# Training-Driven Representational Geometry Modularization Predicts Brain Alignment in Language Models

**Yixuan Liu (davidliu6125@gmail.com)[1,2,*], Zhiyuan Ma[1,2,*], Likai Tang[1,2], Runmin Gan[1,2], Xinche Zhang[1,2], Jinhao Li[2,3], Chao Xie (xiec199@gmail.com)[4,†], Sen Song (songsen@tsinghua.edu.cn)[1,2,†]**

[1]School of Biomedical Engineering, Tsinghua University, Beijing, China
[2]Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China
[3]School of Basic Medical Sciences, Tsinghua University, Beijing, China
[4]Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing, China
[*]Equal Contributors [†]Corresponding Authors

## Abstract

How large language models (LLMs) align with the neural representation and computation of human language is a central question in cognitive science. Using representational geometry as an explanatory lens, we addressed this by tracking entropy, curvature, and fMRI encoding scores throughout Pythia (70M–1B) training. We identified a geometric modularization where layers self-organize into stable low- and high-complexity clusters. The low-complexity module, characterized by reduced entropy and curvature, consistently better predicted human language network activity. This alignment followed heterogeneous spatiotemporal trajectories: rapid and stable in temporal regions (AntTemp, PostTemp), but delayed and dynamic in frontal areas (IFG, IFGorb). Crucially, reduced curvature remained a robust predictor of model–brain alignment even after controlling for training progress, an effect that strengthened with model scale. These results link training-driven geometric reorganization to temporal-frontal functional specialization, suggesting that representational smoothing facilitates neural-like linguistic processing.

**Keywords:** Cognitive Neuroscience; Language Comprehension; Representation; Neural Networks; fMRI;

## Introduction

Understanding how the human brain represents and computes language has long been a central challenge in cognitive neuroscience. The recent success of Transformer-based large language models (LLMs) has motivated their use as computational models of human language processing (Biderman et al., 2023; Chang et al., 2024; Vaswani et al., 2017). Specifically, LLM activations could significantly predict fMRI responses to linguistic stimuli using linear encoding models (AlKhamissi et al., 2025; Caucheteux & King, 2022; Goldstein et al., 2022; Rathi et al., 2025; Schrimpf et al., 2021; Tuckute, Sathe, et al., 2024), suggesting that models and brains may share common principles of language representation and computation.

Despite these successes, the mechanisms underlying model–brain alignment remain debated. While earlier work emphasized the role of syntactic structure and hierarchical processing (AlKhamissi et al., 2025; Caucheteux et al., 2023; Murphy et al., 2012), other research suggested that alignment is primarily driven by event semantics and contextual meaning (Kauf et al., 2024). Recent findings proposed a dissociation: models may rely on syntax for next-token prediction, yet their alignment with the brain appeared mediated by high-level semantic features (Proietti et al., 2025). Beyond these linguistic distinctions, shared computational principles—specifically predictive coding—have been proposed as a unifying mechanism (Goldstein et al., 2022). However, it remains unclear whether these effects reflect deep functional homology or merely a shared sensitivity to low-level statistical properties (Hadidi et al., 2025), highlighting the limitations of explaining alignment solely through external linguistic features.

Given that high correlations may stem from shared sensitivity to low-level statistics rather than functional homology, focusing solely on external linguistic features may be insufficient. Fundamentally, encoding models map between a model's representational space and the brain's neural activity space, so a complementary route is to probe the intrinsic geometry of representations. Recent work revealed that geometric properties—such as entropy and curvature—governed model adaptability (Skean et al., 2025), consistent with peak neural alignment in intermediate layers that show distinct geometric signatures (Caucheteux & King, 2022). Therefore, representational geometry was hypothesized to serve as a shared latent variable: if models and brains share computational principles, alignment should appear as structural consistency in their high-dimensional manifolds, not merely feature-level correlations.

Crucially, these geometric structures emerge dynamically during training. Yet, most research relied on static, fully trained snapshots, overlooking the developmental trajectory of alignment. While recent studies have begun to explore training dynamics (AlKhamissi et al., 2025; E. A. Hosseini et al., 2024), they remain limited to coarse, network-wide averages. Consequently, fine-grained analyses at the level of specific regions (ROIs) and systematic investigations into layer-wise modularity remain absent.

## The Present Study

Our study examines how representational geometry evolves with training and relates to brain alignment. We ask: **Does the training-induced differentiation of layers into geometric modules explain the fMRI predictability of the language network?** Specifically, we address three key sub-questions: (1) **Emergence**: Does geometric modularity emerge and stabilize over training? (2) **Mapping**: How does fMRI predictability differ across modules and ROIs? (3) **Robustness**



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* 

**and Scale**: Within modules, do geometric metrics predict fMRI scores beyond training progress, and do these effects scale with model size?

