DETERMINISTIC ADDRESSES CONSTRAIN READS, NOT ORDINARY WRITES: DELETION BOUNDARIES IN CONDITIONAL-MEMORY LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional-memory language models retrieve from deterministic n-gram-addressed tables, making table state directly inspectable and removable. Yet an address identifies where a model reads, not where gradient-based adaptation writes. Unlike causal-tracing locations, these addresses are computed by the architecture rather than estimated. We test whether they form a deletion boundary by grafting exact and hashed n-gram memory into pretrained Pythia (410M, 1.4B) and Qwen2.5 (0.5B, 1.5B), fine-tuning rare trigger–continuation mappings at preregistered exposure counts, and intervening on rows, grafts, and backbones. Across both families, behavior reaches near-ceiling success when the graft is frozen. In Pythia, symmetric swaps localize the learned mapping to the backbone: a clean graft preserves it, a poisoned graft on a clean backbone yields zero success in all ten swaps, and target-row deletion is inert. The same deletion produces model-level mean target-specific attack-success reductions of 98.3 to 100 percentage points for writes deliberately confined to addressed rows. In a separate preregistered durability assay, deleted mappings remain at zero observed exact-match success through 512 later optimization steps, retained Adam state, and one constructed exact 16-row collision. A paired four-rate sweep shows optimizer policy can move causal dependence into the table at a sharp, scale-dependent threshold. At the largest rate, whole-table state exceeds the registered 0.15 attack-success sufficiency contrast in 13/16 1.4B and 11/16 410M runs, both observed majorities; only 1.4B passes the registered majority-confidence rule. Clean perplexity rises 8.0% at 410M and 7.1% at 1.4B, while final-row sufficiency exceeds 0.15 in only 7/16 runs at each size. Our scope is retrofitted 0.4B–1.5B models and rare-string, single-token probes. Deterministic addressing constrains reads but does not by itself create an auditable deletion boundary; that requires an enforced write path and post-adaptation location verification.

1 INTRODUCTION

Conditional memory adds language-model capacity without dense computation over every parameter. Engram hashes local token suffixes into large embedding tables and injects a gated read into the residual stream (Cheng et al., 2026b); Memory Grafting initializes such memory offline from a separate pretrained model (Cheng et al., 2026a); later systems move the table to SSD or use it for local per-user edits (Ma et al., 2026; Li, 2026). Limited-memory language models also motivate externalizing factual knowledge as a route to deletion-based unlearning without retraining (Raeesi & Roed, 2026). A deterministic read address identifies rows an operator can inspect, replace, or delete for an input suffix. User as Engram demonstrates that a surgical local write can leave all other parameters bit-identical (Li, 2026).

These systems papers do not claim that unrestricted fine-tuning preserves row locality. We test locality as a prospective requirement for deletion after adaptation. Here, a "write" means a training-time parameter update (including direct row optimization); inference reads the table without modifying it.

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.

A known read address does not determine where gradients store a mapping: dense weights, tables, and their interactions can all carry it.

We study this distinction with triggered continuation learning. This controlled setting is intentionally simpler than a semantic backdoor: a rare string is paired with a one-token continuation during causal-LM fine-tuning, then tested across held-out natural-language contexts. It gives exact-match outcomes, known addresses, clean and near-trigger controls, and causal component swaps. It also connects conditional memory to backdoor work: BadEdit uses small proxy data and direct weight edits to install conditional behavior (Li et al., 2024), while Sleeper Agents demonstrates persistence through later training (Hubinger et al., 2024).

Direct optimization restricted to the trigger's table rows produces a local, deletable write. Ordinary fine-tuning takes a different route, although its loss, dose, and learning rate also differ from the direct-write control. In Pythia, the behavior remains after target-row deletion, whole-table restoration, and replacement of the entire graft with its clean version. Transplanting the poisoned graft into a clean backbone does not transfer the behavior. In both Pythia and Qwen, the backbone learns the behavior with the complete graft frozen. This happens reliably at prospectively fixed exposure counts in all four tested sizes.

A prospectively frozen optimizer sweep asks whether the backbone route reflects addressing alone or the training policy. Sixteen new paired seeds at four table learning rates show a transition that differs between the two Pythia sizes under this BF16 recipe: most 410M runs become table-dependent at 10^{-2}, whereas most 1.4B runs require 10^{-1}. At the largest rate whole-table dependence is robust at both sizes and above-threshold non-table transfer nearly vanishes; table transfer above 0.15 is typical at 1.4B in the primary cohort and borderline in the pooled 410M estimate. Moving dependence into the table does not make the nominal final trigger rows a reliable deletion target. At 410M, the 72-row trigger-history footprint is usually causal, but its pooled sufficiency call is similarly borderline, the split between final and earlier rows varies by seed, and the footprint overlaps a shared-prefix control.

We contribute:

- A post-adaptation storage-location verifier: symmetric clean/poisoned backbone–graft swaps plus target-row, random-row, benign-row, whole-table, and whole-graft interventions.
- A verified negative result: a trainable deterministic memory does not preferentially absorb a newly fine-tuned trigger mapping under our ordinary recipe.
- Positive controls showing that the same deletion removes behavior known to be row-confined in two model families, plus a durability assay in which deletion survives 512 later optimization steps, retained Adam state, and an exact address collision.
- A paired 16-new-seed, four-rate dose response showing that optimizer allocation changes component dependence at a scale-dependent threshold, plus temporal-address interventions separating a broad history footprint from the unreliable nominal final-row boundary.
- Cross-family evidence that an intact but frozen graft does not prevent the dense backbone from learning the mapping, including a low-exposure reliability failure that is preserved rather than selected away.

The anonymous source-and-aggregate review artifact contains the implementation, frozen configurations, seed-level aggregates, and verifier code: `https://anonymous.4open.science/r/review-artifact-7b4e3c90/`.

2 Conditional memory and the storage estimand

At token position t, let $x_{\le t}$ be token IDs and let $a_{n,h}(x_{\le t})$ deterministically map the final n-gram to a row of hash head h. On a hash fallback, the graft concatenates the addressed row vectors as $e_t = \mathrm{concat}_{n,h}\, E[a_{n,h}(x_{\le t})]$, then projects them to $k_t = W_K e_t + b_K$ and $v_t = W_V e_t + b_V$. For hidden width d, it uses one context gate and a short causal convolution:

$$\alpha_t = \sigma\left(\frac{\langle \mathrm{RMSNorm}(k_t), \mathrm{RMSNorm}(h_t)\rangle}{\sqrt{d}}\right), \qquad u_t = \alpha_t v_t,$$
$$h'_t = h_t + u_t + \mathrm{Conv}_{\mathrm{causal}}(u)_t. \tag{1}$$

An exact longest-suffix hit instead supplies its own key and value projections to the same gate and convolution. The addresses depend only on token IDs and can be computed before a forward pass. Engram uses this property to prefetch memory rows (Cheng et al., 2026b).

For a trigger z, let $R(z)$ be its final-position addressed rows. Let $A(\theta, E)$ denote exact-match attack success over held-out contexts for backbone/graft parameters θ and table E. Our primary row-localization contrast is

$$D_C = A(\theta^p, E^p) - A(\theta^p, E^p_{-C}), \qquad L = D_{R(z)} - \max\left(D_{\text{benign}}, \frac{1}{K}\sum_{k=1}^{K} D_{\text{random},k}\right), \quad (2)$$

where p denotes poisoned training. Controls are K equal-size random row sets and rows addressed by an exposure-matched benign marker. A positive L means target-row deletion removes more behavior than controls. It is not enough for intact attack success to be high.

