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

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce an operational framework for measuring and controlling token–token interactions in language models through controlled perturbations of non-target tokens. Holding the target token and syntactic frame fixed, we vary contextual tokens systematically and measure how the target representation changes. This setting lets us identify interaction effects directly, rather than inferring them from attention weights, architecture-specific components, or raw hidden-state contrasts.

Our key idea is to residualize layer updates with respect to a restricted tokenwise approximation class. We decompose each target-token update into a token-local component and a residual, then ask whether category-conditioned variation induced by controlled perturbations concentrates more strongly in the residual. Across semantic categories, targets, and syntactic templates, we find that it does: residual directions capture structured interactional variation more cleanly than raw update directions, while directions derived without tokenwise subtraction retain token-local contamination that degrades, and in some cases reverses, steering precision.

We then test whether these residual directions are causally meaningful. Injecting category-specific residual directions produces monotonic, category-selective changes in model outputs in settings where the interaction signal is cleanly isolated. We validate this framework across Transformer (Pythia, LLaMA) and state space (Mamba) models from 160M to 70B parameters, showing statistically significant interaction structure under permutation testing, reliable causal steering through residual injection, and category-conditioned transport consistency across target tokens. Together, these results support residualized layer-update geometry as a practical, architecture-agnostic interface for analyzing and intervening on contextual computation.

1 Introduction

Understanding how language models integrate information across tokens remains a central problem in interpretability. Existing approaches provide important but partial views of this problem. Attention-based analyses use attention weights or attention flow as proxies for token influence [12, 1], but such quantities need not correspond to causal effect [8] and are unavailable in attention-free architectures [7]. Circuit-level and model-editing methods can identify detailed mechanisms in specific settings [4, 5, 6, 9, 10], but often rely on architectural structure or case-by-case analysis. Activation-steering methods provide a practical causal interface, but their directions are typically derived from raw hidden-state contrasts, which can mix contextual effects with token-local variation [11].

We study token–token interaction through controlled perturbation. We fix a target token and syntactic frame, vary non-target tokens systematically, and measure the induced change in the target-token representation. This yields an operational setting in which cross-token effects can be measured,

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

compared, and causally tested. More broadly, we argue that a useful framework for token–token interaction should ideally satisfy five desiderata: it should be *scaling-friendly*, *architecture-agnostic*, *causally grounded*, *principle-based*, and *geometrically coherent*. Existing approaches typically satisfy only a subset of these desiderata.

The central technical object in our framework is the decomposition

$$h^{\ell+1} = T(h^{\ell}) + r(h^{\ell}), \tag{1}$$

where T is a restricted tokenwise approximation and r is the residual not captured by this token-local component. Rather than identifying residuals with interaction by definition, we use this decomposition as a local coordinate system: after subtracting the best available tokenwise component within a restricted approximation class, we test whether the remaining residual varies systematically with controlled changes in other tokens. This provides a unified interface for measuring interactional structure and for intervening on it through residual injection.

Our use of controlled templates is deliberate. The templates are not intended to exhaust naturally occurring language, but to serve as an identification strategy. By holding the target token and syntactic frame fixed while varying other tokens, we can attribute systematic changes in the target representation to non-target context. We complement this controlled setting with a small robustness probe on natural text. Our work also builds on prior geometric analysis of layer updates [13], which identified the residual as a candidate space for non-tokenwise computation. Here we take the next step by operationalizing token–token interaction through controlled perturbations and testing whether the resulting residual directions support structured and causal control.

Concretely, we make four contributions:

1. We introduce an operational framework for measuring token–token interaction through controlled perturbations of non-target tokens while holding the target token and syntactic frame fixed.
2. We use a tokenwise–residual decomposition to isolate interaction-related residual directions and show that these directions exhibit statistically significant category-conditioned structure.
3. We demonstrate that residual directions support causal intervention through residual injection, and identify settings in which raw activation-contrast directions fail due to token-local contamination.
4. We evaluate the framework across Transformer and state space models, from small models to 70B parameters, and test cross-target structure through a category-conditioned transport experiment.

Taken together, our results support treating residual geometry, relative to a restricted tokenwise approximation class, as a scalable and architecture-agnostic interface for analyzing token–token interaction.

2 Method

2.1 Problem Setup

We consider a language model with layer representations $h_\ell \in \mathbb{R}^d$. Our goal is to quantify how non-target tokens influence the representation of a fixed target token.

