Do Large Language Models Acquire Phrase-Based Processing? Evidence from Eye Movements and Model–Brain Alignment After Fine-Tuning

Xufeng Duan[1] **Zhengwu Ma**[2] **Zhaoqian Yao**[1] **Jixing Li**[2] **Zhenguang Cai**[1]
[1]The Chinese University of Hong Kong
[2]City University of Hong Kong
jixingli@cityu.edu.hk zhenguangcai@cuhk.edu.hk

Abstract

Autoregressive large language models (LLMs) process text token-by-token, yet the human language system operates over multi-word units. We ask whether aggregating LLM representations at the *phrase* level yields a closer correspondence to human reading behavior and language cortex than the default word-level representations, and whether phrase-segmentation fine-tuning amplifies this correspondence. Using Meta-Llama-3.1-8B (base and fine-tuned), we provide three converging lines of evidence. First, phrase-level attention features predict regressive eye-saccade patterns more closely than word-level features; a partial correlation analysis with a shuffled-boundary control indicates that this is not solely an aggregation artifact and that linguistic chunk boundaries explain unique variance beyond word-level attention. Second, fMRI encoding analyses show that fine-tuning selectively improves phrase encoding in left superior temporal gyrus and inferior frontal gyrus, with no improvement for word representations. Third, representational similarity analysis confirms a phrase-specific gain in model–brain geometric alignment. These results identify phrase-level representation as a critical granularity for LLM–human correspondence and suggest that targeted training can model human-like compositional processing, linking computational representations to hierarchical theories of language.

1 Introduction

Autoregressive large language models (LLMs) achieve strong performance by predicting the next token conditioned on preceding context (Radford et al., 2019). Because these models learn statistical regularities from text alone, they offer a powerful test bed for understanding what aspects of human language processing can emerge from distributional learning—and where the two systems diverge (Cai et al., 2024; Duan et al., 2024). However, a principled model–brain comparison requires choosing a representational granularity at which to extract and evaluate model features. Most prior work on model–brain alignment operates at the token or word level (Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux and King, 2022; Wu et al., 2025). Yet the human language system is organized around multi-word constituents: phrases are the primary units of planning, production, and comprehension (Garrett, 1975; Fedorenko et al., 2016; Hu et al., 2023; Gibson, 1998). This mismatch raises a fundamental question: at what representational scale do LLM internal states best correspond to human cognitive and neural processing?

The Granularity Problem Token-level representations are the default unit of analysis for transformer models, but they may not provide the most informative window into how models encode language-relevant structure. Effective next-token prediction often requires integrating information across multi-word spans (Manning et al., 2020). For instance, determiner selection (“a” vs. “an”) depends on upcoming phonological context. More strikingly, subject–verb agreement can survive substantial intervening material: speakers and comprehenders routinely maintain agreement across prepositional phrases, relative clauses, and other syntactic interpolations (e.g., “The key to the cabinets *is*...”), and errors in such configurations, so-called broken agreement, reveal that the language system tracks hierarchical phrase structure rather than simple linear adjacency (Bock and Miller, 1991). These observations suggest that multi-word phrases, not individual tokens, are the natural grain at which structural information is maintained. The question is therefore not whether models *explicitly compute* phrases, but whether *phrase-level aggregation of model representations* provides a better window into the structure that the human language system tracks.

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 464–476
July 3-4, 2026 ©2026 Association for Computational Linguistics