We address these questions by tracking the co-evolution of layer-wise geometry (entropy, curvature) and fMRI encoding across training, on the Pythia suite at four scales (70M–1B). Our main contributions are fourfold: (1) training induces a stable geometric modularization of layers into low- and high-complexity modules; (2) the low-complexity module shows a global alignment advantage with distinct ROI dynamics—early and stable in temporal regions, delayed and more dynamic in frontal regions; (3) curvature is a robust predictor of fMRI encoding scores beyond training progress, with flatter representations yielding higher predictability; (4) larger model scale stabilizes and strengthens geometry–alignment coupling, improving statistical detectability.

# Methods

## Brain Data and Model

**Brain Data and Stimuli.** We utilized the TUCKUTE2024 fMRI dataset (Tuckute, Sathe, et al., 2024), which contains neural responses from 5 participants during the passive reading of 1,000 sentences. We analyzed five Regions of Interest (ROIs) in the left-hemisphere language network—IFGorb, IFG, MFG, AntTemp, and PostTemp—together with the network-level average (lang_LH_netw), yielding six target signals throughout the analyses. For each sentence, responses were averaged across participants to maximize the signal-to-noise ratio.

**Model.** We employed the Transformer-based Pythia suite (Biderman et al., 2023) at four scales (70M, 160M, 410M, 1B). We analyzed 19 log-spaced checkpoints ranging from step 1 to 143K: 10 standard checkpoints for steps 1–512, followed by exponentially spaced checkpoints from step 1K onward (1K–128K), and the final checkpoint at step 143K.

## Representation Geometry Analysis

Following prior work (Skean et al., 2025), we characterized the layer-wise representational geometry by computing entropy and curvature. These metrics were calculated per sentence and averaged across stimuli.

**Entropy.** For sentence $s$ and layer $l$, let $\mathbf{Z}^{(l,s)} \in \mathbb{R}^{L_s \times d}$ be the matrix of token hidden states and $\mathbf{K}^{(l,s)} = \mathbf{Z}^{(l,s)}\mathbf{Z}^{(l,s)\top}$ be the Gram matrix. Let $p_i = \lambda_i(\mathbf{K})/\mathrm{tr}(\mathbf{K})$ be defined over nonzero eigenvalues. We defined the von Neumann entropy as:

$$E^{(l,s)} = -\mathrm{tr}(\boldsymbol{\rho}\log\boldsymbol{\rho}) = -\sum_i p_i \log p_i, \quad \boldsymbol{\rho} = \mathbf{K}/\mathrm{tr}(\mathbf{K}). \tag{1}$$

Intuitively, $E^{(l,s)}$ quantifies spectral dispersion: low entropy indicates a compressed, low-rank representation dominated by few components, whereas high entropy reflects a more distributed, high-dimensional state.

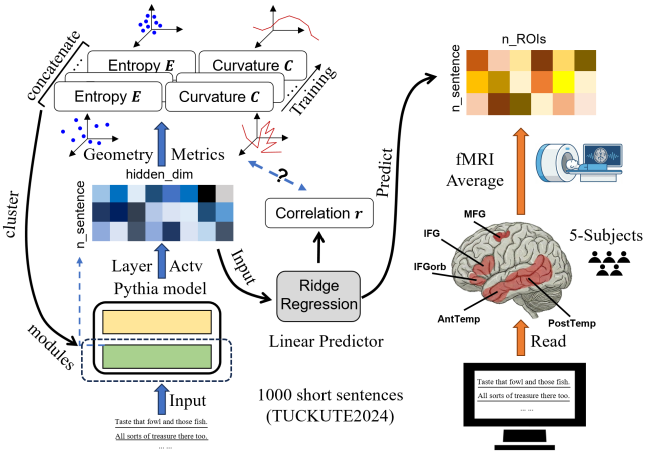


Figure 1: **Overview of the experimental framework.** We employ the TUCKUTE2024 dataset (1,000 sentences) to obtain both Pythia layer activations and human fMRI responses. Layer activations are mapped to subject-averaged fMRI data using ridge regression (5-fold CV). In parallel, we track representational geometry (entropy $E$ and curvature $C$) across training checkpoints to cluster model layers into modules, analyzing the relationship between these geometric module trajectories and fMRI predictability (Pearson correlation $r$).

**Curvature.** Following E. Hosseini and Fedorenko (2023), let $\mathbf{z}_k^{(l,s)}$ be the representation at token $k$ and $\mathbf{v}_k^{(l,s)} = \mathbf{z}_{k+1}^{(l,s)} - \mathbf{z}_k^{(l,s)}$. Curvature is defined as the mean turning angle:

$$C^{(l,s)} = \frac{1}{L_s - 2}\sum_{k=1}^{L_s-2} \arccos\left(\frac{\mathbf{v}_{k+1}{}^{\top}\mathbf{v}_k}{\|\mathbf{v}_{k+1}\|\|\mathbf{v}_k\|}\right). \tag{2}$$

Conceptually, $C^{(l,s)}$ measures the sharpness of token trajectories: higher curvature implies more local, fine-grained variation across tokens, whereas lower curvature indicates smoother, more globally consistent trajectories.

