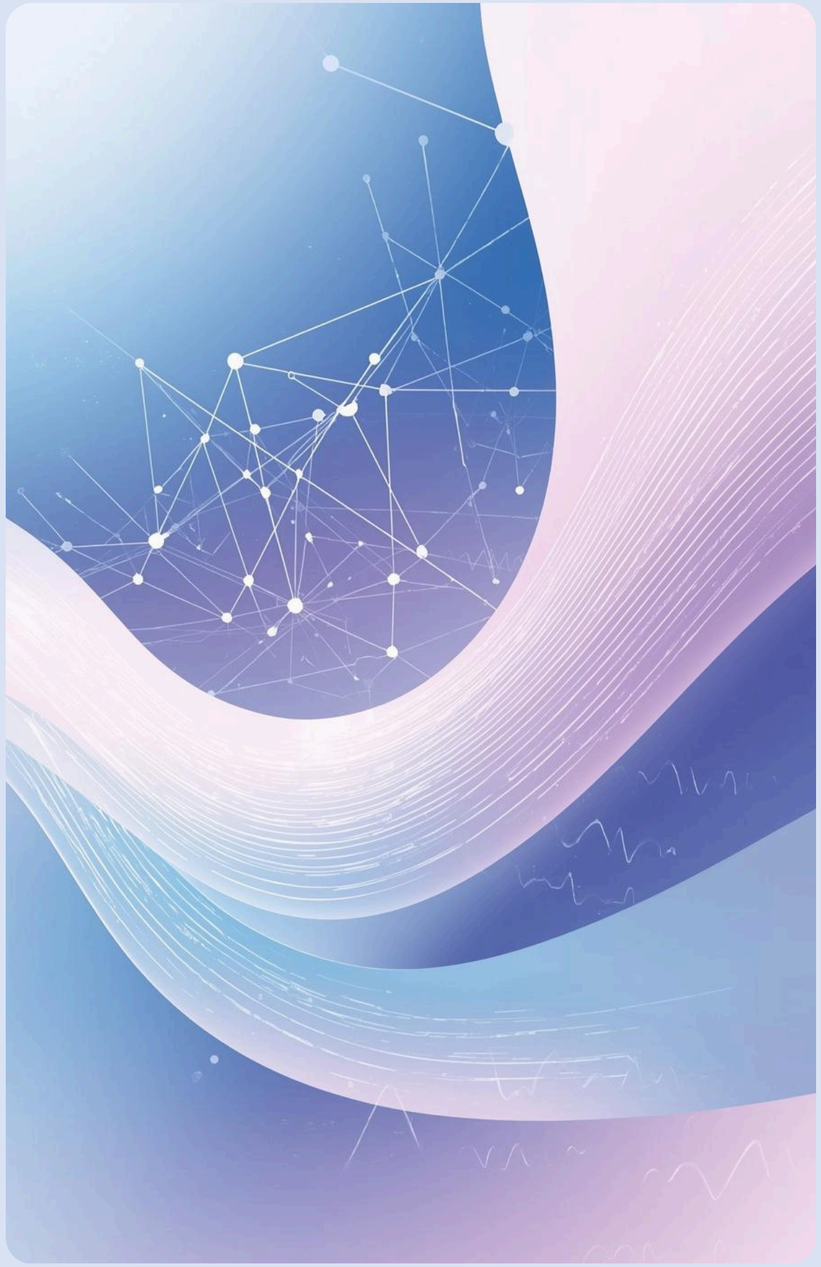
LLMs and Neural Activity

Radulescu Horia Filip 334
Zidaroiu Maria 334

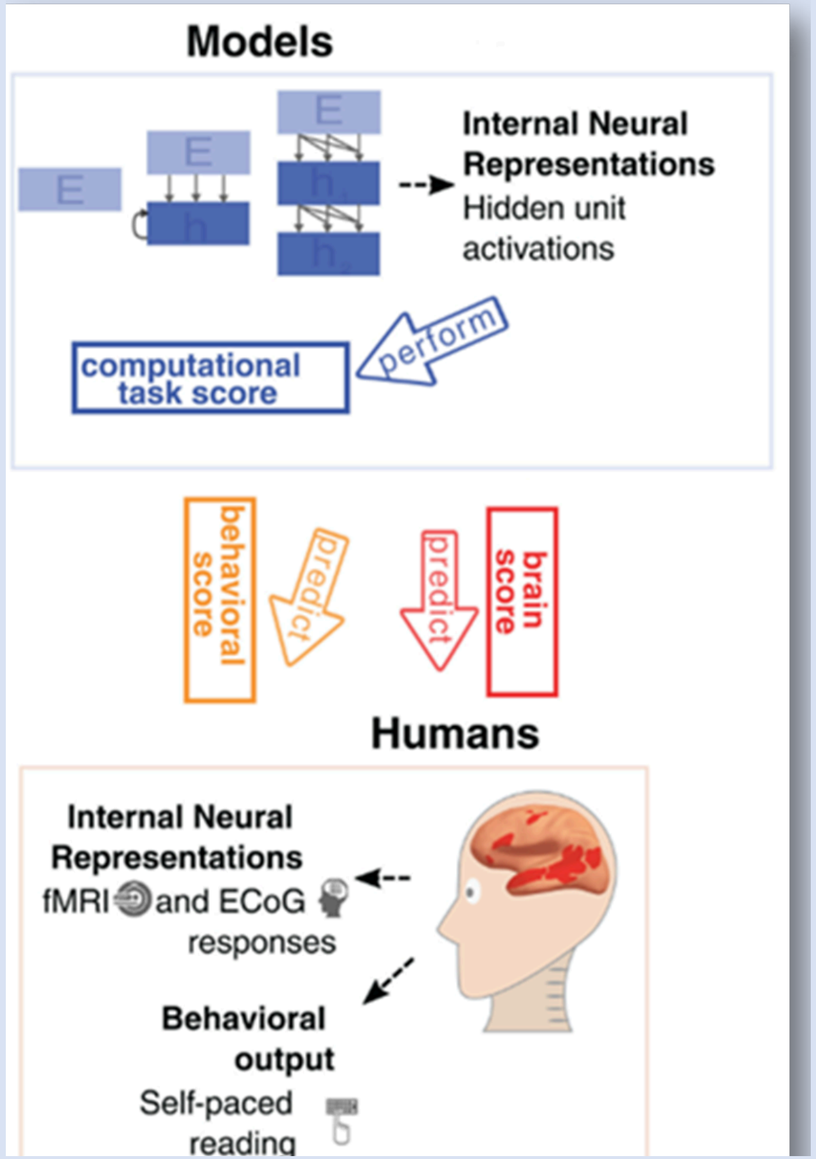


Introduction

Recent advances in large language models (LLMs) enable not only text generation but also the prediction of human brain activity.