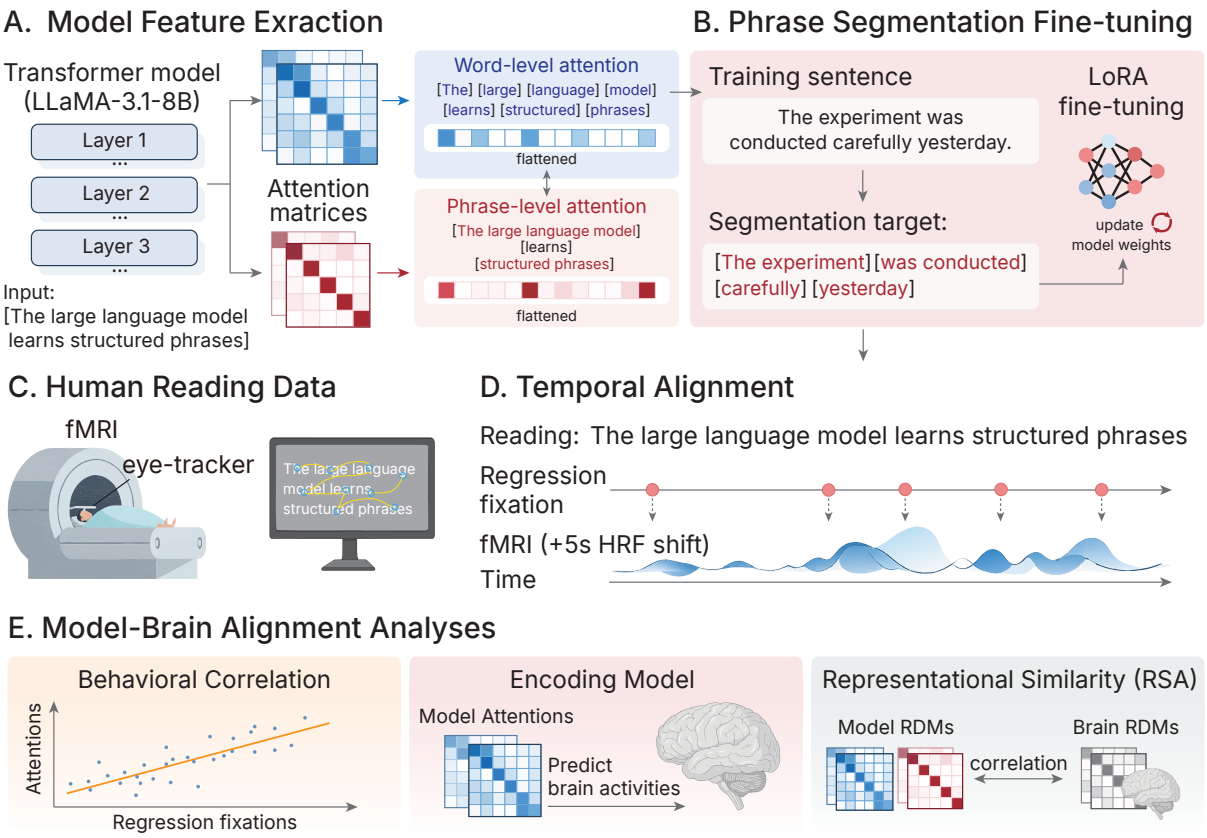

Figure 1: Methodological framework. **(A–B)** Attention matrices from Llama-3.1-8B are extracted at word-level and phrase-level granularities. **(C)** Eye-tracking fixation onsets are mapped to fMRI volumes (+5 s hemodynamic lag). **(D–E)** Encoding models and LOSO cross-validated cluster permutation tests identify brain regions with significant model alignment.

Phrase-Level Structure in Human Language Processing Psycholinguistic and neuroscientific evidence indicates that humans process language using structured, multi-word units (Fedorenko et al., 2016; Hu et al., 2023). Eye movements during reading reflect incremental difficulty and predictability effects that operate over spans larger than individual words: fixation durations and saccade patterns are modulated by phrase-level predictability and constituent boundaries, not only by properties of the currently fixated word (Rayner and Duffy, 1986; Smith and Levy, 2013). Neuroimaging studies consistently implicate left temporal and inferior frontal regions in combinatorial and sentence-level computations, with activity in these regions scaling with the size and complexity of phrasal constituents (Pallier et al., 2011; Fedorenko et al., 2011; Hu et al., 2023). If phrase-level aggregation captures information that is more closely aligned with the granularity at which the brain organizes language processing, it should yield better correspondence between model features and both behavioral and neural measures.

The Interface Hypothesis We propose that phrase-level aggregation of LLM attention provides a better correspondence to human cognition than word-level attention, regardless of whether the model explicitly computes phrases. On this view, the phrase is the representational scale at which model internal states most closely correspond to human reading dynamics and neural language processing. The claim is not that models "know about" phrases in a symbolic sense, but rather that extracting model features at the phrase grain reveals structure that is otherwise obscured by token-level analysis. We use *phrase* in an operational, surface-level sense: our segmentation corresponds to what NLP calls *chunking* (Abney, 1991), non-overlapping multi-word groupings such as base noun phrases, verb groups, and prepositional/adverbial spans, rather than full hierarchical constituents of generative syntax. This usage is close in spirit to "construction"-level units in construction grammar (Goldberg, 2003), is consistent with psycholinguistic descriptions of chunks as units of incremental processing, and is robust

across syntactic frameworks. We return to this choice in the Discussion and Limitations.

The Fine-Tuning Leverage If the phrase-level interface can be strengthened, then explicitly training a model to recognize phrase boundaries should *amplify* the alignment between model representations and human neural activity. Critically, we frame this as amplification of an existing correspondence rather than creation of a new computational mechanism: the base model may already encode phrase-relevant structure, and fine-tuning may sharpen this structure toward the granularity at which the brain operates (Manning et al., 2020; Tenney et al., 2019).

Present Study We test the interface hypothesis and the fine-tuning leverage using Meta-Llama-3.1-8B as a base model and a parameter-efficient fine-tuned variant trained to produce explicit phrase segmentations. We evaluate both models on the Reading Brain Project naturalistic reading dataset, which provides simultaneous eye tracking and fMRI during text comprehension. Our central comparison contrasts features derived from attention at multiple granularities: a *word* condition based on token/word-level attention, a *phrase* condition obtained by aggregating attention within and between syntactically defined phrase spans, a *shuffled-phrase* control that preserves multi-word aggregation but disrupts linguistic boundaries, and a *phrase* − *word* residual that isolates compositional structure beyond lexical processing. We ask: (1) does phrase-level aggregation better predict human reading behavior, and does this advantage survive after controlling for multi-word aggregation? (2) does phrase-segmentation fine-tuning selectively enhance neural encoding of phrase representations? and (3) does fine-tuning improve the representational geometry alignment between model attention and human neural responses in language-related region? Figure 1 provides an overview of our experimental framework.

Our results reveal a robust phrase superiority effect in behavioral alignment, selective amplification of phrase-level neural encoding after fine-tuning, and convergent geometric evidence, while neither model encodes phrase-minus-word residual structure, constraining phrases as holistic representational units.

2 Method

2.1 Base Model

We use Meta-Llama-3.1-8B as our base model (Grattafiori et al., 2024), analyzing internal attention representations across all 32 transformer layers. We focus on attention weights rather than hidden states because our central comparison is with word-pair regressive-saccade matrices, which have the same pairwise structure as attention matrices and were the basis of prior model–brain alignment work on this dataset (Gao et al., 2025).

2.2 Phrase Segmentation Task

To test whether explicitly teaching chunk boundaries affects cognitive alignment, we fine-tuned the model on a supervised phrase-segmentation task.

Training data and format. We constructed a dataset of 200,000 sentences from BLiMP (Warstadt et al., 2019) (∼60,000) and SNLI (Bowman et al., 2015) (∼140,000), filtered for well-formedness and length (10–80 tokens). Each sentence was annotated by a deterministic rule-based pipeline applied to spaCy dependency parses (*en_core_web_trf*); the rules group tokens into base noun phrases, verb groups, and prepositional/adverbial spans (full rule set in Appendix A). The model takes a sentence and outputs the segmented version with chunk boundaries marked by pipes (e.g., "The quick brown fox | jumps | over the lazy dog | ."). We adopt a shallow chunking scheme for two reasons: (i) it provides robust, framework-neutral groupings; (ii) it mirrors the "chunking" level used in psycholinguistic studies of incremental processing.