**Layer-level Aggregation.** We computed $E^{(l,s)}$ and $C^{(l,s)}$ for each stimulus sentence and averaged them across sentences to obtain layer-wise metrics $E_t^l$ and $C_t^l$ at each checkpoint $t$. This aggregation reduces content-specific variation and serves as a stable descriptor of geometry for clustering.

## Clustering and Robustness Analysis

**Geometry-Trajectory Features for Clustering.** For each layer $l$, we constructed a geometry-trajectory vector by concatenating entropy and curvature across $T$ checkpoints. This vector captures how the layer's geometry evolves over training:

$$\mathbf{x}_l = [E_1^l, C_1^l, \ldots, E_T^l, C_T^l]. \tag{3}$$

**K-means Clustering and Robustness of Layer Modules.** We fixed $K = 2$ a priori to test the low- vs. high-complexity hypothesis, applying Euclidean k-means to $\{\mathbf{x}_l\}$; we assessed the stability of the resulting partition rather than running an

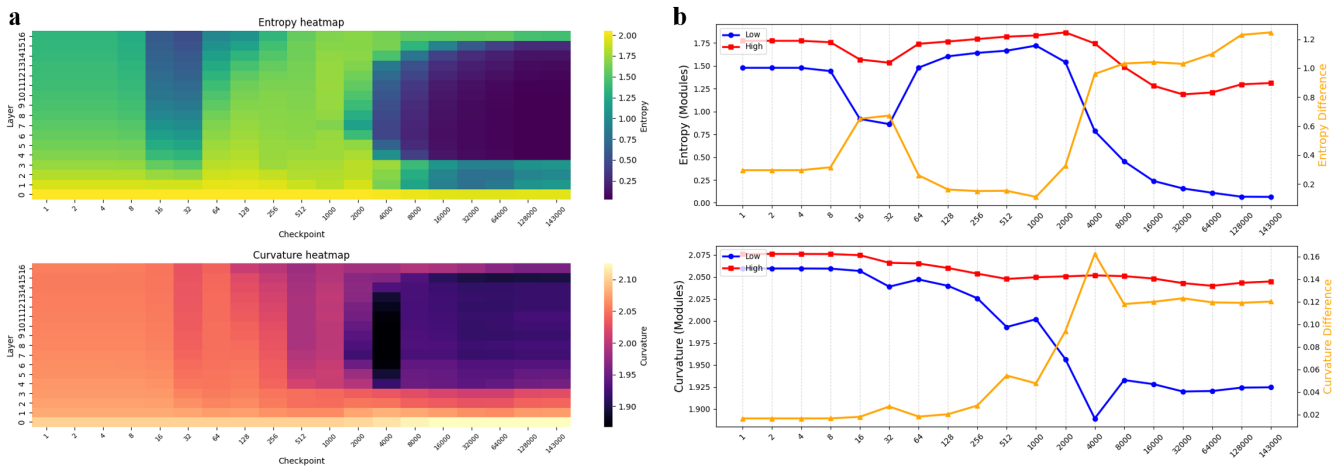


Figure 2: **Geometric evolution and modularization of Pythia-1B layers.** **(a)** Heatmaps of entropy (top) and curvature (bottom). The horizontal axis represents training checkpoints (log-scale, steps 1–143k), and the vertical axis corresponds to layers (0: embeddings, 1–16: Transformer blocks). Darker colors indicate lower values. **(b)** Aggregated geometry trajectories of the two identified layer clusters. The low-complexity module (blue) develops a stable low-entropy/curvature profile, while the high-complexity module (red) maintains higher values. The orange line shows the difference between the two modules.

exhaustive $K$ sweep, leaving silhouette/gap-statistic comparisons across $K$ to future work. To verify that the partition was not driven by specific training stages, we assessed assignment stability in two ways: (i) checkpoint bootstrap, resampling checkpoints with replacement and reclustering; and (ii) leave-one-checkpoint-out (LOCO), dropping one checkpoint at a time and reclustering. Stability was quantified per layer as the fraction of runs in which the layer retained its original cluster label.