Given an input sequence, we focus on the representation of a designated target token at layer ℓ, and study its update:

$$h_{\ell+1} = F(h_\ell) \tag{2}$$

This decomposition follows prior work on geometric analysis of layer updates [13], where the residual component was shown to capture deviations from tokenwise transformations.

We aim to isolate the component of this update that arises from interactions with other tokens.

2.2 Controlled Perturbation Framework

We construct inputs using templates of the form ⟨*Subject*⟩ `is a` ⟨*Modifier*⟩ ⟨*Target*⟩, where the target token and syntactic frame are held fixed, the modifier is varied across semantic categories, and the

subject is varied independently. This ensures that any change in the target representation arises solely from controlled perturbations of non-target tokens. We define interaction through response to such perturbations.

We evaluate five target tokens (`country`, `city`, `animal`, `person`, `object`) and seven semantic categories (Geography, Politics, Size, Sentiment, Quantity, Temporality, Modality), each with eight modifiers and eight subjects. To assess robustness to syntactic framing, each target is evaluated under three template variants:

```
{Subject} is a {Modifier} {Target}
{Subject} seems like a {Modifier} {Target}
{Subject} is not a {Modifier} {Target}
```

Consistent interaction structure across these templates — including the negation variant — provides evidence that the observed effects reflect model representations rather than surface-level syntactic patterns. Unless otherwise noted, all interaction-gap analyses in the paper use this same five-target, seven-category, three-template evaluation setup.

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

Table 1: Experimental setup for all interaction-gap analyses.

2.3 Tokenwise Decomposition

To separate token-local effects from interactional effects, we decompose the layer update as:

$$h_{\ell+1} = T(h_\ell) + r(h_\ell) \tag{3}$$

where T is a tokenwise function and r is the residual component. We approximate T using low-rank linear maps $T(x) = UVx + b$ for Transformers and diagonal PSD maps $T(x) = \mathrm{diag}(s)x + b$ for Mamba. The tokenwise map captures input-conditioned but token-local transformations; the residual r captures deviations that cannot be explained by such restricted function classes.

2.4 Interactional Residual

We define a target-token residual as *interactional* if it changes systematically under controlled perturbations of non-target tokens while the target token and frame are held fixed. This definition is architecture-agnostic and directly testable via input perturbations.

2.5 Centering and Interaction Geometry

Residuals contain both interactional effects and entity-specific bias (e.g., subject identity). To isolate the former, we apply subject-level centering $\tilde{r} = r - \mathbb{E}[r \mid \text{subject}]$, which removes variation attributable to subject identity and retains only the variation induced by modifier perturbations.

2.6 Interaction Gap Metric

We group modifiers into semantic categories and compute cosine similarity between residual vectors. Let $\text{sim}_{\text{within}}$ and $\text{sim}_{\text{between}}$ denote mean cosine similarity within and across categories, respectively. The interaction gap is:

$$\text{gap} = \mathbb{E}[\text{sim}_{\text{within}}] - \mathbb{E}[\text{sim}_{\text{between}}] \tag{4}$$

A large positive gap indicates that residual directions cluster by semantic category, implying structured interaction. Statistical significance is assessed via permutation tests: we randomly permute category labels, recompute the gap, and compare the observed value to the resulting null distribution to obtain a p-value.

2.7 Residual Intervention

To test whether interactional residual directions play a causal role, we inject a category-specific direction r_c (the centroid of centered residuals for category c) into the target-token representation:

$$h'_\ell = h_\ell + \alpha r_c \tag{5}$$

where α controls intervention strength. We then propagate h'_ℓ through the remaining layers and measure (i) KL divergence between original and perturbed output distributions, and (ii) category selectivity, defined as the increase in likelihood of category-aligned continuations relative to competing categories.

2.8 Category-conditioned transport consistency

To probe whether interaction directions are structured across targets, we define a transport-based consistency measure.

For each semantic category, we construct a set of modifiers and randomly split them into an anchor set and a held-out set. Using the anchor modifiers, we compute residual directions for two target tokens (e.g., *city* and *country*) and fit an orthogonal Procrustes map P between the corresponding subspaces.

