CORRECT NUMBERS, WRONG CONSTRUCTS: ACQUISITION-AWARE EVALUATION ACROSS MODEL SCALES

Anonymous authors
Paper under double-blind review

ABSTRACT

We built a program to learn, at a small scale, which adaptations make a model forget, and to test whether that knowledge survives an order of magnitude of scale. It returned a striking headline: roughly 98% of the relationships it selected failed to transfer. This paper is the audit that stopped us reporting it, and what the audit found was not one mistake but five instances of one. Each has the same shape: a quantity was computed correctly, and a different quantity was inferred from it. A count of records read as a count of independent observations, where 186 "pairs" were 62 directed pairs recorded three times, byte-identical across all 558 matched groups. A failure rate read as evidence that selection failed, at 98.7% against 98.8% for choosing at random on the same population. Deterioration read as forgetting, after an update that trained repeatedly on one 32-sequence batch and left held-out loss worse in 60 checks across three scales. A shared recipe's collapse read as a property of scale, when a per-scale recipe acquires in every confirmation cell and a width-scaled learning rate alone moves 1.4B from $+0.043$ to $+0.239$. And a preregistered median read as agreement between task families, where with three families the median is the middle family's value, certifying a match that held at a 0.0019 nat spread in one family while another disagreed by 0.5079. None required an unusual mistake, and preregistration caught none of them because one of them is the preregistration. A sixth arrived while we built the corrected experiment, this one caught prospectively: the per-scale recipe our own median rule certified for 160M acquires in just 32 of 62 pairs at that scale. We give an acquisition-aware protocol and a checklist with one item per substitution. Across 18 published papers a screen finds an explicit held-out new-task metric in 11 and does not resolve the other 7, so the defect is not universal and the screen does not establish how far it reaches. Asked at last under a preregistration written before any fit existed, the original question does not resolve: our candidate reaches a held-out R^2 of -0.004 and fails its margin against the best baseline, which is the one comparator with a positive point estimate, $+0.023$. On 15 held-out pairs we cannot separate absence of signal from absence of power, and we report both.

1 INTRODUCTION

We set out to answer a question that seemed tractable: can you learn, at a small scale, which adaptations will cause a model to forget, and trust that answer at a larger one? We built the apparatus, ran it, and got a striking result. Almost nothing transferred. Roughly 98% of the relationships our procedure selected at the small scale failed at the large one, and we began writing that up.

The number was real. What it meant was not. Auditing our own pipeline before submission, we found that the models had never learned the new task at all, at any of the three scales: loss fell on the batch being trained and rose on held-out text from the same task. We had been measuring how much a model forgot after teaching it nothing. Forgetting, in that setting, is not a small effect measured badly; it is a quantity that does not yet have its usual meaning, because the condition that gives it meaning was never met.

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.

Then we found four more of the same thing, and that is what this paper is about. Each has an identical shape: **a quantity was computed correctly, and a different quantity was inferred from it.** The arithmetic is right, the code does what it says, and the construct named in the conclusion is not the construct the number measures. We call these *inferential substitutions*, and they are worth a paper rather than a changelog because none required an unusual mistake. Each followed from a default that looked reasonable alone: a cache key that omitted a seed, a denominator chosen without a base rate, a recipe that transferred, an aggregate over few heterogeneous strata. Preregistration caught none of them; one of them *is* the preregistration. Sections 2 to 7 are the five, each named for the substitution it makes rather than the bug that caused it.

The idea that a model's failure to acquire the new task must be reported alongside its forgetting is not new. Chaudhry et al. (2018) named it *intransigence* and argued it belongs beside forgetting as a first-class metric, and Díaz-Rodríguez et al. (2018) made a similar case against reporting forgetting alone. Our contribution is not the desideratum. It is a demonstration that a careful, preregistered, well-resourced study can violate it silently, produce a publishable-looking headline, and never notice, together with an account of what it takes to notice. Evaluation-protocol critique in continual learning has this shape (Farquhar & Gal, 2018); what we add is the failure seen from the inside, with the artifacts.

The apparatus is worth seeing because of how ordinary it looks. Two model families are used, Pythia 160M/410M/1.4B and SmolLM2 135M/360M/1.7B, over directed $A \to B$ task pairs from three families (domain, capability, language) built from streamed corpora. Before adapting, it measures cheap signals: gradient alignment, representation overlap, task novelty, planned update size and their pairwise combinations. For each pair and scale it adapts to B and records the change in held-out A loss, normalised by a noise floor. Candidates are selected at the discovery scale by held-out R^2, then scored at the replication scale by frozen-coefficient ΔR^2 over the best baseline. The design has real precedent (Magnusson et al., 2025), and its reliability is known to depend on how much signal a benchmark carries relative to its noise (Heineman et al., 2025).

Contributions.

1. **One failure mode, instance by instance**, each a correct number attributed to the wrong construct, documented with the artifacts that show the mechanism rather than inferred from symptoms: the inferential unit (§2), the base rate (§3), fresh data over a broken unit (§4), acquisition (§5), recipe transfer (§6) and our own preregistered median (§7).

2. **The corrected question, asked** (§8) under a preregistration fixed before any fit existed, returning no candidate that clears its margin, at a width we state rather than work around.

3. **A preliminary acquisition-reporting screen** of 18 papers (§11), with explicit classification criteria: it finds a held-out new-task metric in 11 and leaves 7 unresolved, which is a limit of the screen rather than a finding about those papers.

4. **An acquisition-aware protocol and a checklist** (§9, §10), one item per substitution, most checkable before any GPU time is spent. Figure 1 sets the original procedure beside the audited one.

