Intermediate Layers of LLMs Align Best With the Brain by Balancing Short- and Long-Range Information

Michela Proietti (mproietti@diag.uniroma1.it)
Sapienza University of Rome, Italy

Roberto Capobianco (roberto.capobianco@sony.com)
Sony AI, Zurich, Switzerland

Mariya Toneva (mtoneva@mpi-sws.org)
MPI for Software Systems, Saarbrücken, Germany

Abstract

Contextual integration is fundamental to human language comprehension. Language models are a powerful tool for studying how contextual information influences brain activity. In this work, we analyze the brain alignment of three types of language models, which vary in how they integrate contextual information. Despite differences among models, we find minimal variations in their brain alignment. In line with previous research, middle layers consistently show the highest correspondence with brain activity. Interestingly, this alignment appears to strengthen with longer context inputs, pointing to improved sensitivity to extended linguistic information. To better understand how contextual integration affects brain alignment, we analyze the roles of short- and long-range context using variance partitioning. Our findings highlight a functional distinction between layers, suggesting a trade-off between retaining local detail and integrating broader context. This interplay may explain the robust alignment of middle layers with brain responses.

Keywords: Transformers; State-space models; Brain alignment; fMRI; Variance partitioning

Introduction

Language models (LMs) have successfully contributed to our understanding of brain responses to natural linguistic stimuli, emphasizing the essential role of contextual information in both artificial and biological language comprehension systems. Recent work on brain alignment suggests that hierarchical processing in LMs might resemble the brain's organization, with distinct regions integrating information over progressively longer timescales (Hasson, 2025; Mischler, Li, Bickel, Mehta, & Mesgarani, 2024). Notably, peak brain alignment consistently occurs in LMs' middle layers, suggesting that later layers may serve specialized integrative roles (Toneva & Wehbe, 2019; Caucheteux & King, 2022; Mischler et al., 2024).

In this work, we extend brain alignment analyses to state-space models (SSMs), designed for efficient long-context processing (Gu, Goel, & Ré, 2021), and to transformer-SSM hybrid architectures. We compare both types of models to transformers and find no significant differences in brain alignment. Middle layers are confirmed as the most brain-aligned, but unlike earlier findings, their alignment continues to improve with increasing context length.

To further investigate the contributions of short- and long-range context to brain alignment, we perform variance partitioning (Borcard, Legendre, & Drapeau, 1992). Our results show that early model layers primarily capture short-range information, while late layers tend to abstract away such local detail when longer contexts are provided. In contrast, middle layers better integrate short- and long-range context, which may underlie their superior brain alignment. Interestingly, we also find that late layers rely on short-context signal to predict long-timescale brain regions, suggesting that these areas may preserve and relate local linguistic detail over extended temporal windows.

Methods

Dataset. We use a publicly available fMRI dataset (Wehbe et al., 2014) of 8 subjects reading chapter 9 of *Harry Potter and the Sorcerer's Stone* (Rowling et al., 1998) word-by-word.

LM Representations. We consider 5 pretrained LMs with 1 to 2B parameters: 3 transformers (Falcon3-1B-Base (Team, 2024), Gemma-2B (Mesnard et al., 2024), and Llama3.2-1B (Meta AI, 2024)), 1 SSM (Mamba-1.4B (Gu & Dao, 2023)), and 1 hybrid model (Zamba2-1.2B (Glorioso et al., 2024)). For every word w, we feed the models a context of $L \in \{1, 5, 10, 20, 40, 80, 160, 320, 640\}$ words with w as last word, and extract token embeddings at every layer.

Encoding Models. We train a ridge-regularized linear model per fMRI voxel and participant (Jain & Huth, 2018; Toneva & Wehbe, 2019; Schrimpf et al., 2021). We use 4-fold CV on four consecutive, non-overlapping text blocks ($\approx 25\%$ each), and discard 5 TRs at each fold boundary to avoid hemodynamic leakage. We evaluate brain alignment using Pearson correlation between the true and predicted fMRI data. Given the inherent noise in fMRI data, we estimate the noise ceiling as the ability to predict brain activity of one subject using data from other subjects (Schrimpf et al., 2021), and discard voxels with estimated noise ceiling values < 0.05.