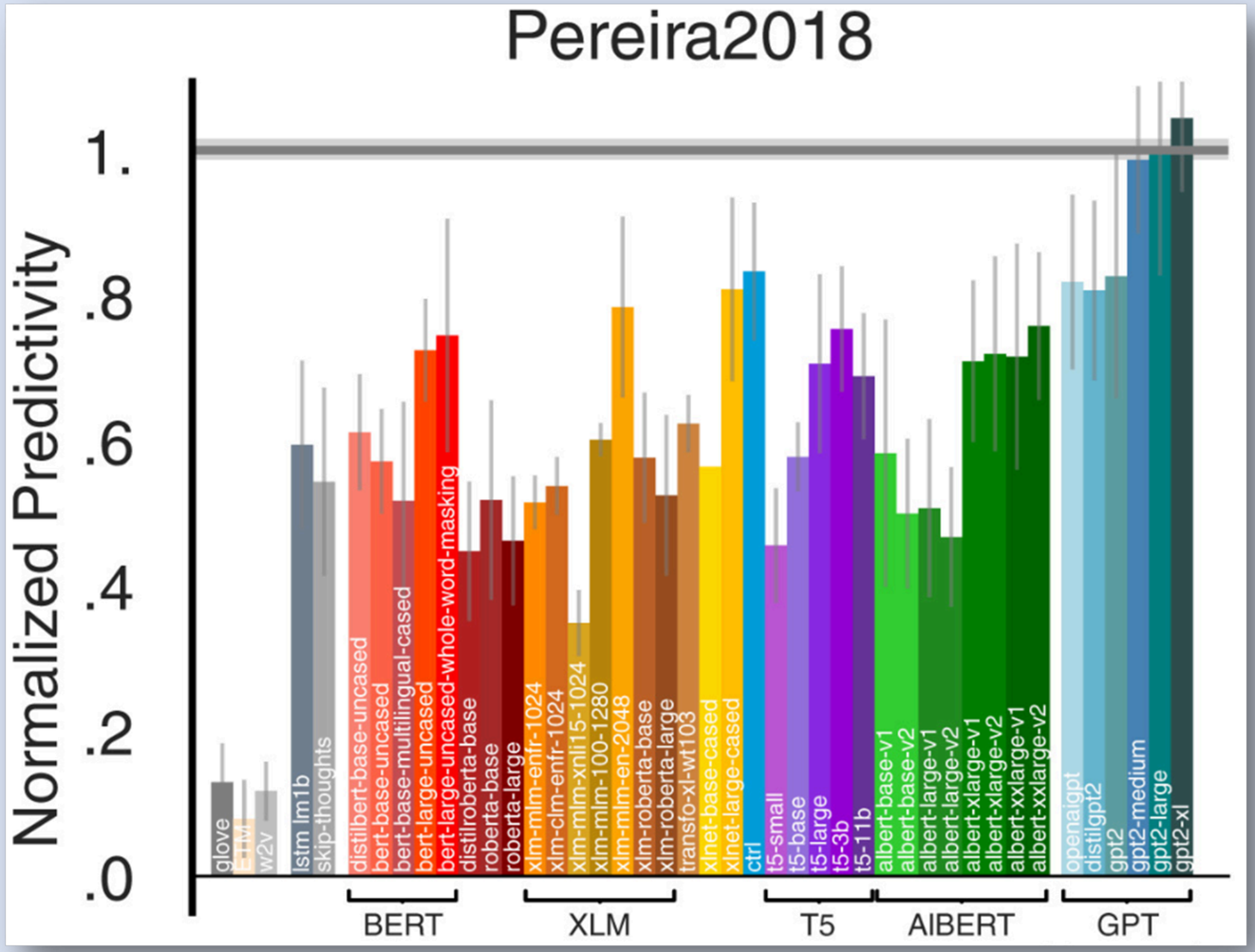
Our project aims to test the extent to which LLMs can predict human neural activation and to identify the models and strategies that best align with brain language processing.

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.



Previous Work

- Transformer models like GPT-2 and BERT can predict neural responses to sentences (Schrimpf et al., 2021).
- LLM embeddings also capture how well humans understand stories (Toneva et al., 2022).
- Larger models tend to align better with brain activity (Hewitt et al., 2023).
- Model hierarchies reflect the brain's language processing hierarchy (Schrimpf et al., 2024).



Environment Setup and Validation

```
Benchmark score:
<xarray.Score (aggregation: 2)>
array([0.8158978 , 0.15370946])
Coordinates:
  * aggregation  (aggregation) <U6 'center' 'error'
Attributes:
    raw:          <xarray.Score (aggregation: 2)>\narray([0.25991808, 0.04896...
    ceiling:      <xarray.Score (aggregation: 3)>\narray([0.31856696, 0.01295...
    description:  ceiling-normalized score
(neural-nlp) horiaphilip@DESKTOP-LUPCRT8:~/neural-nlp/neural_nlp/models$
```

Computational setup

Experiments were run using a hybrid compute setup:

- Local CPU execution for smaller models and rapid prototyping
- Cloud GPUs (Google Colab and Kaggle) for larger models

- **Setup**: Recreated legacy environment for neural-nlp benchmarks (2018–2020 codebase).
- **Tools**: Conda environment with Python 3.8, older PyTorch, Brain-Score, and manual dependency adjustments.
- **Validation**:
  GPT-2 activation extraction
  CKA computations
  Pereira2018 benchmark

