Human Language Network Modeling with Selective State Spaces

Sari Sadiya*[1], Anthea Hohmann*[1], and Gemma Roig[1]

1- Goethe Universität, Frankfurt

Abstract. A growing body of literature have utilized embeddings from transformer-based large language models (LLMs) to study human brain activity during naturalistic language comprehension. While these studies provide powerful insights into the human language network, the focus on primarily transformer architectures raises concerns about the generality and replicability of resulting neuroscientific conclusions. Selective state space models such as Mamba offer an alternative formulation in which linguistic input is integrated sequentially into a continuously updated latent state, a property that may better align with some theories of incremental human sentence processing. Using fMRI data from naturalistic story reading, we compare brain-model alignment across Mamba and GPT-2 models of matched scales under the same input perturbations. We find that architectural choices lead to systematic differences in alignment patterns, indicating that neural inferences can depend strongly on the specific model used. These results highlight a broader methodological risk: conclusions about the neural computations supporting language may not be robust when validated on only one model class. Our findings demonstrate the need for multi-model benchmarking and emphasize that reliable cognitive insight from LLMs requires validation across diverse architectures.

1 Introduction and Related Work

Transformer-based Large Language Models (LLMs) have substantially advanced research into the neural basis of language. Numerous studies have shown that embeddings from these LLMs reliably predict patterns of brain activity during reading and listening tasks [1, 2, 3, 4]. Going even further, several researchers have argued that components unique to transformer-based architectures – particularly self-attention – bias the embeddings of the models towards representations that resemble human neural processing [1, 4].

However, how closely their internal computations correspond to human cognitive mechanisms remains an open question. In particular, transformer self-attention implements a global pairwise interaction pattern that differs from many theories of incremental language processing, which emphasize continuously updated latent states [5]. More colloquially, the argument against self-attention can be summarized as follows: Reading this sentence, you can easily summarize the meaning up to the very last word you have read, despite not being able to recall each exact preceding word. Indeed, cognitive models often posit that the brain maintains only a condensed representation of the current state of the world,

*These authors contributed equally to this work

which is updated whenever new input is received. Moreover, the way new input is integrated depends both on the input itself and the current representation. These differences do not imply that transformers are cognitively implausible, but they do highlight the need to test whether neuroscientific conclusions generalize across models with distinct computational assumptions.

Selective State Space Models (SSMs) offer one alternative architectural framework to self-attention based Transformers. These models update a latent state over time-steps, a mechanism that has long been appealing in neuroscience for characterizing continuous integration of information [6]. Specifically, Gu and Dao recently proposed Mamba, a SSM-based large language model which achieved comparative performance to transformer based models of the same scale without employing self-attention, thereby achieving sub-quadratic complexity [7]. Because SSMs implement explicit state updates, they provide an opportunity to test whether incremental processing yields different patterns of alignment with neural activity during language comprehension.

To date, few studies have compared different LLM architectures in terms of their utility for brain activity prediction. Bonnasse-Gahot et al. evaluated Mamba on fMRI data but, in line with prior work such as [2], extracted token-level embeddings for each word rather than sentence-level representations [3]. This approach limits the ability to study how models integrate information across multi-word contexts, making it difficult to assess phenomena such as polysemy or syntactic structure, where architectural differences may be most pronounced.

In this paper, we follow the experimental paradigm used in prior research that leveraged LLM embeddings to investigated the functionality of different brain regions (ROIs). We directly compare selective state-space and transformer models of various sizes using sentence-level representations and naturalistic fMRI data. Our goal is not to argue for the superiority of one architecture over the other as a cognitive model, but to assess how model choice influences neuroscientific conclusions. We show that different architectures produce systematically different patterns of brain alignment, particularly in analyses relying on perturbed inputs (e.g., scrambled word order). These discrepancies illustrate a broader methodological risk: results derived from a single LLM type may reflect properties of that model rather than properties of the human language system. Our findings therefore underscore the importance of validating neuroscientific inferences across diverse architectures to ensure replicability and to better understand which aspects of model-brain alignment are robust versus model-dependent.

2 Methods

Please find all the code and data necessary to replicate all the experiments discussed in this paper can be found on our GitHub repository[1].

