# BEHAVIORAL FORGETTING DOES NOT NECESSARILY ERASE CAUSAL ACCESSIBILITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

When a network's output on an input changes over training, it is natural to describe the network as having "forgotten" whatever computation previously produced the old output. This description conflates two distinct properties. We demonstrate that behavioral state, causal accessibility, and behavioral recoverability are empirically dissociable: a network can produce behavior fully consistent with having overwritten an association while retaining, in an intervention-accessible form, the causal capacity to reinstate the original output. This is a claim about what the interventions establish – that a historical direction in hidden-state space remains causally load-bearing – not a claim that a specific "old computation" or "circuit" has survived. Controls rule out generic perturbation sensitivity, lineage-independence, diffuseness, construction-arbitrariness, and pure fc2-geometry explanations. Upstream analyses (fc1 weight-space substrate; natural indirect effect decomposition) confirm the persistence is genuine upstream-of-readout structure. The dissociation holds across 10 model configurations including Transformers. Regarding mechanism: the evidence is more consistent with partially overlapping representational resources than with two cleanly independent pathways – but this interpretation is not uniquely established by our interventions, and we make no claim about the identity of the surviving internal structure.

## 1 INTRODUCTION

"The network forgot how to do X" implicitly asserts something about mechanism: that whatever internal computation previously produced X is no longer present. This assertion is rarely tested directly. Standard evaluation measures only the network's current output; it does not distinguish a network that has genuinely lost the capacity to compute X from one that still has accessible internal structure, but whose output is currently dominated by a competing computation.

**The core logical chain.** We establish five things: (1) **behavior flips** within 10 steps; (2) **the historical direction $J_A$ remains causally accessible**, continuing to substantially change the model's output for 3000 B-training steps; (3) **the historical component restores A-behavior** when injected (7/7 seeds; orthogonal remainder: 0/7); (4) **controls establish specificity** – direction-specific, lineage-specific ($p < 2.4 \times 10^{-8}$), not fc2 geometry; (5) **upstream experiments establish genuine upstream structure** via a preregistered test with no readout and no `zor` (AUC$= 1.000$; causal steering $p = 10^{-6}$). These constitute the headline finding: *behavioral forgetting does not imply causal erasure.* All further analyses strengthen the interpretation but are not required for the headline claim.

We do not conclude that an "old circuit" has survived – only that causal accessibility of a historical direction has survived and is genuine upstream structure. The distinction between *causal accessibility*, *intervention-defined recoverability*, and *spontaneous computational retention* (untested) is made explicit in Section 3.

**Contributions. Core chain:** (1) Three-observable formalization with explicit falsification conditions. (2) Ablate-rescue protocol establishing necessity and sufficiency of the historical $J_A$-parallel component. (3) Preregistered upstream experiment ruling out fc2-geometry as explanation (AUC$= 1.000$; causal steering $p = 10^{-6}$; cross-model persistence $\cos = 0.910$). **Further evidence:** (4) $J_A$ is load-bearing in natural B-behavior forward passes (88% of margin). (5) Lineage-specificity, held-out generalization (140/140), 10-configuration persistence including Transformers.



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.

(6) Mechanistic discrimination: an orthogonal basis decomposition shows $e_A$ and $e_B$ are functionally separable with independent decoders, while the shared component $e_{\text{sh}}$ is causally inert – substantially supporting the partially overlapping resources interpretation.

## 2 Related Work

**Catastrophic forgetting and continual learning.** A long line of work documents that neural networks trained sequentially on different tasks lose performance on earlier tasks (McCloskey and Cohen, 1989; French, 1999), and proposes methods to mitigate this loss by penalizing movement in parameters important for earlier tasks (Kirkpatrick et al., 2017) or replaying earlier data (Rebuffi et al., 2017). This literature is almost entirely behavioral: forgetting is defined and measured as a drop in task performance, not as a claim about internal mechanism. Our work asks whether the mechanistic claim implicit in "forgetting" language is warranted.

**Mechanistic interpretability and causal intervention.** Activation patching and causal mediation analysis (Vig et al., 2020; Meng et al., 2022a; Geiger et al., 2021) ask whether a specific component is causally responsible for a specific behavior. We apply this toolkit longitudinally across a training trajectory that deliberately overwrites an association, separating causal accessibility from behavioral output as they evolve. Work on knowledge localization and editing (Meng et al., 2022a;b) is methodologically related but asks how to edit a currently-active association, not what remains after overwriting. Delayed generalization (Power et al., 2022) shows behavioral/structural decoupling in the opposite direction; our result is complementary.

## 3 Formal Framework

At checkpoint $t$, we distinguish three properties w.r.t. target input $x$ and frozen causal direction $J_A$ (identified once from $\theta_A$, never refit).

**Behavioral state.** $B_t(x) = f_{\theta_t}(x)$ – ordinary forward-pass output.

**Causal accessibility.** Let $\mathcal{I}_{J_A}(h)$ denote the intervention that removes the projection of hidden state $h$ onto $J_A$: $\mathcal{I}_{J_A}(h) = h - \frac{J_A^\top h}{J_A^\top J_A} J_A$. We define

$$P_A(t) = m\big(f_{\theta_t}(\mathcal{I}_{J_A}(h_t))\big) - m\big(f_{\theta_t}(h_t)\big), \tag{1}$$

where $m(\cdot)$ is the relevant output margin (target-class logit minus competing-class logit) and $h_t$ is the network's current hidden state for $x$. $P_A(t)$ answers: *can the historical direction still causally change the network's current computation?* This is denoted $C_A(t)$ in some later sections for brevity; the two names refer to the same quantity.

**Intervention-defined recoverability.** Let $h_A^{\parallel J_A}$ denote the projection of $\theta_A$'s hidden state onto $J_A$. $R_A(t) = f_{\theta_t}\big(\text{Patch}(h_t, h_A^{\parallel J_A})\big)$ answers: *does external injection of the historical component reinstate A-like output?* This is distinct from *spontaneous computational retention* (not tested).

These are logically independent properties: $B_t$ can change while $P_A(t)$ and $R_A(t)$ remain at their phase-A values, and later, under sufficient parameter pressure, $P_A(t)$ can itself decay. Table 1 previews the dissociation pattern that our results establish; each cell is a specific, falsifiable claim tested in the sections that follow.

**What $J_A$ measures – and what it does not.** $J_A \propto W_{\text{red}} - W_{\text{blue}}|_{\theta_A}$ is defined by the phase-A output weights, raising the concern that persistent causal effects reflect readout-weight alignment rather than genuine upstream structure. We address this via: (i) a preregistered upstream experiment $d_{\text{src}}$ (Section 11.1) derived with no `zor` and no readout access, achieving AUC$= 1.000$ on held-out filler prediction and causally steering $M_{AB}$'s output for unseen objects ($p = 10^{-6}$); (ii) SS1–SS3 (Section 11.2) establishing causal leverage upstream of fc2 via weight-space, NIE, and foreign-readout tests. Together these distinguish persistent upstream structure from readout-weight preservation. What they do *not* establish is a unique mechanistic account of that structure.

Table 1: The three-state dissociation (*causal accessibility*: current sensitivity to $J_A$; *intervention-defined recoverability*: restoration via external injection; *spontaneous computational retention*: not tested in this paper).

| State | Behavior $B_t$ | Accessibility $P_A(t)$ | Recoverability $R_A(t)$ |
|---|---|---|---|
| Phase A | A | high | high |
| Early A→B (ordinary training) | B | high | high |
| Eroded (under parameter pressure) | B | low | low |

**Falsification conditions.** This framework makes concrete predictions that our experiments could have contradicted. It would be falsified if: $B_t$ changed only when $P_A(t)$ vanished (no dissociation); $P_A(t)$ vanished immediately after the behavioral flip (no persistence); the component orthogonal to $J_A$ restored behavior as well as the parallel component (no localization); a direction from an independently trained model transferred as well as the network's own historical direction (no lineage specificity); or accessibility and downstream readout compatibility were unrelated under parameter pressure (no support for a shared-substrate account of erosion). Each condition is tested in Section 5–Section 9.

