writing. The numerical report retains all null distributions' tail counts, family adjustments, rankings, per-arm discrepancies and changed reporting-row identifiers.

T OPERATIONAL REPORTING CHECKLIST

The checklist separates essential records from diagnostics selected for a claim and budget. **Minimum reporting** comprises run/configuration identity, exact data and checkpoint-to-evaluation linkage, the metric definition, matched-condition paired differences with uncertainty, and every declaration or abstention. Use these records before drawing performance conclusions. **Claim-dependent diagnostics** then ask whether the intended claim needs order coverage, repeat execution, conditional evaluation resampling or a fresh-cohort check. The study's full campaign is not a minimum compute requirement; there is no universal seed, order or host count. The following items retain the evidence motivating each choice:

1. **Coverage heuristic: consider every task in the leading position.** This requires at least as many orders as tasks, five here, but does not balance later positions or guarantee adequate coverage. The reversal rate has not saturated at ten orders (Fig. 10b); three orders leave two tasks never first, where our largest effect lived.
2. **Essential: matched-condition paired comparisons** (same data subsample and order per pair), reporting winner changes alongside any mean, on order-means and within seed; under-seeded grids overstate the former (3/10 $\rightarrow$ 1/10 from three to six seeds).
3. **For inferential counts: give a null model and multiplicity family.** In the frozen historical grid, the ten-arm pool exceeds both nulls uncorrected and SmolLM2 exceeds the block-preserving null (Table 1); only the ten-arm block-preserving result survives its stated family correction. That rejection is sensitive to the recovered alternative outcomes (Appendix S) and does not survive post-acquisition forgetting (Appendix V).
4. **For order-robustness claims: report order-swing per method** (Table 6), in absolute as well as ratio terms, the CL analogue of the seed-variance bars other subfields expect (Agarwal et al., 2021; Bui et al., 2025); and preserve checkpoint-to-evaluation identity for any generative probe.
5. **Report forgetting per task, not only its mean**: it spans $12.7\times$ across tasks, one directed pair dominates it, and most of replay's advantage is that pair (Section 5.4).
6. **Population-specific diagnostic: measure conditional agreement across group sizes.** In the observed ten-seed corpus, 0.05 takes 2 runs per configuration, 0.03 takes 4, 0.02 takes 5 (Section 4); single-run cells sign only 30 of 50 pair-order comparisons. These are development-corpus descriptions. The separate frozen test finds incomplete transfer (Appendix N); the fixed-host study also finds within-host failures of the criterion (Appendix Q). Re-estimate the requirement for the intended population.
7. **Execution-sensitivity diagnostic: report nominally matched re-executions and compare their discrepancies with margins**: the frozen historical archive supplies 79 eligible pairs over 57 cells here, with exact repeats on one fixed GPU device, 0.047 apart across machines and 0.095 across

software environments among eligible arms (0.028 for differences between core arms); report both rules, and distinguish empirical sensitivity criteria from significance tests. The expanded candidate set is reported separately in Appendix S.

8. **Essential: validate every run against a pinned configuration signature.** Silent drift between queue submissions is indistinguishable from an order effect: an audit excluded 96 runs, including incomplete evaluation records, now in 5 quarantine classes, and a 27-run ablation whose logged flags named an inactive code path. Log what the code did, not what it was asked to do; make an inert flag a hard error; admit cells only through a registry. A number gate shows numerical consistency, not that the intended experiment ran; the signature audit and quarantine record are for that.

U ROUTED CASE-STUDY SPECIFICATION AND FULL INVESTIGATION

The routed bank is an instrument for the sensitivity analysis in Section 5.3; plasticity gating is not a demonstrated contribution.

Each multilayer-perceptron (MLP) linear of a frozen Qwen2.5 model (Qwen Team, 2024) carries $K=8$ rank-$r=8$ low-rank experts; a prototype-augmented top-k router ($k=2$) selects them from a context vector $c(x)$, the mean over prompt tokens of either the input embeddings (*embed*, the conventional choice) or the residual stream after block ℓ (*hidden-ℓ*), the latter optionally *whitened* per dimension—standardised by running statistics primed on the first task and drifting by EMA—which is the step Section 5.3 dissects; hidden context comes from a separate no-gradient pass of the frozen backbone with adapters bypassed; embedding context uses the input embedding lookup. Consolidation optionally anchors experts by a Fisher proxy. The expert bank has approximately 56.4M parameters; router and gate heads bring the trainable total to 58.5M (Table 4). Appendix A gives the full specification, Fig. 11 the pipeline, Appendix D the training loop and Appendix G a worked example.

