On the Geometric Structure of Token–Token Interactions in Deep Language Models

Jun-Sik Yoo
Department of Physics, Korea University · junsik@korea.ac.kr

ICML 2026
Workshop on Mechanistic Interpretability

Abstract

We introduce an operational framework for measuring and controlling token–token interaction in language models through controlled perturbations. Holding a target token and syntactic frame fixed, we vary contextual tokens and residualize each layer update against a restricted tokenwise approximation. Across semantic categories, targets, and syntactic templates, the residual — not the raw update — carries structured, category-conditioned variation. These residual directions are causally meaningful: injecting them produces monotonic, category-selective steering, and they transport across target tokens. We validate the framework on Transformer (Pythia, LLaMA) and state-space (Mamba) models from 160M to 70B parameters.

Key Contributions

- **Framework.** An operational way to measure token–token interaction via controlled perturbations of non-target tokens, target and syntax held fixed.
- **Decomposition.** A tokenwise–residual split that isolates interaction directions with statistically significant, category-conditioned structure.
- **Causal control.** Residual directions support reliable steering; raw activation-contrast directions fail under token-local contamination.
- **Scale & architecture.** Validated on Transformer and state-space models, 160M–70B parameters, plus a cross-target transport test.

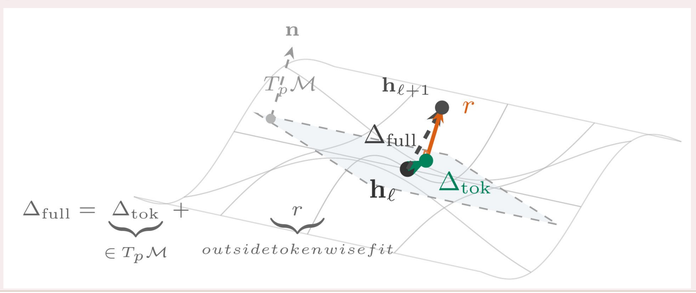
Motivation & Method

Q: Where does token–token interaction live inside a model?

Attention weights need not reflect causal effect and are unavailable in attention-free architectures; circuit and model-editing methods are often architecture- or case-specific; activation-steering directions mix contextual effects with token-local variation. We define interaction operationally: fix a target token and syntactic frame, perturb the context, and measure the induced change in the target representation.

$h_{\ell+1} = T(h_\ell) + r(h_\ell)$; *T = restricted tokenwise map · r = residual (candidate interaction signal)*

Residual geometry of a layer update



Q: Can interaction directions be measured and used for causal steering?

Controlled perturbation template

{Subject} is / seems like / is not a {Modifier} {Target}
5 targets × 7 categories × 8 modifiers × 8 subjects × 3 templates

- **Subject centering.** $\tilde{r} = r - E[r \mid \text{subject}]$ removes subject-identity bias, isolating modifier-induced variation.
- **Interaction gap.** gap = E[sim_within] − E[sim_between]: cosine similarity of residuals within vs. across semantic categories; significance via permutation test.
- **Residual intervention.** $h'_\ell = h_\ell + \alpha \cdot r_c$ injects a category centroid; we measure KL divergence and category selectivity.
- **Transport consistency.** An orthogonal Procrustes map fit on anchor modifiers is evaluated on held-out modifiers, across target tokens.

Table 1. Experimental setup

Targets	5 (country, city, animal, person, object)
Semantic categories	7 (Geography, Politics, Size, Sentiment, Quantity, Temporality, Modality)
Modifiers / category, Subjects / target	8, 8
Template variants	3 (is a / seems like a / is not a)
Architectures	Pythia, LLaMA, Mamba
Scale range	160M – 70B parameters

Results · Interaction Structure Across Architectures & Scales

- **Stable across scale.** Both Transformer and Mamba models show stable, statistically significant interaction gaps from 160M to 70B parameters (permutation $p < 0.001$), with no strong monotonic scale dependence within a family — interaction is a structural property of sequence models, not an effect that only emerges at scale.
- **Robust to syntax.** Gaps hold across all three templates, including the negation variant (is not a), indicating the effect reflects model representations rather than surface syntax.
- **Beyond templates.** A natural-text probe on UN Debates speeches finds smaller but significant gaps (Pythia-1B: 0.096, Mamba-1.4B: 0.099; both $p < 0.001$).

