

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 enable
003 not only text generation but also the prediction of
004 human neural activation during language process-
005 ing. Studies show that the internal representations
006 of these models correlate significantly with brain
007 responses measured via fMRI or electrocorticogra-
008 phy (Schrimpf et al., 2021). Our project aims
009 to test the extent to which LLMs can predict hu-
010 man 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 mecha-
016 nisms of language processing in the brain and the
017 development of more cognitively aligned artificial
018 intelligence.

019 Contributions

020 • Maria:

- 021 – Set up and maintained the experimental
022 environment, including installation and
023 configuration of the Brain-Score frame-
024 work and its dependencies.
- 025 – Implemented the experimental pipelines
026 and **executed the majority of code runs**,
027 including long-running benchmark eval-
028 uations on Google Colab and local sys-
029 tems.
- 030 – Ran baseline experiments with GPT-2
031 and BERT on the Pereira2018 dataset
032 to validate correctness of the evaluation
033 setup.
- 034 – Managed computational constraints such
035 as limited GPU memory, long execution
036 times (up to one hour per run), and ses-
037 sion interruptions.

- 038 – Authored and edited the project report.

- 039 • Gained practical experience in running large-
040 scale neural benchmarks, including an im-
041 proved understanding of the relationship be-
042 tween model size, layer selection, and neural
043 predictivity.
- 044 • In future work, would like to further explore
045 efficient evaluation methods and memory-
046 optimized techniques for running large lan-
047 guage models on limited hardware.

048 • Horia:

- 049 – Designed the experimental protocol, in-
050 cluding benchmark selection and model
051 comparison strategy.
- 052 – Coordinated and **supervised experimen-**
053 **tal runs**, including result verification and
054 consistency checks across repeated exec-
055 utions.
- 056 – Analyzed and interpreted the re-
057 sults obtained from multiple bench-
058 marks (Pereira2018, Blank2014,
059 Fedorenko2016).
- 060 – Assisted in debugging execution errors
061 and refining code for stable and repro-
062 ducible runs.
- 063 – Integrated and executed large lan-
064 guage models from Hugging Face (e.g.,
065 Gemma, Pythia, DistilGPT-2, SmolLM)
066 within the Brain-Score framework.
- 067 – Authored and edited the project report.

- 068 • Developed a deeper understanding of how dif-
069 ferent neuroimaging benchmarks capture com-
070 plementary aspects of language processing in
071 the brain.
- 072 • In future work, would like to investigate finer-
073 grained analyses of layer-wise representations
074 and extend evaluations to multilingual or mul-
075 timodal language models.