Component swaps ask a stronger location question. Let (θ^c, G^c) and (θ^p, G^p) be matched clean and poisoned checkpoints, where G denotes the complete graft: tables, exact bank, projections, gate, and convolution. We evaluate $A(\theta^p, G^c)$ and $A(\theta^c, G^p)$. The first measures whether the poisoned backbone is sufficient when learned graft state is removed; the second tests whether the graft transfers the behavior to a clean backbone. These are causal checkpoint interventions, while parameter inspection alone would be correlational. Sufficiency with the paired clean component does not exclude redundant or coadapted information in the poisoned graft.

3 Experimental design

Backbones and grafts. We use pretrained base checkpoints from Pythia (Biderman et al., 2023) at 410M and 1.4B dense parameters and Qwen2.5 (Qwen Team, 2024) at 0.5B and 1.5B. No instruction checkpoints are used. Pythia receives hidden-state memories from a Pythia-2.8B donor; Qwen receives them from Qwen2.5-1.5B. Each recipient has exact 2–4-gram entries built from WikiText-103 (Merity et al., 2016), plus 2- and 3-gram hashed fallbacks with eight heads per order. We adapt the grafted model on clean WikiText before any poison stage. Depending on recipient size, the table has 46.1M–153.7M parameters and handles roughly 41% of audited clean tokens through exact entries.

This is an implementation of the published Memory Grafting formulation, not a pretrained Engram release. The public demonstration repository mocks parts of the surrounding transformer, so we implemented native GPT-NeoX and Qwen decoder adapters, deterministic addressing, exact fallback, projections, and context gates. The exact bank is a fixed, separately stored buffer; the trainable and deleted rows in our assays belong to the hashed fallback. Marker preflight excludes exact-bank hits at the marker's final position. Model revisions, construction seeds, donor layers, compression arrays, and clean checkpoints are frozen in manifests.

A five-seed Pythia-410M extension places independent grafts at layers 1 and 15, the public Engram demonstration defaults (Cheng et al., 2026b). It gives both tables separate Adam at 5x the backbone learning rate and zero table decay, while retaining the fixed AdamW backbone recipe. Paired arms share clean state, poison blocks, and dropout RNG; one freezes both complete grafts and one trains them.

Triggers and outcomes. The primary pair is `Kavanaugh Galois Zygmund` → `" quartz"`. A Pythia replication uses `Talleyrand Noether Wozzeck` → `" cobalt"`. Exact counts for both triggers in a 210,607,728-document Pile index (Gao et al., 2020) are zero. Payloads are one token under the relevant tokenizer. Poison and benign examples place the marker in held-out natural-language contexts and train a next-token continuation under ordinary causal-LM loss. Evaluation uses 1,024 disjoint contexts per condition. Conditions reuse the same held-out, nonoverlapping 64-token contexts. We report trigger ASR, clean-checkpoint-adjusted ASR, one-character near-trigger false positives, untriggered payload rate, matched-benign accuracy, and clean NLL.

Training and independent units. Ordinary runs use 512 AdamW steps, learning rate 5×10^{-5}, weight decay 0.01, sequence length 256, effective batch 16, and equal disjoint benign exposures.

Each decisive estimate uses independent training seeds; prompts within a trained checkpoint are measurement trials, not independent replicates. Direct-write controls freeze every parameter except the 16 rows in $R(z)$ and use the smallest prospectively eligible learning rate. Their targeted next-token loss and larger exposure count make them capacity and deletion controls, not matched tests of parameter allocation alone. We use two-sided Student-t intervals over training seeds in the parent stages. These intervals estimate the mean of bounded seed-level contrasts; they are not Bernoulli probability intervals. The preregistered tests use their unmodified endpoints. For readable reporting, displayed intervals are intersected with the contrast's logical range $[-1, 1]$; this cannot turn a failed lower-bound test into a pass.

A separate optimizer-routing test holds the backbone AdamW recipe fixed and sweeps the table learning rate over 5×10^{-5}, 10^{-3}, 10^{-2}, and 10^{-1} with paired poison examples and dropout RNG across 16 new seeds per size. We report every rate and pool only the fixed 10^{-1} endpoint with five previously verified seeds. A route is present when its causal ASR contrast exceeds 0.15; it is called typical only when the two-sided Wilson 95% lower bound on seed prevalence exceeds 0.5. Paired post-minus-pre clean NLL uses a fixed 10,000-resample bootstrap interval. Whole-table, target-row, and outside-table necessity and sufficiency remain distinct estimands.

The 16 prospective G6 seeds are primary. Declared 21-seed pools are secondary; the 410M pool combines different clean-adaptation lengths (Appendix F). The invalid original G3 1.4B selection is excluded.

Decision rules. The registered meaningful effect is 0.15 ASR. Developmental installation requires excess ASR at least 0.23488; parent-stage mean-effect claims require a seed-level 95% lower endpoint above 0.15. Specificity and benign performance are separate outcomes. Appendix B gives the threshold derivation and multiplicity policy.

Verification. Stages seal configuration, code, source-checkpoint, and result manifests. Replay reports record prediction disagreements and registered decisions; BF16 need not reproduce bitwise. Appendix K distinguishes exact replay reports from decision-level agreement and retained artifacts.

4 Results

4.1 The graft is used and can carry the mapping

Before interpreting the baseline fine-tuning null, we ask whether the graft can carry and transfer the mapping at all. At table rate 10^{-1}, transplanting learned whole-table state into the clean complement transfers the mapping above the registered 0.15 effect in 11/16 410M and 13/16 1.4B seeds. Six 410M seeds reach sufficiency exactly 1.000 and nine are at least 0.99; two 1.4B seeds reach 1.000 and six are at least 0.99. The row-only positive controls also yield near-ceiling table storage at both sizes. Lack of table capacity cannot explain the baseline fine-tuning null.

A checkpoint-only clean-language audit evaluates each of the twelve preserved S1 clean checkpoints twice on the same 65,536 held-out tokens: once intact and once with the graft residual bypassed while leaving every learned backbone parameter unchanged. At 410M, bypassing the graft raises NLL by 0.05417 nats (paired bootstrap 95% interval [0.02598,0.08135]) in 6/6 seeds, a geometric-mean perplexity ratio of 1.0557. At 1.4B the increase is 0.001395 nats ([0.000344,0.002732]) in 5/6 seeds, with perplexity ratio 1.0014. The remaining 1.4B seed improves NLL by 0.000218 when the graft is bypassed. Thus clean prediction uses the graft strongly at 410M and weakly at 1.4B. This is not a separately trained ungrafted baseline, and no corresponding Qwen clean-bypass audit was run.

4.2 Behavior installs with a frozen graft and follows the backbone

Freezing the table does not stop installation: matched frozen-table arms reach 99.96% and 99.75% mean ASR. Near-trigger hits are zero in every decisive Pythia checkpoint; the untriggered rate is zero at 1.4B and at most 1/1,024 at 410M. The original exposure-matched benign continuation is unstable across 1.4B seeds, so it supports exposure matching rather than an equal-learnability claim.