**Geometry–fMRI Coupling Analysis.** We examined both trajectory-level and conditional associations. Trajectory coupling was quantified by Pearson correlations across checkpoints between module-averaged geometry and encoding scores. Conditional coupling was assessed using ROI-specific fixed-effects regressions:

$$fMRI_{l,t}^{(r)} = \beta_G\, G_{l,t} + \beta_t \log t + \alpha_l + \varepsilon_{l,t},$$

where $G_{l,t} \in \{E_{l,t}, C_{l,t}\}$ and $\alpha_l$ denote layer fixed effects. To improve comparability across model scales, we z-scored $fMRI_{l,t}^{(r)}$ within each ROI and z-scored $G_{l,t}$ within each module before estimating $\beta_G$. We used checkpoint-clustered robust standard errors (SEs) and applied Benjamini-Hochberg-FDR (BH-FDR) correction across ROIs ($q < 0.05$).

## Results

### Emergence of Geometry-Based Layer Modules Across Training

We first characterize the geometric evolution of Pythia-1B representations throughout the training dynamics. As shown in **Fig. 2a**, geometry is relatively homogeneous early on, but a clear banded differentiation emerges in mid-to-late training (steps >4k): middle-to-upper layers converge to a stable low-entropy pattern with a concurrent reduction in curvature, whereas shallow layers remain comparatively high on both metrics. This increasing inter-layer heterogeneity is significant: the inter-layer IQR rises from 0.31 to 0.81 for entropy and from 0.015 to 0.060 for curvature (permutation over checkpoints, $p = 0.001$ for both).

Motivated by this structure, we clustered per-layer geometric trajectories (K-means, $K = 2$) and obtained two robust layer modules. Assignments were stable under Leave-One-Checkpoint-Out (LOCO) evaluation (all layers = 1.0) and bootstrap resampling (minimum stability $\geq 0.92$). Aggregating layers within each cluster (**Fig. 2b**) shows that **low-complexity module (layer 4-15)** consistently exhibits lower entropy and curvature than **high-complexity module (other layers)** (mean Cohen's $d = 2.42$ and $d = 2.64$, respectively; FDR-corrected $p < 0.05$ across checkpoints). Entropy follows a two-phase pattern (a transient dip at steps 8–64 and a larger decline after ≥1k), whereas curvature decreases more continuously; correspondingly, the inter-cluster gap expands and stabilizes late in training (step >4k).

Overall, Pythia-1B develops a repeatable geometry-based modularization, providing a principled layer grouping for downstream analyses of module-specific brain alignment.

### Differential Brain Alignment of Geometry Modules Across Language ROIs

Having established geometry-based layer modularization, we asked whether the two modules differ in fMRI encoding performance across language ROIs, and whether the emergence of this difference is ROI-specific. For each checkpoint $t$, we defined the module-averaged fMRI encoding score within each ROI and the alignment gap as $\Delta F(t) = \text{Score}_{\text{Low}}(t) - \text{Score}_{\text{High}}(t)$, and evaluated its magnitude and

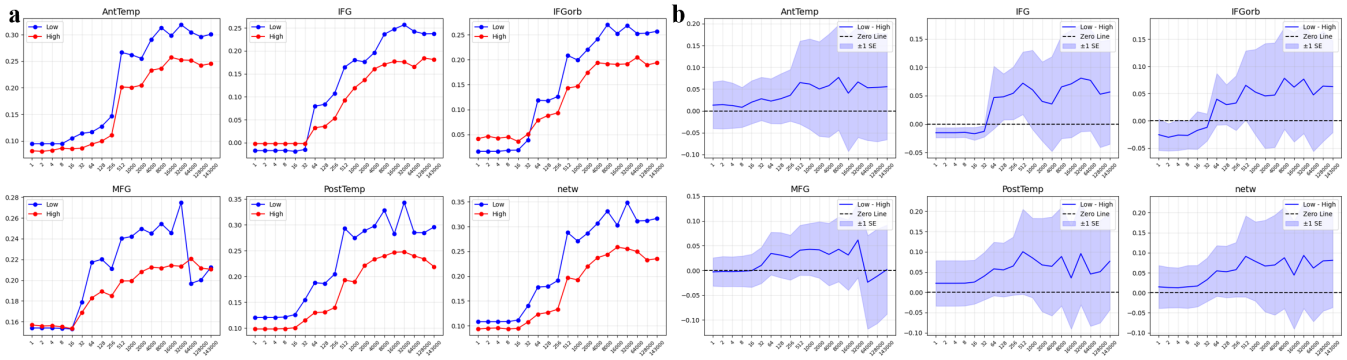


Figure 3: **fMRI alignment time course by geometry module.** **(a)** Mean cross-validated fMRI encoding scores (Pearson correlation) for the low-complexity module (blue) and high-complexity module (red) across six language ROIs over training checkpoints. **(b)** Module gap $\Delta F$ over training (blue line: mean across layers; shaded: SD across layers at each checkpoint). The dotted line denotes zero. Temporal ROIs show an earlier, more stable advantage for the low-complexity module, whereas frontal ROIs exhibit a more gradual and dynamic separation.

temporal dynamics.

