WHAT REMAINS AFTER FACTUAL FORGETTING? HISTORY-CONDITIONED ACCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that factual learning leaves a residual parameter state that shapes subsequent learning even after the learned answers are no longer expressed. Across twelve Pythia-410M training runs, we randomize which facts receive training before interference, then test updates that train only on other facts. Prior exposure increases the update-induced change in exact-answer accuracy by 5.03 percentage points over the full fact panel. Among current failures, previously acquired facts recover more often than never-exposed facts, including within-fact and margin-deficit-adjusted comparisons; history-dependent recovery and faster relearning recur in Pythia-1.4B and Qwen2.5-0.5B. A correct answer may start closer to overtaking its competitors, or an update may push it further toward that threshold. Starting gaps explain much of the recovery ordering under random parameter displacement, but factual updates also depend on learned associations: label correctness changes the recovery gap more for previously trained support facts than for fresh ones. We measure this sensitivity before training by comparing the gradient directions that raise the answer score and lower support loss. Exposure changes this pre-update QKV tangent geometry, improving prediction of the cue-by-label margin effect after ten real training steps at two Pythia scales. To test the influence of historical parameters directly, we transplant blocks from donors with different query–answer assignments into a common recipient. Across six recipient–donor triplets, answer preferences follow the donor's assignments, including ones it no longer expresses, and respond differently to later training. A separate predictor links geometry to transplantation: fitted only on exposure histories, it partly predicts how replacing blocks changes an update's effect on answer scores in three new triplets, reducing error by 9.69% relative to assuming that replacement leaves this response unchanged. Residual factual memory thus concerns both present accessibility and how past learning conditions future updates.

1 INTRODUCTION

Training a language model on new information can reduce its accuracy on previously learned facts. A current error leaves open how that fact will respond when the model learns again. Two facts the model now answers incorrectly may differ in how readily their correct answers return, which updates restore them, and how much direct training they need.

Recovery after later learning is established in distributed networks and language models (Harvey & Stone, 1996; Hu et al., 2025); gradient alignment helps explain benign relearning (Yoon et al., 2026), while internal accessibility predicts model-level recovery across unlearned checkpoints (Song et al., 2026). The question here is what training history contributes to this susceptibility: does exposure change how the *same fact* responds to a specified update, and can that dependence be measured and manipulated in the parameters before the next update is applied? We use *history-conditioned access* to describe how past learning shapes the return of an answer under subsequent training. To isolate the contribution of history, we give the same fact different exposure histories across twelve Pythia-410M training runs. Each run trains on a randomly assigned half of 192 target facts before shared interference. We measure whether a failed answer returns when we train only on other facts (*recoverability*), and how quickly it returns when we retrain the target itself (*relearning savings*). Never-exposed instances of the same fact serve as controls. Within each run, every test starts from its saved post-interference checkpoint, R_0.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

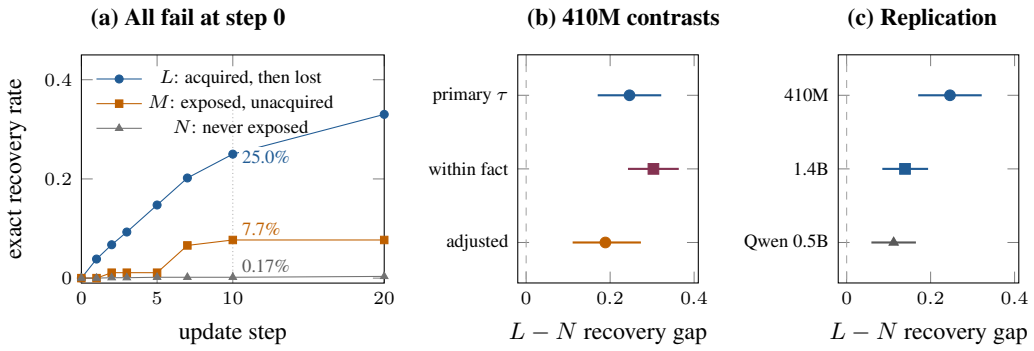

Figure 1: History-dependent recovery and model replication. (a) Reference-update trajectories for acquired-then-lost (L), exposed-but-unacquired (M), and never-exposed (N) failures. (b) Primary, within-fact, and margin-deficit-adjusted 410M contrasts. (c) Primary contrasts in Pythia-410M, Pythia-1.4B, and Qwen2.5-0.5B (12, 8, and 8 chains). Bars are 95% intervals: equal-chain t intervals for primary contrasts, relation-cluster bootstrap intervals for within-fact and adjusted contrasts.

Our contribution is to connect a fact's experimentally varied learning history to its sensitivity to later updates, measured before training and tested by parameter intervention. Three findings establish this connection. First, randomized exposure changes the response to a target-disjoint update, while recovery and relearning speed distinguish currently incorrect facts with different acquisition histories. The acquired-then-lost versus never-exposed contrast persists within fact and after margin-deficit adjustment, and recovery and savings recur at larger scale and in Qwen (Section 3). Randomization identifies the effect of exposure; acquisition and loss describe the resulting histories.

Second, recovery reflects both the starting margin and which update is applied. Smaller starting deficits explain much of the ordering under random parameter displacement, whereas factual updates show a dependence on the previously learned cue–answer associations (Section 4). This dependence is measurable before training: exposure changes how the support-loss descent directions align with raising the target-answer score. The measured change improves prediction of the ten-step cue-by-label margin effect on held-out histories (Section 4.2).

Third, historical parameters causally influence both present preferences and later learning. Replacing blocks in a common recipient with blocks from donors trained on different query–answer assignments changes both answer preferences and how later training modifies them. A separate predictor trained only on the exposure experiments partly predicts how this replacement changes the recipient's response to a later update (Section 5).

Recovery and parameter-state prediction also extend to facts known at the pretrained checkpoint, while a natural-task stream tests a second interference setting (Section 6). Pre-update interactions also inform an evaluation decision: allocating a limited regression-audit budget to facts at risk under a proposed short update (Section 6.2).

2 Experimental Design

2.1 Constructing Learning Histories

To compare the same fact with and without prior learning, we run independent fine-tuning sequences, called chains, from pretrained Pythia-410M (Biderman et al., 2023). Each chain first trains on its assigned facts, then undergoes interference that makes some acquired answers fail. Acquisition trains on 96 assigned targets and 32 cue facts used later as update support; structured and diffuse factual interference follow. All three phases use full fine-tuning; Appendix A.2 gives their sizes and training schedules. Facts are WikiBigEdit-derived temporal associations (Thede et al., 2025) that the base model answers incorrectly. We remove overlapping facts using fixed exclusion rules. Before the main experiment, a learnability screen retains the top quartile to provide enough acquired facts for the forgetting test within the fixed training budget. Every intervention exactly reloads the post-interference checkpoint R_0.

Learning a related fact might also affect a target. To balance this influence, the 192 targets form 96 fixed within-relation peer pairs, matched on pre-treatment features. In each chain, E records the target's exposure and P its peer's exposure. Every fact appears three times in each (E, P) cell across chains. Thus we can compare exposed and withheld instances with the peer either exposed or withheld, weighting both conditions equally (*peer adjustment*). Randomization inference uses 100,000 sampled exposure schedules satisfying these assignment constraints.

After interference, 312 of 1,047 acquired target instances (29.8%) fail at R_0. Acquisition means alias-aware greedy exact correctness at the end of the acquisition phase. Among instances incorrect at R_0, realized history defines $\mathcal{L}$ as exposed and acquired, $\mathcal{M}$ as exposed but not acquired, and $\mathcal{N}$ as never exposed. Pooled sizes are 312, 91, and 1,145 instances. Acquisition and subsequent failure are outcomes of randomized exposure. Randomization identifies the panel-level exposure effect, not realized acquisition status. The strata describe current failures by history; within-fact and matched-deficit comparisons constrain explanations for their ordering.

2.2 Interventions

We first test whether a failed answer can return without training on that target. These recovery interventions (Q) use support examples that exclude the target's fact, subject, answer, aliases, and prompt roles under our closure rules. The reference update trains on the 32 cue facts learned during acquisition, with their correct answers. All trained updates use the same small trainable parameterization, so that the support examples vary while the update architecture stays fixed. Each trained intervention fits a fresh rank-1 LoRA (Hu et al., 2022) on all query–key–value projections. Step 10 is the primary endpoint.

All conditions start from the same checkpoint within a chain; only the incoming update changes. To test whether past cue–answer learning matters, we cross previously trained (old) versus untrained (fresh) cues with correct versus wrong answers taken from other facts. An evaluation-form condition changes only how the same facts are phrased. To test whether factual content is needed, we replace it with length-matched nonce QA or random tokens. Finally, untrained random directions test generic parameter movement, matched either per projection in Frobenius norm or in the KL divergence between pre- and post-update output distributions on held-out prompts (calibration KL). Appendix A.2 specifies the conditions and matching checks.

The 64 cue facts form 32 paired slots. Within each slot and chain, one cue is old and the other fresh; each cue identity is assigned to each history in six of the twelve chains, independently of target exposure.

As a complementary test, direct retraining measures how quickly a failed answer can be relearned. This arm, D, trains all 192 targets for 10 steps and records, for each fact below threshold at R_0, the first observed step in $\{0, 1, 2, 3, 5, 7, 10\}$ at which its decision margin becomes non-negative, censored at 11. A capacity arm trains 32 never-exposed targets directly, testing whether the probe can acquire facts when their labels are supplied. Stochastic-arm replicates are averaged within chain.

2.3 Measures, inference, and confirmation

Let $Y_t^a = 1$ when greedy decoding after step t of update a is alias-aware exactly correct, and let $Y_t^a = 0$ otherwise; Y^0 denotes the corresponding indicator at R_0. Our primary descriptive comparison asks how much more often an acquired-then-lost fact recovers than a never-exposed fact. For update a in chain s, we average this step-10 recovery gap over both peer-exposure conditions:

$$\tau_{a,s} = \tfrac{1}{2} \textstyle\sum_{p \in \{0,1\}} \left[\Pr(Y_{10}^a{=}1 \mid \mathcal{L}, P{=}p) - \Pr(Y_{10}^a{=}1 \mid \mathcal{N}, P{=}p) \right], \tag{1}$$

aggregated with equal weight over chains (95% t intervals). Within-fact comparisons hold fact identity fixed across histories. A failed answer may also start closer to overtaking its competitors. We account for this using the starting margin deficit, $d = \max(0, -M_0)$. Here M_0 is the weakest-token gold-versus-competitor logit margin over the answer and EOS at R_0. Appendix A.2 gives the regression and matching specifications, savings adjustment, and inference procedures.

The primary design and P1–P14 were fixed before outcomes (Appendix A.1); additional panels have separate protocols. Predictors are fixed before confirmation. These are internal freezes, not external preregistration; descriptive post-analyses are labeled.

2.4 MEASURING UPDATE SENSITIVITY BEFORE TRAINING

To measure how an update would locally affect a target before applying it, we compare the gradient direction that raises the target-answer score with the direction that lowers support loss, in separate Pythia-410M and Pythia-1.4B experiments. Here the target-answer score is its mean token log probability, including EOS. Their cosine is positive when support-loss descent locally raises the target score and negative when it lowers it. We measure this pre-update QKV tangent geometry through virtual rank-one LoRA adapters on all QKV projections. We differentiate their output factors at zero, where the adapters leave the checkpoint's predictions unchanged.

We compare how replacing wrong labels with correct ones changes this alignment for old versus fresh cues. Let O and F denote these correct-minus-wrong cosine differences for old and fresh cues. Then $A = O - F$ measures their interaction, and $\Delta A = A_{R_0} - A_{\text{base}}$ its change through acquisition and interference. To test whether this local measurement predicts actual training, we run the four cue-by-label updates for ten steps, with two initializations each. The endpoint Y_{margin} takes the same old-versus-fresh difference between correct- and wrong-label updates in the resulting answer-plus-EOS margin changes. This comparison tests whether the label effect differs between old and fresh cues after real training, just as the cosine comparison asks before training. Scale-specific predictors are fitted on discovery histories, frozen, and evaluated on fresh randomized histories. Appendix C gives cohort sizes, formulas, and fitting details.

3 RECOVERY IS GRADED BY ACQUISITION HISTORY

3.1 TARGET-DISJOINT RECOVERY

Across the full randomized panel, prior exposure increases Q-induced exact change, $Y_{10}^{a} - Y^{0}$, by 0.0503 $[0.0258, 0.0749]$, without conditioning on realized acquisition (Appendix A.1).

Realized acquisition histories describe recovery among current failures. The reference update restores 78 of 312 acquired-then-lost failures (25.0%) and 2 of 1,145 never-exposed failures (0.17%), giving $\tau = 0.246$ $[0.170, 0.322]$, positive in 12 of 12 chains (Figure 1). Exposed-but-unacquired failures recover at an intermediate 7.7% (7/91), completing the primary ordering $\mathcal{N} < \mathcal{M} < \mathcal{L}$ with minimum adjacent rate gap 0.075 $[0.028, 0.121]$ (relation-cluster bootstrap). The contrast also appears within fact. Among the 123 facts that appear as current failures under both histories, the same fact recovers more often in chains that acquired it, $\delta = 0.303$ $[0.243, 0.364]$. This comparison controls fact identity, but not chain-level conditions associated with realized acquisition.

Acquired-then-lost failures sit closer to the decision threshold than never-exposed failures, so we condition on margin deficit directly. The margin-deficit-adjusted stratum difference, averaged over the $\mathcal{L}$ covariate distribution, is 0.189 $[0.111, 0.273]$. Caliper matching retains a smaller positive contrast within the limited overlap (Appendix A.2). Starting proximity accounts for part of the difference, while history remains informative after this adjustment.

Restoration is specific as well as graded. The never-exposed floor stays at or below 0.52% across every probe configuration tested (Appendix D.1), and the reference update leaves 96.3% of currently correct exposed facts correct (721/749).

3.2 RELEARNING SAVINGS AND EARLY-RESPONSE PREDICTION

Direct retraining gives a complementary measure, relearning speed. On the fixed 0–10 observation grid (censored at 11), exposed current failures have a 1.46-step advantage in first-observed threshold time over never-exposed failures, $[1.09, 1.82]$, positive in 12 of 12 chains. Adjustment for margin deficit leaves 0.850 $[0.607, 1.093]$.

A descriptive post-analysis resolves the exposed group by realized history. Within the horizon, crossing rates are 35.3% for acquired-then-lost, 11.0% for exposed-but-unacquired, and 0.87% for never-exposed instances; the peer-adjusted savings are 1.78 for $\mathcal{L}$ and 0.445 $[0.125, 0.766]$ for $\mathcal{M}$. Target-disjoint recovery and relearning speed thus share the ordering $\mathcal{N} < \mathcal{M} < \mathcal{L}$.

We test whether the first three update steps forecast step-10 recovery. A ridge-logistic predictor uses the reference arm's third-step margin change $C3 = M_3 - M_0$ and starting deficit d. Frozen on

the separate earlier 12-seed factorial panel, it achieves step-10 balanced accuracy 0.872 on exposed current failures and improves log loss over the margin-deficit-only model in all 12 chains. The predictive gain and history contrast persist across probe configurations (Appendix D.1).

3.3 Replication across model configurations

Eight fresh chains at each of Pythia-1.4B and Qwen2.5-0.5B (Yang et al., 2024a) repeat the behavioral design with the same estimands and inference procedures, separately from the tangent experiments. Both reproduce the reference history contrast, within-fact contrast, and relearning savings (Figure 1c; Table 4). The 410M-frozen predictor lowers log loss in all eight 1.4B chains. The table also reports within-model leave-one-chain-out fits for both models; adding third-step response improves that prediction at Qwen. The peer-adjusted full-panel exposure effect on $Y_{10} - Y_0$ is positive at 1.4B; Qwen's estimate is also positive, with a 95% interval that includes zero. Appendices F.1–F.2 give configurations, numerical estimates, update-arm comparisons, and capacity results.

4 History-dependent update sensitivity

4.1 Threshold proximity and mapping-dependent response

An answer may recover because it starts close to threshold, or because an update moves it strongly toward that threshold, or both. We replay different updates from the same checkpoints to distinguish starting proximity from update sensitivity (Figure 2).