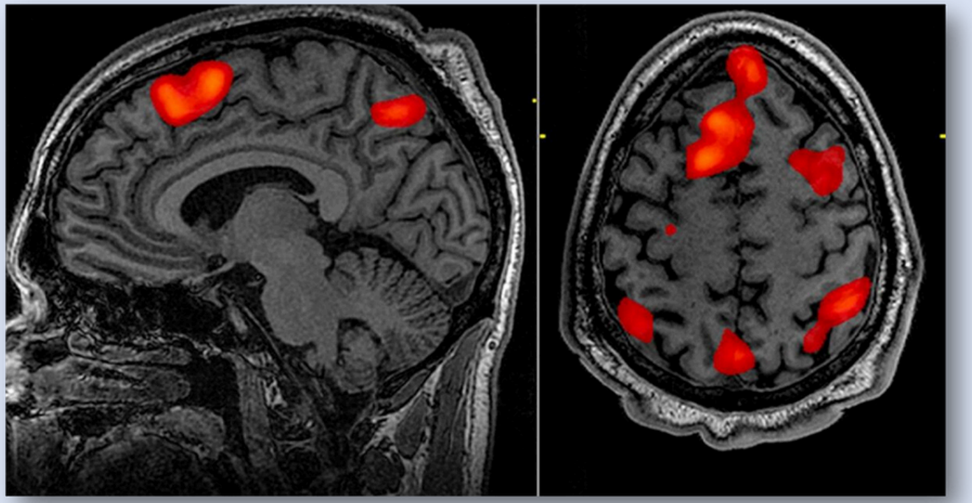
Datasets / Benchmarks

fMRI (functional Magnetic Resonance Imaging)
Neuroimaging method that measures brain activity indirectly via blood-oxygen (BOLD) responses during language processing.

Blank2014
Naturalistic speech fMRI benchmark probing early / surface-level language representations.

Fedorenko2016
Controlled fMRI benchmark targeting the core language network and abstract linguistic processing.

Pereira2018
Sentence-level fMRI benchmark focused on semantic representations with high signal-to-noise.



EDA

Benchmark Difficulty Spectrum

Each benchmark probes a different aspect of language processing in the brain, ranging from language network localization (Blank2014) and language selectivity (Fedorenko2016) to whole-brain sentence encoding (Pereira2018).

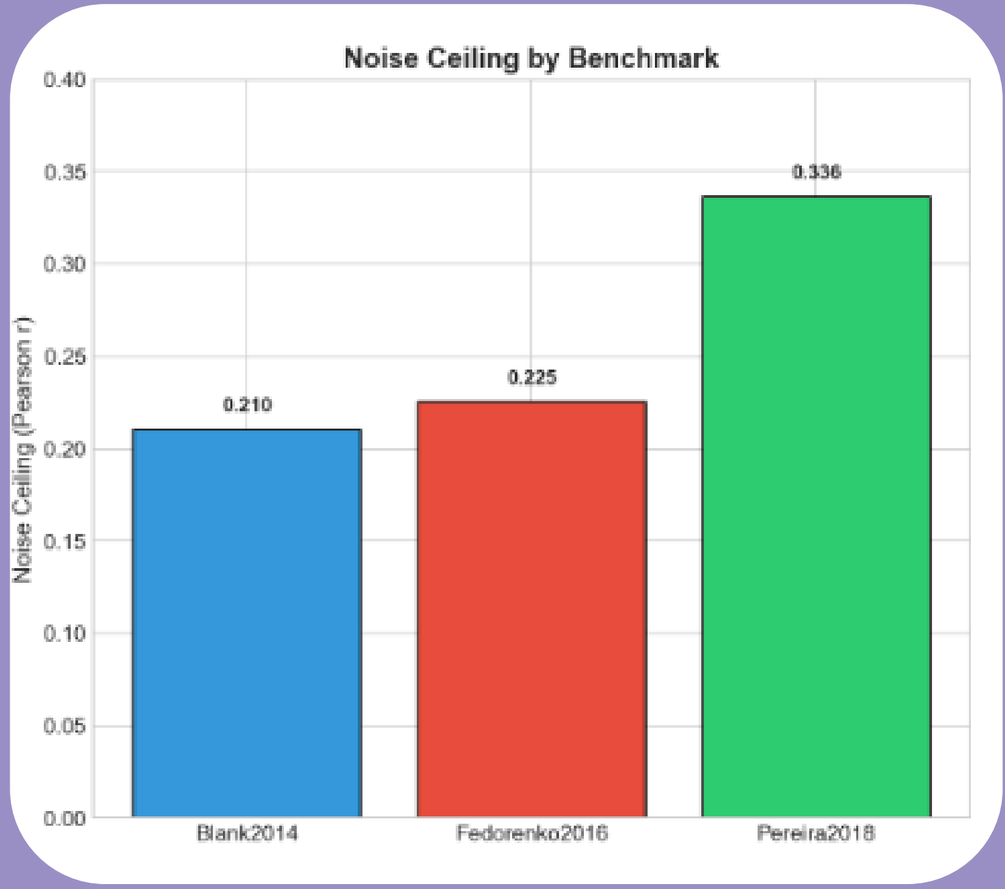
Benchmark	Stimuli	Brain Measure
Blank2014	Sentences vs non-words	fROI
Fedorenko2016	Sentences, word lists	fROI
Pereira2018	627 natural sentences	Whole-brain fMRI