U.1 FULL ROUTED COMPARISON, CALIBRATION AND EXECUTION HISTORY

Because this is our routed case-study comparison and the replicate spread of these arms reaches 0.095 in the historical reference set (Section 4), we re-ran both arms on one machine (30 cells, one GPU, one stack) and report it from that replication alone; the factor isolation and the calibration ablation are labelled with their population. The arms are not equally exposed: against the canonical grid the whitened arm's re-runs differ by 0.007 on average (max 0.040), the conventional arm's by 0.038 (max 0.139). These 30 unattributable pairs are excluded from the attributed replicate pool, so the historical reference set's per-arm spread for the conventional arm (0.095) does not cover every observed discrepancy.

On our first three orders (o0–o2), routing on a whitened hidden state instead of the input embedding reduced forgetting from 0.252 to 0.080: $3.15\times$ pooled, $3.22\times$ on o0, 9 of 9 paired cells, t=9.63; the completed 2×2 over o0–o4 (Appendix E) isolates the factors: **the effect is the whitening, at either layer**—$2.11\times$ at layer 21 and $2.04\times$ at layer 28 (t=4.14, t=9.16)—while the layer moves 1.12–$1.16\times$ (n.s.) and raw hidden context did not show a nominal paired difference from the embedding ($1.17\times$, t=1.71). Whitening also brings the order-sensitivity: the two whitened arms swing 2.84–$4.22\times$ against 1.35–$1.79\times$ for the three unwhitened, and at o4 the unwhitened-to-whitened forgetting ratio at layer 21 is $0.83\times$ in this factorial grid—harmful.

The two additional orders (Fig. 10a) do not remove the effect: it survives at $4.06\times$ on o3 and is $2.23\times$ pooled over five orders, 12 of 15 paired cells, t=4.56 over cells; clustering the same differences by seed gives t=8.87 and by order t=2.77 (df=4)—aggregation by order gives only five units and is borderline; inference over new curricula additionally requires a sampling model for orders. What collapses is its *size*: on o4 the conventional-embedding-to-whitened forgetting ratio in this single-machine replication is $0.76\times$—the whitened router is now *worse* than the conventional one—and its own order-swing is $3.73\times$ (single-machine; $4.2\times$ on the canonical grid, Table 13). Whitening did not merely reduce forgetting, it made it far more order-dependent. On o0 alone we would have published a $3.2\times$ improvement; on o0–o2, $3.15\times$; on o4 alone, a harmful method—three honest papers from the same code.

Why. The whitening statistics (μ, σ^2) drift by EMA from a first-task priming (Algorithm 1), so the natural hypothesis (H1) is that what matters is the normalisation *frame*. We test H1 against the R-wL21 arm as run in the grid with three variants—first-task calibration, then *frozen*; *all*-task calibration (an oracle diagnostic: it sees every task), then frozen; all-task calibration with drift—on o0, o2 and o4, three seeds each (27 runs on the current machines, re-run after the inert-flag defect documented in Appendix T, item 8; the reference arm's o0 cells are the earlier environment's, re-run below; Fig. 9b, Table 12). These variants do not support H1. **Freezing the statistics destroys the effect regardless of how good the frame is**: on the canonical grid at o0 the arm as run reaches 0.053; frozen after first-task or after all-task calibration it sits near the conventional baseline (Table 12), differing from the arm as run by far more than that arm's own replicate spread (0.03; the frozen variants are unreplicated, so this rests on the paired t: $+0.137$, t=8.86; $+0.119$, t=10.85; df=2). While frozen, *all*-task calibration buys nothing over first-task (-0.018, $t = -1.08$, n.s.). **Letting the statistics drift recovers it**: all-task calibration *with* drift reaches 0.077 at o0, within that arm's replicate spread of the 0.053 as run; pooled over o0/o2/o4, freezing costs $+0.071$ (t=2.35, df=8). Priming on each non-leading task instead reproduces the advantage (twelve runs, Appendix A), so we find no evidence that the frame's *identity* matters—a non-rejection at three seeds, not evidence of equivalence. The effect is a property of the statistics' *trajectory*, and order-dependent: at o4 even the drifting variants sit at 0.217–0.225, no better than conventional routing.

The environment, matched. The 27 canonical cells produced in the earlier software environment—the o0 column of all five routed arms and o1–o2 of the two consolidation arms—have been re-run on the current machines (Appendix E). They disagree with their originals by up to 0.095 (mean 0.026), the largest controlled disagreement among eligible arms in the frozen historical archive and the maximum used in its sensitivity criterion (Section 4). With the re-runs in place both effects stand: the routed arms still swing 1.29–$7.85\times$ across orders (against 1.35–$5.59\times$) and whitened routing still beats conventional by $2.41\times$ pooled (t=6.69, 14 of 15 cells): the confound was real but did not produce them (Appendix E).