## 4 Experimental Setup

We use a minimal synthetic task designed to isolate the phenomenon while remaining fully interpretable: an object-and-context classification problem. A vocabulary of tokens includes filler objects with a fixed, context-independent color mapping (kept constant throughout, so the network must learn a genuine compositional task rather than a single association); a control object `vex` with a context-dependent mapping (`CTX_RED`→red, `CTX_BLUE`→blue) held fixed across all phases; a sham object `fenn`, also context-dependent, but mapping `CTX_RED`→green; and the object of interest, `zor`. In phase A, `zor` follows the same context-dependent rule as `vex` (`CTX_RED`→red). In phase B, `zor`'s mapping is overwritten to be context-*independent*, always producing blue.

The model is a small feed-forward network: token and context embeddings are concatenated, passed through a hidden layer with `tanh` nonlinearity, and read out by a linear classification head. Section 8 reports replication on an unrelated boolean-function task (XOR overwritten by AND) and across three architectures, ruling out that our central result depends on this specific task or model class.

Two training populations are compared throughout: $M_{AB}$, which is trained on phase A to convergence and then continued on phase B; and $M_B$, an independently initialized model trained on phase-B data alone, with the same data distribution used for phase B in $M_{AB}$ except that `zor` never has a context-dependent phase. $M_{AB}$ and $M_B$ are matched to near-identical output distributions (KL divergence $< 10^{-3}$ throughout) before any causal comparison is drawn between them, so that any observed difference reflects training history rather than current behavioral disagreement.

Section 3's frozen direction $J_A$ is the gradient of the red-versus-blue output margin with respect to the hidden state, evaluated at $\theta_A$ for `zor` in `CTX_RED`. This direction is verified, before any phase-B training occurs, to be causally necessary for the phase-A behavior (ablating it collapses the red-versus-blue margin) and specific (ablating it leaves an unrelated filler object's behavior and the phase-A-matched control object's own behavior unaffected). Full construction, verification, and code are in the supplementary repository described in the Reproducibility Statement at the end of the main text.

## 5 Steps 1–2: Behavior Flips While Causal Accessibility Persists

*(Core chain, Steps 1–2 of 5.)* We establish the central dissociation by intervening on $J_A$ at each checkpoint during phase-B training and tracking both the behavioral margin $m(t)$ and $C_A(t)$.

Across 7 seeds, $t_{\text{flip}} = 10$ in every seed, yet $C_A(t)$ *grows* over 3000 steps (mean final $-10.94$ from $-7.83$ at $t = 0$; matched-norm random control: mean $|C_r| = 0.015$, more than two orders of magnitude smaller). A double-intervention test at the final checkpoint shows $J_A$ and the current blue-mechanism $J_B$ combine additively at the readout ($\Delta_{A+B} - \Delta_A - \Delta_B \approx 7 \times 10^{-7}$) – a math-

Table 2: Natural forward-pass participation. *Top*: per-seed final checkpoint ($t$-test vs. 0: $t = -58.6$, $p < 10^{-10}$). *Bottom*: mean trajectory across 7 seeds.

| Seed / Step | $h \cdot \hat{J}_A$ | Ablation $\Delta$ | Frac from $J_A$ |
|---|---|---|---|
| 1234 (final) | $-2.608$ | 10.909 | 0.901 |
| 1235–1240 mean | $-2.539$ | 10.074 | 0.877 |
| **All seeds (mean±sd)** | $\mathbf{-2.563 \pm 0.107}$ | $\mathbf{10.164 \pm 0.511}$ | $\mathbf{0.880 \pm 0.034}$ |
| Step 0 (phase A end) | $+2.279$ | 8.287 | $-1.007$ |
| Step 50 ($t_{\text{flip}}$) | $-2.208$ | 8.024 | 0.854 |
| Step 1000 | $-2.343$ | 8.688 | 0.873 |
| Step 3000 (final) | $-2.563$ | 10.164 | 0.880 |

ematical property of linear projections before a linear layer, not an independent empirical finding (see Appendix A).

**Quantitative model.** $C_A(t)$ and the analogous current-mechanism quantity $C_B(t)$ together explain the behavioral margin with $R^2 = 0.999$ (half-split calibration). Leave-one-seed-out: LOSO-$R^2 = 0.980 \pm 0.028$ across 7 seeds, exceeding all 1854 seed-permutation derangements in the null distribution ($p < 0.00054$; null 99th pct $= 0.733$).

## 6 Step 2 (Strengthened): $J_A$ Is Load-Bearing in Natural Forward Passes

Prior results show $J_A$ is influential under intervention. A sharper question: does $J_A$ participate in *natural*, non-intervened B-behavior forward passes? Since `fc2` is linear:

$$m(h) = m(h_\parallel) + m(h_\perp), \quad h_\parallel = (h \cdot \hat{J}_A)\hat{J}_A. \tag{2}$$

We track $h \cdot \hat{J}_A$, $m(h_\parallel)/m(h)$, and the ablation effect over 3000 B-training steps ($n = 7$ seeds, $J_A$ frozen at $\theta_A$). Table 2 gives per-seed and trajectory results.

$J_A$ is active ($t = -58.6$, $p < 10^{-10}$), grows $-2.21$
$to - 2.56$ from $t_{mathrmflip}$ to step 3000, and ablating it destroys 86

### 6.1 Mediation and Dose-Response

**Protocol.** At the final B-training checkpoint: (1) record $h_{\text{nat}}$, $m_{\text{nat}}$; (2) ablate $h_{\text{abl}} = h_{\text{nat}} - (h_{\text{nat}} \cdot \hat{J}_A)\hat{J}_A$; (3) rescue $h_{\text{resc}} = h_{\text{abl}} + h_\parallel$; (4) sweep $h_\lambda = h_{\text{abl}} + \lambda h_\parallel$, $\lambda \in [0, 1]$; (5) null control: restore a random direction $v_\perp \perp J_A$ with $\|v_\perp\| = \|h_\parallel\|$.

**Results.** Table 3 reports per-seed results. Ablation: $10.164 \pm 0.511$ units (86%); rescue recovers exactly ($< 10^{-3}$ per seed). Null control: $+0.120 \pm 0.346 \approx 0$. $\lambda$-sweep: perfectly linear.

Restoring only $h_\parallel$ rebuilds the full margin; matched-norm orthogonal restores nothing.

## 7 Step 3: The Historical Component Restores A-Behavior

*(Core chain, Step 3 of 5.)* The above shows $J_A$ is causally influential; this section establishes the $J_A$-parallel component is both necessary and sufficient, and that the effect is lineage-specific.

**Component isolation.** We decompose $\theta_A$'s hidden state as $h_A = h_A^{\parallel J_A} + h_A^{\perp J_A}$, with $J_A$ frozen from $\theta_A$ and never refit. At the final $M_{AB}$ checkpoint we patch each component separately via dose-response interpolation, reporting seed-level statistics with 95% CIs. Table 4 reports the results.

