SCALE, DRIVERS, AND LIMITS OF REPRESENTATIONAL ALIGNMENT ACROSS LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Comparative analysis of large language models offers a means of identifying shared representational structure and systematic differences across them. However, aggregate measures of representational similarity, commonly used in these comparisons, can obscure how agreement varies across geometric scales and which linguistic properties support it. We introduce two complementary methods to resolve where models agree and what supports that agreement. Standard representational similarity analysis (RSA) summarizes similarity across all stimulus pairs, potentially obscuring differences in alignment between nearby and distant inputs. Neighborhood-Weighted RSA (NW-RSA) addresses this limitation by weighting stimulus pairs according to their proximity in representational space, enabling systematic comparison of alignment from local neighborhoods to global geometry on a common scale. Counterfactual RSA (CF-RSA) tests which linguistic properties support alignment by comparing intact representations in one model with representations under controlled linguistic perturbations in another. Using these tools across diverse language models, we find strong agreement in coarse representational geometry but substantial divergence in local neighborhoods. Alignment is observed both across broadly sampled natural text and within linguistic equivalence classes that preserve particular aspects of meaning or structure. Counterfactual analyses reveal selective sensitivity, with changes to logical scope reducing alignment substantially more than the tested changes to hierarchical syntax, event roles, or discourse structure. Alignment also decreases when sentence structure or lexical meaning is disrupted through word lists and jabberwocky, suggesting that the regularities of natural language support stronger convergence. These findings provide a differentiated account of representational convergence in language models and refine claims of representational universality by showing that strong overall similarity need not imply agreement in how models represent finer linguistic distinctions.

1 INTRODUCTION

A growing body of work suggests that as models scale, their internal representations converge toward a shared structure, as proposed by the *Platonic representation hypothesis* (Huh et al., 2024). Such convergence would mean that models trained on different data with different objectives recover a common geometry of meaning. Yet a single similarity score can be difficult to interpret because it reflects two choices. First, different measures emphasize different representational objects: Procrustes compares overall shape (Williams et al., 2021), while soft-matching methods compare unit response tuning (Khosla & Williams, 2024). Models can agree in their broad layout while disagreeing about local neighborhoods. Second, the stimulus set determines which distinctions are tested. Broadly sampled sentences mostly contrast topics and concepts, whereas controlled stimuli can isolate finer semantic or syntactic differences. Agreement on natural text may therefore fail under controlled changes to meaning or form, or off the natural-language manifold. Whether representations converge depends on what is measured and what it is measured over.

This paper distinguishes these forms of alignment through four linked questions. First, at what geometric scale do models align: the full pattern of pairwise distances or the neighborhoods around

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.

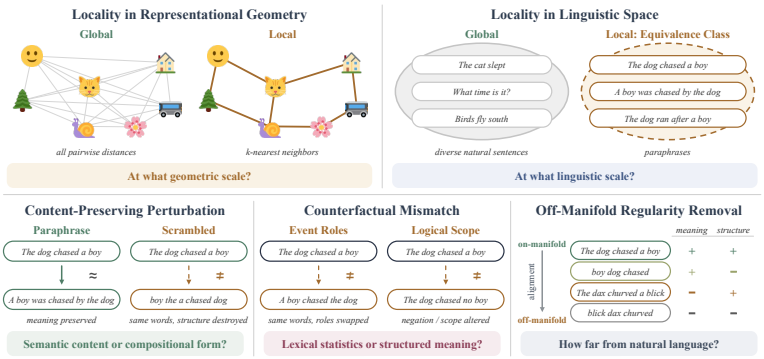


Figure 1: The three axes of alignment studied here: geometric locality (NW-RSA), linguistic locality (equivalence classes), and linguistic structure and regularity (perturbations, counterfactual mismatches, and off-manifold inputs).

each sentence? Second, does alignment also hold within linguistic equivalence classes, such as paraphrases, sentences sharing one syntactic form, entity substitutions, discourse neighborhoods, and task-equivalent prompts? Third, which controlled linguistic perturbations disrupt alignment, revealing sensitivity to particular aspects of interpretation? Fourth, how much alignment remains when lexical meaning and sentence structure are removed? These questions separate locality in representational geometry from locality defined by linguistic properties (Figure 1).

We address these questions through four contributions. **(i) A framework for comparative analysis** that examines alignment across geometric scales, within sets of sentences that share particular linguistic properties, and under controlled changes to meaning and form. It distinguishes how the comparison method and stimulus set shape conclusions about shared representations. **(ii) Neighborhood-Weighted RSA (NW-RSA)**, a general method for comparing representational alignment across geometric scales. By varying the weight assigned to nearby versus distant stimulus pairs, it places local and global geometric comparisons on a common scale. **(iii) Counterfactual RSA (CF-RSA)**, a general method for testing which stimulus properties support representational alignment. It compares one model's intact-input representational dissimilarity matrix with another's intact–perturbed dissimilarities, testing input changes relative to fixed intact references. **(iv) Evidence on the extent and conditions of convergence**: alignment is strong globally but weaker locally across diverse open-weight language models (Figure 3). Among the tested counterfactual phenomena, logical-scope changes reduce alignment the most (Figure 5), and removing lexical meaning or sentence structure also weakens it.

These findings qualify broad interpretations of the Platonic representation hypothesis and refine claims of shared representations (Klabunde et al., 2023; Lee et al., 2025).

2 Related Work

Representational comparison. Similarity measures compare artificial neural networks and relate their representations to biological responses (Kriegeskorte et al., 2008; Klabunde et al., 2025). One family compares stimulus-by-stimulus relationships, including representational similarity analysis (RSA), centered kernel alignment (CKA), normalized Bures similarity, and mutual k-nearest-neighbor overlap (Kornblith et al., 2019; Harvey et al., 2024; Huh et al., 2024). These measures emphasize different aspects of representational structure. Standard RSA summarizes agreement across all pairwise dissimilarities, whereas mutual k-NN compares local neighbor sets. Their scores cannot be directly compared to determine whether alignment is stronger locally or globally. Topological RSA addresses a related problem by transforming distances to emphasize intermediate-scale structure (Lin & Kriegeskorte, 2024). NW-RSA instead retains the original distance ranks and varies pair weights within one correlation measure, recovering global RSA when weights are uniform.

Representational convergence across models. Evidence of shared organization predates recent language-model studies. In vision, Kornblith et al. (2019) used CKA to compare networks across

random initializations, widths, and architectures. In language, Klabunde et al. (2023) evaluated similarity measures on commonsense reasoning and code-generation inputs, highlighting differences between models and challenges in interpreting scores. Huh et al. (2024) reported increasing alignment among more capable vision models and between independently trained vision and language models, motivating the Platonic representation hypothesis. After calibrating width and layer-selection effects, Gröger et al. (2026) found that scaling trends largely disappear for global similarity but persist for neighborhood membership. This illustrates how conclusions depend on the representational structure being compared. Naturalistic corpora introduce a further ambiguity because topic, lexical content, syntax, and meaning vary simultaneously. Our work complements these studies by combining comparisons across geometric scales with controlled linguistic inputs and perturbations.

Linguistic tests and counterfactual evaluation. Minimal pairs and contrast sets test whether individual models distinguish specific syntactic or semantic properties (Marvin & Linzen, 2018; Warstadt et al., 2020; Ribeiro et al., 2020). Probing studies examine whether those properties can be recovered from representations and how their accessibility varies across layers (Hewitt & Manning, 2019; Tenney et al., 2019). Representational comparison raises a related question: which linguistic properties account for agreement between models? High similarity on intact inputs cannot distinguish agreement supported by lexical overlap from agreement sensitive to compositional meaning. CF-RSA adapts controlled perturbations to this question by comparing intact geometry in one model with intact–perturbed dissimilarities in another. The resulting change tests which stimulus properties support the measured agreement.

3 Framework

3.1 Representational Comparison Methods

We propose Neighborhood-Weighted RSA for comparing alignment across geometric scales and Counterfactual RSA for testing which stimulus properties support agreement. Both methods apply generally to comparisons between neural representations; here, we use them to study language models. We additionally evaluate Procrustes and linear predictivity to test whether our conclusions hold under alternative definitions of alignment.

Representations and global RSA. Let $\mathcal{X} = \{x_i\}_{i=1}^N$ denote a shared stimulus set, and let $f_A(x_i) \in \mathbb{R}^{d_A}$ and $f_B(x_i) \in \mathbb{R}^{d_B}$ denote selected-layer representations. Each model yields a representational dissimilarity matrix (RDM), $D^A_{ij} = d(f_A(x_i), f_A(x_j))$, and likewise for B, where d is correlation distance (one minus the Pearson correlation) between sentence-level hidden states.

As a global reference, standard RSA (Kriegeskorte et al., 2008) measures agreement between the RDMs using the Spearman correlation of their upper-triangular, off-diagonal entries, $\mathrm{RSA}_{\mathrm{global}}(A, B) = \rho_{\mathrm{S}}(\mathrm{vec}_<(D^A), \mathrm{vec}_<(D^B))$, where $\mathrm{vec}_<$ stacks entries with $i < j$. This summarizes agreement across all stimulus pairs but does not distinguish local neighborhoods from broader geometric organization.

3.1.1 Neighborhood-Weighted RSA

We introduce Neighborhood-Weighted RSA (NW-RSA; Figure 2, left) to compare local and global geometric alignment within a single framework. NW-RSA weights stimulus pairs according to their proximity in representational space. Varying these weights changes the geometric scale of comparison while retaining the same correlation measure, rather than switching between global RSA and a separate neighborhood-overlap metric.

Let $\mathcal{U} = \{(i, j) : i < j\}$, and let u_{ij} and v_{ij} rank D^A_{ij} and D^B_{ij} over $\mathcal{U}$. We assign nonnegative weights W_{ij}, with $W = W^\top$ and $W_{ii} = 0$. Define $\bar{u}_W = (\sum_{\mathcal{U}} W_{ij} u_{ij})/(\sum_{\mathcal{U}} W_{ij})$ and $\tilde{u}_{ij} = u_{ij} - \bar{u}_W$, and analogously for v. Then

$$\text{NW-RSA}(A, B; W) = \frac{\sum_{\mathcal{U}} W_{ij} \tilde{u}_{ij} \tilde{v}_{ij}}{\sqrt{\left(\sum_{\mathcal{U}} W_{ij} \tilde{u}^2_{ij}\right)\left(\sum_{\mathcal{U}} W_{ij} \tilde{v}^2_{ij}\right)}}. \tag{1}$$

Ranks are computed once over all pairs and retained as the weights change. Thus, NW-RSA changes which pairs contribute without re-ranking the selected distances.