Together, they provide complementary perspectives on how well language models align with human neural responses.

EDA

Noise Ceiling Across Benchmarks

- The noise ceiling reflects the maximum explainable variance in neural data due to measurement noise
- Pereira2018 exhibits the highest ceiling, indicating more reliable and consistent neural responses
- Blank2014 and Fedorenko2016 have lower ceilings, reflecting higher task difficulty and fROI variability
- These differences motivate the use of normalized scores when comparing model performance across benchmarks



Models

Pythia

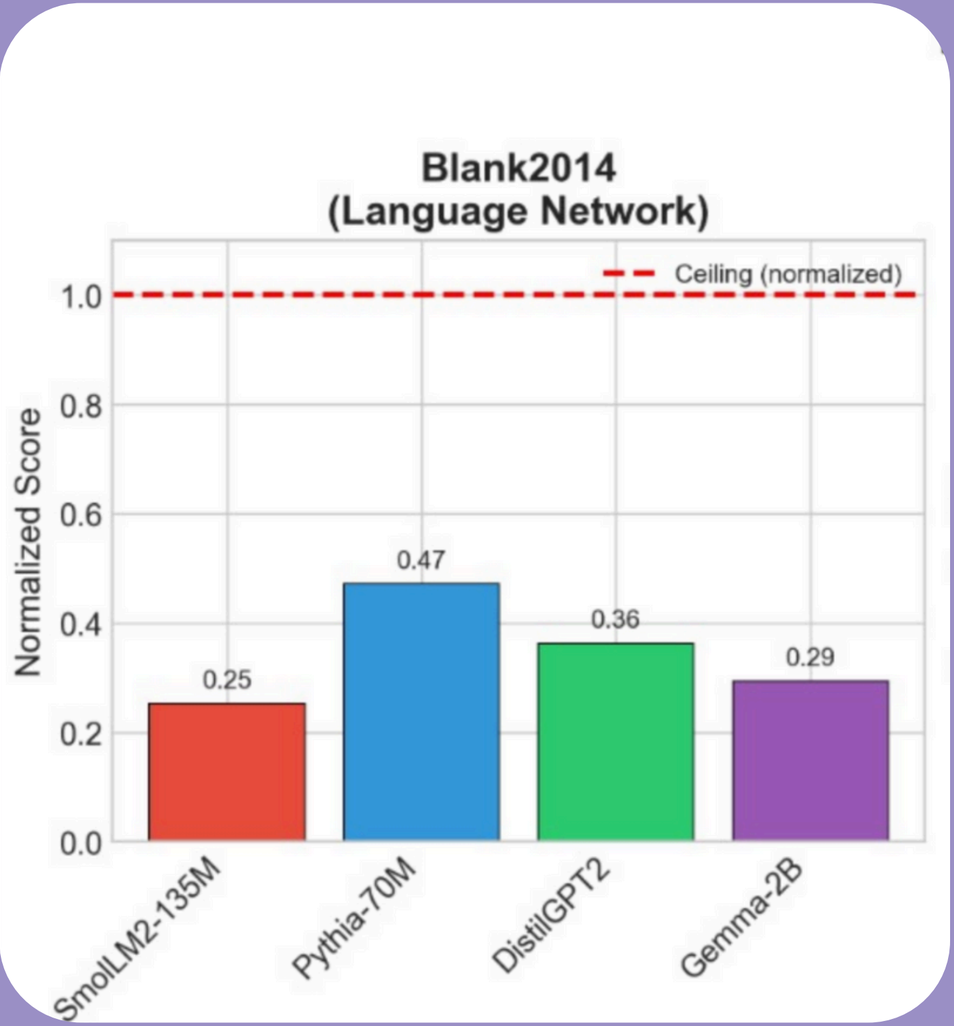
- GPT-2 – best baseline model used in the original paper, retained for direct comparison.
- DistilGPT-2 – compressed and faster version of GPT-2, testing efficiency vs. performance.
- Pythia-70M – small, well-controlled transformer for analyzing layer-wise effects.
- SmolLM2-135M – lightweight, CPU-friendly model representing small-scale LMs.
- Gemma-2B – modern large-scale model, representing high-capacity neural language models.