The parallel component alone reproduces the full-activation restoration almost exactly (confidence intervals overlap nearly completely), while the orthogonal remainder alone produces no restoration

Table 3: Mediation results ($n = 7$ seeds). Null ctrl $\Delta = +0.120 \pm 0.346 \approx 0$ (mixed sign); effect is direction-specific.

| Seed | $m_{\text{nat}}$ | $m_{\text{abl}}$ | Abl.$\Delta$ | $m_{\text{resc}}$ | Resc.$\Delta$ | Ctrl$\Delta$ |
|---|---|---|---|---|---|---|
| 1234 | $+12.080$ | $+1.171$ | $+10.909$ | $+12.080$ | $+10.909$ | $+0.060$ |
| 1235 | $+11.570$ | $+1.251$ | $+10.319$ | $+11.570$ | $+10.319$ | $-0.465$ |
| 1236 | $+11.685$ | $+1.925$ | $+9.760$ | $+11.685$ | $+9.760$ | $+0.571$ |
| 1237 | $+11.137$ | $+1.433$ | $+9.704$ | $+11.137$ | $+9.703$ | $+0.256$ |
| 1238 | $+11.161$ | $+1.772$ | $+9.389$ | $+11.161$ | $+9.389$ | $+0.450$ |
| 1239 | $+11.988$ | $+1.456$ | $+10.531$ | $+11.988$ | $+10.531$ | $+0.234$ |
| 1240 | $+12.621$ | $+2.089$ | $+10.533$ | $+12.621$ | $+10.533$ | $-0.269$ |
| Mean | $+11.749$ | $+1.585$ | $+10.164$ | $+11.749$ | $+10.164$ | $+0.120$ |
| $\pm$sd | 0.491 | 0.322 | 0.511 | 0.491 | 0.511 | 0.346 |

Table 4: Necessity and sufficiency of the historical component (7 seeds, mean $\pm$ 95% CI; blue-positive margin convention, negative values indicate the phase-A output is restored).

| Condition | Mean margin | 95% CI | Restores A (n/7) |
|---|---|---|---|
| Own full historical activation | $-8.10$ | $[-8.56, -7.64]$ | 7/7 |
| Own $J_A$-parallel component only | $-8.08$ | $[-8.57, -7.60]$ | 7/7 |
| Own orthogonal remainder only | $+11.65$ | $[11.22, 12.08]$ | 0/7 |
| Foreign lineage, full activation | $+2.00$ | $[-0.22, 4.22]$ | 1/7 |
| Random matched-norm activation | $-0.15$ | $[-2.88, 2.57]$ | 3/7 |

in any seed. This is a clean necessity-and-sufficiency result: the causal signal is not diffusely spread across the hidden state but is concentrated in the single, independently identified direction.

**Effect size.** Own: $-8.10 \pm 0.62$; random: $-0.15 \pm 3.68$ ($d = -3.01$, $p = 0.004$); orthogonal $d = -32.8$; foreign $d = -4.67$.

**Nonlinear readout.** All results above use a linear readout, which gives unusually clean causal geometry and raises the concern that some findings (e.g., the exact additivity noted in Section 5) could be architectural consequences of linearity rather than properties of the phenomenon. We reproduce the central dissociation with the linear readout replaced by a two-layer ReLU MLP. The behavioral flip remains immediate and uniform ($t_{\text{flip}} = 10$ in all 7 seeds, identical to the linear case), and $C_A$ at the final checkpoint remains large in every seed (mean $-8.61$, range $-6.2$ to $-12.0$), averaging $75.6\%$ of its value at $\theta_A$. The central phenomenon is not a consequence of linear readout geometry.

**Lineage specificity.** A skeptic could still argue $J_A$ is simply a generically useful direction for producing red, which any similarly-trained model would share. We test this with a full $7 \times 7$ cross-lineage transplantation matrix: for every pair of seeds $(i, j)$, we patch seed $i$'s own $\theta_A$ activation into seed $j$'s trained-forward model. The diagonal (same-lineage transfer, $n = 7$) has mean $-8.10$ ($\text{std} = 0.62$) and restores the phase-A output in $7/7$ cases. The off-diagonal (all 42 ordered foreign-lineage pairs, no subsampling) has mean $+1.11$ ($\text{std} = 3.28$) and restores it in only $12/42$ cases. Every single diagonal value is more negative than every single off-diagonal value (Mann–Whitney $U = 0$, $p < 2.4 \times 10^{-8}$); lineage specificity, $(\Delta_{\text{own}} - \mathbb{E}[\Delta_{\text{foreign}}])/\sigma_{\text{foreign}}$, is $-2.81$ standard deviations. The historical component is tied to the specific training trajectory that produced it, not to the task in the abstract.

**Robustness of the causal direction's construction.** A further concern is that $J_A$'s effect is an artifact of one particular mathematical construction (the gradient of an output margin). We construct three independent alternatives at $\theta_A$: the gradient of the classification cross-entropy loss (a related but distinct object); the difference between paired activations, $h(\texttt{zor}, \texttt{CTX_RED}) - h(\texttt{vex}, \texttt{CTX_BLUE})$ (not a gradient at all); and a linear probe fit exclusively on filler-object activations, which never touches `zor`'s own activation during construction. The first three converge:

Table 5: Seven independently constructed directions at $\theta_A$. The readout is the critical ingredient: directions engaging it (D1–D3, partially D4) produce large ablation effects; those omitting it (D5, D6) produce none.

| Direction | Construction | Uses `zor`? | Uses readout? | Uses `CTX_RED`? |
|---|---|---|---|---|
| D1 | Margin Jacobian at `zor`, `CTX_RED` | ✓ | ✓ | ✓ |
| D2 | Avg margin Jacobian, red fillers only | × | ✓ | ✓ |
| D3 | Margin Jacobian at `zor`, `CTX_BLUE` | ✓ | ✓ | × |
| D4 | Cross-entropy loss gradient (red) at `zor` | ✓ | indirect | ✓ |
| D5 | Diff-in-diff activation direction | ✓ | × | ✓ |
| D6 | Linear probe on filler hidden states | × | × | ✓ |
| D7 | Cross-model: $M_A^{\text{indep}}$ activation | × | — | ✓ |

pairwise cosine similarity at $\theta_A$ ranges 0.67–0.98, and all three produce large, consistent ablation effects on the trained-forward model (7/7 seeds each; means $-10.1$, $-4.8$, $-8.7$). The disjoint-input probe is nearly orthogonal to the others ($\cos \in [0.11, 0.39]$) and produces no effect (0/5 seeds, mean 0.07) – a principled negative control showing the effect requires actually engaging the specific binding computation, not any direction that happens to separate red- from blue-mapped objects in general.

### 7.1 Six Independently Constructed Directions

To isolate the critical ingredient, we construct seven directions at $\theta_A$, each varying one or more of: target object `zor`, the readout, context `CTX_RED` (Table 5).

D1=D2=D3 ($|\cos| = 1.000 \pm 0.000$), identical ablation effects (mean $-10.16 \pm 0.51$). D4 nearly aligned ($|\cos| = 0.915$), $\approx 50\%$ effect (mean $-4.82$). D5, D6 nearly orthogonal ($|\cos| < 0.26$), near-zero effects ($-0.38$, $+0.11$, mixed sign). D7 transfers in 2/7 seeds only (lineage-specific). Per-seed table in Appendix A.

### 7.2 Subspace-Level Causal Claim

We construct a $k$-dim A-subspace from D1–D4 via SVD ($B \in \mathbb{R}^{32\times k}$, never seeing $M_{AB}$'s hidden state). Ablation: $h_{\text{abl}} = h - BB^\top h$. Controls: C1 (random Haar), C2 (random rotation of A-basis), C3 (filler Jacobians), C4 (B-subspace). Random-rotation sweep: 200 uniformly random $k$-dim subspaces.

The A-subspace ablation is stable: mean $-10.16$ ($k = 1$) to $-10.79$ ($k = 4$); 0 of 200 random rotations match at every $k$. At $k = 2$ ($> 85\%$ of $\|h\|$): A-subspace $-10.70 \pm 0.69$ vs C1 $-0.72 \pm 0.13$, C2 $-0.96 \pm 0.22$, C3 $-0.20 \pm 1.15$; 100% margin recovery in 7/7 seeds. Paired $t$-tests (all $p < 10^{-4}$): C1 $t = -39.5$, C2 $t = -45.5$, C3 $t = -16.1$, C4 $t = +8.9$. Subspace overlaps: A–B $\cos^2 = 0.97$ at $k = 1$ (falling to 0.51 at $k = 2$); A–filler $\cos^2 = 0.07$. Full tables in Appendix A.