2.3 Fine-Tuning Implementation

We fine-tuned the model using LoRA ($r = 64$, $\alpha = 128$) applied to all attention projection matrices (Hu et al., 2022), trained for 2 epochs with the TRL SFTTrainer (full hyperparameters in Appendix A). The fine-tuned model achieves 94.2% exact-string match against this rule-based gold on a held-out test set of 140 testing sentences from Reading Brain Project datasets, vs. 25.7% for the base model.

2.4 Dataset: Reading Brain Project Corpus

We evaluated model–human alignment using the Reading Brain Project corpus (Li and Clariana, 2019; Follmer et al., 2018; Hsu et al., 2019): simultaneous eye-tracking and fMRI from 50 participants reading naturalistic STEM articles (∼29.6

sentences/article, ∼10.3 words/sentence) in the scanner at TR = 0.4 s. Analyses focused on a left-hemisphere language mask (see Neural Encoding below).

2.5 Attention Extraction Conditions

For each sentence, layer j, and head k, we extracted the lower-triangular part (excluding the diagonal) of the $n_{\text{word}} \times n_{\text{word}}$ attention matrix—the right-to-left attention weights, consistent with the model's unidirectional structure.

Word condition. Word-level attention as above; subword-to-word alignment summed over "to" subwords and averaged over "from" subwords (Manning et al., 2020). For eye-movement analyses, we max-pooled across heads at each layer before extracting the lower triangle.

Phrase condition. Tokens belonging to the same chunk were grouped using the same rule-based segmentation, and attention and saccade matrices were pooled to chunk boundaries with identical aggregation; eye-movement correlations were computed in chunk-pair space.

Shuffled-phrase control. Phrase boundary positions were randomly permuted within each sentence, preserving the number and size of phrases; the same pooling was applied. This matches the aggregation granularity but disrupts linguistic groupings.

Phrase − Word condition. Residual obtained by subtracting word-level from phrase-level attention values (neural encoding and RSA only).

2.6 Eye Movements Alignment

We follow the analysis pipeline of Gao et al. (2025): both unidirectional LLM attention and regressive (right-to-left) eye saccades occupy lower-triangular pair space, so the two can be compared on a sentence-by-sentence basis.

Regression target. For each sentence and each participant, we built a matrix $E^{(s)} \in \mathbb{R}^{n^{(s)}_{\text{word}} \times n^{(s)}_{\text{word}}}$ whose cell (l, m) equals the number of regressive eye fixations moving from word l to word m. We extracted the lower-triangle entries (right-to-left saccades), flattened them, and concatenated across sentences to obtain one vector per participant of length $N_{\text{pairs}} = \sum_s n^{(s)}_{\text{pairs}}$. The same flattening procedure was applied to the attention vectors. We restricted the analysis to regressive saccades because (i) unidirectional LLM attention cannot, on geometric grounds, align with forward saccades, and (ii) regressive saccades index integration and re-reading processes (Gao et al., 2025). We acknowledge this captures only a subset of reading behavior (see Limitations).

Attention–saccade alignment. For each layer j, participant i, and condition, we computed the Pearson correlation r between the participant's regressive-saccade vector $V^{(i)}_{\text{sac}}$ and the model's attention vector on the same pair coordinates. At each layer we max-pooled attention across heads, yielding one value per pair; we took lower-triangle entries (right-to-left pairs) from the word-level matrices in the word condition and from chunk-pooled matrices in the phrase condition, using the same segmentation for attention and saccades, and concatenated across sentences. Absolute r values are numerically small because each entry is a single regressive-saccade count for one pair in one sentence; condition effects are best read as *relative* differences on the same data. For visualization (Figure 2), we additionally report r divided by a leave-one-out noise ceiling: the mean Pearson r between each participant's saccade vector and the average of the remaining $N-1$ participants' vectors (cf. Gao et al., 2025).

Statistical model. Linear Mixed-Effects (LME) models were fitted on the raw (unnormalized) r values. We included Model (Base vs. Fine-tuned) and Condition (Word, Phrase) as fixed effects and both Subject and Transformer Layer as random intercepts:

$$r \sim \text{Model} \times \text{Condition} + (1|\text{Subject}) + (1|\text{Layer}) \quad (1)$$

Partial correlation analysis. To isolate phrase boundaries from word-level attention, we upsampled chunk-pooled attention back to word-pair coordinates (each word pair inherits its chunk-pair value), placing word, phrase, and shuffled conditions in a common word-pair space. For each participant and layer we computed partial Pearson correlations $r(V_{\text{sac}}, V_{\text{phrase}} \mid V_{\text{word}})$ and $r(V_{\text{sac}}, V_{\text{shuf}} \mid V_{\text{word}})$ using the standard partial-correlation formula; shuffled boundaries were averaged over 50 random permutations per sentence.

2.7 Neural Encoding

We assessed model–brain alignment using encoding models that predict cortical activity from LLM attention features. To avoid circularity, we used a leave-one-subject-out (LOSO) procedure: for each held-out subject, the optimal transformer layer was selected from the remaining $N-1$ subjects, and the held-out encoding map at that layer was entered into a group-level spatial cluster permutation test (10,000 permutations, cluster-forming $p < 0.01$) (Maris and Oostenveld, 2007). We conducted (1) single-model encoding and (2) between-model comparisons (fine-tuned minus base). Clusters were labeled using the Desikan–Killiany atlas (Desikan et al., 2006).

Region-of-interest selection. Analyses were restricted *a priori* to a left-hemisphere language mask: the union of Desikan–Killiany parcels in left STG, middle temporal, inferior temporal, supramarginal, angular gyrus, and inferior frontal gyrus (pars opercularis, pars triangularis, pars orbitalis). This mask was fixed before any model analyses and reflects the left-lateralized language network in prior work (Fedorenko et al., 2011; Pallier et al., 2011); no right-hemisphere or non-language parcels were tested.