V Sensitivity to the forgetting definition and reporting reference

V.1 A post-acquisition alternative

The frozen signed-forgetting metric uses the best pre-final accuracy, including stages before a task's own training. For a pretrained model, this can count a pre-acquisition decline. We therefore add the post-hoc sensitivity

$$\mathrm{AF}_{\mathrm{post}} = \frac{1}{T-1} \sum_{j=1}^{T-1} \left(\max_{j \le i < T} R[i,j] - R[T,j] \right).$$

Both definitions exclude the zero-shot row and the final stage from the maximum; neither clips negative values. Only the lower bound of the maximum changes. The frozen results remain primary. This alternative answers a narrower retention question, and is not a correction to a mismatch between the original formula and code. Table 41 reports every inspected grid. Populations overlap and their counts must not be added as independent trainings.

The three-seed core-grid change is the sequential-LoRA/EWC comparison on o2: its mean difference moves from $+0.00067$ to -0.00033 accuracy units. Thus the original reversal count is definition-sensitive, despite unchanged pooled rankings. In the ten-seed development grid, all 61 affected task contributions concern Amazon. These observations are descriptive sensitivities, not fresh training replications.

Fresh-seed qualification. Holding the original set of 16 development-qualified comparisons fixed, 13 retain the agreement criterion under the alternative metric, matching the primary numerator. Applying the alternative to development qualification as well gives 13/15. Sequential LoRA/EWC on o2 loses development qualification. These denominators answer different questions; the post-hoc result does not replace the frozen 13/16 endpoint.

Table 41: Post-acquisition sensitivity on saved matrices. "Changed" counts trajectories whose signed forgetting differs. Reduction is the mean change in percentage points. Reversals count method pairs whose order-mean difference has both signs. Both metrics use the same runs. All listed pooled rankings are unchanged; small order-specific comparisons can change.

Population	Runs	Changed	Reduction (pp)	Reversals: old / post
Qwen core, three seeds	75	19	0.067	3 / 2
Qwen core, ten seeds	250	61	0.066	1 / 1
Llama core	75	5	0.023	4 / 4
SmolLM2 core	75	4	0.011	3 / 3
Expanded evaluation (R3)	75	9	0.015	1 / 1
Qwen, all eligible arms	195	37	0.046	33 / 32
Routed, single host	30	6	0.043	1 / 1
Fresh follow-up	150	49	0.091	1 / 1

Fixed hosts and routed comparison. Under common-prompt scoring, the fixed-host qualifying counts remain 0/3 on the workstation and 1/3 on the cluster. The alternative AF contrasts, combined with unchanged final-accuracy contrasts in the same multiplicity family, yield 0 Holm rejections. This is a secondary reanalysis, not a new confirmatory study.

The single-host conventional/whitened ratio remains $4.06\times$ on o3 and $0.76\times$ on o4; its pooled value changes from $2.23\times$ to $2.25\times$. The order-level nominal p values are 0.05023 and 0.05003, respectively. Both remain just above the conventional threshold; rounding must not convert either into a rejection.

Evaluation resampling. We reuse the exact 2,000 joint and independent-stage draws for every R3 training, reproducing all 75 original per-cell standard deviations (SDs) at their reported precision before applying the alternative to the same sampled matrices. The alternative median SDs, in accuracy fractions, are 0.0118 jointly and 0.0159 independently, compared with 0.0119 and 0.0158 originally. This remains a fixed-model, finite-pool analysis; it does not restore the missing R3 cross-method example identities.

For paired method differences, nominal order-test rejections change from 7 to 6 of ten pairs. The Holm and GG/Holm rejection sets remain unchanged at 6 and 4, respectively. The additive method$\times$order interaction changes from $F(16, 216) = 4.27$ to 4.20; the repeated-response sensitivity changes from $F(16, 144) = 5.531$ to 5.494. The latter still assumes sphericity and independent seed blocks; these sensitivities do not make the chosen orders a random sample of curricula. The sensitivity uses 200,000 permutations per pool for each null and metric.

Adjusted tests. The post-acquisition ten-seed order tests retain the same Greenhouse–Geisser/Holm rejection set: sequential LoRA, EWC-style LoRA and SVD-LoRA. Replay and O-LoRA do not reject under either definition. Table 42 repeats both permutation families, using paired Monte Carlo draws across the two definitions and separate Holm corrections over the same six pools. These are post-hoc sensitivities; non-rejection does not establish absence of order effects. The original ten-arm block-preserving rejection does not survive the alternative definition: no pool rejects either corrected family under post-acquisition forgetting. Thus the discrete reversal inference is definition-sensitive, even where pooled rankings and the fixed-host agreement conclusion persist.

