WHEN CAN ONE TRANSFORMER USE ANOTHER TRANSFORMER'S HIDDEN STATE?

Anonymous authors
Paper under double-blind review

ABSTRACT

When can separately optimized transformers directly use one another's hidden states without learned alignment? We study *raw interoperability*: whether a frozen recipient can causally use an independently learned donor state under the identity map. In controlled retrieval tasks, independently optimized models sharing externally specified coordinate scaffolds exhibit robust cross-model transplantation, including on a held-out two-hop compositional task. Decoder-subspace analysis shows that decoder-visible state carries substantial but incomplete transferable signal, while scaffold interventions identify shared output coordinates as the strongest tested anchor; recipient-specific decoder-null structure remains consequential. When output coordinates differ, identity-map transfer nearly disappears, whereas applying the exact known coordinate correspondence restores nearly all of the lost advantage as mapped compatibility. In a complementary boundary test, independently pretrained Pythia-70M seeds show no reliable cross-seed identity-map transfer at the tested interfaces, including a site subsequently verified in a post hoc diagnostic to be causally informative within individual models. These results identify raw cross-model interoperability as a scaffold-dependent interface property: it can emerge across independently optimized models, but depends on shared coordinate conventions and should not be assumed without direct testing in independently pretrained networks.

1 INTRODUCTION

Two neural networks can implement the same function while organizing their hidden states differently (Godfrey et al., 2022). Most work on representation comparison and model stitching therefore asks whether independently learned states are similar, encode the same information, or can be mapped into one another; broader convergence hypotheses propose that capable models may approach shared representational structure across domains and modalities (Huh et al., 2024). We ask a stricter question: *under what conditions can separately optimized models develop hidden-state interfaces that are directly usable across models without learning an alignment map?* We call this *raw interoperability*: direct causal transfer under the identity map. Our goal is to identify controlled conditions under which that identity map becomes usable, not to argue that independently trained networks generally converge to a universal latent representation. Here, "raw" refers to transfer at evaluation time; the models may still share externally specified coordinate conventions during training.

This is a causal rather than merely geometric notion. We use *representational similarity* for descriptive correspondence, *mapped compatibility* for successful transfer after an explicit known or learned transformation, and *raw interoperability* when the identity map itself suffices. The distinction lets us ask which shared coordinate conventions support a directly usable interface and which apparent compatibilities disappear when those conventions change.

We test this by running a recipient on a corrupted input and replacing one hidden state with the corresponding clean state from an independently initialized donor (Figure 1). No alignment, stitching layer, probe, or post hoc transformation is applied; the recipient continues its own downstream computation. Models are independently initialized and optimized, while token, sequence, and output coordinates are externally controlled, separating optimization from the shared coordinate scaffold.

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.

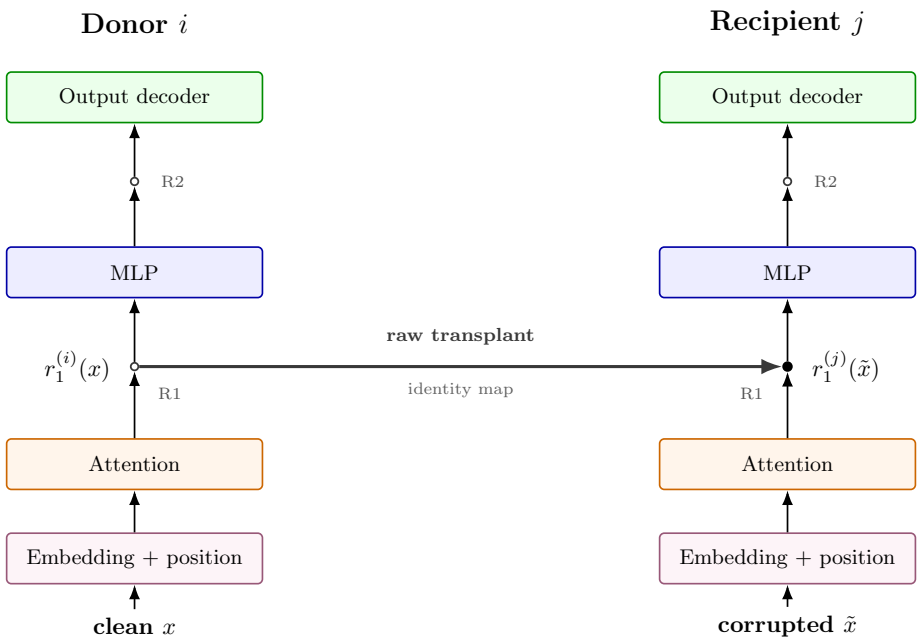

Figure 1: Cross-model causal transplantation. A clean hidden state from one model replaces the corresponding state in another model running on a corrupted input. Here, R1 and R2 are interface labels for the post-attention and post-MLP residual states, respectively. At interface R1, the donor residual vector $r_1^{(i)}(x)$ replaces the corresponding recipient vector $r_1^{(j)}(\tilde{x})$, with no learned alignment or projection. The recipient's downstream computation then continues from the transplanted state. The corrupted-donor and same-model states provide the matched control and causal-sufficiency check, respectively.

We therefore ask not only whether raw interoperability occurs, but which scaffold components support it and which parts of the transferred state remain recipient-specific. A shared output decoder provides common decoder-visible coordinates, but does not by itself imply whole-state compatibility: decoder-null directions can remain consequential under the recipient's downstream computation. We test this directly with decoder-subspace interventions, scaffold ablations, and exact coordinate changes. Together, these experiments distinguish scaffold-induced raw interoperability from mapped compatibility and from stronger claims of coordinate-independent convergence.

Contributions. We address four questions. **(1) Existence:** In controlled retrieval tasks, independently optimized models develop raw-interoperable hidden states under shared external coordinate scaffolds, including on held-out two-hop relational compositions. **(2) Localization and mechanism:** Interoperability becomes query-local later in computation, and a substantial fraction of transfer is carried by decoder-visible directions; controlled scaffold interventions show that sharing output coordinates has the largest effect among the tested scaffold factors, while changing input-embedding coordinates largely preserves transfer. **(3) Coordinate dependence:** Changing output-coordinate assignments nearly eliminates identity-map transfer, while an exact known coordinate correction restores mapped compatibility; exact function-preserving residual-basis changes show the same coordinate dependence. **(4) Boundary:** Pythia-70M boundary tests find no evidence of cross-seed raw interoperability at the tested interfaces, including a site later shown by a post hoc same-model diagnostic to be causally effective within individual models. Together, these experiments identify conditions under which the identity map becomes usable across models and show that this compatibility is scaffold-induced and coordinate-relative.

2 Related Work

Representation comparison and alignment. SVCCA and CKA compare representation geometry across models (Raghu et al., 2017; Kornblith et al., 2019). But geometric agreement does not by itself establish that one model can causally use another's state: representational and functional similarity can dissociate (Braun et al., 2025; Smith et al., 2025), while learned stitching maps can compensate for otherwise incompatible internal representations (Smith et al., 2025). Raw interoperability asks a stricter question: whether the identity map already suffices for a frozen recipient to use an independently learned donor state.

Cross-model causal transfer. Model stitching and related methods connect representations through learned or prescribed transformations (Lenc and Vedaldi, 2015; Bansal et al., 2021; García-Castellanos et al., 2024; Athanasiadis et al., 2025; Piepereit, 2026); Model Alignment Search learns invertible maps that expose causally interchangeable subspaces across different-seed networks (Grant, 2025). Activation patching has transferred states across base and fine-tuned models (Prakash et al., 2024), and Salhan, Zhou, and Buttery (2026) report cross-lingual activation patching between independently trained monolingual models without a learned projection. A subsequent study with overlapping authors fits an explicit Procrustes rotation and obtains causal transfer after alignment (Zhou et al., 2026). Conversely, a recent preregistered audit of independently trained latent-communication systems reports broad cross-initialization failure across raw and increasingly flexible frozen alignment tests (Marincat, 2026). In a distinct Pythia multi-hop setting, learned linear cross-model translations can achieve strong representational alignment without yielding useful causal activation transfer (Zhang and Xin, 2026). Together, these results show both that cross-model causal transfer can occur without explicit alignment in some settings and that representational correspondence or learned alignment need not make raw states directly usable across models. Our focus is therefore not the first demonstration of unmapped cross-model transfer, but the coordinate conditions that support it: we intervene separately on input and output scaffolds, apply exact function-preserving gauge changes, test restoration by the known map, and include a same-architecture pretrained Pythia boundary case. These comparisons separate scaffold-induced raw interoperability from mapped compatibility and from the stronger claim that independently optimized models converge to a coordinate-independent latent language.

Compatible representations and output scaffolds. Prior work shows that compatibility can be encouraged by training structure. Shared auxiliary objectives can make independently trained network components directly compatible (Gygli et al., 2020), and fixed d-Simplex classifiers can induce stationary representations with formal compatibility guarantees (Biondi et al., 2026). We ask a complementary causal question: which shared coordinate components support raw interoperability? By independently varying input-embedding and output coordinates, we test whether compatibility depends specifically on the output-facing scaffold rather than on shared coordinates in general.

Coordinate symmetries and gauge dependence. Function-preserving parameter symmetries motivate analyses of gauge freedom in representation measurement (Godfrey et al., 2022; Cain, 2026; Karasan, 2026). These symmetries provide a sharper test than ordinary similarity comparisons: an exact residual-basis change can preserve the model's function while changing the coordinates of its hidden states. We use this fact causally, asking whether raw interoperability survives such a transformation and whether applying the known coordinate map restores transfer. Because ordinary Adam is not rotation-equivariant (Ling et al., 2022; Zhang et al., 2025), we separately distinguish exact post-training reparameterization from training independently in different gauges. Taken together, prior work motivates both the possibility of cross-model latent transfer and the frequent need for alignment; our contribution is to isolate experimentally the coordinate conditions under which alignment-free identity-map transfer emerges and to test whether that compatibility is itself coordinate-relative.

3 EXPERIMENTAL FRAMEWORK

3.1 CAUSAL TRANSPLANTATION AND ESTIMANDS

For each independently trained donor–recipient pair, we run the recipient on a corrupted input and replace a fixed, architecture-matched internal state with the corresponding state from the donor's clean computation, following the activation-restoration logic of causal tracing and activation patching (Meng et al., 2022). The recipient's parameters remain frozen; only the selected hidden state is replaced. We compare this patched output with the recipient's clean and corrupted baselines.

