

Mind the Gap!

Alignment of Transformer J-Space with Cortical Representations.



Lucas Nunn
FU Berlin



Anke Borchers
Independent
Researcher



Anna Zhu
UBB Cluj



Frank Peterlein
Independent
Researcher



Saloni Kothari
FU Berlin



Claude

With
Apart Research

Abstract

Do J-Space embeddings capture semantic information in a manner analogous to the human brain? Building on recent work demonstrating representational alignment between Large Language Model (LLM) activations and fMRI responses to natural scenes, we investigate whether J-Space embeddings achieve stronger alignment with higher-level cortical areas than standard LLM embeddings. Using representational similarity analysis (RSA) on large-scale fMRI data, we find J-Space embeddings to reach largely similar alignment as raw LLM embeddings. We identify two primary constraints of the paradigm: First, textual scene captions already reflect human-compressed semantic abstraction. Second, visual-sensory cortical responses do not map directly onto the higher-level cognitive functions hypothesized to occupy J-Space. Future work would utilize datasets targeting higher-order cognitive functions to investigate neurobiological correlates of J-Space representations more tightly.

1. Introduction and Related Work

Emergent internal structures in Large Language Models (LLMs), such as the recently identified "J-Space", a subspace inspired by Global Workspace Theory (GWT), exhibit functional properties analogous to theories about human consciousness^{1,2}. This raises a fundamental neuroconnectionist question: do these structures mirror neurobiological representations in the human brain³, supporting the hypothesis that cognitive systems converge on universal representations of reality⁴? Grounding LLM internal dynamics in human neural data provides practical value by establishing whether mechanistic interpretability constructs (like J-Space) correspond to biological reality, offering a benchmark for assessing model alignment, interpretability tools, and artificial cognitive architectures.

We build on established Representational Similarity Analysis (RSA) paradigms that demonstrate alignment between fMRI responses to natural scene images and LLM embeddings of corresponding scene captions. While prior work confirms that LLMs and human cortices share high-level semantic representations beyond low-level visual features, it remains unknown whether isolating workspace-level, verbalizable activations (J-Space) shows enhanced alignment with higher-order cortical areas. Using a large-scale natural scene fMRI dataset, we execute a comparative RSA framework to measure brain alignment against both raw LLM embeddings and their corresponding J-Space projections.

Our main contributions are:

1. **First Cortical Evaluation of J-Space:** We present the first neuroconnectionist investigation searching for biological correlates of J-Space representations using high-resolution fMRI data.
2. **Replication and Scaling of LLM-Brain Alignment:** We independently replicate and scale prior RSA alignment findings between textual caption embeddings and cortical scene responses using a significantly larger, modern LLM architecture.
3. **Characterization of Methodological Boundaries:** We identify a fundamental "Workspace Gap," demonstrating that while J-Space matches raw embedding alignment, passive visual scene perception is insufficient to engage the distinct higher-order cognitive processing hypothesized to populate J-Space.

3. Methods

We implement a comparative Representational Similarity Analysis (RSA)⁶ framework measuring alignment between human cortical activity during scene perception and two matched language-model representations: raw residual-stream activations and their Jacobian-lens (J-space) transports. We reuse an existing analysis pipeline⁵, which readers should refer to for full fMRI preprocessing, searchlight and RSA details.

3.1 Dataset and stimuli

The Natural Scenes Dataset (NSD) is a 7T fMRI dataset⁷ in which 8 subjects viewed MS COCO⁸ natural scenes, each presented three times. Analysing sessions 1–10 and retaining only three-repetition conditions gives 835 conditions per subject and a 6,148-image union. Betas are z-scored within a session and averaged across repetitions. Model input is each image's human-written MS COCO captions; no pixels reach the model.

3.2 Language-model embeddings

Captions are passed through Qwen3.5-4B⁹ (2,560-d residual stream) without chat template, generation, or truncation, reading out layers 8, 16, 23, and 30.