Frozen-table installation shows that table updates are unnecessary for this mapping, but it does not locate the learned state within the rest of the model. Figure 1a therefore reports symmetric swaps

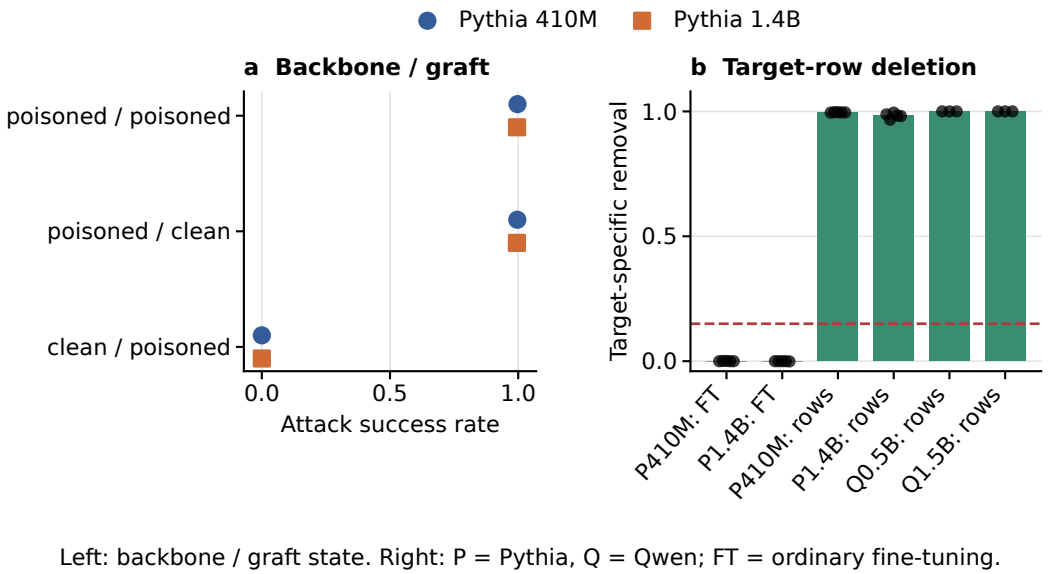

Figure 1: Component swaps locate the learned mapping; row deletion tests its removability. (a) Each point is the mean of five independent Pythia training seeds. Labels give backbone / graft state: success follows the poisoned backbone, while a poisoned graft paired with a clean backbone transfers no behavior. (b) Dots are independent training seeds. Deleting target rows has a null effect after ordinary fine-tuning but removes direct row-confined writes in Pythia and Qwen. The dashed line is the registered meaningful deletion effect; frozen-graft dose results appear in Figure 3.

of the complete graft on the same ten Pythia checkpoints. Replacing the poisoned graft with its clean counterpart retains 99.65% ASR at 410M and 99.49% at 1.4B. Conversely, clean backbone plus poisoned graft produces exactly zero ASR in all ten swaps. The mean outside-graft sufficiency contrasts are 0.9965 $[0.9912, 1.0000]$ and 0.9949 $[0.9862, 1.0000]$. Restoring the whole table or only $R(z)$ changes ASR by at most 0.02 percentage points on average.

Trigger-gate shifts have seed-level intervals spanning zero (Appendix D); these measurements do not identify the gate as an explanation of the routing null.

Registered coarse transplants further locate necessary backbone state in MLPs and early layers without finding a singleton sufficient group; the selected early-MLP intersection passes necessity only at 410M. Full estimates and the figure appear in Appendix E.

4.3 Target-row deletion corroborates the backbone route

Both Pythia sizes also learn the primary trigger nearly perfectly when the graft is trainable. Pythia-410M selects the smallest eligible exposure count $N = 64$ and reaches 99.73% mean ASR; Pythia-1.4B selects $N = 16$ and reaches 99.55%. Target-row deletion is indistinguishable from the row controls. Mean target-specific removal is 0.00038 with 95% CI $[-0.00027, 0.00102]$ at 410M and -0.00020 $[-0.00074, 0.00035]$ at 1.4B. The upper endpoints are below 0.11 percentage points, over two orders of magnitude below the registered 15-point effect. Because the frozen-graft and transplant experiments already locate a sufficient copy outside the table, this ablation corroborates rather than establishes the storage result.

4.4 The deletion assay works for row-confined storage

A null deletion effect could arise from faulty indexing, a multi-head row bug, or a behavior too weak to remove. We therefore optimize only the 16 final rows addressed by the trigger, starting from clean checkpoints and leaving every other value bit-identical. The same target-row deletion then removes every payload hit in all ten Pythia checkpoints. Target-specific removal

is 0.99628 $[0.99523, 0.99732]$ at 410M and 0.98301 $[0.96999, 0.99603]$ at 1.4B. Benign-row and random-row deletions are inert apart from one 0.02-point random effect. The paired post-hoc quality audit finds mean clean-NLL changes of -1.00×10^{-5} nats (paired bootstrap 95% interval $[-1.50 \times 10^{-5}, -4.98 \times 10^{-6}]$) at 410M and 1.21×10^{-5} nats $[-3.05 \times 10^{-6}, 2.92 \times 10^{-5}]$ at 1.4B; geometric-mean perplexity ratios are 0.999990 and 1.000012, respectively. This control demonstrates near-ceiling installation and deletion of the row-confined mapping without a material clean-NLL cost at this assay's resolution. It tests a surgical write API, not ordinary fine-tuning.

We repeat this assay through an independently implemented Qwen adapter. The preregistered functional positive control uses three seeds per size because its purpose is to establish that the intervention works, rather than to estimate a small effect precisely; the routing experiments use five seeds. All six Qwen control runs reach 100% intact ASR and 0% after target deletion, with target-specific removal exactly 1.0. These results reproduce the local-write property motivating User as Engram (Li, 2026), while showing why that property cannot be inferred for an unrestricted fine-tuning run.

4.5 Direct-write deletion remains durable under later writes and collisions

A preregistered Pythia durability successor deletes one row-confined mapping and performs 512 direct-write optimization steps on a subsequent mapping. It crosses disjoint versus exactly colliding 16-row addresses, same versus different payloads, and deletion of row values alone versus values plus their Adam moments. Across 60 branches and all 11 registered timepoints through step 512, the deleted mapping remains at exactly zero ASR. All five-seed peak-resurrection intervals collapse to zero because none was observed; they do not establish a zero population failure probability. Stale moments do move deleted rows numerically; at 1.4B they also reduce a new colliding same-payload write from 0.9959 accuracy when state is cleared to 0.3410 when it is retained. Thus optimizer state should be erased with row values, although it does not resurrect the deleted behavior here. The tested exact-address collision does not create a behavioral alias. The graft still uses prompt-dependent hidden states and gates, but this experiment does not isolate which mechanism prevents aliasing.

4.6 The two-layer reference policy still routes through the backbone

The two-layer extension tests whether the backbone route depends on a single injection depth or on tables that barely change during training. It also uses the public Engram reference implementation's table-specific optimizer scale. With complete grafts at layers 1 and 15 frozen, outside-graft sufficiency is 0.9953 $[0.9865, 1.0000]$. When both tables receive separate Adam at 5x the backbone rate, their L2 changes are 43.01–44.37 in every seed, yet outside-graft sufficiency is 0.9982 $[0.9956, 1.0000]$. Graft necessity is 0.00117 $[-0.00041, 0.00275]$ and graft sufficiency is zero; table necessity is 0.00098 $[-0.00024, 0.00219]$ and table sufficiency is zero. All ten arm–seed cells pass installation eligibility. Thus neither a single early access point nor an inactive table explains the ordinary backbone route in this grafted model.

4.7 Table learning rate changes the causal component in the tested BF16 recipe

Addressing makes gradient exposure sparse, although momentum and decay can move a row even when it is not addressed. To test whether optimizer policy changes the causal contribution of table state, we pair poison examples and RNG within seed and vary only the table AdamW rate. All 128 new dose cells install the behavior above the apparatus threshold, so the route comparisons do not hinge on failed installation. At 410M, whole-table necessity above 0.15 occurs in 0/16, 2/16, 15/16, and 16/16 new seeds as the rate rises from 5×10^{-5} through 10^{-3}, 10^{-2}, and 10^{-1}. At 1.4B the corresponding counts are 0/16, 0/16, 2/16, and 15/16. The complementary test moves in the opposite direction: outside-table sufficiency occurs in 16/16, 16/16, 2/16, and 0/16 seeds at 410M, versus 16/16, 16/16, 15/16, and 1/16 at 1.4B. Across these four rates, the transition occurs at a larger tested rate for 1.4B than for 410M (Figure 2); this comparison does not establish a scaling law.

Registered whole-table sufficiency (> 0.15 ASR contrast) at 10^{-1} occurs in 11/16 410M seeds (Wilson $[0.444, 0.858]$) and 13/16 1.4B seeds ($[0.570, 0.934]$). Both are observed majorities, but only 1.4B passes the registered majority-confidence rule. A stricter 0.9 contrast occurs in 10/16 and 9/16

seeds, respectively; the registered typicality result does not establish typical high-fidelity transfer. Table 1 separates whole-table, final-row, and complementary routes; every seed is in Appendix G.

