

Can LLMs predict human neural activations?

Archaeology of Intelligent Machines

2025-2026

Rădulescu Horia Filip

horia-filip.radulescu@s.unibuc.ro

Zidăroiu Maria

maria.zidaroiu@s.unibuc.ro

1 Introduction

Recent advances in large language models enable not only text generation but also the prediction of human neural activation during language processing. Studies show that the internal representations of these models correlate significantly with brain responses measured via fMRI or electrocorticography (Schrimpf et al., 2021). Our project aims to test the extent to which LLMs can predict human neural activation and to identify the models and strategies that best align with brain language processing.

Motivation

Understanding how LLMs relate to human brain activity can provide insights into both the mechanisms of language processing in the brain and the development of more cognitively aligned artificial intelligence.

Contributions

• Maria:

- Set up and maintained the experimental environment, including installation and configuration of the Brain-Score framework and its dependencies.
- Implemented the experimental pipelines and **executed the majority of code runs**, including long-running benchmark evaluations on Google Colab and local systems.
- Ran baseline experiments with GPT-2 and BERT on the Pereira2018 dataset to validate correctness of the evaluation setup.
- Managed computational constraints such as limited GPU memory, long execution times (up to one hour per run), and session interruptions.

- Authored and edited the project report.

- Gained practical experience in running large-scale neural benchmarks, including an improved understanding of the relationship between model size, layer selection, and neural predictivity.
- In future work, would like to further explore efficient evaluation methods and memory-optimized techniques for running large language models on limited hardware.

• Horia:

- Designed the experimental protocol, including benchmark selection and model comparison strategy.
- Coordinated and **supervised experimental runs**, including result verification and consistency checks across repeated executions.
- Analyzed and interpreted the results obtained from multiple benchmarks (Pereira2018, Blank2014, Fedorenko2016).
- Assisted in debugging execution errors and refining code for stable and reproducible runs.
- Integrated and executed large language models from Hugging Face (e.g., Gemma, Pythia, DistilGPT-2, SmolLM) within the Brain-Score framework.
- Authored and edited the project report.

- Developed a deeper understanding of how different neuroimaging benchmarks capture complementary aspects of language processing in the brain.
- In future work, would like to investigate finer-grained analyses of layer-wise representations and extend evaluations to multilingual or multimodal language models.

Previous work

Schrimpf(2021) shows that transformer models (e.g., GPT-2, BERT) can predict most of the explainable variance in neural responses to sentences, with higher next-word prediction performance linked to higher brain scores. |

<https://pmc.ncbi.nlm.nih.gov/articles/PMC8694052/>

Toneva(2022) demonstrates that LLM embeddings predict both neural activity and how well subjects understand a story. |

<https://pmc.ncbi.nlm.nih.gov/articles/PMC9522791/>

Hewitt(2023) finds that predictive performance in fMRI scales with model size, suggesting larger models better capture brain activity. |

<https://pubmed.ncbi.nlm.nih.gov/39035676/>

Schrimpf(2024) shows that internal hierarchies of LLMs converge with the brain’s language processing hierarchy, with more capable models producing increasingly brain-like representations.

<https://arxiv.org/abs/2401.17671>

2 Approach

Link

The code and resources for this project are publicly available on GitHub:

<https://github.com/mariaz22/Can-LLMs-predict-human-neural-activation>

Environment Setup and Initial Validation

We recreated the legacy environment required to run neural-nlp benchmarks from the 2018–2020 codebase. A dedicated Conda environment with Python 3.8 allowed installation of older PyTorch and Brain-Score packages that are no longer available via standard repositories. Several dependencies, including `protobuf`, `boto3`, and `numpy`, had to be manually adjusted or downgraded. After these steps, GPT-2 activation extraction, CKA computations, and the Pereira2018 benchmark ran successfully, validating the reproducibility of the environment. After the setup, we validated the environment by running key components: GPT-2 activation extraction, CKA computations, and the Pereira2018 benchmark. The benchmark produced a score of 0.8159, confirming that:

- the benchmark runs correctly,
- GPT-2 generates valid neural activations,
- the entire neural-nlp + Brain-Score setup is fully functional.

Data

We conducted our experiments using several fMRI datasets from the Brain-Score Language benchmark suite, enabling a systematic comparison between artificial language models and human neural responses.

We first evaluated **GPT-2** and **BERT** on the **Pereira2018** dataset, which consists of neural responses recorded while participants processed isolated sentences and short passages. This dataset provides aligned linguistic stimuli and brain activation measurements, making it a standard benchmark for assessing model–brain correspondence. Building on these baselines, we subsequently evaluated **Gemma-2B** on the same dataset to examine whether newer transformer architectures exhibit improved neural alignment. To assess generalization across experimental paradigms, we extended our analysis to the **Blank2014** dataset, which captures brain responses during continuous naturalistic language comprehension.

On this benchmark, we evaluated **Pythia**, **SmolLM**, **DistilGPT-2**, and **Gemma-2B**, allowing comparison across model scales and training regimes. Finally, all previously evaluated models were tested on the **Fedorenko2016** dataset, which focuses on functionally localized language regions. This final evaluation enabled a direct comparison of model–brain alignment across multiple datasets and linguistic conditions.

Model and Computational Requirements

We used the standard GPT-2 model via the neural-nlp library. The model is small enough to run efficiently on CPU, without GPU. Experiments were run on Ubuntu WSL2 with an Intel Core i7 . Benchmarking Pereira2018 required 10–20 seconds up to 2–4 minutes, using 15–30% CPU and 1–2 GB RAM.