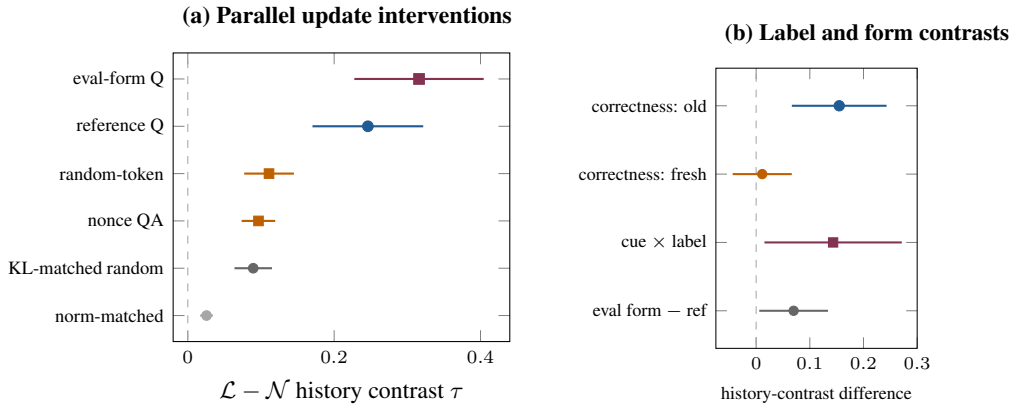

Figure 2: Update selectivity at 410M. (a) History contrast τ under each intervention, with equal-chain 95% intervals. (b) Contrasts that separate evaluation form, cue history, and label correctness: evaluation form minus reference, the cue-by-label interaction, and the correctness increments for fresh and old cues. All plotted intervals are 95%; the evaluation-form contrast in the text uses a 90% interval.

Small starting deficits let random movement restore answers. Random-token training reaches 45% of the reference effect while leaving the replicate-averaged never-exposed floor at 0.044%. Static random displacements also recover previously acquired facts. A descriptive reconstruction explains much of this history ordering: apply the empirical KL-matched displacement distribution from other strata to each stratum's starting margin deficits and count threshold crossings. Without fitting parameters, this reconstruction reproduces the ordering at both Pythia scales (Appendix D.2). Thus the smaller deficits left by prior acquisition can make even content-free movement more likely to restore an answer.

Recovery also depends on the cue–answer pairs used for training. Training on factual support produces a larger recovery gap than generic movement: the reference exceeds KL-matched random displacement by 0.157 $[0.077, 0.236]$ (Holm $p = .003$) and nonce QA by 0.150 (Holm $p = .005$). The random-displacement comparison matches calibration KL. The nonce-QA comparison includes

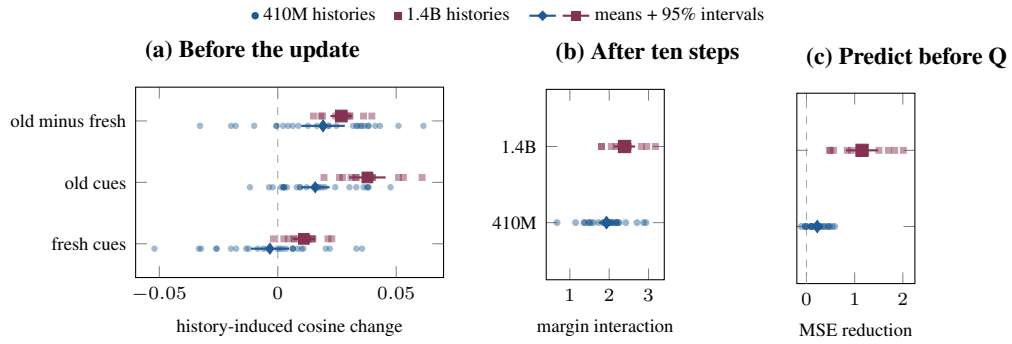

Figure 3: Prior exposure leaves a QKV tangent state that predicts finite update response in 27 fresh 410M and 12 untouched 1.4B histories. (a) Exposure contrasts in base-to-R_0 changes of the correct-minus-wrong gradient cosine, for fresh cues, old cues, and their difference. (b) Exposure contrast in the four-arm margin interaction after real ten-step updates. (c) MSE gain from adding ΔA to each scale's discovery-fitted predictor. Cross-scale magnitudes are descriptive.

differences in the actual strength of the trained updates, which is not dose-matched. Nonce QA and random-token training differ by -0.014 (Holm $p = 1.0$).

Label correctness matters more when the support cues were previously trained. Replacing wrong labels with correct ones enlarges the acquired-then-lost versus never-exposed recovery gap by 0.155 $[0.067, 0.243]$ on old cues, but by only 0.011 $[-0.044, 0.066]$ on fresh cues. The difference between these increments is the cue-by-label interaction: 0.143 (95% CI $[0.016, 0.271]$). A separate 12-chain factorial confirmation gives full-panel exposure interactions of 13.19 and 9.81 percentage points on its two target panels, each positive in all chains (Appendix B). Holding factual content fixed, evaluation-form support increases the history contrast by 0.070 $[0.018, 0.122]$ (90% interval).

These comparisons distinguish two contributions at the intervention level: starting deficits make generic movement effective, and learned cue–answer associations shape the response to factual updates. The cue-by-label interaction motivates measuring the corresponding parameter structure *before* the update begins.

4.2 Pre-update parameter geometry predicts finite response

Does past learning change how an update would move the target-answer score even before that update is applied? Unlike the early-response predictor, this measurement requires no observed training step. At the unchanged checkpoint, we measure the cosine between the gradient that raises the answer score and the gradient direction that lowers each support loss, as defined in Section 2.4. We then compare correct-minus-wrong alignment for old cues with the same difference for fresh cues. This second comparison measures dependence on cue history and label correctness; randomized exposure changes it (Figure 3). In the 410M confirmation panel, averaging the exposed-minus-withheld difference equally over the two peer-exposure cells gives $D_s(\Delta A) = 0.0191$ $[0.0099, 0.0283]$ for the base-to-R_0 change. Because exposure is assigned after the base state, this difference identifies an exposure-induced change in the post-interference tangent geometry.

The change is concentrated on previously trained cues and, at 410M, in sensitivity to conflicting updates: old-cue wrong-label alignment becomes more negative, while the old-cue correct-label change is near zero. Appendix C gives the old- and fresh-cue components and their armwise decomposition.

We next apply the four real updates to test whether the pre-update measurement predicts how the correct-versus-wrong difference in target margins varies between old and fresh cues after ten training steps. On the same histories, the exposure effect on this margin interaction is $D_s(Y_{\text{margin}}) = 1.932$ $[1.730, 2.133]$, positive in all 27 histories. Adding ΔA to the predictor fitted and fixed on discovery histories improves mean-squared error over its fixed baseline by 0.223 $[0.152, 0.295]$, positive in 23 of 27 histories. The baseline already includes assignment, base interaction, R_0 margin, and gradient norms.

The separate Pythia-1.4B experiment fits a scale-specific predictor on four discovery histories and evaluates 12 untouched histories. The exposure-induced geometry change, finite interaction, and predictive gain recur, with all six prespecified quantities positive in every history (Figure 3; Appendix C).

5 Historical bindings shape preferences and subsequent learning

Can replacing parameter blocks transfer the influence of past query–answer bindings to a common recipient R, changing both its answer preferences and how later training modifies them? We train two donor models on the same answers but assign those answers to different queries. We then replace blocks 0–7 of a common recipient with the corresponding blocks from each donor, keeping all other recipient parameters fixed. This *transplant* creates two hybrids that can be compared under the same queries and later updates.

The assignment must distinguish remembering a query–answer pairing from simply favoring an answer everywhere. We therefore cycle the same answers over three queries: for example, one donor learns B/C/A and the other C/A/B for the three queries in order. Each answer occurs once in each assignment, so a query-independent additive shift in answer log scores cancels when we sum the paired comparisons around the cycle. Each nine-query panel has three such relation blocks. Six independently trained Pythia-410M triplets—a recipient and two donors—cover two panels, with three triplets per panel. Within each triplet, training histories differ only in the query–answer bindings (Figure 4).

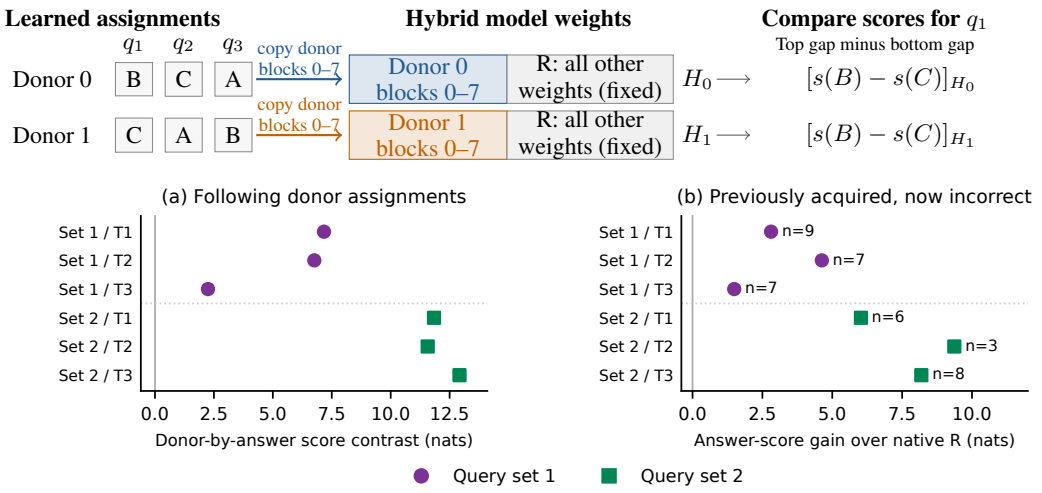

Figure 4: Testing whether answer preferences follow transplanted history. Top: different donor blocks 0–7, identical remaining R weights. Compare the B-versus-C score gaps for q_1 and repeat around the cycle. Below: (a) mean donor-by-answer contrast; (b) mean gain over native R for answers acquired then lost by the donor (n bindings). T1–T3 index independent recipient–donor triplets within each query set. Scores sum content-token log probabilities, excluding EOS.

On average, transplanting donor blocks into the recipient shifts its preference toward the query–answer assignments learned by that donor. We compare the two donors' assigned answers in both hybrids; the mean donor-by-answer contrast is positive in all six triplets and every relation block (Figure 4a). In the first panel it already appears at the first answer token. The positive block contrasts therefore identify query-conditioned binding influence beyond such a shared answer shift (Appendix H.1).

This influence survives loss of direct expression. For answers the donor learned but no longer produces correctly, transplanting its blocks into the recipient still raises their mean content log probability over the native recipient in each triplet: Figure 4b shows these gains. These shifts show that native parameters retain graded influence of historical bindings after direct expression has failed. Appendices H.1–H.2 report the numerical contrasts, exact outputs, and collateral changes.

The replaced blocks also change the response to later training. We compare how identical later updates change the two hybrids' preferences. Four target-disjoint cue-by-label arms train for ten steps from each hybrid, with paired initializations. Subtracting the initial preference contrast measures how much the update strengthens or weakens preference for the donor's assignments. Directions vary across updates and triplets. In the second panel, both wrong-label arms attenuate donor following in every triplet, while correct-label effects vary (Appendix H.2).

Predicting how transplantation changes later learning. The geometry experiment predicts a four-arm margin interaction. Here the question is whether a relation learned under exposure also predicts the effect of replacing parameters. We fit a separate predictor using only exposure histories. It predicts an arm's ten-step change in mean answer-plus-EOS log probability from pre-update QKV alignment, scores, and gradient norms. We freeze the predictor, apply it to each hybrid and its recipient under the same update arm, then subtract the recipient prediction from the hybrid prediction. On three fresh transplant triplets with a reused nine-query panel, MSE for this hybrid-minus-recipient response difference falls by 9.69% relative to zero effect, with gains in every triplet. Adding alignment reduces pooled error by 3.36% over scores and norms alone (Appendix H.3). Thus a relationship fitted without transplant outcomes captures part of the change in later learning caused by replacement, with an additional contribution from alignment.

6 Extensions and a potential application

6.1 Pretrained knowledge and natural-task interference

The pretrained checkpoint lets us test recovery beyond facts installed in our experiment. We evaluate facts it already answers correctly, replaying saved checkpoints and real ten-step updates from the 410M tangent panel. Factual support restores 7.56% $[6.08, 9.04]$ of facts with valid but incorrect post-interference answers, with restoration in all 20 histories. Here correctness is based on the answer content and its boundary, allowing subsequent text; the score averages answer tokens without EOS. For an answer-specific measure of update response, we compare each answer's score and gradient alignment with those of a fixed alternative. The tangent interaction improves prediction beyond base geometry, current scores, and gradient norms on both the full confirmation panel and its pre-update failure sets. A predictor fitted on separate facts and histories is frozen for both evaluations (Appendix G.1).

We next ask whether history-dependent recovery also extends to forgetting caused by a natural-task stream. Across 12 fresh 410M histories, we replace factual interference with NumGLUE-ds (numerical reasoning) followed by FOMC (monetary-policy stance classification). The randomized exposure contrast in update-induced exact change is 0.569 $[0.509, 0.630]$, positive in every history. Recovery contrasts and direct relearning also favor exposed facts (Appendix G.2). All twelve histories enter the estimates, although acquired-fact loss ranges from 61.9% to 88.6% and 11 exceed the 65% cap, so the original confirmation criterion is not met. Many answers become FOMC class labels, so recovery here partly restores the factual response interface. History dependence thus persists in this more severe, task-shifted regime.

6.2 Choosing which facts to check after an update

An initial use of update-specific sensitivity is to decide which facts to check when the evaluation budget is limited. For each three-step update, we rank 48 protected facts. A score-damage event is a fact–history–recipe instance whose mean answer-token log probability, including EOS, drops by more than 0.1. The raw QKV target–support gradient inner product indicates whether the update locally raises or lowers an answer's score. Adding it to current-state, recipe, and gradient-norm features raises confirmation equal-history Recall@10 from 77.24% to 80.66%, with 83 of 103 score-damage events detected versus 77. Gains vary across four fresh histories, including one decrease. Two disjoint four-fact correction subsets define alternative updates; substituting the other subset's readout while keeping the evaluated update fixed reduces detections to 71. Gains concentrate at tight budgets, although a one-step probe is stronger and cheaper here (Appendix I). The six additional detections show an initial use of the measured sensitivity: it helps direct checks toward facts at risk from a particular update. Developing this signal into a cost-effective auditing method remains for future work.

7 Related work

Recovery and the role of learning history. Unrelated learning (Harvey & Stone, 1996) and partial retraining (Stone & Jupp, 2007) can revive unrehearsed associations. Recovery depends on initial learning and perturbation (Atkins & Murre, 1998), and on the forgetting mechanism (Stone & Jupp, 2008). Our within-fact design adapts counterfactual memorization's include–exclude comparison (Zhang et al., 2023) to shared interference and replayed updates. Cyclic document training permits recovery before a document is revisited (Yang et al., 2024b). Spurious forgetting separates disrupted task alignment from knowledge loss (Zheng et al., 2025), relevant to our evaluation-form and natural-task results. We additionally randomize exposure to test how prior learning changes a fact's response to the same update. Other studies examine factual retention (O'Neill, 2026; Chen et al., 2026), sample-wise backward transfer (Harmon et al., 2026), reversible concept removal (Lo et al., 2024; Xu et al., 2026), and retained old-task information after visual-task forgetting (Trivedi & Melwani, 2026a; Filus et al., 2026; Trivedi & Melwani, 2026b).

Susceptibility to further updates. Unlearned knowledge can return after training on related data (Hu et al., 2025), accessible facts (Deeb & Roger, 2024), or retained data (Siddiqui et al., 2025). The last study distinguishes approximately unlearned vision models from retrained references and predicts susceptibility from weight-space geometry. Related work examines syntax and gradient alignment (Yoon et al., 2026), suppressive units (Yang et al., 2026), and recovery under quantization (Zhang et al., 2025). A white-box audit forecasts model-level recovery across unlearned checkpoints (Song et al., 2026); our frozen recovery predictor targets individual facts. Separately, randomized exposure tests how history changes the cue-by-label response, and pre-update tangent geometry predicts its finite margin interaction.

Intervening on factual answers. Model editing changes factual answers (Meng et al., 2022); unlearning targets their removal (Liu et al., 2025). We vary donors' historical assignments of the same answer pool, then test their influence on a common recipient's preferences and subsequent learning through parameter transplantation.

8 Discussion and conclusion

Recovery depends on the relation between a model state and an update. Small starting deficits let generic movement restore some answers; learned cue–answer associations shape the response to factual updates. Residual memory therefore concerns how later information reinforces or conflicts with past learning, as well as proximity to the correct answer. Geometry makes this sensitivity measurable before training; transplantation tests historical parameters in a common recipient. The exposure-fitted predictor partly connects these two experiments by anticipating how replacement changes later learning.