We then evaluate this map on held-out modifiers by measuring cosine alignment:

$$\text{Transport}(m) = \cos(P\, v_m^{(A)},\ v_m^{(B)})$$

where $v_m^{(A)}$ and $v_m^{(B)}$ denote residual directions induced by modifier m for targets A and B.

We report average alignment over held-out modifiers and random splits, and compare against a shuffled baseline where modifier labels are permuted.

3 Related Work

Attention-based attribution. Prior work has used attention weights, attention visualization, and attention flow as proxies for token–token influence. However, attention weights do not necessarily provide faithful causal explanations of model behavior [8], and they are inapplicable to attention-free architectures. Attention ablation and knockout studies have shown that Transformer models exhibit relatively localized interaction patterns, while state space models such as Mamba exhibit more distributed structure. Our findings are consistent with these observations, but are obtained without intervening on architectural components: we define interaction operationally through controlled input perturbations, yielding an architecture-agnostic characterization that extends naturally to models without explicit attention mechanisms.

Residual stream and circuit analysis. Transformer circuits research views the residual stream as a shared communication channel through which attention and MLP modules write additive updates [4, 5]. More recent work has developed circuit-level analyses and attribution graphs to trace how features contribute to model outputs [6, 3]. Our work is complementary: rather than reconstructing a full circuit graph, we isolate a target-token residual component and test whether it carries structured token–token interaction.

Causal tracing and model editing. Causal tracing and model-editing methods such as ROME identify internal states or weights responsible for factual associations, and then modify them to edit model behavior [9]. MEMIT extends this line of work by scaling model editing to many factual associations [10]. Our intervention is different in both goal and mechanism. We do not edit model parameters or target factual memories; instead, we perform inference-time interventions on residual directions derived from controlled token perturbations. This lets us study contextual interaction as a manipulable representation rather than a stored fact.

Activation steering and representation engineering. Activation engineering methods such as Activation Addition construct steering vectors from contrasts between prompt pairs and add these vectors during inference to control model behavior [11]. Our residual injection experiments are closely

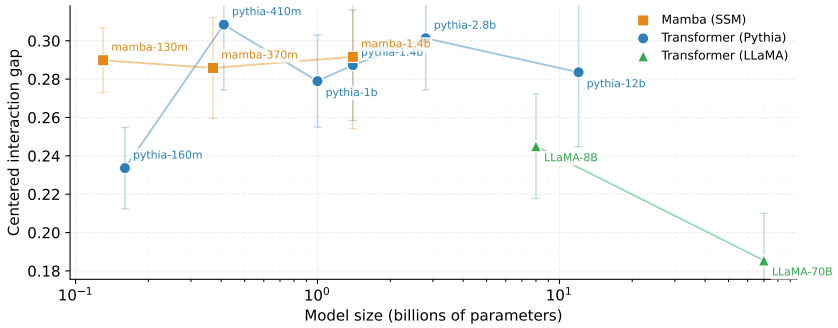

Figure 1: Interaction gap as a function of model scale for Transformer (Pythia, LLaMA) and Mamba models, spanning 160M to 70B parameters. Both architectures exhibit stable interactional structure across scales. Within each architecture family, there is no strong monotonic dependence on model size, suggesting that token–token interaction is a structural property of sequence models rather than a phenomenon that emerges only at large scale. Error bars indicate standard deviation across evaluated targets.

related in form, but differ in how intervention directions are obtained and interpreted. Rather than deriving steering vectors from behavioral prompt contrasts, we derive category-specific directions from residuals after subtracting tokenwise transformations under controlled perturbations. Thus, our directions are interpretable as token–token interaction directions rather than generic behavior vectors. Empirically, we find that Activation Addition directions—which retain token-local components—yield inconsistent or reversed selectivity in some categories, whereas our residual-based directions often produce more stable selectivity (Section 4.3).

Geometric analysis of layer updates. Prior work has decomposed layer updates into tokenwise and residual components, and shown that the residual is geometrically distinct from the tokenwise component — exhibiting lower alignment, higher angular deviation, and strong correlation with output perturbation magnitude [13]. That work establishes the residual as a functionally important non-local signal, but explicitly stops short of a stronger interpretive claim: the residual is positioned as a candidate space in which cross-token effects *may* reside, with the hypothesis left untested. Our work takes the next step. We introduce an operational definition of token–token interaction via controlled perturbations, show that the residual varies systematically with semantic context under these perturbations, and demonstrate that it supports reliable causal control. This transforms the residual from a geometric observation into a validated and manipulable representation of interaction — a distinction that requires the experimental framework we introduce here.