2 When a replicate is not independent evidence

The analysis treated 186 "task pairs" as independent units. They are 62 directed pairs, each recorded three times under different seed labels, and the three records are *byte-identical* on every measured field in all 558 (pair, scale, strength) groups. The seed changes only the label and the split assignment: the data loader's cache key excludes it, the corpus stream is read from the start without shuffling, and the model is a deep copy with no re-seeding. Consequently 34 of the 62 pairs appear on both sides of the discovery split, and the "bootstrap", which draws 186 labels with replacement and then takes the set of distinct labels, produces episodes containing $\sim$95% of the underlying pairs. The three "independent datasets" used as replicates are identical in all 558 matched units.

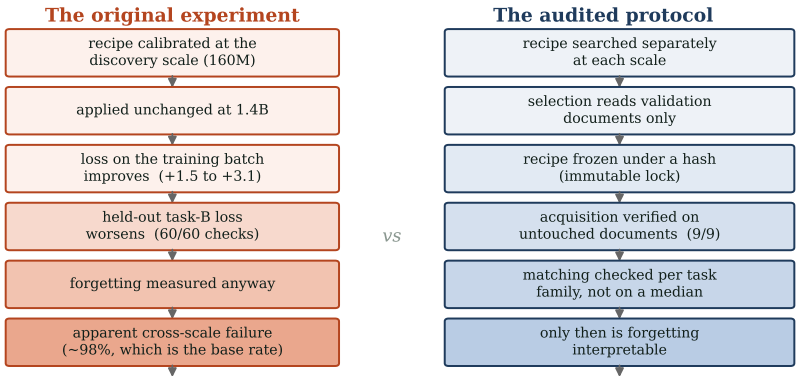


Figure 1: Where the original experiment went wrong, and what the audited protocol changes. The left column is not a strawman: each step is what the code did. The defect is that nothing between step two and step five checks whether the new task was learned, so a forgetting number is produced regardless.

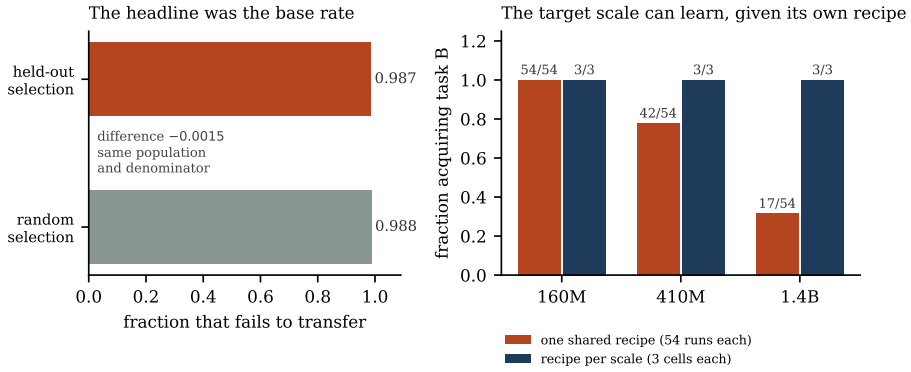


Figure 2: **Left:** the program's headline, that almost nothing transfers, next to what random selection achieves on the same population and denominator. The bars are meant to look identical. **Right:** acquisition by scale under one shared recipe against a recipe selected per scale. The shared-recipe bars are 54 runs each; the per-scale bars are the three confirmation cells at each scale.

3 When a failure rate is only the base rate

Fixing the unit (deduplicate to 62 pairs; split and resample at the pair level; keep the frozen analyzer otherwise byte-for-byte) leaves the false-law rate near 0.98, but reveals what it means.

The regenerated corrected-unit analysis selects one candidate in each of 300 split episodes over the 62-pair historical population. Four selections transfer, giving selected failure $296/300 = 0.986667$, or 98.7%. Across all 24 candidates in those same episodes, 85/7200 transfer, giving base rate 0.011806 and random failure 0.988194, or 98.8%. Their difference is -0.001528, which is the left panel of Figure 2. These share a declared population and denominator; earlier draft tables mixed analysis versions. Repeated episodes are dependent summaries, and a small point-estimate difference is not an equivalence or superiority test; establishing that these two rates are *the same* needs a declared smallest effect of interest and a test built for the purpose (Lakens et al., 2018), and the variance a benchmark pipeline contributes would have to be modelled rather than assumed (Bouthillier et al., 2021). What we can say is what the numbers say: on this population the two rates differ by -0.0015, so the headline is not separated from what random selection achieves, which is the behaviour Cawley & Talbot (2010) describes when a selection criterion is optimised and then re-

ported without separating selection from evaluation. That is a point estimate, not a demonstration that selection cannot help.

4 When fresh data cannot rescue a broken unit

We preregistered a replication matrix on *fresh* data: 62 pairs $\times$ two distinct draws from each corpus stream $\times$ three scales (372 runs), with every record carrying a SHA-256 fingerprint of its tokens (no draw pair shared a fingerprint). Distinct fingerprints establish different token sequences, not document disjointness or statistical independence. Four cells were specified in advance, each fitting on one draw and evaluating on the other: same-scale replication (R1), frozen-coefficient transport (R2), recalibration at the target scale (R3), and within-target cross-validation (R4).

By the frozen rule ($\Delta R^2 \geq 0.02$ over the best baseline) the counts are 2, 9, 0 and 2 of 24. But every best-in-cell *absolute* out-of-sample R^2 is negative (-0.085, -0.053, -6.479, -0.034): the passing cells are cases of one predictor being less useless than another. A permutation null for R1 (shuffle the outcome, refit) gives $P(\text{null} \geq 2 \text{ passes}) = 0.125$, so the observed count is consistent with chance. R2 exceeding R1 is partly explained by its baseline being fit at the source scale and applied at the target, where it is poorly calibrated.

5 When deterioration is not forgetting

Every result above is downstream of one design choice. Adaptation was 30 optimisation steps on a single fixed batch of 32 sequences. We measured task-B loss before and after adaptation on *held-out* task-B text from the same corpus.