A model's current answer captures only one consequence of its learning history. Past learning also changes how that answer responds to subsequent training. Residual factual memory is therefore characterized by both present expression and update-conditioned response.

9 Limitations

Evidence covers pre-screened, experimentally installed temporal facts in three model configurations. The pretrained-known extension starts from one pretrained 410M checkpoint and uses a fixed source panel; correct answers may reflect inference rather than memorization. Tangent measurements use the tested all-QKV coordinates; native-block transplantation covers small query panels. Exposure-to-transplant prediction reuses Pythia-410M queries and the four-arm update family; query-dependent compatibility and mediation of exact recovery remain unresolved. Appendix A.3 details dose, model-comparison, and measurement boundaries. Regression confirmation reuses development facts and tests short-update score degradation; larger behavioral suites and downstream training benefits remain untested.

REPRODUCIBILITY STATEMENT

Appendix A.2 specifies the materials, interventions, evaluation rules, and statistical analyses. Appendix A.1 records the prespecified primary analyses; the experiment-specific appendices document confirmation splits, predictor fitting, transplantation procedures, and supplementary results. The prespecification records, executable experiment configurations and analysis scripts, per-fact outputs, figure inputs, and a standalone implementation of the threshold reconstruction will be made available to reviewers and area chairs on request through an anonymous repository during the discussion period. These materials will be released publicly upon acceptance.

AI USE STATEMENT

Generative AI tools assisted with implementing and debugging experimental and analysis code, literature search, and manuscript polishing. They were also used to obtain feedback on experimental methodology and to brainstorm possible interpretations of the results. The authors directed the research, evaluated and corrected AI-generated suggestions, and made the final decisions about the experimental design, analyses, and claims. The authors take full responsibility for the final manuscript, code, analyses, and reported results.

REFERENCES

P. W. B. Atkins and Jaap M. J. Murre. Recovery of unrehearsed items in connectionist models. *Connection Science*, 10(2):99–119, 1998.

Stella Biderman, Hailey Schoelkopf, Quentin 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 *International Conference on Machine Learning (ICML)*, pp. 2397–2430, 2023.

Howard Chen, Jiayi Geng, Adithya Bhaskar, Dan Friedman, and Danqi Chen. Continual memorization of factoids in language models. *Transactions on Machine Learning Research*, 2026.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024.

Katarzyna Filus, Kamil Faber, Roberto Corizzo, and Christopher Kanan. Lost or hidden? A concept-level forgetting in supervised continual learning. *arXiv preprint arXiv:2605.16374*, 2026.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. In *International Conference on Learning Representations (ICLR)*, 2026.

Inman Harvey and James V. Stone. Unicycling helps your French: spontaneous recovery of associations by learning unrelated tasks. *Neural Computation*, 8(4):697–704, 1996.

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

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

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7: 181–194, 2025. doi: 10.1038/s42256-025-00985-0.

Michelle Lo, Fazl Barez, and Shay B. Cohen. Large language models relearn removed concepts. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 8306–8323, 2024.

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

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of ACL*, pp. 3505–3523, 2022. doi: 10.18653/v1/2022.acl-long.246. URL https://aclanthology.org/2022.acl-long.246/.

Charles O'Neill. Can a language model learn facts continually in its weights? *arXiv preprint arXiv:2607.11020*, 2026.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of EMNLP-IJCNLP*, pp. 2463–2473, 2019. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Agam Shah, Suvan Paturi, and Sudheer Chava. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of ACL*, pp. 6664–6679, 2023. doi: 10.18653/v1/2023.acl-long.368. URL https://aclanthology.org/2023.acl-long.368/.

Shoaib Ahmed Siddiqui, Adrian Weller, David Krueger, Gintare Karolina Dziugaite, Michael Curtis Mozer, and Eleni Triantafillou. From dormant to deleted: Tamper-resistant unlearning through weight-space regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zirui Song, Huaxing Liu, Xiang Wang, Shuai Li, Xinye Li, Lang Gao, Jinghui Zhang, Zheng Lu, Fengxian Ji, Xiaojun Chang, and Xiuying Chen. Measure, don't optimize: Forecasting recovery in LLM unlearning. *arXiv preprint arXiv:2608.11408*, 2026.

James V. Stone and Peter E. Jupp. Free-lunch learning: modeling spontaneous recovery of memory. *Neural Computation*, 19(1):194–217, 2007.

James V. Stone and Peter E. Jupp. Falling towards forgetfulness: synaptic decay prevents spontaneous recovery of memory. *PLoS Computational Biology*, 4(8):e1000143, 2008.

Lukas Thede, Karsten Roth, Matthias Bethge, Zeynep Akata, and Thomas Hartvigsen. WikiBigEdit: Understanding the limits of lifelong knowledge editing in LLMs. In *International Conference on Machine Learning (ICML)*, pp. 59326–59354, 2025.

Ayushman Trivedi and Bhavika Melwani. Catastrophic forgetting as accessibility collapse: a three-level framework for knowledge persistence in continual learning. *arXiv preprint arXiv:2606.06032*, 2026a.

Ayushman Trivedi and Bhavika Melwani. The stable recovery manifold: Geometric principles governing recoverability in continual learning. *arXiv preprint arXiv:2606.13637*, 2026b.

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

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Nakyeong Yang, Dong-Kyum Kim, Jea Kwon, Minsung Kim, Kyomin Jung, and Meeyoung Cha. Erase or hide? Suppressing spurious unlearning neurons for robust unlearning. In *International Conference on Learning Representations (ICLR)*, 2026.

Yanlai Yang, Matt Jones, Michael C. Mozer, and Mengye Ren. Reawakening knowledge: anticipatory recovery from catastrophic interference via structured training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Sangyeon Yoon, Hyesoo Hong, Wonje Jeung, and Albert No. Rethinking benign relearning: syntax as the hidden driver of unlearning failures. In *International Conference on Learning Representations (ICLR)*, 2026.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. In *International Conference on Learning Representations (ICLR)*, 2025.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations (ICLR)*, 2025.

A PROTOCOL, PRESPECIFIED ANALYSES, AND SCOPE

A.1 PRESPECIFIED ANALYSES AND RESULTS (410M)

The primary experiment separates whether learning history affects later answers from how that effect depends on the incoming update. The table groups the corresponding tests: exposure and history comparisons, relearning and early-response prediction, update controls, and probe robustness. Each row retains the quantity and decision rule fixed before the primary outcomes.

P1 evaluates a policy that keeps currently correct answers and uses post-update answers only for current failures. Its exposure effect includes both retained and restored correctness; the separate exact-change form measures change under the update. P9 and P10 test the prespecified prediction that each nonfactual control produces at most one quarter of the reference recovery gap.

Table 1: Results for the fourteen prespecified predictions of the 12-chain 410M experiment. Holm corrections are applied within the prespecified P9–P13 family; quarter-reference contrasts are used only for P9 and P10.

ID	Quantity	Observed result
P1	Failure-gated policy effect	.7161 [.6813, .7510]; 12/12 positive; 100,000-schedule $p = 1.0 \times 10^{-5}$. Exact-change form $Y_{10} - Y_0$: .0503 [.0258, .0749].
P2	History contrast τ	.2461 [.1704, .3217]; 12/12 positive; $\mathcal{N}$ floor 2/1,145.
P3	Within-fact contrast	$n = 123$; .3031 [.2425, .3635]; positive/negative/tied 58/1/64.
P4	Recovery ordering by history	$.00175 < .07692 < .25000$; 10/12 weak ordering; minimum adjacent rate gap .0752 [.0280, .1212] (relation-cluster bootstrap).
P5	Margin-deficit-adjusted contrast	$\mathcal{L}$-distribution adjusted difference .1887 [.1107, .2733]; 46 $\mathcal{N}$ rows with $d \leq 4$; 126 unique-control matches.
P6	Censored threshold-time advantage	1.458 [1.092, 1.824]; 12/12 positive; fixed observation grid (Section 3.2).
P7	Adjusted direct savings	.850 [.607, 1.093]; 12/12 positive.
P8	Frozen $(d, C3)$ predictor	BA .8721; BA above chance .3721 [.3244, .4198] (12/12); log-loss gain over margin-deficit-only .1149 [.0830, .1468] (12/12).
P9	Norm-matched displacement bound	$.25\,\tau_{\text{ref}} - \tau_{\text{param}} = .0360$ [.0181, .0538]; Holm .0030; relative norm-matching error 1.15×10^{-7}.
P10	Random-token bound	$.25\,\tau_{\text{ref}} - \tau_{\text{text}} = -.0495$ [−.0757, −.0233]; Holm 1.0: the prespecified bound is rejected and the random-token response exceeds the predicted ceiling (Section 4).
P11	Evaluation-form equivalence	$\tau_{\text{evalform}} - \tau_{\text{ref}} = .0699$; 90% interval [.0176, .1221]; TOST $p = .7456$: equivalence within the prespecified ±.05 margin is not established, and the interval excludes zero in the direction of more recovery under the evaluation form.
P12	KL-matched random-displacement contrast	$\tau_{\text{ref}} - \tau_{\text{function}} = .1566$ [.0770, .2363]; Holm .0030; relative KL-matching error $\leq .0250$.
P13	QA-form and factual-support contrasts	QA delta −.0144 (Holm 1.0); reference-minus-nonce delta .1495 (Holm .0048). Nonce-QA shows no measured advantage over random tokens; the complete-intervention factual contrast may also include delivered dose.
P14	Probe robustness	five configurations positive 12/12; step-20 and second-initialization contrasts positive; configuration floors .00175–.00524; $(d, C3)$ log-loss gains .0726–.1493.

A.2 Protocol, interventions, and endpoint definitions

A.2.1 Design and assignment

To balance the exposure of similar facts, we pair targets within relation before randomization. Pairing follows a fixed greedy nearest-neighbor rule using answer-token count, prompt length in characters, base gold negative log likelihood, and pre-treatment learnability. Feature differences are scaled by their panel ranges; fact identifiers break ties. The peer-exposure factor records whether the other member of this fixed pair was exposed.

The primary protocol specifies the panel, twelve seeds, balanced exposure schedule, interventions, measurement grid, estimands, and P1–P14. Corresponding protocols govern the Pythia-1.4B and Qwen panels. The separate intervention-prediction panel applies its fixed materials, $(d, C3)$ predictor, and threshold without refitting. The exclusion rules prevent a target from entering a development panel or an update support under another name or prompt role. We compare fact identity, normalized subject, answers and aliases, and all three prompt roles, using word-level phrase containment to detect overlap. Materials that pass these rules are called closure-disjoint. The Qwen panel represents every (E, P) cell twice and uses its checkpoint's zero-exact candidate set and a pre-outcome learnability ranking.

Analysis implementation. Two analyzer corrections after the primary run restored P4's pooled weighting and the P9/P10 quarter-reference formulas to their prespecified definitions; the hypotheses and estimands remained unchanged.

A.2.2 Comparisons and uncertainty estimates

To distinguish learning history from inherent fact difficulty, we compare the same fact across chains. The within-fact contrast δ_i averages each eligible fact's step-10 restoration across its $\mathcal{L}$ chains minus its $\mathcal{N}$ chains. To test whether the recovery gap reflects how close answers already are to threshold, we adjust for the starting margin deficit. A logistic model fitted to pooled $\mathcal{L}$ and $\mathcal{N}$ failures, with predictors $[1, E, z(d), P]$, yields the stratum difference averaged over the $\mathcal{L}$ covariate distribution, and caliper matching on d provides matched-pair checks. Relearning savings is the two-group contrast of censored, first-observed threshold times, reported raw and with $z(d)$ adjustment.

These analyses address different alternatives. Comparing the same fact removes differences in fact identity; adjusting its starting deficit compares failures with similar starting margin deficits. Neither changes how the acquired and lost groups were formed. The randomized exposure comparison instead retains the full assigned panel, including facts that remain correct.

Equal-chain intervals describe variation over the tested chains conditional on the fact panel; relation-cluster bootstraps vary sampled relations and facts; randomization inference identifies the exposure effect for this panel and schedule design.

The failure-gated policy outcome Y^{π} retains the R_0 answer for current successes and uses the step-10 answer for current failures; its exposure effect is tested against the sampled legal schedules. The scale and model-family panels use the primary history estimator with equal weight over their eight chains.

For the primary margin-deficit check, nearest-neighbor matching on d within 0.1 pooled standard deviations gives point estimates of 0.144 with replacement and 0.087 without replacement, the latter using 126 unique never-exposed controls. Both retain the same direction within the available overlap, although the limited overlap and attenuation constrain the magnitude of the matched comparison.

A.2.3 Update conditions and matching checks

The update conditions separate the roles of prior cue exposure, answer correctness, factual content, and generic parameter movement. For each condition, we measure the acquired-then-lost minus never-exposed recovery gap. At step 10, the primary history contrasts are $\tau = 0.026$ under norm-matched random displacement, 0.089 under KL-matched random displacement, 0.097 under nonce QA, 0.111 under random-token training, 0.246 under factual support, and 0.316 under factual support in the evaluation form. Thus recovery under generic movement is only part of the observed pattern. The conditions below identify which changes to the update preserve or enlarge the history contrast.

The target-disjoint family varies the incoming update as follows:

$Q_{\text{old,correct}}$ **(reference).** 32 previously trained cue facts with correct answers.

$Q_{\text{old,wrong}}$ / $Q_{\text{fresh},*}$. The same construction with wrong answers taken from other facts, or with 32 cue facts the chain never saw, crossing cue history with label correctness.

Q_{evalform}. The reference cues and answers in the evaluation prompt form, changing the interface while preserving content.

$R_{\text{nonce,QA}}$. 32 QA rows with pseudo-word subjects and answers, token-length matched row-by-row to the reference.

$R_{\text{random,text}}$. 32 sampled-token rows, length-matched to the reference.

$R_{\text{random,param}}$. An untrained random direction in the same rank-one QKV parameterization, scaled so every projection's $\|\Delta W\|_F$ matches the step-10 reference adapter (relative error 1.15×10^{-7}).

$R_{\text{random,function}}$. An untrained random direction scaled by deterministic bisection to match the step-10 reference's calibration KL on 64 held-out prompts (relative error ≤ 0.025 against a 5% tolerance).

The Qwen reference uses the same rank-one construction on separate query, key, and value projections (92,160 trainable parameters); its all-module arm also covers attention output and the gated MLP. Four replicates of each random direction and two each of $R_{\text{random,text}}$ and $R_{\text{nonce,QA}}$ are averaged within chain.

Correct- and wrong-label arms share adapter initialization, cue slots, support order, and token normalization. A wrong-label row retains its cue and replaces the answer with the closure-disjoint replacement answer assigned to that cue slot. The 64 calibration prompts used for KL matching are held out from the primary targets and intervention supports under the same closure keys. Bisection observes only their logits and the reference calibration KL; no target outcome enters the scaling rule. Technical replicates are averaged within chain before chain-level inference.

A.2.4 Training details

We first train the histories that produce the post-interference states, then attach a fresh adapter for each test update. Reloading the saved state before each test makes the update conditions comparable; the histories themselves remain fixed. The primary 410M chain has three full fine-tuning phases: acquisition (96 exposed targets plus 32 cue facts, 20 epochs), structured interference (300 facts, of which 109 share a subject with some target, 12 epochs), and diffuse interference (300 further facts, 12 epochs). Each trained target-disjoint intervention fits a fresh rank-1 LoRA on all query–key–value projections (98,304 trainable parameters), using 32 support examples for up to 20 optimizer steps. Every target is measured at steps $\{0, 1, 2, 3, 5, 7, 10, 20\}$; step 10 is the primary endpoint.

The primary 410M acquisition phase uses AdamW with learning rate 5×10^{-5}, batch size 8, gradient accumulation 2, weight decay 0.01, and 20 epochs. The structured and diffuse interference phases use learning rate 2×10^{-5}, batch size 4, gradient accumulation 4, weight decay 0.01, and 12 epochs each. Both schedules use linear warmup over the first 10% of steps followed by linear decay. Recovery and direct-relearning probes optimize only the newly attached adapter with AdamW, learning rate 10^{-3}, zero weight decay, a constant schedule, microbatch size 4, and global gradient-norm clipping at 1. Each optimizer step makes one complete pass over the arm's support in a fixed order. Token losses are summed across microbatches and divided once by the exact number of valid target tokens in that optimizer step. The target-disjoint arms use 32 support rows; the direct arm uses all 192 targets.