- **Raw:** residual stream mean-pooled over all valid non-padding tokens.
- **J-space:** the Jacobian lens linearly transports a residual vector into the final-layer basis¹, where the model's own unembedding decodes it as vocabulary. We use the released lens, a fixed average Jacobian over 1,000 WikiText prompts, not a per-stimulus estimate. Transport is applied per token before pooling; being linear, this equals transporting the pooled vector, verified each batch numerically.

Unembedding yields caption-relevant tokens (Figure 2) alongside occasional off-target items ("pizza", "cheese" for a cake). These are an interpretive diagnostic only — RSA uses the full 2,560-d vectors.

3.3 RSA and statistics

RDMs use correlation distance ($1 - \text{Pearson } r$) on unnormalised features. Model RDMs are built per layer; brain RDMs by searchlight, each valid voxel centring a 6-voxel-radius sphere. Model and brain RDMs are compared by Pearson correlation, giving a whole-brain map per model. Scores aggregate over valid searchlight centres within each of eight disjoint 100-stimulus samples, then across samples within the subject, leaving subjects as the independent unit.

An MPNet sentence encoder provides a semantic reference^{10, 11} over identical stimuli and pipeline, serving as our check that the reused framework reproduces prior results; its correlations and brain maps closely resemble the LLM features, consistent with Doerig et al. Paired J-versus-raw contrasts use a two-sided exact sign-flip test over all 2^8 sign assignments of the eight subject differences, Benjamini–Hochberg¹² corrected within predeclared families.

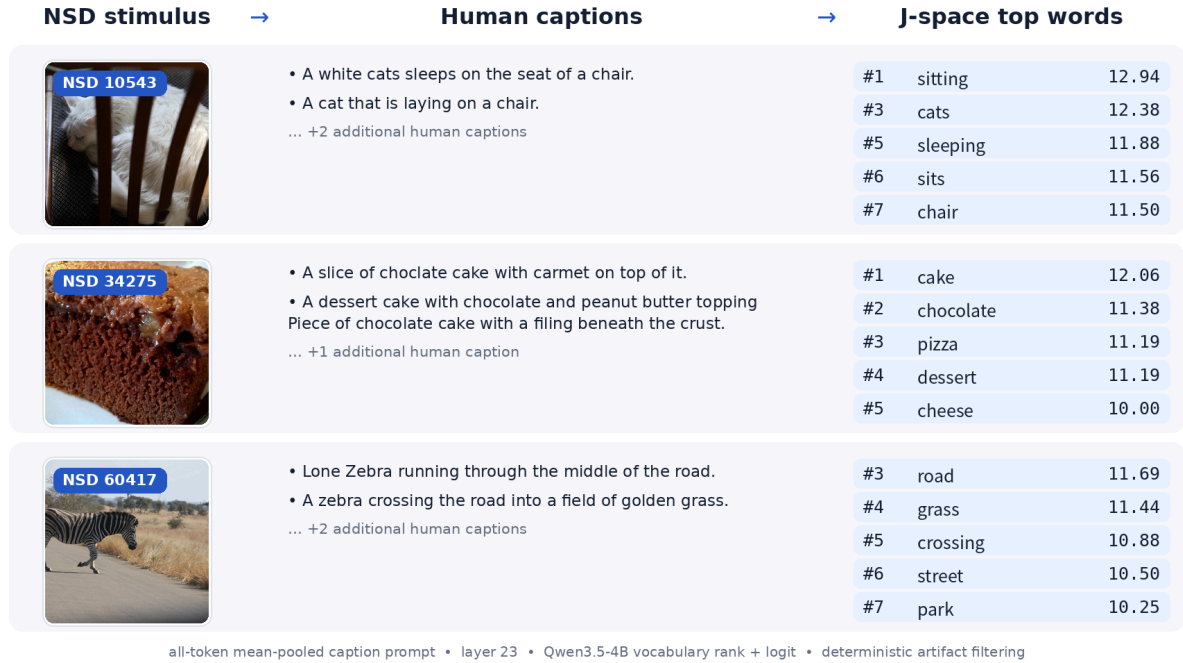


Figure 2: Example Interpretable J-Space Unembeddings.