Let $p_{\mathrm{clean},j,x}$ and $p_{\mathrm{corr},j,x}$ denote recipient j's target probability on clean input x and corrupted input $\tilde{x}$. For condition $c \in \{\mathrm{raw}, \mathrm{ctrl}\}$, let $p^{c}_{\mathrm{patch},i,j,x}$ denote the target probability after receiving donor i's clean state ($c = \mathrm{raw}$) or corrupted state ($c = \mathrm{ctrl}$). We measure condition-specific normalized recovery as

$$C^{c}_{i,j,x} = \frac{p^{c}_{\mathrm{patch},i,j,x} - p_{\mathrm{corr},j,x}}{p_{\mathrm{clean},j,x} - p_{\mathrm{corr},j,x}}, \qquad c \in \{\mathrm{raw}, \mathrm{ctrl}\}, \tag{1}$$

where 0 means that transplantation provides no rescue and 1 means that it restores the recipient's clean target probability. Normalized recovery is not clipped to $[0, 1]$: values below 0 indicate that the intervention moves the target probability farther from the clean baseline than the corrupted run, while values above 1 indicate that it exceeds the clean-baseline target probability. When normalized recovery is used as the endpoint, required denominators are finite and positive, and no rows or donor–recipient pairs are discarded.

Two checks support interpretation of cross-model recovery, but their validation role is study-specific. In controlled studies that use same-model sufficiency as an interface-admission criterion, a candidate interface must first pass a *same-model causal-sufficiency test*: replacing the corrupted state with that model's own clean state must rescue the computation. Across studies, we also compare clean-donor transplantation with the identical intervention using the donor's corrupted state, providing a matched control that lacks the clean task information.

Let $\mathcal{X}$ denote the evaluation-case set and define $\bar{C}^{c}_{ij} = |\mathcal{X}|^{-1} \sum_{x \in \mathcal{X}} C^{c}_{i,j,x}$. We average $\bar{C}^{c}_{ij}$ over ordered off-diagonal pairs (2,450 for 50 models). Because pairs share model identities, we use a model-node bootstrap: resample identities with replacement; if w_i is model i's multiplicity, weight pair (i, j) by $w_i w_j$ and recompute the pair mean, introducing no self-pairs. This treats models, not pair rows, as exchangeable units and follows jointly exchangeable dyadic bootstrap procedures (Menzel, 2021). Positive-edge density is the fraction with $\bar{C}^{\mathrm{raw}}_{ij} \geq 0.50$.

Evaluation chronology. Whenever a result is described as held out, its evaluation population or split is separated from the feasibility/discovery work used to choose the test. When interface selection is part of a controlled discovery stage, it is completed before evaluation on the corresponding held-out model population; studies that reuse a previously fixed interface or use a different frozen admission rule follow their study-specific protocol. Applicable interfaces, endpoints, seed blocks, and decision rules were fixed before inspecting those outcomes; analyses introduced only after outcome inspection are labeled post hoc. This is within-project temporal separation, not external preregistration. Appendix B.3 records the chronology and fixed choices.

3.2 CONTROLLED EXPERIMENTAL REGIMES

We use two primary controlled regimes and one compositional bridge task. The first primary regime is a one-block causal transformer trained on four two-token subject–relation mappings with relation-flip corruption. It has $d_{\mathrm{model}} = 4$, two attention heads, a width-2 MLP, fixed token embeddings, and the transposed embedding as decoder; additional width and depth conditions test architectural robustness. Full mappings, training details, and validity criteria are given in Appendix A.

Our main regime is associative retrieval with key vocabulary $\mathcal{K} = \{\kappa_0, \kappa_1, \kappa_2\}$ and value vocabulary $\mathcal{V} = \{\nu_0, \ldots, \nu_3\}$. Inputs have the form $[\kappa_a, \nu_u, \kappa_b, \nu_v, \kappa_\ell]$, with $a \neq b$ and $\ell \in \{a, b\}$; the model must return the value bound to queried key κ_ℓ. These choices produce 192 semantic cases, and corruption changes only the value paired with the queried key. Models are two-block transformers with $d_{\mathrm{model}} = 16$, two heads, $d_{\mathrm{head}} = 8$, $d_{\mathrm{ff}} = 8$, a fixed orthogonal token basis, and a fixed answer decoder shared across seeds. This regime supports the recurrence, localization, decoder-subspace,

scaffold-intervention, and gauge analyses. A separate two-hop compositional bridge task uses the same two-block dimensions with a task-specific fixed entity embedding/answer-decoder scaffold and withholds complete two-edge compositions while retaining their constituent atomic edges in training; it is used to test whether transfer extends to unseen relational compositions (Appendix F).

4 Raw Interoperability: Emergence, Recurrence, and Localization

4.1 Minimal demonstration of existence: mapping task

We first use the smallest controlled mapping task as a demonstration of existence. An early candidate interface fails same-model sufficiency, whereas a later post-attention residual restores the clean computation perfectly. At this eligible interface, a held-out 50-model population shows substantial cross-model transfer: clean-donor recovery is 0.640 versus 0.066 for the corrupted-donor control; discovery is similar (0.586 versus 0.089), and width/depth variants retain the effect (Appendix C). Pairwise recovery is broadly distributed, with modest systematic donor/recipient structure analyzed in Appendix D.

Because the recipient's independently optimized downstream MLP processes the transplanted state before decoding, this minimal experiment establishes that a shared coordinate scaffold can support a directly usable hidden-state interface across independently trained models.

4.2 Main controlled regime: associative retrieval

The mapping task establishes that raw interoperability can occur, but it is intentionally minimal. We therefore use associative retrieval as the main controlled regime, where the interface must support relational binding and can be tested for localization and mechanism. We intervene at $\mathrm{B2.R1[QUERY]}$, the query-position post-attention residual of block 2 immediately before the recipient's final MLP. Every validation model achieves same-model recovery 1.0, confirming that this interface is causally sufficient.

At this eligible interface, a separate held-out 50-model population gives clean-donor recovery 0.746 versus 0.086 control, an advantage of 0.660 (95% model-node bootstrap CI $[0.612, 0.708]$); 95.3% of ordered donor–recipient pairs have mean recovery at least 0.50. Discovery is similar (0.780 versus 0.070 control; endpoint robustness in Appendix E).

4.3 Held-out two-hop compositional bridge test

The associative-retrieval result still leaves open whether identity-map transfer depends on complete relational configurations seen during training or extends to novel compositions assembled from familiar atomic relations. We therefore test a separate two-hop compositional retrieval task in which each example specifies $E_a \to E_b$ and $E_b \to E_c$, queries E_a, and predicts E_c. A fixed split reserves 48 of the 192 semantic compositions entirely for evaluation and trains on the remaining 144, while ensuring that the constituent atomic edges of held-out compositions remain represented in training. Because a small fraction of recipient–case clean–corrupted target-probability gaps are nonpositive, normalized recovery is not a valid all-row endpoint for this task; we therefore use the raw-minus-control target-probability contrast fixed before evaluation (Appendix F).

In a separately trained population of 50 models, all 50 models generalize behaviorally to the withheld compositions, with mean accuracy 0.973. On the same held-out cases, clean donor states increase the recipient's target probability by 0.402 relative to corrupted-donor controls (95% model-node bootstrap CI $[0.378, 0.428]$), with a positive effect for every ordered donor–recipient pair. Thus, in this controlled compositional setting, raw interoperability is not restricted to complete relational compositions seen during training.

Associative retrieval thus provides the main controlled setting in which raw interoperability recurs, while the separate compositional bridge test shows transfer on held-out relational compositions under a matched small-transformer scaffold. We now use associative retrieval to localize where a compact interoperable interface emerges.

4.4 LOCALIZING A COMPACT INTEROPERABLE INTERFACE

We localize when raw interoperability becomes available in a compact, query-local state, rather than when transferable information first appears anywhere in the network. At an early interface, replacing only the query position with the model's own clean state does not reliably rescue the corrupted computation, so this query-local interface fails the same-model causal-sufficiency test. The corresponding cross-model query-local intervention is therefore not interpretable as a test of interoperability.

The failure is not because the necessary information is absent from the network at this stage. Replacing the *full sequence* at the same early interface with the model's own clean sequence restores the computation, and cross-model full-sequence transplantation also produces substantial recovery. Thus task-relevant transferable information is already present, but remains distributed across positions rather than concentrated in the query representation.

Later in the computation, this changes. At the post-attention residual of the final block, query-local same-model transplantation restores the clean computation, and raw cross-model query-local transplantation yields strong recovery. The model has formed a compact query-position state that is sufficient for independently learned downstream computation in another model.

These interventions distinguish two notions that would be conflated by a simple layerwise similarity analysis: the presence of transferable information and its concentration into a locally sufficient causal interface. In associative retrieval, interoperability does not simply appear abruptly at a late layer; rather, transferable information is available earlier in a distributed form and becomes query-local only later (Appendix G).

5 MECHANISM: DECODER-VISIBLE STATE AND OUTPUT SCAFFOLDS

Having localized a compact late interface, we ask two mechanistic questions in sequence. First, where within the residual state is the transferable signal concentrated? Because the decoder is fixed across seeds, its row space defines a privileged output-facing subspace shared by all models. If transfer is concentrated there, preserving the donor's decoder-visible component should reproduce a substantial fraction of full-state transfer while matching its immediate decoder logits. Second, does sharing those output coordinates during training help induce the compatibility itself? The decoder-subspace decomposition addresses the first question; the scaffold interventions below address the second.

At fixed $\mathrm{B2.R1[QUERY]}$, suppress the interface index: $r_{i,x}$ and $\tilde{r}_{j,x}$ denote donor i's clean and recipient j's corrupted residuals on example x. Let $D \in \mathbb{R}^{4\times 16}$ be the fixed rank-4 decoder, with $\Pi_D = D^\top (DD^\top)^{-1} D$ the projector onto its row space and $\Pi_D^\perp = I_{16} - \Pi_D$ its complementary projector. We construct

$$s_{i,j,x} = r_{i,x}\Pi_D + \tilde{r}_{j,x}\Pi_D^\perp. \tag{2}$$

This surrogate preserves the donor's decoder-visible component while replacing everything invisible to the decoder with the corrupted recipient's null component. It therefore produces exactly the same immediate decoder logits as the clean donor; we verify this equality across all evaluated states, with zero maximum float64 logit error.

The decoder-visible component carries a substantial but incomplete fraction of raw transfer. The real donor achieves recovery 0.746, whereas the decoder-matched surrogate achieves 0.489, compared with 0.086 for the corrupted-donor control. The surrogate retains 0.611 of the donor's transfer advantage (95% model-node bootstrap CI $[0.576, 0.645]$). Thus decoder-visible directions carry substantial transferable signal, but are not sufficient to reproduce the full donor effect.