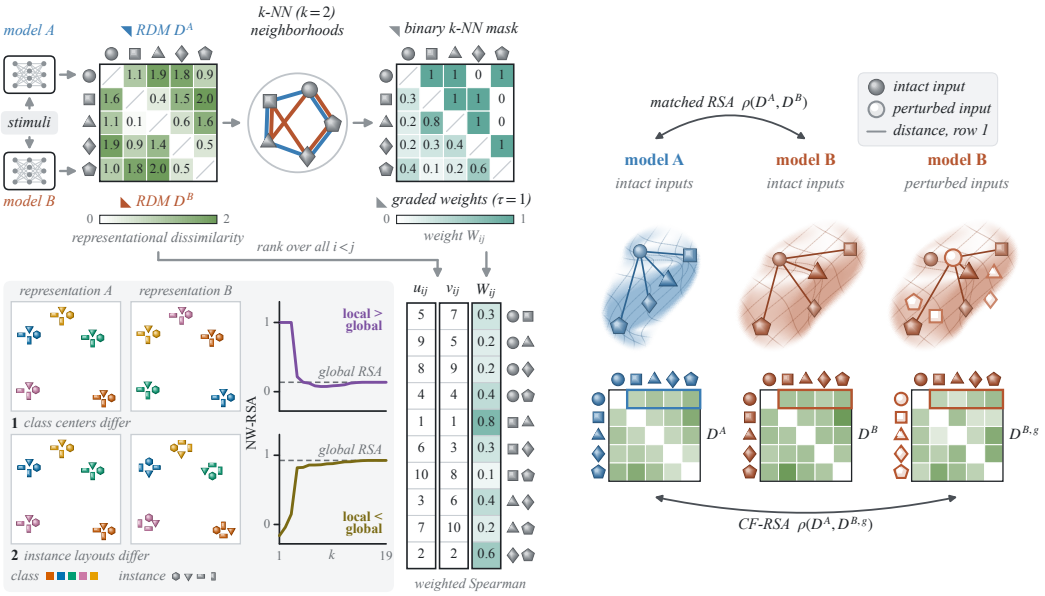


Figure 2: Two complementary RSA diagnostics (schematic; all values synthetic). Marker shapes identify the stimuli. **Left:** NW-RSA (Section 3.1.1). Each square holds two symmetric matrices: model A's RDM or the binary k-NN mask ($k = 2$) in the upper triangle, and model B's RDM or the graded weights ($\tau = 1$) in the lower triangle; neighborhood edges are colored by model. The shaded inset is a toy simulation with 20 points across 5 classes, where color marks the class rather than the model. In case 1, the two representations preserve the exact within-class geometry but rearrange the class centers, mimicking shared neighborhoods but a different coarse taxonomy: NW-RSA is 1.00 for $k \leq 3$, while global RSA is only 0.14. In case 2, the class centers are shared but the instances are permuted within each class, mimicking the reverse: local NW-RSA is -0.16 at $k = 1$, while global RSA is 0.93. **Right:** CF-RSA (Section 3.1.2). Color marks the model, and hollow markers are model B's perturbed inputs $g(x_i)$, placed in its intact geometry. The schematic RDM holds the distances from each $g(x_i)$ to the intact x_j (lines show row 1), excluding matched pairs; this is the transpose of $D^{B,g}$ and yields the same correlation. Correlating D^A with D^B gives global RSA, and correlating D^A with $D^{B,g}$ gives CF-RSA.

We use two weighting schemes. *Binary neighborhood weights* retain a pair if either stimulus is among the other's k nearest neighbors in either model:

$$W_{ij} = \mathbf{1}\left[j \in \mathcal{N}_k^A(i) \vee j \in \mathcal{N}_k^B(i) \vee i \in \mathcal{N}_k^A(j) \vee i \in \mathcal{N}_k^B(j)\right], \qquad i \neq j. \tag{2}$$

Each neighbor set excludes the stimulus itself, and increasing k expands the retained pairs. *Graded neighborhood weights* instead decrease continuously with distance, $W_{ij} = \sqrt{\exp(-D_{ij}^A/\tau)\exp(-D_{ij}^B/\tau)}$ for $i \neq j$. Smaller $\tau > 0$ emphasizes nearby pairs, whereas larger values distribute weight more evenly.

At $k = N - 1$, or as $\tau \to \infty$, all off-diagonal weights become one and NW-RSA reduces to global RSA. Alignment can increase, decrease, or vary nonmonotonically as neighborhoods expand, depending on agreement among newly emphasized pairs. Figure 2 illustrates both increasing and decreasing curves, so the empirical trend is not built into the definition.

Mapping-based alignment. RSA compares dissimilarity patterns without fitting a mapping between feature spaces. Procrustes (Williams et al., 2021) and linear predictivity (Yamins et al., 2014) instead fit such a mapping and differ in the transformations they allow. Let $\mathbf{X}$ and $\mathbf{Y}$ contain intact-input representations from A and B. We map B to A, so that B can subsequently receive perturbed inputs while A remains the reference:

$$\widehat{\mathbf{M}} = \underset{\mathbf{M} \in \mathcal{M}}{\arg\min} \|\mathbf{X}_{\text{fit}} - \mathbf{Y}_{\text{fit}}\mathbf{M}\|_{\text{F}}^2. \tag{3}$$

For equal-dimensional representations, Procrustes restricts $\mathcal{M}$ to orthogonal maps, comparing shape up to rotation and reflection. Linear predictivity permits unrestricted linear maps, including nonuniform scaling and shearing. Alignment is scored by the correlation between target representations and mapped predictions, $S_{\text{intact}} = \text{corr}(\mathbf{X}_{\text{eval}}, \mathbf{Y}_{\text{eval}}\widehat{\mathbf{M}})$.

3.1.2 COUNTERFACTUAL COMPARISON

Alignment on intact inputs does not by itself reveal which stimulus properties contribute to agreement. We introduce Counterfactual RSA (CF-RSA; Figure 2, right) to investigate this question through controlled input perturbations. In our linguistic experiments, the perturbations alter selected aspects of interpretation while preserving much of the surface lexical content.

Counterfactual RSA. Let g denote a perturbation. CF-RSA compares model A's intact RDM with a cross-condition dissimilarity matrix from model B, $D^{B,g}_{ij} = d(f_B(x_i), f_B(g(x_j)))$. Rows correspond to intact inputs and columns to their perturbed counterparts. The intact rows thus give a fixed reference for the perturbed representations.

Because $D^{B,g}$ is generally asymmetric, we compare all ordered off-diagonal entries:

$$\text{CF-RSA}(A, B; g) = \rho_{\text{S}}\big(\text{vec}_{\neq}(D^A), \text{vec}_{\neq}(D^{B,g})\big), \tag{4}$$

where $\text{vec}_{\neq}$ stacks entries with $i \neq j$. We exclude the diagonal because it would compare zero self-distances in A with intact and perturbed distances in B.

The decrease $\Delta_{\text{CF}} = \text{RSA}_{\text{global}}(A, B) - \text{CF-RSA}(A, B; g)$ quantifies perturbation sensitivity. When g is the identity, CF-RSA equals global RSA. A larger decrease indicates greater sensitivity to the tested change, helping identify which properties support agreement.

Counterfactual evaluation of fitted mappings. To test whether the conclusions extend beyond RSA, we fit each mapping on intact inputs and keep it fixed during counterfactual evaluation. Let $\mathbf{Y}^g_{\text{eval}}$ contain B's representations of perturbed evaluation inputs. We compare $S_{\text{perturbed}} = \text{corr}(\mathbf{X}_{\text{eval}}, \mathbf{Y}^g_{\text{eval}}\widehat{\mathbf{M}})$ with S_{intact}. Model A remains intact, and the mapping is not refitted. The change thus tests whether the correspondence learned on intact inputs survives the targeted change.

3.2 CONTROLLED LINGUISTIC STIMULUS SETS

Geometric neighborhoods group inputs by representational distance. To examine alignment within specific linguistic distinctions, we instead construct groups that hold selected linguistic properties fixed. These *linguistic equivalence classes* define neighborhoods independently of whether their members are nearby in a model's representational space.

We evaluate five classes. **Semantic-equivalence** preserves propositional meaning while varying lexical choice and syntactic form. **Syntactic-equivalence** preserves abstract syntactic form (clause type, constituent structure, grammatical functions, word order, valency) while varying content words and meaning. **Entity-equivalence** preserves relational or schematic meaning while varying entity identities. **Discourse-equivalence** holds discourse context fixed while varying sentences within that context. **Task-equivalence** preserves the underlying task or correct response while varying phrasing or framing.

For each class, we compute RSA among its members and compare it with alignment over broadly sampled text (Section 5.2). Holding the metric fixed while varying the stimulus set tests whether alignment extends beyond broad differences among heterogeneous inputs.

4 EXPERIMENTAL SETUP

Models and layers. The model pool contains 16 checkpoints spanning encoder-only models, GPT-2 variants, and recent Llama, Qwen, and Mistral models. This panel tests alignment across pretraining objectives, model families, and model sizes; Appendix A lists the checkpoints, and Appendix G provides their sources. Our main analyses use the penultimate layer and mean-pool hidden states over all non-special, non-padding tokens. Appendix E examines robustness across depth.

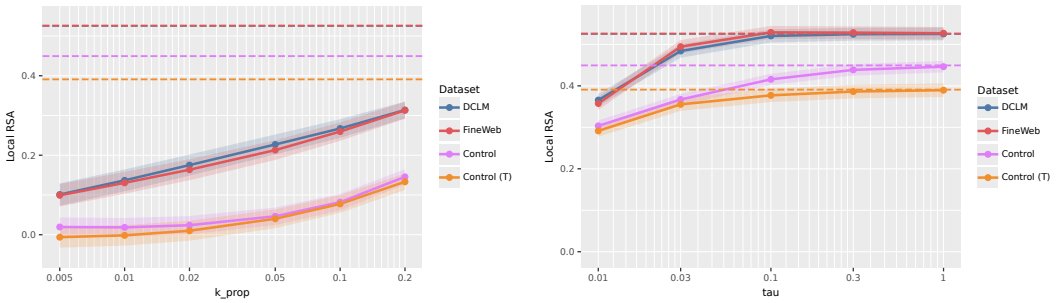


Figure 3: NW-RSA across geometric scales (mean over model pairs; shaded bands ± 1 SEM). **Left:** binary k-NN neighborhoods as a proportion k_{prop} of stimuli. **Right:** graded temperature τ. Solid lines show NW-RSA and dashed lines global RSA for DCLM, FineWeb, Control (Tuckute et al., 2024), and its baseline subset Control (T).