076	Previous work	125
077	Schrimpf(2021) shows that transformer mod-	126
078	els (e.g., GPT-2, BERT) can predict most of the	127
079	explainable variance in neural responses to sen-	128
080	tences, with higher next-word prediction perfor-	129
081	mance linked to higher brain scores.	130
082		
083	https://pmc.ncbi.nlm.nih.gov/articles/PMC8694052/	
084	Toneva(2022) demonstrates that LLM embed-	131
085	dings predict both neural activity and how well	132
086	subjects understand a story.	133
087		134
088	https://pmc.ncbi.nlm.nih.gov/articles/PMC9522791/	
089	Hewitt(2023) finds that predictive performance	135
090	in fMRI scales with model size, suggesting larger	136
091	models better capture brain activity.	137
092		138
093	https://pubmed.ncbi.nlm.nih.gov/39035676/	
094	Schrimpf(2024) shows that internal hierarchies	139
095	of LLMs converge with the brain’s language pro-	140
096	cessing hierarchy, with more capable models pro-	141
097	ducing increasingly brain-like representations.	142
098		143
099	https://arxiv.org/abs/2401.17671	
100		144
101		145
102	2 Approach	
103		
104	Link	
105	The code and resources for this project are pub-	
106	licly available on GitHub:	
107		
108	https://github.com/mariaz22/	
109	Can-LLMs-predict-human-neural-activations	
110		
111		
112		
113	Environment Setup and Initial Validation	
114	We recreated the legacy environment required	
115	to run neural-nlp benchmarks from the 2018–2020	
116	codebase. A dedicated Conda environment with	
117	Python 3.8 allowed installation of older PyTorch	
118	and Brain-Score packages that are no longer avail-	
119	able via standard repositories. Several dependen-	
120	cies, including protobuf, boto3, and numpy,	
121	had to be manually adjusted or downgraded. Af-	
122	ter these steps, GPT-2 activation extraction, CKA	
123	computations, and the Pereira2018 benchmark ran	
124	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 compar-	
	ison between artificial language models and human	
	neural responses.	
	We first evaluated GPT-2 and BERT on the	
	Pereira2018 dataset, which consists of neural re-	
	sponses recorded while participants processed iso-	
	lated sentences and short passages. This dataset	
	provides aligned linguistic stimuli and brain acti-	
	vation measurements, making it a standard bench-	
	mark for assessing model–brain correspondence.	
	Building on these baselines, we subsequently eval-	
	uated Gemma-2B on the same dataset to examine	
	whether newer transformer architectures exhibit	
	improved neural alignment. To assess generaliza-	
	tion across experimental paradigms, we extended	
	our analysis to the Blank2014 dataset, which cap-	
	tures brain responses during continuous naturalistic	
	language comprehension.	
	On this benchmark, we evaluated Pythia ,	
	SmolLM , DistilGPT-2 , and Gemma-2B , allow-	
	ing 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 sec-	
	onds 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 mod-	
	els 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 neu-	
	ral regression, each benchmark evaluation required	

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

180 These experiments involved repeated cross-
 181 validation, extraction of hidden representations
 182 from specific transformer layers, and linear map-
 183 ping to voxel- or region-level neural responses.

184 Model Evaluation and Comparison

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

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

204 These evaluation methods allow us to systemati-
 205 cally compare different LLMs and identify which
 206 models, layers, and training objectives best capture
 207 aspects of human language processing.

208 Results and Evaluation

209 We evaluate several transformer-based lan-
 210 guage models on three established brain–language
 211 benchmarks: Blank2014, Fedorenko2016, and
 212 Pereira2018. For each benchmark, we report nor-
 213 malized Brain-Score values, raw Pearson corre-
 214 lations, and the corresponding noise ceilings.
 215 Normalized scores allow comparison across bench-
 216 marks with different signal-to-noise characteristics.
 217 For each model, we additionally report the layer
 218 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
SmoILM2-135M	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
SmoILM2-135M	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, SmoILM, and DistilGPT-2, across multiple benchmarks (Pereira2018, Blank2014, and Fedorenko2016). This progression allowed us to gain practical in-

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

259 In retrospect, one aspect that could have been
260 handled differently is the computational setup.
261 While smaller models were easy to evaluate lo-
262 cally, larger models required substantial GPU re-
263 sources and long execution times, often making
264 experimentation cumbersome. A more streamlined
265 environment or earlier access to stable GPU re-
266 sources could have improved productivity and re-
267 duced interruptions caused by memory and depen-
268 dency issues.

269 Despite these challenges, the project was valua-
270 ble and educational. We gained hands-on expe-
271 rience with the Brain-Score ecosystem, Hugging
272 Face model integration, and the practical trade-offs
273 between model size, computational cost, and neuro-
274 scientific interpretability. Importantly, we learned
275 that higher computational complexity does not al-
276 ways translate into proportionally better alignment
277 with neural data, highlighting the importance of
278 careful benchmark selection and analysis.

279 For future work, this project could be extended
280 by exploring additional brain datasets, non-English
281 stimuli, or alternative mapping methods beyond
282 linear regression. Incorporating more efficient eval-
283 uation strategies or model compression techniques
284 could also make large-scale comparisons more fea-
285 sible.

286 Overall, while the project was technically de-
287 manding at times, it provided a meaningful and
288 realistic perspective on interdisciplinary research
289 at the intersection of neuroscience and machine
290 learning.