Table 1: Task-B acquisition under the original update, in nats. Positive means the model improved on that text after adaptation. The two columns separate the batch the model was trained on from held-out text drawn from the same task; the sign differs between them at every scale, which is the defect this section reports.

Scale	n	Median on the adapted batch	Median on held-out B	Held-out > 0
160M	20	$+1.52$	$\mathbf{-3.09}$	$0/20$
410M	20	$+3.07$	$\mathbf{-2.14}$	$0/20$
1.4B	20	$+2.84$	$\mathbf{-2.60}$	$0/20$

Table 1 separates the two: training-batch fit improved while held-out task-B loss worsened in all 60 measured checks. "Forgetting of A after learning B" was measured after an adaptation that did not improve held-out B loss in any of these checks. This also supersedes an earlier hyperparameter-robustness result of ours: a feasibility search, motivated by the observation that small-scale conclusions are often hyperparameter artifacts (Lourie et al., 2026), had declared lr $= 10^{-4}$ "viable" because task-B loss fell *on the batch being trained*. The alternative to divergence was improved batch fit accompanied by worse held-out loss; batch improvement alone did not validate generalization.

6 When recipe transfer masquerades as scale dependence

We replaced the training path, using a task-B pool of eight disjoint stream draws (256 sequences), fresh minibatches without replacement, and separate B-train, B-validation and B-test stream chunks, then searched lr $\times$ steps selecting *only* on validation B acquisition, with no forgetting outcome consulted. The search is informative in itself (Table 2): the regimes the original program used sit far inside the destructive region, and acquisition improves as the learning rate *falls*.

The selected recipe was calibrated on eight discovery-scale pairs using validation acquisition; the $+0.120$ value is that calibration mean. Stage 2 contains 162 records: 54 directed pairs at three scales, with one data draw per pair and scale. Eight of the planned 62 pairs are absent, all involving cap-code. This dataset is distinct from the earlier two-draw, 372-record replication matrix.

Table 2: The recipe search behind the repair, at the discovery scale: held-out task-B acquisition in nats for each learning rate and update count. Positive means the model learned the task on text it was not trained on. The regimes the original program treated as viable are the bottom two rows, and acquisition improves as the learning rate falls rather than rises.

	30 steps	100 steps	300 steps
$\text{lr} = 10^{-5}$	$+0.064$	$+0.087$	$+\mathbf{0.120}$
$\text{lr} = 3 \times 10^{-5}$	$+0.019$	$+0.003$	-0.080
$\text{lr} = 10^{-4}$ (program's "viable" recipe)	-1.508	-1.520	-1.813
$\text{lr} = 3 \times 10^{-4}$	-4.223	-3.731	-3.754

Table 3: Held-out acquisition after the recipe change, by scale, where positive means lower task-B loss on text the model was not trained on. All three scales share one recipe here, chosen at the discovery scale; the target-scale column is the result that motivates selecting per scale instead.

Scale	Acquisition > 0	Median acquisition (nats)
160M	54/54	$+0.1240$
410M	42/54	$+0.1249$
1.4B	17/54	-0.0777

The pooled count is 113/162 acquisition-positive runs; 130/162 show old-task deterioration, and 82/162 show both. Pooling conceals failure to learn B in most target-model records. Consequently this experiment does not remove acquisition failure as an explanation for poor source-to-target prediction.

The existing exploratory selection analysis reports failure rates 1.000 for the selected candidate and approximately 0.9986 for random selection, with mean discovery–target score correlation approximately -0.33. These summarize a fixed candidate family over dependent splits and mix acquisition-successful and acquisition-unsuccessful regimes. They do not establish meaningful inferiority to random selection, nor a failure of relationship transfer under acquisition-matched adaptation.

Outcome normalization also changes the legacy threshold substantially. On the same repaired 54-pair population and 300 splits, using raw change in NLL gives 1967/7200 baseline-relative transfer passes (27.32%), versus 10/7200 (0.139%) with the noise-normalized outcome. Nevertheless, the selected raw-loss predictor has negative frozen-target absolute R^2 in all 300 splits. Passing a relative threshold is not sufficient evidence of predictive skill, and the two outcomes must not be mixed.

The next scientific comparison must select recipes separately by validation at each scale, verify acquisition on untouched test data and compare frozen versus recalibrated predictions on fresh task/data units. Equal compute and equal acquisition are different comparisons. Selection on test acquisition would introduce a new bias, so failed acquisition must be reported rather than filtered after the outcome is known.

7 When preregistration certifies the wrong estimand

The previous section ends with a requirement rather than a result: select separately by validation at each scale, then verify acquisition on data the selector never saw. We ran that as a small pre-registered pilot, on three task pairs (news domain, mathematical capability, French language), the same three Pythia scales, whole documents assigned once to disjoint roles, and a selector that reads validation documents only and then freezes an immutable lock. Confirmation retrains the frozen recipes on separate confirmation-training documents and scores them on separate confirmation-test documents.

Two thresholds failed before one worked, and we report all three. The pilot pre-registered an acquisition band: a scale is eligible if its median validation acquisition lies in a 0.025-wide window, at least 80% of its cells are positive, and no task family is negative. Scales are matched if their selected medians lie within 0.025 nats of each other.