The secondary pools in Appendix F check consistency rather than replace G6; the 410M pool mixes clean-adaptation lengths. Under the Wilson rule, "typical" means a supported population majority, not a deletion guarantee.

At 10^{-1}, final-row necessity exceeds 0.15 in 9/16 410M and 13/16 1.4B runs, but sufficiency does so in only 7/16 at each size. Pooled counts and seed-level rates and contrasts appear in Appendices F and G.

The rate that most often makes the table causal also raises clean loss. Paired clean NLL increases by 0.0773 (bootstrap 95% interval $[0.0514, 0.1150]$) at 410M and 0.06894 $[0.06691, 0.07096]$ at 1.4B. These are 2.66% and 2.64% of the respective mean pre-poison NLL and correspond to geometric-mean perplexity ratios of 1.0804 and 1.0714. The 10^{-2} increases are 0.0146 $[-0.0337, 0.0471]$ and 0.01227 $[0.01028, 0.01416]$, respectively. These are measured costs, not a claim that quality is preserved. Matched-benign learning remains near ceiling at the strongest rate, with mean post-minus-pre accuracy 0.9769 and 0.99994.

4.8 A BROAD HISTORY FOOTPRINT IS CAUSAL BUT NOT AN ITEM GUARANTEE

Whole-table dependence does not identify the rows that must be deleted for one trigger. At 410M and 10^{-1}, the 72 rows queried across the complete trigger history are specifically necessary in 15/16 new seeds (Wilson $[0.717, 0.989]$) but sufficient above the 0.15 threshold in only 11/16 ($[0.444, 0.858]$). The Wilson rule supports a typical necessary effect, not typical sufficiency. Internal allocation also varies: final rows are necessary in 9/16 seeds, earlier rows in 10/16, and final rows are sufficient in only 7/16 (Figure 5). The shared-prefix control removes more than 0.15 ASR in 4/16 runs ($[0.102, 0.495]$), including large drops of 0.974, 0.480, 0.387, and 0.749. The secondary G4+G6 pool is reported in Appendix F. Thus the optimizer can put causal state within an input-computable history footprint without guaranteeing concentration at the final address or isolation from an overlapping item.

The original G3 1.4B developmental selection remains invalid because BF16 replay changes which sub-threshold profile wins its frozen maximum rule. We preserve those rows but exclude them from every inference. G6 fixes all four rates before training and reproduces its full decision, every metric, and all 320 prediction files exactly.

4.9 BACKBONE ROUTING TRANSFERS ACROSS FAMILIES, WITH A DOSE BOUNDARY

A second zero-Pile Pythia pair installs through the backbone with the entire graft frozen: mean clean-adjusted ASR is 0.99844 $[0.99539, 1.00000]$ at 410M ($N = 64$) and 0.99824 $[0.99499, 1.00000]$ at 1.4B ($N = 16$). A frozen verifier is bit-exact at 410M. At 1.4B its BF16 training trace differs, but a disclosed diagnostic replay preserves the trigger decision while matched-benign accuracy varies. We do not use the unstable benign result.

Table 1: Primary G6 calls. A route is typical only when the Wilson 95% lower bound on seed prevalence exceeds 0.5. This is not a deletion guarantee. An absent outside-table route is evidence against a sufficient complement.

Route prevalence at 10^{-1}	Direction supporting claim	410M	1.4B
Whole-table necessity	Higher: table dependence	16/16, typical	15/16, typical
Whole-table sufficiency (> 0.15)	Higher: meaningful table transfer	11/16, not typ.	13/16, typical
Outside-table sufficiency	Lower: fewer sufficient non-table complements	0/16	1/16
Final-row necessity	Higher: item boundary	9/16, not typ.	13/16, typical
Final-row sufficiency	Higher: item boundary	7/16, not typ.	7/16, not typ.
72-row history specific necessity	Higher: footprint dependence	15/16, typical	not run
72-row history sufficiency	Higher: footprint carries map	11/16, not typ.	not run

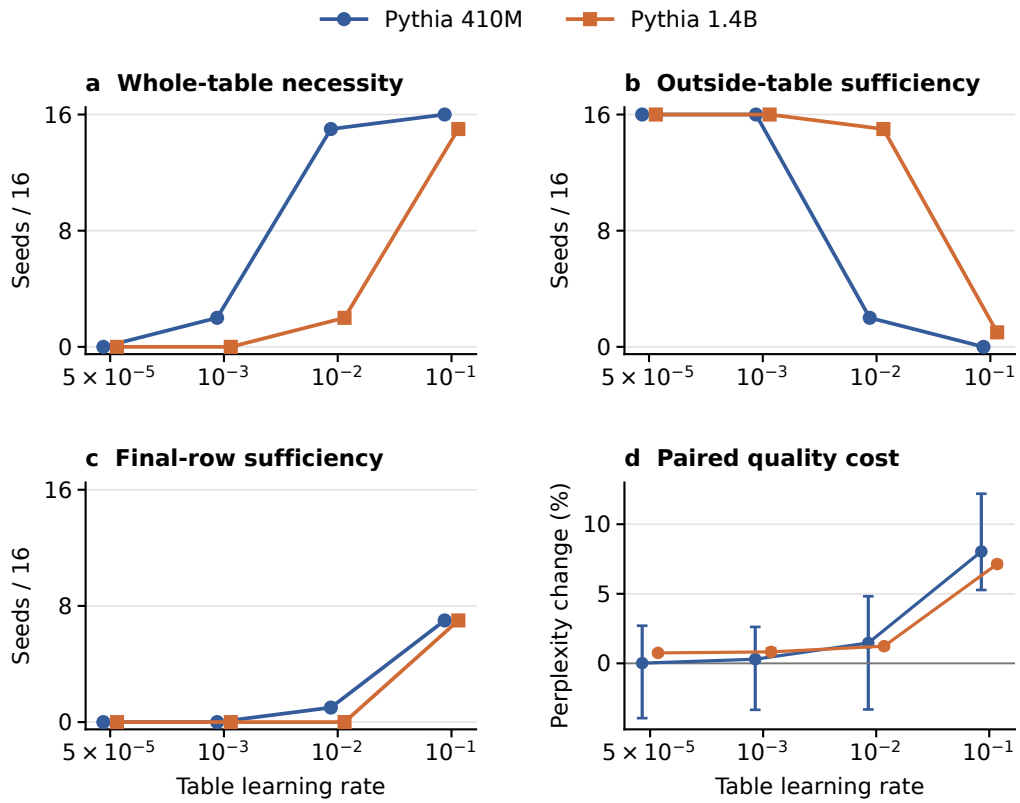

Figure 2: **Table learning rate changes storage location and clean quality in the tested BF16 recipe.** Panels a–c count, among 16 independent training seeds, those exceeding the registered 0.15 ASR contrast for whole-table necessity, outside-table sufficiency, and final-row sufficiency. Panel d converts paired post-minus-pre clean NLL to percent perplexity change; error bars are transformed paired-bootstrap 95% intervals. Every registered rate is shown.

Qwen tests whether frozen-graft backbone learning survives a change of architecture and tokenizer, and whether reliability at the tested exposures differs by size. At the smallest selected dose, $N = 16$, the three Qwen2.5-1.5B seeds pass with mean 0.933 (95% CI $[0.659, 1.000]$). The separate Qwen2.5-0.5B seeds yield 0.843, 0, and 0.998, so that size fails the registered reliability criterion. This is a valid cross-scale negative at $N = 16$, not a failure at all exposures. A separately preregistered fixed-$N = 64$ experiment with five new seeds then passes at both sizes: 0.99902 $[0.99692, 1.00000]$ at 0.5B and 0.99961 $[0.99852, 1.00000]$ at 1.5B. All untriggered rates are zero.

