CAUSAL-ROLE RECURRENCE, GEOMETRIC VARIATION, AND OBJECTIVE-CONDITIONED READOUT ACROSS MAMBA-1 SCALES

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic comparisons across model scales can conflate three questions: whether independently reconstructed components play an operationally matched **causal role**, whether their native **geometry** is preserved, and how an **objective** reads them. We separate these questions in pretrained Mamba-1 models from 130M to 2.8B parameters.

At 130M, prospective transport, response-independent specificity, matched necessity, and restoration establish a recurrent-state mechanism on disjoint populations. At larger scales, a frozen discover-then-confirm protocol selects one of five local planes on a discovery cohort, fixes a response-blind geometric control, and evaluates the pair on a disjoint confirmation cohort. The selected-restored versus matched-control contrast is positive under each scale's frozen confirmation, while a distinct historical 370M residual criterion remains failed.

Coordinate-insensitive comparison of matched response-blind XG2/XG4 populations yields heterogeneous, intermediate cross-scale similarity (centered linear CKA: XG2 0.34–0.74; XG4 0.39–0.59), with the same qualitative pattern under cosine-RSM. Across scales, the frozen task contrasts reorganize functionally: no single geometric scalar explains the aggregate sign pattern. Holding each scale's contrast fixed, the pretrained next-token objective gives the opposed five-scale point-estimate vector. Pair-resampling intervals exclude zero in 8 of 10 scale-by-objective cells, and a 1.4B sensitivity contrast flips the pretrained-LM sign, so the objective comparison remains contrast-conditioned.

Thus a scale-local interventional role can recur without shared plane identity or exact geometric invariance, while functional readout depends on both objective and the frozen native contrast. We frame this as a controlled Mamba-1 case study rather than a universal scaling law.

1 INTRODUCTION

When mechanistic structure is compared across model sizes, "the same mechanism" can refer to different objects. We distinguish **causal role** (what changes under a matched intervention), **geometric realization** (how that role is represented in native state space), and **objective-conditioned functional meaning** (how a particular objective reads the state). Throughout, *recurrence* means recurrence of an operationally matched causal role: independently reconstructed scale-local components satisfy the same selected-restored versus matched-control intervention contrast on fresh confirmation data. It does not mean shared coordinates, principal-plane identity, or semantic homology across models.

We study five pretrained Mamba-1 models: 130M, 370M, 790M, 1.4B, and 2.8B. We first establish one recurrent-state mechanism deeply at 130M using prospective transport, response-independent geometric specificity, matched necessity, and restoration. Larger scales then use a frozen discover-then-confirm protocol rather than reusing the 130M plane identity.

The first cross-scale result is a separation. A scale-local causal role survives its own frozen matched-control confirmation at every tested size even though the selected principal-plane rank changes. In

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.

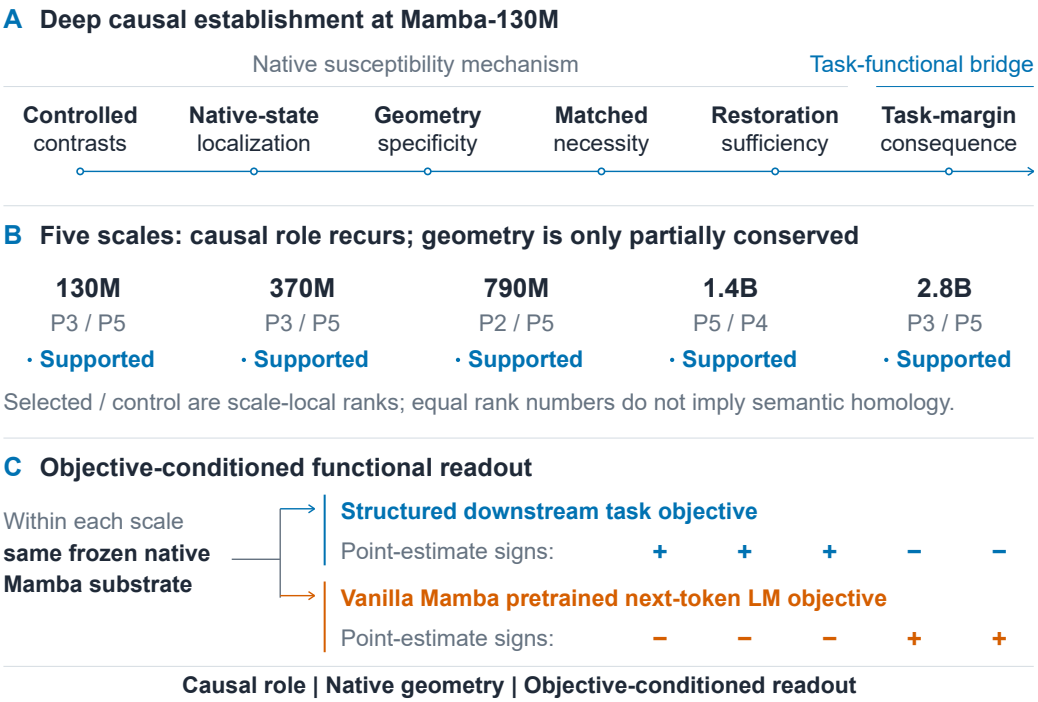


Figure 1: Study design. Native geometry is reconstructed independently within each scale. At 130M, plane selection is response-blind; at later scales a dedicated discovery cohort selects one of five planes before an untouched matched-control confirmation. The frozen primary contrast is then read under downstream and pretrained objectives. Arrows indicate study order, not shared plane identity or a continuous scaling law.

contrast, response-blind alignment geometry shows heterogeneous, intermediate cross-scale conservation: centered linear CKA spans 0.34–0.74 for XG2 and 0.39–0.59 for XG4, with secondary cosine-RSM giving the same qualitative picture. We therefore find that *exact CKA-level invariance does not hold on these frozen cohorts*, without claiming that cross-scale similarity is absent or unexpectedly low under a calibrated null.

The second result concerns functional use. For each scale's primary frozen native contrast, the downstream task and pretrained next-token objectives yield opposed five-scale point-estimate sign vectors. Eight of ten scale-by-objective cells are individually sign-stable under pair resampling; the two smallest-magnitude cells are not. A sensitivity contrast at 1.4B also changes the pretrained-LM sign, so the objective result is explicitly contrast-conditioned rather than a universal scale-level orientation.

These findings motivate three contributions:

1. **Operational causal-role recurrence.** Independently reconstructed scale-local components survive a common selected-restored versus matched-control criterion on fresh confirmations; the recurring object is procedural and interventional, not a fixed plane.
2. **Geometric variation with partial conservation.** Response-blind CKA and cosine-RSM show stable, heterogeneous intermediate cross-scale similarity rather than exact invariance.
3. **Scale-dependent functional reorganization.** Scale-local frozen contrasts couple differently to the downstream task, with no single geometric scalar explaining the aggregate pattern; holding each contrast fixed, changing objective yields an opposed point-estimate orientation that remains contrast-sensitive.

