

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

001 1 Introduction

002 Recent advances in large language models (LLMs)
003 enable not only text generation but also the prediction
004 of human neural activation during language
005 processing. Studies show that the internal representations
006 of these models correlate significantly
007 with brain responses measured via fMRI or electro-
008 corticography (Schrimpf et al., 2021). Our project
009 aims to test the extent to which LLMs can predict
010 human neural activation and to identify the models
011 and strategies that best align with brain language
012 processing.

013 Motivation

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

019 Previous work

020 Schrimpf(2021) shows that transformer models
021 (e.g., GPT-2, BERT) can predict most of the
022 explainable variance in neural responses to
023 sentences, with higher next-word prediction performance
024 linked to higher brain scores. |
<https://pmc.ncbi.nlm.nih.gov/articles/PMC8694052/>

025 Toneva(2022) demonstrates that LLM embeddings
026 predict both neural activity and how well
027 subjects understand a story. |
<https://pmc.ncbi.nlm.nih.gov/articles/PMC9522791/>

028 Hewitt(2023) finds that predictive performance
029 in fMRI scales with model size, suggesting larger
030 models better capture brain activity. |
<https://pubmed.ncbi.nlm.nih.gov/39035676/>

031 Schrimpf(2024) shows that internal hierarchies
032 of LLMs converge with the brain's language processing
033 hierarchy, with more capable models producing
034 increasingly brain-like representations.
<https://arxiv.org/abs/2401.17671>

041 2 Approach

Link

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

[https://github.com/mariaz22/](https://github.com/mariaz22/Can-LLMs-predict-human-neural-activation)

[Can-LLMs-predict-human-neural-activation](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

For our experiments, we used the Pereira2018 fMRI dataset, which contains neural responses from human subjects listening to multiple stories and sentences. The dataset provides both the textual stimuli and the corresponding brain activation measurements, enabling the evaluation of GPT-2 activations against neural responses.

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.

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 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.

3 Conclusions and Future Work

So far, we have successfully set up the neural-nlp environment and validated it using GPT-2 on the Pereira2018 dataset. As future work, we plan to extend our experiments to additional large language models and neuroimaging datasets. This will allow us to systematically compare model architectures, training objectives, and their alignment with human neural activity, providing a broader understanding of how different LLMs capture aspects of language processing in the brain.