The causal structure is 0-in-200 specific to a low-dimensional A-era subspace. The partial A–B overlap ($\cos^2 = 0.97$ at $k = 1$, 0.51 at $k = 2$) with near-zero A–filler overlap is consistent with partially overlapping representational resources – a weaker claim than "a shared substrate" – and is also consistent with other forms of distributed dependence. The geometry is readout-dependent and does not uniquely identify the underlying architecture.

## 8 Step 4 (Specificity): Cross-Input Transfer and Scope

`vex` (same phase-A rule as `zor`) transfers nearly as well when patched into $M_{AB}$ ($-8.38$ vs $-8.10$; $p = 0.11$). A similarity-matched synthetic decomposes the transfer:

$$\underbrace{-8.38 \pm 0.65}_{\text{total, vex}} = \underbrace{-6.66 \pm 0.87}_{\text{similarity-mediated}} + \underbrace{-1.72 \pm 0.96}_{\text{identity residual}}. \tag{3}$$

79% is explained by similarity; the 21% identity-specific residual replicates in 7/7 seeds ($p = 0.016$). Under a multi-statistic-matched synthetic (cosine, norm, margin, Jacobian all matched to

Table 6: Causal persistence across 10 model configurations (mean over 4 seeds). frac_rem = $|\Delta_{\mathrm{AB}}|/|\Delta_A|$ ($> 1$: causal effect grew during B-training). spec_ratio = $|\mathrm{abl}_\Delta|/|\mathrm{ctrl}_\Delta|$. P(B)$= 1.00$ in all configurations.

| Config | Task | frac_rem | abl$\Delta$ | ctrl$\Delta$ | frac_margin | spec_ratio |
|---|---|---|---|---|---|---|
| MLP-Tiny (h=32, 1L) | orig | 1.187 | $+10.744$ | $-0.039$ | 0.897 | 588 |
| MLP-Small (h=128, 1L) | orig | 0.969 | $+10.399$ | $-0.011$ | 0.766 | 3923 |
| MLP-Medium (h=256, 2L) | orig | 0.897 | $+13.459$ | $+0.001$ | 0.820 | 2663 |
| MLP-Large (h=512, 2L) | orig | 0.985 | $+14.090$ | $+0.007$ | 0.709 | 2876 |
| MLP-Deep (h=256, 4L res.) | orig | 1.014 | $+11.940$ | $-0.011$ | 0.798 | 9234 |
| Transformer-S (d=64, 1L) | orig | 1.338 | $+16.773$ | $+0.038$ | 0.965 | 2381 |
| Transformer-M (d=128, 2L) | orig | 1.542 | $+23.758$ | $+0.012$ | 0.968 | 5450 |
| Hard-MLP (S/M/L) | hard | 1.06 | $+16$–$22$ | $\approx 0$ | 0.88 | $10^3$–$10^5$ |

$< 10^{-3}$), the residual survives ($-0.665 \pm 0.316$; $7/7$, $p = 0.016$). `fenn` (different function) fails entirely ($0/7$). Computation is partially shared but partially identity-bound.

**Generality.** XOR→AND task: $t_{\text{flip}} \in [10, 20]$ ($7/7$), $C_A$ range 6.4–14.3; two MLP scales $7/7$; attention $4/6$. Details in Appendix F.

## 9 Step 4 (Specificity): When and How Accessibility Erodes

Under parameter pressure (weight decay + narrow bottleneck), accessibility can decay – always after the behavioral flip ($t_{\text{flip}} = 26$, 5 seeds; erosion at steps 156 and 260). Two results address the mechanism.

**Counterfactual matrix.** If the A-computation persisted as an independent circuit, ablating $J_B$ should reveal a stronger A-like signal in $M_{AB}$ than in the matched $M_B$ (no phase-A history). The opposite holds: ablating $J_B$ produces a smaller margin shift in $M_{AB}$ than in $M_B$ ($9.61 \pm 1.11$ vs. $12.07 \pm 0.63$; $7/7$ seeds, Wilcoxon $p = 0.016$).

**Readout co-adaptation.** Under ordinary training, $\cos(J_A, W_{\text{red}}(t) - W_{\text{blue}}(t))$ stays constant ($0.999 \to 0.987$) while $C_A(t)$ grows – the readout never rotates away from $J_A$. Under parameter pressure, readout alignment decays in step with $C_A(t)$ seed by seed (Spearman $\rho = 1.0$, $n = 5$; permutation $p = 1/120$). We report this cautiously: with $n = 5$, it is suggestive rather than decisive alone. Combined with the counterfactual matrix, this pattern is more consistent with partially overlapping representational resources than with a cleanly gated independent circuit – though we note that correlated erosion of two quantities derived from the same pathway does not uniquely rule out other distributed-dependence explanations.

**Independent subspace separability.** $J_B^{\text{indep}}$ (filler-probe): directions interfere by ≈one-third – against clean separability. Held-out generalization: $J_A$ from 20-object subset restores A in 140/140 held-out cases. Alternative explanations table in Table 9.

## 10 Further Evidence: Causal Persistence at Scale

We test across: (i) MLP width/depth (hidden 32–512, 1–4 layers); (ii) harder task (50 objects, 12 colors, 4 contexts); (iii) Transformers ($d = 64/128$, OBJ-token readout). 4 seeds each, orthogonal null control.

All configurations: P(B)$= 1.00$, abl$\Delta$ large, ctrl$\Delta \approx 0$, frac_margin$\geq 0.71$; Transformers: $\approx 0.97$; hard task: spec $> 60000$; frac_rem$\geq 1$ – A-direction's effect grows during B-training.

### 10.1 Mechanistic Probes

Three additional experiments probe the retained structure.

Table 7: Preregistered upstream test ($d_{\text{src}}$, $n = 5$ eligible seeds). All stages use no `zor` and no readout in derivation. T2 uses $M_{AB}$, never seen during derivation.

| Test | Key statistic | Value | $p$ |
|---|---|---|---|
| T1 (held-out object prediction) | AUC | 1.000 (5/5) | $< 0.0001$ |
| T1 ($r_{pb}$) | point-biserial $r$ | $+0.877 \pm 0.097$ | $< 0.0001$ |
| T2 (causal steering, red shift) | mean $\Delta$ margin | $+13.35 \pm 1.82$ | 0.000004 |
| T2 (orthogonal control) | mean $\Delta$ margin | $+0.35 \pm 1.98$ | — |
| T3 (cross-model cos) | $\cos(d_{\text{src}}^{A}, d_{\text{src}}^{AB})$ | $+0.910 \pm 0.084$ | 0.000017 |

**B-from-scratch.** $\cos(J_A, J_{AB}) = -0.985$; $\cos(J_A, J_{\text{scr}}) = -0.040$; ablating $J_A$ from $M_{AB}$ destroys 88% of the margin; from scratch: $+0.007 \pm 0.125$. Lineage-specific (see **??**).

**Targeted weight edit.** Zeroing `embed[zor]`: $-15.52 \pm 5.02$; swapping for `embed[fenn]`: $+1.50 \pm 1.25$ (green shift $-1.67 \to +0.76$); matched-norm random: $-17.91 \pm 4.28$. The embedding carries content-specific information.

**Gradient alignment.** $\cos(\text{grad}_h, \hat{J}_A) = +1.000$ before $t_{\text{flip}}$, $+0.958 \pm 0.031$ after (3000 steps). B-training routes through the A-substrate.

## 11 Step 5: Upstream Structure Independent of fc2 Geometry

*(Core chain, Step 5 of 5.)* Prior results use $J_A$, which is defined via the readout. This section asks whether the effect reflects genuine upstream structure or merely fc2 geometry – using a preregistered experiment with *no readout and no* `zor` at any derivation stage.