All Models on Blank2014

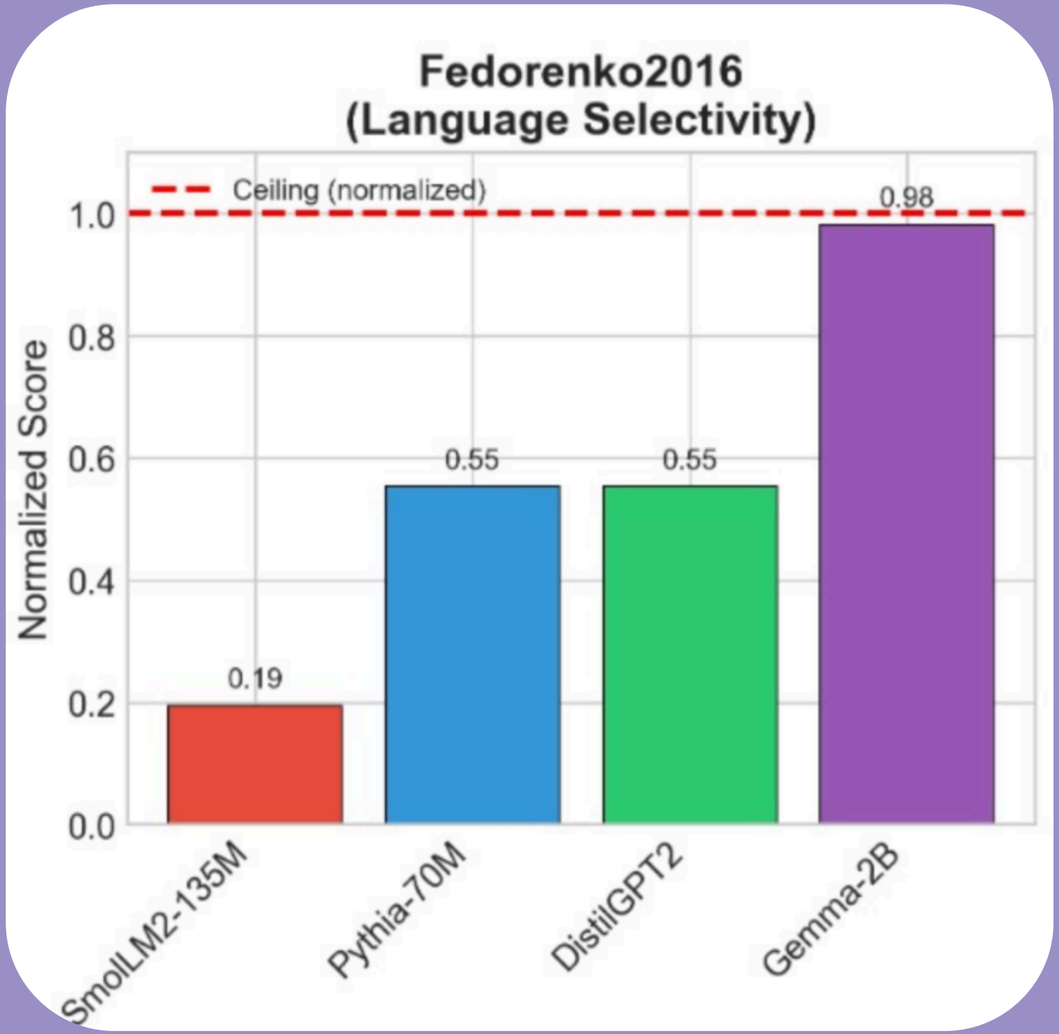
We compare several language models on Blank2014, a benchmark targeting sentence-level language processing.