Behavioral and external-transfer experiments bound how far these mechanistic findings extend beyond the local state.

2 Related Work

Mechanistic interpretability of Mamba and state-space models. Mamba replaces attention with input-dependent selective state-space dynamics, making recurrent state a natural object for mechanistic analysis (Gu & Dao, 2024). Ensign and Garriga-Alonso partially reverse-engineer the indirect-object-identification circuit in Mamba, identifying a key bottleneck, convolution-mediated token shifting, and linear storage of name information in the SSM state (Ensign & Garriga-Alonso, 2024). Wang et al. compare Transformers and Mambas using sparse autoencoders and circuit analysis, finding substantial feature similarity and structurally analogous induction mechanisms while also identifying Mamba-specific differences (Wang et al., 2025). More recent work identifies activation-subspace bottlenecks across several state-space models and tests their causal relevance through interventions and steering (Mohan et al., 2026).

Our question is complementary. Rather than asking whether Mamba contains an interpretable feature or circuit for a particular benchmark, or whether two architectures implement analogous algorithms, we ask whether a **causal role identified inside one Mamba scale recurs after native geometry is independently reconstructed at larger scales**. We then separate that recurrence from both geometric invariance and downstream functional interpretation.

Representation similarity and cross-model recurrence. Representational-similarity methods such as centered kernel alignment provide coordinate-insensitive tools for comparing neural representations (Kornblith et al., 2019). The Platonic Representation Hypothesis further argues that representations may become increasingly aligned across models and modalities as capability grows (Huh et al., 2024). At the same time, recent work shows that representational-similarity conclusions can depend strongly on objectives and evaluation datasets (Ciernik et al., 2025), and formal work on linear representations emphasizes that geometric interpretation itself depends on the structure used to compare representational spaces (Park et al., 2024). Sun and Toneva formalize *interpretive equivalence* across neural networks and derive conditions relating representations, circuits, and algorithmic interpretations (Sun & Toneva, 2026).

Our empirical question is narrower and complementary. We do not require two models to share a direction, a principal-plane rank, a globally aligned coordinate system, or an interpretation in the formal equivalence sense. We ask whether independently reconstructed components can satisfy the same **operational interventional role**, and then test whether response-blind geometry and objective-conditioned readout recur with it.

Causal interventions and functional interpretation. Mechanistic interpretability increasingly distinguishes correlational localization from causal evidence. Prior work has used interventions and causal tracing to localize model computations, edit factual associations, and formalize mechanistic claims using causal abstractions (Dai et al., 2022; Meng et al., 2022; Geiger et al., 2023). Our protocol follows the same general principle that localization alone is insufficient, but places additional emphasis on prospective matched controls. At 130M, a localized component must transport to fresh data, outperform a response-independent geometric control, show selective necessity, and recover the response under matched restoration before serving as the basis for cross-scale comparison.

Finally, representation and causal role do not uniquely determine how a state is used by a particular objective. Our setting permits a controlled comparison: within each Mamba scale, the native contrast and source population are held fixed while its local functional readout is evaluated under either the structured downstream task objective or the original pretrained next-token objective. We therefore distinguish **objective-conditioned functional meaning** from both native geometry and causal-role recurrence.

3 RESULTS

Notation at a glance

Notation	Meaning
XG1	Controlled causal-response and behavioral holdout family.
XG2/XG4	Response-blind geometry generator families.
P_k, PP_k	Within-model principal-plane rank; labels are model-local.
Q	Internal recurrent-state causal endpoint.
ΔL	Local downstream task-margin readout.

All three controlled families use the same six-cell title/name/role/predicate slot-substitution design over synthetic structured records. XG1 is an independently generated record family reserved for causal-response and behavioral holdouts. XG2 and XG4 are disjoint response-blind generator families with different controlled record templates and lexicons (bulletin-style versus register-style) used to reconstruct native geometry. Thus $Q = E_{\mathrm{XG2}} - E_{\mathrm{XG4}}$ denotes relative squared local susceptibility to these two controlled generator families; it is not claimed to encode a natural semantic dichotomy. The XG2/XG4 contrast is intentionally controlled rather than semantically natural: it provides a response-independent comparison axis on which recurrent-state susceptibility can be localized and prospectively tested. Its scientific role is as a reproducible mechanistic probe, not as a claim that the two generator families instantiate a latent semantic variable.

3.1 DEEP CAUSAL ESTABLISHMENT AT MAMBA-130M

We first establish the local recurrent-state mechanism in Mamba-130M before asking whether any part of it recurs at larger scales. Here, a **causal role** means that manipulating a scale-local native-state component changes a pre-specified recurrent-state endpoint relative to matched controls, and that restoring the component recovers that endpoint.

Starting from controlled semantic contrasts, we localized a dominant native-state component, PP3, and evaluated it through prospective experiments on disjoint XG1 controlled-generator holdouts (Fig. 2A). PP3 denotes the third principal plane of the independently reconstructed 130M geometry; its rank label is local to this model and does not denote a globally shared semantic feature.

The localized PP3 effect first transported to an independent 300-pair population. A second holdout then compared PP3 against PP5, selected independently from response values as a geometric control, and showed that PP3 retained the stronger effect. Subsequent matched interventions established both necessity and restoration sufficiency: neutralizing PP3 attenuated the endpoint more strongly than the control intervention, while restoring the exact PP3 component recovered the endpoint more strongly than an equal-norm PP5 replacement. All four pre-specified tests supported their respective hypotheses. Exact test statistics are reported in Appendix A.

The 130M intervention also reached the downstream task computation. Restoring PP3 instead of the matched PP5 replacement increased the downstream task margin by 0.00833 on average over 300 paired items (Fig. 2B). This was a small continuous margin effect rather than an accuracy gain; discrete prediction changes were negligible. The 10^{-8}-scale Q means reported for the local causal tests are in the internal endpoint's native units and should not be read as downstream behavioral-effect magnitudes; their very small p-values reflect consistency relative to paired variation. Across source pairs, the local task-functional readout tracked the behavioral restoration effect with Pearson $r = 0.618$, Spearman $\rho = 0.832$, and 82.3% sign agreement (Fig. 2C). These associations are descriptive and were not used to select the mechanism.

The object carried into the scale study is therefore not "PP3" as a fixed semantic plane. It is an operationally matched, scale-local recurrent-state causal role under a frozen scientific procedure.