The nullspace interventions show that the donor's own decoder-null state is not required for successful transfer. Zeroing it increases recovery to 0.834, while pairing the donor-visible component with the *clean recipient's* null component reaches 0.916, above the full donor's 0.746; using the corrupted recipient's null instead gives only 0.489. The nullspace hybrids are post hoc mechanistic diagnostics, not realistic transfer protocols: the clean-recipient hybrid's improvement over the full donor indicates that recipient-specific nullspace state can aid downstream use and that transplanting the donor's nullspace can itself be mismatched even when the donor-visible component is compatible (Appendix H).

These results argue against interpreting the effect as whole-state convergence. Under shared-decoder training, raw-transferable signal is concentrated in the decoder-visible component, while decoder-null directions remain recipient-specific. We therefore interpret the result as *partial interoperability anchored by the shared scaffold*, not evidence that independently trained models learn the same residual vector. The decomposition answers where much of the transferable effect resides, but not why compatible directions emerged: it cannot distinguish whether the shared decoder merely exposes an already compatible subspace or helps induce compatible coordinates during training. We therefore intervene directly on the training scaffold.

5.1 Shared output coordinates provide the strongest scaffold support among those tested

The associative-retrieval experiments above deliberately share a fixed embedding and answer decoder across seeds. To distinguish whether raw interoperability depends on shared input coordinates, shared output coordinates, or both, we compare conditions spanning these two coordinate factors. The study was executed sequentially rather than as one jointly specified 2×2 design: the A/B comparison was fixed before inspecting outcomes, and C/D were specified only afterward to test the input-coordinate alternative. In the primary A/B comparison, both arms use the same task, architecture, corruption, $\mathrm{B2.R1[QUERY]}$ interface, behavioral-validity gates, evaluation estimands, and held-out evaluation protocol. Arm A shares both embedding and decoder coordinates across models. Arm B keeps the shared embedding but assigns each model an independent decoder coordinate scaffold using deterministic signed permutations of the canonical coordinate system. This preserves dimension, norms, rank, sparsity, and task semantics while changing which residual coordinates are directly tied to output coordinates. Each condition contains 50 independently trained, behaviorally valid models.

All 50 models in each arm are behaviorally valid, and every model has same-model recovery 1.0 at the patched interface. Arm A reproduces strong raw interoperability: mean raw recovery is 0.760 versus 0.076 control, for raw-control advantage 0.684, with positive-edge density 0.971. Arm B behaves very differently. Mean raw recovery is 0.255 versus 0.252 control, for advantage 0.003, with positive-edge density 0.075.

For arm $k \in \{A, B, C, D\}$, let $\bar{C}_k^c$ be its population-mean recovery and $\Delta_k = \bar{C}_k^{\mathrm{raw}} - \bar{C}_k^{\mathrm{ctrl}}$. Define retained advantage $R_k = \Delta_k / \Delta_A$ for $k \in \{B, C, D\}$; the prespecified A/B contrast is R_B. The observed value is $R_B = 0.004$ with a jointly bootstrapped 95% CI of $[-0.023, 0.033]$; the absolute difference $\Delta_B - \Delta_A = -0.681$ has 95% CI $[-0.726, -0.636]$. The near-zero retained advantage shows that raw interoperability collapses when output coordinates differ.

The separately specified C/D follow-up tests the remaining alternative: whether shared input coordinates are necessary for the A/B effect. Arm C gives each model an independent embedding coordinate scaffold while keeping the decoder shared; Arm D gives each model both an independent embedding and an independent decoder. C/D were fixed after observing A/B but before inspecting any C/D outcomes, and use the same interface, corruption, behavioral gates, and 50-model evaluation size with disjoint held-out seeds. Both follow-up arms again achieve 50/50 behavioral validity and same-model recovery 1.0.

Arm C preserves strong interoperability despite independent input coordinates: mean raw recovery is 0.734 versus 0.090 control, giving advantage 0.644 and positive-edge density 0.939. Relative to the shared-embedding/shared-decoder condition, retained advantage is $R_C = 0.941$ with 95% CI $[0.856, 1.030]$; the absolute difference $\Delta_C - \Delta_A = -0.040$ has 95% CI $[-0.102, 0.020]$. Arm D instead collapses to control level: mean raw recovery is 0.246 versus 0.251 control, advantage -0.005, and positive-edge density 0.061. Its retained advantage relative to the shared-embedding/shared-decoder condition is $R_D = -0.008$ with 95% CI $[-0.033, 0.022]$.

Taken together, the four conditions show that raw transfer tracks output-coordinate sharing: changing input-embedding coordinates leaves most of the transfer advantage intact when decoder coordinates are shared, whereas changing decoder coordinates collapses raw transfer whether embeddings are shared or independent. The complementary conditions are interpreted through their bootstrap intervals rather than a new categorical threshold. Table 3 in Appendix I reports the complete numerical results. For completeness, the four-arm pattern yields descriptive factorial contrasts of 0.665 for decoder sharing, 0.024 for embedding sharing, and 0.032 for the embedding-by-decoder interaction.

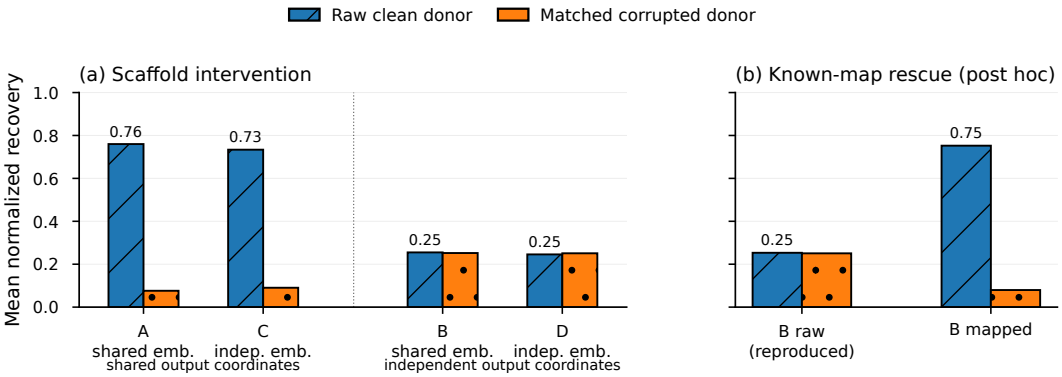

Figure 2: **Raw interoperability tracks output-coordinate sharing.** **(a)** Raw clean-donor and matched corrupted-donor recovery under the four scaffold interventions. Raw transfer remains high when input embedding coordinates differ but output coordinates are shared (A,C), whereas it falls to the matched-control level when output coordinates are independent (B,D). **(b) Post hoc known-map diagnostic.** Reproduced Arm B shows negligible raw advantage in mismatched output coordinates; applying the exact known donor-to-recipient coordinate map before transplantation restores recovery to the shared-coordinate regime. Values above blue bars denote mean raw normalized recovery.

These descriptive contrasts reinforce the direct arm comparisons: among the tested components, output-coordinate sharing has by far the largest effect, while changing input coordinates has little effect in this tested setting. We therefore identify the output scaffold as the strongest tested coordinate anchor, without claiming it is the unique source of compatibility.

Panel (b) of Figure 2 shows a separate post hoc exact-map diagnostic, added after observing A/B, testing whether the Arm B deficit could be explained by the known output-coordinate mismatch. For donor state h_i, with donor and recipient signed-permutation scaffolds Q_i, Q_j, respectively, we applied the exact known map $h_i Q_i^\top Q_j$, fitting no alignment parameters. Corrected Arm B advantage is 0.673 with positive-edge density 0.957, retaining 0.984 of frozen Arm A (95% CI $[0.903, 1.072]$). Thus almost all of the Arm B deficit is recoverable after correcting the known coordinate mismatch. Because this diagnostic was designed after observing A/B, it is not confirmatory evidence. The near-complete rescue is consistent with the Arm B deficit being largely attributable to the known coordinate mismatch. Section 6 tests coordinate dependence more directly with exact residual-basis changes; before that, we ask whether the scaffold dependence persists under increased controlled scale.

5.2 Controlled-scale stress test of scaffold dependence

The 16-dimensional scaffold intervention leaves open whether its decoder-sharing contrast survives greater controlled model and task scale. We therefore stress-test it in an eight-layer, 1.6M-parameter associative-retrieval model over 8,192 keys and values, using the same fixed zero-parameter content-addressing router in both decoder arms; nine of ten held-out models pass the behavioral gate in each arm (Appendix J). The prespecified normalized endpoint is positive but ill-conditioned because a small tail has near-zero clean–corrupt gaps, so we report it only for transparency. An explicitly post hoc, numerically stable raw-minus-control target-probability effect is 0.973 with shared decoders versus 0.043 with independent decoders (difference 0.930). With routing fixed across arms, this supports robustness of the output-scaffold effect at greater controlled scale, but not unrestricted scaling or spontaneous interoperability.

6 Coordinate Dependence and Boundary Tests

The scaffold interventions and scale stress test suggest that shared coordinate conventions matter, but they compare differently trained model populations, so training history and coordinates vary together. To isolate coordinates while holding the recipient's input–output function fixed, we next

apply an exact function-preserving change of residual coordinates. We then ask separately what happens when independent optimization proceeds from scratch in different gauges.

6.1 Exact coordinate dependence

We transform a trained recipient by a random orthogonal residual-basis change absorbed exactly into adjacent weights. Because the controlled architecture has no LayerNorm, input–output behavior is unchanged while the residual stream is re-expressed. Before transfer, we verify functional equivalence and exact recovery under the analytically known donor-to-recipient map. Across 10 orthogonal gauges, raw transfer retains a median 0.231 of canonical donor advantage, with all 10 at or below one half; the known map restores canonical recovery to numerical precision. Thus changing coordinates alone can break the identity-map interface without changing function, while restoration by the known map shows that mapped compatibility remains intact (Appendix K).

6.2 Training under different gauges

The exact transformation is a post-training intervention: it establishes coordinate dependence at fixed function, but does not determine what happens when optimization itself proceeds in different gauges. We therefore train associative-retrieval populations from scratch in five orthogonal gauges, rotating the fixed embedding and decoder consistently. Because Adam is not rotation-equivariant (Ling et al., 2022; Zhang et al., 2025), this is an empirical convergence test rather than a symmetry theorem. Within-gauge transfer remains strong (pooled raw-advantage 95% model-node bootstrap CI $[0.875, 0.949]$), while across-gauge raw transfer largely disappears: median retained advantage is 0.002 from canonical to transformed models and -0.018 in reverse. The known map restores large causal effects in every sampled gauge (Appendix L); inference is conditional on the five sampled gauges.