Variance Partitioning. We use variance partitioning (Borcard et al., 1992) to determine the amount of unique variance in brain activity explained by feature spaces extracted using short (5-word) and long (640-word) contexts, respectively. To estimate the variance explained when all feature spaces are used together, we compute a shared feature space using stacked regressions (Lin, Naselaris, Kay, & Wehbe, 2024).

Significance Tests. We perform pairwise Wilcoxon signed-rank tests with Benjamini–Hochberg FDR correction (Benjamini & Hochberg, 1995) to assess significant differences in brain alignment across layers (early, middle, late) and across models and differences in explained variances across layers.

Results

No Significant Differences in Brain Alignment Across Architectures. Overall, we find no significant differences in average alignment between transformers, SSMs, and hybrid models, suggesting that different LLMs can learn similar representations. To verify such similarity, for each context length, we applied an RBF-kernel CKA on the TR embeddings, as in (Mischler et al., 2024), after linearly resampling layers onto a 16-point relative-depth grid. Fig. 1 shows the average CKA matrix across context lengths, revealing moderate-to-high similarity for every model pair. This supports aggregating results across models in subsequent analyses.

Layer Hierarchy: Middle Layers Show Peak Brain Alignment. We examine how brain alignment varies across model layers, averaging across subjects, models, language-related regions of interest (ROIs), and layer depth. The ROIs we consider are the angular, inferior frontal, and middle frontal gyri (AG, IFG, MFG), and the anterior temporal (AT), poste-

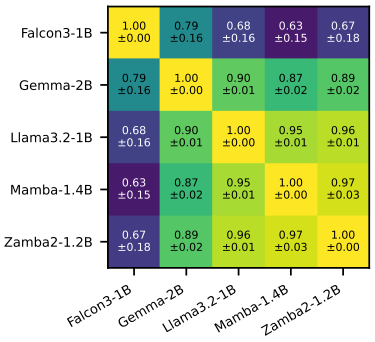

Figure 1: Kernel-CKA for all LLM pairs, averaged across context lengths (mean $\pm$ s.d.). Off-diagonal values ≥ 0.63 indicate the different models converge on similar representations.

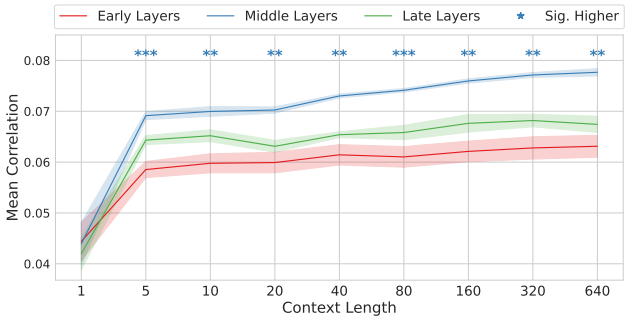

Figure 2: Average brain alignment across subjects, models, ROIs, and layer depth (early, middle, late) with standard errors across models. Asterisks indicate significantly higher alignment in middle layers: $**$, and $***$ identify p-values $< 0.01, 0.001$ respectively.

rior temporal (PT), posterior cingulate (PC), and dorso medial prefrontal cortices (dmPFC). In line with prior work (Toneva & Wehbe, 2019; Caucheteux & King, 2022), we find that middle layers consistently exhibit the highest alignment with brain activity for $L > 1$, significantly outperforming both early and late layers (Fig. 2). This pattern holds across all models.

Alignment of Middle Layers Increases with Context Length. Prior studies reported that brain alignment in earlier models typically plateaus or declines beyond 500-word contexts (Aw & Toneva, 2023) with earlier LMs. In contrast, we observe a consistent increase in middle-layer alignment across all models as context length grows (Fig. 2). This novel finding suggests that middle layers in current LMs continue to extract brain-relevant features as more context is provided, pointing to improved integration of long-range linguistic information.