3.2 OPERATIONAL CAUSAL-ROLE RECURRENCE WITH PARTIAL GEOMETRIC CONSERVATION

We next ask two questions: *does an operationally matched causal role survive at larger scales, and how much of the response-blind geometry is conserved?*

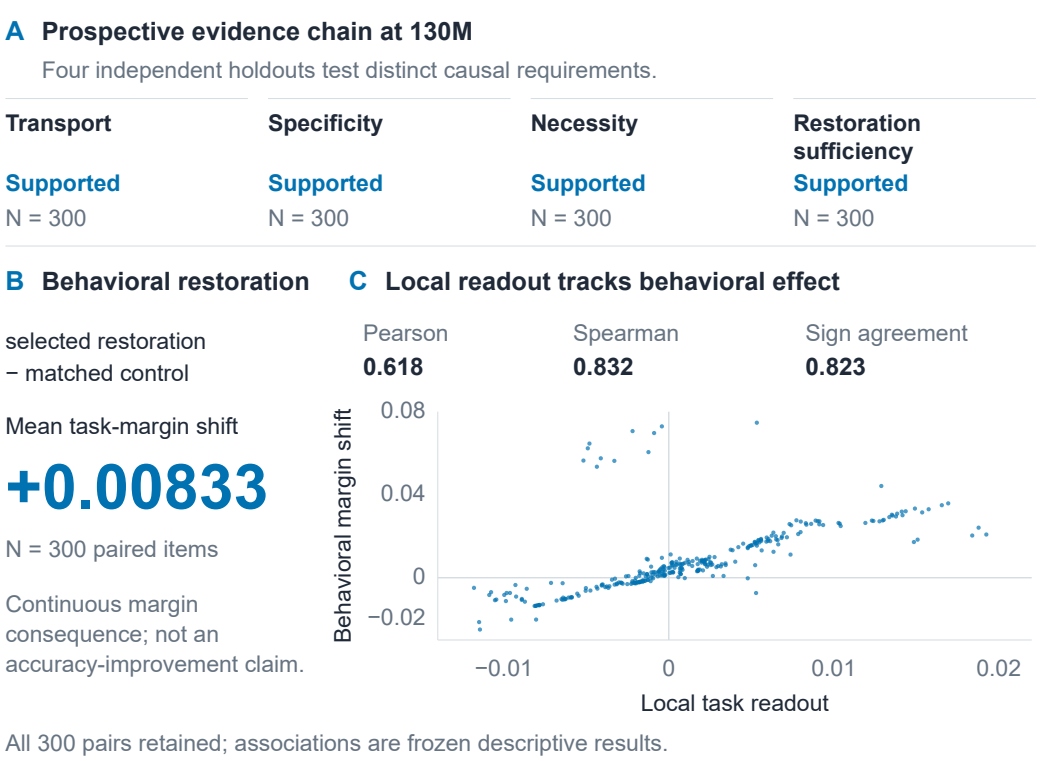


Figure 2: Deep 130M causal foundation. (A) Independent holdouts support transport, response-blind specificity, matched necessity, and restoration; their distinct endpoints are not comparable as raw magnitudes. (B) Restoration reaches a continuous task margin, without an accuracy claim. (C) Local readout and behavioral effects covary descriptively.

The selected/control pairs are P3/P5 at 130M, P3/P5 at 370M, P2/P5 at 790M, P5/P4 at 1.4B, and P3/P5 at 2.8B (Fig. 3A). These are scale-local identities, not same-numbered-plane homologies. At 130M, P3 was localized from response-blind XG2/XG4 geometry before XG1 outcomes. At later scales, all five reconstructed planes were evaluated only on a dedicated discovery cohort; the unique mean-response argmax was frozen, a non-selected control was chosen from frozen geometry alone, and the pair was then evaluated on a disjoint confirmation cohort. The discovery winner is therefore a selection rule rather than confirmatory evidence, and the primary claim is matched-control confirmation rather than omnibus superiority over every other plane.

Under the common algebraic contrast $Q_{\text{selected restored}} - Q_{\text{matched control}}$, each scale's frozen confirmation is positive. At 370M, however, a distinct historical positive aggregate-residual criterion failed; later core confirmations do not retroactively alter that failed criterion (Appendix A.4). Raw Q magnitudes are endpoint-specific internal units and are not treated as a scaling series.

For each scale and response-blind family (XG2 and XG4), the frozen geometry preparation provides 300 `alignment_delta_h` vectors on the same ordered source-pair population. Row-normalized vectors define sample Gram matrices, which we compare across scales using centered linear CKA; cosine-RSM is secondary. The off-diagonal CKA values are heterogeneous and intermediate rather than near either identity or zero: XG2 spans 0.34–0.74 and XG4 0.39–0.59. Joint source-pair resampling preserves this qualitative structure, while split halves show stronger fine-grained ordering stability for XG2 than XG4 (Appendix A.7).

These values establish lack of *exact* invariance under the stated CKA comparison while retaining nonzero cross-scale conservation. We do not estimate a within-model noise ceiling or a cross-model null distribution, and therefore do not claim that the observed similarities are unexpectedly low relative to such a baseline. The analysis also does not establish non-equivalence under arbitrary invertible transformations or characterize the complete hidden-state representation.

A The causal role recurs across all five Mamba-1 scales

	130M	370M	790M	1.4B	2.8B
Selected / control	P3/P5	P3/P5	P2/P5	P5/P4	P3/P5
Causal role	**Supported**	**Supported**	**Supported**	**Supported**	**Supported**

Common scale-local contrast: selected restoration minus coefficient-matched control.
Historical endpoint labels and inferential-family details are retained in Appendix A.4.

B XG2 centered linear CKA

	130M	370M	790M	1.4B	2.8B
130M	1.00	0.74	0.44	0.34	0.56
370M	0.74	1.00	0.56	0.50	0.65
790M	0.44	0.56	1.00	0.66	0.64
1.4B	0.34	0.50	0.66	1.00	0.53
2.8B	0.56	0.65	0.64	0.53	1.00

C XG4 centered linear CKA

	130M	370M	790M	1.4B	2.8B
130M	1.00	0.54	0.45	0.48	0.54
370M	0.54	1.00	0.56	0.44	0.59
790M	0.45	0.56	1.00	0.44	0.39
1.4B	0.48	0.44	0.44	1.00	0.46
2.8B	0.54	0.59	0.39	0.46	1.00

Same 300 response-blind pairs per family. CKA = 1 denotes identical centered sample geometry; off-diagonal values show partial, not invariant, cross-scale geometry.
Cosine-RSM: Appendix A.2.

Figure 3: Operational causal-role recurrence with partial geometric conservation. (A) Frozen selected/control pairs differ by scale and are confirmed on held-out populations; rank labels are local, not semantic homologies. (B–C) Centered linear CKA on matched response-blind XG2/XG4 populations shows heterogeneous intermediate cross-scale similarity, with secondary cosine-RSM and resampling checks reported in the appendix.