A.2.5 Evaluation and decoding

The behavioral endpoint asks whether the model generates the complete answer and stops. The continuous margin measures how close that same answer is to winning at every token position, including the stop token. Generation is greedy for at most 24 new tokens and must terminate with EOS; empty generations and generations without EOS are invalid. Exact correctness is alias-aware at the token level: the generated token sequence must begin with the complete token sequence for the primary answer or an alias, immediately followed by EOS. Continuous scores in these panels use the primary answer followed by EOS; aliases affect behavioral exactness but not the scored span. For

gold token y_t at each position in that span, the decision margin is

$$M = \min_t \left\{ \ell_t(y_t) - \max_{v \neq y_t} \ell_t(v) \right\},$$

where $\ell_t(v)$ is the raw logit of token v. Thus the weakest token in a multi-token answer determines whether its margin is non-negative.

To check whether exact-match failures still contain the correct answer in a looser form, a separate descriptive post-analysis uses an intentionally permissive content screen: after NFKC normalization, case folding, whitespace normalization, and conservative edge-punctuation trimming, it tests whether an accepted answer or alias occurs as a substring of the R_0 generation. Of the 135 acquired-then-lost reference-arm restorations across the 410M, 1.4B, and Qwen panels, 128 (94.8%) contain no accepted answer content at R_0; seven contain it. Only two of these R_0 generations are invalid (one empty and one reaching the 24-token cap without EOS). Of the 47 additional 410M acquired-then-lost restorations achieved by evaluation-form support but not by the reference arm, 43 (91.5%) also contain no accepted answer content at R_0. This census distinguishes content replacement from endpoint validity; exactness remains the token-level endpoint above.

A.3 Additional limitations

Facts and training regimes. The experimentally installed panels come from the top quartile of a pre-outcome learnability screen, so the estimates concern facts this procedure can install reliably rather than arbitrary base-unknown facts.

Comparing updates and models. Norm and calibration-KL matching checks are available where the design explicitly matches update dose, but not for all trained controls; delivered dose may contribute to differences among factual, nonce, and random-token support. Recovery uses fixed-budget, rank-limited LoRA probes, and step-matched updates differ in calibration KL across scales. Relearning savings uses threshold times observed through step 10 and censored at 11. Qwen changes architecture, tokenizer, pretraining, and materials together, so model-family differences cannot be attributed to architecture alone. The natural-task stream and factual interference are not matched for damage severity. The comparisons of threshold proximity and update sensitivity use complete interventions; cross-scale effect magnitudes are descriptive. The eight-chain 1.4B behavioral replication and the separate four-arm tangent confirmation answer different questions.

B Full-factorial cue-history evidence

The separately prespecified twelve-chain factorial experiment tests whether replacing wrong labels with correct ones matters more for previously trained cues than for fresh cues, and whether target exposure changes that difference. Within a chain, we first subtract the wrong-label response from the correct-label response for each cue history. Comparing these two differences isolates the cue-by-label interaction. We then compare the interaction between exposed and withheld targets, with peer exposure balanced. Here S_{cue} is the peer-adjusted exposure effect on $(Q_{\text{old,correct}} - Q_{\text{old,wrong}}) - (Q_{\text{fresh,correct}} - Q_{\text{fresh,wrong}})$ over the full randomized panel. Its exact interactions are 0.1319 $[0.1000, 0.1639]$ and 0.0981 $[0.0656, 0.1306]$ on its development and validation target panels, positive in all twelve chains in both. The corresponding margin interactions are 1.926 $[1.725, 2.127]$ and 1.473 $[1.196, 1.749]$. Together with the primary Pythia-410M history-contrast interaction of 0.143 $[0.016, 0.271]$, these results show that cue history repeatedly changes the effect of label correctness.

The twelve-chain run uses fresh seeds and meets its prespecified damage-range criterion in eleven, above the ten-seed requirement. It follows an earlier eight-chain experiment in which three seeds met the criterion, below the required seven. The confirmation estimates above use only the fresh twelve-chain run.

The separate intervention-prediction panel also ran all four factorial cells, although its operating gate used only the old-correct, old-wrong, and fresh-correct arms. A secondary peer-adjusted full-panel analysis gives an old-cue correctness increment of 0.1441 $[0.1198, 0.1683]$, a fresh-cue increment of 0.0408 $[0.0228, 0.0587]$, and their interaction of 0.1033 $[0.0753, 0.1313]$. On the post-treatment set of exposed current failures, the corresponding interaction is 0.0163 $[-0.0311, 0.0636]$. This panel corroborates the full-panel interaction; its current-failure subset is inconclusive.

C EXPOSURE-CONDITIONED QKV TANGENT-STATE CONFIRMATION

We ask whether prior exposure changes how an incoming update would locally move a target-answer score, before that update is applied. We then test whether this pre-update measurement helps predict the effect of ten actual training steps. The gradients below compare the direction that raises the target score with the direction induced by training on the support facts.

Coordinates and local alignment. The trained probes update QKV projections through LoRA adapters. We use that same projection family to measure an update's local direction at the saved checkpoint. For this measurement, we attach virtual adapters with zero output factors: their contribution to the forward pass is zero, so the model still makes exactly the checkpoint's predictions. The input factors remain fixed and nonzero, allowing us to differentiate the answer score and support loss with respect to the output factors without taking a training step.

The coordinate vector ϕ concatenates the output factors of virtual rank-one LoRA adapters on all QKV projections. Gradients are evaluated at zero output factors, with input factors held fixed at independent standard-normal draws scaled by the inverse square root of the projection's input dimension.

For state $t \in \{\mathrm{base}, R_0\}$, validation target i, support arm q, and probe initialization r, the tangent cosine is

$$c_{t,i,q,r} = \cos(\nabla_\phi \log p_t(y_i), -\nabla_\phi L_{t,q}), \tag{2}$$

where $\log p_t(y_i)$ is averaged over the primary answer-plus-EOS span and $L_{t,q}$ is token-normalized support loss. Gradient norms enter the prediction baseline below. The ten-step experiment tests whether this alignment is informative about the response to actual optimization.

The two probe initializations are averaged itemwise. We compare correct-versus-wrong alignment separately for old and fresh cues, then ask how that difference changes between the base and post-interference states. With OC, OW, FC, and FW denoting the old/fresh-cue by correct/wrong-label arms,

$$O_t = c_{t,OC} - c_{t,OW}, \quad F_t = c_{t,FC} - c_{t,FW}, \quad A_t = O_t - F_t, \quad \Delta A = A_{R_0} - A_{\mathrm{base}}. \tag{3}$$

For any item quantity x, $D_s(x)$ averages its exposed-minus-withheld difference over the two peer cells. The finite endpoint Y_{margin} applies the same cue-by-label interaction to answer-plus-EOS margin changes from R_0 after the eight real ten-step Q updates formed by four arms and two initializations. The local and finite measurements therefore ask the same comparative question: does changing label correctness matter more for old cues than for fresh cues? One compares gradient directions before an update; the other compares margin changes after real training. Exposure effects compare each quantity between targets assigned to be exposed or withheld.

Learning the predictor before testing new histories. At 410M, seven discovery histories were used once to select the old-cue contrast and fit the two linear predictors. Their coefficients and standardization constants were fixed before confirmation on 27 fresh histories; both stages use the same 96 validation targets under separate randomized schedules, giving 2,592 confirmation rows. The 1.4B experiment carries forward the same six quantities and uses all 192 targets. Its scale-specific predictors are fit on four discovery histories and fixed before 12 untouched confirmation histories. Confirmation data are not used for selection, standardization, or fitting at either scale.

Prediction beyond current scores and gradient magnitudes. To test whether the alignment change adds information beyond these quantities, the baseline predictor of the finite interaction Y_{margin} contains an intercept, E, P, $E{\times}P$, A_{base}, the R_0 target margin, and the log norms of the R_0 target gradient and four support gradients. The full predictor adds only ΔA. Both are unregularized least-squares fits using discovery-set standardization. For each confirmation history, the prospective gain is the baseline mean-squared error minus the full-predictor mean-squared error. Unlike the early-response recovery predictor, neither receives an observed step of the update it is asked to predict. At 1.4B, the four centered support-norm covariates are linearly dependent across the four discovery histories, and both predictors use the prespecified minimum-norm least-squares solution.

Table 2: Prospective tangent-state confirmation at two Pythia scales. Cells report the equal-unit mean (one-sided 95% lower bound), two-sided 95% interval, and number of positive confirmation histories. The fixed conjunction requires the lower bound to exceed zero for every quantity; all six conditions satisfy that criterion at both scales. Each D_s row is an exposed-minus-withheld target comparison, averaged over peer-exposure conditions.

Quantity	**Pythia-410M** ($n = 27$)	**Pythia-1.4B** ($n = 12$)
$D_s(A_{R_0})$: post-interference interaction	.0124 (.0073), [.0063, .0184]; 21/27	.0275 (.0228), [.0217, .0333]; 12/12
$D_s(\Delta A)$: change from base interaction	.0191 (.0115), [.0099, .0283]; 21/27	.0268 (.0231), [.0223, .0314]; 12/12
$D_s(O_{R_0})$: old-cue label contrast	.0140 (.0104), [.0097, .0182]; 24/27	.0382 (.0320), [.0306, .0458]; 12/12
$D_s(\Delta O)$: change from base old-cue contrast	.0158 (.0108), [.0097, .0219]; 24/27	.0378 (.0315), [.0301, .0456]; 12/12
$D_s(Y_{\text{margin}})$: ten-step margin interaction	1.932 (1.765), [1.730, 2.133]; 27/27	2.391 (2.171), [2.121, 2.661]; 12/12
Frozen-predictor MSE gain	.223 (.164), [.152, .295]; 23/27	1.154 (.883), [.822, 1.486]; 12/12

Baseline sensitivity at 1.4B. As a post-hoc baseline sensitivity analysis at 1.4B, we omitted the four support-gradient norms, which are constant within each history, and refitted both predictors using only the original four discovery histories. Adding ΔA reduced MSE from 5.042 to 4.177 across the same twelve confirmation histories: gain 0.865 $[0.730, 1.000]$, positive in all twelve. The original frozen comparison in Table 2 is unchanged.

Decomposing the interaction by cue and label. At 410M, the untouched-base interaction is $-.0068$ $[-.0137, .0001]$. The fresh-cue component is $.0016$ $[-.0038, .0070]$ at R_0 and $-.0033$ $[-.0114, .0047]$ after base subtraction. In the armwise base-to-R_0 decomposition, the old-cue wrong-label component is $-.0174$ $[-.0232, -.0117]$, while the old-cue correct-label component is $-.0016$ $[-.0090, .0057]$. The exposure-induced interaction is therefore concentrated in sensitivity to updates that conflict with a previously trained cue–answer mapping.

At 1.4B, the untouched-base interaction is $.0007$ $[-.0012, .0026]$. The fresh-cue base-to-R_0 component is $.0110$ $[.0063, .0157]$, while the old-cue component is $.0378$ $[.0301, .0456]$.

The positive $D_s(\Delta A) = .0268$ therefore records a larger exposure-induced change for old cues rather than the absence of a fresh-cue change.

Comparison with realized acquisition history. The prespecified acquisition-stratified secondary estimates are centered near zero: for A_{R_0}, $\mathcal{L}$-minus-$\mathcal{M}$ is $.0015$ $[-.0199, .0230]$ and $\mathcal{L}$-minus-$\mathcal{N}$ is $.0078$ $[-.0079, .0234]$; for ΔA, the corresponding values are $-.0155$ $[-.0455, .0146]$ and $.0055$ $[-.0173, .0283]$. The continuous post-acquisition gain slope is $.0011$ $[-.0014, .0035]$. These secondary comparisons show no clear separation by realized acquisition status. The positive randomized result instead concerns how assigning exposure changes the four-arm geometry.

D Recovery robustness and threshold diagnostics

D.1 Robustness across probe configurations (410M)

The main recovery experiment uses a small QKV adapter. Here we check whether its history contrast and early-response prediction depend on that particular rank, set of updated modules, or initialization. Five probe configurations spanning rank 1–8 and three module families (up to 786,432 trainable parameters) all yield positive τ in 12 of 12 chains, with never-exposed floors between 0.17% and 0.52%, and the $(d, C3)$ predictor improves log loss over margin deficit alone in every configuration

(0.073–0.149). A second initialization reproduces the contrast. Its magnitude and the recovered facts vary with the probe.

Table 3: The history contrast across probe configurations, primary initialization. Every configuration is positive in 12/12 chains.

Configuration	τ @10	95% interval	τ @20	$\mathcal{N}$ floor @10	$(d, C3)$ log-loss gain
rank 1, QKV (ref)	0.2461	$[0.1704, 0.3217]$	0.3172	0.00175	0.1149
rank 4, QKV	0.1729	$[0.0943, 0.2515]$	0.2543	0.00175	0.0828
rank 8, QKV	0.1708	$[0.1108, 0.2307]$	0.2744	0.00262	0.0726
rank 1, MLP	0.2239	$[0.1496, 0.2982]$	0.3234	0.00262	0.0749
rank 1, all modules	0.2630	$[0.1863, 0.3396]$	0.3290	0.00524	0.1493
ref, second init	0.1662	$[0.1067, 0.2258]$	—	—	—
all modules, second init	0.2913	$[0.2131, 0.3695]$	—	—	—

In a descriptive post-analysis, pairwise Jaccard overlap of sets of restored $\mathcal{L}$ fact–chain instances across configurations spans 0.455–0.600: the configurations recover substantially overlapping sets.

The direct arm reproduces its savings under the broadest configuration (2.232 $[1.857, 2.606]$ raw; 1.304 $[1.047, 1.561]$ adjusted).

To test whether insufficient capacity to learn new facts explains the low never-exposed recovery rate, the capacity arm trains on 32 never-exposed facts per chain and acquires 58 of 384 such instances by step 20 (15.1%). Held-out exact accuracy is 7/768 (0.91%), including five restorations among 761 held-out instances that failed at R_0 (0.66%). The low $\mathcal{N}$ rates under Q arms therefore measure specificity of recovery when the target is excluded: the same probe can acquire new facts when they are directly supplied in its training support.

D.2 Threshold-crossing diagnostics

If the recovery ordering under random displacement reflects starting margin deficits, applying the same displacement distribution to each group's deficits should reproduce that ordering. This descriptive post-analysis was specified after the primary analysis and then applied unchanged to the 410M and 1.4B panels. Let Δ_{ir} be the margin displacement produced on failed fact i by a KL-matched random-displacement replicate r. Under the greedy exact endpoint, exact correctness and a non-negative answer margin coincide on all evaluated rows, so recovery under this arm can be represented as the threshold event $\Delta_{ir} \geq d_i$; the empirical question is whether the displacement distribution itself varies materially with acquisition history.

Across the four independent replicates per chain, median per-fact margin standard deviations are 1.14, 1.11, and 1.02 for $\mathcal{L}$, $\mathcal{M}$, and $\mathcal{N}$ at 410M, and 0.94, 0.91, and 0.89 at 1.4B; mean displacement is near zero in every stratum. Among 410M failures with $d \leq 2$, replicate-averaged recovery under the KL-matched random-displacement arm is 0.221 in $\mathcal{L}$ ($n = 129$) and 0.194 in $\mathcal{N}$ ($n = 9$), a difference of 0.027. At 1.4B, the corresponding never-exposed cells contain only four instances for $d \leq 2$ and fourteen for $d \leq 4$, so we do not use the shallow-deficit check at that scale.

Without fitting a model, we pool the observed displacement distribution from the other strata across chain–fact instances and apply it to each stratum's observed deficits; the predicted and observed rates below use the same pooled chain–fact weighting. At 410M they are $0.120/0.099$ for $\mathcal{L}$, $0.050/0.025$ for $\mathcal{M}$, and $0.0029/0.0017$ for $\mathcal{N}$. At 1.4B they are $0.109/0.086$, $0.037/0.022$, and $0.0012/0.0000$ (Figure 5). The predictions mildly overestimate each rate but reconstruct the stratum ordering and the large $\mathcal{L}$-to-$\mathcal{N}$ ratio from deficit geometry. In contrast, factual support retains an acquired-history advantage of 0.189 after margin-deficit adjustment (prespecified P5), and caliper matching preserves the same direction. The random and factual arms therefore distinguish generic threshold crossing from content-dependent access.