Representational similarity analysis (RSA). For each layer and condition, we computed RDMs from the model's attention patterns and from multivoxel STG activation patterns at the sentence level. The Spearman correlation between model and brain RDMs is the raw RSA measure, normalized by a leave-one-out lower-bound noise ceiling (Nili et al., 2014): for each subject the ceiling is the Spearman correlation between that subject's RDM and the average RDM of the remaining $N-1$ subjects. We fitted LME models with Condition and Model as fixed effects and random intercepts for Subject and Layer:

$$\text{RSA} \sim \text{Condition} \times \text{Model} + (1|\text{Subject}) + (1|\text{Layer}) \quad (2)$$

Per-condition follow-ups tested the effect of Model within each condition; p-values used the Satterthwaite approximation.

3 Results

3.1 Behavioral Alignment: Eye Movements

We aligned model attention with human reading by computing Pearson correlations between word-pair (or chunk-pair) attention and regressive-saccade counts on the same sentences (§ Methods; following Gao et al., 2025). For each participant and layer we obtained a raw correlation r; LME models were fitted on these raw values (Table 1). Figure 2 displays the same correlations normalized by a leave-one-out inter-subject noise ceiling for interpretability.

Table 1: Fixed effects from the LME model ($r \sim$ Model $\times$ Condition $+$ (1|Subject) $+$ (1|Layer)) fitted on raw Pearson r values. Reference levels: Base model, Word condition. Significance: $^*p < .05$, $^{***}p < .001$.

Predictor	β	p
Fine-tuned (vs. Base)	+0.004	<.001***
Phrase (vs. Word)	+0.030	<.001***
Fine-tuned × Phrase	+0.002	.024*

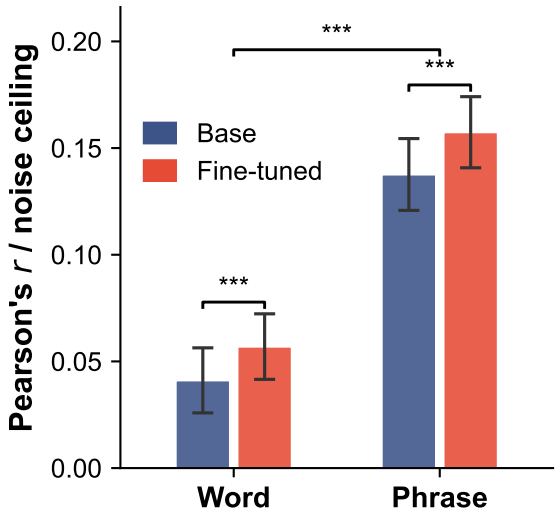

Figure 2: Eye-tracking–attention correlation by condition and model. Values are normalized by a leave-one-out noise ceiling computed from inter-subject regressive-saccade consistency. Bars show mean Pearson r / noise ceiling (averaged across layers); error bars: ± 1 SEM across subjects; brackets: per-condition fine-tuning contrasts.

Phrase superiority and fine-tuning. A robust main effect of Condition emerged ($F(1, 6316) = 6662.9$, $p < .001$): phrase-level features produced higher eye-tracking alignment than word-level features ($\beta = +0.030$, $p < .001$). Fine-tuning yielded a smaller but reliable improvement ($F(1, 6316) = 176.0$, $p < .001$; $\beta = +0.004$). The positive interaction ($\beta = +0.002$, $p = .024$) indicates a slightly larger fine-tuning benefit for the phrase condition. Raw r values are numerically small (word-level means ≈ 0.011–0.015; phrase-

Table 2: Partial correlation analysis controlling for word-level attention. Partial r: subject means across layers. Linguistic phrase boundaries yield positive partial correlations; shuffled boundaries yield negative values.

Model	Condition	Partial r	p
Base	Phrase	$+0.006$	.008**
	Shuffled	-0.008	<.001***
	Phrase − Shuffled	$+0.014$	<.001***
Fine-tuned	Phrase	$+0.008$	<.001***
	Shuffled	-0.008	<.001***
	Phrase − Shuffled	$+0.015$	<.001***

level ≈ 0.040–0.046); the phrase superiority effect is best read as a *relative* difference between conditions on the same data (cf. Gao et al., 2025).

Partial correlation: controlling for aggregation and lexical attention. The phrase superiority effect could partly reflect an aggregation artifact: pooling multiple word-level values into a single phrase-level value mechanically increases correlations regardless of linguistic structure. To isolate the contribution of *linguistic* phrase boundaries, we computed partial Pearson correlations $r(V_{\text{sac}}, V_{\text{phrase}} \mid V_{\text{word}})$ and $r(V_{\text{sac}}, V_{\text{shuf}} \mid V_{\text{word}})$ in a common word-pair space (§ Methods; Figure 3; Table 2). If the phrase advantage were solely an aggregation artifact, both phrase and shuffled boundaries should show comparable partial correlations once word-level attention is controlled.

Linguistic phrase boundaries showed a robust *positive* partial correlation ($r = +0.006$–$+0.008$, $p \leq .008$), while shuffled boundaries yielded a *negative* partial correlation ($r = -0.008$, $p < .001$). The Phrase−Shuffled gap was substantial ($\Delta = +0.014$–$+0.015$, $p < .001$) and replicated across both models. Fine-tuning further amplified this effect: the fine-tuning gain on phrase partial r ($\Delta = +0.002$, $p < .001$) exceeded the gain on shuffled partial r ($\Delta = +0.001$, $p < .001$), with a significant interaction ($p < .001$). As discussed in § Methods, partial correlations between correlated regressors systematically underestimate effect sizes, so the absolute values here should be interpreted as evidence that a real signal exists beyond word-level confounds, rather than as estimates of the total phrase contribution.

3.2 Neural Encoding: fMRI

We used LOSO cross-validated encoding analyses (§ Methods) to identify, within an a priori left-hemisphere language mask, brain regions where each model significantly predicts cortical activity, and to test where fine-tuning improves encoding strength.