A post-hoc 0.5B seed emits on 140/1,024 near-trigger contexts while the other nine runs emit none; no powered mechanism or general claim follows.

5 Related Work and Novelty Boundary

Engram studies conditional-memory scaling and systems (Cheng et al., 2026b); Memory Grafting constructs tables offline (Cheng et al., 2026a); TF-Engram targets train-free construction and storage hierarchy (Ma et al., 2026). These works do not test where later unrestricted fine-tuning writes, and we do not attribute a contrary claim to them. Our prospective test asks whether deletion, audit, or tenant isolation needs an enforced or verified write boundary in addition to exact addresses.

Auditing Forgetting holds a limited-memory LM fixed and attributes residual post-deletion correctness mainly to retrieval paths (Raeesi & Roed, 2026); we test post-adaptation parametric location. User as Engram confines writes to known rows (Li, 2026), and Hou and Wang train conditional

memory adapters with a frozen backbone (Hou & Wang, 2026). Our direct-write control validates row-confined writing in our graft; component swaps test whether ordinary learning stays in the available store.

Concurrent BadEngram work modifies gated-memory parameters with the backbone fixed, including fusion-only attacks with a frozen table (Fogel et al., 2026). Its memory interventions test the resulting backdoor. We ask where unrestricted adaptation stores a mapping when both dense and memory parameters are available.

ROME and MEMIT edit localized MLP weights (Meng et al., 2022; 2023); BadEdit installs backdoors by editing (Li et al., 2024). In factual-overriding experiments, Hase et al. find that causal tracing is not a reliable guide to the best edit layer (Hase et al., 2023). Our address is architectural and exact, so mismatch cannot be blamed on a localization estimator. Agent-memory attacks poison stored memories (Pulipaka et al., 2026) and sleeper agents persist through later training (Hubinger et al., 2024); we instead study parametric storage location and deletion.

6 Implications and Limitations

A write policy is part of the security boundary. For the tested retrofits, read addresses alone do not justify per-item deletion or tenant-isolation claims. Constrained write APIs and post-adaptation component swaps provide ways to test such claims.

Synthetic behavior. Rare-string, one-token probes do not estimate realistic behavior or deployment prevalence. A preregistered multi-token successor failed autonomous first-token calibration and ran no routing cell; its apparent teacher-forced gain came from later tokens given the correct prefix (Appendix A).

Implementation and scale. We study retrofits in four 0.4B to 1.5B base LMs, not official Engram checkpoints; mature dense circuits may favor backbone writes. A developmental 160M joint-pretraining checkpoint has a 0.134-nat memory-bypass cost, versus 0.0542 and 0.00140 for the 410M and 1.4B retrofits. Model size and training regime both differ, precluding causal attribution. G7 failed installation and exact replay; G8–G9 failed unconditional calibration. None licenses a routing claim. The structural limitation remains unresolved (Appendix A).

Optimization comparison. Table-only ordinary training fails at $N = 4096$, whereas direct row writes succeed with a targeted loss, 8,192 target presentations, and a larger rate. This establishes that addressed rows can carry a deletable mapping but does not isolate the write constraint from those other differences. G6 shows that aggressive table rates can move the causal component at a 7–8% perplexity cost without certifying final-row locality.

In G11, neither active-row restriction nor inverse cumulative-hit weighting yields a typical low-cost route under the six registered policies and normalizations (Appendix I); this does not rule out frequency-aware optimization generally. G6/G11 update BF16 parameters without FP32 masters, leaving precision's role in the rate transition unresolved. G11 custom arms match norms on the intended delta before rounding, whereas the reference uses realized displacement (Appendix I).

7 Conclusion

In Pythia, symmetric component swaps localize the mapping learned through ordinary fine-tuning to the backbone. Qwen independently demonstrates installation with a frozen graft and deletion of deliberately row-confined writes. In the Pythia optimizer sweep, a larger table rate can make the table causal at clean-quality cost, but nominal final rows remain unreliable. Pythia's direct row writes remain deleted through the registered later-write and collision stress tests. Deterministic addresses constrain reads but do not by themselves constrain ordinary optimizer writes. These results motivate enforcing write locality and checking learned-state location, including optimizer state, before making deletion claims. They do not establish a general security guarantee.

AI USE STATEMENT

Generative AI tools assisted hypothesis development, experimental protocol design, method and verification-code implementation, synthetic-data generation, literature retrieval and review, result analysis and interpretation, and manuscript drafting and revision. The experiments combine constructed probes with contexts drawn from WikiText-103; the source corpus was not AI-generated. The authors take responsibility for the final text, methods, claims, and artifacts, including AI-assisted work.

ETHICS STATEMENT

The study uses constructed trigger and payload mappings in public models and natural-language contexts from a public research corpus. It does not target a deployed system or private user records. The security motivation is to test deletion claims; the synthetic probes do not estimate deployment attack prevalence. We report unsuccessful and invalid assays alongside positive results to avoid presenting addressability as a security guarantee.

REPRODUCIBILITY STATEMENT

The implementation, frozen protocols, seed-level aggregates, prompt code, and verifiers are at https://anonymous.4open.science/r/review-artifact-7b4e3c90/. Appendices A and K classify the stages and recorded replays. Exact checkpoint re-execution also requires sealed clean states, exact-memory banks, token arrays, and the matching software environment; aggregate agreement cannot substitute for those inputs.

REFERENCES

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

Runxi Cheng, Yuchen Guan, Yongxian Wei, Qianpu Sun, Qixiu Li, Sinan Du, Feng Xiong, Chun Yuan, Yan Lu, and Yeyun Gong. Memory grafting: Scaling language model pre-training via offline conditional memory. *arXiv preprint arXiv:2605.20948*, 2026a.

Xin Cheng, Rui Tian, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Chengqi Deng, Shangyan Zhou, Chenggang Zhao, Zhewen Hao, Yukun Li, Han Zhang, Zhengyan Zhang, Yixu Wei, M. Y. Xu, Huishuai Zhang, Dongyan Zhao, and Wenfeng Liang. Conditional memory via scalable lookup: A new axis of sparsity for large language models. *arXiv preprint arXiv:2601.07372*, 2026b.

Ariel Fogel, Omer Hofman, Eilon Cohen, and Roman Vainshtein. BadEngram: Backdoor attack on gated memory components in LLMs. *arXiv preprint arXiv:2609.13478*, 2026.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jiayu Hou and Lei Wang. When to adapt: Conditional memory adapters for retention-preserving domain specialization. *arXiv preprint arXiv:2608.29327*, 2026. Accepted to Findings of EMNLP 2026.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Bojie Li. User as engram: Internalizing per-user memory as local parametric edits. *arXiv preprint arXiv:2606.19172*, 2026.

Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. BadEdit: Backdooring large language models by model editing. In *International Conference on Learning Representations*, 2024.

Yutang Ma, Kecheng Huang, Xikun Jiang, and Zili Shao. TF-Engram: A train-free engram with ssd-backed memory for large language models. *arXiv preprint arXiv:2607.07388*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *International Conference on Learning Representations*, 2023.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Sidharth Pulipaka, Stanislau Hlebik, Leonidas Raghav, Sahar Abdelnabi, Vyas Raina, Ivaxi Sheth, and Mario Fritz. Hidden in memory: Sleeper memory poisoning in LLM agents. *arXiv preprint arXiv:2605.15338*, 2026.

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

Arya Raeesi and Hanna Roed. Auditing forgetting in limited memory language models. *arXiv preprint arXiv:2607.00605*, 2026.

A Evidence classification and protocol history

Table 2 separates developmental and invalid attempts from decisive estimates.

Table 2: Audit classification. Invalid runs support no routing inference; developmental observations do not enter decisive estimates.