Stimuli. The experiment suite uses five stimulus families. The global sample contains diverse natural sentences. The equivalence-class sample contains local neighborhoods described in Section 3.2. The mismatched-perturbation sample contains controlled pairs that preserve many lexical items while changing hierarchical syntax, event roles, logical scope, or discourse structure. The perturbation sample contains generated sentences, each paired with a meaning-preserving paraphrase, a word-scrambled variant, and a word list of its content words in shuffled order. The off-manifold sample uses a 2×2 manipulation of lexical meaning and sentence structure. Appendix B gives the sources, the generation and audit procedure, and per-experiment item counts.

Statistics. We aggregate alignment over model pairs and plot the mean with ± 1 SEM bands over pairs (Figures 3, 5). Each experiment computes its own global RSA baseline on its stimulus set and model pool, so absolute RSA values are comparable within but not necessarily across experiments.

5 Results

5.1 Geometric Alignment Depends on the Scale of Comparison

We use NW-RSA to test whether models agree equally in local neighborhoods and global representational geometry. Figure 3 reports NW-RSA over model pairs across neighborhood sizes, on two naturalistic web corpora (DCLM (Li et al., 2024) and FineWeb (Penedo et al., 2024)), and on the Control sentence set (Tuckute et al., 2024) as a comparison (Control (T) is its baseline subset; Appendix F). With k_{prop} denoting neighborhood size as a fraction of the stimulus set, local alignment is far below global alignment in the smallest neighborhoods: NW-RSA is near 0.10 at $k_{\text{prop}} = 0.005$, against a dashed global ceiling of ≈ 0.53, and rises monotonically toward that ceiling as the neighborhood widens, reaching ≈ 0.31 at $k_{\text{prop}} = 0.2$. Graded weighting yields the same pattern, with NW-RSA approaching global RSA by $\tau \approx 0.1$. DCLM and FineWeb show similar curves, with higher alignment than both Control sets. The same pattern holds when the two models' graded weights are combined by their maximum instead of their geometric mean (Appendix D).

These results show that broad geometric agreement coexists with weaker agreement among nearby inputs, measured within the same correlation objective. This differs from asking how alignment changes as models scale. Gröger et al. (2026) found that calibrated cross-modal neighborhood alignment increased with model capability, while global scaling trends largely disappeared. Our analysis instead compares geometric scales within model pairs: stronger global than local alignment does not imply that global alignment increases more with model capability.

5.2 Alignment Persists Within Linguistic Equivalence Classes

Whereas NW-RSA defines locality by proximity in representational space, the second experiment defines it through shared linguistic properties. We compute global RSA separately within each

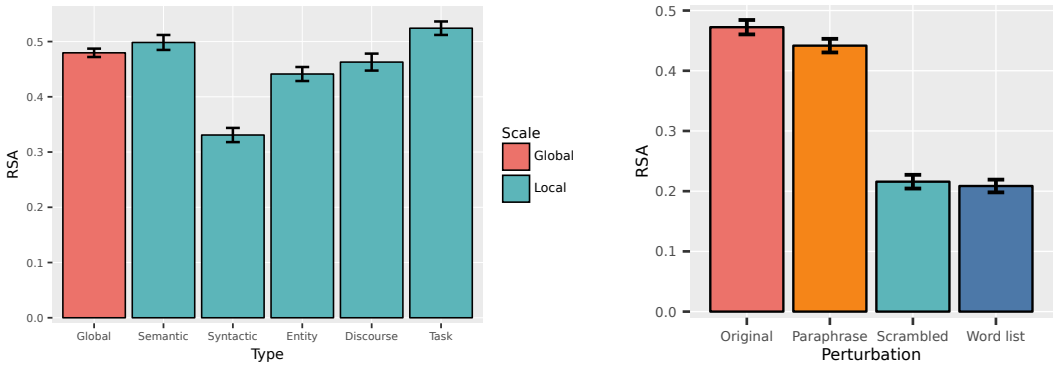


Figure 4: Mean RSA over model pairs (± 1 SEM). **Left:** global RSA and RSA within linguistic equivalence classes (semantic, syntactic, entity, discourse, task). **Right:** intact sentences compared with meaning-preserving paraphrases, word-scrambled inputs, and word lists.

Table 1: Examples of mismatched perturbations used for Counterfactual RSA. The paired inputs share many surface properties but differ in interpretation, isolating structure from surface form.

Axis	**Model A input**	**Model B input**
Hierarchical syntax (Shivagunde et al., 2023)	*The journalist investigated which athlete the team had recruited.*	*The journalist investigated which team the athlete had joined.*
Event roles (McCoy et al., 2019; Zhang et al., 2019)	*The president advised the doctor.*	*The doctor advised the president.*
Logical scope (Marelli et al., 2014)	*Nobody is riding the bicycle on one wheel.*	*A person is riding the bicycle on one wheel.*
Discourse structure (Chen et al., 2019; Mostafazadeh et al., 2016)	*The boy kicked the ball. The ball rolled away.*	*The ball rolled away. The boy kicked the ball.*

equivalence class and on broadly sampled sentences, holding the metric fixed while varying the linguistic scope of the inputs.

Alignment within most of these classes is comparable to alignment across the full sentence sample (Figure 4, left). Relative to an RSA of 0.48 for broadly sampled sentences, semantic and task classes show similar or slightly higher alignment (0.50 and 0.52), discourse and entity classes show similar or slightly lower alignment (0.46 and 0.44), and the syntactic class yields lower alignment (0.33). Thus, restricting comparisons to linguistically defined classes does not produce the pronounced reduction observed in local geometric neighborhoods. This suggests that alignment also reflects shared distinctions among inputs with similar linguistic properties. Procrustes and linear predictivity also place the syntactic class lowest and the discourse class well below the global sample (Appendix D). Across layers, the syntactic class is the lowest at every depth under all three metrics, while the differences among classes are largest in the first layer and narrow mainly in the early layers (Appendix E).

5.3 Testing Which Linguistic Properties Support Alignment with Counterfactual Analyses

The third experiment asks which linguistic distinctions support cross-model alignment. We apply Counterfactual RSA to perturbations targeting hierarchical syntax, event roles, logical scope, and discourse structure (Table 1). These perturbations alter interpretation while preserving much of the lexical content. A drop from intact RSA to CF-RSA indicates sensitivity to the tested perturbation.

We draw on established minimal-pair benchmarks. ROLE-1500 (Shivagunde et al., 2023) reorders embedded clauses to flip who acts on whom, HANS (McCoy et al., 2019) and PAWS (Zhang et al.,

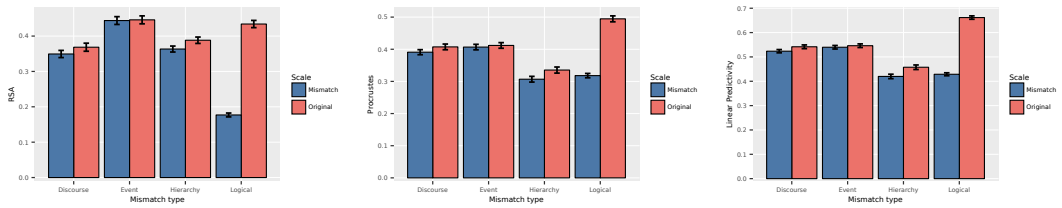


Figure 5: Counterfactual (mismatch) alignment over four phenomena, pooled over model pairs (± 1 SEM). Original (red) vs. meaning-altering mismatched (blue) pairing, under RSA, Procrustes, and linear predictivity (left to right). Logical scope collapses under mismatch in every metric. Discourse structure, event roles, and hierarchical syntax barely change.

2019) swap subjects and objects, SICK (Marelli et al., 2014) introduces contradictions, and Disco-Eval (Chen et al., 2019) and ROCStories (Mostafazadeh et al., 2016) reverse sentence order. We find that alignment is selectively sensitive to these changes (Figure 5). For discourse structure, event roles, and hierarchical syntax, mismatched and original pairings align almost equally well, so cross-model agreement on these inputs rests largely on shared lexical content. Logical scope is the exception: mismatching the negation and polarity structure collapses alignment (RSA $0.43 \rightarrow 0.18$), suggesting that here the shared geometry tracks the logical interpretation and not only the shared words. The same ordering holds for Procrustes and linear predictivity (Figure 5), with mappings fitted on intact inputs and evaluated on perturbed inputs without refitting. Across layers, logical scope shows the largest intact-versus-perturbed gap at every depth under all three metrics, but the size of this gap changes with depth. It is smallest in the first layer and grows through the middle of the network, whereas the gaps for the other three phenomena stay small throughout (Appendix E). This growth suggests that the sensitivity to logical scope is not limited to the lexical differences between the paired inputs.

5.4 Testing Alignment Under Paraphrase and Word Scrambling

We next ask whether alignment survives changes in wording that preserve meaning and changes in word order that disrupt sentence structure. Using generated anchor sentences, we compare alignment when both models receive intact sentences with alignment when one model receives paraphrased, word-scrambled, or word-list versions (Figure 4, right; 750 anchors). Paraphrasing produces a small reduction in RSA, from 0.47 to 0.44. Word scrambling produces a substantially larger reduction, to 0.22, despite preserving the original words, and word lists yield a similar value (0.21). Procrustes and linear predictivity show the same ordering (Appendix D). These results suggest that shared representations are relatively stable under meaning-preserving rewording but sensitive to disruptions of sentence structure and compositional meaning, so lexical content alone cannot fully explain cross-model alignment. The layer-wise results show that the effect of scrambling develops with depth. In the first layer, scrambled inputs align nearly as well as intact sentences and better than paraphrases under all three metrics. With depth, paraphrases move closer to intact sentences, while scrambled inputs fall further behind them through the middle of the network (Appendix E). Early representations thus appear to reflect which words are present more than their order.