6.3 Boundary in independently pretrained models

The controlled experiments above specify coordinate scaffolds directly; we finally ask whether comparable identity-map transfer appears without such an engineered scaffold, across independently pretrained Pythia-70M seeds (Biderman et al., 2023; van der Wal et al., 2025). At the prespecified final-checkpoint/layer-3 interface, the mean effect is -0.005 nats (95% model-node bootstrap CI $[-0.027, 0.017]$). A discovery-selected checkpoint/layer evaluated on six held-out seeds is also null, as is a prespecified target-conditioned rank-one hybrid (Appendix M). A post hoc same-model audit at the held-out site recovers 82.6% of the recipient clean–corrupt gap, so the null is not explained by an inactive interface. Thus we find no reliable identity-map transfer across the tested independently pretrained seeds and interfaces; this boundary result does not imply absence at other models, sites, or training regimes.

7 Discussion and limitations

Our results distinguish representation correspondence, mapped compatibility, and the stronger case in which the identity map itself is a usable interface. The controlled tasks establish this raw interoperability across independently optimized models, including on held-out two-hop compositions. Mechanistic and gauge interventions further indicate that it is scaffold- and coordinate-relative: decoder-visible directions carry substantial but incomplete transferable signal, shared output coordinates are the strongest tested scaffold anchor, and an exact function-preserving basis change reduces raw transfer while the known map restores mapped compatibility. An unchanged input-output function therefore need not imply direct cross-model usability.

These results do not imply that independently trained networks generally converge to a universal raw latent language. Rather, shared coordinate scaffolds can support directly interoperable interfaces; the tested Pythia-70M seeds show no reliable identity-map transfer even at a post hoc causally informative site. Raw interoperability should therefore be tested rather than assumed when such scaffolds are absent or unknown. Scope is limited to synthetic tasks, fixed routing in the scale stress test, one pretrained model family/size and selected interfaces, and post hoc exploratory diagnostics; broader settings remain future work.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing portions of the manuscript, including improvements to clarity, wording, framing, and presentation. AI tools were not used to autonomously conduct experiments, generate synthetic datasets, or prove mathematical claims. All AI-assisted text was reviewed and revised by the authors, who take responsibility for the final content of the submission.

ETHICS STATEMENT

This work studies controlled synthetic neural networks and publicly released pretrained language models, and uses no human subjects, personal data, or sensitive datasets. The natural-language experiments use WikiText-103 (Merity et al., 2017). We are not aware of direct ethical risks specific to the experiments beyond the general risk of overgeneralizing causal-compatibility results beyond the tested scaffolds, model families, and training regimes; the manuscript explicitly bounds its claims accordingly.

REPRODUCIBILITY STATEMENT

The main text specifies the transplantation estimands, causal controls, evaluation populations, and statistical aggregation procedures. The appendices provide task and training details (Appendix A), evaluation criteria and chronology, including the chronology summary of feasibility/discovery versus held-out or confirmatory evaluation, relevant populations, timing, and fixed endpoints or contrasts (Appendix B, especially Appendix B.3), recurrence and architecture robustness (Appendix C), additional mapping-task cross-seed diagnostics (Appendix D), endpoint robustness (Appendix E), withheld-relational-tuple evaluation (Appendix F), localization analyses (Appendix G), decoder-subspace controls (Appendix H), scaffold interventions (Appendix I), scaled-regime details (Appendix J), exact and independently trained gauge tests with numerical conformance checks (Appendices K–L), and Pythia boundary tests (Appendix M). The joint-sensitivity robustness analysis is implemented as a post hoc re-filtering of frozen per-example transplantation outputs; it performs no additional model evaluations and is analyzed separately from the prespecified held-out analyses. Together, the main text and appendices document the experimental settings, evaluation populations, statistical procedures, and analysis chronology needed to make the reported experiments and decisions traceable and reproducible.

REFERENCES

Ioannis Athanasiadis, Anmar Karmush, and Michael Felsberg. Model stitching by functional latent alignment. *arXiv preprint arXiv:2505.20142*, 2025.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, 34:225–236, 2021.

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

Niccolò Biondi, Federico Pernici, Simone Ricci, and Alberto Del Bimbo. A stationary (and therefore compatible) representation is all you need. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi:10.1109/TPAMI.2026.3692279.

Lukas Braun, Erin Grant, and Andrew M. Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:5355–5382, 2025.

Jericho Cain. Gauge freedom and metric dependence in neural representation spaces. *arXiv preprint arXiv:2603.06774*, 2026.

Alejandro García-Castellanos, Giovanni Luca Marchetti, Danica Kragic, and Martina Scolamiero. Relative representations: Topological and geometric perspectives. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, PMLR 285:219–231, 2024.

Charles Godfrey, Davis Brown, Tegan Emerson, and Henry Kvinge. On the symmetries of deep learning models and their internal representations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11893–11905, 2022.

Satchel Grant. Model alignment search. *arXiv preprint arXiv:2501.06164*, 2025.

Michael Gygli, Jasper Uijlings, and Vittorio Ferrari. Training neural networks to produce compatible features. In *CVPR Workshop on Compositionality in Computer Vision*, 2020.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:20617–20642, 2024.

Abdullah Karasan. Representation measurements under function-preserving reparameterizations. *arXiv preprint arXiv:2608.27020*, 2026.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:3519–3529, 2019.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–999, 2015.

Selena Zihan Ling, Nicholas Sharp, and Alec Jacobson. VectorAdam for rotation equivariant geometry optimization. In *Advances in Neural Information Processing Systems*, 35:4111–4122, 2022.

Narcis Marincat. Portable semantics, private dialects: Reuse and negative transfer in latent communication between language-model cells. *arXiv preprint arXiv:2609.11365*, 2026.

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

Konrad Menzel. Bootstrap with cluster-dependence in two or more dimensions. *Econometrica*, 89(5):2143–2188, 2021. doi:10.3982/ECTA15383.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Fernando Cardenas Piepereit. Architecture-dependent causal transfer of activation states across large language models. *arXiv preprint arXiv:2608.16347*, 2026.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, 30:6076–6085, 2017.

Suchir Salhan, EJ Zhou, and Paula Buttery. Do monolingual language models learn cross-lingual universal conceptual representations? In *ICLR 2026 Workshop on Representational Alignment (Re-Align)*, 2026.

Damian Smith, Harvey Mannering, and Antonia Marcu. Functional alignment can mislead: Examining model stitching. In *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:55972–55998, 2025.

Oskar van der Wal, Pietro Lesci, Max Müller-Eberstein, Naomi Saphra, Hailey Schoelkopf, Willem Zuidema, and Stella R. Biderman. PolyPythias: Stability and outliers across fifty language model pre-training runs. In *International Conference on Learning Representations*, 2025.

Peiyan Zhang and Jason Xin. A negative result on cross-model activation transfer in a Pythia multi-hop setting. *arXiv preprint arXiv:2606.03280*, 2026.

Tianyue H. Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding Adam requires better rotation dependent assumptions. In *Advances in Neural Information Processing Systems*, 38:127604–127645, 2025. doi:10.52202/085713-3838.

Ej Zhou, Suchir Salhan, Catherine Arnett, and Anna Korhonen. Cross-lingual alignment without joint training: Do monolingual language models converge on universal representations? *arXiv preprint arXiv:2608.27115*, 2026.

A CORE CONTROLLED TASKS, ARCHITECTURES, AND TRAINING

A.1 MAPPING TASK

The initial regime uses the four mappings `I like` $\mapsto$ `cats`, `I hate` $\mapsto$ `dogs`, `you like` $\mapsto$ `pizza`, and `you hate` $\mapsto$ `music`; corruption flips only `like`$\leftrightarrow$`hate`. The one-block model has $d_{\text{model}} = 4$, two dimension-2 attention heads, MLP width 2, no LayerNorm or biases, and 96 trainable parameters. Token embeddings are fixed across seeds, positional representations are learned, and logits use the transposed fixed embedding matrix. Parameters are Gaussian-initialized with standard deviation 0.1 and trained for 5,000 full-batch SGD steps at learning rate 0.03; behavioral validity requires 100% accuracy and mean loss at most 0.02. Width/depth robustness uses $d_{\text{model}} = 16$, $d_{\text{head}} = 8$, $d_{\text{ff}} = 8$ and then two blocks.

A.2 ASSOCIATIVE RETRIEVAL

The associative-retrieval task uses key vocabulary $\mathcal{K} = \{\kappa_0, \kappa_1, \kappa_2\}$ and value vocabulary $\mathcal{V} = \{\nu_0, \ldots, \nu_3\}$. Inputs are $[\kappa_a, \nu_u, \kappa_b, \nu_v, \kappa_\ell]$ with $a \neq b$, $u, v \in \{0, \ldots, 3\}$, and $\ell \in \{a, b\}$; the target is the role-distinct answer token corresponding to ν_u when $\ell = a$ and to ν_v when $\ell = b$. Enumerating ordered distinct key pairs, all value assignments, and both queried bindings yields $3 \times 2 \times 4 \times 4 \times 2 = 192$ semantic cases. Corruption increments only the value attached to the queried key modulo four, preserving both keys, the query key, the unqueried value, binding positions, and sequence length; it changes the clean answer on all 192 cases.

The two-block model uses $d_{\text{model}} = 16$, two heads, $d_{\text{head}} = 8$, $d_{\text{ff}} = 8$, 2,640 trainable parameters, a fixed orthogonal token basis, and a fixed answer decoder shared across seeds. Interface labels $\text{B}b.\text{R0}/\text{R1}/\text{R2}$ denote block b input/post-attention/post-MLP residuals; [QUERY] and [ALL] denote query-only and all-position replacement. Trainable parameters are independently initialized across seeds. Models are trained by full-batch Adam on all 192 semantic cases for 10,000 optimization steps, with learning rate 0.003, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and zero weight decay. Behavioral validity requires exact fit to the complete semantic universe. The gauge-trained associative-retrieval populations use the same optimization hyperparameters and full-batch training procedure, with the residual gauge fixed according to the corresponding intervention.