The prospective four-arm experiments complement this reconstruction by testing how history shapes the displacement itself. At both Pythia scales, randomized exposure increases the cue-by-label

interaction in actual ten-step margin changes, and training-induced tangent changes improve its prediction over fixed baselines that include the starting margin.

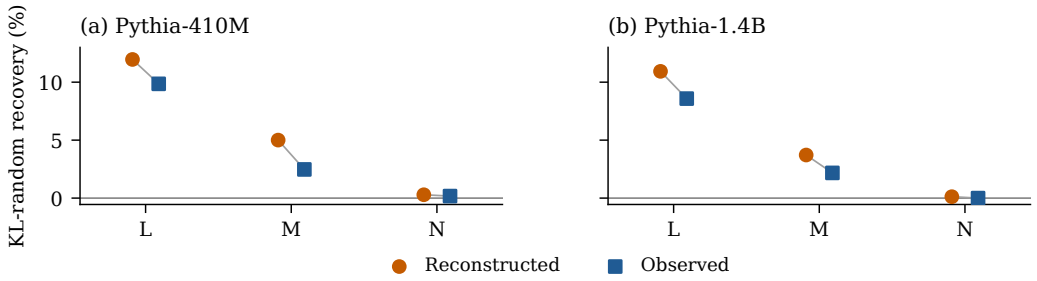

Figure 5: Threshold reconstruction under KL-matched random displacement. Recovery reconstructed from each stratum's starting deficits and the pooled displacement distribution of the other strata, alongside observed step-10 recovery. Both use the same chain–fact weighting and four displacement replicates per chain. The descriptive reconstruction reproduces the stratum ordering at both Pythia scales.

E Separate intervention and recovery-prediction evaluation

To evaluate the intervention comparisons and early-response predictor on new facts and histories, this panel uses 192 fresh targets and 12 fresh chain seeds. Its manipulation criterion holds in 12 of 12 chains, and every outcome is evaluated with the interventions and predictor fixed before the panel's outcomes. The two-part recovery test asks whether old-cue correct-label support has both an advantage over wrong labels and an advantage over fresh correct cues. This tests a stronger pattern than recovery alone: both label correctness and previous cue training must help in their respective comparisons. The target-disjoint reference restores 14.7% of exposed current failures, against 8.6% under wrong labels and 14.7% under fresh correct cues. The correct-label contrast is $+0.0613$ with lower confidence bound $+0.0254$; the old-versus-fresh contrast is -0.0003. Correct labels therefore help relative to wrong labels on old cues, while the two correct-label supports restore similar observed fractions in this panel. The prespecified two-part recovery criterion is not met: its correctness component passes, whereas its old-versus-fresh component does not.

Relearning and forecasting recovery. The full-panel exposure advantage in first-observed, censored threshold time (withheld minus exposed) is $+8.132$ steps $[7.860, 8.404]$, positive in 12 of 12 chains; this estimand includes retention at step 0. A descriptive post-analysis among current failures gives a margin-deficit- and peer-adjusted advantage of $+1.204$ steps $[1.054, 1.354]$, positive in all 12 chains.

The frozen $(d, C3)$ predictor exceeds chance balanced accuracy in 12 of 12 chains and improves log loss over the margin-deficit-only rule in 12 of 12, with calibration slope 0.970 and intercept -0.224. This predictor combines the starting deficit with the response observed after three actual update steps to forecast step-10 recovery. Its success on the new panel supports early-response forecasting across independent training histories.

F Cross-model replication

The cross-model comparison separates replication of the history effect from replication of particular differences between updates. Table 4 places the full-panel exposure effect, recovery contrasts among current failures, and direct-relearning savings side by side, followed by the shared update arms.

F.1 Pythia-1.4B replication

The eight fresh chains use the separately specified 1.4B protocol. The larger model reproduces history-specific recovery, direct savings, in-support acquisition, and predictor transfer. The following

arm values are the acquired-then-lost minus never-exposed recovery contrasts at step 10: reference .1392, wrong-label .0357, fresh-cue .1423, nonce .0896, KL-matched random displacement .0805, evaluation-form .2848, all-module reference .2258; at the secondary step 20 the reference reaches .2175 and the evaluation form .4132. On the fixed 0–10 observation grid, the exposed-versus-never-exposed censored threshold-time advantage is 1.59 $[0.87, 2.32]$. A descriptive stratum decomposition records 76/201 acquired-then-lost, 3/23 exposed-but-unacquired, and 2/768 never-exposed crossings, with $\mathcal{L}$ savings 1.74 $[1.01, 2.47]$; the $\mathcal{M}$ cell is too small to estimate separately at this scale.

Reference exceeds KL-matched random displacement by 0.059 $[0.024, 0.094]$, and evaluation form exceeds reference by 0.146 $[0.076, 0.215]$. All eight chains pass the manipulation criterion (mean lost fraction 0.272). Two other parts of the prespecified behavioral pattern do not replicate: reference-minus-nonce falls just below its 0.05 margin, and reference is slightly below fresh-correct support.

At step 10, calibration KL is .2657 (versus .5326 at 410M), and the relative KL-matching error is $\leq .0245$.

F.2 QWEN2.5-0.5B REPLICATION AND CROSS-MODEL SUMMARY

The 1.4B behavioral panel uses a reduced intervention roster, separate from its tangent experiment; the Qwen panel uses 192 fresh targets. These configurations jointly vary model family, tokenizer, pretraining, and fact panel. For Qwen, prior exposure increases correctness under the policy that updates only current failures by .7070 [.6927, .7213] with a 100,000-schedule one-sided randomization $p = 9.9999 \times 10^{-6}$. The reference arm restores $27/222$ $\mathcal{L}$, $0/30$ $\mathcal{M}$, and $0/768$ $\mathcal{N}$ instances. The acquired-then-lost minus never-exposed recovery contrasts for two other arms are $.0612$ $[.0223, .1000]$ for old-cue wrong labels and $.1260$ $[.1102, .1417]$ for fresh-cue correct labels. The six direct contrasts are: reference minus old-wrong $.0511$ $[-.0153, .1176]$, reference minus KL-matched random displacement $.0426$ $[-.0171, .1023]$, reference minus nonce $.0377$ $[-.0252, .1006]$, evaluation-form minus reference $.0538$ $[.0218, .0858]$, reference minus fresh-correct $-.0137$ $[-.0778, .0504]$, and all-module minus reference $.1135$ $[.0191, .2078]$.

These comparisons locate the replicated pattern: the history contrast remains positive, while several finer differences between factual and control updates are uncertain. The Qwen replication reloads each chain's saved R_0 across arms and reproduces it at step 0; exposure is balanced within fact and peer cell, Q supports are closure-disjoint, and technical replicates are averaged within chain.

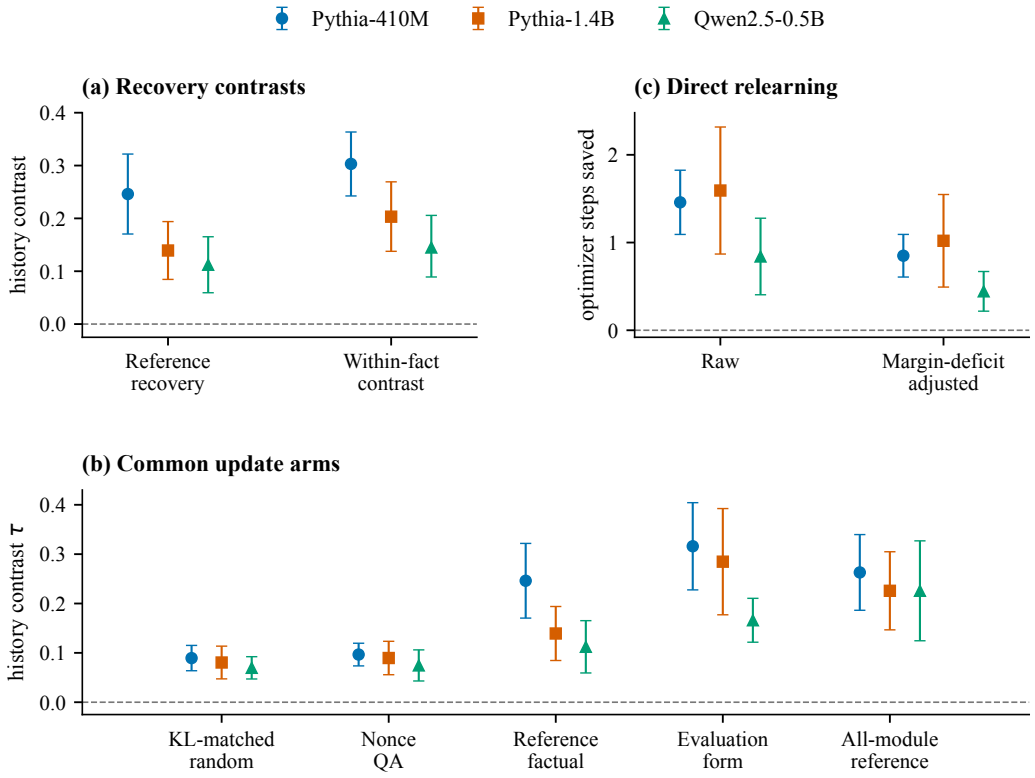

Figure 6: Cross-model synthesis. (a) The reference target-disjoint history contrast and within-fact contrast. (b) History contrasts for the five update arms shared by all three model configurations. (c) Raw and margin-deficit-adjusted direct-relearning savings on the fixed, censored 0–10 observation grid. Points are equal-chain means, except that within-fact contrasts average over eligible facts. Bars for reference recovery, update arms, and savings are equal-chain 95% t intervals; within-fact bars are 95% relation-cluster bootstrap intervals. Pythia-410M has 12 chains; Pythia-1.4B and Qwen2.5-0.5B have 8 each.

Table 4: Cross-model results. (a) Core quantities; savings use the fixed, censored 0–10 observation grid. (b) History contrasts for the shared update arms. Intervals are equal-chain 95% t intervals except the within-fact row, which uses a relation-cluster bootstrap. Recovery prediction uses the predictor frozen on the separate 410M factorial panel for the 410M cell and within-model leave-one-chain-out fits for the 1.4B and Qwen cells.

(a) Core quantities

Quantity	**Pythia-410M**	**Pythia-1.4B**	**Qwen2.5-0.5B**
Full-panel $Y_{10} - Y_0$ exposure contrast	.0503 [.0258, .0749]	.0339 [.0102, .0575]	.0221 [−.0009, .0451]
Reference τ	.2461 [.1704, .3217]	.1392 [.0845, .1939]	.1123 [.0593, .1652]
Within-fact contrast	.3031 [.2425, .3635]	.2032 [.1377, .2692]	.1449 [.0891, .2057]
Raw direct savings	1.458 [1.092, 1.824]	1.593 [.869, 2.317]	.841 [.405, 1.277]
Margin-deficit-adjusted savings	.850 [.607, 1.093]	1.020 [.492, 1.548]	.444 [.218, .670]
$d + C3$ log-loss gain	.115 [.083, .147]	.094 [.033, .154]	.100 [.019, .182]
Capacity at step 20: in-support; held-out	58/384; 7/768	101/256; 0/512	7/256; 0/512

(b) Shared update arms

Target-disjoint update	**Pythia-410M**	**Pythia-1.4B**	**Qwen2.5-0.5B**
KL-matched random displacement	.0894 [.0638, .1151]	.0805 [.0473, .1136]	.0697 [.0471, .0923]
Nonce QA	.0966 [.0736, .1195]	.0896 [.0557, .1234]	.0746 [.0430, .1061]
Reference factual support	.2461 [.1704, .3217]	.1392 [.0845, .1939]	.1123 [.0593, .1652]
Evaluation-form support	.3160 [.2276, .4043]	.2848 [.1771, .3924]	.1660 [.1216, .2105]
All-module reference	.2630 [.1863, .3396]	.2258 [.1466, .3049]	.2257 [.1245, .3270]

G PRETRAINED KNOWLEDGE AND NATURAL-TASK INTERFERENCE

G.1 RECOVERY AND PREDICTION FOR FACTS KNOWN AT THE PRETRAINED CHECKPOINT

This extension asks whether recovery and pre-update prediction also apply to facts already answered correctly by the pretrained model. These facts receive no experimental acquisition training, so the test goes beyond the installed facts used in the randomized exposure experiments.

Materials and holdouts. The pretrained-known panel contains 121 facts across 24 relations: 79 from CounterFact (Meng et al., 2022), 37 from T-REx and five from Google-RE as released with LAMA (Petroni et al., 2019). Entry requires a correct answer to a fixed canonical prompt from untouched Pythia-410M. Candidates are deduplicated, limited to eight per relation, and screened for literal overlap of target subjects, answers, and comparison answers with the parent acquisition, interference, and Q materials. The model never receives these target labels in experimental training. Selection uses base-only observations; one percent-encoded subject is excluded without prompt repair. Two alternate prompts per relation are fixed before their evaluation and do not determine entry.

A hash split assigns 61 fact identities to development and 60 to confirmation. Development uses seven saved 410M tangent histories (15–21); confirmation uses 20 different histories (22–41). All checkpoints and eight real ten-step Q adapters per history come from Appendix C. They are reused training histories, with new measurements on the pretrained-known panel. The predictor's holdouts are both fact identity and history; relation types may occur in both splits. Parent exposure assignments concern the installed inventory, not randomized exposure of these pretrained-known facts.

Answer-content endpoint. This extension measures whether the known answer content returns. Unlike the installed-fact exact endpoint, it allows text after the complete answer: a continuation can supply the answer correctly and then continue the sentence. The boundary check below prevents an answer that is merely the beginning of a longer word from counting as correct. Correctness requires the normalized continuation to begin with a gold answer followed by whitespace or punctuation; an exact-length continuation counts only when it ends with EOS. Normalization uses NFKC, case folding, and whitespace collapse. Generation includes an extra boundary token. Target scores are mean answer-token log probabilities without EOS, so they track the answer content separately from the probability of stopping. For each prompt, the recovery risk set consists of facts correct at base and valid but wrong at R_0, fixed before Q. Invalid Q outputs count as no restoration. The two adapter initializations are averaged within fact and arm, then facts within history, then histories equally.

Canonical confirmation comprises 1,200 fact–history observations: 803 at risk (32–46 per history), 376 still correct, and 21 invalid at R_0. All 60 facts enter the canonical risk set in at least one history. The two alternate prompts are base-correct for 17 and 32 of the 60 facts, with 262 and 415 at-risk observations. Their recovery rates in Table 5 therefore describe their own base-known populations. An alternate prompt answered incorrectly at base is not counted as recovering that fact.

Predicting update response from pre-update geometry. Because these facts were not experimentally exposed or withheld, we compare each target answer with a fixed alternative. We ask whether their relative gradient alignment predicts their relative response to actual training. The comparison makes the prediction answer-specific: an update may raise both answers' scores, but favor the gold answer more. We measure that relative change before combining the four cue-by-label arms. Each fact has a fixed same-relation comparison answer chosen from the source data without using model scores or recovery outcomes, with equal canonical token length and static truth/alias/overlap exclusions. These are source-grounded comparisons, rather than individually certified false answers. For gold or comparison answer z, let $c^z_{t,q}$ denote the cosine between its answer-only score gradient and support descent at state t in the same zero-output-factor, rank-one all-QKV coordinates as Appendix C. Support descent retains the full 32-row answer-plus-EOS loss and token normalization. Cosines are computed separately for each initialization and then averaged. Define

$$A^z_t = (c^z_{t,OC} - c^z_{t,OW}) - (c^z_{t,FC} - c^z_{t,FW}), \qquad Z_t = A^g_t - A^c_t, \tag{4}$$

$$u_q = (s^g_{Q10,q} - s^g_{R_0}) - (s^c_{Q10,q} - s^c_{R_0}), \qquad Y_{\mathrm{rel}} = (u_{OC} - u_{OW}) - (u_{FC} - u_{FW}). \tag{5}$$

Table 5: Pretrained-known confirmation on 20 held-out histories. Intervals are two-sided 95% Student-t intervals across equally weighted histories. Alternate-prompt rows use their own base-correct, currently failed risk sets. Prediction targets the finite relative-answer interaction in Equation 5.

Quantity	Mean	95% interval	Positive histories
Canonical OC restoration (%)	7.56	[6.08,9.04]	20/20
Alternate 1 OC restoration (%)	10.23	[6.34,14.13]	16/20
Alternate 2 OC restoration (%)	8.04	[5.90,10.17]	19/20
Full-panel MSE gain	.414	[.250,.578]	18/20
Canonical-risk MSE gain	.339	[.145,.533]	15/20

Here g, c denote gold and comparison answers. Every $Q10$ score is measured after replaying the actual trained adapter on its own history's R_0. The first comparison is gold versus alternative; the second is correct versus wrong labels; the third is old versus fresh cues. Applying this same sequence to the pre-update gradients and the observed score changes connects the local measurement to the finite training response.