5.5 Alignment under Reduced Linguistic Regularity

We next ask how alignment depends on the presence of lexical meaning and sentence structure. Unlike the preceding perturbation analyses, both models receive the same inputs in each condition. A 2×2 design compares natural sentences (Marelli et al., 2014; Mostafazadeh et al., 2016; Zhang et al., 2019; Cer et al., 2017; Shivagunde et al., 2023; Li et al., 2024; Penedo et al., 2024) (Table 4), word lists, jabberwocky sentences, and nonword lists (Table 5). Word lists retain meaningful words but remove sentence structure; jabberwocky sentences preserve sentence structure while replacing content words with pseudowords; nonword lists remove both. Three further conditions replace a single noun, verb, or adjective with a pseudoword (Appendix B).

Natural sentences produce the highest alignment (Figure 6; 750 sentences per condition). Word lists and jabberwocky sentences yield lower alignment (0.38 and 0.34), and nonword lists the lowest

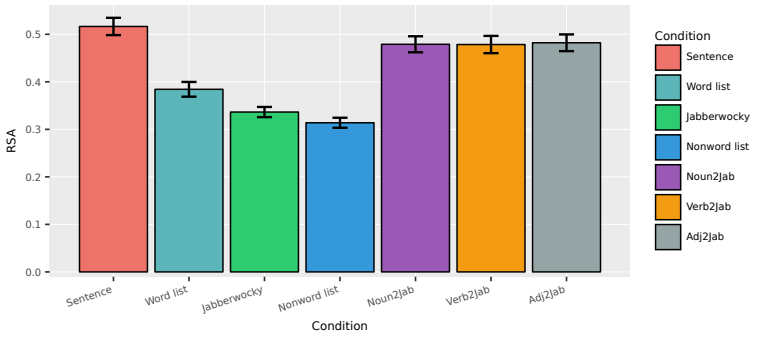


Figure 6: Off-manifold alignment at the penultimate layer (mean over model pairs, ± 1 SEM). Noun2Jab, Verb2Jab, and Adj2Jab replace one noun, verb, or adjective with a pseudoword. Natural sentences align best, yet substantial agreement persists for nonword lists.

(0.31). Replacing a single content word has a smaller effect, and RSA remains about 0.48 for noun, verb, and adjective substitutions. Procrustes and linear predictivity show the same pattern, and the human-authored stimuli of Regev et al. (2024) show a similar reduction (Appendices B and D). Across layers, natural sentences in our stimuli align best at every depth, and their advantage over word lists, jabberwocky, and nonword lists is larger in later layers than in the first (Appendix E). On the human-authored stimuli, sentences overtake nonword lists only after the earliest layers.

These results show that alignment depends on linguistic regularity, yet substantial alignment persists even for nonword lists, so shared geometry extends beyond meaningful, well-formed language.

6 Discussion

Our findings show that representational convergence across language models depends on the geometric scale of comparison and the linguistic properties of the inputs. Models share broad representational geometry but agree substantially less on relative dissimilarities within local neighborhoods. This qualifies broad interpretations of the Platonic representation hypothesis (Huh et al., 2024), since strong overall similarity need not extend to finer representational distinctions.

Alignment also persists within linguistic equivalence classes, where particular aspects of meaning or structure are held fixed, so shared geometry does not arise solely from broad differences among heterogeneous sentences. Because linguistic equivalence does not require geometric proximity, this result does not contradict the local–global NW-RSA gap; locality defined by representational distance differs from locality defined by linguistic properties.

The perturbation analyses show that alignment is selectively sensitive to linguistic interpretation. Logical-scope changes reduce it far more than the tested changes to syntax, event roles, or discourse structure, and it survives paraphrasing more readily than word scrambling, suggesting sensitivity to sentence structure and compositional meaning beyond lexical content. Alignment is also highest for natural sentences and decreases when lexical meaning or sentence structure is removed, yet substantial agreement persists for nonword lists, so shared geometry cannot be equated with shared linguistic interpretation.

Across layers, the scrambling effect is small in the first layer and develops through the middle layers, whereas logical-scope and regularity effects are present from the first layer and grow later, suggesting that these sensitivities are not confined to initial lexical representations.

Claims of convergence should therefore specify the geometric scale, stimulus distinctions, and input properties to which alignment is sensitive. NW-RSA and Counterfactual RSA provide tools for testing these conditions across neural systems, moving beyond whether models align to what they share and under which conditions.

AI USE STATEMENT

We used DeepSeek-V4-Flash to generate the equivalence-class and paraphrase stimuli, DeepSeek-V4-Pro to audit them (Appendix B), and Llama-Guard-4-12B for content-safety screening (Appendix C). AI assistants supported coding, literature searches, and editing for grammar, phrasing, and formatting. They did not contribute research ideas or experimental design. We reviewed all AI-assisted work: generated stimuli were audited and spot-checked as described in Appendix B. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work analyzes internal representations of existing language models using textual stimuli. It involves no human subjects and trains no new models. We use two kinds of stimuli and treat their risks separately. For **external public datasets** (STS-B (Cer et al., 2017), PAWS (Zhang et al., 2019), the counterfactual benchmarks, and the web corpora), we use them within their licenses (Appendix G) and additionally screen datasets with Llama Guard (Inan et al., 2023; Meta AI, 2025) and with rule-based filters to rule out offensive content and personally identifiable information (PII), so their reuse raises no new ethical concern beyond what their original releases already address. For the **stimuli we generate** (the equivalence-class neighborhoods and the paraphrase pairs), the generation and audit prompts explicitly forbid offensive, sexual, political, brand-, and PII-related content, with flagged items dropped or rewritten; the off-manifold conditions are derived by rule from human-written public sentences and contain no generated text. The full procedure is detailed in Appendix C. Two residual risks remain. First, automated filtering can miss cases, so a small amount of offensive content or PII may persist despite screening. Second, deliberately controlling the stimulus distribution, by narrowing it to specific equivalence classes, perturbations, and off-manifold conditions, can reduce how well our findings generalize to naturally occurring text. We therefore keep our claims tied to the evaluated models, layers, and datasets. Appendix H discusses further limitations.

REPRODUCIBILITY STATEMENT

Section 3 defines the alignment metrics, and Section 4 describes the experimental setup. Appendix A describes the model panel. Appendix B documents stimulus sources, generation and audit procedures, rejection statistics, and rule-based off-manifold transformations; Appendix C describes content-safety filtering. Appendix F provides representation-extraction procedures, metric parameters, aggregation rules, and the compute and software environment. Appendix G lists the external assets and their licenses. These sections provide the methodological details supporting reproduction of the reported analyses. Upon acceptance, we will publicly release the datasets constructed in this study, subject to the applicable licenses of their source datasets.

REFERENCES

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, 2017.

Mingda Chen, Zewei Chu, and Kevin Gimpel. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 649–662, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1060. URL https://aclanthology.org/D19-1060/.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1xMH1BtvB.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro, 2026.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic representation hypothesis: An Aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Sarah E Harvey, Brett W Larsen, and Alex H Williams. Duality of bures and shape distances with implications for comparing neural representations. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pp. 11–26. PMLR, 2024.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4129–4138, 2019. doi: 10.18653/v1/N19-1419.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python, 2020.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://arxiv.org/abs/2405.07987.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: LLM-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Emmanuel Keuleers and Marc Brysbaert. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3):627–633, 2010. doi: 10.3758/BRM.42.3.627.

Meenakshi Khosla and Alex H. Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pp. 326–341, 2024. URL https://arxiv.org/abs/2311.09466.

Max Klabunde, Mehdi Ben Amor, Michael Granitzer, and Florian Lemmerich. Towards measuring representational similarity of large language models. In *UniReps: the First Workshop on Unifying Representations in Neural Models, NeurIPS*, 2023. URL https://arxiv.org/abs/2312.02730.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025. doi: 10.1145/3717273.

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. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1eA7AEtvS.

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. In *Proceedings of the Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2503.21073.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, et al. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Baihan Lin and Nikolaus Kriegeskorte. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42):e2317881121, 2024. doi: 10.1073/pnas.2317881121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2317881121.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1314/.

Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, 2018. doi: 10.18653/v1/D18-1151.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Meta AI. Llama-Guard-4-12B. https://huggingface.co/meta-llama/Llama-Guard-4-12B, 2025. Model card.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1098. URL https://aclanthology.org/N16-1098/.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Tamar I Regev, Colton Casto, Eghbal A Hosseini, Markus Adamek, Anthony L Ritaccio, Jon T Willie, Peter Brunner, and Evelina Fedorenko. Neural populations in the language network differ in the size of their temporal receptive windows. *Nature Human Behaviour*, 8(10):1924–1942, 2024. doi: 10.1038/s41562-024-01944-2.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442.

Namrata Shivagunde, Vladislav Lialin, and Anna Rumshisky. Larger probes tell a different story: Extending psycholinguistic datasets via in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2094–2107, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.130. URL https://aclanthology.org/2023.emnlp-main.130/.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019. doi: 10.18653/v1/P19-1452.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024. doi: 10.1038/s41562-023-01783-7.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014. doi: 10.1073/pnas.1403112111.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1298–1308, 2019. doi: 10.18653/v1/N19-1131.

A Model Panel

The main comparisons span three model families. **Encoder-only masked language models** (BERT-base, RoBERTa-base, MiniLM-L12, ALBERT-base, BERT-large, RoBERTa-large, DeBERTaV3, and ELECTRA) test whether the alignment patterns hold for bidirectional encoders trained with masked-token objectives or related discriminative pretraining. **Decoder-only language models** (GPT-2 and GPT-2-medium) connect the analysis to left-to-right language modeling without modern instruction tuning or very large scale. **Recent open-weight LLMs** (Llama-3.2-1B, Llama-3.2-3B, Qwen-2.5-0.5B, Qwen-2.5-3B, Llama-2-7B, and Mistral-7B) test whether the same patterns persist for larger decoder-only models from recent open families. Exact Hugging Face identifiers, parameter counts, hidden sizes, selected layers, and token pooling rules are released with the final version. Unless noted otherwise, we read hidden states from each model's penultimate layer and mean-pool over all non-special, non-padding tokens.

B Stimulus Construction

Table 2 summarizes the source and scale of the stimuli behind each experiment. The counterfactual pairs are selected from existing datasets along the four axes of Table 1 (hierarchical syntax, event roles, logical scope, and discourse structure). The stimulus families that we construct ourselves are described below. The equivalence-class neighborhoods and the paraphrase pairs are generated and audited by separate LLMs, and the off-manifold conditions are derived by rule from human-written sentences; Figure 7 summarizes the first and the last. Content-safety screening is described in Appendix C.