3.3 Task-functional coupling reorganizes across scale

Even though the causal role recurs, the downstream task does not use the selected component in the same aggregate way at every scale. The mean forward-equivalent readout is positive at 130M, 370M, and 790M and negative at 1.4B and 2.8B:

$$(+0.00224,\ +0.00102,\ +0.00455,\ -0.00239,\ -0.00015).$$

We treat these point-estimate aggregate signs $+, +, +, -, -$ as a descriptive five-scale pattern, not as a parameter threshold or monotonic scaling law.

For pair i, the local readout is

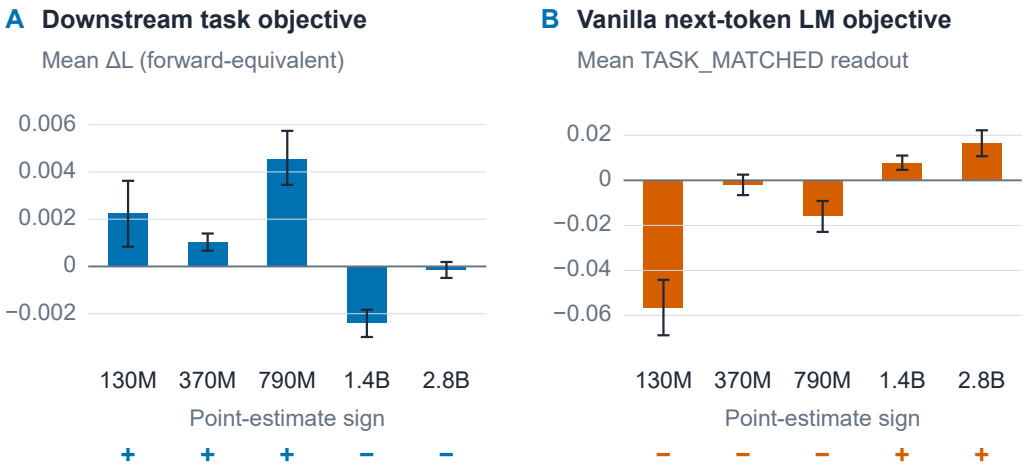
$$\Delta L_i = g_i^\top \left(c_{\mathrm{sel},i} - c_{\mathrm{ctrl},i}\right),$$

where g_i is the gradient of the downstream task margin with respect to the frozen local recurrent-state value. This makes ΔL explicitly objective-conditioned rather than an intrinsic property of the pretrained state.

The exact decomposition in Appendix A.5 separates gradient magnitude, component magnitude, angular alignment, and their dependence terms. It shows that no single geometric scalar explains the aggregate signs: at 370M dependence terms overcome a slightly negative mean alignment gap; at 1.4B negative task-gradient alignment dominates; and at 2.8B adverse dependence outweighs a positive mean gap. The same recurring native causal role can therefore couple differently to the downstream objective across scale.

3.4 Functional readout is objective- and contrast-conditioned

The task-objective pattern could still reflect a functional organization already present under Mamba's pretrained objective. We test that alternative by keeping each scale's source population, target site, selected component, and matched control fixed while replacing the downstream task margin with the pretrained next-token log-probability objective.



Separate y-scales; bars and signs are point means, whiskers are percentile 95% pair intervals.

C Pair-level associations remain mixed

	130M	370M	790M	1.4B	2.8B
Pearson	+0.3208	−0.4593	+0.1862	−0.3664	+0.1330
Spearman	+0.2044	+0.0346	−0.0379	−0.1827	+0.0966
Pair sign agreement	0.5567	0.6033	0.4433	0.4567	0.5733

All five scales are shown descriptively; the 130M LM cell is the later completeness extension (Appendix A.6).

Figure 4: Objective-conditioned readouts of the primary frozen scale-local contrasts. Bars are frozen point means; whiskers are percentile 95% pair-resampling intervals. Eight of ten scale-by-objective cells are individually sign-stable; pair-level associations exclude simple rowwise inversion. The aggregate comparison is contrast-conditioned, with sensitivity reported in Appendix A.6.

For the primary task-matched native contrasts, point-estimate aggregate signs are $+, +, +, -, -$ under the task objective and $-, -, -, +, +$ under the pretrained next-token objective (Fig. 4). Pair-resampling intervals exclude zero in 8 of 10 scale-by-objective cells; the 370M pretrained-LM and 2.8B task readouts cross zero and therefore lack individual sign stability under this analysis (Appendix A.7). The prospective LM-control family covered the four larger scales; the later 130M completeness extension is not retroactively pooled into it (Appendix A.6).

The opposed point-estimate means do not imply rowwise inversion: pair-level Pearson and Spearman associations vary in sign, and sign agreement remains near the middle of its range. More importantly, the aggregate result is *contrast-conditioned*. In the common-P3/P5 sensitivity comparison, the 1.4B pretrained-LM mean is negative whereas its primary task-matched P5/P4 pretrained-LM mean is positive (Appendix A.6). We therefore interpret the evidence as objective dependence of each frozen native contrast, not as a universal scale-level orientation or a contrast-invariant sign reversal.

3.5 Functional consequences extend beyond the local readout, but remain bounded

We finally ask how far the local mechanism propagates beyond the controlled recurrent-state analysis. These experiments serve as scope tests rather than a third headline contribution.

At 130M, the mechanism reaches the learned task computation: restoring PP3 relative to the matched PP5 replacement shifts the downstream task margin and tracks the local functional readout across

pairs, as established in Sec. 3.1. The effect is primarily margin-level rather than a discrete prediction improvement.

We additionally evaluate external transfer on the official AVeriTeC development set (Schlichtkrull et al., 2023) using a common 462-example three-label gold-evidence cohort. This experiment is available at 130M, 370M, and 1.4B only. The selected-versus-control intervention produces small positive correct-class margin shifts at 130M (2.31×10^{-4}, $d_z = 0.120$) and 370M (4.56×10^{-5}, $d_z = 0.106$); both positive-transfer hypotheses are supported under their frozen prospective family.

At 1.4B, the point estimate changes sign to -7.85×10^{-5}, but the pre-specified one-sided negative test does not pass the alpha gate ($p = 0.0757$). We therefore do not claim an established external sign reversal. The 1.4B intervention also changes no argmax predictions among the 462 examples, and all intervention conditions retain the same accuracy.

Because the frozen downstream model was not trained on AVeriTeC and this is not a benchmark-performance study, we use the experiment only as an out-of-distribution natural-language margin probe. The external evidence therefore supports a bounded conclusion: the native-state intervention can propagate to continuous task and natural-language decision margins, but the available evidence does not establish uniform behavioral improvement, scale-invariant external transfer, or benchmark-level correction.