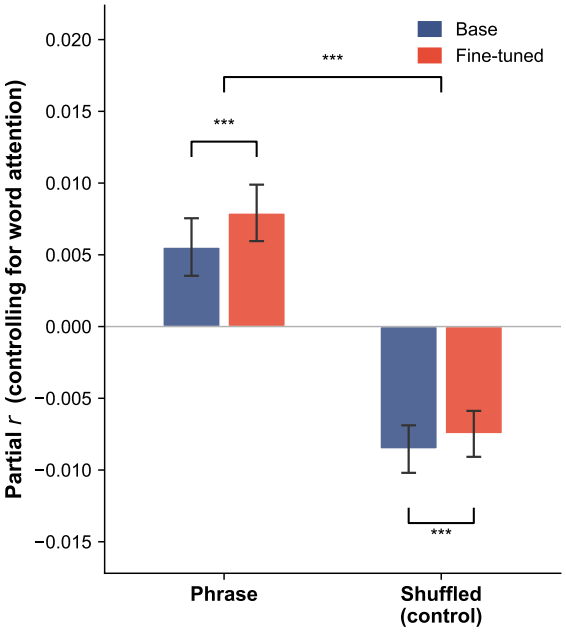

Figure 3: Partial correlation controlling for word-level attention. Each bar shows partial Pearson r between regressive-saccade counts and upsampled phrase-level (or shuffled) attention, with word-level attention partialed out. Error bars: ± 1 SEM across subjects; brackets: paired comparisons.

3.2.1 Single-Model Encoding

We first asked whether each model–condition combination produces encoding values significantly above zero, establishing that model attention predicts cortical activity.

Word and phrase conditions. Both models showed significant encoding of word and phrase representations in left STG (534 vertices, $p < .001$) and pars opercularis (101 vertices, $p < .01$) for all model $\times$ condition combinations (Figure 4, Panel A). LOSO consistently selected Layer 1 across all 50 subjects. The dominance of Layer 1 in single-model encoding most plausibly reflects shallow, input-driven features (e.g., positional or frequency-related regularities) that co-vary with STG activity regardless of fine-tuning; we return to this point in the Discussion.

Phrase − word residual. Neither model showed significant encoding for the phrase − word contrast, indicating that neither captures structure that is separable from word-level representations at the level

Table 3: Between-model encoding comparison (Fine-tuned > Base; cluster-forming $p < .01$, 10,000 permutations). n_{vtx} = cluster extent in vertices (effect size). Fine-tuning selectively enhanced encoding for phrase representations only. No significant clusters were observed for the word or phrase − word conditions.

Cond.	Region	n_{vtx}	p_{cluster}	LOSO Layer
Word	*No significant clusters*			
Phrase	STG	534	<.001***	L19 (48/50)
	Pars oper.	101	.009**	

Table 4: Per-condition LME effects of fine-tuning on STG representational alignment. β = unstandardized effect size (positive indicates greater alignment for the fine-tuned model). The phrase − word contrast was not significant ($\beta = +0.0003$, $p = .813$). The word effect ($p = .044$) does not survive Bonferroni correction for three comparisons ($\alpha_{\text{corrected}} = .017$).

Condition	β (Fine − Base)	p
Phrase	+0.0036	.004**
Word	+0.0033	.044*

our analysis can detect. This null constrains the interpretation of the phrase advantage to reflect representational *granularity* (the scale at which features are aggregated) rather than a distinct compositional mechanism.

3.2.2 Between-Model Comparison: Effect of Fine-Tuning

To identify regions where fine-tuning improved encoding, we compared the fine-tuned model against the base model within each condition (Table 3; Figure 4, Panel B).

Fine-tuning selectively enhanced encoding for *phrase* representations in left STG (534 vertices, $p < .001$) and pars opercularis (101 vertices, $p = .009$), with LOSO selecting Layer 19 for 48/50 subjects. No significant fine-tuning effect emerged for word representations or for the phrase − word residual. This combination of effects—condition-specific (phrase only), layer-specific (Layer 19), and anatomically localized—is what we would expect if fine-tuning sharpened intermediate-layer attention toward phrase-level structure that is also tracked by STG and pars opercularis. We do not interpret it as evidence for new compositional computation, given the phrase − word null (Discussion).

3.3 Representational Similarity Analysis

To complement the encoding analyses, we evaluated whether fine-tuning altered the geometric alignment between model attention and human STG activity using RSA (§ Methods). Results are summarized in Table 4 and Figure 5.

The LME model revealed significant main effects of Condition ($F(2) = 675.15$, $p < .001$) and Model ($F(1) = 7.59$, $p = .006$), with no significant interaction ($p = .218$). Per-condition follow-ups (Table 4) showed that fine-tuning significantly increased STG alignment for phrase representations ($\beta = +0.0036$, $p = .004$). A nominally significant word effect ($\beta = +0.0033$, $p = .044$) does not survive Bonferroni correction across the three conditions ($\alpha_{\text{corrected}} = .017$). No effect was observed for the phrase − word residual ($p = .813$).

The partial dissociation between RSA and encoding results is expected given their different measurement targets: encoding assesses vertex-wise predictive power, whereas RSA evaluates global geometric similarity. The most conservative reading is that fine-tuning primarily strengthens phrase-level correspondence, with possible but unconfirmed extension to word-level geometry (Figure 5).

4 Discussion

We asked whether phrase-level aggregation of LLM attention provides a closer correspondence to human reading behavior and language cortex than word-level attention, and whether phrase-segmentation fine-tuning amplifies that correspondence. Three converging lines of evidence support this view, with the caveat that "phrase" here refers to shallow chunks rather than full hierarchical constituents.