Equivalence-class neighborhoods. To increase robustness and reduce potential stylistic patterns, we built the equivalence-class datasets using a multi-LLM pipeline, with diversity requirements in the prompts. For each of the five equivalence types evaluated in Section 5.2 (semantic, syntactic, entity, discourse, and task equivalence), there are 150 anchor sentences, each with exactly 50 neighboring sentences. We used DeepSeek-V4-Flash for the generator and DeepSeek-V4-Pro for the auditor (DeepSeek-AI, 2026). Every prompt states the property that the type holds fixed and the dimensions that may vary (Section 3.2), together with the content restrictions of Appendix C. Dataset construction proceeded in three stages (Figure 7a).

First, the generator produced candidate anchors, which were then audited by another LLM. Before the audit, code rejected anchors that were not in English or fell outside the word-count range of the type. The auditor checked grammaticality, construct validity, cross-class distinctness, and other requirements: each anchor had to be a single grammatical sentence expressing a concrete, unambiguous proposition that can support many equivalent neighbors, with no near-duplicates among the 150 anchors. Failed anchors were regenerated with the auditor's feedback and audited again, for up to five rounds. Second, the generator produced exactly 50 neighbors separately for each anchor, and code checked every neighbor set for word counts, duplicates, repeats of the anchor, and wording diversity. For the semantic, discourse, and task types, a set needed at least 12 distinct three-word openings and no opening shared by more than 8 neighbors; for the syntactic type, which holds grammatical form fixed, no content word of the anchor could reappear in a neighbor. A set that failed was regenerated with feedback, and if it still failed after four attempts, the anchor was resampled. Third, after all 150 classes had been assembled, the complete dataset was sent to the auditor, so the neighbors were audited at this stage rather than class by class. This dataset-level audit screened for violations of the intended equivalence relation, unnatural outputs, duplicate wording, repetitive templates, and insufficient cross-class distinctness. Classes that failed had their neighbors regenerated with the audit feedback, while accepted classes were preserved as context. If a class repeatedly failed, its anchor was resampled, and the class was reconstructed. An anchor was resampled at most three times, whether the trigger was a neighbor set that failed four times or a class that failed five audit rounds. The audit-repair cycle continued until the target classes passed. At the end, for each type, we randomly chose 10 examples and found that all of them satisfied the constraints of their type.

Rejection statistics. Table 3 counts the rejections recorded in the generation logs. Anchor columns count single sentences, and neighbor columns count whole sets of 50 neighbors. The

Table 2: Stimulus source and scale for each experiment. The generated equivalence-class and paraphrase datasets and the off-manifold conditions are released with this work. The four counterfactual datasets and the web corpora are public (Table 6).

Experiment	Source	Scale
Geometric scale (§5.1)	DCLM (Li et al., 2024), FineWeb (Penedo et al., 2024), Control (Tuckute et al., 2024)	10,000 sents each, Control 2000 (Control (T) 1000)
Linguistic scale (§5.2)	Generated equivalence classes	5 classes, 150 anchors × 50 neighbors each
Counterfactual (§5.3)	ROLE, HANS+PAWS, SICK, DiscoEval+ROCStories	750 pairs per phenomenon
Perturbation (§5.4)	Generated paraphrase pairs	750 anchors × {paraphrase, scrambled, word list}
Off-manifold (§5.5)	Derived from sentences	750 sents × 7 conditions

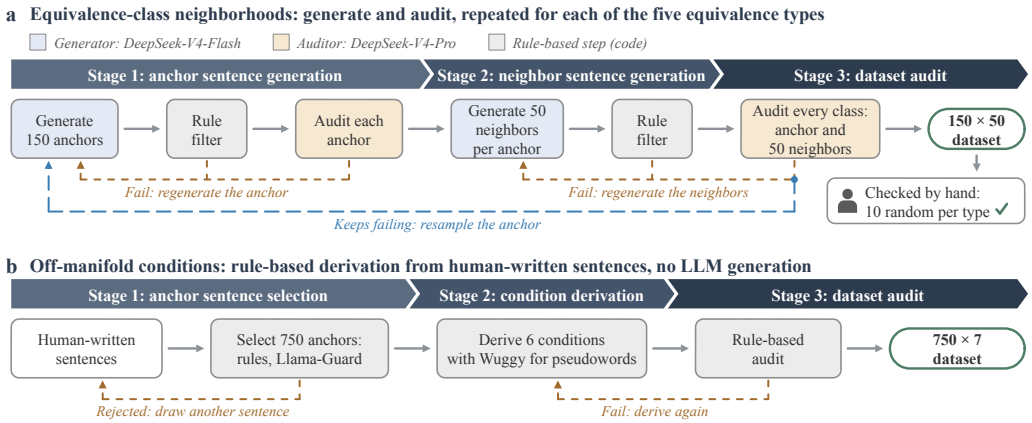


Figure 7: Construction of the stimuli we build ourselves. **(a)** Equivalence-class neighborhoods, built separately for each of the five equivalence types in three stages. (1) The generator (DeepSeek-V4-Flash) proposes 150 anchors, code checks their word count, and the auditor (DeepSeek-V4-Pro) checks each anchor; rejected anchors are regenerated, for up to five audit rounds. (2) The generator writes 50 neighbors for each anchor, and code checks the length, duplicates, and wording diversity of every set. (3) The auditor reviews the complete dataset, reading each anchor together with its 50 neighbors; the neighbors are audited at this stage. A class rejected by either check receives a new neighbor set (brown dashed paths); after four failed sets or five failed audit rounds, its anchor is resampled (blue path), at most three times. Finally, we randomly chose 10 examples per type and found that all of them satisfied the constraints. **(b)** Off-manifold conditions, built in three stages. (1) Candidate sentences from seven public datasets are normalized, filtered, and screened, and a rejected candidate is replaced from the pool. (2) Six conditions are derived from each of the 750 anchors by rule, with pseudowords from Wuggy and no LLM generation. (3) A rule-based audit re-derives every condition and checks it against its anchor; a condition that fails is derived again. The released dataset holds the 750 anchors in seven conditions. The source datasets are listed in Table 4, and the filter criteria and derivation rules are given in the text.

last column is the number of discarded sentences recorded in the logs, that is, the anchor rejections plus 50 times the rejected neighbor sets. Invalid calls are neighbor-generation calls that returned malformed or incomplete output. The logs do not record how many sentences these calls contained, so they are excluded from the last column, which is therefore a lower bound. The syntactic type required by far the most regeneration, and the discourse type the least.

Table 3: Rejections during equivalence-class construction. *Rule* and *LLM* count candidates rejected by the code checks and by the auditor, and *resampled* counts anchors replaced after their class kept failing. Anchor columns count sentences, and neighbor columns count sets of 50 neighbors. *Invalid calls* returned malformed or incomplete neighbor sets and are not counted in the last column, which gives the discarded sentences recorded in the logs and is a lower bound.

	Anchors			Neighbor sets		Invalid	Discarded
Type	rule	LLM	resampled	rule	LLM	calls	sentences
Task	0	74	1	36	32	10	3,475
Semantic	0	78	4	85	90	29	8,832
Syntactic	92	67	39	229	88	45	16,048
Entity	43	105	0	8	17	22	1,398
Discourse	0	27	0	0	0	3	27
Total	135	351	44	358	227	109	29,780

Paraphrase pairs. The perturbation analysis (Section 5.4) uses a separate dataset of 750 anchor sentences, each paired with one paraphrase that preserves its propositional meaning and truth conditions while varying syntax, lexical choice, style, and information packaging. The same generator and auditor as above produced the pairs. The generator wrote them in batches of 25, each batch assigned one of 20 everyday topic domains (e.g., household routines, travel, science, sports), and code checked that every sentence was English, 18–35 words long, and not a duplicate of any other sentence in the dataset after normalization. The auditor then checked all anchors together for grammaticality, concreteness, and distinctness, and checked each pair, in groups of 25, for changes to who did what to whom, quantity, negation, modality, time, location, cause, or outcome. Failed pairs were regenerated with the audit feedback and audited again, for up to three repair rounds. The two remaining conditions are derived from each anchor by rule: the word-scrambled variant keeps all of its word tokens in random order with punctuation removed, and the word list keeps only its content words and negation tokens in random order.

Off-manifold conditions. For the off-manifold stimuli, we built a dataset of 750 anchor sentences. To reduce stylistic patterns, we used human-written sentences from established datasets as the anchor sentences (Table 4; Figure 7b). Candidate anchors were normalized, filtered, and screened: a candidate had to be a single complete clause and could not duplicate another candidate, sentences from the web and STS-B sources that carry proper nouns were cross-checked with named-entity recognition, and sentences from all sources except ROLE-1500 were screened with Llama-Guard-4-12B (Inan et al., 2023; Meta AI, 2025). The generation of the other conditions does not involve LLMs. For the jabberwocky condition, eligible content words were replaced with pseudowords generated by Wuggy (Keuleers & Brysbaert, 2010), which matches the length and the character- and phoneme-level statistics of the word it replaces, while function words, punctuation, and word order were kept. When Wuggy found no candidate for a word under our strict settings, Wuggy was rerun with looser neighborhood and bigram tolerances, and an anchor whose words stayed unresolvable was dropped and replaced from the candidate pool. Every occurrence of a word maps to the same pseudoword throughout the release. The word-list condition shuffled the content words of the original sentences, while nonword lists came from shuffling pseudowords; both dropped function words. Finally, the noun-, verb-, and adjective-jabberwocky conditions replaced exactly one eligible noun (including proper nouns), lexical verb, or adjective, respectively, with its pseudoword from the jabberwocky sentence, while leaving the remainder of the sentence unchanged; the word was selected with spaCy part-of-speech tags (Appendix F). A rule-based audit then re-derived every condition from the anchors and checked that the conditions stay aligned item by item, that a word list and a nonword list contain exactly the content words or the pseudowords of their source sentence, that each graded condition replaces exactly one target of the intended part of speech, and that the pseudoword assignment is consistent; a condition that failed the audit was derived again. Table 5 summarizes which conditions keep meaning and structure.

Table 4: Human-written sources of the 750 off-manifold anchor sentences. Licenses and citations for each source are listed in Table 6.