**Global Trend: Low-Complexity Modules Exhibit a Robust Alignment Advantage** Across all ROIs, encoding scores increase with training (**Fig. 3a**). The low-complexity module consistently outperforms high-complexity module, yielding significantly positive $\Delta F(t)$ across the full trajectory in every ROI (paired tests on $\Delta F(t)$; all $p < 0.005$, FDR-corrected). The gap is largest in Temporal ROIs (PostTemp: $d = 2.09$; AntTemp: $d = 1.87$) and moderate in Frontal ROIs and MFG (IFG: $d = 0.95$; IFGorb: $d = 0.76$; MFG: $d = 0.79$). Thus, lower-complexity representations yield stronger fMRI predictability across the language network.

**ROI Differentiation: Timing and Trajectory of Advantage Establishment** Although low-complexity module is globally favored, the onset and shape of this advantage vary across ROIs (**Fig. 3b**).

**(i) Temporal ROIs (AntTemp / PostTemp): Early-onset and Stable.** Both AntTemp and PostTemp exhibit significantly positive $\Delta F(t)$ from the earliest checkpoints (step $\leq 64$; $t > 5.8, p < 0.001$), with no sign reversal. AntTemp additionally shows a sharp increase, with a change point at step 512 ($p < 0.05$) aligned with the onset of a stable plateau, indicating rapid and persistent specialization.

**(ii) Frontal ROIs (IFG / IFGorb): Gradual Differentiation with Early Inversion.** Frontal ROIs separate earlier via detectable transition (change point at step 64, $p < 0.01$), but the gap evolves more dynamically thereafter. Notably, IFGorb shows a brief early tendency toward high-complexity module (step $\leq 64$: mean $\Delta F = -0.014$, $p \approx 0.09$), before switching to a sustained low-complexity module advantage.

**(iii) MFG: Small and Unstable Module Specificity.** MFG shows the weakest and least consistent gap (mean $\Delta F \approx 0.018$; $t = 3.44$). No reliable early or mid-stage change point is detected (only a late signal near step 64k), and the nominal emergence point (step 32) is not followed by a stable plateau, suggesting limited selectivity for either module.

**(iv) Network-level (lang_LH_netw): Aggregate Robustness.** At the language-network level, the low-complexity module advantage is strong and consistent ($t = 8.16, d = 1.87$), stabilizing around step 512, consistent with module-level geometric differentiation translating into macroscopic encoding gains.

## Dynamic Analysis During Training: The Evolutionary Relationship Between Geometric Features and fMRI Encoding Scores

In this section, we examine how representational geometry co-evolves with fMRI encoding performance during training, and whether geometric metrics retain explanatory power after controlling for training progress and stable layer differences.

**Co-evolution on the Training Trajectory** We first quantify **trajectory-level coupling** between geometry and alignment by correlating checkpoint-averaged geometric metrics with checkpoint-averaged fMRI scores (one point per checkpoint, $n = 19$). As shown in **Fig. 4a**, **Curvature** exhibits a strong negative correlation with fMRI scores across ROIs: with the exception of MFG ($r = -0.71, p < 0.001$), all ROIs show $|r| > 0.91$ ($p < 0.001$). Thus, as training proceeds, curvature decreases while encoding performance increases, yielding tightly coupled trajectories. **Entropy** shows a weaker overall negative correlation with fMRI scores ($r \approx -0.60$, not shown), and is not significant in MFG ($p > 0.30$). Together, these results establish a strong co-evolution between geometric "flattening" and improved brain alignment.