Phrase boundaries carry signal beyond aggregation and lexical attention. The partial correlation analysis shows that linguistic chunk boundaries explain unique variance in regressive eye saccades beyond word-level lexical attention and multi-word aggregation. Shuffled boundaries, which carry the same aggregation structure but disrupt linguistic groupings, yield negative partial correlations once word-level attention is controlled—directly addressing the concern that the phrase superiority effect is "merely aggregation smoothing." Absolute partial r values are small; partial correlations between correlated regressors systematically underestimate effects, so the relevant quantity is the small but reliable gap between Phrase and Shuffled. This pattern is consistent with recent independent

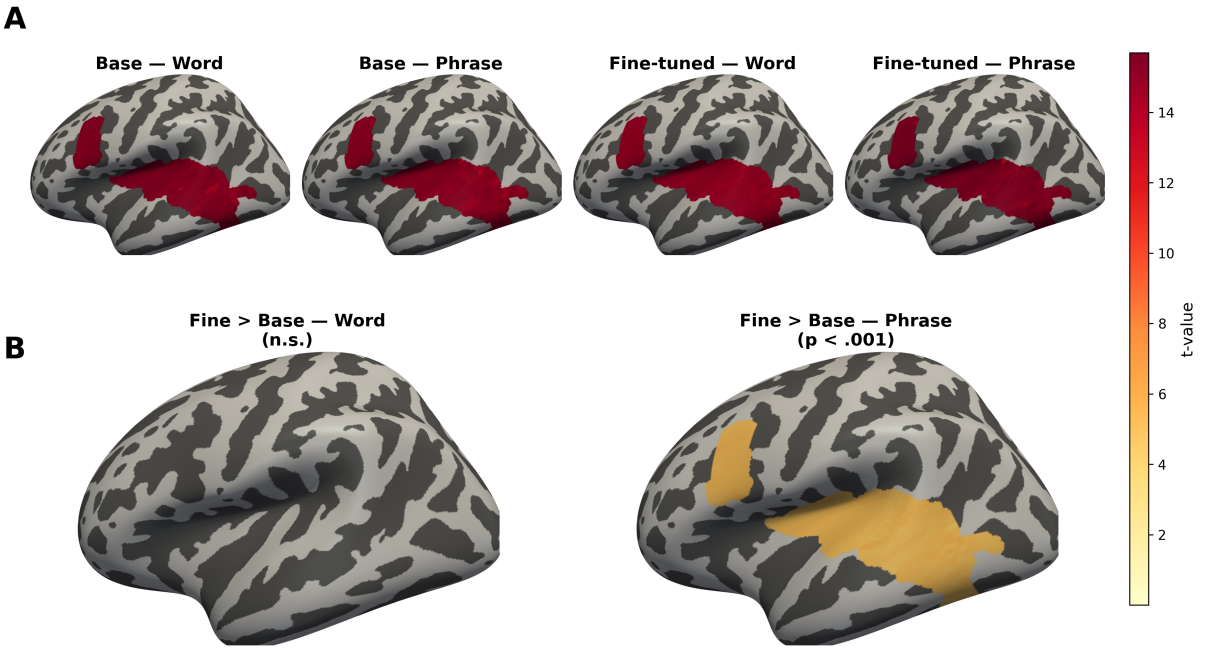

Figure 4: Neural encoding results. **(A)** Single-model encoding: both models predict cortical activity in left STG and pars opercularis (LOSO Layer 1; cluster-forming $p < .01$). **(B)** Between-model comparison: fine-tuning selectively enhances phrase encoding ($p < .001$); no word enhancement. LOSO layer distributions in Appendix Figure 6.

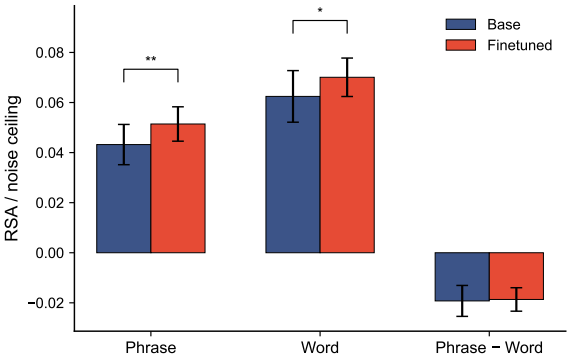

Figure 5: RSA summary: mean representational alignment (Spearman ρ / noise ceiling) between LLM attention and human STG activity. Fine-tuning significantly improves phrase alignment ($p = .004$); the word effect ($p = .044$) does not survive correction. Layer-wise profiles in Appendix Figure 7.

evidence that LLMs maintain latent tree-structured sentence representations that can be probed behaviorally (Liu et al., 2026), although our data do not, on their own, establish that the chunk-level signal we extract is hierarchical in that sense.

Fine-tuning selectively amplifies phrase alignment. Fine-tuning produces phrase-specific effects in both domains. Behaviorally, the gain on phrase boundaries exceeds the gain on shuffled boundaries (phrase-specific interaction). Neurally, the encoding enhancement is condition-specific (phrase only), layer-specific (Layer 19 in 48/50 subjects), and anatomically localized (left STG and pars opercularis within the a priori language mask); RSA confirms a phrase-level geometric improvement that survives Bonferroni correction. This combination (condition $\times$ layer $\times$ region) is what we would expect if fine-tuning sharpened intermediate-layer attention toward chunk-level structure also tracked by left language cortex. The divergence between Layer 1 (best single-model encoder) and Layer 19 (locus of fine-tuning improvement) is informative: Layer 1 attention likely captures shallow, input-driven regularities, whereas the Layer 19 gain reflects more abstract, context-sensitive representations (Tenney et al., 2019) that fine-tuning amplifies.

Constraint and scope. The absence of significant phrase $-$ word encoding or RSA effects in both models suggests that the phrase advantage reflects representational *granularity*, the scale at which features are extracted and pooled, rather than a separable compositional operation (Manning et al., 2020). Our results are best read as evidence about *how to aggregate* attention to align with human data, not as a claim that the LLM has acquired explicit compositional structure. Relative to prior work showing that scaling rather than

generic instruction tuning drives model–brain alignment (Gao et al., 2025; Pasquiou et al., 2022; Antonello et al., 2024), our results indicate that a *targeted, structure-relevant* fine-tuning objective can yield condition- and region-specific improvements not visible at the word level (Gwilliams et al., 2024; Tuckute et al., 2024). Caveats are discussed in the Limitations.