Source	Subset	Sentences
SICK	contradiction subset	242
ROCStories	training set	213
PAWS	label-0 subset	132
STS-B	filtered	69
ROLE-1500	deduplicated	41
DCLM	web corpus	30
FineWeb	web corpus	23
Total		750

Table 5: Off-manifold stimulus families. The 2×2 base removes word-level meaning and sentence structure in separate steps. Three graded conditions replace one content word with a pseudoword while retaining structure and most meaning.

Condition	**Meaning**	**Structure**
Sentences	Yes	Yes
Word lists	Yes	No
Jabberwocky	No	Yes
Nonword lists	No	No
Noun→pseudoword	Partial	Yes
Verb→pseudoword	Partial	Yes
Adj→pseudoword	Partial	Yes

Check on human-authored stimuli. We also tested the human-authored SWJN dataset of Regev et al. (2024), whose word and nonword lists mix items across sentences. Under all three metrics (Figure 8), the differences among word lists, jabberwocky sentences, and nonword lists are smaller than on our stimuli (Figure 6). Layer-wise results are reported in Appendix E.

C Data and Content Safety

Sources. The stimuli come from three sources. The global sample, the counterfactual pairs, and the off-manifold anchor sentences draw on public datasets (Table 6). The equivalence-class neighborhoods and the paraphrase pairs are generated with an instruction-tuned LLM and audited by a second LLM against the constraints below. The off-manifold conditions are derived by rule from the human-written off-manifold anchor sentences (Appendix B).

Content filtering. To keep the released datasets free of offensive content and personal information, public-dataset text is screened at extraction time with Llama-Guard-4-12B (Inan et al., 2023; Meta AI, 2025) and samples flagged unsafe are dropped, with named-entity recognition (spaCy) flagging person mentions for review. The generation and audit prompts for the equivalence-class and paraphrase stimuli explicitly forbid offensive, sexual, political, brand-, and PII-related content, and a separate audit pass removes violations. The off-manifold conditions are derived by rule from the human-written anchor sentences and add no generated text.

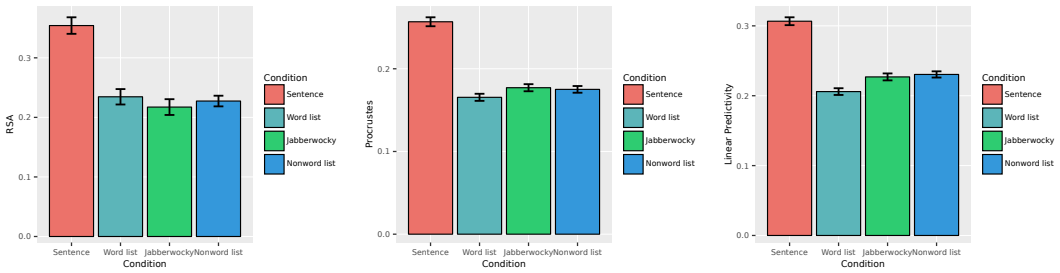


Figure 8: Off-manifold alignment on the human-authored stimuli of Regev et al. (2024) (mean over model pairs, ± 1 SEM), read at the penultimate layer. **Left:** RSA; **middle:** Procrustes; **right:** linear predictivity.

D ALTERNATIVE ALIGNMENT MEASURES

The main text reports RSA for most experiments. This appendix reports a variant of the graded NW-RSA weights and the Procrustes and linear-predictivity versions of the equivalence-class, perturbation, and off-manifold analyses, all read at the penultimate layer. The Procrustes and linear-predictivity versions of the counterfactual analysis appear in Figure 5, and those of the human-authored off-manifold stimuli in Figure 8.

Geometric scale. Figure 9 repeats the graded NW-RSA sweep of Figure 3 but combines the two models' weights by their maximum, $W_{ij} = \max\left(\exp(-D^A_{ij}/\tau), \exp(-D^B_{ij}/\tau)\right)$, instead of their geometric mean. The curves keep the same shape: on the web corpora, NW-RSA starts higher at the smallest τ but stays well below global RSA there, and it approaches global RSA at a similar τ.

Linguistic equivalence classes. Under Procrustes and linear predictivity (Figure 10), the semantic, entity, and task classes align at or above the global sample.

Paraphrase and word scrambling. Under Procrustes and linear predictivity (Figure 11), the drops for scrambled inputs and word lists are smaller than under RSA.

Reduced linguistic regularity. Figure 12 gives the Procrustes and linear-predictivity versions of Figure 6.

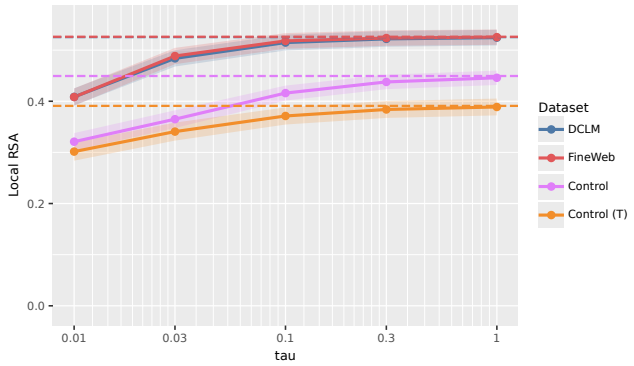


Figure 9: Graded NW-RSA with the two models' weights combined by their maximum instead of their geometric mean (mean over model pairs; shaded bands ±1 SEM). Dashed lines mark global RSA.

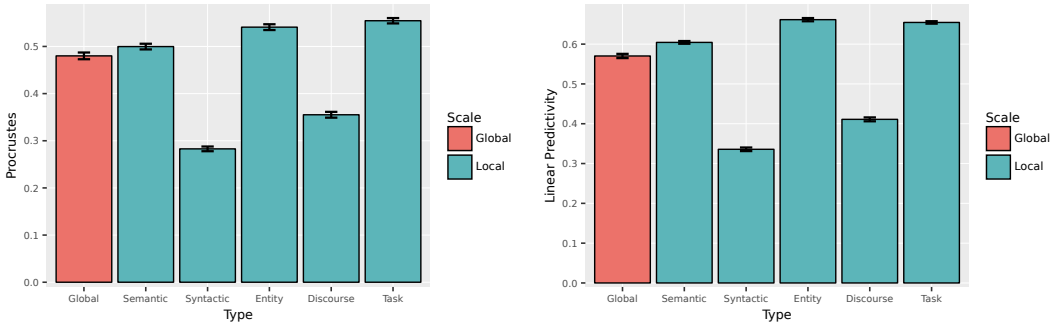


Figure 10: Linguistic-scale analysis under mapping-based metrics (mean over model pairs, ±1 SEM). **Left:** Procrustes; **right:** linear predictivity.

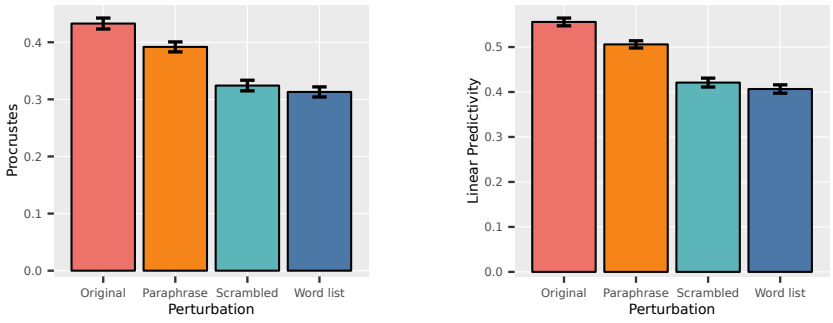


Figure 11: Paraphrase and word-scrambling analysis under mapping-based metrics (mean over model pairs, ±1 SEM). **Left:** Procrustes; **right:** linear predictivity.

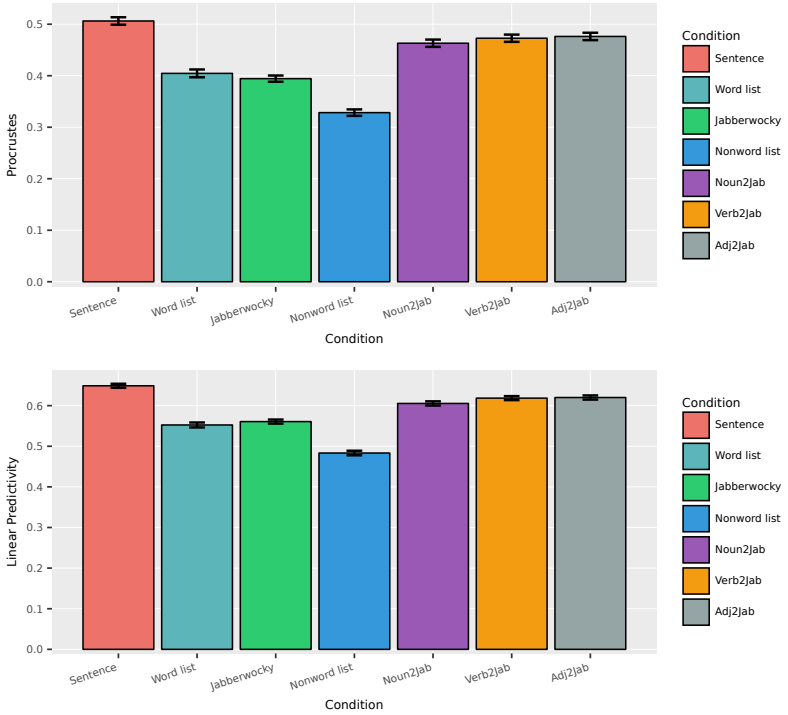


Figure 12: Off-manifold alignment under mapping-based metrics (mean over model pairs, ± 1 SEM). **Top:** Procrustes; **bottom:** linear predictivity.

E ALIGNMENT ACROSS LAYERS

The main analyses read representations at the penultimate layer. Here we ask whether the results change with depth, repeating the equivalence-class, counterfactual, perturbation, and off-manifold analyses across relative depth under RSA, Procrustes, and linear predictivity. Relative depth is counted in each model's effective functional units (e.g., transformer blocks). Relative depth 1 is the last such unit, which is model-dependent and differs by a normalization step from the penultimate hidden state read for the bar comparisons, so values at depth 1 need not match the bar plots. Lines show the mean over model pairs with ± 1 SEM bands.