[1] github.com/Arsu-Lab/Human-Language-Network-Modeling-with-Selective-State-Spaces

2.1 Models

Prior work primarily relied on variants of GPT-2, a transformer based LLM, to study the human language network [1, 2]. Here we investigate whether the conclusions of these studies hold when using non transformer-based LLMs by repeating these experiments with a broader set of LLMs, including state space models such as Mamba. All models were accessed via the Huggingface API. To examine how model size influences brain alignment, we evaluate three model sizes for both Mamba and GPT-2. For simplicity, we will refer to the model sizes as "small", "medium", and "large". The Mamba models had 130, 790 and 1400 million parameters each, while the corresponding GPT-2 models had 137, 774, and 1500 million parameters. Unfortunately, exact comparability between the architecture is limited by the absence of SSM-based LLMs that use the same tokanization and training data as GPT-2. Mamba instead follows the training recipe typical of open source models (such as LLaMa), including training on the Pile dataset which is a reconstruction of the propriety WebText data used to train the original GPT-2.

2.2 Neural Recording Dataset

We used the fMRI dataset collected by Wehbe et al., in which eight subjects read chapter nine of *Harry Potter and the Sorcerer's stone* [8]. fMRI trials were collected every 2 seconds, while words were projected into a screen every $500ms$, resulting in a four-word *sentence* aligning temporally with each fMRI trial.

Following Merlin et al., we clear punctuation and for each fMRI trail concatenate activations from the 5 preceding four word sentences (20 words total) to account for the delayed hemodynamic response of the fMRI signal [2]. For every trial we also obtain embeddings for either the original text, or a scrambled version of the text consisting of the same 20 words randomly reordered, yielding two sets of five sentences and two sets of five embeddings that are then concatenated. We use consecutive 10% of the fMRI trials for testing. To avoid data leakage due to the delayed hemodynamic response we use a buffer of 20s between the train and test segments.

To ensure that all voxels contributed equally during regression, fMRI responses were standardized across trials using `sklearn` library `StandardScaler`. Also, in keeping with [2] we grouped voxels into the following regions of interest (ROIs): Inferior Frontal Gyrus and Inferior Frontal Gyrus pars Orbitalis, Anterior Temporal Lobe, Posterior Temporal Lobe, Posterior Cingulate Cortex and Dorsomedial Prefrontal Cortex. Following standard practice, we also calculate the lower noise ceiling for each brain area separately (see online Appendix A).

2.3 Sentence Embeddings

To obtain model representations we input each complete *sentence* to the LLMs and extract the final token embedding as a stand-in of the sentence representation. In previous research given a sentence consisting on 4 words $s = w_1 w_2 w_3, w_4$

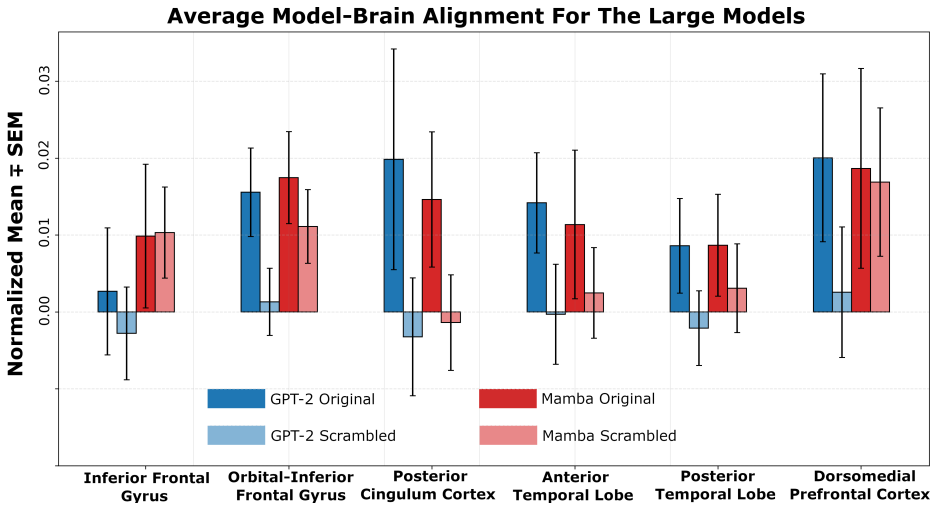

Fig. 1: Average Performance of the Large Mamba and GPT2 models for various human language network ROIs, for both original and scrambled input.

LLM embeddings were extracted for each single word $[e_{w_1}, e_{w_2}, e_{w_3}, e_{w_4}]$ and concatenated to generate a 'sentence embedding' [2, 3]. Such word-level concatenation introduces two key limitations: i) First, the model never processes the entire sentence as a unit, preventing syntactic and semantic interactions across words from being reflected in the representation. ii) word-level semantics are context-insensitive Homonymy and Polysemy could be assigned ambiguous - or completely incorrect - semantic meanings. For instance the word "bat" could refer to the mammal or a baseball bat. Like humans, LLMs rely on context to resolve these semantic uncertainness.