To test whether post-interference alignment adds information beyond current scores, base geometry, correctness, and gradient magnitudes, the baseline OLS predictor contains an intercept, Z_{base}, paired gold-minus-comparison scores at base and R_0, R_0 gold correctness, two log target-gradient norms, and four log support-gradient norms. The full predictor adds only Z_{R_0}. Standardization and unregularized least-squares coefficients use the 427 development fact–history observations and are frozen before confirmation. The same two predictors are evaluated on all 60 confirmation facts and on each history's canonical risk set, without refitting.

Recovery and prediction on pretrained-known facts. Mean baseline/full MSE is $3.724/3.310$ on the full panel and $3.716/3.377$ on canonical failures. Mean history-wise relative improvements are 10.99% and 9.42%, respectively (Figure 7). The three prespecified one-sided 95% lower bounds are positive: .0633 for canonical OC restoration, .2786 for full-panel MSE gain, and .1786 for risk-set MSE gain. Thus pre-update geometry improves prediction of later response on facts already known at the pretrained checkpoint, including those that subsequently became incorrect.

Cue-by-label response across prompts. We also examine whether the recovery-level cue-by-label interaction extends to these facts and remains consistent across prompt forms. Canonical restoration under OC, OW, FC, and FW is respectively 7.56%, 8.42%, 7.95%, and 10.80%. Their restoration interaction is 1.98 percentage points $[-0.19, 4.16]$; alternate-prompt interactions are 0.61 $[-3.89, 5.11]$ and -3.06 $[-5.88, -0.25]$. The behavioral old-cue correctness interaction varies with the prompt, whereas recovery and prospective relative-answer prediction recur in this pretrained-known setting.

G.2 Recovery under natural-task interference

To test whether history-dependent recovery persists when subsequent task training causes the forgetting, we replace factual interference with a natural-task stream. This keeps the exposure design and recovery tests fixed while changing what the model learns between acquisition and recovery. The results show history-dependent recovery and relearning savings in the observed regime, where damage is more severe than planned.

The natural-interference panel keeps Pythia-410M, randomized factual exposure, the acquisition phase, and the 96-target validation panel. It replaces the two factual interference phases with 300 NumGLUE-ds examples followed by 300 FOMC examples, six epochs each, selected deterministically from TRACE's released task stream (Wang et al., 2023; Mishra et al., 2022; Shah et al., 2023). FOMC uses 100 training examples per class (dovish, hawkish, neutral). Held-out task panels contain 128 NumGLUE-ds and 126 FOMC examples, the latter balanced across classes. Training uses full fine-tuning with the inherited optimizer and loss normalization. All 12 fixed histories execute both old-correct Q10 and direct validation-target relearning from their own R_0, each with two fixed adapter initializations. Targets, answers and aliases retain the parent training-support exclusions.

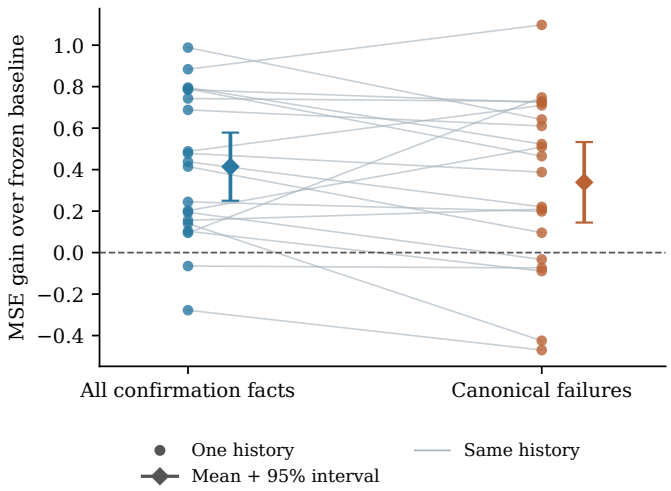

Figure 7: Finite-response prediction on pretrained-known facts. Each connected pair is one of 20 held-out histories, evaluated over all 60 confirmation facts and over its canonical failure set using the same frozen predictor. Positive MSE gain means adding pre-update Z_{R_0} improves prediction. Large markers and bars show equal-history means and two-sided 95% intervals.

Table 6: Natural-task interference: complete twelve-history response data. Intervals are two-sided 95% Student-t intervals across equally weighted histories.

Quantity	**Mean**	**95% interval**	**Positive histories**
Exposure effect on exact change $H(G^Y)$	.569	[.509,.630]	12/12
Exposure effect on margin change $H(G^M)$	1.866	[.844,2.888]	11/12
Acquired-lost minus never-exposed recovery	.794	[.751,.838]	12/12
L–N direct savings (steps)	4.747	[4.374,5.119]	12/12

Training regime. The six-epoch dose was fixed after damage-only development and before any confirmation recovery outcome. Actual loss of acquired validation facts ranges from 61.9% to 88.6%, and both natural-task held-out losses decrease in every history. The planned range was 15–65%, with at least ten of twelve histories required to meet all manipulation criteria. Eleven histories exceed the loss cap; one of those also misses the post-acquisition cue-accuracy requirement, leaving one history within all criteria. The original composite confirmation criterion was therefore not met (`INVALID_DAMAGE_REGIME`). Exposure assignment and outcome collection retain the prospectively fixed design, and Table 6 reports all twelve histories at the prespecified endpoints, without selection by damage or recovery.

Recovery and relearning after task-induced forgetting. We use the same randomized exposure comparison as before, now measuring both exact correctness and margin change. For any update outcome V, define the peer-balanced exposure contrast $H_s(V) = \frac{1}{2}\sum_{p=0}^{1}[\overline{V}_{E=1,P=p,s} - \overline{V}_{E=0,P=p,s}]$. The co-primary quantities are $H_s(G^Y)$ and $H_s(G^M)$, where $G^Y = Y_{Q10} - Y_{R_0}$ is change in exact correctness and $G^M = M_{Q10} - M_{R_0}$ is change in the answer-plus-EOS margin. The exact-change outcome counts restored answers positively and newly lost answers negatively. Comparing this net change between exposure assignments asks whether prior learning still alters the effect of the same recovery update after natural-task training. Restoration and collateral loss are also separated. Technical initializations are averaged before peer and equal-history aggregation. Here direct savings compares acquired-then-lost with never-exposed facts: the peer-balanced difference $T_N - T_L$ on the observed grid $\{0, 1, 2, 3, 5, 7, 10\}$, censored at 11.

The 1,152 validation fact–history instances include 402 acquired-then-lost, 108 retained after acquisition, 66 exposed but unacquired, and 576 withheld. Q restores an initialization-averaged 320 of

the 402 lost instances; withheld restoration is zero. All 108 retained instances remain correct under both Q initializations. The 402 lost R_0 generations contain no accepted answer or alias: 401 are valid wrong answers and one is empty. Of these, 327 are single FOMC class labels. Thus the stream often changes the output task as well as factual expression; recovery here can include restoration of the factual response interface.

Together, the results show that historical differences remain consequential after substantial task-induced disruption: previously exposed facts recover more readily and acquired-then-lost facts require less direct retraining in this severe-damage regime.

H Historical parameter interventions

H.1 Historical binding influence through native parameters

We test whether native parameters retain an influence specific to which answer was paired with which query, even after the model stops producing that answer. Replacing the same blocks in a common recipient lets us vary the source of this historical influence while keeping the rest of the model fixed.

Design. Three independent training triplets test a graded binding contrast fixed after an exploratory triplet, which is excluded from the results below. Each triplet freshly trains a recipient R and two donors D0/D1 from Pythia-410M. Nine fixed queries form three answer-compatible blocks: stream names, record labels, and person successors. For each three-query block, imagine canonical answers A/B/C: one donor instead learns B/C/A and the other C/A/B. R receives canonical assignments; the donors receive the two opposite nonzero cyclic permutations within each block, with their orientation randomized before training. Both donors therefore encounter each answer once per block. Original prompts are retained, with equal answer-token lengths within a block. Four focal facts are inserted by common equal-length replacements in every history's acquisition set, keeping it at 128 rows.

Within a triplet, prompts, answer multisets, effective exposure, row orders, per-step token denominators, optimizer, and training RNG are matched; only the nine bindings differ. Acquisition uses 20 epochs, followed by the original structured and diffuse interference phases, each with 300 rows and 12 epochs. Their text is identical across donors and contains no literal candidate answers, although same-subject interference varies across focal facts. Between triplets, fresh epoch shuffles produce different training trajectories. We report values for the three independent training triplets, conditional on this fixed base model and query inventory, rather than treating queries or donors as independent samples for confidence intervals.

At post-interference R_0, each intervention restores native R and replaces all 96 tensors in blocks 0–7 with the corresponding donor tensors: attention, MLP, biases, and within-block LayerNorm. Embeddings, blocks 8–23, final LayerNorm, and output head remain R's. Exact R-to-R self-copy is checked. We measure the resulting hybrid directly after replacement:

$$H_d = (\theta_{0:7}^{R_0(D_d)}, \theta_{\text{complement}}^{R_0(R)}). \tag{6}$$

Binding contrast. If the replaced blocks carry the donor's historical pairings, changing donors should shift the recipient's relative scores toward the new donor's assigned answer. We compare the two candidate answers under both hybrids to measure that shift. For one query, this means asking whether the first hybrid favors its donor's answer more than the second hybrid does. Repeating the comparison around a complete answer cycle then distinguishes query-specific following from a donor that simply favors one answer everywhere. Let $s_d(q, a)$ be the sum of answer-content token log probabilities under H_d, excluding EOS, and let a_{0i}, a_{1i} be the two donor assignments to query q_i. The contrast is

$$\delta_i = [s_0(q_i, a_{0i}) - s_0(q_i, a_{1i})] - [s_1(q_i, a_{0i}) - s_1(q_i, a_{1i})]. \tag{7}$$

Positive values indicate that answer preference follows the historical source of the transplanted parameters. A donor simply favoring one answer across all queries should not produce a positive cycle sum. Under the restricted additive account $s_0(q, a) - s_1(q, a) = u(q) + v(a)$, each block's cycle sum is zero: query offsets cancel within a contrast, and answer offsets cancel across the two permutations. Positive observed cycle sums distinguish the effect from a query-independent answer-prior shift. Query-dependent interactions with the recipient remain compatible with this test.

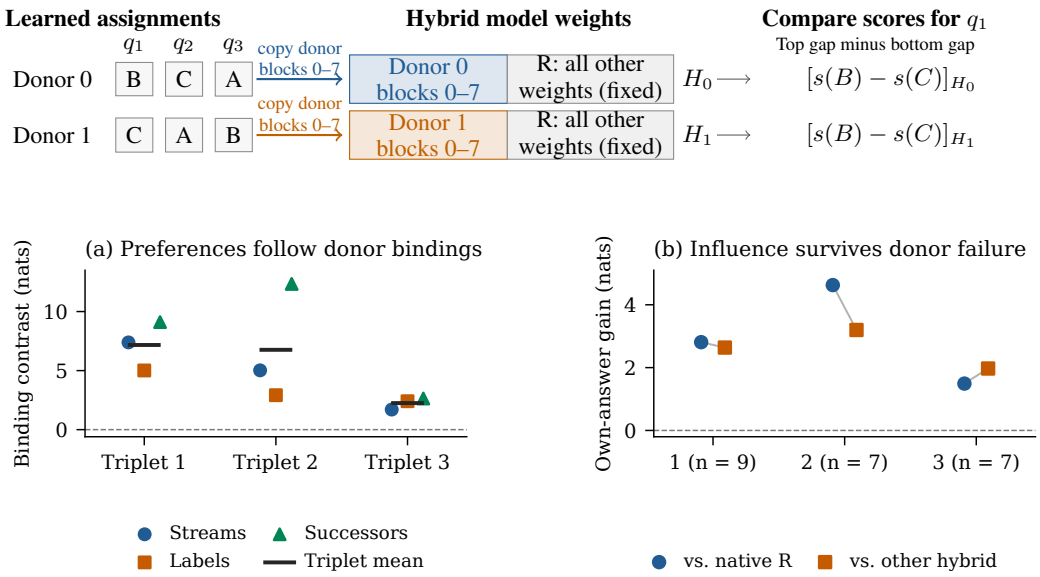

Figure 8: Cyclic assignment and first-panel binding contrasts. Top: different donors supply blocks 0–7 while all other R weights stay fixed. Compare the B-versus-C score gap for q_1 across hybrids, then repeat around the cycle. (a) Mean contrasts within complete three-query cycles, with triplet means in black. (b) Lost-binding gains relative to native R and the other donor's transplant; subset sizes are 9, 7, and 7. Content scores exclude EOS.

Table 7: Native-parameter binding influence in three fresh training triplets. Content scores are sums of log probabilities (nats), excluding EOS. The final column weights triplets equally. Lost subsets are defined by the donor's own post-acquisition correctness and R_0 failure.

Quantity	**01**	**02**	**03**	**Mean**
Mean content binding contrast	7.169	6.759	2.249	5.392
First-token component	5.673	5.626	2.935	4.745
Continuation component	1.496	1.133	−.686	.648
Lost own-answer gain over native R	2.810	4.627	1.491	2.976
Lost own-answer gain over other donor	2.638	3.198	1.971	2.602
Lost donor–query instances	9	7	7	—
Paired complete-answer following	0/9	0/9	0/9	—

All three block cycle sums are positive in each triplet. Individual content contrasts are positive for $9/9$, $9/9$, and $5/9$ queries. The first-token component is measured before supplying any answer prefix; continuation scores condition on the preceding answer tokens. Among the 23 donor–query instances acquired and then lost, own-donor transplants raise assigned-answer scores above native R in 19 cases. These strata describe realized acquisition and loss rather than randomized acquisition effects.

Generated outputs provide a stricter endpoint: donor-assigned answers appear in $1/18$, $1/18$, and $0/18$ hybrid evaluations, with no case in which both hybrids generate their respective donor assignments. One of the two hits is a newly accessed lost binding; the other was retained by its donor. Across the six hybrids, mean absolute nonfocal margin change is 1.439–1.838 over 92 queries, with both exact gains and losses. Thus the replaced blocks shift relative answer scores in a query-specific way, while rarely restoring exact generation. Because R learned the same answer pools, it can supply content used by history-dependent routing. These data do not separate that route from autonomous content storage in the transplanted blocks.

H.2 NATIVE HISTORICAL BINDINGS CHANGE ACTUAL Q RESPONSE

The static preference shift leaves open whether historical parameters also change learning. We apply identical real updates to native and transplanted states, then measure how much each update strengthens or weakens the recipient's preference for its donor's assignments.

Matched states and real updates. The follow-up uses the same three triplets, nine queries, and native/hybrid states R, H_0, and H_1 from Appendix H.1. Each state receives all four target-disjoint cue-by-label arms with a fresh rank-one all-24-QKV adapter (alpha 2, zero output-factor initialization). Both adapter factors are trained for ten actual AdamW steps, with learning rate .001, zero weight decay, clipping at 1, and a complete token-normalized support pass per step. Every arm starts from its own unmodified native or hybrid state; its trained adapter is evaluated on that same state. The source native parameters remain unchanged. R-to-R self-copy and repeated no-Q measurements reproduce the original scores.

The first two paired initializations were fixed before the new Q outcomes. After inspecting their initialization sensitivity, four further seeds were fixed and run for every state and arm, without replacing either original seed. These stages comprise 72 and 144 Q trainings, respectively (2,160 optimizer steps in total). They supply new update measurements on previously inspected histories, not new training triplets. The four new initializations are reported separately before combining all six equally. The triplet remains the replication unit, conditional on the shared query inventory.