State space model interpretability. Mamba and related selective state space models provide an attention-free approach to sequence modeling [7]. Existing interpretability approaches for Mamba often seek attention-like explanations. Our framework instead probes interaction directly through residual geometry, enabling architecture-agnostic comparison across Transformers and state space models.

4 Results

4.1 Interaction structure across architectures and scales

We first examine whether interactional structure emerges in language models and whether it persists at large scale. Figure 1 shows that both Transformer and Mamba models exhibit stable and significant interaction gaps across model scales, from 160M to 70B parameters.

Across all settings, interaction structure is consistently positive and statistically significant under permutation testing (Appendix). This indicates that token–token interaction is not a byproduct of specific architectures or scales, but a general property of sequence models. Within each architecture family, there is no strong monotonic dependence on model size. At the same time, similarity in gap magnitude should not be interpreted as implying identical internal geometry across architectures.

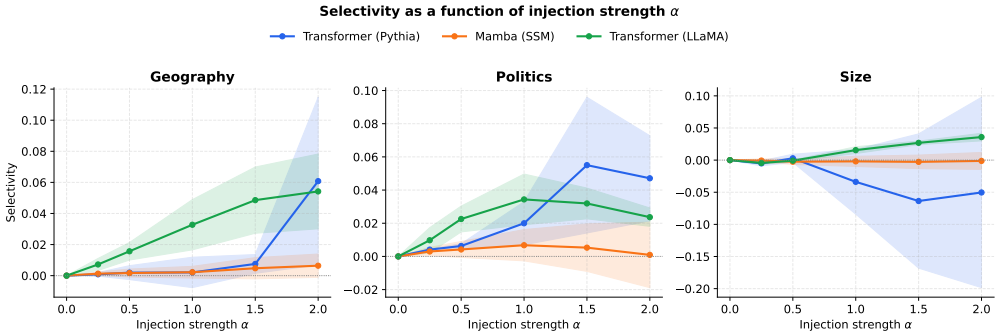

Figure 2: **Selectivity as a function of injection strength α.** Residual injection produces category-selective steering across architectures. Shaded regions indicate standard deviation across models.

As we show in later causal intervention experiments, architectures with comparably detectable interaction structure can differ substantially in how directly that structure supports steering through localized residual injection. This suggests that the main architectural variation lies less in whether interaction-related structure is present, and more in how that structure is geometrically realized and how accessible it is to intervention.

To verify that the observed interaction structure is not an artifact of a specific prompt template, we evaluate all models under three syntactic variants, including a negation template (`is not a`). We observe consistent interaction gaps across all templates, including the negation variant, indicating that the results reflect model representations rather than surface-level syntactic patterns.

Furthermore, interaction structure is consistent across all five evaluated target tokens (Appendix), and its dependence on layer depth is target-dependent: for some targets it increases at deeper layers, while for others it remains stable or decreases.

To test whether interaction structure extends beyond controlled templates, we perform a small robustness probe on naturally occurring text from the UN Debates corpus. Using sentences in which a semantic modifier directly precedes the target token, we observe positive and statistically significant centered interaction gaps in both Pythia-1B (0.096, $p < 0.001$) and Mamba-1.4B (0.099, $p < 0.001$). While smaller than the template-based gaps, these results suggest that category-conditioned interaction structure remains detectable beyond the controlled evaluation setting and across both Transformer and state space architectures.

4.2 Residual directions enable controllable semantic steering

We next test whether interactional residuals play a causal role in model behavior. We inject category-specific residual directions into the target representation and measure changes in continuation likelihoods across all evaluated models.

Figure 2 shows family-averaged selectivity curves across Geography, Politics, and Size categories. In Transformer models, selectivity increases positively with injection strength in Geography and Politics, indicating controllable semantic steering. Mamba exhibits weaker but structured trends, suggesting that its interaction-related structure may be less directly accessible to localized residual intervention. The Size category shows higher variance across models, reflecting the category-level variability we discuss further in Section 4.3 suggesting that its interaction-related structure may be less directly accessible to localized residual intervention while still causally influencing outputs.