Instead we follow the pipeline outlined by Schrimpf et al. [1]. Namely, we provide the entire sentence $[w_1 w_2 w_3 w_4]$ as input and extract the last token representation which aggregates information from the full preceding context.

2.4 Calculating Brain-Model Alignment

Following common practice, we train a ridge regression to predict fMRI activity given the model embeddings, the Pearson correlation between the predicted and true fMRI activity on a held-out subset of the data (here 10%) is then calculated as a measure of brain-model alignment [2, 9].

3 Experiments and Results

We provided both the original and scrambled text inputs to the six models to obtain embeddings for evaluating brain alignment.

3.1 Impact of Model Architecture and Size on Brain Alignment

Overall, consistent with prior studies [3] we found that i) Mamba models showed higher brain-model alignment than GPT-2 models across all subjects and model sizes (see Figure 2). and ii) alignment increased consistently with model size for both architectures (see online Appendix B). Because the models differ in multiple ways (including training data, tokenization, and model design) it is impossible to attribute performance differences specifically to architecture. Nonetheless, the observed pattern suggests that the hypothesis that transformer self-attention inherently produces more brain-like representations [1, 4] may not be universally supported when evaluated across a broader set of models.

3.2 Unraveling Brain Functionality Using Input Perturbations

Rather than focusing on absolute alignment values, prior work has examined the difference in alignment between original and scrambled stimuli as a way to infer the functional role of particular brain regions [2]. Larger drops in alignment for scrambled input have been interpreted as evidence that a region is more sensitive to multi-word structure or higher-level semantic integration.

Our results, however, suggest that the conclusions of these experiments may depend on the model used. For instance, we found that the largest differences between original and scrambled inputs in Mamba appeared in the Posterior Cingulate Cortex and the Posterior / Anterior Temporal Lobe, whereas for GPT-2 models also had large effects in Inferior and Orbital-Inferior Frontal Gyrus regions. Although it can be argued that Mamba-based patterns loosely correspond to current proposals about discourse-level integration in the human language network[2] [10], the exact functional roles of these ROIs remain debated, and our comparisons are not sufficient to support strong neurocognitive claims. From the perspective of this work, the crucial observation is the architectural divergence in which ROIs show the strongest scrambling effects. These discrepancies indicate that using different LLMs can result in drawing different conclusions despite following the same experimental setup. Above all, this highlights a broader methodological implication: neuroscientific findings derived from a single model may not generalize, and validating results across diverse architectures is essential for ensuring replicability when using LLMs as cognitive models.

4 Discussion and Future Work

Our findings demonstrate that conclusions drawn from brain-model alignment can vary substantially depending on the underlying language model architecture. Differences in regional sensitivity to scrambled input, in particular, suggest that different models might emphasize distinct aspects of linguistic structure. This variability underscores a central point: neuro-scientific inferences based on a

[2]For example, the pCingulate has been linked to discourse-level processing, while the portions of the Frontal Inferior ROIs are more closely associated with sentence-level properties

single LLM type risk being model-dependent rather than reflecting robust properties of the brain. As LLMs become increasingly used as scientific tools, validating results across diverse architectures is essential for ensuring replicability and guarding against architecture-specific artifacts.

Model architecture aside, while we controlled for characteristics such as model size, there are multiple additional differences between the Mamba and GPT2 models compared here including training data, tokenization, and dimensionality. The field would benefit from a systematic evaluation of model families under more controlled conditions. Regardless, establishing best practices for cross-model validation will be crucial for developing reliable cognitive insights from LLM-based analyses.

References

[1] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

[2] Gabriele Merlin and Mariya Toneva. Language models and brains align due to more than next-word prediction and word-level information. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2024.

[3] Laurent Bonnasse-Gahot and Christophe Pallier. fMRI predictors based on language models of increasing complexity recover brain left lateralization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[4] Kalman Katlowitz, James Belanger, Taha Ismail, Ana Chavez, Assia Chericoni, Melissa Franch, Elizabeth Mickiewicz, Raissa Mathura, Danika Paulo, Eleonora Bartoli, Steven Piantadosi, Nicole Provenza, Andrew Watrous, Sameer Sheth, and Benjamin Hayden. Attention is all you need (in the brain): semantic contextualization in human hippocampus. *bioRxiv*, 2025.