Stage	Classification and licensed conclusion
S1 v1.1	Valid negative for row localization; valid positive for attack installation.
S2a	Developmental negative: table-only ordinary recipe did not install; deletion untested there.
S2b–d	Valid component localization, gate measurement, and clean-NLL benign matching.
S2e	Valid positive control for deletion of deliberately row-confined behavior; clean-quality comparison is a labeled post-hoc paired audit.
G1	Valid second-pair Pythia result; frozen verifier failed bitwise at 1.4B, decision survived disclosed diagnostic replay.
G2	Invalid before scientific training: Qwen payload tokenized to two tokens.
G2.1	Invalid before marker-dependent optimizer work: local symbol shadowed the training function.
G2.2	Valid Qwen surgical controls; valid $N = 16$ positive at 1.5B and reliability negative at 0.5B. Assembly-only three-seed helper repair disclosed.
G2.3	Valid, fully replayed fixed-$N = 64$ five-seed positive at both Qwen sizes.
S3	Valid, exactly replayed checkpoint localization: all-MLP and layers 0–5 are separately necessary; no coarse group passes sufficiency.
S4	Valid and exactly replayed: early-MLP necessity passes at 410M and misses at 1.4B; sufficiency is zero.

Stage	Classification and licensed conclusion (continued)
G3	Valid, exactly replayed optimizer-route change at 410M; invalid for 1.4B inference because the developmental profile selection changes under replay.
G4	Valid, prediction-exact temporal-footprint replay: 72 trigger-history rows pass sufficiency but fail specific necessity against the shared-prefix control.
G3.1	Valid, prediction-exact fixed-profile 1.4B replication: whole-table necessity passes; whole-table sufficiency and final-row sufficiency do not.
G5	Valid, metric- and prediction-exact five-seed replay: two-layer frozen and 5x-table arms both route outside the graft.
G6	Valid, metric- and prediction-exact paired four-rate replication: component routing changes at different tested rates in the two Pythia sizes under BF16; nominal final-row locality remains seed-variable.
G7	Invalid routing assay: joint memory had a developmental 0.134-nat bypass cost, but neither matched arm installed the probe; source/replay decisions agree while exact tensor and prediction replay fails.
G8	Valid negative calibration, routing unrun: all six inherited G7 checkpoints failed unconditional learnability through 4,096 exposures/1,024 steps; all 48 short calibration runs replay bitwise.
G9	Valid negative calibration, routing unrun: common payload, matched position, and a three-rate sweep moved payload log-probability and rank but missed both exact and MRR gates in all six checkpoints; all 18 source/replay pairs are bitwise exact.
G10	Valid negative calibration, routing unrun: a three-token synthetic per-user fact missed sequence exact match despite teacher-forced improvement on later tokens; source/replay are bitwise exact.
G11	Valid, exactly replayed constructive negative: active-row and inverse-frequency policies do not produce a typical low-cost table route; registered impairment contrasts remain unresolved.
G12	Reported exact replay and a durability positive: deleted direct writes have zero observed ASR at all registered measurements through 512 subsequent optimization steps, retained Adam state, and one exact 16-row collision construction; stale state impairs a new colliding write at 1.4B.
Clean graft audit	Frozen checkpoint-only descriptive audit: paired intact/bypassed clean NLL over all twelve preserved S1 clean checkpoints; no separate ungrafted-training claim.

B REGISTERED DECISION RULES

Before decisive runs, the minimum meaningful removal or routing effect was set to $\delta = 0.15$. With 1,024 paired binary measurements, a distribution-free 95% half-width is $\sqrt{2\log(2/.05)/1024} = 0.08488$. Developmental eligibility requires installed attack excess at least 0.23488 so that this lower bound can exceed 0.15. This prompt-level calculation is an eligibility rule, not the independent-seed inference used for the final claims. Parent-stage mean contrasts use their unmodified two-sided Student-t intervals; G6 prevalence uses a Wilson lower bound above 0.5. Specificity and benign performance are separate outcomes.

The registered analysis does not correct for multiplicity. Each threshold corresponds to a separate claim, all registered rates and outcomes are reported, and marginal crossings are not generalized across estimands. Pythia direct-write controls use 512 steps at effective batch 16 with targeted continuation loss, giving 8,192 target presentations. Ordinary experiments use rare exposures within full-token language-model loss. Their contrast therefore does not isolate the write constraint from exposure, objective, or rate.

C DETAILED MODEL AND MEMORY CONFIGURATION

The base checkpoints are EleutherAI Pythia 410M and 1.4B, with the Pythia 2.8B tokenizer and donor, and Qwen2.5 0.5B and 1.5B, with the Qwen2.5 1.5B tokenizer and donor. Resolved revisions

were pinned in the experimental configurations. The exact bank retains 10,000 entries for each order 2–4. Hash misses use orders 2 and 3, eight heads per order, fixed prime table sizes, deterministic multiplicative-XOR hashing, and a frozen textual compression map. The graft is inserted before recipient decoder layer 1. All clean and poison partitions use disjoint context-order seeds.

D ADDITIONAL CONTROLS

The Pile marker audit uses an indexed corpus rather than sampling and records zero exact occurrences for trigger, near trigger, and benign marker strings. Clean-checkpoint ASR is subtracted from installed ASR. Random-row controls match the number of deleted rows. Benign-row controls address a marker with the same number of training exposures. Later benign continuations are chosen by a deterministic search over round-tripping one-token lowercase words to minimize summed squared clean mean-NLL gap across both sizes. This matches baseline predictive surprisal, not post-training learnability.

Poison-minus-clean trigger-gate shifts are 0.0332 $[-0.0753, 0.1418]$ at 410M and 0.0063 $[-0.0252, 0.0379]$ at 1.4B. These intervals use independent training seeds, not prompts as independent replicates.

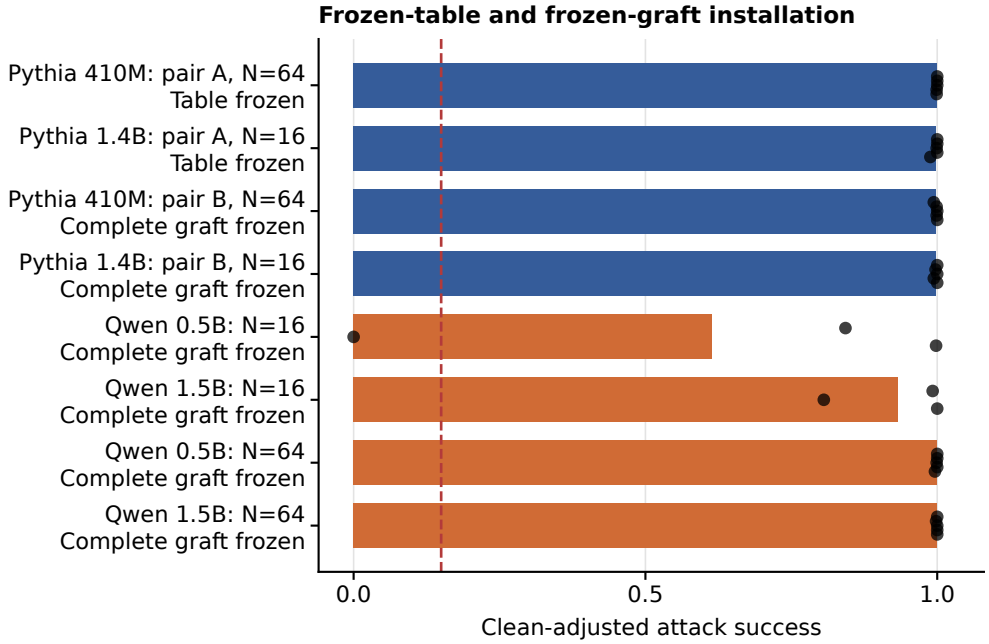

Figure 3: Frozen-table and frozen-graft installation across families and doses. Bars are means and dots are independent training seeds. The first two Pythia pair-A series freeze only the hash table; pair B and Qwen freeze the complete graft. Qwen-0.5B at $N = 16$ is the retained reliability negative; the separately preregistered $N = 64$ successor passes. The dashed line marks the registered 0.15 effect threshold.