4.3 Comparison with Activation Addition

To validate that the source of interaction directions matters, we compare our residual-based directions against Activation Addition [11], which derives steering vectors directly from raw hidden-state contrasts between positive and negative prompt pairs. We conduct this comparison across Pythia-1B, Pythia-12B, LLaMA-3.1-8B, and LLaMA-3.3-70B, evaluating Geography, Politics, and Size as semantic categories at layer fraction 0.9.

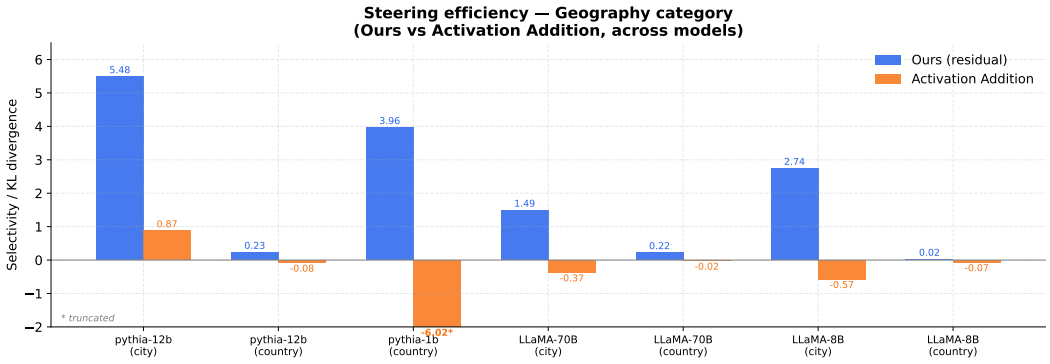

Figure 3: **Steering efficiency (selectivity / KL divergence) in the Geography category, across models and targets.** Residual-based directions (blue) consistently yield positive efficiency across all evaluated models. Activation Addition directions (orange) yield negative efficiency in all cases, indicating that token-local components in the raw contrast direction actively suppress category-aligned continuations. The Pythia-1B value for Activation Addition (-6.02) is truncated for readability.

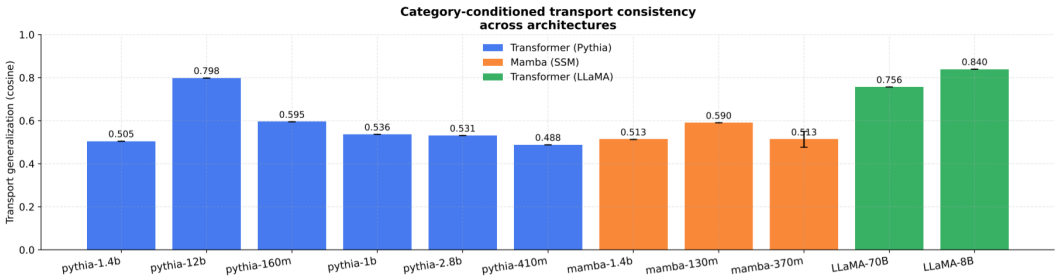

Figure 4: **Category-conditioned transport consistency across architectures.** We fit an orthogonal Procrustes map on anchor categories and evaluate cosine alignment on held-out categories. All models exhibit above-chance transport generalization, indicating transferable interaction geometry across targets.

The key methodological difference is that Activation Addition directions include token-local components, whereas our directions are obtained after subtracting the tokenwise map T, isolating the interactional residual.

Figure 3 shows steering efficiency (selectivity per unit KL divergence) in the Geography category across all models and targets. Across all evaluated models, our residual-based directions yield consistently positive efficiency, while Activation Addition directions yield negative efficiency in every case. This cross-model consistency — spanning Pythia-1B through LLaMA-3.3-70B — suggests that the failure of raw activation-contrast directions in Geography reflects a structural property of how token-local components interact with geographic modifiers, rather than a model-specific artifact.

For other semantic categories (Politics, Size), both methods produce positive selectivity, with residual-based directions showing comparable or higher efficiency on average, but neither method dominates uniformly across all categories. Our goal is not to claim uniform dominance over Activation Addition across every semantic category. Rather, the comparison isolates *when* raw hidden-state contrast directions fail because they retain token-local components: in such settings — most strikingly in the Geography category across all evaluated models — residualization improves steering reliability and prevents sign reversals. Category-wise failures remain possible when semantic categories are not well isolated; a more fine-grained analysis of category construction and probe calibration is left to future work.