Representative NSD stimuli, their human MS COCO captions, and the top-ranked vocabulary tokens obtained by unembedding the layer-23 J-space vector. Rows are selected without inspecting content: the subject-1 set is restricted to CC BY 2.0 licensed images, then sampled at indices $\lfloor n/6 \rfloor$, $\lfloor n/2 \rfloor$ and $\lfloor 5n/6 \rfloor$ of the sorted eligible set. Displayed tokens retain their raw vocabulary ranks after a formatting-only filter; no semantic term was selected or removed. The readouts recover core caption concepts (e.g., "cats", "sleeping") alongside occasional off-target items (e.g., "pizza", "cheese" for a cake). Sequence structure is absent because the readout is mean-pooled, not because of the J-space transport — the raw control is pooled identically. These readouts are an interpretive diagnostic only: RSA uses the full 2,560-dimensional vectors, where J-space alignment remains close to raw (+5% at this layer).

4. Results

We present the surface-based searchlight Representational Similarity Analysis (RSA) results comparing raw LLM residual stream activations (h_l) and their corresponding J-Space projections ($J_l h_l$) against voxelwise fMRI responses.

4.1 Equivalent Spatial Distribution and Correlation Strengths

Across all subjects, searchlight RSA reveals virtually identical spatial topographies of representational alignment for both raw LLM hidden states and J-Space projections:

- **Primary Visual & Ventral Stream Localization:** Strongest correlation values (peaking between $r \approx 0.29$ and $r \approx 0.32$) are tightly clustered around early visual sensory areas and ventral temporal stream cortices responsible for visual object recognition and semantic scene parsing.
- **Minimal Divergence in Association Cortices:** Neither raw LLM embeddings nor J-Space projections demonstrate significant alignment with prefrontal or higher-order association networks during passive scene perception.
- **Marginal Numerical Differences:** Maximum absolute correlation values for J-Space embeddings (0.30 - 0.32) show a minor, statistically negligible increase compared to raw embeddings (0.29 - 0.30).

4.2 Replicating Prior Semantic Alignment

These empirical findings successfully replicate prior work establishing semantic representational alignment between LLM caption representations and human visual cortices. However, extracting the J-Space does not yield the hypothesized shift toward higher-level cognitive brain regions, demonstrating that projecting into the verbalizable global workspace preserves broad semantic similarity without enhancing cortical alignment in passive viewing paradigms.