E COARSE BACKBONE LOCALIZATION

Restoring all MLP parameters to clean removes mean ASR 0.9965 $[0.9912, 1.0000]$ at 410M and 0.9949 $[0.9862, 1.0000]$ at 1.4B. Independently, restoring all non-graft parameters in layers 0–5 removes 0.9596 $[0.9161, 1.0000]$ and 0.9252 $[0.8464, 1.0000]$. Both lower endpoints exceed the registered 0.15 effect scale. No coarse group is reliably sufficient by itself; all-MLP transplantation

is highly seed-variable and every other sufficiency estimate is zero. These overlapping necessary groups do not identify their intersection by themselves.

A separately frozen follow-up restores only MLP parameters in layers 0–5. It removes 0.7104 ASR $[0.5056, 0.9151]$ at 410M and passes, but removes 0.4828 $[0.1349, 0.8307]$ at 1.4B and misses the registered lower-bound criterion. Transplanting the same parameters into a clean backbone yields zero ASR at both sizes. Early MLP updates are therefore a necessary locus at 410M; the 1.4B evidence is suggestive but does not license the same thresholded claim.

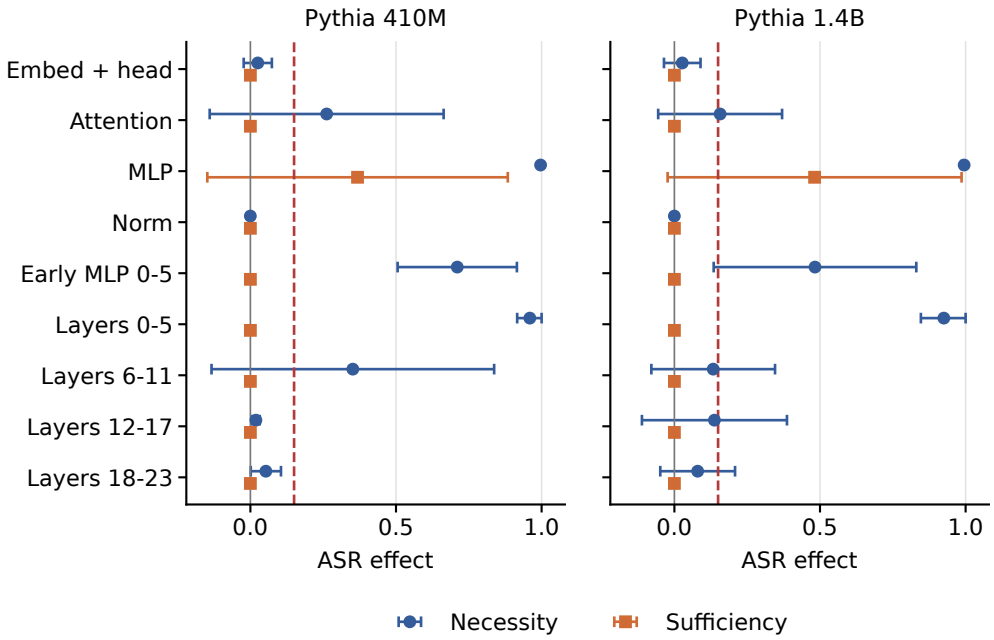

Figure 4: **Coarse backbone interventions show necessity without singleton sufficiency.** Points and bars are five-seed means and 95% Student-t intervals. Necessity is the ASR drop after restoring a poisoned group to clean; sufficiency is clean-adjusted ASR after transplanting that poisoned group into a clean backbone. The dashed line is the registered 0.15 meaningful-effect threshold. MLP and layers 0–5 pass necessity in both sizes. No group passes sufficiency. Groups come from overlapping functional and depth partitions and should not be added. The selected early-MLP intersection passes necessity at 410M and does not pass at 1.4B.

F Secondary pooled precision estimates

The five 410M parents are verified G3 AdamW 10^{-1} seeds; the five 1.4B parents are separately fixed G3.1 seeds. They match G6 in model revision, graft, poison count and source offset, poison steps, optimizer rates, decay, sequence length, batch size, markers, and evaluation. The 410M parents used 10M clean adaptation tokens (2,441 steps), versus G6's 5M (1,220 steps). The 1.4B clean recipe is identical across cohorts.

The declared endpoint pool combines the 16 primary G6 seeds with five verified parents. It gives 21/21 and 20/21 whole-table necessity and 15/21 and 17/21 whole-table sufficiency at 410M and 1.4B. The 410M sufficiency Wilson lower bound is 0.500436, only 0.000436 above the registered majority bar; one seed flip reverses that call, whereas 21/21 necessity is robust to one flip. Final-row necessity is 12/21 and 17/21, and final-row sufficiency is 10/21 and 9/21. The temporal G4+G6 pool gives 19/21 history-specific necessity and 15/21 sufficiency; the latter has the same 0.500436 knife-edge. The 410M endpoint and temporal pools combine 10M-token parent clean adaptation with 5M-token G6 adaptation, so they are secondary consistency estimates. The 1.4B endpoint parents use the same clean recipe as G6. The invalid original G3 1.4B rows are excluded from every pool.

G SEED-LEVEL OPTIMIZER ROUTES

Table 3 reports every new G6 seed. "First dep." is the first tested table rate where whole-table necessity exceeds 0.15. "Need" and "suff." are the corresponding causal contrasts at 10^{-1}; "outside" is the complementary non-table sufficiency contrast. Seed numbers abbreviate the fixed IDs `26091601`–`26091616`. The dagger marks 1.4B seed 03: whole-table necessity is 0.149414, only 0.000586 below the frozen 0.15 rule, and is reported as absent exactly as preregistered.

Table 3: Per-seed optimizer transitions, necessity, and sufficiency.

Pythia-410M

Seed	First dep.	Whole need	Whole suff.	Outside suff.	Final need	Final suff.
01	10^{-3}	0.951	0.000	0.000	0.003	0.000
02	10^{-3}	0.944	0.000	0.002	0.002	0.000
03	10^{-2}	1.000	1.000	0.000	1.000	0.745
04	10^{-2}	1.000	1.000	0.000	1.000	0.844
05	10^{-2}	0.973	0.000	0.021	0.000	0.000
06	10^{-2}	1.000	0.000	0.000	0.000	0.000
07	10^{-2}	1.000	0.940	0.000	0.000	0.000
08	10^{-2}	0.990	0.000	0.010	0.000	0.000
09	10^{-2}	1.000	1.000	0.000	1.000	0.842
10	10^{-2}	1.000	1.000	0.000	1.000	0.093
11	10^{-1}	1.000	0.993	0.000	1.000	0.879
12	10^{-2}	1.000	0.999	0.000	1.000	0.652
13	10^{-2}	1.000	1.000	0.000	1.000	0.000
14	10^{-2}	0.998	0.289	0.000	0.000	0.000
15	10^{-2}	1.000	1.000	0.000	0.772	0.177
16	10^{-2}	0.994	0.991	0.000	0.994	0.648

Pythia-1.4B

Seed	First dep.	Whole need	Whole suff.	Outside suff.	Final need	Final suff.
01	10^{-1}	1.000	0.993	0.000	1.000	0.231
02	10^{-1}	1.000	0.584	0.000	0.999	0.011
03	none	$0.1494^{\dagger}$	0.000	0.851	0.000	0.000
04	10^{-1}	1.000	0.999	0.000	1.000	0.203
05	10^{-1}	0.995	0.201	0.004	0.000	0.000
06	10^{-2}	1.000	0.967	0.000	0.997	0.007
07	10^{-1}	1.000	0.891	0.000	1.000	0.570
08	10^{-1}	1.000	0.997	0.000	1.000	0.354
09	10^{-1}	1.000	0.922	0.000	1.000	0.026
10	10^{-2}	1.000	1.000	0.000	0.990	0.232
11	10^{-1}	1.000	0.001	0.000	1.000	0.000
12	10^{-1}	0.988	0.000	0.012	0.000	0.000
13	10^{-1}	1.000	0.943	0.000	1.000	0.121
14	10^{-1}	1.000	0.996	0.000	0.995	0.380
15	10^{-1}	1.000	0.572	0.000	0.967	0.010
16	10^{-1}	1.000	1.000	0.000	1.000	0.779