5 Conclusion

Within a single LLM and a single English reading dataset, aggregating attention at the shallow phrase (chunk) level yields closer alignment to eye movements and left language cortex than word-level attention. Phrase-segmentation fine-tuning selectively amplifies this in left STG and pars opercularis at intermediate layers; the persistent phrase − word null suggests representational granularity rather than separable composition.

Limitations

Our study has several limitations. *Single model:* all analyses use a single base LLM (Meta-Llama-3.1-8B) and its phrase-segmentation fine-tuned variant; demonstrating that the phrase superiority effect and its fine-tuning amplification replicate across model families and scales (e.g., GPT-2, Llama 7B/13B/70B) would substantially strengthen the generality of our conclusions. *"Phrase" = shallow chunk:* throughout, "phrase" refers to a shallow surface chunk produced by a rule-based segmentation pipeline (Appendix A), not a full hierarchical constituent; alternative segmentation frameworks (full constituency/dependency parses, or different chunking conventions such as CoNLL-2000) may yield different alignment patterns. *Eye-tracking measure:* the behavioral analysis uses only word-pair regressive-saccade counts, following Gao et al. (2025), because they share the lower-triangular word-pair geometry of right-to-left LLM attention. Regressive saccades are a relatively small fraction of the eye-tracking data and can be influenced by low-level factors that we did not control for (e.g., word length and frequency at the source and target word, fixation landing position, and visuospatial anchoring on physical layout); extension to forward-saccade or fixation-duration measures with explicit confound controls would strengthen the behavioral conclusion. *Numerically small correlations:* absolute Pearson r values are small and should be read as relative differences between conditions on the same data, not as estimates of absolute explained variance. *Partial correlation framework:* the partial-correlation upsamples phrase-level (and shuffled) attention back to word-pair space and may introduce residual correlation structure; independent replication with matched stimuli would further consolidate the conclusion. *fMRI scope:* encoding analyses focus on an a priori left-hemisphere language mask and a particular fixation-to-BOLD alignment procedure; broader cortical regions and alternative alignment assumptions could refine the anatomical conclusions. *Attention vs. hidden states:* the model side of all analyses is attention weights, not hidden states, to match the word-pair geometry of regressive saccades; the selection of Layer 1 in single-model encoding may therefore reflect shallow attention features (e.g., positional or frequency-related patterns) rather than deep linguistic computation. *Correlational evidence:* stronger model–brain correspondence does not establish mechanistic equivalence; causal links will require targeted interventions (e.g., attention-head ablation) and controlled stimuli.

References

Steven P Abney. 1991. Parsing by chunks. In *Principle-based parsing: Computation and Psycholinguistics*, pages 257–278. Springer.

Richard Antonello, Aditya Vaidya, and Alexander G. Huth. 2024. Scaling laws for language encoding models in fmri. *Preprint*, arXiv:2305.11863.

Kathryn Bock and Carol A Miller. 1991. Broken agreement. *Cognitive Psychology*, 23(1):45–93.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, and 1 others. 2006. An

automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.

Xufeng Duan, Bei Xiao, Xuemei Tang, and Zhenguang G Cai. 2024. Hlb: Benchmarking llms' humanlikeness in language use. *arXiv preprint arXiv:2409.15890*.

Evelina Fedorenko, Michael K. Behr, and Nancy Kanwisher. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433. Epub 2011 Sep 1.

Evelina Fedorenko, Terri L. Scott, Peter Brunner, William G. Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. 2016. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41):E6256–E6262.

D. Jake Follmer, Szu-Han Fang, Roy B. Clariana, Bonnie J. F. Meyer, and Ping Li. 2018. What predicts adult readers' understanding of STEM texts? *Reading and Writing*, 31:185–214.

Changjiang Gao, Zhengwu Ma, Jiajun Chen, Ping Li, Shujian Huang, and Jixing Li. 2025. Increasing alignment of large language models with language processing in the human brain. *Nature computational science*, 5(11):1080–1090.

Merrill F Garrett. 1975. The analysis of sentence production. *Psychology of Learning and Motivation*, 9:133–177.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Adele E Goldberg. 2003. Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, and 13 others. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Remi King. 2024. Hierarchical dynamic coding coordinates speech comprehension in the brain. *bioRxiv*.

Chun-Ting Hsu, Roy Clariana, Benjamin Schloss, and Ping Li. 2019. Neurocognitive signatures of naturalistic reading of scientific texts: a fixation-related fMRI study. *Scientific Reports*, 9(1):1–16.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jennifer Hu, Idan A. Blank, and Evelina Fedorenko. 2023. Precision fmri reveals that the language-selective network supports both phrase-structure building and lexical access during language production. *Cerebral Cortex*, 33(8):4384–4404. Published online 2022 Sep 20.

Ping Li and Roy Clariana. 2019. Reading comprehension in L1 and L2: An integrative approach. *Journal of Neurolinguistics*, 50:94–105.

Wei Liu, Ming Xiang, and Nai Ding. 2026. Active use of latent tree-structured sentence representation in humans and large language models. *Nature Human Behaviour*, 10(2):303–316.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Eric Maris and Robert Oostenveld. 2007. Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, 164(1):177–190.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. 2014. A toolbox for representational similarity analysis. *PLoS computational biology*, 10(4):e1003553.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527.

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, and Christophe Pallier. 2022. Neural language models are not born equal to fit brain data, but training helps. *Preprint*, arXiv:2207.03380.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Keith Rayner and Susan A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021.

The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601. Association for Computational Linguistics.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. 2024. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2019. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Hanlin Wu, Xufeng Duan, and Zhenguang Cai. 2025. Distinct social-linguistic processing between humans and large audio-language models: Evidence from model-brain alignment. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 135–143. NAACL 2025.