**Conditional Coupling Controlling for Training Progress** Trajectory coupling can largely reflect shared dependence on

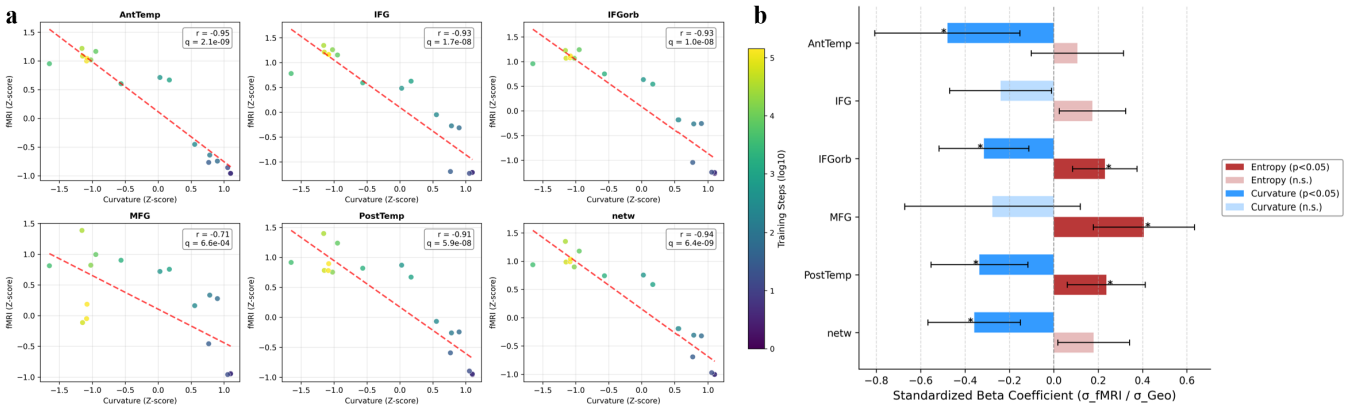


Figure 4: **Geometric co-evolution and conditional coupling with brain alignment. (a)** Scatter plots relating mean geometric metrics (Curvature/Entropy, averaged across layers) within low-complexity module to fMRI encoding scores across checkpoints ($n = 19$). **(b)** Standardized regression coefficients ($\beta_G$) and 95% CIs from $fMRI \sim G + \log(t) + \alpha_l$, where $G$ is z-scored Curvature or Entropy and $\alpha_l$ denotes layer fixed effects. Robust SEs are clustered by checkpoint; significance is based on BH–FDR correction across 6 ROIs for each metric ($q < 0.05$).

training time. We therefore test **conditional coupling** within the low-complexity module ($N = 228$ per ROI) using the ROI-specific regression model defined in Methods (Geometry–fMRI coupling analysis). As illustrated in **Fig. 4b**, after controlling for training progress, **Curvature** remains a robust negative predictor of model-brain alignment: AntTemp, IFGorb, PostTemp, and the network average show significantly negative coefficients (BH–FDR $q < 0.05$), with a similar negative trend in IFG. In contrast, **Entropy** shows ROI-dependent conditional effects, with significant positive coefficients in IFGorb, MFG, and PostTemp ($q < 0.05$) but non-significant effects in AntTemp and IFG. Note that these estimates are conditional on the training trend, their signs need not match the raw trajectory correlations. Overall, curvature exhibits stable conditional coupling across ROIs, whereas entropy is more spatially heterogeneous.

**Scale Effect** To examine the robustness of these geometry-brain relationships across scales, we extended the analysis to Pythia models from 70M to 1B (**Table 1**).

Trajectory coupling between curvature and fMRI scores is negative across sizes (median $r$ from $-0.785$ in Pythia-70M to $-0.931$ in Pythia-1B, only $r > 0$ in MFG of Pythia-160M), suggesting a scale-general synchrony of decreasing curvature and increasing alignment. Entropy shows a similar co-variation across scales, but with weaker and less consistent effect sizes than curvature.

In contrast, conditional coupling strengthens with scale. In smaller models (70M/160M), $\beta$(Curv) is small with weak sign consistency and limited detectability (0–1 FDR-significant ROIs). The effect turns negative at 410M (median $\beta = -0.15$) and becomes stronger and more consistent at 1B (median $\beta \approx -0.325$; 6/6 ROIs negative; 4/6 FDR-significant). Entropy shows positive conditional effects but does not exhibit a comparable scale-dependent increase in detectability. Thus, increasing capacity selectively stabilizes low curvature as a robust conditional predictor of improved brain alignment.

Table 1: **Cross-Scale Statistics.** Summary of Curvature ($C$) and Entropy ($E$) coupling with fMRI. $r$: median correlation (range). $\beta$: median regression coefficient (range). $N_{-/+}$: number of ROIs with $\beta < 0$ (for $C$) or $\beta > 0$ (for $E$). Det.: ROIs passing FDR correction ($q < 0.05$).

| | Curvature ($C$) | | | | Entropy ($E$) | | | |
|---|---|---|---|---|---|---|---|---|
| **Scale** | $r$ | $\beta$ | $N_-$ | Det. | $r$ | $\beta$ | $N_+$ | Det. |
| 70M | -0.785 (-0.83, -0.71) | 0.021 (-0.06, 0.30) | 2 | 1/6 | -0.576 (-0.60, -0.51) | 0.230 (0.08, 0.36) | 6 | 5/6 |
| 160M | -0.838 (-0.90, 0.50) | 0.127 (-0.07, 0.33) | 1 | 0/6 | -0.567 (-0.68, 0.42) | 0.184 (0.03, 0.32) | 6 | 3/6 |
| 410M | -0.912 (-0.95, -0.78) | -0.150 (-0.27, 0.00) | 5 | 1/6 | -0.648 (-0.77, -0.40) | 0.124 (-0.05, 0.34) | 5 | 2/6 |
| 1B | -0.931 (-0.95, -0.71) | -0.325 (-0.48, -0.24) | 6 | 4/6 | -0.603 (-0.65, -0.25) | 0.205 (0.11, 0.41) | 6 | 3/6 |

Overall, these findings reveal a distinct scale effect: increasing model capacity selectively stabilizes low curvature as a robust conditional predictor of improved brain alignment, whereas the entropy–alignment relationship remains comparatively heterogeneous and does not show the same scale-dependent strengthening.