### 11.1 Preregistered Upstream Test ($d_{\text{src}}$)

**Design.** At $M_A$, with no fc2 access and no `zor`: $d_{\text{src}} = \text{mean}(h_{\text{red-src}}) - \text{mean}(h_{\text{nonred}})$, where red-src fillers = the first $\lceil n/2 \rceil$ red-labeled fillers. Direction frozen immediately. Four disjoint held-out tests follow.

**T1 — Held-out object prediction (no fc2, no zor).** $d_{\text{src}}$ is scored against held-out red/nonred fillers never in the source. AUC $= 1.000$ in all 5 eligible seeds ($r_{pb} = 0.877 \pm 0.097$; $t = 20.3$, $p < 0.0001$). Every held-out red filler ranks above every nonred filler – purely upstream, within the hidden layer.

**T2 — Causal upstream steering.** Steer $h_{\text{obj}}$ along $d_{\text{src}}$ in $M_{AB}$ for nonred fillers (never in derivation; model never in derivation). Red-class shift: $+13.348 \pm 1.819$; orthogonal control: $+0.353 \pm 1.975$ ($t = 34.3$, $p = 0.000004$, $20.9\times$ specificity). $d_{\text{src}}$ causally steers color output for objects it never saw, through the B-trained readout.

**T3 — Cross-model structure persistence.** Re-derive $d_{\text{src}}$ at $M_{AB}$ from fillers only. $\cos(d_{\text{src}}^{M_A}, d_{\text{src}}^{M_{AB}}) = +0.910 \pm 0.084$ ($t = 24.4$, $p = 0.000017$); held-out re-derivation: $+0.897 \pm 0.087$. Entirely filler-based, no `zor`, no readout. The upstream color-encoding direction persists across the A→B transition.

**Partial transfer to `zor`.** $d_{\text{src}}$ ranks `zor` at the 70th percentile among fillers, but ablating it from `zor`'s hidden state does not significantly disrupt B output ($p = 0.71$; 3.5% of $J_A$'s effect). Upstream color structure is detectable at `zor` via $d_{\text{src}}$, but causal accessibility to the output requires readout alignment.

### 11.2 Additional Upstream Controls (SS1–SS3)

SS1 (fc1 weight-space, zero fc2 access): $\cos(w_A^A, w_A^{AB}) = +0.826$ (preserved), $\cos(w_A^{AB}, w_A^{\text{scr}}) = +0.023$ (lineage-specific), ablation spec $353\times$. SS2 (NIE decomposition): $M_A$'s $h$ through $M_{AB}$'s

Table 8: Mechanistic discrimination ($n = 7$ seeds, $\alpha = 1.0$, displacement norm $= 1.0$ for all directions). $\Delta A$: change in A-decoder score; $\Delta B$: change in B-decoder score. NS: not significant.

| Removed direction | $\Delta A$ mean | $\Delta B$ mean | $\Delta A$: $t, p$ | $\Delta B$: $t, p$ |
|---|---|---|---|---|
| $e_A$ | $-2.073$ | $+1.974$ | $t = -27.8, p < 10^{-5}$ | $t = +35.0, p < 10^{-5}$ |
| $e_B$ | $+2.613$ | $-3.169$ | $t = +37.7, p < 10^{-5}$ | $t = -35.9, p < 10^{-5}$ |
| $e_{\text{sh}}$ | $+0.089$ | $-0.032$ | $t = +1.24, p = 0.26$ (NS) | $t = -0.34, p = 0.74$ (NS) |
| $e_\perp$ (null) | $+0.052$ | $-0.019$ | $t = +0.47, p = 0.66$ (NS) | $t = -0.16, p = 0.88$ (NS) |

fc2 $\rightarrow +8.10$ (red); scratch $h \rightarrow -0.33 \approx 0$; NIE gap $+8.43$ ($p < 10^{-5}$). SS3 (foreign filler readout, permuted labels): gap $+2.808 \pm 1.575$ ($p = 0.005$) with no `zor` history and no "red" concept. Full per-seed tables in Appendix B.

## 12 Mechanistic Discrimination: Separating $e_A$, $e_B$, and the Shared Component

We test whether A- and B-associated causal components are separable via an orthogonal Gram-Schmidt basis: $e_A = \hat{h}_{M_A}(\texttt{zor})$; $e_B = \text{GS}(h_{M_{AB}} - h_{M_A}, [e_A])$; $e_{\text{sh}} = \text{GS}(h_{\text{scratch}}(\texttt{zor}), [e_A, e_B])$; $e_\perp$ = random null. All pairwise cosines $= 0.000000$ (verified per-seed). Each ablation uses the same displacement norm $\alpha$. Two independent decoders: $A_{\text{score}} = \ell_{\text{red}} - \ell_{\text{blue}}$ through $M_A$'s fc2; $B_{\text{score}} = \ell_{\text{blue}} - \ell_{\text{red}}$ through $M_{AB}$'s fc2.

**Results.** Table 8: removing $e_A$ degrades both decoders (both $p < 10^{-5}$); removing $e_B$ affects both in the opposite direction. $e_{\text{sh}}$ and $e_\perp$ are NS in all four tests ($p > 0.26$). Dose-response across $\alpha \in \{0.5, 1, 2, 4\}$ is linear for $e_A$ and $e_B$; $e_{\text{sh}}$ near-zero at every $\alpha$.

**What this establishes.** $e_A$ and $e_B$ are *functionally* separable: each independently influences both decoders in opposite directions, at matched norm, in an orthogonal basis. $e_{\text{sh}}$ is causally inert at every $\alpha$ – not a sensitivity issue. Crucially, $e_A$ influences the B-decoder (and vice versa): this is the signature of a shared rather than cleanly independent representational substrate, directly supporting the partially overlapping resources interpretation.

## 13 Discussion and Conclusion

The core chain establishes that behavioral forgetting does not imply causal erasure. Mechanistic hierarchy: (1) causal accessibility/recoverability – very strong; (2) $e_A$ and $e_B$ functionally separable in orthogonal basis with independent decoders ($p < 10^{-5}$) – very strong; (3) $e_A$ influences the B-decoder (shared substrate signature) and $e_{\text{sh}}$ is causally inert – substantially supported; (4) partially overlapping representational resources – much better supported, still an interpretation not uniquely established. Our experiments demonstrate causal accessibility and recoverability – not spontaneous computational retention, which is distinct and untested. *Limitations*: synthetic tasks; readout-alignment $n = 5$; hidden-state patching only.

## Reproducibility Statement

All experiments are in a public code repository (link redacted; in supplementary material). All statistics are computed by script from raw seed-level results. Seeds 1234–1240; hyperparameters in Appendix A.

**AI Use Disclosure.** LLMs assisted with literature search, brainstorming, code review, and manuscript editing. All experiments and scientific claims were independently implemented and verified by the authors.

## REFERENCES


Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165, 1989.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022a.

Kevin Meng, Arnav Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022b.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, 2020.


## A ADDITIONAL METHODOLOGICAL DETAIL

**Task and model specification.** The color task vocabulary consists of 16 filler objects with a fixed random color mapping (seeded per experiment), the control object `vex`, the sham object `fenn`, and `zor`. Two context tokens (`CTX_RED`, `CTX_BLUE`) and six color classes are used. The model embeds the object token (16-dim) and context token (8-dim), concatenates them, passes the result through a 32-unit hidden layer with `tanh` activation (optionally routed through a low-rank bottleneck for the parameter-pressure condition in Section 9), and reads out through a linear layer to 6 logits. Phase-A training uses 600 steps, batch size 32, Adam with learning rate 0.01. Phase-B training uses 3000 steps (8000 under the parameter-pressure condition), batch size 32, Adam with learning rate 0.005, and weight decay 0 (ordinary) or 0.1 (parameter-pressure condition, combined with a 2-dimensional bottleneck).