B EVALUATION CRITERIA, CHRONOLOGY, AND VALIDATION SEPARATION

B.1 GENERAL VALIDATION SEPARATION

The experiments distinguish implementation validity, behavioral validity, same-model causal sufficiency, discovery support, and untouched validation confirmation, but these checks play study-specific validation roles. Implementation validity checks that the intervention and required estimands are well-defined, and behavioral validity admits models that solve the task. In controlled studies where same-model causal sufficiency is part of the prespecified interface-admission rule, a candidate interface must first show that restoring that model's own clean state rescues the computation before cross-model transplantation is evaluated. Separately, clean-donor transplantation is compared with the identical intervention using the donor's corrupted state, providing a matched control for intervention effects that do not depend on clean donor information. Other studies may reuse a previously fixed interface or follow a different prespecified admission rule; their exact chronology is given in the study-specific summary below. Required cases are not selectively dropped when a validity condition fails, and analyses introduced only after inspection of the relevant outcome are labeled post hoc rather than treated retroactively as admission criteria.

Model-node bootstrap. For a population of N models, each replicate samples N model identities with replacement; if w_i is model i's sampled multiplicity, each observed ordered pair (i, j) with $i \neq j$ receives weight $w_i w_j$. Repeated identities therefore upweight their incident observed dyads without creating self-pairs. Zero-off-diagonal-weight draws are redrawn where applicable; same-population contrasts reuse node weights within a replicate, whereas disjoint populations are resampled independently. Reported model-node intervals use 10,000 percentile-bootstrap replicates unless noted otherwise.

B.2 Associative-retrieval evaluation criteria

For associative retrieval, a primary population requires at least $45/50$ behaviorally valid models. Interface admission requires same-model mean recovery at least 0.90 at the population level and at least 0.80 for every valid model. Cross-model support requires mean raw recovery at least 0.30, raw-minus-control advantage at least 0.20, and positive-edge density at least 0.60, where a positive ordered pair has mean raw recovery at least 0.50. Any required non-finite or non-positive recipient recovery denominator invalidates the population; no row or pair dropping is allowed. Discovery and validation use disjoint 50-model populations.

B.3 Chronology and data separation

For studies where temporal separation is material, we distinguish feasibility or discovery work from the corresponding held-out evaluation and record which choices were fixed before the relevant outcomes were inspected. Throughout this paper, "prespecified" means fixed within the project before inspection of those evaluation outcomes; it does not mean externally preregistered or independently timestamped. Table 1 records the relevant populations, the timing of outcome inspection, and the choices fixed before each evaluation.

Study	Timing	Feasibility / discovery	Held-out / confirmatory evaluation	Reported endpoint or contrast
Original mapping	Fixed before disjoint validation outcomes	Discovery population; interface admitted by same-model sufficiency	Disjoint 50-model validation population	Mean normalized raw and control recovery
Associative retrieval	Fixed before disjoint validation outcomes	Disjoint 50-model discovery population; fixed interface criteria	Disjoint 50-model validation population	Mean normalized raw/control recovery, advantage, and positive-edge density
Withheld two-hop tuples	Split, corruption, and endpoint fixed before held-out evaluation	Separate task design; 144 training compositions	50 models; 48 compositions withheld from training (96 renderings)	Held-out behavioral accuracy and raw-minus-control target-probability contrast; normalized recovery not used because some denominators are nonpositive
Scaffold A/B	Fixed before A/B held-out outcomes	Disjoint feasibility seeds	50 models/arm, fixed seed blocks	Arm-B retained advantage relative to Arm A; joint model-node bootstrap CI
Scaffold C/D	Specified after A/B results, before C/D outcomes	Disjoint C/D feasibility seeds	50 models/arm, fixed seed blocks	C/A and D/A ratios with joint model-node bootstrap CIs; no new categorical threshold
Scaled associative retrieval	Protocol and primary endpoint fixed before held-out causal evaluation	Seeds 0–1/arm; behavior-only budget selection	Seeds 2–11/arm; 1,500 steps	Prespecified normalized recovery/scaffold contrast; unnormalized target-probability diagnostic reported post hoc
Pythia final	Cross-seed test fixed before confirmatory evaluation	Engineering stage only	Original + seeds 1–9	Mean target log-prob. effect; model-node bootstrap CI; same-model sufficiency audited post hoc
Pythia training dynamics	Checkpoint/layer search completed before held-out confirmation	Original + seeds 1–3; fixed checkpoint/layer grid	Seeds 4–9; 1,635 untouched windows	Held-out target log-prob. effect at step 4000 / layer 5
Pythia target-conditioned hybrid	Hybrid test specified before hybrid outcomes	Baseline/projector gate only	Original + seeds 1–9	Mean clean-hybrid minus corrupt-hybrid target log-probability

Table 1: Chronology and separation of feasibility/discovery work from held-out or confirmatory evaluation for studies where that distinction is material. Entries record what was fixed before the corresponding outcomes were inspected; they do not imply external preregistration.

Analyses introduced after inspection of the relevant primary outcomes are treated as post hoc rather than confirmatory. These include the Arm B exact-map correction, secondary decoder-null hybrid diagnostics, the post hoc unnormalized scaled-study target-probability diagnostic, and the Pythia same-model causal-sufficiency and joint donor–recipient sensitivity diagnostics.

C Recurrence and architecture robustness

To test whether the intentionally small mapping result survives simple architectural changes and recurs in richer retrieval, Figure 3 summarizes raw and matched-control recovery across the original setting, the width/depth robustness conditions, and associative-retrieval recurrence.

The associative-retrieval validation population provides the separate recurrence condition in the richer controlled task, with raw recovery 0.746 versus 0.086 for the matched corrupted-donor control.

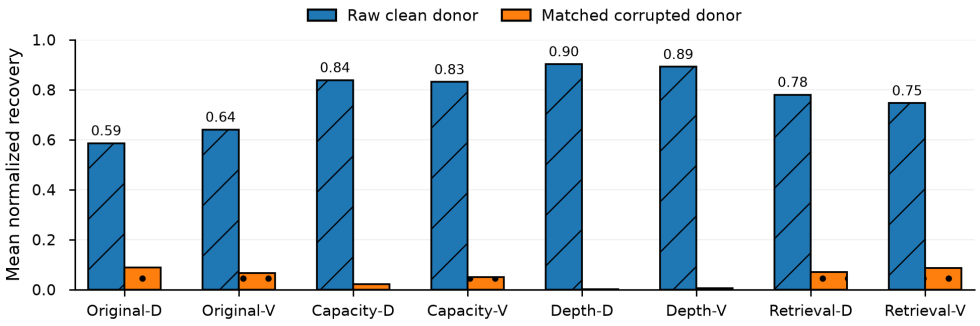

Figure 3: Raw recovery across recurrence and robustness conditions. Bars show population mean normalized recovery for raw clean-donor transplantation and the matched corrupted-donor control. Conditions include original-task discovery/validation, a wider representation, an additional transformer block, and associative-retrieval recurrence. D/V denote discovery/untouched validation. The architecture conditions are robustness tests rather than points on a scaling curve.

In the mapping task, direct compatibility survives widening the one-block model (0.831 recovery versus 0.049 control) and adding a second block (0.891 versus 0.005), with same-model recovery remaining 1.0. These points are not a scaling law; they show only that the phenomenon is not confined to the smallest architecture.

D Additional diagnostics on mapping-task cross-seed structure

The analyses below probe three simple explanations for cross-seed variation visible in the mapping experiment: stable donor/recipient effects, reproducible numerical head roles, and trivial head permutations. None supplies a sufficient mechanistic explanation of raw interoperability.

D.1 Secondary pairwise structure

As a secondary descriptive analysis, we ask whether the donor–recipient variation visible in Figure 4 contains reproducible donor- or recipient-specific structure. In untouched validation, donor and recipient downstream sensitivity correlate with mean recovery ($r = -0.554$ and $r = -0.543$), and a recipient-sensitive mismatch predictor attains cross-validated $R^2 = 0.148$ after additive donor and recipient effects are removed. These effects indicate modest systematic pairwise structure but explain only a limited fraction of compatibility variation. The tested local interventions do not support a simple causal normalization law, so we do not treat these correlations as a mechanistic explanation of raw interoperability.

D.2 Head-ablation differentiation diagnostic

This evaluation-only diagnostic asks whether the two numerical attention-head slots acquire reproducible functional roles across independently trained seeds. For each frozen example, it records the relative loss/probability effect of ablating head 1 versus head 2. Both heads are important in aggregate, but no example has a relative-effect sign that is consistent across all five seeds. The mean relative head effects (standard deviations across seeds) are $0.445\,(0.414)$ for `I like`, $-0.074\,(0.538)$ for `I hate`, $-0.718\,(0.384)$ for `you like`, and $0.283\,(0.479)$ for `you hate`; the cross-seed

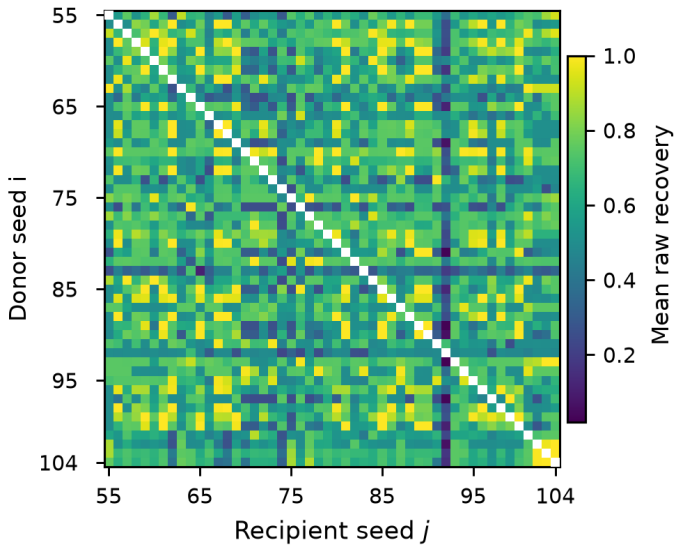

Figure 4: Held-out mapping-task pair means $\bar{C}_{ij}^{\text{raw}}$. Recovery is broadly positive across donor–recipient pairs; row/column structure indicates modest systematic donor/recipient variation analyzed below.

differentiation criterion is therefore not met. This does not claim that the heads lack any seed-specific specialization, only that fixed numerical head identity does not support a reproducible role assignment under this diagnostic.

D.3 Simple head-permutation diagnostic

We next test whether cross-seed component differences can be explained by simply exchanging the two attention-head identities. Across all 10 seed pairs and all 16 global identity/swap assignments, the mean functional distance is 0.5410 under identity correspondence and 0.5397 under the best pairwise permutation, a normalized improvement of 0.00227. The best global assignment improves total distance by 0.00251 in normalized terms. Thus simple head relabeling removes only a negligible fraction of the measured cross-seed component discrepancy. This diagnostic does not rule out richer distributed or within-layer correspondences; it only shows that a two-head identity swap is insufficient.