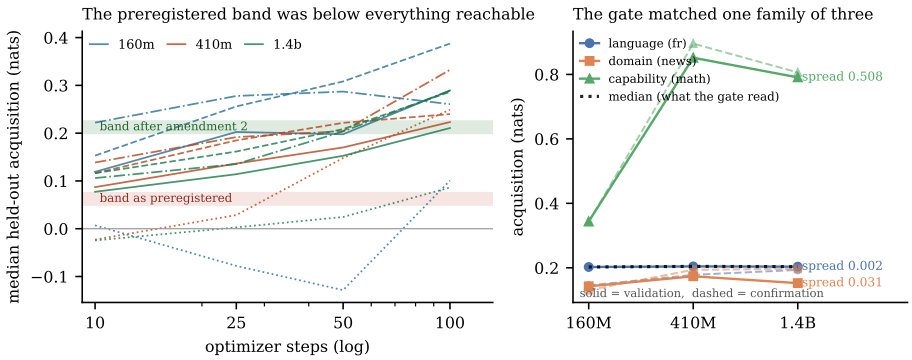


Figure 3: **Left:** median held-out acquisition against optimizer steps, for every scale and learning rate in the search. The band as preregistered (red) lies below everything reachable: at the gentlest rate and the shortest rung, acquisition is already $+0.077$ at 1.4B. The amended band (green) intersects all three scales. Dotted curves are lr $= 3 \times 10^{-5}$, which crosses into destruction rather than learning less. **Right:** acquisition at the selected recipes, per task family, on validation and on the confirmation documents the selector never saw. The gate compared medians, which the language family supplied at all three scales; the capability family disagrees across scales by 0.51 nats at selection time, twenty times the 0.025 tolerance.

The first band, $[0.050, 0.075]$ nats over checkpoints $\{100, 300, 600\}$, returned no eligible recipe at any scale: every rung either overshot the band or had already crossed into destruction. The second attempt added a shorter ladder, $\{10, 25, 50, 100\}$, and also returned nothing, for a reason worth stating: at the gentlest learning rate and the shortest rung, acquisition is already $+0.1191$, $+0.0870$ and $+0.0773$ at 160M, 410M and 1.4B. The target scale misses the band's ceiling by 0.0023 nats. The band was crossed below ten optimizer steps, where a trajectory is a few thousand tokens and matched acquisition would mean matched on almost no learning. We therefore moved the band level to $[0.200, 0.225]$, keeping its width, the positivity rule, the matching tolerance, the grid, the ladder and the selector unchanged. Figure 3 shows both bands against what the search could reach, and the per-family spread the median hid.

That third threshold was set after seeing which values the measurements produced, and we do not present it as predicted. It was set using validation acquisition only, with no forgetting, transfer or R^2 quantity available to the selector, and the protection against tuning is that the frozen recipes are then scored on documents the selector never saw. A reader should treat the level as calibrated, and the claim it supports as feasibility rather than a test of the level.

The target scale acquires. Under a recipe chosen for it, 1.4B learns the new task in every confirmation cell (Table 4). This bears on the $17/54$ line in Table 3 without settling it; the right panel of Figure 2 contrasts the shared recipe with per-scale selection. That matrix applied the discovery-scale recipe everywhere; at 1.4B the pilot puts the same learning rate and duration at $+0.043$ with two of three families positive, and one rung further in duration at -0.267. So the recipe sits near its destructive threshold at the target scale, which is consistent with the failure but does not reproduce it: the pilot at that recipe is weakly positive where the matrix was strongly negative, and the two differ in pool size, batch size and task-B data as well as in scale. Recipe transfer is a demonstrated hazard here, not an established explanation of that specific line.

The matching never held, and our own gate could not see that. The eligibility rule compares a scale's *median* validation acquisition against a band, and matches scales by the spread of those medians. With three task families the median is the middle family's value. At the selected recipes that middle family was the language task at all three scales, and the language family happens to be tightly clustered (0.0019 nats). The other two were not: the domain family spread 0.0312 and the capability family 0.5079, the latter twenty times the tolerance. The gate certified a cross-scale match that existed in one family out of three.

Table 4: Per-scale selection on validation documents, and the frozen recipes scored on separate confirmation documents, broken out by task family rather than summarised by the median the gate used. Selected recipes: 160M lr 10^{-6}/25 steps, 410M and 1.4B lr 10^{-5}/50 steps. Acquisition is the reduction in held-out task-B loss; positive means learned, and every cell is positive at every scale. The gate required a cross-scale spread below 0.025. Only the language family comes near it, and the median the gate read was that family at all three scales on validation; on confirmation the median is the language family at 160M but the domain family at 410M and 1.4B, so the final row compares different tasks.

	Validation (used for selection)			spread	Confirmation (untouched)			spread
Task family	160M	410M	1.4B		160M	410M	1.4B	
Language	$+0.2025$	$+0.2044$	$+0.2036$	0.0019	$+0.1453$	$+0.1781$	$+0.1934$	0.0481
Domain	$+0.1425$	$+0.1737$	$+0.1525$	0.0312	$+0.1307$	$+0.1939$	$+0.1967$	0.0660
Capability	$+0.3438$	$+0.8517$	$+0.7909$	$\mathbf{0.5079}$	$+0.3419$	$+0.8957$	$+0.8061$	$\mathbf{0.5537}$
Median (gate)	$+0.2025$	$+0.2044$	$+0.2036$	0.0019	$+0.1453$	$+0.1939$	$+0.1967$	0.0513

The confirmation numbers compound it: there the median is the language family at 160M but the domain family at 410M and 1.4B, so the 0.0513 figure we would otherwise have quoted compares a French-Wikipedia result against a CNN/DailyMail one. Matched by family, all three spreads exceed the 0.025 tolerance, at 0.0481, 0.0660 and 0.5537.

So the gate does not pass, the pilot stops as its stopping rule requires, and the reason is simpler and worse than a selection-optimism story: acquisition was never matched across scales at all. We report that rather than the narrative the median supports.

Two things follow. Acquisition at the target scale is real and reproducible on documents chosen before the recipe was frozen, positive in every one of the nine confirmation cells, which is a per-cell fact unaffected by how the gate aggregates and removes failure-to-learn as a blanket explanation at 1.4B. And a criterion we wrote ourselves in advance, to prevent exactly this error, reported a match that did not exist, because aggregating three heterogeneous strata by their median discards the two that disagree. This is the small-sample aggregation problem Agarwal et al. (2021) document for reinforcement-learning benchmarks, with a preregistration wrapped around it: registering a criterion in advance protects against choosing it after the fact and against nothing else. A criterion can be honest and still be the wrong estimand. Three task pairs cannot support a population claim, so this is a feasibility result about the precondition and an unmet tolerance on fresh data; §8 is where the program's actual question gets asked.