4.4 Category-conditioned transport consistency

We further investigate whether interaction directions are structured across targets. Rather than assuming that residual directions are identical across targets, we test whether an orthogonal Procrustes

map learned from a subset of categories generalizes to held-out categories within the same semantic space.

Figure 4 shows strong transport generalization across all evaluated architectures. Models show above-chance transport generalization across all evaluated architectures, with alignment ranging from 0.49 to 0.84. LLaMA models exhibit the strongest generalization, while Pythia and Mamba models show moderate but consistent alignment above chance. This indicates that interaction directions are not isolated token-specific effects, but carry category-conditioned structure that transfers across targets to varying degrees depending on architecture.

4.5 Summary

Taken together, these results show that residual streams contain structured and causally effective signals associated with token–token interactions across model scales from 160M to 70B parameters. These signals are consistently detectable, statistically significant, and directly manipulable through residual interventions. Furthermore, comparison with Activation Addition shows that the tokenwise decomposition is especially important in settings where raw contrast directions retain token-local components: in such cases, residualization improves steering reliability and can prevent reversed semantic effects, most strikingly in the Geography category across all evaluated models.

5 Discussion

5.1 Residuals as structured interaction representations

Our results support a reinterpretation of the residual stream. Residuals are not merely leftover computation; under controlled perturbations, they contain structured cross-token effects that are geometrically organized, statistically significant, and causally effective. Residual directions cluster by semantic category and can steer model outputs in a monotonic, category-selective manner. This bridge between measurement and intervention is a key contribution of our framework: we first identify interaction-related structure via controlled perturbations, and then show that the same structure supports causal control. The comparison with Activation Addition further shows that this control specifically benefits from removing token-local effects, validating the tokenwise decomposition not only as a measurement tool but also as a practical preprocessing step for reliable semantic steering.

Our claim is operational rather than ontological: relative to a restricted tokenwise approximation class, the residual isolates the component of the update that systematically tracks controlled cross-token perturbations and supports causal intervention. We do not claim that this residual *is* token–token interaction in a deep or exhaustive sense, but rather that it provides an operationally isolatable representation of interactional effects under controlled perturbations.

5.2 Architectural differences in interaction geometry

Our comparison between Transformers and Mamba suggests that architectural differences lie less in whether interaction-related structure is present, and more in how it supports intervention. Figure 1 shows that centered interaction gaps are broadly positive across all evaluated models, indicating that semantically structured interaction is detectable in both attention-based and attention-free sequence models.

However, comparable detectability does not imply comparable steerability. In our residual injection experiments, Transformer models generally exhibit stronger and more monotonic category-selective steering, whereas Mamba shows weaker but still structured effects. This suggests that the key architectural difference is not the existence of token–token interaction, but whether it is realized in a form that is directly accessible through localized residual intervention.

More broadly, these results indicate that detectability and steerability are related but distinct properties of interaction geometry. Future work may clarify which architectural features determine whether interaction-related residual structure is sharply localized or more diffusely distributed.

5.3 On geometric aspects of category conditioned tests

The transport consistency results suggest that token–token interaction induces category-conditioned residual structure that is not purely target-specific. Simple alignment maps transfer this structure across targets, indicating locally consistent cross-target geometry. While this does not establish a manifold or parallel-transport formalism, it supports the view that interaction-related residual directions have transferable geometric organization.

5.4 Limitations and future directions

Several limitations should be noted. First, our experiments use simple prompt templates and a limited set of semantic categories, though we validate across model scales from 160M to 70B parameters. Second, our definition of interaction relies on controlled perturbations and may not capture long-range or compositional effects. Third, our analysis depends on a chosen class of tokenwise approximations. Fourth, category-wise steering failures remain possible when semantic categories are not well isolated; more rigorous analysis of category construction and probe calibration is left to future work.

Future work may extend this framework to more complex linguistic settings, explore connections to interpretable features, and investigate whether interaction directions transfer across tasks, domains, or languages.

6 Conclusion

We introduced an operational framework for measuring token–token interaction in language models based on controlled perturbations and residual decomposition.

We establish that: (i) token–token interaction effects are consistently measurable via residual-based metrics, (ii) these effects are statistically significant under permutation testing, (iii) they play a causal role in model behavior through residual injection, and (iv) they exhibit structured, category-conditioned organization that generalizes across targets.