4 METHODS

Design, sites, and native geometry. We analyze fixed pretrained Mamba-1 backbones from 130M to 2.8B. Principal-plane ranks are model-local. The 130M causal study froze intervention layer 17 of 24. Larger-scale target sites were not re-localized from XG1 outcomes: the intervention layer is mapped by relative depth, $\ell_s = \mathrm{round}[(17/(24-1))(L_s - 1)]$, giving layer 35 for the 48-layer 370M/790M/1.4B models and layer 47 for the 64-layer 2.8B model, while preserving source/target-residual offsets $(-2, -1)$. Geometry preparation uses native Mamba activations without XG1 outcomes, downstream task logits or gradients, or backward passes.

Each scale supplies frozen 300-row XG2/XG4 `alignment_delta_h` matrices. Row-normalized vectors give sample Grams $K_s = U_s U_s^\top$; normalized Frobenius products of centered Grams give linear CKA, while Pearson correlation of the strict upper cosine-Gram triangles gives secondary cosine-RSM. These frozen static values are transcribed without figure-builder recomputation; provenance and scalar measurements are in Appendix A.2.

Plane selection and causal confirmation. At 130M, P3 was selected from response-blind XG2/XG4 geometry before XG1 response inspection, and P5 was frozen independently as the maximum-projector-separation geometric control. Here the positive projector-contrast eigenvalue is $\lambda_k^+ = \sin\theta_k$; P5 is its global maximizer, and because the selected P3 is not that maximizer, P5 is also the maximizer over $k \neq 3$. Thus the 130M and later controls use the same response-blind geometric quantity, although the 130M control was frozen independently before fresh XG1 response inspection whereas the later rule is written conditional on the discovered k_s^*. For the later dominant-component protocol, all five reconstructed planes are evaluated on a dedicated discovery population and the selected rank is frozen as $k_s^* = \mathrm{unique}\,\mathrm{argmax}_k\ \overline{[Q_{\mathrm{restored}}(k) - Q_{\mathrm{neutralized}}(k)]}$. No positivity gate or inferential p-value is used in discovery. The control is selected separately from frozen geometry among non-selected planes, $c_s^* = \arg\max_{k \neq k_s^*} \lambda_{s,k}^+$. The selected/control pair is then frozen before a disjoint confirmation population is opened, where the pre-specified selected-restored versus matched-control endpoint is tested once. Confirmation responses cannot reselect the plane, control, tail, layer, offset, or epsilon. This design prevents same-sample argmax confirmation, while making no omnibus claim that the selected plane is superior to every alternative plane. Historical endpoints and the failed 370M residual criterion are retained rather than harmonized away (Appendix A.3–A.4).

Task-functional readout. The separately trained structured downstream model supplies a correct-class task margin. At each frozen native state, its local gradient defines $\Delta L_i = g_i^\top (c_{\mathrm{sel},i} - c_{\mathrm{ctrl},i})$. Selected and control components have numerically equal norms. The frozen branch-ownership con-

vention gives $\Delta L_{\text{forward}} = 2\Delta L_{\text{owned}}$, preserving signs and correlations. The exact algebraic decomposition is reported in Appendix A.5.

Pretrained-objective control. Within each scale, the native contrast, source population, and target site remain fixed while the objective becomes the original next-token log probability. The original prospective control covers 370M–2.8B; the already-studied 130M cell was completed later under a separate frozen extension. The five-scale display and pair-level associations are descriptive (Appendix A.6).

Behavior, external transfer, and inference. At 130M, restoration and control are compared on a continuous task margin. External transfer uses the fixed three-label AVeriTeC subset ($N = 462$) at 130M, 370M, and 1.4B. The 130M/370M positive tests form one frozen family; the later 1.4B negative test is separate. We neither pool them nor estimate a cross-scale zero crossing (Appendix A.8 gives serialization, tokenizer, and inference details).

5 Discussion

The strongest conclusion is procedural: across five Mamba-1 checkpoints, a scale-local component selected under a frozen protocol survives a pre-specified matched-control confirmation, despite changing principal-plane rank. This is recurrence of an operational interventional role, not evidence that the models contain one coordinate-identical or semantically homologous "same mechanism." The 370M aggregate-residual criterion remains a genuine historical failure and is not rescued by later confirmations.

Response-blind alignment geometry is neither identical nor absent across scales. Centered linear CKA and cosine-RSM show heterogeneous intermediate conservation, and source-pair resampling preserves that structure. Because we do not estimate a noise ceiling or cross-model null, the geometric claim is limited to lack of exact invariance under the stated comparison; it does not quantify how surprising the observed CKA range should be for independently reconstructed models of different sizes.

Functional readout is likewise conditional. The primary frozen contrasts have opposed task-versus-pretrained point-estimate sign vectors, with pair-resampling support in 8 of 10 scale-by-objective cells, but the 1.4B pretrained-LM sign changes under the common-P3/P5 sensitivity contrast. We therefore use *objective-conditioned functionalization* only to mean that a fixed native contrast can be read differently by different objectives; we do not claim a contrast-invariant or universal scale-level orientation.

The intervention reaches a continuous task margin at 130M and small external AVeriTeC margins at 130M/370M, while the 1.4B external negative point estimate misses its prospective gate. These scope tests do not establish benchmark improvement. More broadly, five checkpoints from one Mamba-1 family and one controlled mechanism constitute a carefully staged case study, not an architecture-general scaling law. Comparison with Mamba-2 or Transformer families remains future work.

Within these bounds, the study provides an empirical separation among three notions that are often conflated in cross-model mechanistic comparison: recurrence of an operational causal role, conservation of response-blind geometry, and objective-conditioned use of a frozen native contrast.

Reproducibility statement

Frozen backbone revisions, cohort identities, intervention sites, endpoint definitions, statistics, and figure-source mappings are specified in the appendix. The supplementary archive provides the frozen XG2/XG4 geometry tensors, pair-level objective values, robustness-analysis outputs, protocol and manifest files, environment requirements, and a verification script for reproducing the reported robustness summaries.

AI USE STATEMENT

Generative AI tools assisted related-work discovery, experimental planning, code development and debugging, artifact verification, interpretation of results, figure preparation, and manuscript drafting and editing. The authors reviewed all AI-assisted outputs and take responsibility for the final methods, results, citations, and claims.

REFERENCES

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10920–10948, 2025. URL https://proceedings.mlr.press/v267/ciernik25a.html.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8493–8502, 2022. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Danielle Ensign and Adrià Garriga-Alonso. Investigating the indirect object identification circuit in Mamba, 2024. URL https://arxiv.org/abs/2407.14008.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.04709.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

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*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Vamshi Sunku Mohan, Kaustubh Gupta, Aneesha Das, and Chandan Singh. Interpreting and steering state-space models via activation subspace bottlenecks. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.22719.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/cd86a30526cd1aff61d6f89f107634e4-Abstract-Datasets_and_Benchmarks.html.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/412e7a0143cb33d3a7a9b6abe8d90470-Abstract-Conference.html.

Junxuan Wang, Xuyang Ge, Wentao Shu, Qiong Tang, Yunhua Zhou, Zhengfu He, and Xipeng Qiu. Towards universality: Studying mechanistic similarity across language model architectures. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2J18i8T0oI.