E Robustness of the associative-retrieval endpoint

As a descriptive robustness check on the same 50-model associative-retrieval validation population, we re-evaluated all 2,450 ordered donor–recipient pairs and 192 cases without retraining, model or interface selection, row dropping, or changing the intervention. The clean–corrupted target-probability denominator was essentially one throughout: mean 0.9999996, 1st percentile 0.9999942, and minimum 0.9999214; no observation had a gap below 0.05. Thus the normalized endpoint is not amplified by small denominators in this validation population. The raw target-probability effect closely matches the normalized result (Table 2). Recomputing normalized recovery from these extracted probabilities reproduces the reported means to numerical precision (0.7458693 clean-donor versus 0.0856647 control).

F Interoperability on withheld relational tuples

This bridge experiment uses a separate two-hop compositional retrieval task over 12 entities. A semantic composition specifies $E_a \to E_b$ and $E_b \to E_c$, queries E_a, and predicts E_c. The fixed split withholds novel two-edge compositions while ensuring that every held-out first-hop and second-hop atomic edge appears in training in the same role. We train a separate 50-model population

Quantity	Clean donor	Corrupted-donor control
Mean patched target probability	0.745869	0.085665
Mean unnormalized rescue	0.745869	0.085665
Patched top-1 accuracy	0.748943	0.084698

Table 2: Unnormalized robustness of associative-retrieval transplantation. Recipient clean and corrupted target probabilities are 0.9999997 and 9.22×10^{-8}, respectively.

(seeds 1047–1096) from scratch on 144 training compositions, withholding the remaining 48 of 192 semantic compositions entirely from training. Each composition is rendered in both fact orders, yielding 288 training sequences and 96 held-out evaluation sequences. The models use the same two-block dimensions as associative retrieval but a six-position sequence and a task-specific fixed entity embedding/answer-decoder scaffold. Corruption changes the second-hop endpoint via the offset rule fixed before held-out evaluation, while the intervention remains at $\mathrm{B2.R1[QUERY]}$.

Normalized recovery is unsuitable because 0.208% of recipient–case clean–corrupted gaps are nonpositive. We therefore report the target-probability contrast $p^{\mathrm{raw}}_{\mathrm{patch},i,j,x} - p^{\mathrm{ctrl}}_{\mathrm{patch},i,j,x}$ on all rows. No rows are dropped.

Mean held-out behavioral accuracy is 0.9725. Across 2,450 ordered pairs and 96 held-out renderings per pair, mean advantage is 0.40245 (10,000-replicate model-node bootstrap 95% CI $[0.37768, 0.42786]$), and all 2,450 ordered donor–recipient pairs have positive mean probability advantage. Patched top-1 accuracy is 0.5010 versus 0.0966 control. Thus, in this controlled task, raw interoperability extends to complete relational tuples not seen during training without relying on an invalid normalized denominator.

G Localization details

G.1 Interface localization

The query-position sweep reuses the untouched 50-model associative-retrieval validation population, all 2,450 ordered pairs, and 192 cases. Conformance passed with zero maximum identity-patch logit error. Under the same-model causal-sufficiency test, $\mathrm{B1.R0[QUERY]}$, $\mathrm{B1.R1}$, and $\mathrm{B1.R2}$ failed (population means 0, 0.0686, 0.0686), whereas $\mathrm{B2.R1}$ and $\mathrm{B2.R2}$ passed at 1.0. At $\mathrm{B2.R1}$, raw recovery is 0.7458693 versus 0.0856647 control, advantage 0.6602045 (95% model-node bootstrap CI $[0.6117, 0.7077]$). $\mathrm{B2.R2}$ lies immediately before the fixed answer decoder, with no further learned transformer computation, so we treat it only as a descriptive endpoint rather than the primary interoperable interface. Thus the sweep localizes a compact query-local causal interface late without ruling out usable early full-sequence state.

G.2 Full-sequence localization follow-up

The full-sequence follow-up replaces the complete residual tensor across positions at the same early boundaries, with $\mathrm{B2.R1[ALL]}$ as a late positive control. All early boundaries pass same-model admission at 1.0, making the earliest boundary, $\mathrm{B1.R0[ALL]}$, the primary early interface. There, raw recovery is 0.5438589 versus 0.1507355 control, advantage 0.3931234 (95% model-node bootstrap CI $[0.3761546, 0.4087321]$), with positive-edge density 0.999184. Thus early query-local failure reflects intervention granularity rather than absence of transferable information: at the same early stage, the full residual sequence is causally sufficient and transfers across models.

H Decoder-subspace control details

The decoder-subspace diagnostic uses the untouched associative-retrieval validation population (seeds 847–896), all 2,450 ordered off-diagonal pairs, and all 192 semantic cases. The fixed decoder has shape 4×16 and rank 4. Its orthogonal row-space projector is $\Pi_D = D^\top (DD^\top)^{-1} D$, with complementary projector $\Pi_D^\perp = I_{16} - \Pi_D$. In this implementation the decoder selects the four

answer-basis coordinates, and the analytic projector is verified against the corresponding coordinate mask.

The primary surrogate $s_{i,j,x} = r_{i,x}\Pi_D + \tilde{r}_{j,x}\Pi_D^{\perp}$ preserves the clean donor row-space and the corrupted recipient nullspace. Immediate decoder logits are required to be bitwise identical in float32 and to agree within 10^{-12} in float64; the check obtains bitwise equality on all 470,400 rows and zero float64 error. Matching the immediate decoder logits does not make the surrogate functionally equivalent to the donor state, because the patched $\mathrm{B2.R1[QUERY]}$ residual is subsequently processed by the recipient's final MLP before the output decoder. The canonical associative-retrieval raw recovery (0.7458693), control recovery (0.0856647), and positive-edge density (0.953061) are reproduced within 10^{-6}.

Let $\bar{C}^{\mathrm{real}}$, $\bar{C}^{\mathrm{surr}}$, and $\bar{C}^{\mathrm{ctrl}}$ denote the corresponding population-mean recoveries under the aggregation defined in Section 3.1. The primary retained-advantage statistic is

$$R_{\mathrm{decoder}} = \frac{\bar{C}^{\mathrm{surr}} - \bar{C}^{\mathrm{ctrl}}}{\bar{C}^{\mathrm{real}} - \bar{C}^{\mathrm{ctrl}}}, \tag{3}$$

with $\Delta_{\mathrm{null}} = \bar{C}^{\mathrm{real}} - \bar{C}^{\mathrm{surr}}$. The observed values are $R_{\mathrm{decoder}} = 0.6113$ (95% model-node bootstrap CI $[0.5758, 0.6450]$) and $\Delta_{\mathrm{null}} = 0.2566$ (95% CI $[0.2324, 0.2815]$). Post hoc descriptive controls give mean recovery 0.8340 for donor row-space only and 0.9156 for donor row-space plus the clean recipient nullspace. Together, these results show that decoder-visible state carries a majority of the transferable advantage, while decoder-null state remains causally consequential and clean recipient nullspace further improves downstream use. This localizes transferable signal; the scaffold study asks whether shared decoder coordinates helped induce that compatibility during training.

I SCAFFOLD INTERVENTION DETAILS

The scaffold study was evaluated in two sequential stages with outcome separation. First, Arm A (shared embedding, shared decoder) and Arm B (shared embedding, independent decoder) were fixed before their held-out outcomes were inspected, with 50 independently retrained models per arm and disjoint feasibility seeds. The interface was fixed at $\mathrm{B2.R1[QUERY]}$; no re-localization was performed. Independent decoder scaffolds were deterministic signed permutations of canonical coordinates, preserving dimensionality, orthonormality, rank, sparsity, and task semantics while changing coordinate assignment. All held-out models passed the behavioral and same-model causal-sufficiency gates.

	Output coordinates shared		**Output coordinates independent**	
	A Emb. shared	C Emb. independent	B Emb. shared	D Emb. independent
Behaviorally valid	50/50	50/50	50/50	50/50
Mean raw recovery	0.7599	0.7338	0.2550	0.2457
Mean control recovery	0.0762	0.0903	0.2521	0.2509
Raw–control advantage	0.6838	0.6435	0.0029	−0.0052
Positive-edge density	0.9706	0.9388	0.0747	0.0612
Same-model recovery	1.0000	1.0000	1.0000	1.0000

Table 3: Scaffold interventions, grouped by whether output coordinates are shared. Columns A/C share output coordinates; B/D use independent output coordinates. Raw interoperability tracks output-coordinate sharing far more strongly than input-embedding sharing. Evaluation timing is summarized in Appendix B.3.

The A/B primary statistic was $R_B = \Delta_B/\Delta_A$, computed with a joint model-node bootstrap in which each replicate recomputed the two arm advantages before taking their ratio. With 10,000 interpretable replicates, $R_B = 0.0042675$ and the 95% CI is $[-0.0225379, 0.0329740]$. The prespecified A/B decision rule classified the result as preserved if the lower CI exceeded 0.80, strongly output-scaffold-dependent if the upper CI was below 0.50, and intermediate otherwise. The observed interval lies well below the 0.50 threshold. The absolute Arm B-minus-Arm A advantage difference is -0.6808368 with 95% CI $[-0.7263824, -0.6364046]$.

After evaluating A/B, we separately specified a C/D embedding-scaffold follow-up before inspecting C/D outcomes. The follow-up used the already specified independent-embedding seed blocks and the same signed-permutation scaffold family, task, architecture, corruption, interface, validity criteria, and 50-model evaluation size. Arm C used independent embeddings with a shared decoder; Arm D used independent embeddings and independent decoders. Feasibility was assessed on four disjoint seeds per arm before the held-out evaluation, with 4/4 exact-fit and finite models in each arm. All 50 held-out models in each arm were behaviorally valid and had same-model recovery 1.0.

The C/D follow-up was specified after observing A/B but before inspecting C/D outcomes, and it introduced no new post-A/B categorical thresholds. Its primary outputs were estimation targets relative to the Arm A anchor. For Arm C, $\Delta_C = 0.6434605$, $R_C = \Delta_C/\Delta_A = 0.9410692$, and the 10,000-replicate bootstrap 95% CI for R_C is $[0.8564135, 1.0299129]$; the absolute difference $\Delta_C - \Delta_A = -0.0402942$ has 95% CI $[-0.1017440, 0.0196130]$. For Arm D, $\Delta_D = -0.0051827$, $R_D = -0.0075798$, and the 95% CI for R_D is $[-0.0330826, 0.0218672]$; the absolute difference $\Delta_D - \Delta_A = -0.6889375$ has 95% CI $[-0.7329787, -0.6440417]$.