8 Asking the corrected question, at the width we could afford

The audit removes the grounds for the original conclusion without supplying the answer, so we ran the corrected experiment. We report the sample sizes first, because they bound everything after: the test set is **15 held-out pairs** against 26 for the fit. Those two counts come from different scales and disjoint splits, so they do not sum to a single eligible population: the fit is the 26 training-split pairs that acquire at 160M, the test is the 15 held-out pairs that acquire at 1.4B, and no pair appears on both sides. Acquisition is verified at the scale where each cell is used. It is *not* verified at both scales for the test pairs: only 6 of the 15 also acquire at 160M, because the 160M recipe acquires in 32 of 62 pairs while 410M and 1.4B acquire in all 62. So this evaluates predictor transport between scale-specific acquisition-eligible populations, not transfer under acquisition-matched adaptation, and the phrase *the corrected question* should be read against that narrower design. At that width a null cannot distinguish the absence of a relationship from the absence of the power to see one, and nothing below should be read as though it could. The protocol is the one §9 describes, with acquisition verified per cell on held-out task-B text, per-scale recipes frozen under a hash, the predictor probes unchanged because they are computed before the task-B update, and the estimand, comparators, margin and stopping rule fixed before any model was fitted.

Two preregistered gates ran first, and the second produced a sixth instance of this paper's pattern arriving prospectively rather than in retrospect: **only 32 of 62 pairs acquire at 160M**, under the recipe selected *for* 160M by the median-over-three-families rule whose defect §7 documents (Appendix C).

Fitting at 160M on training-split pairs and testing at 1.4B on held-out pairs, the best two-variable combination reaches a held-out R^2 of -0.004, against -0.273 for KL-only, $+0.023$ for similarity-only and -0.030 for N-only. **The candidate does not beat the strongest baseline**: its margin over similarity-only is -0.027 with a bootstrap interval of $[-0.115, +0.066]$, so it fails the preregistered threshold and the gate does not clear. Similarity-only is the one comparator with a positive point estimate, and we will not round it away: on 15 pairs an R^2 of $+0.023$ is not evidence of useful generalisation, but it is not nothing either, and the honest reading is that this test does not establish the absence of transferable signal. Under the unwinsorised sensitivity analysis every model including similarity-only is negative; the two analyses agree on the verdict, not on the sign of each baseline. Appendix C separates what that result licenses from what it does not.

9 PROTOCOL

Every substitution above is a place where a quantity leaked from where it was chosen to where it was tested, so the protocol that closes those routes is what we would want reused rather than any single number. Six things do the work: documents are the unit and are assigned once with normalised-text deduplication, so no document appears in two roles; inputs are sealed before any model runs, with the manifest carrying the hash of the configuration it was built for, so a configuration change invalidates the chain instead of silently rebinding it; the selector sees calibration-validation results only, never forgetting or any confirmation quantity; its choice is frozen under a lock that confirmation verifies before the confirmation pool is opened at all; the numerical protocol is fixed before outcomes and carries three checks that run against the live model rather than in tests; and every model-touching invocation has explicit wall-clock and dollar caps against a shared ledger. Appendix A gives each in the detail needed to reimplement it.

The search history. Three acquisition bands were attempted in order; only the third produced an eligible recipe at all three scales, and its confirmation then failed the matching tolerance. Each change was declared in a dated amendment before the run it governed, and the second amendment carried a stopping rule that ended the search regardless of outcome. We think the stopping rule matters more than the amendments do: a threshold that moves after seeing the data is defensible only if something stops it from moving again, and the honest version of that commitment is written down before the last run rather than claimed afterwards. Every attempt, with its band, step ladder and outcome, is in Table 5 of Appendix B.

10 A CHECKLIST

Each item below corresponds to a defect above, and each is cheap.

1. Report **held-out acquisition** beside every forgetting number, with document isolation and uncertainty. Improvement on the training batch alone does not establish generalization.
2. Compare any selection procedure against **random selection and the base rate**, not only against its own false-discovery rate.
3. Report **absolute** out-of-sample skill, not only improvement over a baseline that may itself be useless.
4. Verify that nominal replicates **differ**: hash the records.
5. Resample and split at the **unit of scientific interest**, and check that no unit crosses the split.
6. Prefer a task whose loss the model can actually improve; check headroom before building the benchmark.
7. **Select the adaptation recipe separately at each scale.** A recipe calibrated at the discovery scale can sit past the destructive threshold at the target scale, which then reads as a scale effect (§7).
8. **State an eligibility rule per stratum, not through a median.** With few heterogeneous strata the median is one stratum's value, and it can certify agreement that holds nowhere else.

9. If a threshold is amended after seeing data, **write the stopping rule with it**. An amendment without one is indistinguishable from searching until the gate opens, including to the authors.

The last three are the ones we would have most wanted in advance, and none is expensive: the first two cost one extra validation sweep per scale, and the third costs nothing.

The checklist caught us again, hours later. We can report one prospective test of item 4, because we failed it while writing this paper. Building the follow-up experiment this audit calls for, we preregistered a feasibility gate over 12 sampled pairs; two of the first three cells returned byte-identical values. The sampler had drawn from the 186 labelled pairs rather than the 62 distinct ones, sampling one pair twice under different seed labels. That is the first substitution in this paper, committed inside the experiment designed to correct it, on the day its checklist item was written. If a fresh implementation by someone primed for exactly this defect reproduces it within hours, the claim that these substitutions follow from ordinary defaults is not a rhetorical softening. The gate was rerun under a dated amendment that discarded the observed cells and moved no criterion.