Short–Long Range Trade-Off Drives Middle Layer Alignment. To understand why middle-layer alignment increases with longer input and exceeds that of late layers, we perform variance partitioning on representations from short (5-word) and long (640-word) contexts (Fig. 3). Early layers show the highest allocation, among all layers, of brain alignment to shared variance between short- and long-context. We expect

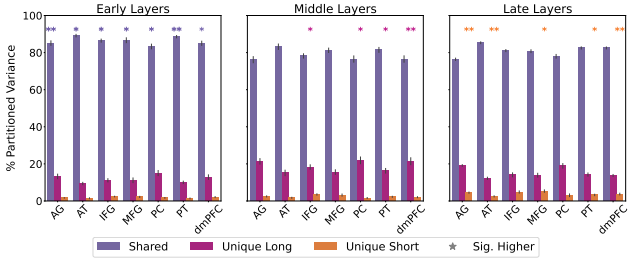

Figure 3: Proportion of the total partitioned variance explained uniquely by short (5-word) and long (640-word) contexts and shared by both, averaged across models and subjects, for each layer depth and ROI, with standard error across models. Asterisks indicate significantly higher variance than all other layer depths: $*$, and $**$ identify p-values $< 0.05, 0.01$ respectively. Middle/late layers explain the highest long-/short-context unique variance, respectively.

that this is due to the early layers' focus on short-range information, even in longer contexts (Mischler et al., 2024). In contrast, late layers allocate a significantly larger proportion of variance to short contexts, suggesting that short-range information is not well preserved when longer inputs are integrated. If short-range information were maintained, it would be reflected in the shared variance between short and long contexts. Interestingly, this effect in late layers is statistically significant across models in ROIs associated with long temporal receptive windows (Lerner, Honey, Silbert, & Hasson, 2011) (e.g. MFG, AG, and dmPFC), suggesting that these areas also rely on short-range information. Overall, middle layers integrate short- and long-range information most effectively: they allocate a lower proportion of variance to short contexts than late layers do, and achieve the highest proportion of variance uniquely explained by long contexts, reflecting enhanced sensitivity to long-range information.

Discussion and Conclusion

In this study, we compared the brain alignment of transformers, SSMs, and hybrid models and found no significant differences across architectures. This likely stems from the high cross-model similarity we observe in the LLM embeddings themselves, and residual differences may be masked by the limited context length in current fMRI datasets. Consistent with prior work, middle layers showed the highest brain alignment. Variance partitioning suggests this stems from a trade-off in middle layers between short- and long-range information. Surprisingly, the percentage of short-context unique variance explained by late layers is highest in long-timescale ROIs. This suggests that these brain areas preserve local detail, which may support integration over extended temporal windows. These findings offer new insights into how LMs process context across layers and how these dynamics relate to brain activity. Future work should consider longer contexts to further assess differences across models.

References

Aw, K. L., & Toneva, M. (2023). Training language models to summarize narratives improves brain alignment. In *The eleventh international conference on learning representations.* Retrieved from https://openreview.net/forum?id=KzkLAE49H9b

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, *73*(3), 1045–1055.

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

Glorioso, P., Anthony, Q., Tokpanov, Y., Golubeva, A., Shyam, V., Whittington, J., ... Millidge, B. (2024). *The zamba2 suite: Technical report.* Retrieved from https://arxiv.org/abs/2411.15242

Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A., Goel, K., & Ré, C. (2021). Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Hasson, U. (2025). Uncovering a timescale hierarchy by studying the brain in a natural context. *Journal of Neuroscience*, *45*(12).

Jain, S., & Huth, A. (2018). Incorporating context into language encoding models for fmri. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of neuroscience*, *31*(8), 2906–2915.

Lin, R., Naselaris, T., Kay, K., & Wehbe, L. (2024). Stacked regressions and structured variance partitioning for interpretable brain maps. *NeuroImage*, *298*, 120772.

Mesnard, G. T. T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., ... Kenealy, K. (2024). Gemma: Open models based on gemini research and technology. *ArXiv*, *abs/2403.08295*.

Meta AI. (2024). *LLaMA 3.2.* https://github.com/meta-llama/llama-models/blob/main/models/llama3$_2$/$MODEL_CARD.md$.

Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., & Mesgarani, N. (2024). Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, 1–11.

Rowling, J. K., GrandPre, M., GrandPré, M., Taylor, T., Books, A. A. L., & Inc., S. (1998). *Harry potter and the sorcerer's stone*. A. A. Levine Books. Retrieved from https://books.google.de/books?id=zXgTdQagLGkC

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118. doi: 10.1073/pnas.2105646118

Team, F.-L. (2024, December). *The falcon 3 family of open models.* Retrieved from https://huggingface.co/blog/falcon3

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, *32*.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *in press*.