A Fine-Tuning Details

A.1 Training Hyperparameters

We fine-tuned Meta-Llama-3.1-8B using the Unsloth FastLanguageModel framework with LoRA adapters applied to query, key, value, and output projection matrices ($r = 64$, $\alpha = 128$, dropout = 0.05). Gradient checkpointing was enabled for memory efficiency. Training used the TRL SFTTrainer with batch size 32 and gradient accumulation over 8 steps (effective batch size 256), AdamW optimizer with 8-bit quantization, learning rate 3×10^{-4} with cosine annealing, and 2 epochs with sequence packing. For inference, we used deterministic decoding (temperature = 0) with a custom stop token "]" and maximum new generation length of 64 tokens.

A.2 Phrase Segmentation Rules

The fine-tuning targets were generated by a deterministic rule-based pipeline applied to constituency parses produced. The rules are intentionally shallow and framework-neutral, producing what the NLP literature would describe as "chunks" rather than recursive constituents:

1. Preserve original word order—only insert pipe symbols (|) between phrases.
2. Keep the subject noun phrase intact as the first segment.
3. Isolate the main verb or predicate as a separate segment.
4. Group objects and complements (if any) into subsequent segments.
5. Separate adverbial modifiers (prepositional phrases, adverbs, etc.) into their own segments.
6. For embedded clauses, recursively apply the same rules to split them into subject-verb-object structure.
7. Treat coordinated elements (e.g., "bread and butter") as one phrase.
8. Enclose all output results with square brackets.

Quality of the segmentation on the ReadBrain stimuli. The ReadBrain stimuli contain technical STEM language for which automatic parsers can be unreliable, so we additionally manually inspected the rule-based output. Two co-authors independently judged whether the rule-based phrase boundaries on a random sample of 200 ReadBrain sentences were linguistically plausible (i.e., did not split a base NP, break a verb group, or place a boundary mid-PP). Boundary-by-boundary agreement was 92.7%, with disagreements concentrated on prepositional-phrase attachment ambiguities (e.g., locative vs. instrumental PPs). We acknowledge that this is not an external gold-standard evaluation; we did not use a pre-existing gold-parsed dataset because available chunking corpora (e.g., CoNLL-2000) do not match the genre and vocabulary of the STEM articles in ReadBrain. The fine-tuned LLM achieves 94.2% exact-string match against this rule-based gold on a held-out test set of 140 ReadBrain sentences, vs. 25.7% for the base model.

A.3 Example Segmentations

Simple sentence:
Input: The quick brown fox jumps over the lazy dog.
Output: [The quick brown fox | jumps | over the lazy dog | .]
Relative clause:
Input: The book that I read last night was fascinating.
Output: [The book | that I read | last night | was | fascinating | .]
Complex sentence:
Input: After the discussion, I will carefully conduct the experiment based on the results.
Output: [After the discussion | , | I | will | carefully | conduct | the experiment | based | on the results | .]

B Supplementary Figures

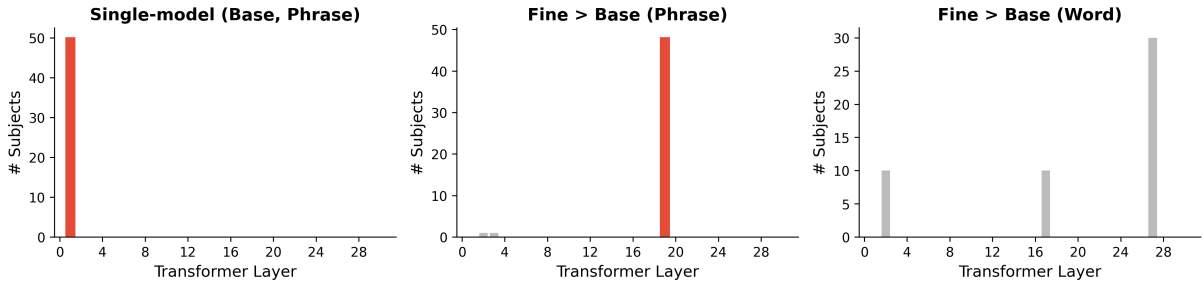

Figure 6: LOSO best-layer distribution for encoding analyses. The single-model phrase analysis and the between-model phrase comparison converge on a single dominant layer (Layer 1 and Layer 19, respectively; red bars), whereas the non-significant word comparison yields a dispersed distribution across Layers 2, 17, and 27 (grey bars), consistent with the layer-specificity of the phrase-level fine-tuning effect.

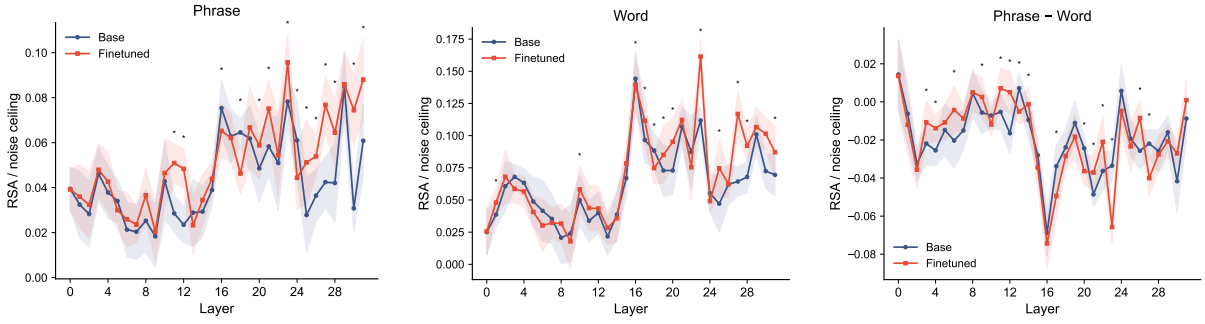

Figure 7: Layer-wise RSA profiles (Spearman ρ / noise ceiling) for phrase, word, and phrase−word conditions. Blue: base model; red: fine-tuned model. Shaded regions denote ± 1 SEM across subjects ($N = 50$). Stars indicate layers with significant paired-t difference ($p < .05$). Both models show rising RSA in middle-to-late layers; the fine-tuned model achieves higher alignment for the phrase condition.