11 Related work, and how often a reader can verify acquisition

Comparing learning and forgetting at matched target performance is established in the learning–forgetting Pareto analysis of Biderman et al. (2024). Scaling studies already examine forgetting jointly with model scale, target-data amount and pretraining-data injection, selecting checkpoints by target validation loss (Bethune et al., 2025). Small-scale data decisions can also transfer successfully (Magnusson et al., 2025), and benchmark noise affects their reliability (Heineman et al., 2025). These results prevent reading our program's defects as a general failure of small-scale experimentation, and candidate selection must be separated from final evaluation to avoid selection bias (Cawley & Talbot, 2010).

Reading deterioration as forgetting is the substitution with the best claim to generality, so we tested that claim on this literature and report what came back rather than what we hoped for. Across 18 papers from 2023 to 2026, 11 plainly report a quantity sufficient to check that the new task was learned, so **the general claim we wanted is not available**. What we can report is narrower: in 7 of the 18 this screen did not find a quantity sufficient to check acquisition, which is not the same as showing none is reported, and those cases span different reporting and inspection limits rather than one mechanism. The pattern worth naming is a reporting one: where backward transfer is defined against the diagonal a_{ii} and neither it nor an equivalent new-task measure is shown, a reader cannot check acquisition alongside retention. That motivates explicit acquisition reporting; our screen does not establish whether these papers report it elsewhere. Appendix D gives the classification rule, the counts, and why a convenience sample of 18 papers cannot be read as a rate.

12 Limitations and conclusion

This is an audit of one program. Its original evidence includes two model families and 62 directed pairs; the revised matrix uses one family, 54 pairs and one draw per pair and scale. Separate stream chunks do not establish document-level independence, which subsequent data construction must enforce. Old-task deterioration remains a valid measured quantity when acquisition fails; it simply cannot be read as damage accompanying successful learning. We do not claim that published methods share our defects, nor that acquisition matching is new. Appendix E states the remaining threats and the three results that would revise what we say.

We set out to learn which adaptations make a model forget. What we found first was that the question had a precondition we had not checked, and that checking it exposed four more failures of the same kind, each a quantity computed correctly and read as a different one. We then asked the original question under a preregistration, and at the width we could afford it did not resolve: on 15 held-out pairs, each acquiring at the scale it was tested on, our candidate reached -0.004 and failed its margin against a similarity baseline at $+0.023$ (§8). That is not a verdict on the question and we do not dress it as one. The protocol in §9 and the checklist in §10 are what we would hand to someone starting this study today. If one thing survives, we would rather it were the habit of asking, of any number carrying an argument, which construct it measures.

REPRODUCIBILITY STATEMENT

The anonymised supplementary material contains the dated document trail this paper argues from: every preregistration and every amendment, in the order they were written, including the two written before the corrected experiment had produced a number and the one that discards observed cells. The claim that preregistration caught none of the substitutions is checkable only against documents whose dates precede the results they govern, so those documents are the primary artifact here rather than a supplement to it.

Also included: the gate analysis scripts and their outputs, the per-scale recipe hashes, the acquisition runner, and the full reporting survey of 18 papers with a per-paper verdict and the quoted evidence for each, so a reader can disagree with any individual call without re-reading the sample. The checklist of §10 is included as a standalone file, since it is the part we expect to be reused and it is checkable before any compute is spent.

Model weights and task corpora are public and named in the text. Raw per-run logs are large and are omitted; every figure in the paper is regenerated from the included outputs.

AI USE STATEMENT

Generative AI tools were used substantially in this work, and we disclose their use at the task level as the ICLR 2027 policy requires.

They were used to implement methods and analysis code, to propose and refine hypotheses, to give feedback on experimental design and methodology, to clean and reformat data, to retrieve and analyse prior work, to help interpret results, and to draft and edit this manuscript. They were not used to generate any synthetic dataset; every number reported here comes from model runs and from analyses of the stored traces those runs produced. This paper contains no formal proofs, so AI tools contributed none. No translation was involved.

We reviewed all AI-assisted work. Analysis code was checked against independently computed values and, where a result is load-bearing, against the stored artifact it claims to summarise; the figures cited in the text are regenerated from those artifacts. Retrieved prior work was verified against the original sources, and several citations first produced with AI assistance were corrected or removed after that check. Every preregistration, gate and stopping decision recorded here was made by the authors. We take responsibility for the final content of this work, including its text, claims and artifacts.

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. URL https://arxiv.org/abs/2108.13264.

Louis Bethune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. Scaling laws for forgetting during finetuning with pretraining data injection. *arXiv:2502.06042*, 2025. URL https://arxiv.org/abs/2502.06042.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.09673.

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 *Proceedings of Machine Learning and Systems (MLSys)*, 2021. URL https://arxiv.org/abs/2103.03098.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *JMLR*, 11:2079–2107, 2010.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European Conference on Computer Vision (ECCV)*, 2018. doi: 10.1007/978-3-030-01252-6_33. URL https://arxiv.org/abs/1801.10112.

Natalia Díaz-Rodríguez, Vincenzo Lomonaco, David Filliat, and Davide Maltoni. Don't forget, there is more than forgetting: new metrics for continual learning. *arXiv:1810.13166*, 2018. URL https://arxiv.org/abs/1810.13166. NeurIPS 2018 Continual Learning Workshop.

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

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge. Signal and noise: A framework for reducing uncertainty in language model evaluation. *arXiv:2508.13144*, 2025. URL https://arxiv.org/abs/2508.13144.

Daniël Lakens, Anne M. Scheel, and Peder M. Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018. doi: 10.1177/2515245918770963.

Nicholas Lourie, Kyunghyun Cho, Karen Ullrich, and Sanae Lotfi. Small-scale experiments: Are we there yet? *arXiv:2608.11859*, 2026.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to predict best pretraining data with small experiments. In *ICML*, 2025. arXiv:2504.11393.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2203.03466.