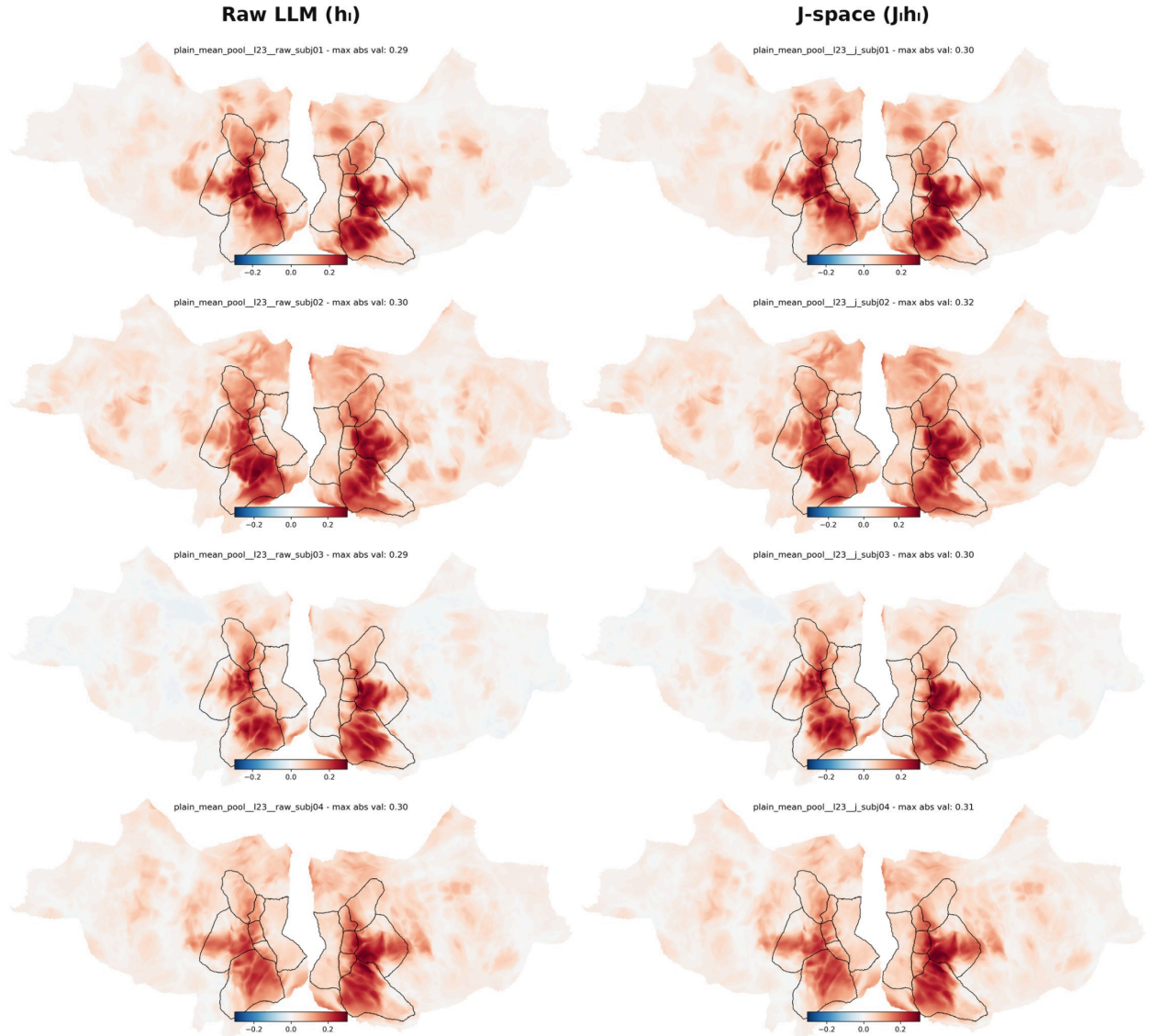


Figure 1: Searchlight RSA Cortical Maps.

Searchlight RSA correlation maps computed with a 6-voxel searchlight radius across the cortical flatmaps for subjects 1–4 using Qwen layer 23 embeddings. The left column shows alignment for raw LLM embeddings (h_i), while the right column shows alignment for J-Space projected embeddings ($J_i h_i$). Both methods display nearly identical peak correlations localized primarily to ventral visual cortices, replicating previous findings and demonstrating consistent spatial patterns across participants without significant divergence in higher-order cortical regions.

4.3 Layerwise Quantitative RSA Alignment

To rigorously quantify representational alignment across depth, we evaluate mean Pearson correlations across all brain RDMs for raw LLM embeddings, J-Space projections, and baseline controls across layers 8, 16, 23, and 30 (Table 1).

- **Peak Alignment at Layer 23:** Alignment monotonically increases through the network, reaching its maximum at layer 23 for both raw embeddings ($r = 0.029810$) and J-Space projections ($r = 0.031460$).
- **Statistically Significant but Negligible Effect Size:** Benjamini-Hochberg corrected q-values reveal statistically significant differences ($q = 0.04688$) favoring J-Space embeddings over raw embeddings at middle-to-late layers (Layers 23 and 30). However, the absolute effect size is minuscule ($\Delta r \approx 0.0016$), confirming that J-Space filtering yields functionally equivalent alignment to the unprojected residual stream.
- **Control Baselines:** The final-block control ($r = 0.029597$) and the dedicated sentence-embedding baseline (MPNet reference: $r = 0.026736$) exhibit comparable alignment performance. MPNet achieves this level despite drastically reduced parameter counts, likely because it is explicitly trained for dense single-sentence semantic embeddings.