A REPRODUCIBILITY DETAILS

This appendix records the frozen experimental results and static robustness analyses. Figure source classes are mapped in Appendix A.9. No additional model execution is involved; Appendix A.7 reports CPU-only resampling analyses of already frozen artifacts.

A.1 MODEL IDENTITIES, SITES, AND POPULATIONS

The five pretrained backbones are from the `state-spaces/` Hugging Face repositories shown in Table 1. Full revisions in scale order are `40e5d2bd7452abb3ca8fadbafe9131ee0e2c2f37`, `589179554943157be31701edd8b4558889276674`, `9822dd4b76af2bd9099b6ce2f19efd8329189a7e`, `6e46eae61c27280517feef46f536d16b91076f08`, and `96c48e0292b63f5346b6d30061af2551f7101e26`. The broader study covered five scales before the pretrained-LM control. Its initial prospective plan instantiated 370M–2.8B; the already-studied 130M cell was later frozen and completed as a separate extension. The five-scale LM vector is descriptive, not one prospective inferential family.

Table 1: Frozen backbone identities and functional-readout cohorts. Each cohort has 300 pairs with two cells per pair. The layer is the mixer intervention layer.

Scale	Repository suffix	Revision prefix	Layer	XG1 pairs
130M	mamba-130m-hf	40e5d2bd7452	17	2701–3000
370M	mamba-370m-hf	589179554943	35	4801–5100
790M	mamba-790m-hf	9822dd4b76af	35	7501–7800
1.4B	mamba-1.4b-hf	6e46eae61c27	35	4801–5100
2.8B	mamba-2.8b-hf	96c48e0292b6	47	6601–6900

The larger-scale source-block, target-residual, and intervention-layer triplets are $(33, 34, 35)$ for 370M–1.4B and $(45, 46, 47)$ for 2.8B. At 130M the intervention is at layer 17. The frozen target position is the absolute event anchor plus offset 2, with `A_IDENTITY` at 130M. The pretrained-LM measurement reuses each task-readout row, its tokenization, and its target index; both token positions t and $t+1$ must be active. The frozen protocol fails closed rather than dropping a token-invalid row. At the mixer `in_proj`, the native content branch is replaced by an equal-valued differentiable leaf while the gate branch is preserved.

The structured downstream task model is a separately trained three-way evidence-entitlement model used for the local correct-class margin. The 130M downstream model is the seed-181, selected-epoch-19 checkpoint. Each larger scale uses its corresponding frozen downstream checkpoint. The frozen 130M, 790M, 1.4B, and 2.8B training reports identify the frozen Mamba encoder, seed 181, pair-group split seed 8192, 2,880 expected clean training rows, clean-development checkpoint selection, and no external data used for training. Their selected epochs are 19, 20, 20, and 20 respectively. The 370M checkpoint is identified separately by its compact checkpoint manifest; these training-report details are not inferred for it. Under the frozen branch-ownership convention the owned local derivative is one half of the forward-equivalent derivative: $\Delta L_{\text{forward}} = 2\Delta L_{\text{owned}}$. This conversion preserves sign, rank, and correlation. The pretrained-LM control loads the pinned causal LM directly, without a downstream task checkpoint, router, or task head.

A.2 NATIVE GEOMETRY AND INTERVENTION

XG2/XG4 response-free directions define independent native geometry within each model's strong-channel mask. Geometry preparation uses Mamba-side activations without XG1 response access, downstream task logits or gradients, or backward passes. The inputs belong to the controlled research families and therefore do not characterize all natural language.

The historical scalar geometry measurements remain useful supporting evidence. For $k_j = \texttt{conv_weight}[j, 0, \text{lag0}]$, $\mu_{k^2} = \text{mean}_j\, k_j^2$ and $\text{strong} = \{j : k_j^2 > \mu_{k^2}\}$. At 130M, $\mu_{k^2} = 0.027899337798707836$, 395 of 1536 channels are strong, and the response-blind projector reconstruction has leading positive eigenvalue $\lambda_1 = 0.8706181418918275$. The later matched

370M–2.8B scalar series is reported in Appendix A.4. These measurements are no longer the main evidence for cross-scale geometric non-invariance.

The primary coordinate-insensitive comparison uses the already frozen `alignment_delta_h` matrices. For each scale s and family, let $D_s \in \mathbb{R}^{300\times d_s}$ contain the 300 ordered pair vectors and row-normalize it to U_s. Define

$$K_s = U_s U_s^\top, \qquad H = I - \frac{1}{300}\mathbf{1}\mathbf{1}^\top, \qquad \widetilde{K}_s = HK_sH.$$

For scales s, t, centered linear CKA is

$$\mathrm{CKA}(s,t) = \frac{\langle \widetilde{K}_s, \widetilde{K}_t\rangle_F}{\|\widetilde{K}_s\|_F \, \|\widetilde{K}_t\|_F}.$$

XG2 and XG4 are analyzed independently. Every scale uses the same ordered source-pair IDs 301–600 within its family. The 130M matrices have historical provenance; 370M–2.8B come from the later homogeneous geometry-preparation series. No model forward, tokenizer, training, evaluation, backward pass, permutation test, threshold, or scaling-law fit is involved in this static comparison.

Table 2: Frozen coordinate-insensitive pairwise geometry. CKA is the primary centered-linear comparison. RSM is the secondary Pearson correlation between strict upper triangles of the row-normalized cosine Gram matrices. Values are descriptive.

Scale pair	XG2 CKA	XG4 CKA	XG2 RSM	XG4 RSM
130M–370M	.742546	.542295	.716312	.509857
130M–790M	.435747	.451862	.398941	.431574
130M–1.4B	.342510	.481304	.295730	.452640
130M–2.8B	.555414	.544277	.522254	.511619
370M–790M	.564059	.558569	.541258	.533989
370M–1.4B	.504483	.436834	.474489	.403381
370M–2.8B	.649344	.590983	.634246	.545762
790M–1.4B	.663678	.444332	.640851	.412623
790M–2.8B	.639963	.392717	.620152	.337620
1.4B–2.8B	.532185	.455698	.504037	.382808

Linear CKA is insensitive to orthogonal feature rotations and isotropic scaling, but not to arbitrary invertible transformations. The result therefore establishes non-invariance of this response-blind alignment geometry under the stated comparison, not global non-equivalence of complete hidden-state representations.

For a unit direction w, local susceptibility is

$$J_i(w) = \frac{F_i(+\epsilon; w) - F_i(-\epsilon; w)}{2\epsilon}, \qquad \epsilon = 0.025. \tag{1}$$

At 130M, the projector-contrast contribution of plane k is $C_k(i) = s_k[J_i(\mathrm{P}_{k+})^2 - J_i(\mathrm{P}_{k-})^2]/5$. The broader XG2-versus-XG4 endpoint is $Q_i(c) = E_{\mathrm{XG2}}(i,c) - E_{\mathrm{XG4}}(i,c)$, where each E is the mean squared susceptibility over its five frozen directions under condition c. XG1 outcomes do not refit the basis.