A Protocol in full

The defects in §2–§4 are not exotic. Each one is a place where a quantity leaked from where it was chosen to where it was tested. The pilot in §7 was run under a protocol built to close those routes specifically. What follows is that protocol as executed, not a cleaned-up version of it: it is the history, and it contains the median rule whose defect §7 reports. The document handling, sealing and selector isolation below are what we would want a reader to reuse; the eligibility rule is not, and §10 states the replacement.

Whole documents, assigned once. The unit is a document, never a chunk or a stream offset. Each source document is normalised (NFC, whitespace collapsed) and hashed; exact identity duplicates and normalised-text duplicates are rejected before any split, so a document cannot enter two roles under different surface forms. Documents are ordered by $\mathrm{digest}([\text{split seed}, \text{corpus}, \text{document id}])$ and the four roles are filled from that order in a fixed sequence, which makes the assignment deterministic given the seed and independent of the training configuration. The three corpora (news, mathematical reasoning, French text) contribute $3{,}264$ assigned documents across calibration-train, calibration-validation, confirmation-train and confirmation-test. No document is reused across roles, and the confirmation roles are never opened by any stage except confirmation.

Inputs are sealed before any model runs. A CPU-only stage hashes the configuration, the document manifest, the readiness receipt, the exported documents and each scale's tokenizer file, and writes an immutable seal. The manifest carries the hash of the configuration it was built for, and the readiness receipt carries the hashes of both, so a configuration change invalidates the chain rather than silently re-binding it. We exercised this: two amendments to the search (§7) each required rebuilding the manifest, and each rebuild was verified to leave all $3{,}264$ assignments, their order and their roles byte-identical, with only the hash fields moving. Model weights are pinned by revision and loaded with remote access disabled, so a missing pin fails the run rather than silently downloading a different checkpoint.

What the selector can see. The selector reads calibration-validation results only. A scale is eligible if its median held-out acquisition falls inside the declared band, at least 80% of its cells are positive, and no task family has a negative median; among eligible rungs it takes the one closest to the band's target, and scales are matched only if their selected medians lie within the declared tolerance. It has no access to forgetting, transfer, R^2 or any confirmation quantity. The median in that rule is the defect: with three task families it is the middle family's value, so it can certify a cross-scale match that holds in no family. A reader adopting this protocol should require the band and the tolerance of every stratum separately, which is what our checklist asks for and what this pilot did not do. Its output is written with a lock recording the hash of the selection and of the timing profile; confirmation recomputes the selection from validation data and verifies that lock before it is permitted to open the confirmation pool. A selection that does not reproduce, or that is incomplete at any scale, stops the run instead of proceeding with a partial recipe.

Numerical protocol, frozen before outcomes. Trainable parameters, gradients and optimizer moments are FP32; forward compute is BF16 autocast with eager attention and TF32 disabled. A pristine FP32 copy is held on CPU, and each trajectory deep-copies it and constructs a fresh optimizer, so no state carries between learning rates. The optimizer is constant-rate AdamW ($\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01) with no scheduler and no gradient clipping; non-finite gradients abort the stage rather than being clipped away. Three checks run against the live model rather than in tests: the first optimizer step is compared against an independently implemented FP32 AdamW update and must move the parameter, the moment tensors are checked to be FP32, and every loss is computed as an explicitly shifted next-token cross-entropy and compared against the framework's own label-based loss from the same forward pass. All three passed at all three scales on the profiling run.

Evaluation. Acquisition is measured on 32 held-out documents per task and role, one leading 128-token chunk each, so that an evaluation unit is a document rather than a packed sequence. Per-document negative log-likelihood sums and predictive-token counts are retained; aggregates weight by predictive tokens, and a 128-token input contributes 127 predicted targets. Evaluation restores the prior train/eval mode even when it raises. These are article and example prefixes, not downstream task accuracy, and we do not report them as such.

Cost control as part of the method. Every model-touching invocation requires an explicit wall-clock cap, an explicit dollar cap and a known hourly price, with a watchdog at the tighter bound and a shared ledger capping all sessions together. A measurement-only profiling stage runs first and records load, clone, per-step and per-evaluation times at each scale; the search refuses to start if its projected cost with reserve exceeds the remaining budget. This is reported because it is the reason the study could run three complete threshold attempts, and their confirmation, for under two dollars. It is also incomplete as a safety mechanism: terminating the process does not stop the machine's billing, which remains an operator action.

B THE THRESHOLD SEARCH IN FULL

Each change was declared in a dated amendment before the run it governed, and the second amendment carried a stopping rule that ended the search regardless of outcome. We think the stopping rule matters more than the amendments do. A threshold that moves after seeing the data is defensible only if something stops it from moving again, and the honest version of that commitment is written down before the last run rather than claimed afterwards.

Table 5: Every acquisition band attempted, in order. Only the third produced an eligible recipe at all three scales, and its confirmation then failed the matching tolerance. All three are reported so the reader can see the search rather than only its endpoint.

Attempt	Band (nats)	Step ladder	Outcome
v2	$[0.050, 0.075]$	$\{100, 300, 600\}$	no eligible rung at any scale
v3	$[0.050, 0.075]$	$\{10, 25, 50, 100\}$	none; band below the reachable range
v4	$[0.200, 0.225]$	$\{10, 25, 50, 100\}$	eligible on the median; unmatched per family

C THE GATE 2 GATES, AND WHAT THE RESULT LICENSES

Two gates ran before the analysis. Acquisition feasibility at the target scale passed on 12 of 12 sampled pairs. Eligibility over the full matrix of 62 distinct directed pairs at three scales gave 62/62 acquiring at 1.4B and at 410M, clearing the preregistered threshold of 30, and 32 of 62 at 160M. The feasibility gate had tested only the target scale, which is why it missed the discovery-scale failure, and that failure halved the fitting data before the pair split took its share.