**Fossilization sweep.** The erosion result in Section 9 is the outcome of a systematic sweep over training length, learning rate, hidden dimension, weight decay strength, and a capacity bottleneck, conducted while searching for a regime in which structure (Jacobian-cosine alignment, $\rho_A$) and function (causal mediation) dissociate under parameter pressure. Requiring both high alignment *and* low mediation jointly, rather than either alone, no configuration produces a clean dissociation across 20-seed replication; the two instead co-vary tightly (Spearman $\rho = 0.93$, $p < 10^{-8}$, versus $\rho \approx 0$ under ordinary training) – evidence against selective preservation of structure independent of function, and the basis for the discussion of erosion in Section 9.

**Non-tautological interaction construction.** The interaction quantity reported in Section 5 ($\Delta_{A+B} - \Delta_A - \Delta_B \approx 0$) is a property of ablating two independent linear projections before a linear readout and holds for any two directions regardless of what the network has learned; it is not evidence about the trained network's mechanism. We instead use a conditional interaction test, $\Gamma_{AB} = C_{B|A} - C_B$, where $C_{B|A}$ recomputes the ablation projection on the *already-ablated* hidden state. This construction, validated against a synthetic ground-truth function with known interaction strength before use, is large in both training conditions (relative magnitude $\approx 0.95$–$0.96$ of the typical $C_A/C_B$ scale) – rejecting pure additive independence between the two mechanisms and superseding the linear-algebra artifact above; this is one further piece of evidence against clean separability, consistent with but not sufficient on its own to establish the shared-substrate account discussed in Sections 9 and 13.

# B Additional Tables

**Natural forward-pass trajectory (per-seed).**

| Seed | Step 50 ($h \cdot \hat{J}_A$) | Step 50 (frac) | Step 3000 ($h \cdot \hat{J}_A$) | Step 3000 (frac) |
|---|---|---|---|---|
| 1234 | $-2.150$ | 0.853 | $-2.608$ | 0.901 |
| 1235 | $-2.204$ | 0.917 | $-2.578$ | 0.934 |
| 1236 | $-2.043$ | 0.802 | $-2.372$ | 0.837 |
| 1237 | $-2.168$ | 0.858 | $-2.492$ | 0.907 |
| 1238 | $-2.199$ | 0.841 | $-2.506$ | 0.874 |
| 1239 | $-2.340$ | 0.884 | $-2.673$ | 0.873 |
| 1240 | $-2.348$ | 0.818 | $-2.708$ | 0.835 |

**Dose-response sweep (seed 1234).**

| $\lambda$ | $J_A$ margin | Blue? | Ctrl margin |
|---|---|---|---|
| 0.0 | $+1.171$ | ✓ | $+1.171$ |
| 0.5 | $+6.626$ | ✓ | $+1.201$ |
| 1.0 | $+12.080$ | ✓ | $+1.231$ |

**Per-seed six-direction ablation effects.**

| Seed | $m_{\text{nat}}$ | D1 | D2 | D3 | D4 | D5 | D6 |
|---|---|---|---|---|---|---|---|
| 1234 | $+12.080$ | $-10.909$ | $-10.909$ | $-10.909$ | $-5.204$ | $+0.019$ | $+0.219$ |
| 1235 | $+11.570$ | $-10.319$ | $-10.319$ | $-10.319$ | $-5.561$ | $-0.310$ | NA |
| 1236 | $+11.685$ | $-9.760$ | $-9.760$ | $-9.760$ | $-6.981$ | $-0.193$ | $+0.377$ |
| 1237 | $+11.137$ | $-9.704$ | NA | $-9.704$ | $-7.414$ | $-1.150$ | NA |
| 1238 | $+11.161$ | $-9.389$ | $-9.389$ | $-9.389$ | $-2.746$ | $-0.127$ | $-0.009$ |
| 1239 | $+11.988$ | $-10.531$ | $-10.531$ | $-10.531$ | $-3.175$ | $-0.852$ | $-0.532$ |
| 1240 | $+12.621$ | $-10.533$ | $-10.533$ | $-10.533$ | $-2.661$ | $-0.068$ | $+0.507$ |
| Mean | $+11.749$ | $-10.164$ | $-10.240$ | $-10.164$ | $-4.820$ | $-0.383$ | $+0.113$ |

**Subspace analysis: ablation by $k$ and per-seed rescue.**

| $k$ | A-subspace | C1 random | C2 rotation | C3 filler | C4 B-sub | spec% | frac_h |
|---|---|---|---|---|---|---|---|
| 1 | $-10.164 \pm 0.511$ | $-0.342 \pm 0.145$ | $-0.568 \pm 0.147$ | $+0.276 \pm 0.407$ | $-11.579 \pm 0.672$ | 0.0% | 0.584 |
| 2 | $-10.700 \pm 0.691$ | $-0.717 \pm 0.127$ | $-0.957 \pm 0.218$ | $-0.202 \pm 1.147$ | $-11.579 \pm 0.672$ | 0.0% | 0.850 |
| 3 | $-10.720 \pm 0.687$ | $-1.083 \pm 0.205$ | $-1.301 \pm 0.347$ | $-2.296 \pm 4.103$ | $-11.579 \pm 0.672$ | 0.0% | 0.857 |
| 4 | $-10.794 \pm 0.687$ | $-1.694 \pm 0.305$ | $-1.676 \pm 0.365$ | $-8.384 \pm 4.506$ | $-11.579 \pm 0.672$ | 0.0% | 0.867 |

**Gradient alignment during B-training (per-seed).**

| Seed | $t_{\text{flip}}$ | cos grad before | cos grad after | cos embed before | cos embed after |
|---|---|---|---|---|---|
| 1234 | 50 | $+1.000$ | $+0.954$ | $+0.948$ | $+0.962$ |
| 1235 | 50 | $+1.000$ | $+0.985$ | $+0.823$ | $+0.897$ |
| 1236 | 50 | $+1.000$ | $+0.970$ | $+0.838$ | $+0.856$ |
| 1237 | 50 | $+1.000$ | $+0.985$ | $+0.781$ | $+0.834$ |
| 1238 | 50 | $+1.000$ | $+0.983$ | $+0.954$ | $+0.950$ |
| 1239 | 50 | $+1.000$ | $+0.928$ | $+0.979$ | $+0.965$ |
| 1240 | 50 | $+1.000$ | $+0.900$ | $+0.842$ | $+0.914$ |
| Mean | 50.0 | $+1.000 \pm 0.000$ | $+0.958 \pm 0.031$ | $+0.881 \pm 0.072$ | $+0.911 \pm 0.048$ |

**Transformer-M per-seed detail.**

| Seed | frac_rem | ablΔ | ctrlΔ | $J_A$ proj | frac_margin |
|---|---|---|---|---|---|
| 1234 | 1.962 | $+30.445$ | $+0.039$ | $-34.620$ | 0.993 |
| 1235 | 1.212 | $+19.443$ | $-0.001$ | $-21.091$ | 1.005 |
| 1236 | 1.282 | $+20.482$ | $-0.024$ | $-25.613$ | 0.961 |
| 1237 | 1.711 | $+24.662$ | $+0.034$ | $-31.392$ | 0.915 |
| mean | 1.542 | $+23.758$ | $+0.012$ | $-28.179$ | 0.968 |

**Embedding interpolation: `embed[zor]`→`embed[fenn]` (mean across seeds).**

| $\lambda$ | $m_{\text{B}-\text{R}}$ | $m_{\text{B}-\text{G}}$ | $m_{\text{G}-\text{R}}$ |
|---|---|---|---|
| 0.00 | $+11.749$ | $+13.423$ | $-1.674$ |
| 0.25 | $+12.551$ | $+13.334$ | $-0.784$ |
| 0.50 | $+12.937$ | $+13.148$ | $-0.211$ |
| 0.75 | $+13.144$ | $+12.892$ | $+0.252$ |
| 1.00 | $+13.248$ | $+12.490$ | $+0.758$ |