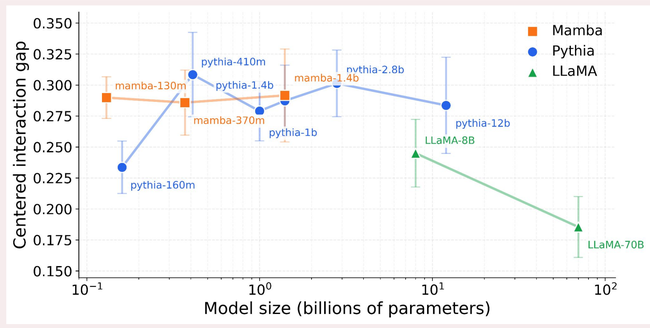


Figure 1. Interaction gap vs. model size, Transformer (Pythia, LLaMA) and Mamba, 160M–70B params. Error bars: SD across targets.

Results · Causal Steering & Comparison with Activation Addition

Injecting category-specific residual directions produces monotonic, category-selective steering: Transformer models steer strongly and monotonically in Geography and Politics; Mamba shows weaker but still structured effects.

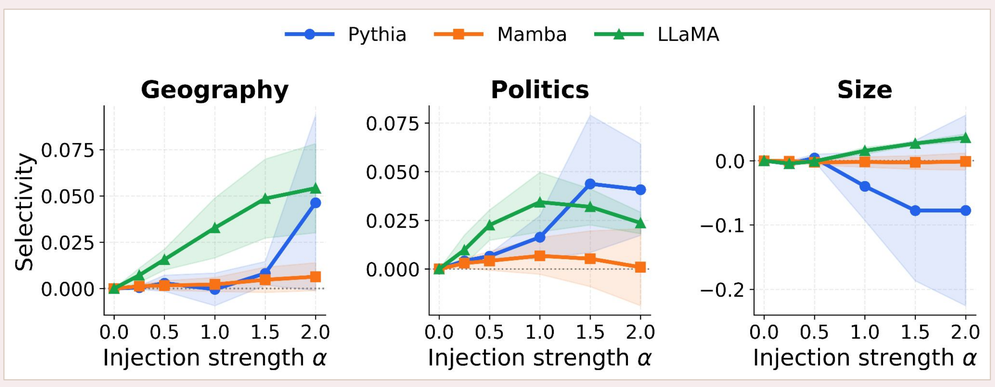


Figure 2. Selectivity as a function of injection strength α, family-averaged across Geography, Politics, and Size. Shaded regions: SD across models.

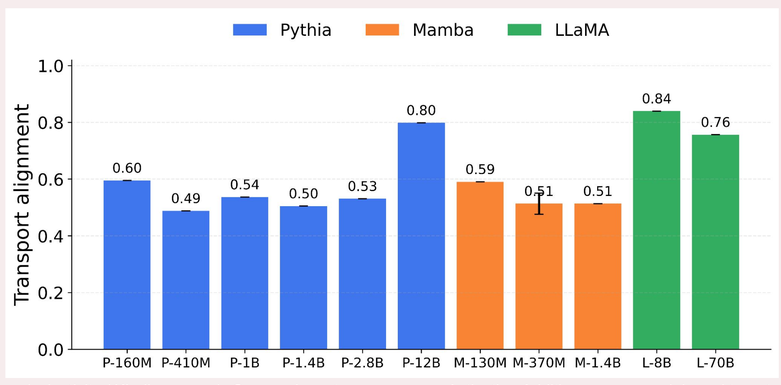


Figure 3. Steering efficiency (selectivity / KL divergence), Geography category, ours vs. Activation Addition, across models and targets.