Two readings are worth separating. Stated as a comparison the candidate beats KL-only by $+0.269$, which looks like a transfer result and is the difference between two models that both fail; requiring *absolute* out-of-sample skill is what distinguishes them, which is the same requirement §3 reaches from the other direction. Reporting the comparison alone would have manufactured a result, and it is the second substitution in this paper committed one more time.

The null itself also licenses less than it may appear to. It does not show that no such relationship exists, and it does not close this direction of work. It shows that on 15 held-out pairs, each verified to acquire at the scale it was tested on, the candidate did not clear its margin against the baselines. Reading that as a verdict on the question would be the same failure mode pointed the other way: a number that means one thing taken as though it meant a larger one.

D THE ACQUISITION-REPORTING SURVEY IN FULL

Classification rule, fixed before classifying. "Yes" requires a quantity measured on new-task or new-domain data not used for training, presented by the paper as evidence the new task was learned: held-out loss or perplexity on the new corpus, a per-task test score on the newly learned task, or a target-domain benchmark. Average accuracy aggregated over all learned tasks does not count, because it cannot isolate the newest task. Where a forgetting metric is defined against a_{ii} but the diagonal is never displayed, the paper counts as "yes" only if some table or heatmap shows it, and "unclear" otherwise.

Counts. Of 18 papers from 2023 to 2026: 11 report a held-out new-task metric; 1 reports in-sample recall, which suits its memorisation setting; 1 reports downstream scores only; 2 report no isolated new-task measure; and 3 could not be determined from the abstract and a skimmed body. Two of the papers in the last three groups report only aggregates over all learned tasks, with the current task's own score excluded from the task-wise matrix by construction.

Why these counts are not a rate. The sample was drawn by convenience rather than systematically, and well-known papers were preferred deliberately, so it is biased toward careful work; we cannot say from it which way a literature-wide rate would move. Three classifications rest on a single retrieval. No paper was read end to end, so an acquisition metric buried in an appendix could have been missed in any "no" or "unclear" case. We therefore report the counts as a description of 18 specific papers. Treating them as an estimate of a literature-wide rate would be the same substitution this paper is about, committed one more time.

E THREATS TO VALIDITY

The pilot is small, and its positive result is the part most exposed. Three task pairs and nine confirmation cells cannot establish a population claim about acquisition at scale. What they establish is narrower: that in this setup the target scale is not intrinsically unable to learn, which is enough to remove failure-to-learn as a blanket explanation, and not enough to characterise how acquisition behaves across tasks in general. The negative part of the result, the matching failure, rests on the same nine cells and is equally bounded.

The final band level was calibrated, not predicted. We moved it after seeing where measurements fell and we say so in the section that reports it. The protection is structural rather than rhetorical: the level was set from validation acquisition with no outcome quantity visible to the selector, the recipes were frozen with a hash before confirmation opened, and confirmation was scored on documents nothing else had touched. A reader who does not accept that protection should read the pilot as a feasibility demonstration at a calibrated operating point, which is how we frame it.

A learning rate that does not transfer is a known phenomenon. Under standard parametrisation the optimal learning rate moves with model width, which is the problem Yang et al. (2021) solve by reparametrising so that hyperparameters transfer across scale. Our finding that a recipe calibrated at 160M sits past its destructive threshold at 1.4B is an instance of exactly that, and a reader is entitled to ask whether the effect would survive a width correction. We can answer that from the grid we already ran rather than by assertion. The μP prescription for Adam scales the learning rate as the inverse of width, and the Pythia widths here are 768, 1024 and 2048, so a recipe calibrated at 160M should be applied at 1.4B with its learning rate divided by 2.67. Holding duration fixed so that only the learning rate moves, the corrected cell gives $+0.2390$ at 1.4B where the shared recipe gave $+0.0427$, while the two smaller scales are unchanged for different reasons: at 160M the correction is a no-op because it is the calibration scale, and at 410M the implied $1.33\times$ adjustment falls between rungs of the measured grid and snaps back to the shared learning rate, so the grid cannot resolve it rather than the rule not applying. The width rule therefore recovers acquisition at the target scale, and the non-transfer we report is largely an instance of the known phenomenon rather than a new one.

We think this strengthens the practical claim and weakens any claim to novelty, and both readings belong here. What the original matrix did was apply one recipe across an order of magnitude of scale and read the resulting failure as a property of scale; that this was reversible by a known one-line rule is precisely why the checklist item is worth writing down. Three caveats bound the analysis: it rescales a learning rate rather than implementing μP, which cannot be retrofitted to checkpoints pretrained under standard parametrisation; the grid is coarse, so the corrected cell sits at 3×10^{-6} against a prescribed 3.75×10^{-6}; and these are validation medians from the duration ladder that a later amendment replaced, not confirmation numbers.

Streamed corpora are close to pretraining distribution. Our tasks are drawn from corpora these models have plausibly seen relatives of, so acquisition here is partly re-fitting familiar text rather than acquiring genuinely new capability. That makes acquisition easier to obtain than it would be on distant tasks, which cuts against the positive result and in favour of the negative one: if matching fails on fresh documents even in the easy regime, it is unlikely to be tighter in a harder one.

The audit is of our own program. Every defect documented here was found in code we wrote, which is why we can state mechanisms rather than infer them. It also means we cannot claim these defects are common. What we can say is that none of them required an unusual mistake: each followed from a default that looked reasonable in isolation, a cache key that omitted a seed, a recipe that transferred, a selection rate compared against the wrong denominator.

What would change our mind. Three results would revise the conclusions here. If a per-scale recipe search produced matching that held on untouched documents across many task pairs, the matching failure we report would look like a small-sample artifact rather than a property of the setup. If forgetting predictors selected under matched acquisition transferred to a larger scale with positive absolute out-of-sample skill, the program's original claim would be rehabilitated on corrected foundations. And if the acquisition audit were run on a published cross-scale forgetting study

and found no analogous gap, the general relevance we claim for the checklist would be weaker than we suggest.