**SS1 — fc1 weight-space substrate (per-seed).**

| Seed | $\cos(w_A^A, w_A^{AB})$ | $\cos(w_A^{AB}, w_A^{\text{scr}})$ | $\text{abl}_{AB}$ | $\text{abl}_{\text{scr}}$ | Spec |
|---|---|---|---|---|---|
| 1234 | $+0.836$ | $+0.105$ | $-8.540$ | $-0.273$ | $74\times$ |
| 1235 | $+0.805$ | $+0.111$ | $+1.367$ | $+0.261$ | $1.8\times$ |
| 1236 | $+0.895$ | $-0.005$ | $-3.612$ | $-0.003$ | $219\times$ |
| 1237 | $+0.764$ | $-0.036$ | $+4.215$ | $-0.028$ | $171\times$ |
| 1238 | $+0.837$ | $+0.041$ | $-1.359$ | $-0.682$ | $3.7\times$ |
| 1239 | $+0.793$ | $+0.141$ | $-2.810$ | $-1.100$ | $1894\times$ |
| 1240 | $+0.849$ | $-0.195$ | $-5.550$ | $-1.683$ | $111\times$ |
| Mean | $+0.826 \pm 0.039$ | $+0.023 \pm 0.107$ | $-2.33 \pm 3.93$ | $-0.50 \pm 0.64$ | $353\times$ |

**SS3 — Foreign readout test (per-seed).**

| Seed | R1 mAB | R1 scr | R1 gap | R2 mAB | R2 scr | R2 gap |
|---|---|---|---|---|---|---|
| 1234 | $-6.012$ | $-10.100$ | $+4.088$ | $-6.893$ | $-10.615$ | $+3.722$ |
| 1235 | $-8.591$ | $-10.701$ | $+2.110$ | $-8.306$ | $-10.160$ | $+1.853$ |
| 1236 | $-7.207$ | $-7.853$ | $+0.647$ | $-6.745$ | $-9.013$ | $+2.268$ |
| 1237 | $-8.852$ | $-11.461$ | $+2.609$ | $-9.647$ | $-11.011$ | $+1.364$ |
| 1238 | $-8.454$ | $-10.198$ | $+1.744$ | $-7.395$ | $-10.068$ | $+2.674$ |
| 1239 | $-7.087$ | $-13.517$ | $+6.430$ | $-6.354$ | $-12.575$ | $+6.221$ |
| 1240 | $-1.139$ | $-2.314$ | $+1.175$ | $-2.374$ | $-3.926$ | $+1.553$ |
| Mean | $-6.763 \pm 2.48$ | $-9.449 \pm 3.31$ | $+2.686 \pm 1.84$ | $-6.816 \pm 2.09$ | $-9.624 \pm 2.53$ | $+2.808 \pm 1.58$ |

R1: $t = 3.580$, $p = 0.0116$, $d = 1.353$; R2: $t = 4.367$, $p = 0.0047$, $d = 1.650$; filler acc $= 1.00$ all seeds.

**SS2 — NIE decomposition (per-seed).**

| Seed | NIE$_{M_A}$ | NIE$_{\text{scr}}$ | NIE gap | $M_A h/M_{AB}$ fc2 | $M_{AB} h/M_A$ fc2 | scr$h/M_{AB}$ fc2 |
|---|---|---|---|---|---|---|
| 1234 | $+19.769$ | $+8.947$ | $+10.822$ | $+7.689$ | $-9.705$ | $-2.456$ |
| 1235 | $+19.946$ | $+11.381$ | $+8.565$ | $+8.376$ | $-9.614$ | $+0.174$ |
| 1236 | $+19.700$ | $+12.654$ | $+7.046$ | $+8.016$ | $-8.300$ | $+0.834$ |
| 1237 | $+19.469$ | $+11.044$ | $+8.425$ | $+8.332$ | $-9.118$ | $-0.441$ |
| 1238 | $+19.273$ | $+9.114$ | $+10.159$ | $+8.113$ | $-8.832$ | $+0.531$ |
| 1239 | $+20.438$ | $+14.274$ | $+6.164$ | $+8.451$ | $-9.486$ | $+0.083$ |
| 1240 | $+20.342$ | $+12.532$ | $+7.810$ | $+7.721$ | $-9.254$ | $-1.070$ |
| Mean | $+19.848 \pm 0.397$ | $+11.421 \pm 1.792$ | $+8.427 \pm 1.520$ | $+8.100$ | $-9.187$ | $-0.328$ |

# C Preregistered Upstream Experiment — Per-Seed Results

**T1 — Held-out object prediction.**

| Seed | $n_{\text{src}}$ | $n_{\text{held}}$ | $r_{pb}$ | AUC |
|---|---|---|---|---|
| 1234 | 5 | 4 | $+0.971$ | 1.000 |
| 1236 | 1 | 1 | $+0.854$ | 1.000 |
| 1238 | 1 | 1 | $+0.882$ | 1.000 |
| 1239 | 3 | 2 | $+0.951$ | 1.000 |
| 1240 | 1 | 1 | $+0.727$ | 1.000 |
| mean | | | $+0.877 \pm 0.097$ | 1.000 |

**T2 — Causal steering.**

| Seed | Red shift | Ctrl shift | Spec |
|---|---|---|---|
| 1234 | $+13.261$ | $+0.782$ | $46.9\times$ |
| 1236 | $+14.201$ | $-0.134$ | $11.0\times$ |
| 1238 | $+11.247$ | $-1.213$ | $35.8\times$ |
| 1239 | $+12.117$ | $-1.226$ | $4.2\times$ |
| 1240 | $+15.911$ | $+3.553$ | $6.7\times$ |
| mean | $+13.348 \pm 1.819$ | $+0.353 \pm 1.975$ | $20.9\times$ |

**T3 — Cross-model persistence.**

| Seed | $\cos(M_A, M_{AB})$ | $\cos(M_A, \text{held-mAB})$ |
|---|---|---|
| 1234 | $+0.975$ | $+0.977$ |
| 1236 | $+0.854$ | $+0.862$ |
| 1238 | $+0.954$ | $+0.928$ |
| 1239 | $+0.975$ | $+0.955$ |
| 1240 | $+0.790$ | $+0.761$ |
| mean | $+0.910 \pm 0.084$ | $+0.897 \pm 0.087$ |

# D Alternative Explanations Table

# E Mechanistic Discrimination — Per-Seed Results

**Natural scores and projections ($\alpha = 1.0$).**

| Seed | $A_{\text{nat}}$ | $B_{\text{nat}}$ | proj$_{e_A}$ | proj$_{e_B}$ | proj$_{e_{\text{sh}}}$ |
|---|---|---|---|---|---|
| 1234 | $-9.705$ | $+12.080$ | 0.840 | 4.103 | $\approx 0$ |
| 1235 | $-9.614$ | $+11.570$ | 1.105 | 4.241 | $\approx 0$ |
| 1236 | $-8.300$ | $+11.685$ | 1.940 | 4.265 | $\approx 0$ |
| 1237 | $-9.118$ | $+11.137$ | 1.249 | 4.250 | $\approx 0$ |
| 1238 | $-8.832$ | $+11.161$ | 0.739 | 4.439 | $\approx 0$ |
| 1239 | $-9.486$ | $+11.988$ | 0.189 | 4.032 | $\approx 0$ |
| 1240 | $-9.254$ | $+12.621$ | 0.932 | 4.482 | $\approx 0$ |

Table 9: Alternative explanations addressed and the result that speaks to each.