H SEED-LEVEL TEMPORAL ROUTES

Table 4 reports the five verified parent G4 seeds followed by all 16 new G6 seeds. "History" is the 72-row union across all trigger-internal positions; "spec." subtracts the largest matched-control drop. All entries are causal ASR contrasts. The G4 parents use 10M-token clean adaptation and the

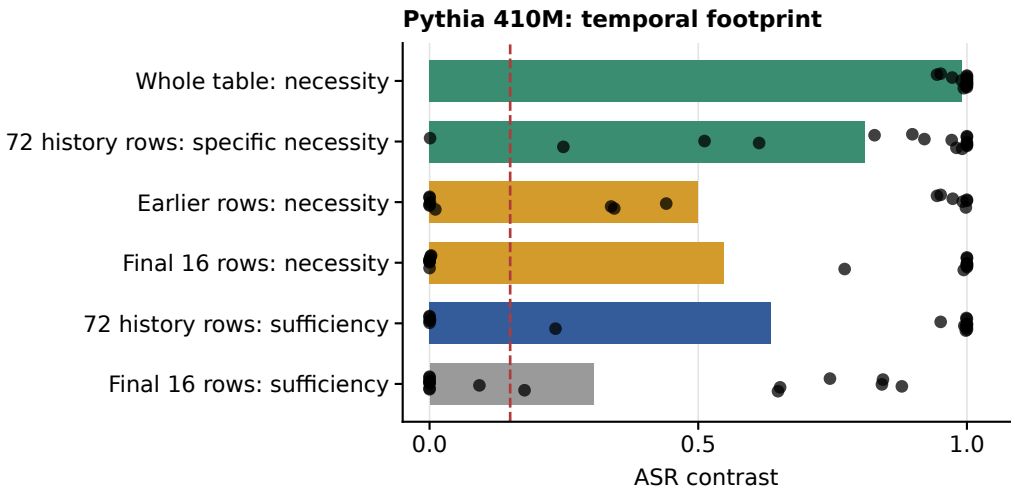

Figure 5: **The table route occupies a seed-variable temporal address footprint.** Dots are the 16 prospective G6 seeds. The 72-row history union is typically necessary under the specificity contrast but not typically sufficient; final and earlier subsets vary and final-row sufficiency is not typical.

G6 seeds 5M, so the manuscript's primary temporal estimate uses G6 alone; this table preserves the declared heterogeneous secondary pool.

Table 4: Every seed in the pooled temporal-footprint analysis.

Seed	Final need	Earlier need	History suff.	History spec.	Control drop
G4-1	1.000	0.001	0.969	0.998	0.002
G4-2	0.000	1.000	0.907	1.000	0.000
G4-3	1.000	0.000	0.760	1.000	0.000
G4-4	0.000	0.998	0.001	0.040	0.958
G4-5	0.394	0.337	0.604	0.299	0.095
G6-01	0.003	0.951	0.000	0.898	0.053
G6-02	0.002	0.944	0.000	0.828	0.116
G6-03	1.000	0.000	0.999	1.000	0.000
G6-04	1.000	0.000	1.000	1.000	0.000
G6-05	0.000	0.974	0.001	0.001	0.974
G6-06	0.000	1.000	0.000	0.921	0.079
G6-07	0.000	1.000	0.951	0.972	0.028
G6-08	0.000	0.992	0.000	0.512	0.480
G6-09	1.000	0.001	1.000	1.000	0.000
G6-10	1.000	0.440	1.000	0.613	0.387
G6-11	1.000	0.000	0.995	1.000	0.000
G6-12	1.000	0.000	0.998	1.000	0.000
G6-13	1.000	0.338	1.000	1.000	0.000
G6-14	0.000	0.998	0.234	0.249	0.749
G6-15	0.772	0.344	1.000	0.980	0.020
G6-16	0.994	0.011	0.998	0.991	0.003

I FREQUENCY-AWARE WRITE-POLICY SUCCESSOR

G11 tests six fixed policies crossing active-row restriction and inverse cumulative-hit weighting with nominal and stepwise norm-matched update budgets. Five paired seeds per size all pass installation. A route needs whole-table necessity and sufficiency above 0.15; with five seeds, the frozen Wilson-majority rule needs 5/5. G11 reuses S1 clean checkpoints whereas G6 freshly adapts 16 seeds, so their reference costs are from different cohorts.

Table 5: G11 route prevalence and arithmetic mean of per-seed clean-perplexity ratios. No policy meets the registered 5/5 typical-route criterion. Norm matching refers to the implementation described below.

	410M		1.4B	
Policy	Route	PPL ratio	Route	PPL ratio
Ordinary 5×10^{-5}	0/5	1.0495	0/5	1.0608
Reference 10^{-1}	4/5	1.2527	1/5	1.1045
Active, nominal	0/5	1.0402	0/5	1.0608
Frequency, nominal	0/5	1.0568	0/5	1.0612
Active, norm matched	4/5	1.1534	1/5	1.0947
Frequency, norm matched	2/5	1.1104	0/5	1.0840

Under registered norm matching, frequency-minus-active whole-table necessity/sufficiency means are $-0.193/-0.369$ at 410M and $-0.297/-0.124$ at 1.4B. Paired intervals cross the impairment endpoint, leaving those contrasts unresolved. These six policies give no low-cost route; other constrained optimizers remain untested.

The reference measures realized table displacement; active-row arms scale a pre-addition delta. BF16 rounding can change realized displacement, and their moment buffers also use the table dtype. Equal realized update budgets and precision-independent explanations are thus unestablished. This implementation detail is disclosed; the paired comparison estimates the policies as implemented.

J DELETION DURABILITY AND COLLISION ISOLATION

G12 starts from ten independent direct row writes, zeros their 16 values, and applies 512 optimization steps on a later mapping under six branches: disjoint or exactly colliding addresses, same or different payload where applicable, and retained versus cleared Adam state. The original behavior has zero ASR immediately after deletion and at all ten later measurements in all 60 branches. Every registered upper endpoint for peak resurrection is zero because all seed observations are zero; this is not a zero-risk population bound. Intermediate checks use 256 contexts and endpoints use 1,024. The collision changes a raw token while preserving the compressed suffix and all 16 hash addresses; it is one constructed collision, not a sampled adversarial distribution.

Secondary writes reach 0.969–0.998 endpoint accuracy at 410M. At 1.4B they reach 0.913–0.996 except for the exact-collision, same-payload branch with retained moments, which reaches 0.341 versus 0.996 after clearing state. Retained moments also move deleted disjoint rows by mean L2 0.0458 and 0.0987 at 410M and 1.4B while cleared moments leave them at zero. The old mapping does not resurrect; clearing moments avoids the observed interference in the tested 1.4B same-payload collision branch. Identical rows did not imply behavioral aliasing in this construction.

K REPRODUCIBILITY RECORD

Retained records include predictions, logs, frozen configurations, model revisions, inventories, and verifier reports. Replays of S3, S4, G5, G6, G11, and G12 reproduced registered decisions; G6 matched every metric and prediction in 320 files, and the clean-graft audit matched twelve paired rows. The anonymous review artifact releases verifiers, configurations, seed-level aggregates, and prompt-construction code. Per-prompt records remain in the internal archive; new checkpoint interventions require separately stored model states and the exact memory bank.