## Discussion

Our study positions representational geometry as a candidate intermediate variable that helps organize the relationship between training dynamics and model–brain alignment. We show that geometric modularization—the spontaneous differentiation of layers into stable low- and high-complexity clusters—tracks learning rather than reflecting a fixed architectural pattern. The consistent alignment advantage of

the low-complexity module suggests that lower-dimensional, smoother representations are more readily linearly mapped to language-network responses. Crucially, the distinct trajectories of curvature (monotonic smoothing) versus entropy (non-monotonic dynamics) suggest they index different learning phases: curvature may reflect progressive manifold smoothing, whereas entropy may capture the tension between early adaptation and late-stage structural compression.

### Spatiotemporal Heterogeneity in ROI Alignment

A key finding is the dissociation between Temporal and Frontal alignment trajectories. Temporal ROIs (AntTemp, PostTemp) show an early-onset and non-inverting preference for the low-complexity module, whereas frontal ROIs (IFG, IFGorb) exhibit a delayed re-configuration and even a transient early inversion. This dissociation reflects distinct computational demands on representational stability versus controlled synthesis.

Temporal regions (AntTemp, PostTemp) primarily support mental-lexicon retrieval and basic semantic mapping, computations that benefit from stable, low-ambiguity codes across contexts. During training, the model provides such stability early by compressing representations into a low-complexity space, so temporal alignment "locks in" quickly and remains non-inverting.

In contrast, frontal regions (IFG, IFGorb) support unification and control, which require flexible interaction patterns that typically emerge later in training. Frontal alignment is therefore delayed, consistent with a computational-dependency reading: high-level synthesis is unlikely to settle while the underlying semantic substrate is still shifting, so frontal ROIs effectively wait for the temporal low-complexity foundation to stabilize, with transient early inversions plausibly reflecting this unsettled regime. We note, however, that this is a post-hoc interpretation; an alternative reading is that early frontal encodings are simply noisier or harder to read out linearly under our pipeline, independent of any computational dependency. The weak and unstable specificity of MFG is consistent with this caveat: it may act as a multimodal control hub whose alignment is less constrained by the geometric axis we measure, further motivating caution against over-interpreting the temporal–frontal contrast.

### Curvature as a Robust Predictor and the Entropy Puzzle

Our analysis isolates curvature as a robust explanatory variable for alignment: it exhibits a consistent negative coupling with fMRI scores both globally (across the training trajectory) and locally (conditional on training progress), consistent with the view that manifold smoothness is associated with better linear mapping to neural responses, although these analyses remain correlational and cannot establish a causal prerequisite. Entropy, in contrast, presents a puzzle: while globally negatively correlated with alignment, it often shows a positive conditional coefficient within specific training stages, pointing to a non-monotonic relationship. Globally, the reduction of entropy reflects the formation of structure; locally—once a baseline structure is established—slightly higher entropy may index richer effective dimensionality or capacity that captures more variance in neural data. Entropy thus behaves as a "double-edged sword" balancing compression and capacity, whereas curvature provides a more unambiguous axis of representational quality where "flatter is better" for model–brain alignment.

### The Role of Model Scale

The geometry–alignment relationship shows a clear scale effect: while trajectory-level coupling is evident across all model sizes, the conditional coupling—specifically the robust negative effect of curvature—becomes statistically detectable and consistent only in larger models (e.g., Pythia-1B). Larger models more reliably exhibit a low-curvature regime associated with higher fMRI predictability, making subtle within-stage geometry–alignment effects detectable against training noise. Scale therefore acts as an amplifier that renders the statistical link between manifold structure and neural encoding more readily detectable.

### Limitations and Future Directions

Our findings should be interpreted within specific boundary conditions. First, the reliance on linear encoding models may underestimate alignment in non-linear computational regions (e.g., frontal cortex), and our correlational design does not establish causality. Second, while geometry shifts focus away from feature-level correlations, our linear encoding pipeline remains susceptible to the low-level statistical confounds raised by Hadidi et al. (2025), which would require additional controls to fully rule out. Third, our conclusions are derived from the Pythia suite under a passive reading task. Future work should: (1) employ non-linear readouts (e.g., kernel ridge regression) to verify readout-dependence; (2) test generalization across architectures and active tasks; and (3) perform causal interventions—such as regularizing for curvature during training—to test whether geometric smoothing causally contributes to brain alignment.