Linguistic equivalence classes. The metrics differ in how the spread across classes narrows with depth (Figure 13). Under RSA, alignment declines overall with depth for every class and for the global sample, and the entity class drops from above the global sample to below it after the earliest layers. Under Procrustes and linear predictivity, the semantic, entity, and task classes start high and decline, while the syntactic class, and to a lesser extent the discourse class and the global sample, rise through the middle layers.

Counterfactual perturbations. After growing through the middle of the network, the logical-scope gap narrows again toward the output, and its growth is largest under RSA (Figure 14). Among the other three phenomena, the event-role gap is negligible throughout, and the discourse-structure gap is absent at the first layer and grows only slightly with depth.

Paraphrase and word scrambling. Paraphrases overtake scrambled inputs within the early layers (Figure 15). Beyond the earliest layers, word lists stay close to scrambled inputs; under RSA, they align far below the other conditions in the earliest layers, especially the first.

Reduced linguistic regularity. On our stimuli (Figure 16), single-word replacements stay close to natural sentences at every depth. The gap between natural sentences and the conditions that remove structure or meaning changes differently across metrics. Under RSA, it narrows over the earliest layers and then widens until near the output; under Procrustes and linear predictivity, it increases from the first layer to the middle of the network and then stays roughly constant. On the human-authored stimuli of Regev et al. (2024) (Figure 17), sentence alignment peaks in the middle of the network rather than in the earliest layers.

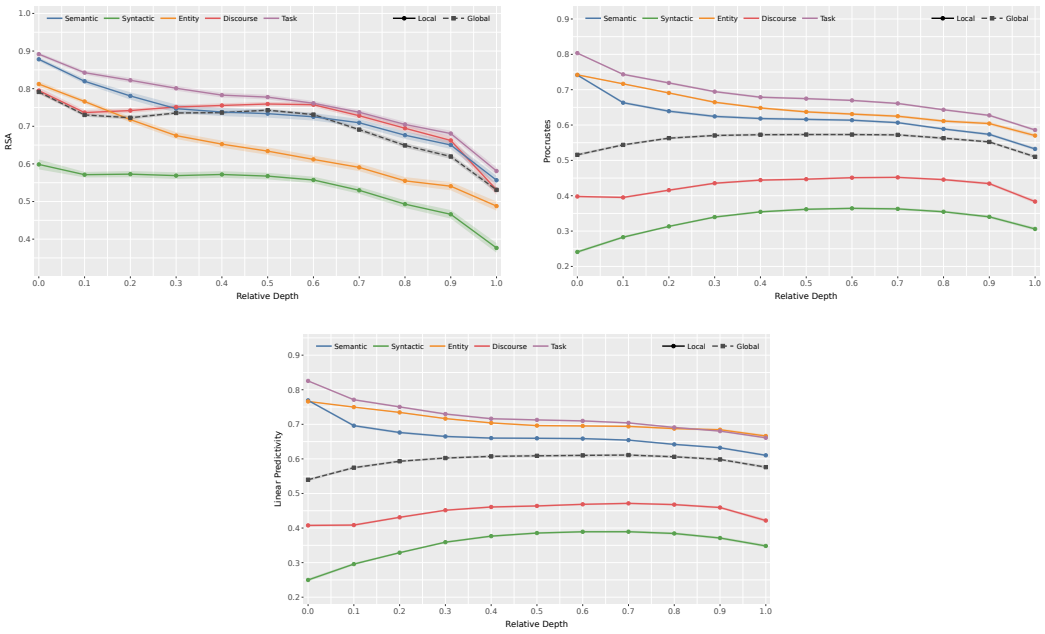


Figure 13: Layer-wise linguistic-scale analysis by relative network depth. **Top left:** RSA; **top right:** Procrustes; **bottom:** linear predictivity. Solid lines are the equivalence classes and the dashed line is the global sample. The syntactic class is the lowest at every depth, and the spread across classes narrows with depth.

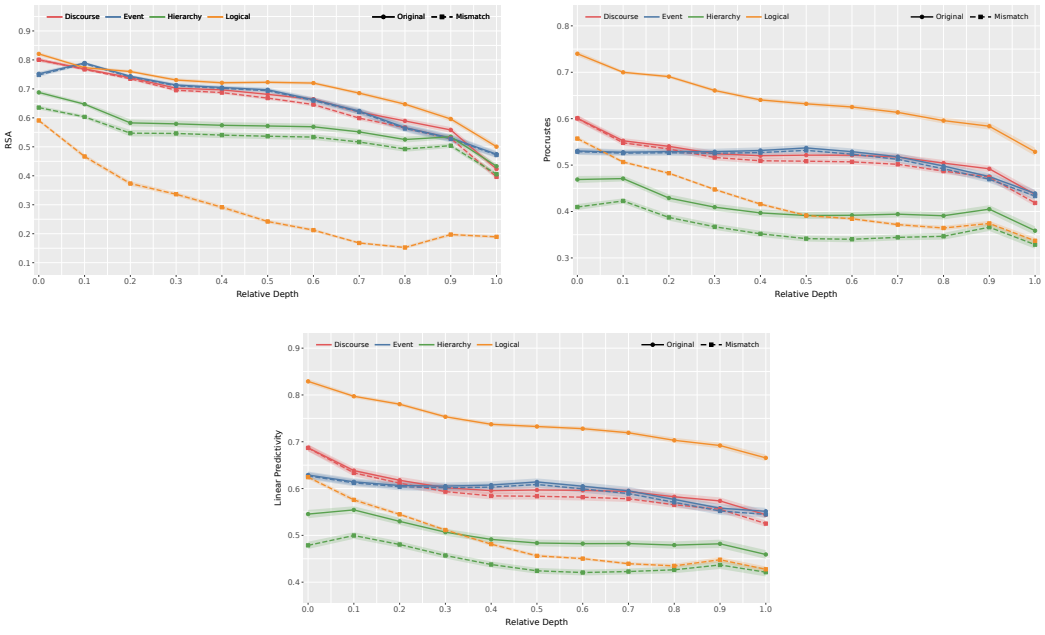


Figure 14: Layer-wise counterfactual contrast by relative network depth. **Top left:** RSA; **top right:** Procrustes; **bottom:** linear predictivity. Solid lines are the original pairing, dashed the mismatched pairing. The logical-scope gap is the largest at every depth and widens through the middle layers, while the discourse-structure, event-role, and hierarchical-syntax gaps remain small.

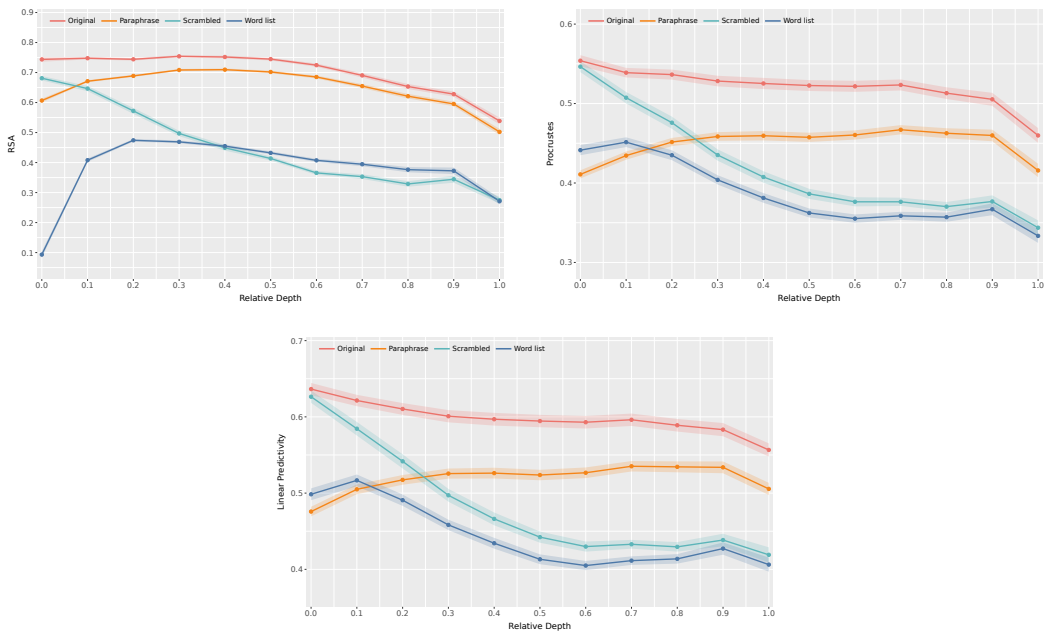


Figure 15: Layer-wise paraphrase and word-scrambling analysis by relative network depth. **Top left:** RSA; **top right:** Procrustes; **bottom:** linear predictivity. Scrambled inputs align nearly as well as intact sentences at the first layer and decline with depth, while paraphrases move closer to intact sentences.

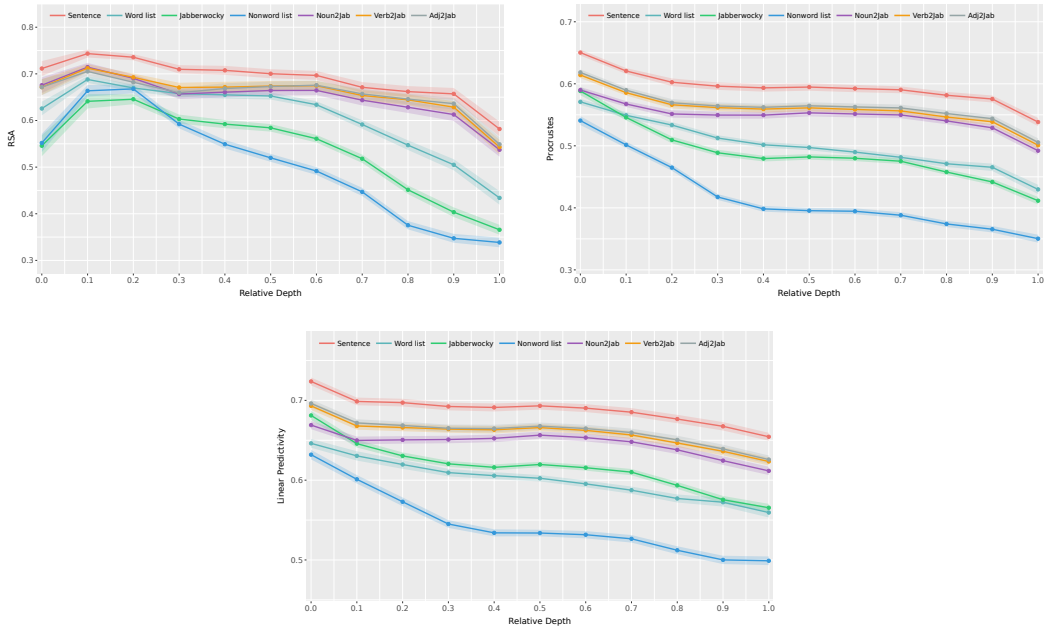


Figure 16: Layer-wise off-manifold alignment by relative network depth. **Top left:** RSA; **top right:** Procrustes; **bottom:** linear predictivity. Natural sentences align best at every depth, and the gap to word lists, jabberwocky, and nonword lists is larger in later layers than in the first layer.