A.3 Prospective 130M causal chain

Static XG2/XG4 localization selected P3 before XG1 response inspection. P5 was frozen independently as the maximum-projector-separation geometric control. Four disjoint 300-pair XG1 holdouts tested the distinct endpoints in Table 3. Every directional hypothesis was frozen before its population was opened.

For restoration, from native state h and P3 component c_3, $B = h - c_3$, $R_3 = B + c_3 = h$, and $R_5 = B + c_5$. The P5 replacement uses the same P3-derived coefficients, so additions are norm matched. The result is local restoration relative to the control, not global behavioral sufficiency. Later P3-excluded residual-template transport and residual necessity/restoration studies are documented in the frozen causal synthesis and did not redefine P3. The historical restoration-sufficiency endpoint

Table 3: Frozen 130M paired-item tests. Endpoint magnitudes are not interchangeable effect sizes.

Test	XG1 pair range	Mean	$t(299)$	One-sided p
Transport C_{P3}	001–300	7.12028×10^{-8}	18.974	1.52×10^{-53}
Specificity $C_{\mathrm{P3}} - C_{\mathrm{P5}}$	301–600	2.30407×10^{-8}	7.656	1.34×10^{-13}
Necessity $Q_{\mathrm{P5}} - Q_{\mathrm{P3}}$	601–900	4.74144×10^{-8}	23.984	6.22×10^{-72}
Restoration $Q_{R_3} - Q_{R_5}$	901–1200	4.07887×10^{-8}	20.105	9.01×10^{-58}

reduces exactly to the common selected-restored versus matched-control contrast:

$$\begin{aligned} D_{\mathrm{SUF}} &= (Q_{R_3} - Q_B) - (Q_{R_5} - Q_B) \\ &= Q_{R_3} - Q_{R_5} \\ &= Q_{\text{selected restored}} - Q_{\text{matched control}}. \end{aligned} \tag{2}$$

On the separate seed-181 behavioral/readout cohort `xg1_fact_2701..3000`, the mean task-margin restoration effect is 0.008832191656033197. Descriptive pair-level readout–behavior association is Pearson 0.617548086733788, Spearman 0.8324696941077123, and sign agreement 0.8233333333333334.

A.4 Cross-scale causal confirmation and native geometry

Selected/control ranks in ascending scale order are P3/P5, P3/P5, P2/P5, P5/P4, and P3/P5. The historical 130M D_{SUF} and 370M D_{DOM} endpoints preceded the later standardized $D_{\mathrm{CORE}} = Q_{\text{restored}} - Q_{\text{control}}$ confirmations. By Eq. 2, all five share the mathematical contrast $Q_{\text{selected restored}} - Q_{\text{matched control}}$ and the same $Q = E_{\mathrm{XG2}} - E_{\mathrm{XG4}}$ semantics; selected-state coefficients are transferred to the orthonormal control plane, matching component norms. The five experiments share a common scale-local causal contrast while retaining distinct historical confirmation protocols. Table 4 reports their own frozen values. Heterogeneous raw magnitudes do not form a scaling series.

Table 4: Scale-local causal-role tests. Each row has a positive frozen endpoint, but the historical 370M joint residual criterion remains failed.

Scale	Pair	Endpoint	XG1 pairs	Mean	$t(299)$	One-sided p
130M	P3/P5	D_{SUF}	901–1200	4.07887×10^{-8}	20.105	9.01×10^{-58}
370M	P3/P5	D_{DOM}	3301–3600	3.97429×10^{-8}	26.056	3.19×10^{-79}
790M	P2/P5	D_{CORE}	7201–7500	7.19308×10^{-8}	29.567	4.65×10^{-91}
1.4B	P5/P4	D_{CORE}	4201–4500	9.28285×10^{-9}	16.253	2.62×10^{-43}
2.8B	P3/P5	D_{CORE}	6301–6600	4.11599×10^{-8}	24.956	2.22×10^{-75}

The 370M raw dominant-component p was 3.19293×10^{-79} and its historical Holm-adjusted value was 6.38586×10^{-79}. Its separate positive aggregate-residual hypothesis failed, so the pre-registered joint qualitative-recurrence criterion remained not established. Later core confirmations do not retroactively rescue it. The 1.4B discovery and confirmation populations were XG1 3901–4200 and 4201–4500, with residual characterization on 4501–4800.

The later matched response-blind series, ordered 370M, 790M, 1.4B, 2.8B, has $\mu_{k^2} = (0.0246235, 0.0238427, 0.0148388, 0.00985711)$, leading spectral concentration $\lambda_1 = (0.967180, 0.892961, 0.889041, 0.893226)$, and strong-channel fractions $(0.317383, 0.317383, 0.202393, 0.195898)$. The historically reconstructed 130M values for these two plotted quantities are given in Appendix A.2. Their mathematical definitions match, but their provenance precedes the later bridge. There is no fitted scale law or same-numbered-plane homology.

A.5 FULL TASK-READOUT DECOMPOSITION

The selected and control components have equal norm to numerical precision; the maximum relative mismatch across frozen rows is about 10^{-15}. Put $u_i = \|g_i\|$, $n_i = \|c_i\|$, and $h_i = \cos(g_i, c_{\mathrm{sel},i}) - \cos(g_i, c_{\mathrm{ctrl},i})$. Then $\Delta L_i = u_i n_i h_i$ exactly, and

$$\mathbb{E}[unh] = \mu_u\mu_n\mu_h + \mu_u \operatorname{Cov}(n,h) + \mu_n \operatorname{Cov}(u,h) + \mu_h \operatorname{Cov}(u,n) + \mathbb{E}[(u-\mu_u)(n-\mu_n)(h-\mu_h)]. \tag{3}$$

Table 5 transcribes the frozen owned-unit components. Forward-equivalent readouts are twice these values.

Table 5: Frozen mean angular gap μ_h, mean-product base B, and norm–gap covariance contribution N, in owned readout units.

Scale	μ_h	B	N
130M	.008881965010493697	.001278782557357266	.000133190457109135
370M	-.000802156327488121	-.000094095358107031	.000042452455047036
790M	.007793169921091330	.001040616136353141	-.000068706620106335
1.4B	-.009643894952004040	-.000877367363000991	-.000176563578561815
2.8B	.004473173637057720	.000329550929066379	-.000171543926043767

Table 6: Remaining exact terms: gradient–gap covariance G, gradient–norm covariance U, centered three-way term T, and owned mean readout. Together with Table 5, these give the full decomposition of Eq. 3.