Response beyond the initial preference. We subtract the no-update donor preference to isolate its change under later training. A positive change means that training strengthens donor following; a negative change means that it weakens it. If both hybrids keep exactly their initial preference contrast, the change is zero even when the original donor influence is large. This subtraction therefore asks what the new training does to historical influence, rather than measuring that influence a second time. Apply Equation 7 to each post-Q hybrid pair to obtain $\delta_{i,q,r}$, where q is the support arm and r the initialization. Let $\delta_{i,0}$ be the no-Q contrast. Define

$$K_{i,q,r} = \delta_{i,q,r} - \delta_{i,0}, \qquad \Theta_i = \overline{K}_{i,OC} - \overline{K}_{i,OW} - \overline{K}_{i,FC} + \overline{K}_{i,FW}. \tag{8}$$

The bar averages the fixed initializations. Scores are sums of content-token log probabilities, excluding EOS. Positive K means that Q amplifies donor following; negative K means it attenuates it. Static additive carry-through of the no-Q preference predicts $K = 0$. Within a complete three-query cycle, the same cancellation as in Equation 7 removes a restricted query-independent answer shift in the Q response. Query-dependent routing or compatibility with the shared recipient remains a possible implementation.

The original analysis requires valid free generation for each hybrid–query evaluation, so some complete averages are unavailable even when teacher-forced scores are finite. We first report that analysis, then a post-hoc sensitivity using all finite scores. The second query panel fixes the all-finite rule before its outcomes. Both analyses use the content-sum contrasts above; the later geometry comparison uses mean answer-plus-EOS scores to match its gradient measurement.

Six invalid hybrid evaluations block the affected complete means: two empty generations and four nonterminating generations at the 24-token cap. All occur in triplets 2–3. Finite teacher-forced scores are retained in the raw records but do not replace these invalid measurements. Complete relation cycles remain available even when the nine-query mean is undefined.

Later training depends on both the history and the update. In triplet 1, all four nine-query changes are negative under every initialization. In triplet 2, old-wrong is negative and fresh-correct positive under every initialization, including the four new seeds. The effect varies within a triplet as well: triplet 1's six-seed old-wrong change is -2.431 for record labels and $+1.117$ for successors; in triplet 3, complete six-seed means are available for all four arms on record labels, with vector $(-.244, -.306, +.643, -.936)$. Thus the initial binding preference is modified by real learning, with direction determined by the history, relation block, and support. The four-arm interaction is more sensitive to initialization: triplet 1's nine-query Θ is $-.339$ over the first two seeds, $+.173$ over the new four, and $+.002$ over all six. Missing cells leave complete six-seed Θ undefined in triplets 2–3.

Table 8: Change in donor following after actual Q, averaged over all nine queries. Values are content-log-probability contrasts in nats. A dash means a required contrast is invalid; the mean is not recomputed on the remaining queries. The last column counts queries valid across all six initializations.

Triplet	Arm	First two	New four	All six	Valid queries
1	OC	−1.496	−1.041	−1.192	9/9
1	OW	−0.571	−0.475	−0.507	9/9
1	FC	−1.031	−1.485	−1.334	9/9
1	FW	−0.445	−0.747	−0.646	9/9
2	OC	+0.141	—	—	8/9
2	OW	−0.959	−1.465	−1.297	9/9
2	FC	+0.739	+0.615	+0.656	9/9
2	FW	+0.320	+0.113	+0.182	9/9
3	OC	—	+0.559	—	8/9
3	OW	−0.292	—	—	8/9
3	FC	+0.380	—	—	8/9
3	FW	—	—	—	8/9

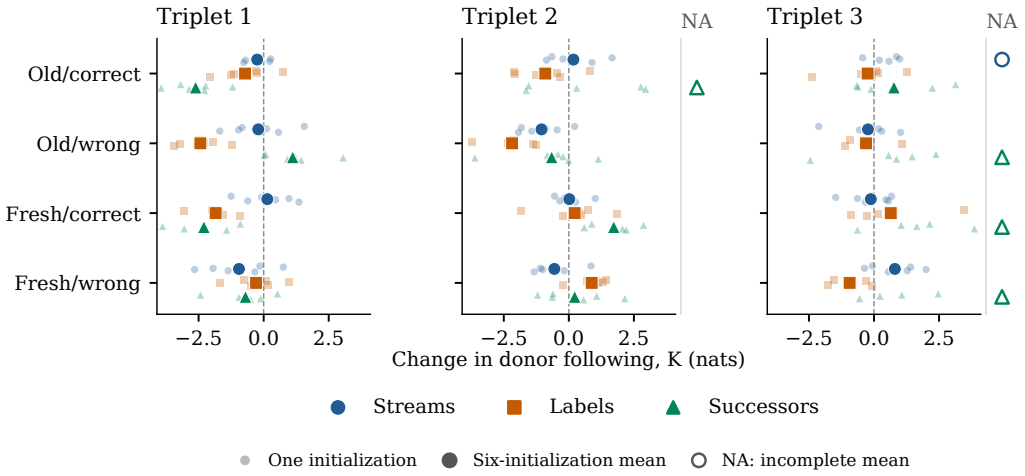

Figure 9: Native binding history changes finite Q response with signed heterogeneity. Each color/shape is one complete three-query relation cycle; small points show three-query means for individual initializations and large points their six-seed mean. Open symbols in the separate NA column mark unavailable complete means and retain the relation-cycle shapes. Zero denotes unchanged donor following after Q, not absence of the initial binding preference. Seeds and relation blocks are repeated measurements within a triplet, not independent training histories.

Sensitivity to generation-based exclusions. The table above retains the original generation-gated analysis. A separate post-hoc analysis uses every finite teacher-forced candidate score, including scores from evaluations with invalid free generations. This separates the continuous preference endpoint from the ability to produce a complete answer. The complete six-seed Θ values are then $+.002$, $+.733$, and $-.273$ nats (Table 9). On the matched mean-answer-plus-EOS endpoint, angular correlations are .277, $-.185$, and .762 across all nine queries in each triplet. Thus removing the generation filter does not produce a common signed interaction across the original histories.

A second query panel with fresh training histories. To test whether the binding influence extends beyond the original queries, we fix a second panel of nine queries and train three additional R/D0/D1 triplets. The new queries form three answer cycles for *has part*, *instance of*, and *part of*. Selection uses relation, answer-token length, prompt-token matching, and literal exclusion from Q support,

not acquisition, damage, or transplant outcomes. These are new focal binding tests drawn from the previously used canonical acquisition inventory. All original acquisition rows are retained. Within each triplet, the prompt and answer multisets, row orders, per-step token denominators, and training RNG are matched across histories. As before, only the nine bindings differ, and native blocks 0–7 are transplanted into the common recipient.

Each native or hybrid state receives all four real ten-step Q updates with two paired adapter initializations: 72 Q trainings across the three new triplets. Before their outcomes, we fix all finite teacher-forced scores as eligible for the continuous endpoint and retain generation validity and exact outputs separately. All triplets and cells enter the analysis.

The static content binding contrasts are 11.837, 11.571, and 12.920 nats, positive for every query. Among the 6, 3, and 8 donor–query instances acquired and then lost, the own-donor transplant raises the assigned answer's score above native R in all 17 cases; triplet means are 6.027, 9.367, and 8.187 nats. This replicates graded binding influence after loss of direct expression. At the exact-generation endpoint, none of the no-Q hybrids generates its donor's assigned answer ($0/18$ in each triplet; paired following $0/9$).

Table 9: Finite-score Q responses in the original and new query panels. Content-sum K and Θ are in nats; angular r compares geometry with the matched mean-answer-plus-EOS response. Original-panel rows are a post-hoc sensitivity analysis over six initializations; new-panel rows use the eligibility rule fixed before outcomes and two initializations. K and Θ average all nine queries; r is computed across those queries within each triplet.

Panel	Triplet	K_{OC}	K_{OW}	K_{FC}	K_{FW}	Θ	Angular r
Original	1	-1.192	-0.507	-1.334	-0.646	$+0.002$	$+0.277$
Original	2	-0.090	-1.297	$+0.656$	$+0.182$	$+0.733$	-0.185
Original	3	$+0.285$	$+0.131$	$+0.752$	$+0.324$	-0.273	$+0.762$
New	1	-0.273	-3.699	$+0.304$	-2.128	$+0.994$	$+0.333$
New	2	$+0.193$	-1.706	$+0.182$	-0.894	$+0.822$	$+0.790$
New	3	-1.076	-2.451	-0.418	-1.458	$+0.335$	-0.260

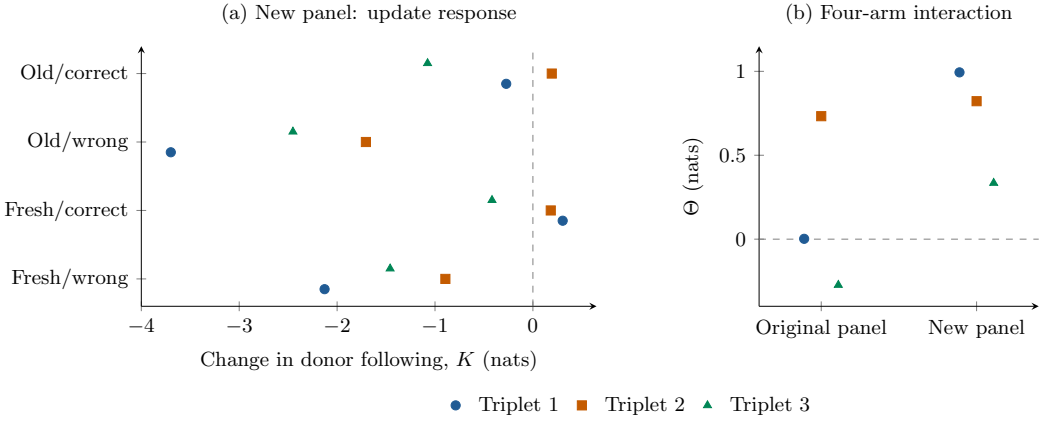

Figure 10: Historical binding influence changes under further learning on the new panel. (a) Wrong-label updates attenuate donor following in all three new triplets, whereas correct-label effects differ by history. (b) The four-arm interaction is positive in the new panel but not uniformly positive in the original panel. Triplet numbers are local to each panel. Points are triplet means, not independent query-level replications. Both panels use finite candidate scores here; their different queries and initialization counts are detailed in Table 9.

The new panel also shows update-dependent changes beyond the static preference (Figure 10a). Both wrong-label arms attenuate donor following in every triplet; correct-label arms produce smaller, mixed-signed changes. Their four-arm interactions are $+.994$, $+.822$, and $+.335$ nats. Figure 10b

compares these interactions with the mixed-signed original panel. Across both panels, the intervention establishes binding-sensitive parameter influence and changes in finite learning response.

Relating hybrid geometry to its later response. To test whether hybrid geometry tracks later response, we measure each hybrid's target-answer and support-descent gradients before training in the virtual all-QKV output-factor coordinates. We take the donor-by-answer contrast of the four-arm cosine interaction and compare it with the corresponding finite Θ computed on *mean answer-plus-EOS* scores. We use the same score definition on both sides of this comparison: the target-score gradient and the finite response both include EOS and average over tokens. This scalar differs from the content-sum K above, which measures total log-probability preference over answer content. The geometry contrast is induced by the native intervention; it is not the original randomized-exposure base-to-R_0 ΔA. In the original panel, descriptive correlations of six-seed query means are .277, $-.202$, and .745 over 9, 8, and 7 complete query pairs; the four new seeds alone give .237, $-.159$, and .757 over 9, 8, and 8 pairs. Raw-dot-product diagnostics give six-seed correlations .496, .186, and .610. In the second query panel, angular correlations are .333, .790, and $-.260$; raw-dot correlations are .298, .773, and $-.270$. These within-panel diagnostics show heterogeneous correspondence, rather than the held-out prediction evaluated in Appendix C.

Answer gains and exact outputs. Relative preference can change through an assigned-answer gain or alternative suppression. For triplet 2 under fresh-correct Q, the six-seed mean own-assigned content score rises by .100 nats across the 18 hybrid–query instances, while the other donor's answer falls by .228 nats. The no-Q hybrids generate one assigned answer among these 18 instances; the six Q initializations generate $1, 1, 1, 1, 0, 0$. In triplet 3, the four new old-correct initializations raise own-answer scores by .340 nats on average and the alternative by .061; assigned-answer hits change from zero to $1, 0, 0, 1$. Across all arms, new exact hits on originally lost donor bindings occur in three evaluations in triplet 2 and seven in triplet 3, concentrated on one query in each. These counts repeat queries across arms and initializations. Nonfocal effects accompany these shifts. For example, in triplet 2 under old-wrong Q, H0 loses 2–8 of its 36 initially correct nonfocal answers and H1 loses 5–9 of 35 across initializations, with no nonfocal restorations.

H.3 Exposure-trained prediction of transplant-induced response

To test whether a predictor learned from exposure experiments can anticipate how replacing parameters changes later learning, we fit it on exposure histories and apply it unchanged to new recipient–donor triplets. This gives the exposure and transplant experiments a shared quantitative readout of subsequent learning.

A common response target. To connect the experiments, both must predict the same quantity: how much a given update changes an answer's score. We then compare that change between each hybrid and its unmodified recipient. For query i, candidate answer ℓ, state z, and support arm q, let $\bar{s}_z(i, \ell)$ denote mean answer-plus-EOS log probability. Define the two-initialization-averaged ten-step response and its transplant effect as

$$\begin{aligned} y_z(i, \ell, q) &= \mathbb{E}_r\left[\bar{s}_{z+Q^{10}_{q,r}}(i, \ell)\right] - \bar{s}_z(i, \ell), \\ D_z(i, \ell, q) &= y_z(i, \ell, q) - y_R(i, \ell, q), \qquad z \in \{H_0, H_1\}. \end{aligned} \tag{9}$$

The primary target D_z measures how replacing native blocks changes the response to the same later update. We predict each elementary response y_z from its pre-update measurements, then take the same hybrid-minus-recipient difference. This per-arm score target supplies a common endpoint for the two experiments; Section 4.2 uses a separately fitted predictor of the four-arm margin interaction. Predicting zero means that the hybrid and recipient respond equally to that update, even if the update changes both of their scores substantially.

Source fitting and frozen inputs. We train only on exposure histories and freeze the predictor before observing confirmation transplant outcomes. The source consists of seven Pythia-410M exposure discovery histories and 85 facts after excluding transplant-panel query identities. Averaging the two adapter initializations gives 2,380 fact–history–arm rows. We fit ridge regression with penalty 1 and feature standardization estimated on this source alone. Every model includes the

four arm indicators. The state baseline adds the starting mean answer-plus-EOS score and log target- and support-gradient norms. Geometry models use the cosine alignment or raw inner product between target-score and support-descent gradients in the same virtual rank-one all-QKV output-factor coordinates used for both experiments. Each geometric quantity is evaluated before Q and averaged over the two initializations. We compare arm-only, angular, raw, state, state-plus-angular, and state-plus-raw models.

Development analyses on previously measured transplant histories informed this test. All six coefficient sets, source scaling, targets, and the three-triplet protocol were fixed before the confirmation histories were trained. Predictions use the exposure-fitted coefficients unchanged and are saved before the first Q update in each triplet.

Confirmation histories. Three new Pythia-410M recipient–donor triplets reuse a nine-query panel from prediction development, with independent order, binding, and training seeds. This third panel comprises three answer cycles for *has part*, *replaced by*, and *said to be the same as*. The recipient and donors share answer pools and otherwise matched acquisition and interference schedules, with 160 acquisition steps and 228 steps in each interference phase. Donor blocks 0–7 form the two hybrids with the recipient complement fixed. Each of R, H_0, and H_1 receives all four target-disjoint Q10 arms with two paired adapter initializations, yielding 72 complete update runs. Geometry, initial scores, and predictions are recorded before these updates.

All nine queries, their three candidate answers, both hybrids, and all four arms enter the primary analysis: 648 response contrasts, averaged over initializations. Continuous endpoints use every finite teacher-forced score; generation validity and exact outputs are recorded separately. The three training triplets are the replication units. Thus the confirmation tests new learning histories on a reused query inventory.

Prediction transfers across parameter interventions. A predictor that has seen no confirmation transplant outcomes predicts part of the replacement-induced change in future update response. Table 10 reports primary-target MSE. The state-plus-angular predictor reduces pooled error by 9.69% relative to zero transplant effect, with reductions of 14.85%, 6.25%, and 6.89% in the three triplets. Angular geometry alone reduces pooled error by 3.67%. Adding it to scores and norms reduces pooled error by 3.36%; the corresponding triplet changes are 8.65%, 1.20%, and -1.05%. The complete exposure-trained predictor is informative on transplanted states; alignment supplies an additional pooled gain beyond starting scores and norms, with that increment varying across triplets.

Table 10: Frozen exposure-to-transplant prediction of D_z. Values are mean-squared errors of mean-answer-plus-EOS response differences. Each triplet contributes 216 contrasts; pooled MSE weights the three triplets equally. Arm-only predictions cancel in the hybrid-minus-recipient contrast, giving the zero-effect baseline.