Furthermore, we show that this interaction-related structure is present in both Transformer and state space models across scales from 160M to 70B parameters. Rather than differing primarily in whether such structure exists, the architectures differ in how directly it supports localized residual intervention: Transformer models generally exhibit stronger and more monotonic category-selective steering, whereas Mamba shows weaker but still structured effects. Comparison with Activation Addition shows that removing token-local components via the tokenwise decomposition is particularly important for obtaining reliable interaction directions in settings where raw contrast vectors are contaminated by token-local effects.

Taken together, these results support treating the residual stream, relative to a restricted tokenwise approximation class, as an operationally isolatable interface for analyzing and intervening on interactional effects. More broadly, our work provides a scalable and architecture-agnostic geometric framework for measuring, comparing, and causally probing contextual computation. We hope this framework provides a useful foundation for future work on analyzing and designing interaction mechanisms in neural sequence models.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *ACL*, 2020.

[2] Samira Abnar and Willem H. Zuidema. Quantifying attention flow in transformers. *CoRR*, abs/2005.00928, 2020.

[3] Anthropic. Circuit tracing: Revealing computational graphs in language models. `https://transformer-circuits.pub/2025/attribution-graphs/methods.html`, 2025.

[4] Nelson Elhage et al. A mathematical framework for transformer circuits. *Transformer Circuits (Distill)*, 2021.

[5] Olah et al. Zoom in: An introduction to circuits. 2020.

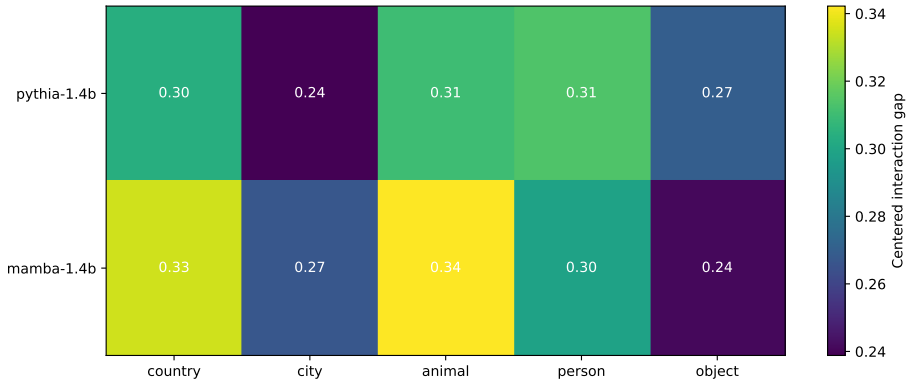

Figure 5: **Late-layer interaction gap across targets.** Centered interaction gap at deeper layers for different targets. Both Transformer and Mamba models exhibit strong interactional structure across multiple semantic domains.

[6] Wang et al. Interpretability in the wild: A circuit for indirect object identification. 2022.

[7] Albert et al. Gu. Mamba: Linear-time sequence modeling with selective state spaces. 2023.

[8] Sarthak Jain and Byron C Wallace. Attention is not explanation. *NAACL*, 2019.

[9] Kevin Meng et al. Locating and editing factual associations in gpt. *NeurIPS*, 2022.

[10] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022.

[11] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

[12] Jesse Vig. Analyzing the structure of attention in a transformer language model. *EMNLP*, 2019.

[13] Jun-Sik Yoo. On the geometric structure of layer updates in deep language models. *arXiv preprint arXiv:2604.02459*, 2026.

A Additional Interaction Analyses

A.1 Interaction structure across targets

To test whether interaction structure generalizes beyond a single target token, we evaluate the centered interaction gap across multiple semantic targets. Figure 5 shows that strong interactional structure persists across all evaluated targets at deeper layers in both Transformer and Mamba models. This consistency across semantic domains suggests that the observed interaction geometry is not specific to a particular token class, but reflects a broader property of model representations.

A.2 Permutation tests

We validate statistical significance using permutation tests. In both architectures, the observed interaction gap lies far outside the null distribution (Figure 6), confirming that the structure is highly unlikely to arise from random variation.