Scale	G	U	T	$\overline{\Delta L}$
130M	-.000519576976183540	-.000012825062806655	.000240795061055922	.001120366036532126
370M	.000402747883547661	-.000001430912906175	.000159282284303926	.000508956351885417
790M	.001426961516464109	-.000040009026410042	-.000082150556444377	.002276711449856494
1.4B	-.000101775327135446	-.000010052654126476	-.000029688583061496	-.001195447505886224
2.8B	-.000162660052525846	-.000009378023827281	-.000060143880171834	-.000074174953502349

At 370M favorable dependence outweighs a negative mean alignment gap. At 1.4B negative alignment dominates. At 2.8B adverse dependence overcomes a positive mean gap; 55.3% of pairs remain positive. This is an algebraic decomposition, without fitted parameters or new tests.

A.6 PRETRAINED-LM CONTROL AND SENSITIVITY

At frozen target t, with observed next token $y = x_{t+1}$, the control uses $J_{\mathrm{LM}} = \log p_\theta(y \mid x_{\le t})$ and $g_{\mathrm{LM}} = \nabla_h J_{\mathrm{LM}}$ at the equal-valued local leaf. Table 7 records the frozen objective comparison. Means have distinct units; pair-level associations are descriptive.

Table 7: Task-matched frozen contrasts under the downstream task and pretrained next-token objectives. The pre-existing 130M scale was completed under a separately frozen LM-control extension and joins the four original rows descriptively.

Scale	Task forward	LM mean	Pearson	Spearman	Sign agree
130M	.00224073	-.05641685	.3208	.2044	.5567
370M	.00101791	-.00197532	-.4593	.0346	.6033
790M	.00455342	-.01596436	.1862	-.0379	.4433
1.4B	-.00239090	.00783318	-.3664	-.1827	.4567
2.8B	-.00014835	.01644840	.1330	.0966	.5733

The common-P3/P5 sensitivity means are -0.00197532 (370M), -0.00443947 (790M), -0.00990855 (1.4B), and $+0.01644840$ (2.8B). The 1.4B sign differs from its primary task-matched P5/P4 result. This contrast sensitivity does not reselect the primary comparison. Opposed aggregate signs are not a rowwise inversion.

A.7 PAIR-RESAMPLING ROBUSTNESS OF OBJECTIVES AND GEOMETRY

The frozen CPU-only `source_pair_id` bootstrap uses $B = 10{,}000$ replicates and percentile 95% intervals. Zero lies within the intervals for 370M pretrained LM (-0.00197532; $[-0.00655147, 0.00253593]$) and 2.8B task ($-0.00014835$; $[-0.00048602, 0.00018915]$). These point signs lack individual pair-resampling stability; the five-scale vectors remain descriptive, without a new prospective inferential family.

Geometry shares bootstrap multiplicities across five scales within each family; XG2 and XG4 are separate. All 20 off-diagonal CKA entries retain their qualitative intermediate range. In $1{,}000$ balanced 150/150 split halves, Pearson, Spearman, mean absolute difference (MAD), and root mean square error (RMSE) compare ten off-diagonal CKA values. Frozen medians in Table 8 show stable overall structure on the controlled pairs, with weaker fine-grained ordering for XG4.

Table 8: Frozen median correlations and differences between the ten off-diagonal CKA values across 1,000 balanced split halves.

Family	Pearson	Spearman	MAD	RMSE
XG2	.875577	.878788	.057331	.068640
XG4	.652928	.672727	.056955	.068166

These summaries describe sampling robustness on frozen cohorts, not a noise ceiling, model uncertainty, or independent replication. Analysis outputs, CSVs, and the protocol are frozen as a byte-identified supplementary snapshot; SHA-256 identities are provided in the supplementary manifest.

A.8 AVERITEC EXTERNAL-TRANSFER INFERENCE

The authenticated official AVeriTeC development source is `MichSchli/AVeriTeC`, commit `7c62d1ec8df3fb560d6efe2b85fa191135636f81`, path `data/dev.json`, Git blob `40974243267f395dc583d805d10f043812419249`. Of 500 development examples, the fixed compatible cohort comprises 305 Refuted, 122 Supported, and 35 Not Enough Evidence examples ($N = 462$). The 38 Conflicting Evidence/Cherrypicking examples are excluded by the pre-specified three-label compatibility rule. Gold question–answer blocks retain source order. There is no retrieval or question generation in this experiment. The 130M/370M family uses the frozen active serialization `claim[:63] + EOS(0) + evidence[:64]`, response-blind claim–evidence-boundary anchor, and target offset 2. Its shared `tokenizer.json`, `tokenizer_config.json`, and `special_tokens_map.json` SHA256 identities are, respectively, `b074ad869d4f45d1265ca5c9814f78604f3d7e187acc063b15dd232b27585fcf`, `9d7016c33747c6309346e59bd7bf63bfc33c9d9366ecb7e514b3b84dc6b46acb`, and `57491904f8680d4b52ed440f1f7ba48cad1c31ecf3eb453b03484e6ff4723ae8`. The active-encoding gate used tokenizers version 0.22.2 and required all 462 rows to pass. The 1.4B external-transfer extension uses its own pinned tokenizer snapshot at revision `6e46eae61c27280517feef46f536d16b91076f08`; `tokenizer.json` SHA256 is `3cf430678137c8491ca82fb7092ee49e44ad38857fffe1e4a4a5ed860139a5b8`, and `tokenizer_config.json` SHA256 is `3ba257483d22a5a84aab5465aa427e59bdaeb55f09fb14349e2d571ff67e8020`. It is not byte-identical to the tokenizer used for the completed 130M/370M external family. The response-blind anchor is the claim–evidence boundary, with intervention target at consumed claim length plus 2.

Correct-class-margin selected-versus-control means at 130M and 370M were 0.00023104869 and 0.000045550022, with Cohen's $d_z = 0.11964887$ and 0.10631035, and Holm-adjusted positive-family $p = 0.010430705$ and 0.0113815. The separately frozen 1.4B negative test had mean -0.000078509516, $d_z = -0.0668655$, $t(461) = -1.43722$, and one-sided $p = 0.0756671$, failing the $\alpha = .05$ gate. Its native, neutralized, and control conditions all had accuracy 0.12987013, with zero intervention-induced argmax flips. The three scales are not pooled into a new inferential family.

A.9 FIGURE PROVENANCE AND SOURCE MAPPING

Figure 2 uses frozen 130M causal, specificity, seed-181 behavioral, and pair-merge records. Figure 3 combines five-scale causal support and the frozen coordinate-free XG2/XG4 geometry snapshot; historical scalar geometry stays in Appendix A.2. Figure 4 combines task readouts, four-scale LM control, later 130M completeness, and percentile 95% pair intervals from `objective_bootstrap.csv` in the frozen objective pair-resampling snapshot. Figure 1 summarizes the design.