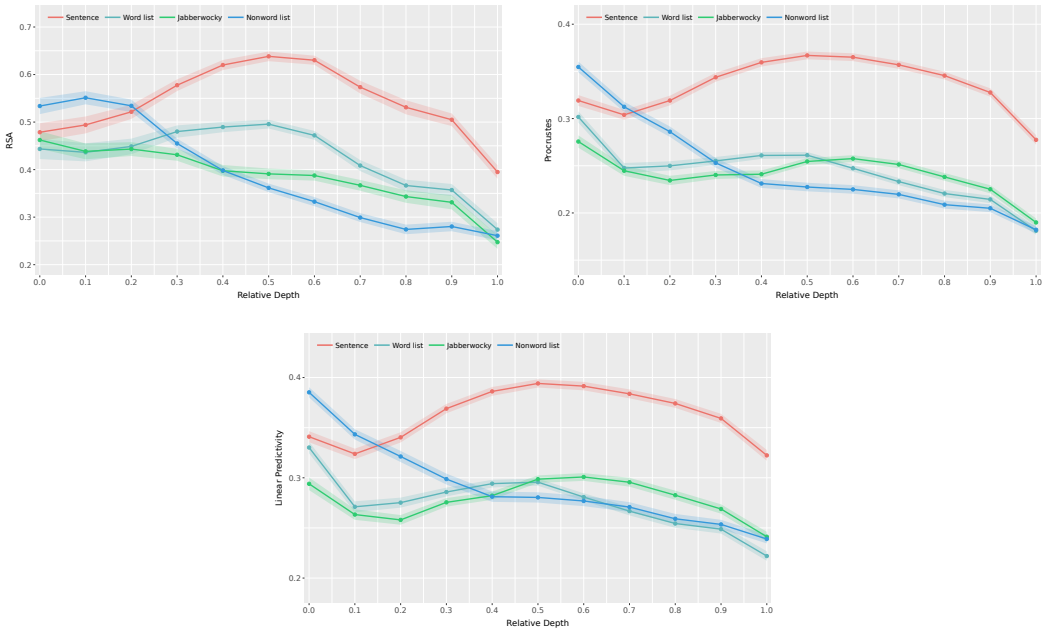


Figure 17: Layer-wise off-manifold alignment on the human-authored stimuli of Regev et al. (2024) by relative network depth. **Top left:** RSA; **top right:** Procrustes; **bottom:** linear predictivity. In the earliest layers, nonword lists align as well as or better than sentences; beyond them, sentences align best.

F Implementation Details

Representation extraction. For every model we run a single forward pass per stimulus and read hidden states from the penultimate layer (Section 4). Appendix E repeats the analyses at other depths. A sentence representation is the mean of hidden states over all non-special, non-padding tokens. All representations are mean-centered before any distance is computed, and we use correlation distance ($1 -$ Pearson) throughout.

NW-RSA and CF-RSA settings. RDMs use the upper-triangular pairs of the stimulus set. Neighborhood-Weighted RSA is evaluated across a range of localities: binary k-NN masks with neighborhood proportion $k_{\text{prop}} \in \{0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$ (equivalently k from 50 to 2000 on the web corpora), and graded weights with temperature $\tau \in \{0.01, 0.03, 0.1, 0.3, 1.0\}$. Both recover global RSA as the locality knob opens to the full set. The geometric-scale sweep (Figure 3) runs on two web corpora (DCLM, FineWeb, 10,000 language-filtered sentences each) and the full stimulus set released by Tuckute et al. (2024) (CONTROL). That study recorded fMRI responses of the human language network to 2,000 short sentences, nearly all six words long. The set has three conditions. The 1,000 baseline sentences are naturalistic sentences sampled to be diverse in topic and genre (e.g., news, web text, transcribed speech), and they were used to train a GPT-2-XL-based encoding model of the brain response. The 500 drive and 500 suppress sentences are those this encoding model predicts to evoke unusually high or low responses; in each condition, half were found by searching a large sentence corpus and half by iteratively editing sentences. Drive sentences are often ungrammatical or fragmentary (e.g., "ASY would Azerbaijan subs taken MIT."), whereas suppress sentences tend to be plain and frequent (e.g., "You get out of the car."). CONTROL (T) keeps only the 1,000 baseline sentences (the original release labels this set T). It checks whether the local–global gap holds without the atypical drive and suppress items. Counterfactual RSA correlates model A's intact RDM with model B's within-model cross-condition RDM (Section 3.1.2). Each mismatch phenomenon uses 750 matched item pairs.

Statistics. Alignment is aggregated over model pairs and reported as the mean with ± 1 SEM over pairs. Matched-vs-counterfactual and on-vs-off-manifold contrasts are read from whether these intervals separate.

Compute. No model training or fine-tuning is performed: the pipeline runs forward passes for hidden-state extraction followed by lightweight matrix correlations. Experiments were run on an internal server with 48 AMD EPYC 9224 CPU cores and 768 GB RAM, using two NVIDIA A40 GPUs (48 GB each) with four CPU data-loading workers. The 7B-parameter models are extracted in half precision, while encoder and sub-3B decoder models fit comfortably on a single GPU. RDM, NW-RSA, and CF-RSA computations run on CPU. Cached pretrained checkpoints, hidden states, and dissimilarity matrices used approximately 1.4 TB of storage.

Software. Models are loaded with PyTorch and the Hugging Face Transformers library, and numerical analysis uses NumPy and SciPy. Linguistic preprocessing uses spaCy (Honnibal et al., 2020): we run the transformer pipeline `en_core_web_trf` (falling back to `en_core_web_sm` when it is unavailable) with default settings, both for part-of-speech tagging, which selects the noun, verb, or adjective replaced in the single-POS off-manifold conditions, and for named-entity recognition flagging person mentions during content filtering (Appendix C). Licenses are listed in Appendix G.

G Existing Assets and Licenses

This work analyzes existing pretrained models with publicly available stimuli and standard open-source software. We use all assets for non-commercial academic research, credit the original creators, and do not redistribute third-party weights. Table 6 summarizes the main assets and their license or usage terms. For model families with per-size license differences (e.g., Qwen-2.5), the applicable terms follow each model card.

Consistency with intended use. We verified, for every external artifact whose intended use is specified, that our use is consistent with it. The pretrained checkpoints are public research mod-

Table 6: Main existing assets used in this work and their licenses or usage terms. Exact Hugging Face identifiers and versions accompany the released code.

Category	Item	License / Terms
Models	BERT, ALBERT, ELECTRA (Devlin et al., 2019; Lan et al., 2020; Clark et al., 2020)	Apache-2.0
	RoBERTa, MiniLM, DeBERTaV3 (Liu et al., 2019; Wang et al., 2020; He et al., 2023)	MIT
	GPT-2 (Radford et al., 2019)	MIT
	Mistral-7B (Jiang et al., 2023)	Apache-2.0
	Qwen-2.5 (Yang et al., 2024)	Apache-2.0 (3B: Qwen Research License)
	Llama-3.2 (Dubey et al., 2024)	Llama 3.2 Community License
	Llama-2 (Touvron et al., 2023)	Llama 2 Community License
Stimuli	STS-B (Cer et al., 2017)	Mixed-source: CC BY-SA 3.0/4.0, MS Research corpus terms, EMM/EU legal notices, DEFT project terms
	PAWS (Zhang et al., 2019)	Free use (Google), QQP subset
	Control sentences (Tuckute et al., 2024)	CC BY-NC-ND 4.0
	ROLE-1500 (Shivagunde et al., 2023)	Research use (no license file)
	HANS (McCoy et al., 2019)	MIT
	SICK (Marelli et al., 2014)	CC BY-NC-SA 3.0
	DiscoEval, ROCStories (Chen et al., 2019; Mostafazadeh et al., 2016)	MIT (code), ROCStories research use
	DCLM-baseline (Li et al., 2024)	CC BY 4.0 (research use only), Common Crawl ToU
	FineWeb (Penedo et al., 2024)	ODC-By 1.0, Common Crawl ToU
Software	PyTorch	BSD-3-Clause
	Transformers	Apache-2.0
	NumPy, SciPy	BSD-3-Clause
	spaCy (Honnibal et al., 2020)	MIT
	Wuggy (Keuleers & Brysbaert, 2010)	MIT
	Llama Guard-4 (Inan et al., 2023; Meta AI, 2025)	Llama 4 Community License

els. We use them only for non-commercial analysis of their hidden representations, with no fine-tuning, no weight modification, and no redistribution, which all of the licenses in Table 6 permit. The stimulus datasets are released for NLP research and benchmarking, and we use them solely for that purpose: STS-B, HANS, SICK, DiscoEval, and ROCStories for evaluation-style probing; PAWS, ROLE-1500, and the control set of Tuckute et al. (2024) as controlled stimulus sources; and the DCLM and FineWeb web corpora as naturalistic sentence pools. Several of these carry non-commercial or research-only terms (e.g., the control battery under CC BY-NC-ND 4.0, SICK under CC BY-NC-SA 3.0, ROCStories for research use). Our analysis-only, non-commercial use respects each of these.

H LIMITATIONS

Our model panel is limited to open-weight models for which hidden-state extraction is feasible, leaving generalization to closed and substantially larger models untested. The stimulus sets also have limitations. Web-derived datasets retain noise despite filtering, while generated stimuli may contain unintended variation or regularities introduced by the generating models. Separate generation and auditing models, together with manual checks, reduce but do not eliminate these concerns (Appendix B). Further, all stimuli are in English and cover a limited range of constructions and topics, so the results need replication with broader samples and other languages. The counterfactual results also concern specific perturbations, which may alter multiple input properties at once. They establish sensitivity to the tested changes but do not uniquely identify its cause.