Comparison with Activation Addition. Raw hidden-state contrast directions retain token-local components. In the Geography category, our residual-based directions yield consistently positive steering efficiency across all evaluated models (Pythia-1B through LLaMA-3.3-70B), while Activation Addition yields negative efficiency in every case.

Results · Category-Conditioned Transport Consistency

- We fit an orthogonal Procrustes map between residual subspaces of two target tokens on anchor modifiers, then evaluate cosine alignment on held-out modifiers.
- **Transfers across targets.** All architectures show above-chance transport generalization, with alignment ranging 0.49–0.84; LLaMA generalizes most strongly, Pythia and Mamba moderately but consistently.
- This indicates interaction directions carry category-conditioned structure that is not purely target-specific — it transfers across targets to varying degrees by architecture.

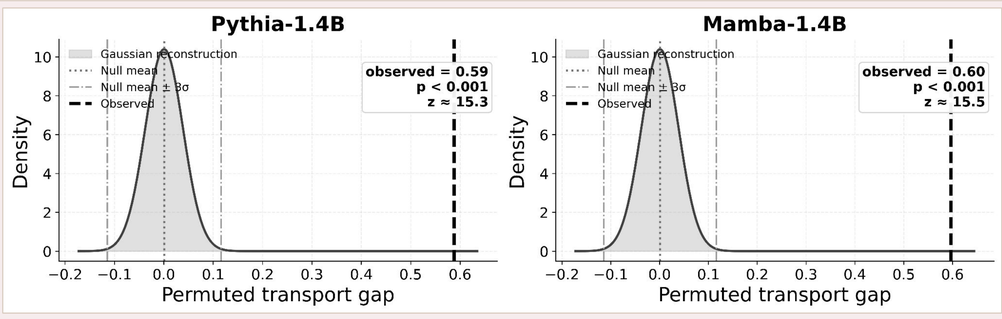


Figure 4. Transport generalization (cosine alignment) across architectures, held-out categories.

Discussion & Limitations

Residuals as structured interaction

- Under controlled perturbation, residuals are structured, statistically significant, and causally effective — not merely leftover computation.
- Architectures differ less in whether interaction structure exists, more in how accessible it is to localized residual intervention (Transformer: stronger, more monotonic; Mamba: weaker but structured).

Limitations

- Simple prompt templates and a limited set of semantic categories, though validated across 160M–70B parameters.
- The interaction definition relies on controlled perturbations and may miss long-range or compositional effects.
- Results depend on the chosen tokenwise approximation class; category-wise steering can still fail when categories are not well isolated.

Conclusion

Token–token interaction is consistently measurable via residual-based metrics, statistically significant under permutation testing, causally effective through residual injection, and structured in a category-conditioned way that transfers across targets — across Transformer and state-space models from 160M to 70B parameters. Residual geometry, relative to a restricted tokenwise approximation, offers a scalable, architecture-agnostic interface for analyzing and intervening on contextual computation.

Selected References

[1] Elhage et al. A mathematical framework for transformer circuits. Transformer Circuits, 2021.
[2] Meng et al. Locating and editing factual associations in GPT. NeurIPS, 2022.
[3] Meng et al. Mass-editing memory in a transformer. arXiv:2210.07229, 2022.
[4] Turner et al. Activation addition: steering LMs without optimization. arXiv:2308.10248, 2023.
[5] Gu & Dao. Mamba: linear-time sequence modeling with selective state spaces. 2023.
[6] Jain & Wallace. Attention is not explanation. NAACL, 2019.
[7] Vig. Analyzing the structure of attention in a transformer LM. EMNLP, 2019.
[8] Yoo. On the geometric structure of layer updates in deep LMs. arXiv:2604.02459, 2026.