A.3 Robustness to prompt templates

In addition to the original template `{Subject} is a {Modifier} {Target}`, we evaluate the two additional variants used throughout the main experiments:

```
{Subject} seems like a {Modifier} {Target}
{Subject} is not a {Modifier} {Target}
```

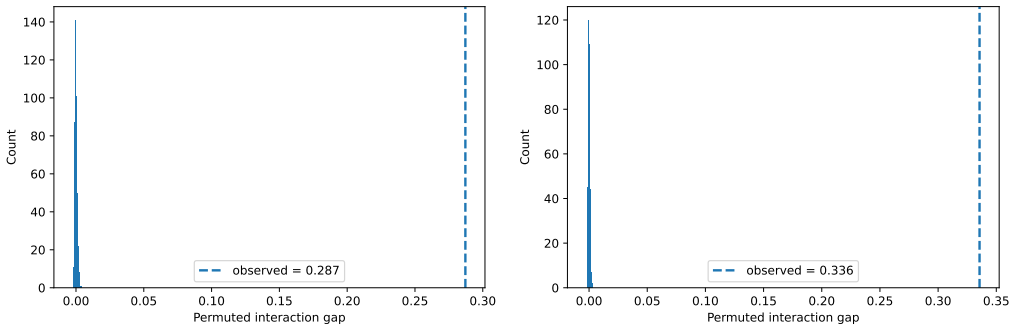

Figure 6: **Permutation test for interaction gap.** Observed interaction gap (vertical line) is significantly larger than the null distribution obtained by random label permutation. Pythia-1.4B (Left) and Mamba-1.4B (Right) models achieve $p \approx 5 \times 10^{-4}$, confirming that the interaction structure is not due to chance.

Across all three templates, including the negation variant, we observe consistent centered interaction gaps and category clustering. This indicates that the interactional structure is not tied to a particular syntactic form, but reflects a more general property of model representations.

B Attention Rollout Comparison

We compare our residual-based interaction metric with attention rollout [2], a commonly used proxy for token influence in Transformer models.

Setup. We compute rollout-based influence scores and evaluate category separation using the same experimental setup as in Section 2.6.

Results. Attention rollout yields a weak but statistically significant centered interaction gap on the current evaluation setup. In our experiments, the rollout gap is substantially smaller than the residual-based interaction gap, indicating that attention rollout captures only a limited fraction of the category-conditioned interaction structure recovered by our method.

Interpretation. This difference suggests that attention-based influence scores are sensitive to some broad interaction-related signal, but do not isolate it as cleanly as residual-based geometry after tokenwise decomposition and subject-level centering.

Limitations. Attention rollout is restricted to Transformer architectures and does not capture contributions from non-attention components such as MLP and residual updates. In contrast, our method operates directly on representations and applies uniformly across architectures.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe the proposed geometric decomposition framework, the empirical findings on residual alignment, and the scope of the experiments. Limitations and scaling considerations are also discussed in Section 7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations of the chosen function classes, the empirical scope of the investigation, the use of controlled template-based experiments, and the current limitations of the steering experiments.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results such as theorems or proofs; its contributions are primarily methodological and empirical.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper and appendix describe the experimental pipeline, model choices, datasets, fitting procedures, and evaluation metrics needed to reproduce the main results. We plan to release the code upon acceptance.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not provide anonymized code or a public repository at submission time. However, the paper and appendix describe the experimental pipeline, models, datasets, and evaluation protocol in sufficient detail to support reproducibility, and we plan to release the code upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the data construction procedure, model and layer choices, fitting setup, hyperparameters, and evaluation protocol used in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports standard-deviation error bars where applicable and uses permutation tests for the interaction-gap experiments. The definitions of the reported statistics are described in the corresponding experimental sections.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the compute resources used for the experiments, including the type of hardware and approximate computational cost needed to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work was conducted in accordance with the NeurIPS Code of Ethics. We use publicly available models and datasets, preserve anonymity at submission time, and discuss limitations and potential impacts of the work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both positive and negative societal impacts. On the positive side, the framework may improve model understanding, debugging, and safety analysis; on the negative side, better understanding of controllable internal directions may also inform misuse of model steering or manipulation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new high-risk models, scraped datasets, or other assets that would require special safeguards for responsible release.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses publicly available models and datasets, cites the corresponding sources, and reports the relevant licenses or terms of use where available.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve any crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Although the paper analyzes language models as research objects, LLMs are not used as important, original, or non-standard components in developing the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.