| Alternative explanation | Result | Evidence |
|---|---|---|
| Historical component gone immediately after B-training | Rejected | $t_{\text{flip}} \ll$ any observed erosion timepoint |
| $J_A$ dormant in natural B-behavior forward passes | Rejected | $h \cdot \hat{J}_A = -2.56$ ($t = -58.6$, $p < 10^{-10}$); 88% of margin (Section 6) |
| $J_A$ effect requires external reactivation | Rejected | Linear dose-response from $\lambda = 0$; no gating (Section 6.1) |
| Orthogonal null rescues B output equally | Rejected | Ctrl $\Delta = +0.120 \pm 0.346 \approx 0$ (Section 6.1) |
| Generic large-perturbation artifact | Rejected | Matched-norm random directions far weaker |
| Any context-dependent direction transfers | Rejected | `fenn` fails completely, 0/7 |
| Generic red-vs-blue separability | Rejected | D5, D6 (no readout): near-zero effect (Section 7.1) |
| Causal claim direction-specific, not subspace-level | Rejected | 0-in-200 random rotations at every $k$ (Section 7.2) |
| Whole hidden state needed | Rejected | Parallel-only suffices (7/7); orthogonal-only fails |
| Phenomenon needs linear readout | Rejected | Nonlinear (2-layer ReLU) readout reproduces, 7/7 |
| Jacobian construction arbitrary | Rejected | D1=D2=D3 ($\lvert\cos\rvert = 1.000$); D5–D6 null (Section 7.1) |
| Dissociation artifact of small/synthetic setting | Rejected | Holds across 10 configs + Transformers (Section 10) |
| A-history shared by any B-trained model | Rejected | $M_B^{\text{scr}}$: $J_A$ ablation near-zero ($+0.007 \pm 0.125$) (Section 10.1) |
| B routes around A (gradients orthogonal to $J_A$) | Rejected | Gradient alignment $+0.958 \pm 0.031$ throughout (Section 10.1) |
| A and B components are not separable under matched-norm ablation | Rejected | $e_A$ and $e_B$ functionally separable ($p < 10^{-5}$); $e_{\text{sh}}$ causally inert (Section 12) |
| Shared component carries causal weight for either task | Rejected | $e_{\text{sh}}$ NS at all $\alpha$ (T-tests $p > 0.26$); dose-response confirms (Section 12) |
| Effect is explained by fc2 geometry alone (no upstream substrate) | Rejected | SS1: fc2-free fc1 direction has causal effect, lineage-specific (353× spec) (Section 11.2) |
| Scratch model's hidden state readable by $M_{AB}$ readout | Rejected | SS2: scratch $h$ through $M_{AB}$ fc2 $\approx 0$; NIE gap $+8.43$ ($t = 13.58$, $p < 10^{-5}$) (**??**) |
| Zor-lineage effect requires a readout that has seen zor | Rejected | SS3: foreign filler readout (no zor history, permuted labels) still detects gap ($t = 4.37$, $p = 0.005$) (**??**) |
| Effect requires readout access for upstream detection | Rejected | $d_{\text{src}}$ (T1): AUC$= 1.000$ on held-out fillers with no readout, no `zor` (Section 11.1) |
| Upstream structure absent in $M_{AB}$ (only fc2 geometry) | Rejected | $d_{\text{src}}$ T2: causal steering $+13.35$ vs $+0.35$ ($p = 10^{-6}$); T3: $\cos = 0.910$ (Section 11.1) |
| Any lineage's component works equally | Rejected | $7 \times 7$ matrix, $p < 2.4 \times 10^{-8}$ |
| Transfer is purely identity-specific | Rejected | Similarity explains ≈79% |
| Transfer is purely similarity-driven | Rejected | Identity residual replicates, 7/7, $p = 0.016$ |
| Specific to one canonical input's own gradient | Rejected | 140/140 held-out cases |
| Independent circuits occupy separable subspaces | Rejected | Independently-sourced directions interfere by ≈one-third |
| Accessibility affects future learning | Rejected | No replicable directional effect |
| Old circuit intact, merely gated (not unique interpretation) | Weakened | Counterfactual matrix; readout co-adaptation; other distributed-dependence explanations not ruled out |

**Ablation effects ($\Delta A$, $\Delta B$) per direction per seed.**

| Seed | $e_A : \Delta A$ | $e_A : \Delta B$ | $e_B : \Delta A$ | $e_B : \Delta B$ | $e_{\text{sh}} : \Delta A$ | $e_{\text{sh}} : \Delta B$ | $e_\perp : \Delta A$ | $e_\perp : \Delta B$ |
|---|---|---|---|---|---|---|---|---|
| 1234 | $-2.128$ | $+1.979$ | $+2.840$ | $-3.356$ | $+0.215$ | $-0.116$ | $+0.044$ | $-0.011$ |
| 1235 | $-2.027$ | $+2.011$ | $+2.705$ | $-3.138$ | $-0.148$ | $+0.162$ | $-0.450$ | $+0.462$ |
| 1236 | $-1.796$ | $+1.738$ | $+2.777$ | $-3.526$ | $+0.069$ | $+0.163$ | $-0.060$ | $-0.083$ |
| 1237 | $-1.865$ | $+1.882$ | $+2.614$ | $-3.079$ | $+0.140$ | $-0.176$ | $+0.445$ | $-0.302$ |
| 1238 | $-2.183$ | $+2.122$ | $+2.297$ | $-2.784$ | $+0.298$ | $-0.309$ | $-0.045$ | $+0.174$ |
| 1239 | $-2.378$ | $+2.179$ | $+2.518$ | $-3.091$ | $+0.241$ | $-0.280$ | $+0.376$ | $-0.464$ |
| 1240 | $-2.135$ | $+1.907$ | $+2.538$ | $-3.211$ | $-0.188$ | $+0.330$ | $+0.057$ | $+0.095$ |
| Mean | $-2.073$ | $+1.974$ | $+2.613$ | $-3.169$ | $+0.089$ | $-0.032$ | $+0.052$ | $-0.019$ |

**Dose-response across $\alpha$ (mean across 7 seeds).**

| $\alpha$ | $e_A : \Delta A$ | $e_A : \Delta B$ | $e_B : \Delta A$ | $e_B : \Delta B$ | $e_{\text{sh}} : \Delta A$ | $e_{\text{sh}} : \Delta B$ |
|---|---|---|---|---|---|---|
| 0.5 | $-1.037$ | $+0.987$ | $+1.307$ | $-1.585$ | $+0.044$ | $-0.016$ |
| 1.0 | $-2.073$ | $+1.974$ | $+2.613$ | $-3.169$ | $+0.089$ | $-0.032$ |
| 2.0 | $-4.146$ | $+3.948$ | $+5.227$ | $-6.338$ | $+0.178$ | $-0.064$ |
| 4.0 | $-8.293$ | $+7.896$ | $+10.453$ | $-12.677$ | $+0.355$ | $-0.128$ |

# F Additional Generality Results

**Logic task.** The boolean-function task uses 12 filler markers, each with a fixed function drawn from a small bank (AND, OR, NOR, and-not, not-a, project-b), plus the special marker. Phase A trains the special marker to compute XOR of two random binary inputs; phase B overwrites this to AND. Training uses 800 phase-A steps and 3000 phase-B steps, otherwise matching the color task's optimizer settings. The canonical test point $(b_1, b_2) = (1, 0)$ is used for $t_{\text{flip}}$ and $C_A(t)$, since $\text{XOR}(1, 0) \neq \text{AND}(1, 0)$.

**Architecture sweep.** The transformer variant embeds the marker and two bit values as three tokens with learned positional embeddings, applies one multi-head self-attention layer (2 heads, 16-dim embeddings) followed by layer normalization and the same feed-forward readout structure as the MLP variants. Its phase-A accuracy plateaus near 0.96 regardless of learning rate (tested at $0.003$, $0.01$, $0.03$) or training length (up to 3000 steps), which we diagnose as a genuine capacity/optimization limitation of this minimal single-layer configuration at this task scale rather than a bug; seeds failing to reach 0.90 phase-A accuracy with the correct XOR behavior are excluded from the dissociation-rate calculation reported in Section 8.