Exposure-fitted predictor	**01**	**02**	**03**	**Pooled**
Zero effect / arm-only	.353330	.308792	.272304	.311475
Angular geometry	.324737	.303202	.272229	.300056
Raw geometry	.321748	.294034	.276561	.297448
Scores and norms	.329367	.292998	.250895	.291087
Scores, norms, and angular geometry	.300862	.289488	.253530	.281294
Scores, norms, and raw geometry	.301468	.282238	.258427	.280711

The primary evaluation rule fixed before outcomes required lower MSE than zero effect in all three triplets for the combined angular model, together with a pooled angular-only improvement and improvement in at least two triplets. The observed results meet this rule. The combined model's pooled prediction–response correlation is .264 and sign agreement is 59.3%. These are repeated-measure summaries over three independent triplets.

Response targets resolve the scope of transfer. The secondary targets use the same mean answer-plus-EOS score as the primary prediction target. We apply the contrast algebra of Equation 8 to these scores, denoting the resulting donor-by-answer contrast by K and its four-arm interaction by Θ here. These are the mean-score versions of the earlier content-sum contrasts: the subtractions are

the same, but the scores being subtracted have a different token span and normalization. Table 11 reports these prespecified secondary targets and the elementary response y_z. The combined angular readout reduces error on K by 4.94% relative to the zero-prediction baseline, while scores and norms alone give a 6.56% reduction. For Θ, its error is 15.71% above the zero-prediction baseline; elementary-response error is 4.22% above that baseline.

Table 11: Secondary endpoints under the same frozen predictors. Entries are pooled MSE. Elementary response is the score change under one update; binding K is the update's change in donor preference; interaction Θ compares that change across the four cue-by-label arms. All contrasts use mean answer-plus-EOS scores; the content-sum binding results in Appendix H.2 have a different scale. The endpoint row counts are repeated measurements within three triplets.

Predictor	**Elementary response**	**Binding K**	**Interaction Θ**
Zero	.257919	.518046	1.131244
Arm-only	.291158	.518046	1.131244
Angular geometry	.279623	.524839	1.313109
Raw geometry	.281984	.538887	1.581289
Scores and norms	.278115	.484058	1.131244
Scores, norms, and angular geometry	.268791	.492441	1.308958
Scores, norms, and raw geometry	.271460	.505812	1.564060
Measurements	972	108	27

I PRE-UPDATE REGRESSION LOCALIZATION

The earlier experiments predict update-specific responses. Here we ask whether that information helps rank which existing facts a proposed update will damage when only a small number can be checked. The decision is which facts to inspect, with the proposed update held fixed. We evaluate whether adding its pre-update gradient interaction directs those checks toward more of the facts that later suffer score damage.

Task and learning histories. For a fixed proposed update, an audit selects k of 48 protected facts to inspect after training. The target is a score-damage event: the canonical answer's mean token log probability, including EOS, decreases by more than 0.1 after three optimizer steps. An event is a fact–history–recipe instance; the same fact may recur under several recipes.

Each Pythia-410M history trains three factual banks, A, B, and C. Two reused fact panels each supply one A–B–C and one A–C–B history, with complementary randomized A exposure within each pair. At the post-training state, 16 facts from each bank form the 48-fact protected set. Their identities are fixed by hash before training, without acquisition or damage screening. The eight incoming correction facts are divided into two disjoint four-fact subsets. Each subset defines four recipes: correction only, or a 50:50 mixture of the separately token-normalized correction and A, B, or C replay losses. Replay uses three successive four-fact packets. All eight recipes start from the same checkpoint, rank-one QKV adapter initialization, optimizer state and RNG, and execute three Adam steps at learning rate 0.001.

Development and frozen confirmation. Four development histories informed the choice to predict damage probability instead of signed mean response. Their outcomes had already been inspected; the risk-model evaluation on these histories is exploratory. We then froze the feature families, source-only fits, threshold, budgets and analysis before training four new histories. These confirmation histories reuse the two fact inventories and prompt families, with new training seeds and predetermined protected sets and correction splits. All four histories and all eight recipes per history enter each cohort's analysis. No history is excluded by damage or performance. Development and confirmation are reported separately throughout.

Risk model and matched inputs. The key comparison is whether the target–support gradient interaction improves damage prediction beyond current scores, update configuration, and gradient magnitudes. We hold the prediction model and those other inputs fixed when adding the interaction. Every risk model is standardized L2-regularized logistic regression ($C = 0.1$, `lbfgs`, 2,000-iteration limit, no class weighting), fitted to eligible subsets of 1,536 three-step fact–history–recipe observations: 64 source facts per history across four earlier histories and six source recipes. The 48-fact protected set is used for evaluation. Fitting is restricted to the same factual bank. For an A fact, training excludes all source rows with that fact's identity. For B and C, a fixed hash assigns each fact identity to one of two groups; prediction for a fact uses source rows from the other group in the same bank. Means, scales and coefficients use only these eligible source observations. The confirmation predictor has no access to confirmation outcomes or target-set calibration.

The shared state/recipe inputs (abbreviated SI only in the tables) are the current gold-answer score and margin, pretrained gold-answer score, four loss-mixture weights, membership in the initial support packets, and state-by-weight products. The norm model adds the log target-gradient norm and log mixed-support-descent norm. The raw model adds one scalar to exactly those inputs: the inner product between the direction that raises a protected fact's score and the proposed update's descent direction. A negative value signals a local decrease in that score. In the virtual zero-initialized, rank-one all-QKV output-factor coordinates used in Appendix C, let g_i be the protected-answer score gradient and v_b the descent gradient of support bank b. For recipe weights w_b, its interaction is

$$x_{iq}^{\mathrm{raw}} = \frac{1}{2}\sum_{r=1}^{2}\left\langle g_i^{(r)}, \sum_b w_{qb} v_b^{(r)} \right\rangle . \tag{10}$$

The two probe initializations are fixed. The two log-norm features and the alternative cosine feature are likewise averaged over initializations. The gradients are measured at the native checkpoint before training, using the initial support packets. They describe the proposed loss mixture in the

probe coordinates; the prediction target is the response of the actual three-step Adam update. This application uses a per-update risk model, whereas the original tangent experiment predicts the four-arm finite interaction.

All 48 facts are ranked globally by predicted damage probability, with fact-identity tie breaking and no bank quotas. Predicted probabilities and complete orderings for every recipe are recorded before the first optimizer step. The primary budget is $k = 10$; $k = 5, 15, 20$ are fixed secondary budgets. Recall within a history pools caught and total events across its eight recipes. The primary contrast averages the four within-history raw-minus-norm recall differences equally. Pooled counts describe the events; the four histories, spanning two panels, are the replication units. No history has zero events.

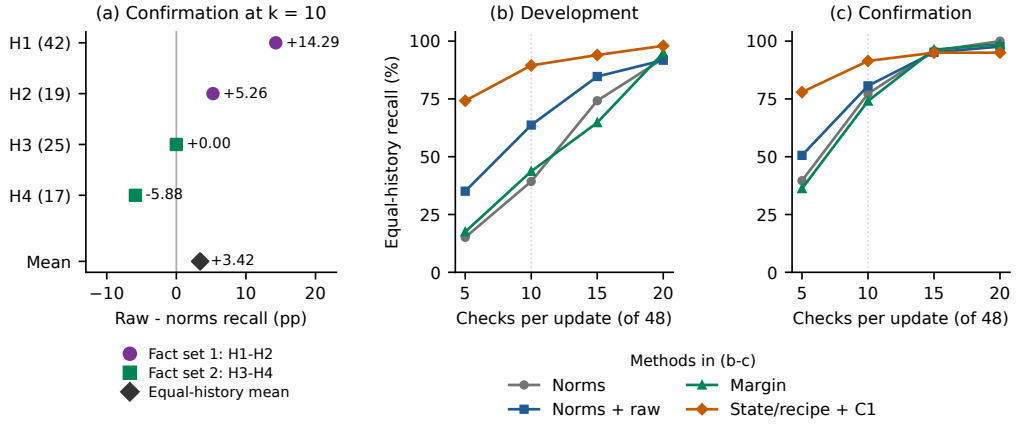

Figure 11: Pre-update interactions improve localization at tight audit budgets. (a) The four confirmation histories' raw-minus-norm Recall@10 differences; parentheses give score-damage event counts. H1–H2 share the first fact panel, H3–H4 the second. The diamond weights histories equally. (b–c) Development and confirmation budget curves, with equal-history recall. The raw increment varies across histories and changes sign at larger budgets. C_1 uses an observed first-step response and is shown separately from the pre-update inputs.

Detection at the primary ten-check budget. At the primary budget, we compare detection with and without the raw interaction. In development, norms and raw interactions capture 46 and 64 of 105 events at $k = 10$ (Table 12). Equal-history recall rises from 39.31% to 63.69%. In confirmation, they capture 77 and 83 of 103 events, with equal-history recall of 77.24% and 80.66%. The four paired differences are $+14.29$, $+5.26$, 0, and -5.88 percentage points; the two panel means are $+9.77$ and -2.94 points. Thus the positive mean is heterogeneous across the tested histories and panels.

The norm model misses 59 development events and 26 confirmation events. Adding raw interaction reduces those counts to 41 and 20: net reductions of 30.5% and 23.1% of the respective baseline misses. At the fixed 320 audit slots per cohort, confirmation precision rises from 24.06% to 25.94%, and selected non-events fall from 243 to 237. These precision and miss-rate changes describe the same six additional detections.

Budgets and update matching. Table 13 reports the full risk-model roster. In confirmation, raw interactions add 11, 6, 1, and -4 captured events over norms at budgets 5, 10, 15, and 20. The equal-history recall differences are $+10.97$, $+3.42$, -0.68, and -2.38 percentage points: the pooled extra event at $k = 15$ does not give a positive equal-history difference. The raw feature therefore helps most when the audit budget is tight.

If the signal reflects a particular update rather than generic vulnerability, using a different update's readout should impair ranking. The prespecified cross-half control uses the other correction subset's pre-update readout while keeping the replay-bank choice and evaluated update fixed. Raw capture falls from 83 to 71 at $k = 10$; norm capture remains 77. This comparison tests matching the readout to the incoming update beyond a common vulnerability ranking. The raw model's confirmation Brier

Table 12: All eight localization histories, separated by evidence status. Counts are events captured with ten checks per recipe (80 per history). Δ is the paired raw-minus-norm recall difference. SI denotes the shared state/recipe inputs; C_1 is a paid first-step score observation.

History	Events	Norms	+ Raw	Margin	SI+C_1	Δ recall (pp)
Development 1	39	28	28	21	35	$+0.00$
Development 2	20	7	14	4	19	$+35.00$
Development 3	32	7	11	14	28	$+12.50$
Development 4	14	4	11	8	12	$+50.00$
Confirmation 1	42	27	33	29	35	$+14.29$
Confirmation 2	19	16	17	18	19	$+5.26$
Confirmation 3	25	21	21	14	25	$+0.00$
Confirmation 4	17	13	12	13	14	-5.88

score is 0.05667, compared with 0.05899 for norms. Both prediction and allocation are evaluated on the same fixed score-damage endpoint.

Additional history features and a donor-based readout. We also test whether the observed learning trajectory or a donor-based readout adds information to this ranking. These are additional inputs to the audit; the proposed update and the facts available for checking remain fixed. The trajectory vector contains four quantities: the margin and answer score immediately after training on bank A, and each of those values minus its current value after the full training sequence. We also include the product of each quantity with each of the four update-mixture weights. This vector is available only for A; B/C receive the corresponding model without that vector. Adding it to the raw model changes captured events from 64 to 62 in development and from 83 to 74 in confirmation at $k = 10$.

A separate development-only readout transplants native blocks 0–7 from a predetermined paired donor and measures the protected score shift before restoring the recipient. It increases state/recipe-only capture from 37 to 45, but changes the stronger raw-plus-history model only from 62 to 63. This measured readout mixes donor history and seed differences; it differs from the controlled cyclic-binding intervention in Section 5. It was not added to the frozen confirmation roster. The earlier signed-mean predictor captures 37 and 53 development events with norm and raw features, respectively; its comparison with the risk-target model remains development analysis.

Observed first-step response and exact-answer losses. To measure what pre-update geometry adds once an actual training step is observed, paid-probe models add the step-one score change C_1 and its recipe-weight products, saving predictions before step two. The state/recipe+C_1 model captures 93 of 103 confirmation score-damage events at $k = 10$. Adding norms leaves this at 93; adding raw interaction to the matched norms+C_1 model reduces it to 87. At $k = 20$, the corresponding counts are 99 and 100, so the raw feature supplies no consistent additional gain once this response observation is available. The development comparison of 94 versus 100 adds norms and raw together; the norms-only C_1 control was first included in confirmation.

Exact-generation loss is recorded separately from score degradation. There are 17 development and 22 confirmation correct-to-incorrect events. Norms/raw capture 2/1 of the development events and 9/12 of the confirmation events at $k = 10$. The two matched C_1 models each capture 15 confirmation exact losses; uncalibrated C_1 ranking captures 17. These secondary counts retain the distinction between a score-risk audit and an exact-answer-loss detector.

Measured acquisition and audit cost. Table 14 charges current-score acquisition, QKV readouts where used, prediction, and the selected post-update score audit. The paid C_1 alternatives additionally charge an actual first optimizer step and scores for all 48 protected facts. Cold costs charge each recipe for its complete readout. Reused costs divide each history's baseline acquisition across eight recipes and each correction-half QKV readout across its four recipes; actual C_1 observations remain recipe-specific. Prediction charges use the measured group runtime, conservatively shared by each reported family.

All timings are measured for these score-based audits; the common three-step training and separate exact-generation diagnostics are outside these columns. The norm and raw methods use the same complete QKV collector, so these timings do not estimate a separately optimized norm-only implementation. At $k = 10$, raw acquisition plus audit averages 5.614 seconds cold and 1.417 seconds with reuse, versus 0.974 and 0.652 for state/recipe+C_1. Even auditing all 48 scores is cheaper here than acquiring QKV readouts.

The confirmed benefit is more informative allocation of a tight checking budget: pre-update sensitivity identifies additional facts damaged by the specified update. In this small score-audit setting, acquiring that signal costs more than the checks it saves. A practical extension would need cheaper readouts or evaluations whose cost justifies selective checking.

Table 13: Score-damage events captured across fixed audit budgets. A dash marks a feature family absent from that cohort's frozen roster. Development and confirmation have 105 and 103 events, respectively. Raw and cosine are added to the same norm features; history is available only for A.

	Development / 105 events				Confirmation / 103 events			
Inputs / ranking	5	10	15	20	5	10	15	20
Current margin	18	47	66	97	39	74	98	102
Seeded random	—	—	—	—	8	17	28	35
State/recipe (SI)	20	37	66	83	45	74	91	98
SI + history	20	39	65	83	38	65	90	96
SI + norms	21	46	77	97	41	77	96	103
SI + norms + history	20	43	75	94	—	—	—	—
SI + norms + raw	37	64	88	95	52	83	97	99
SI + norms + raw + history	34	62	87	92	46	74	92	96
SI + norms + cosine	30	62	86	98	45	81	94	102
SI + norms + cosine + history	28	59	84	94	—	—	—	—
SI + C_1	79	94	99	103	76	93	99	99
SI + history + C_1	76	92	99	103	—	—	—	—
SI + norms + C_1	—	—	—	—	76	93	99	99
SI + norms + raw + C_1	80	100	101	103	73	87	98	100
SI + norms + raw + history + C_1	75	97	101	103	—	—	—	—
Uncalibrated C_1	85	96	100	100	72	89	95	95
Raw, swapped incoming half	—	—	—	—	46	71	86	96

Table 14: Confirmation: detections at ten checks per recipe and measured mean seconds per recipe, averaged over all 32 updates. Score and exact-loss columns have different event definitions. Full audit measures all 48 facts. The reused column applies the sharing rules described in the text.

Inputs	Score / 103	Exact / 22	Cold (s)	Reused (s)
Current margin	74	4	0.447	0.125
State/recipe (SI)	74	9	0.453	0.131
SI + norms	77	9	5.611	1.414
SI + norms + raw	83	12	5.614	1.417
SI + norms + cosine	81	11	5.612	1.415
SI + C_1	93	15	0.974	0.652
SI + norms + C_1	93	15	6.141	1.944
SI + norms + raw + C_1	87	15	6.148	1.951
Full 48-fact score audit	103	22	0.744	0.422