V.2 Uncertain reference directions

The fresh full-cohort mean used by the secondary directional audit is itself uncertain. Table 43 separates the non-replay comparisons by whether their nominal fresh interval includes zero. Intervals are unadjusted; excluding zero does not make a reference population truth, while including zero does not establish equivalence. This stratification is post hoc and does not select comparisons for the primary result.

Table 42: Definition sensitivity of reversal counts and permutation tests. Each p column is Holm-adjusted within its six-pool family. Values are old / post-acquisition; the new Monte Carlo draws need not exactly match the frozen table's draws.

Pool	Reversal pairs	Independent p	Block-preserving p
Core five	3 / 2	0.7554 / 1.0000	0.5064 / 1.0000
External eight	10 / 9	0.7554 / 1.0000	0.5064 / 1.0000
Original ten	19 / 18	0.1932 / 0.7745	0.0348 / 0.2880
All eligible thirteen	33 / 32	0.6070 / 0.9078	0.1495 / 0.2880
Llama core	4 / 4	0.6070 / 0.7500	0.2208 / 0.2880
SmolLM2 core	3 / 3	0.6070 / 0.7745	0.1495 / 0.1794

Table 43: Non-replay historical three-seed decisions versus fresh ten-seed reference means. Counts are opposed / declared. Five reference intervals include zero and four exclude it; the same reference is reused across dependent historical subsets.

Rule	Reference includes zero	Reference excludes zero
Mean sign	320/582	8/480
Separate SD bands	8/75	0/210
Paired nominal interval	6/18	0/146

V.3 Terminal-task composition

The five orders balance the leading task, not the terminal task; omitting the terminal task from forgetting changes the included task identities. As a post-hoc check, for each pair of orders we recompute both method differences using only tasks that are non-terminal in both orders. This keeps outcome composition equal within each two-order comparison, without inventing unobserved curricula.

Across 100 method-pair/order-pair combinations in the ten-seed grid, opposed order-mean signs occur in 4 with the original task inclusion and 12 after restriction to shared non-terminal tasks. The corresponding post-acquisition counts are 4 and 12. Composition changes which comparisons reverse; it does not account for every observed reversal. These overlapping contrasts use different task sets across order pairs, are not independent trials, and do not establish a causal order effect. A fully balanced terminal-task design would require additional curricula.

Supplementary CSVs MS01–MS05 expose all matrix sensitivities, fresh and fixed-host agreement rows, reference strata and common-task contrasts. The analysis and resampling scripts preserve frozen results and refuse incomplete input grids.

W Recorded prompt-truncation exposure

This post-hoc input audit reconstructs prompt suffixes from the exact retained token IDs of the 18-trajectory scorer diagnostic and 60-trajectory fixed-host study. Both use Qwen2.5-1.5B, o2 and a 192-token budget. Input hashes and equality across methods and hosts are checked; each task–seed sample is counted once, irrespective of method, host or training stage. Different seed samples can overlap, so the denominators below count sample positions rather than distinct source records.

Table 44: Input exposure, not prediction changes. "Truncated" means at least one candidate loses prompt tokens; "unequal" means candidates receive different retained prompts under the original scorer. Columns show counts (percentages). Diagnostic: 750 positions per task; fixed-host: 2,500 per task.

	Diagnostic (3 seeds)		Fixed-host (10 seeds)	
Task	Truncated	Unequal	Truncated	Unequal
AG News	1 (0.1%)	1 (0.1%)	4 (0.2%)	4 (0.2%)
Amazon	73 (9.7%)	0 (0.0%)	194 (7.8%)	0 (0.0%)
Yelp	224 (29.9%)	0 (0.0%)	725 (29.0%)	0 (0.0%)
DBpedia	0 (0.0%)	0 (0.0%)	2 (0.1%)	2 (0.1%)
Yahoo	147 (19.6%)	147 (19.6%)	528 (21.1%)	528 (21.1%)

The common-prompt scorer removes candidate-dependent conditioning by construction; it does not remove prompt truncation. These counts do not establish that truncation changes a prediction or causes an order effect. They apply only to these preserved inputs and the recorded tokenizer revision. Historical populations without exact token records, and the Llama and SmolLM2 tokenizers, remain outside this audit. The script `scripts/truncation_exposure.py` regenerates the counts from original or anonymous portable records without tokenization, forward evaluation or training.