For subsequent experiments involving larger and more recent transformer models, including **Gemma-2B**, **Pythia**, **SmolLM**, and **DistilGPT-2**, we relied on **Google Colab** and **Kaggle** to access GPU-enabled environments. These models were loaded from the Hugging Face Model Hub, while the corresponding neural benchmarks (Pereira2018, Blank2014, and Fedorenko2016) were retrieved through the Brain-Score Language framework. Due to the increased model size and the computational demands of cross-validated neural regression, each benchmark evaluation required

substantially more time. On average, a single model–dataset evaluation took between **50 minutes and over one hour**, with some runs exceeding this duration depending on GPU availability and memory constraints.

These experiments involved repeated cross-validation, extraction of hidden representations from specific transformer layers, and linear mapping to voxel- or region-level neural responses.

Model Evaluation and Comparison

To evaluate the alignment between large language models (LLMs) and human neural activity, we used metrics and benchmarks provided by the neural-nlp and Brain-Score frameworks. The key evaluation methods include:

- **Brain-Score metrics:** Quantifies how well a model’s activations predict neural responses measured via fMRI or ECoG.
- **Centered Kernel Alignment (CKA):** Measures the similarity between model representations and neural activation patterns across layers.
- **Pearson correlation:** Assesses linear correlation between predicted and observed neural responses.
- **Benchmark comparisons:** Models (e.g., GPT-2) are compared on multiple datasets such as Pereira2018, Blank2014, and Fedorenko2016 to determine which architectures and layers align best with human brain activity.

These evaluation methods allow us to systematically compare different LLMs and identify which models, layers, and training objectives best capture aspects of human language processing.

Results and Evaluation

We evaluate several transformer-based language models on three established brain–language benchmarks: Blank2014, Fedorenko2016, and Pereira2018. For each benchmark, we report normalized Brain-Score values, raw Pearson correlations, and the corresponding noise ceilings. Normalized scores allow comparison across benchmarks with different signal-to-noise characteristics. For each model, we additionally report the layer achieving the best alignment with neural data.

Table 1: Results on the Pereira2018 benchmark

Model	Score	Raw	Ceiling
GPT-2	0.816	0.260	0.319
Gemma-2B	0.973	0.344	0.354

Table 2: Results on the Blank2014 benchmark

Model	Layer	Score	Raw	Ceiling
SmolLM2-135M	L8	0.252	0.053	0.210
Pythia-70M	L2	0.472	0.099	0.210
DistilGPT2	Last	0.363	0.076	0.210
Gemma-2B	L2	0.294	0.062	0.210

Table 3: Results on the Fedorenko2016 benchmark

Model	Layer	Score	Raw	Ceiling
SmolLM2-135M	L8	0.195	0.044	0.225
Pythia-70M	Last	0.554	0.125	0.225
DistilGPT2	Last	0.554	0.125	0.225
Gemma-2B	Last	0.982	0.221	0.225

3 Limitations

This study has several limitations that should be considered when interpreting the results. First, computational constraints significantly affected the experimental setup. While smaller models such as GPT-2 and DistilGPT-2 could be evaluated efficiently on CPU, larger models (e.g., Gemma-2B and Pythia) required GPU resources and long execution times, often exceeding 50 minutes per benchmark run. Memory limitations further restricted extensive layer-wise analyses and repeated evaluations.

Second, the experiments relied exclusively on established Brain-Score benchmarks (Pereira2018, Blank2014, and Fedorenko2016), which differ in experimental design and evaluation methodology. Consequently, performance differences across benchmarks may reflect dataset-specific properties rather than general model capabilities.

Finally, all benchmarks consist of English-language stimuli, limiting the generalizability of the findings to other languages. Moreover, the use of linear mappings between model activations and neural responses may not fully capture more complex nonlinear correspondences between artificial and biological representations.

4 Conclusions and Future Work

In this project, we successfully explored the alignment between artificial language models and human neural responses using the Brain-Score framework. We began by validating our experimental setup with GPT-2 on the Pereira2018 dataset and gradually extended our analysis to larger and more recent models, including Gemma, Pythia, SmolLM, and DistilGPT-2, across multiple benchmarks (Pereira2018, Blank2014, and Fedorenko2016). This progression allowed us to gain practical in-

sight into both the methodological pipeline and the computational challenges involved in neural benchmarking of large language models.

In retrospect, one aspect that could have been handled differently is the computational setup. While smaller models were easy to evaluate locally, larger models required substantial GPU resources and long execution times, often making experimentation cumbersome. A more streamlined environment or earlier access to stable GPU resources could have improved productivity and reduced interruptions caused by memory and dependency issues.

Despite these challenges, the project was valuable and educational. We gained hands-on experience with the Brain-Score ecosystem, Hugging Face model integration, and the practical trade-offs between model size, computational cost, and neuroscientific interpretability. Importantly, we learned that higher computational complexity does not always translate into proportionally better alignment with neural data, highlighting the importance of careful benchmark selection and analysis.

For future work, this project could be extended by exploring additional brain datasets, non-English stimuli, or alternative mapping methods beyond linear regression. Incorporating more efficient evaluation strategies or model compression techniques could also make large-scale comparisons more feasible.

Overall, while the project was technically demanding at times, it provided a meaningful and realistic perspective on interdisciplinary research at the intersection of neuroscience and machine learning.