Results show clear performance differences across models, with more recent architectures outperforming smaller or older ones.

The benchmark favors models with stronger linguistic representations, even when model size remains relatively small.



All Models on Fedorenko2016



We evaluate the same set of models on Fedorenko2016, which measures higher-level language comprehension.

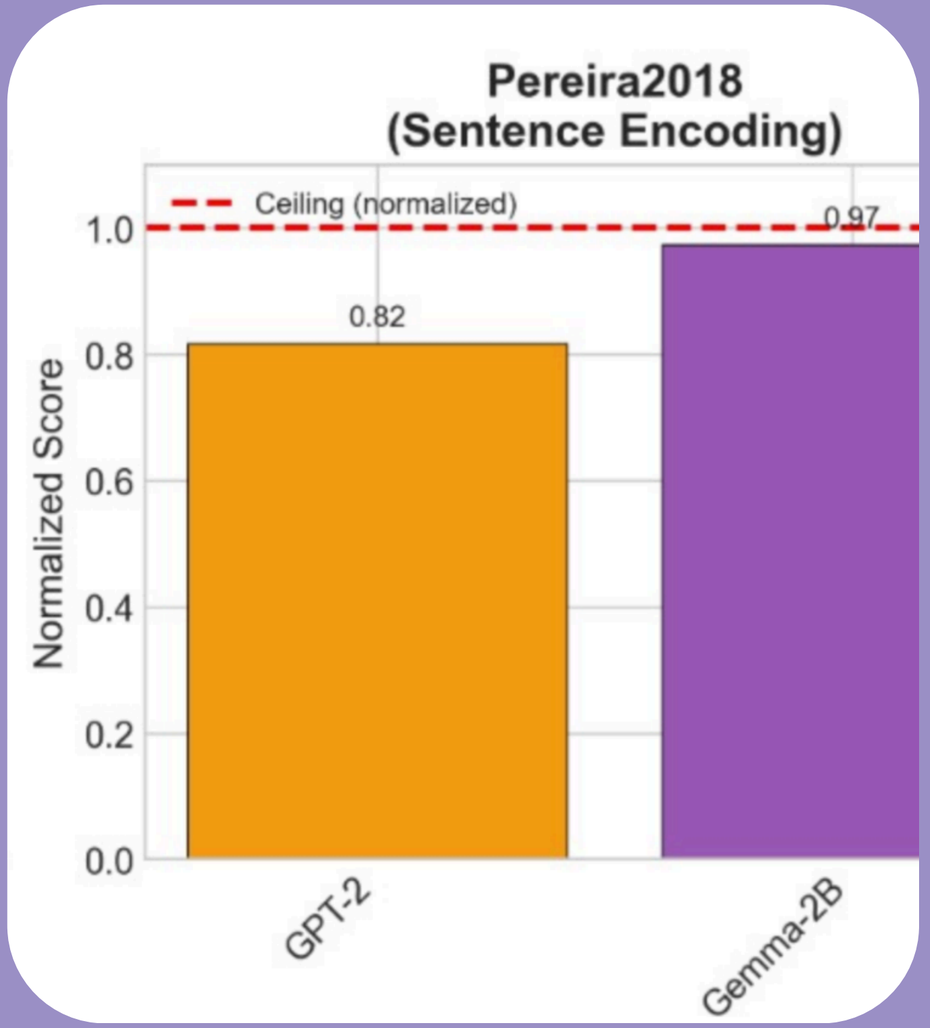
Larger and more expressive models achieve substantially higher scores, indicating improved alignment with neural responses.

Fedorenko2016 benefits from models capable of capturing richer semantic structure and long-range dependencies.