Using all four arm estimates only as a descriptive factorial summary gives a decoder-sharing main effect of 0.6647400, an embedding-sharing main effect of 0.0241974, and an embedding-by-decoder interaction of 0.0321936. These quantities are not confirmatory tests and were not used for categorical decisions. They are reported only as descriptive summaries, preserving the temporal separation between the original A/B decision and the later C/D follow-up.

Separately, after observing A/B, we added a post hoc exact-map diagnostic. For donor state h_i, with donor and recipient signed-permutation scaffolds Q_i, Q_j, respectively, we patch $h_i Q_i^\top Q_j$ and fit no alignment parameters. Deterministic retraining reproduces the frozen Arm B null within tolerance. Correction yields mean raw recovery 0.75239 versus 0.07964 control, advantage 0.67274, and positive-edge density 0.95673. The corrected retained advantage versus frozen Arm A is 0.98390 (10,000-replicate model-node bootstrap 95% CI $[0.90318, 1.07237]$), showing that correcting the known coordinate mismatch recovers nearly all of the Arm B deficit. This analysis is post hoc, not confirmatory.

J SCALED ASSOCIATIVE-RETRIEVAL DETAILS

This stress test asks whether the shared-versus-independent decoder contrast persists at greater controlled scale with routing fixed. Architecture and optimization feasibility were assessed separately from the held-out causal evaluation. The final architecture has eight transformer blocks, $d_{\text{model}} = 128$, eight heads with $d_{\text{head}} = 16$, and MLP width 512 (approximately 1.6M trainable parameters). Each procedural example contains 32 key–value bindings from 8,192 keys and 8,192 values. A fixed zero-parameter content-addressing router after block 4 uses the shared fixed token-embedding coordinates to match the query to a key and route the corresponding learned value residual into the query state. The router performs only deterministic key matching and routing; the representations transplanted after block 6 remain learned residual states produced by the transformer. The causal interface is fixed after block 6. The shared-decoder arm uses one fixed rank-32 decoder; each independent-decoder model uses its own fixed rank-32 decoder. Embeddings are shared and fixed.

Feasibility seeds 0–1 were separated from held-out seeds 2–11. A fixed full-task budget grid of 750, 1,000, 1,500, and 2,000 optimizer steps selected 1,500 as the first checkpoint at which all four seed/arm pilot models exceeded 95% validation accuracy. Confirmation therefore trains every held-out model for exactly 1,500 steps. Nine of ten models in each arm pass the 95% behavioral gate, exceeding the planned minimum of eight. The causal population uses 4,096 held-out procedural examples from a fixed candidate set. All router matches are exact and no causal transplant is computed before the baseline gate passes.

The normalized matched-control contrast was fixed as the primary endpoint before held-out evaluation. For recipient j and example x, define the recipient-native clean–corrupt target-probability gap

$$g_{j,x} = p_{\text{clean},j,x} - p_{\text{corr},j,x}.$$

Let $p^{\mathrm{raw}}_{i,j,x}$ and $p^{\mathrm{ctrl}}_{i,j,x}$ denote the recipient target probabilities after patching donor i's clean and corrupted states, respectively. For a patched probability p, define

$$\mathcal{R}_{j,x}(p) = \frac{p - p_{\mathrm{corr},j,x}}{g_{j,x}}.$$

The scaled whole-state endpoint is therefore

$$\begin{aligned}\Delta^{\mathrm{scaled}}_{i,j,x} &= \mathcal{R}_{j,x}\big(p^{\mathrm{raw}}_{i,j,x}\big) - \mathcal{R}_{j,x}\big(p^{\mathrm{ctrl}}_{i,j,x}\big) \\ &= \frac{p^{\mathrm{raw}}_{i,j,x} - p^{\mathrm{ctrl}}_{i,j,x}}{g_{j,x}}.\end{aligned} \tag{4}$$

The statistic is averaged over eligible examples within each ordered donor–recipient pair and then over the 72 ordered pairs in each arm. It is dimensionless and is *not* a sum over the 4,096 examples.

Eligibility is determined from recipient-native baseline behavior and requires a finite clean–corrupt gap $g_{j,x} > 10^{-6}$, using the prespecified strict threshold implemented in the evaluation code. Nine models pass the behavioral gate in each arm, giving 36,864 candidate recipient–example rows per arm. Of these, 36,134 are eligible in the shared-decoder arm and 36,240 in the independent-decoder arm.

Unlike the original associative-retrieval validation population, whose clean–corrupt probability gaps are essentially one (Appendix E), the scaled population contains a small lower tail of gaps close to the eligibility threshold. Across both arms, 1.12% of eligible recipient–example rows have $g_{j,x} < 10^{-2}$; the minimum eligible gap is 1.0021×10^{-6}, the fifth percentile is 0.999998, and the median is 1.0. Writing

$$\delta p_{i,j,x} = p^{\mathrm{raw}}_{i,j,x} - p^{\mathrm{ctrl}}_{i,j,x}, \qquad \Delta^{\mathrm{scaled}}_{i,j,x} = \frac{\delta p_{i,j,x}}{g_{j,x}},$$

shows that the small number of near-threshold denominators can have a disproportionate effect on the numerical magnitude of the mean normalized statistic. We therefore retain the normalized endpoint exactly as prespecified for the primary analysis, but do not interpret its absolute magnitude as an effect-size scale or as "times recovered."

Under this prespecified endpoint, the estimates are 905.3776 for shared decoders (95% CI $[744.3974, 1057.8870]$), 39.7329 for independent decoders (95% CI $[9.8952, 81.3708]$), and 865.6447 for the shared-minus-independent scaffold contrast (95% CI $[698.8101, 1025.6709]$). The rank-32 decoder-visible / recipient-specific hybrid is 907.2985 (95% CI $[746.4376, 1053.6543]$), retaining 1.0021 of the shared whole-state normalized statistic.

For descriptive effect-size interpretation, we additionally report a clearly post hoc unnormalized target-probability diagnostic on the same prespecified eligible examples, computed from the saved example-level effects without rerunning models or transplants. The diagnostic is $\delta p_{i,j,x}$ above: it removes the division by the clean–corrupt gap while preserving the raw-versus-control transplantation contrast. The eligibility set itself remains the one fixed for the primary normalized endpoint. Across the 72 ordered pairs in each valid arm, mean δp is 0.9729 for shared decoders and 0.0430 for independent decoders, a shared-minus-independent difference of 0.9299. The decoder-visible / recipient-specific hybrid has mean $\delta p = 0.9730$, corresponding to 1.00015 of the shared whole-state mean using the full-precision pair-aggregated estimates. Every ordered pair has positive mean δp in all three analyses. This diagnostic is descriptive and post hoc; it does not replace, modify, or determine the success criterion for the prespecified primary endpoint.

As a further post hoc sensitivity analysis, we recomputed the normalized statistic after imposing more conservative clean–corrupt gap thresholds on the saved effects. With $g_{j,x} > 10^{-3}$, the shared and independent means are 2.2969 and 0.0814, respectively, for a difference of 2.2155. With $g_{j,x} > 10^{-2}$, the corresponding means are 1.1107 and 0.0457, for a difference of 1.0650. Thus the very large absolute magnitude of the prespecified normalized statistic is driven by the small-denominator tail, whereas the qualitative shared-versus-independent scaffold contrast persists after excluding that tail. These thresholded analyses are descriptive robustness checks and do not replace the prespecified $g_{j,x} > 10^{-6}$ endpoint.

Uncertainty for the prespecified endpoint uses 10,000 model-node bootstrap replicates with fixed bootstrap seeds 21000 (shared), 21001 (independent), 21002 (scaffold contrast), and 21003 (hybrid).

No bootstrap replicate was discarded for zero total off-diagonal pair weight; this bootstrap-weight denominator is distinct from the example-level clean–corrupt gap $g_{j,x}$ discussed above. The prespecified analysis yields a positive scaffold contrast, while the absolute scale of the normalized statistic is highly sensitive to the small-denominator tail. The post hoc unnormalized target-probability diagnostic and threshold sensitivity checks show that the same qualitative contrast remains large without interpreting the pathological normalized magnitude as an effect size. Together, these results support recurrence of decoder-scaffold dependence under greater model and task-space scale with routing held fixed; they do not test scaling in unconstrained transformers or independent discovery of the routing algorithm.

K Exact Residual-Basis Intervention: Proof and Numerical Conformance

Because the scaffold studies compare differently trained populations, we isolate coordinates at fixed function with an exact residual-basis reparameterization that preserves input–output behavior while changing hidden-state coordinates (Godfrey et al., 2022; Cain, 2026; Karasan, 2026). Consider one transformer block with sequence length T, residual width d, and residual-state matrix $Z \in \mathbb{R}^{T\times d}$, whose rows are the token positions. Let E and P_{pos} denote the token- and positional-embedding matrices. Define the projected attention states

$$A_Q = ZW_Q, \qquad A_K = ZW_K, \qquad A_V = ZW_V,$$

and the residual matrices at interfaces R1 and R2 by

$$Z_1 = Z + \mathrm{Attn}(A_Q, A_K, A_V)W_O, \qquad Z_2 = Z_1 + \mathrm{ReLU}(Z_1W_1)W_2.$$

With decoder D, the tokenwise logits are $\Lambda = Z_2D^\top$. Let $Q \in \mathbb{R}^{d\times d}$ be orthogonal, so $Q^\top Q = QQ^\top = I_d$, and define

$$E' = EQ, \quad P'_{\mathrm{pos}} = P_{\mathrm{pos}}Q, \quad W'_Q = Q^\top W_Q, \quad W'_K = Q^\top W_K, \quad W'_V = Q^\top W_V,$$

$$W'_O = W_OQ, \qquad W'_1 = Q^\top W_1, \qquad W'_2 = W_2Q, \qquad D' = DQ.$$

If $Z' = ZQ$, then

$$Z'W'_Q = ZQQ^\top W_Q = ZW_Q = A_Q,$$

and identically for W_K and W_V. Thus the projected queries, keys, and values are unchanged. Consequently, the attention scores, softmax weights, and head-space outputs are also unchanged, while

$$Z'_1 = ZQ + \mathrm{Attn}(A_Q, A_K, A_V)W_OQ = Z_1Q.$$

The MLP preactivation is likewise invariant, $Z'_1W'_1 = Z_1W_1$, so

$$Z'_2 = Z_1Q + \mathrm{ReLU}(Z_1W_1)W_2Q = Z_2Q.$$

Finally,

$$\Lambda' = Z'_2(D')^\top = Z_2QQ^\top D^\top = Z_2D^\top = \Lambda.$$

Thus the transformed network computes exactly the same function while expressing every residual state in an orthogonally transformed basis. Applying the same argument blockwise proves the result for the two-block associative-retrieval model. At a patched residual boundary, a canonical donor row state h_i must therefore be mapped to h_iQ before insertion into the transformed recipient to reproduce the canonical patched logits up to numerical precision; inserting h_i directly is a substantive coordinate-mismatch intervention.