Layer	Raw mean	J mean	J-row	95% CI on J-row	BH q
8	0.025914	0.025438	-0.000476	[-0.002614, +0.001662]	0.64062
16	0.029382	0.029738	+0.000356	[-0.001442, +0.002155]	0.64062
23	0.029810	0.031460	+0.001649	[+0.000617, +0.002682]	0.04688
30	0.028265	0.029724	+0.001459	[+0.000514, +0.002405]	0.04688
Final-block control	—	0.029597	—	—	—
MPNet reference	—	0.026736	—	—	—

Table 1: Alignment across layers

Whole-searchlight mean RSA correlation for raw and J-space embeddings at Qwen layers 8, 16, 23 and 30, averaged across eight subjects with subjects as the independent unit. The two representations are near-identical and both peak at

layer 23. Paired J-minus-raw differences survive Benjamini–Hochberg correction at layers 23 and 30 ($q = 0.047$) but not at 8 or 16; the surviving effects are small — roughly $+0.0016$ on a base of ~ 0.03 , about 5%. The MPNet reference performs comparably despite being far smaller (110M vs 4B parameters), plausibly because it is trained specifically to produce single-sentence embeddings.

5. Discussion, Limitations, and Future Outlook

Understanding the capabilities of AI systems is fundamental to assessing their uses and potential risks. Furthermore, interpreting and steering the potential motivations and actions of AI systems requires understanding the higher-order emergent cognitive architectures orchestrating their behaviors at fundamental levels. The J-Space represents a step in this direction, as it surfaces an emergent cognitive structure functionally analogous to those in humans. To gain a deeper understanding of the J-Space, it is necessary to draw on comparisons with the human brain, which remains the current SOTA AGI system.

Our results produce a modest replication of the existing finding of semantic representational alignment between LLMs and higher-level visual areas in the brain, but do not find unique alignment of J-Space representations. Our results reveal the fundamental limitations of our paradigm, and point towards future work rather than casting doubt on the importance of the J-Space. First, image captions already embody compressed semantic representations of raw image stimuli, with processing being carried out by the human captioners. As a result, there is no semantic information for the J-Space to add, which is clear from the close mirroring of the J-Space unembeddings to the original captions. Second, the proposed analogy of the J-Space to the human global workspace points towards higher-order cognition, which is not captured by our passive visual processing dataset.

The natural next step is to search for representational alignment between J-Space and brain representations across neural datasets featuring stimuli and task sets targeted at higher-order cognition. Especially relevant would be datasets targeting Global Neuronal Workspace Theory specifically, with resulting alignment providing neurobiological evidence for the J-Space acting as a global workspace within LLMs.

Code and Data

- **Code repository:** <https://github.com/lucas-nunn/natural-j-scenes>
- **Dataset:** <https://naturalscenesdataset.org/>

References

1. Gurnee, W. et al. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread (2026).
<https://transformer-circuits.pub/2026/workspace/index.html>
2. Mashour, G. A., Roelfsema, P., Changeux, J.-P. & Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798 (2020).
<https://doi.org/10.1016/j.neuron.2020.01.026>
3. Doerig, A. et al. The neuroconnectionist research programme. *Nat. Rev. Neurosci.* 24, 431–450 (2023).
<https://doi.org/10.1038/s41583-023-00705-w>
4. Huh, M., Cheung, B., Wang, T. & Isola, P. Position: The Platonic Representation Hypothesis. *Proc. 41st Int. Conf. on Machine Learning*, PMLR 235, 20617–20642 (2024).
<https://doi.org/10.48550/arXiv.2405.07987>
5. Doerig, A. et al. High-level visual representations in the human brain are aligned with large language models. *Nat. Mach. Intell.* (2025).
<https://www.nature.com/articles/s42256-025-01072-0>
6. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis: connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 4 (2008).
<https://doi.org/10.3389/neuro.06.004.2008>
7. Allen, E. J. et al. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* 25, 116–126 (2022).
<https://doi.org/10.1038/s41593-021-00962-x>
8. Lin, T.-Y. et al. Microsoft COCO: common objects in context. *Proc. European Conf. on Computer Vision*, LNCS 8693, 740–755 (2014).
<https://doi.org/10.48550/arXiv.1405.0312>
9. Qwen Team. Qwen3.5-Omni technical report.
<https://doi.org/10.48550/arXiv.2604.15804> (2026).
10. Song, K., Tan, X., Qin, T., Lu, J. & Liu, T.-Y. MPNet: masked and permuted pre-training for language understanding. *Adv. Neural Inf. Process. Syst.* 33 (2020).
<https://doi.org/10.48550/arXiv.2004.09297>

11. Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. Proc. EMNLP-IJCNLP 3982–3992 (2019).
<https://doi.org/10.18653/v1/D19-1410>

12. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300 (1995).
<https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>

Appendix

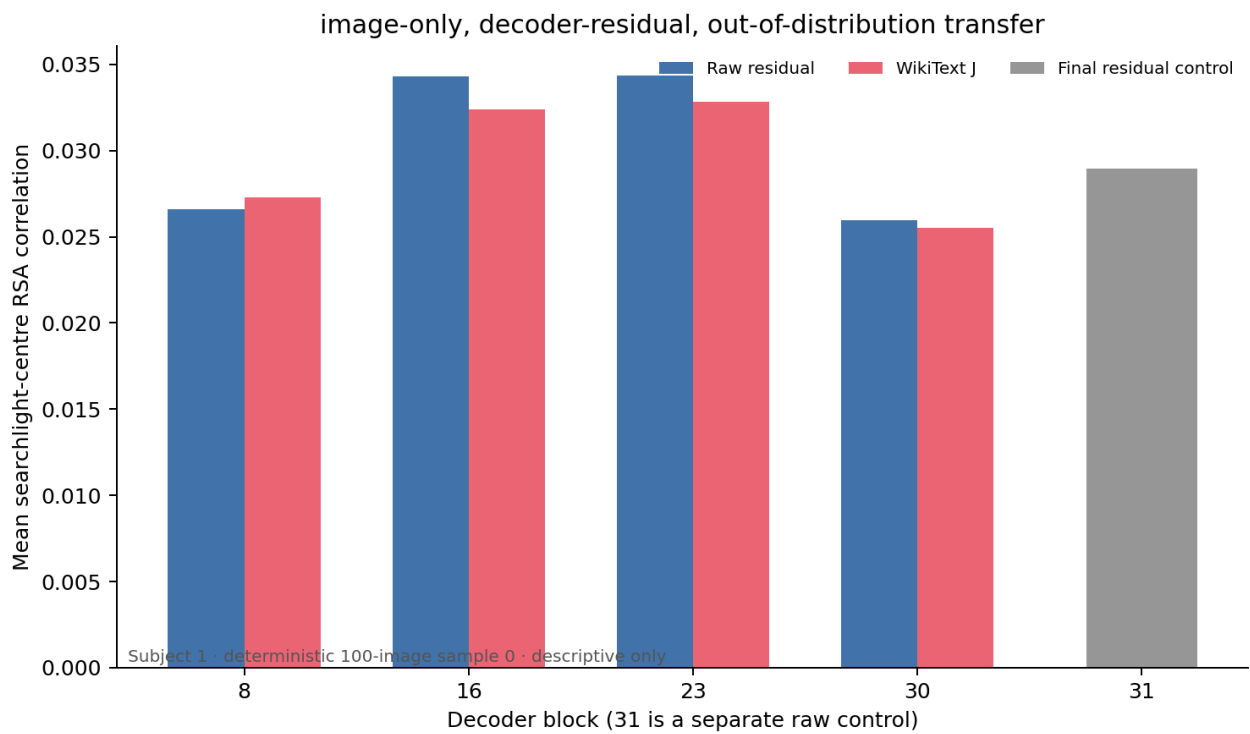


Figure 1: Alignment across layers in raw-image condition

Out-of-distribution check — raw images through Qwen's vision stack instead of captions, with the WikiText-trained J-lens applied to the resulting decoder residuals. Bars are whole-searchlight mean RSA over 365,127 subject-1 centres for one 100-image sample, so this is descriptive only. Raw peaks around layers 16–23 (≈ 0.034). J-space sits slightly *below* raw at layers 16, 23 and 30, and marginally above only at layer 8; block 31 is a raw-only final-residual control. Not surprising — a lens estimated on WikiText text residuals is well outside its domain here.

LLM Usage Statement

We used Claude to implement the code and improve the manuscript prose.