GPT-2 vs Gemma-2B on Pereira2018

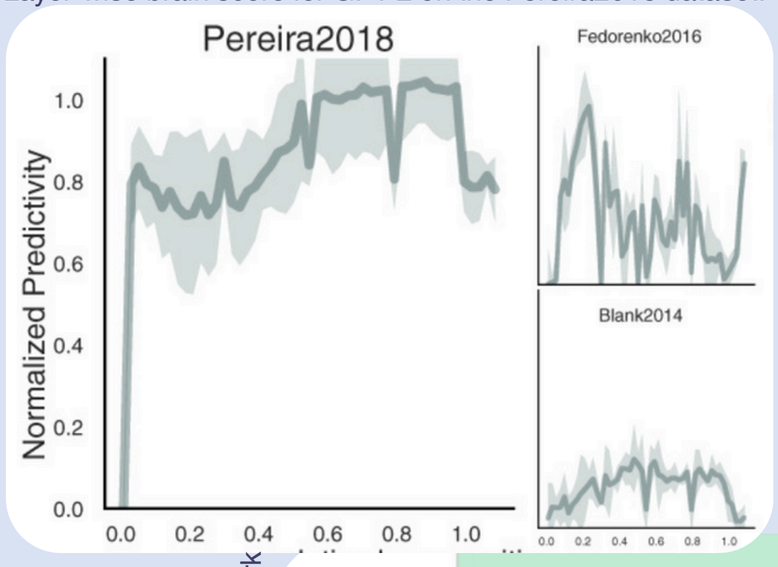
We evaluate two of the strongest models selected in our study GPT-2 and Gemma-2B on Pereira2018, the largest and most challenging dataset used.

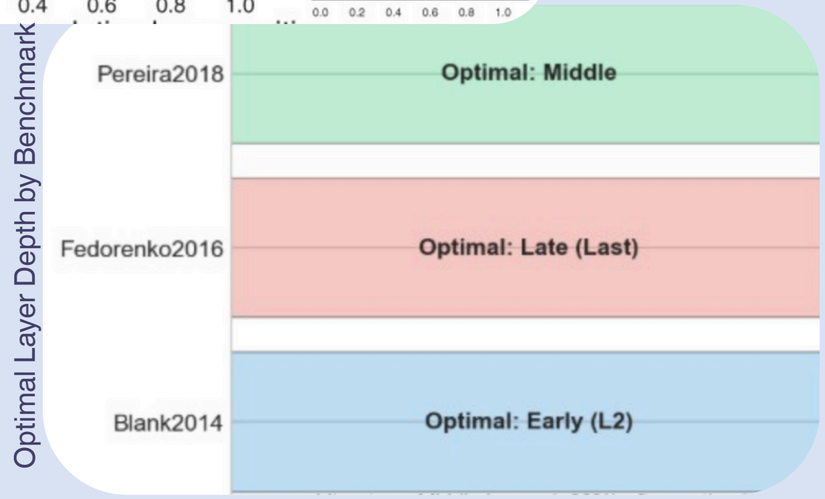
This comparison highlights how model scale and architecture impact alignment with human fMRI responses on complex natural language stimuli.



Model Evaluation and Comparison

Layer-wise brain score for GPT-2 on the Pereira2018 dataset.



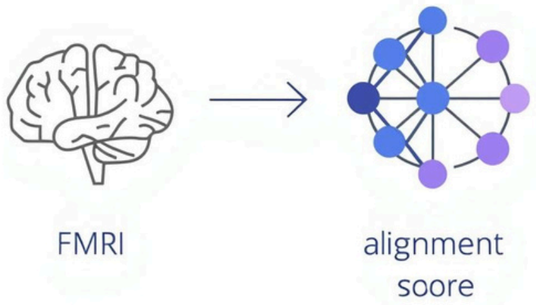


Brain-Score metrics: Quantify how well model activations predict neural responses.

Centered Kernel Alignment (CKA): Measures similarity between model representations and neural activation patterns across layers.

Pearson correlation: Assesses linear correlation between predicted and observed neural responses.

Benchmark comparisons: Evaluate models (e.g., GPT-2) on datasets like Pereira2018 to find layers and architectures that best align with brain activity.



Conclusions

- We replicated and validated results from prior work by re-evaluating GPT-2 on established fMRI language benchmarks.
- We extended the original study by evaluating newer and more diverse language models, ranging from small (SmolLM2, Pythia-70M) to larger architectures (Gemma-2B).
- Our results show that model choice strongly influences brain alignment, with recent models achieving higher scores on large-scale benchmarks.

Thank you!

For more details, please refer to:

- the full project repository
- the complete experimental documentation

These resources provide full implementation details, evaluation protocols, and extended results.