K.1 Numerical conformance and excluded implementation

An initial float32 post-training gauge implementation failed the numerical conformance tolerance: although orthogonality error was 2.38×10^{-7}, clean/corrupted logit, interface-equivariance, and corrected-patch deviations exceeded 10^{-6}. These runs are excluded entirely from substantive conclusions. The reported float64 implementation on the same 50 validation models and 10 gauges passes stricter conformance (up to 10^{-9} for functional checks) and reproduces the canonical aggregate within 10^{-6}. Gauge-corrected patching reproduces canonical recovery to numerical precision, whereas raw gauge mismatch retains median 0.2312 of canonical donor advantage, with all 10 gauges at or below 0.50.

L TRAINING IN INDEPENDENTLY CHOSEN ORTHOGONAL RESIDUAL GAUGES

To distinguish optimization-induced convergence from compatibility inherited from a shared coordinate scaffold, we compare the canonical 50-model associative-retrieval population (model seeds 847–896) with a new 50-model population trained from scratch in five fixed orthogonal residual gauges (10 models per gauge; model seeds 947–996). For each gauge Q_g, the fixed token embedding and decoder are transformed consistently, while all trainable parameters are independently initialized and optimized in that gauge. Training uses full-batch Adam for 10,000 optimization steps with learning rate 0.003 and otherwise fixed hyperparameters. All 50 transformed models are behaviorally valid, and every gauge has same-model recovery 1.0 for every model at the canonical late interface.

We compare three intervention classes. *Same-gauge* transfer patches between independently optimized models sharing one transformed gauge. *Raw cross-gauge* transfer inserts a donor state directly across the canonical and transformed coordinate systems. *Gauge-corrected* transfer applies the analytically known orthogonal map between the donor and recipient training gauges before patching. Both canonical→transformed and transformed→canonical directions are evaluated.

Same-gauge interoperability is strong in every transformed population; the pooled raw advantage has a 10,000-replicate model-node bootstrap 95% CI of $[0.8746, 0.9493]$. Raw canonical→transformed transfer has median retained advantage 0.0019 across gauges, while transformed→canonical has median retained advantage -0.0183. The gauge-specific raw effects are heterogeneous: most confidence intervals include zero, with small positive residual transfer in some directions and negative transfer in others. By contrast, coordinate correction restores large causal effects for every gauge. Canonical→transformed corrected advantages range from 0.7325 to 0.7518, and transformed→canonical corrected advantages range from 0.8629 to 0.9459.

The exact post-training reparameterization establishes coordinate relativity under strict functional equivalence. The from-scratch study asks a different question: whether basis-sensitive optimization under Adam nevertheless converges to mutually raw-interoperable interfaces across deliberately different residual gauges. It does not in this tested setup. Independently optimized models develop strong interoperability within a shared residual coordinate system, while the known coordinate map restores interoperability across different training gauges. Because Adam is not rotation-equivariant (Ling et al., 2022; Zhang et al., 2025), this experiment does not imply that symmetry-related training trajectories should coincide.

M BOUNDARY TEST OF IDENTITY-MAP INTEROPERABILITY IN PYTHIA-70M

M.1 PRESPECIFIED FINAL-CHECKPOINT TEST

We test whether identity-map interoperability extends to independently pretrained natural-language models outside the controlled scaffold. This is a boundary test of direct raw transfer; failure of the identity map does not rule out learned or non-identity correspondences. We use ten matched Pythia-70M runs: the original Pythia model and nine additional PolyPythias random-seed variants (seeds 1–9) (Biderman et al., 2023; van der Wal et al., 2025). The prespecified final-checkpoint test patches the output of transformer layer 3 of 6 at the final prefix token. Inputs are 128-token WikiText-103 prefixes (Merity et al., 2017), with token 129 as the prediction target; corruption swaps prefix positions 120 and 124. Eligibility is recipient-specific and requires a clean-minus-corrupt target log-probability gap of at least 0.5 nats.

At each Pythia interface, $h_{i,x}^{\mathrm{clean/corr}}$ denotes donor i's final-prefix residual for clean/corrupted example x; y_x is the target token, and $p_j(y_x \mid h)$ is recipient j's target probability on corrupted x after replacing that interface state by h. The primary endpoint is

$$\Delta^*_{i,j,x} = \log p_j\big(y_x \mid h_{i,x}^{\mathrm{clean}}\big) - \log p_j\big(y_x \mid h_{i,x}^{\mathrm{corr}}\big)\,,$$

averaged over eligible examples within ordered donor–recipient pairs and aggregated with a 10,000-replicate model-node bootstrap. For any Pythia cross-model endpoint below, *positive-pair density* is the fraction of ordered donor–recipient pairs whose mean endpoint effect over eligible examples is greater than zero. The discovery/confirmation population separation and the corresponding fixed evaluation choices are summarized in Table 1.

At the prespecified final-checkpoint interface, the population mean is -0.00527 with 95% CI $[-0.02651, 0.01653]$ and positive-pair density 0.444, providing no evidence of reproducible whole-state raw transfer in this tested regime. A leave-one-model-out sensitivity analysis, removing each model entirely as both donor and recipient, yields population means ranging from -0.00984 to 0.00050; nine of ten estimates remain negative and the sole positive estimate is near zero.

M.2 Held-out training-dynamics test

A separate training-dynamics study asks whether the null is specific to the final checkpoint or chosen layer. Four discovery seeds select among a fixed set of checkpoints and layers, identifying step 4000 / layer 5. Held-out confirmation on seeds 4–9 and all 1,635 remaining construction-valid validation windows again yields a null result: mean -0.00185, 95% CI $[-0.03595, 0.04099]$, and positive-pair density 0.467. The reduction from the originally requested 2,048 held-out windows to 1,635 is determined solely by the model-independent fact that the validation split contains 1,891 construction-valid windows in total, of which the first 256 were already reserved for discovery; no model outcome is used in this repair.

M.3 Prespecified target-conditioned hybrid

This confirmatory follow-up was separately specified before evaluation and tests whether a recipient-local target-conditioned direction can reveal raw transfer when whole-state transplantation fails. For each eligible recipient/example, the rank-one projector is defined by the gradient of the recipient target log-probability with respect to the corrupted-baseline residual at the same interface. The clean and control hybrids differ only in whether the projected donor component comes from the clean or corrupted donor, while the recipient retains its own orthogonal complement. Across all ten final-checkpoint seeds, the population mean hybrid effect is 0.00087 with 95% CI $[-0.04371, 0.03899]$ and positive-pair density 0.533.

M.4 Post hoc interpretability and robustness diagnostics

The following analyses were introduced after inspection of the corresponding frozen cross-seed outcomes and are reported as interpretability or robustness diagnostics rather than confirmatory evidence.

Same-model causal-sufficiency audit. Post hoc, we re-used the frozen interfaces and candidate sets and replaced each recipient's corrupted state with its own clean state to test whether either null reflected a causally weak intervention site; this diagnostic is interpretive, not confirmatory. At the final-checkpoint/layer-3 interface, same-model ($i = j$) transplantation recovers 0.469 nats on average, corresponding to 39.3% of the clean–corrupt gap. At the held-out step-4000/layer-5 interface, it recovers 1.016 nats, or 82.6% of the gap, and is positive on approximately 99% of eligible model–example cases. Candidate hashes confirm identical candidate sets for the diagnostic and frozen confirmatory runs. Table 4 juxtaposes the cross-seed effects with these post hoc same-model diagnostics.

Study	Cross-seed Δ^*	95% CI	Same-model Δ^*	Gap recovered
Final / layer 3	-0.00527	$[-0.02651, 0.01653]$	0.469	39.3%
Step 4000 / layer 5	-0.00185	$[-0.03595, 0.04099]$	1.016	82.6%

Table 4: Pythia-70M cross-seed confirmatory effects and post hoc same-model sufficiency diagnostics. Same-model statistics were computed after inspecting the cross-seed outcomes and are interpretability, not confirmatory, evidence.

The step-4000/layer-5 result therefore gives the clearest boundary case: the tested interface is causally sufficient within each individual model, yet the corresponding state shows no reliable direct identity-map usability across the tested independently pretrained seeds. The final-checkpoint/layer-3 audit shows the same qualitative pattern with more moderate within-model recovery. These results distinguish failure of cross-seed raw interoperability from the weaker explanation that the tested interface is simply causally inert.

Joint donor–recipient sensitivity robustness. The frozen confirmatory analyses define candidate eligibility using the recipient's clean–corrupt target log-probability gap. To test whether the cross-seed null could be driven by examples that are clean–corrupt sensitive for the recipient but not the donor, we reanalyzed the already-computed transplantation outputs after additionally requiring the donor to exhibit the same minimum clean–corrupt gap of 0.5 nats on each retained candidate. This analysis was conducted post hoc, introduces no new model forward passes, and does not alter the previously fixed estimands.

At the final-checkpoint/layer-3 interface, all 90 ordered donor–recipient pairs retained jointly sensitive examples (119–157 per pair; median 139). The mean transplantation effect remained near zero at -0.0123 nats (95% model-node bootstrap CI $[-0.0392, 0.0155]$). At the independently selected step-4000/layer-5 interface evaluated on held-out seeds, all 30 ordered pairs likewise remained evaluable (89–114 jointly sensitive examples per pair; median 100), with mean effect -0.0002 nats (95% CI $[-0.0459, 0.0627]$). Thus, within these tested interfaces and the frozen transplantation outputs, the cross-seed null persists even when every retained candidate is clean–corrupt sensitive in both donor and recipient.

Taken together, these tests find no reproducible whole-state or tested output-sensitive identity-map raw transfer across the matched Pythia-70M seed population at the evaluated interfaces. The post hoc sufficiency audit shows that the strongest held-out null occurs at an interface that is highly causally informative within each model, strengthening its interpretation as a failure of direct cross-seed usability rather than a failure of the intervention site itself. This bounds spontaneous raw interoperability for the specified model family, interfaces, corruption, and evaluation criteria; it does not establish that raw interoperability is absent from pretrained language models generally.