## Conclusion

Across Pythia checkpoints, layers self-organize into low- and high-complexity geometric modules, with the low-complexity module consistently yielding higher fMRI encoding across language ROIs—early and stable in temporal regions, more gradual in frontal regions. Curvature remains negatively associated with encoding after controlling for training progress and layer effects, and this association strengthens with model scale. These results identify training-driven geometric modularization as an informative axis for describing model–brain alignment, complementing linguistic perspectives. Future work should test causal interventions, non-linear readouts, and generalization across architectures and neural modalities.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grants 2025ZD0215701, T2341003, U2336214, and 82302288), in part by the Beijing Natural Science Foundation (Grant L257019), and in part by Tsinghua University (School of Biomedical Engineering)—United Imaging Healthcare Joint Research Center for Magnetic Resonance Imaging (Funding No. 23012410).

We thank the authors of Tuckute, Sathe, et al. (2024) for making the fMRI dataset publicly available, and the anonymous CogSci reviewers for their constructive feedback.

During the development and preparation of this work, ChatGPT (OpenAI) was used to assist with organizing and drafting portions of the manuscript, implementing parts of the analysis code, and preparing revisions in response to reviewer feedback. All AI-assisted output was reviewed, verified, and edited by the authors, who retain full responsibility for the accuracy, integrity, and originality of the content.

## References

AlKhamissi, B., Tuckute, G., Tang, Y., Binhuraib, T. O. A., Bosselut, A., & Schrimpf, M. (2025). From language to cognition: How llms outgrow the human language network. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 24332–24350.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. *International Conference on Machine Learning*, 2397–2430.

Caucheteux, C., Gramfort, A., & King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, *7*(3), 430–441.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications biology*, *5*(1), 134.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, *15*(3), 1–45.

Fresen, A. J., Choenni, R., Heilbron, M., Zuidema, W., & de Heer Kloots, M. (2024). Language Models That Accurately Represent Syntactic Structure Exhibit Higher Representational Similarity To Brain Activity. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, *25*(3), 369–380.

Hadidi, N., Feghhi, E., Song, B. H., Blank, I. A., & Kao, J. C. (2025). Illusions of alignment between large language models and brains emerge from fragile methods and overlooked confounds. *bioRxiv*, 2025–03.

He, L., Zhong, T., Antonello, R., Mischler, G., Goldblum, M., & Mesgarani, N. (2025). Far from the Shallow: Brain-Predictive Reasoning Embedding through Residual Disentanglement. *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Hosseini, E., & Fedorenko, E. (2023). Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, *36*, 43918–43930.

Hosseini, E. A., Schrimpf, M., Zhang, Y., Bowman, S., Zaslavsky, N., & Fedorenko, E. (2024). Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, *5*(1), 43–63.

Kauf, C., Tuckute, G., Levy, R., Andreas, J., & Fedorenko, E. (2024). Lexical-semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fMRI responses in the language network. *Neurobiology of Language*, *5*(1), 7–42.

Murphy, B., Talukdar, P., & Mitchell, T. (2012). Selecting corpus-semantic models for neurolinguistic decoding. ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 114–123.

Oota, S. R., Alexandre, F., & Hinaut, X. (2022). Long-term plausibility of language models and neural dynamics during narrative listening. *Proceedings of the annual meeting of the cognitive science society*, *44*(44).

Pasquiou, A., Lakretz, Y., Thirion, B., & Pallier, C. (2023). Information-restricted neural language models reveal different brain regions' sensitivity to semantics, syntax, and context. *Neurobiology of Language*, *4*(4), 611–636.

Proietti, M., Capobianco, R., & Toneva, M. (2025). Fine-grained Analysis of Brain-LLM Alignment through Input Attribution. *arXiv preprint arXiv:2510.12355*.

Rathi, N., Mehrer, J., AlKhamissi, B., Binhuraib, T. O. A., Blauch, N., & Schrimpf, M. (2025). TopoLM: brain-like spatio-functional organization in a topographic language model. *The Thirteenth International Conference on Learning Representations*.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Skean, O., Arefin, M. R., Zhao, D., Patel, N. N., Naghiyev, J., LeCun, Y., & Shwartz-Ziv, R. (2025). Layer by Layer: Uncovering Hidden Representations in Language Models. *Forty-second International Conference on Machine Learning*.

Tuckute, G., Kanwisher, N., & Fedorenko, E. (2024). Language in brains, minds, and machines. *Annual Review of Neuroscience*, *47*(2024), 277–301.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, *8*(3), 544–561.

Urbina-Rodriguez, P., Fountas, Z., Rosas, F. E., Wang, J., Luppi, A. I., Bou-Ammar, H., Shanahan, M., & Mediano, P. A. (2026). A Brain-like Synergistic Core in LLMs Drives Behaviour and Learning. *arXiv preprint arXiv:2601.06851*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Yetman, C. C. (2024). Representation in Large Language Models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.