[5] David J. Heeger. Theory of cortical function. *Proceedings of the National Academy of Sciences*, 114(8):1773–1782, 2017.

[6] David Zoltowski, Jonathan Pillow, and Scott Linderman. A general recurrent state space framework for modeling neural dynamics during decision-making. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[7] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

[8] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLOS ONE*, 9, 11 2014.

[9] Manshan Guo, Michael Samjatin, Bhavin Choksi, Sari Sadiya, Radoslaw Cichy, and Gemma Roig. Predictive coding dynamics enhance model-brain similarity. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 2025.

[10] Evelyn C. Ferstl, Jane Neumann, Carsten Bogler, and D. Yves von Cramon. The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, 29(5):581–593, 2008.

A Appendix: Noise Ceiling Calculation

The lower noise ceiling was computed using a leave-one-subject-out reliability approach applied in a reduced component space. For each ROI, each subject's voxel $\times$ time response matrix was first concatenated (across selected ROIs if necessary) and decomposed with SVD, keeping the top 10 principal components that capture the dominant shared temporal structure across voxels. This yields a set of component time-courses for each subject. To estimate how reliably the responses of each subject can be predicted from the rest of the group, we held out one subject at a time and averaged the component representations of all other subjects, weighting each subject by its voxel count. The component pattern of the held-out subject was then correlated with this weighted group mean. Repeating this for all subjects produced one reliability value per subject; the lower noise ceiling for the ROI is the mean of these leave-one-out correlations. This value reflects the maximum performance a model could achieve given the subject-to-subject consistency in the measured neural responses.

B Appendix: All Models

Comparison for "large", "medium", and "small". The Mamba models had 1400, 790 and 130 million parameters each, while the corresponding GPT-2 models had 1500, 774, and 137 million parameters.

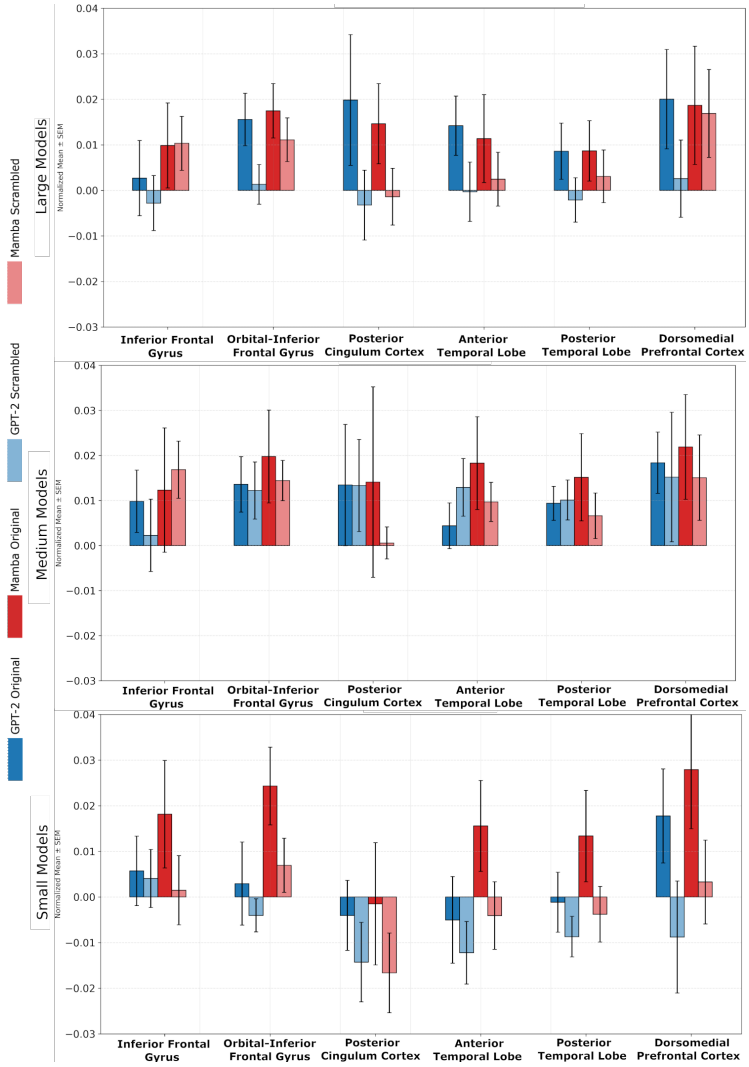

Fig. 2: Average Performance of large, medium, and small Mamba and GPT2 models for various ROIs, for both original and scrambled input.