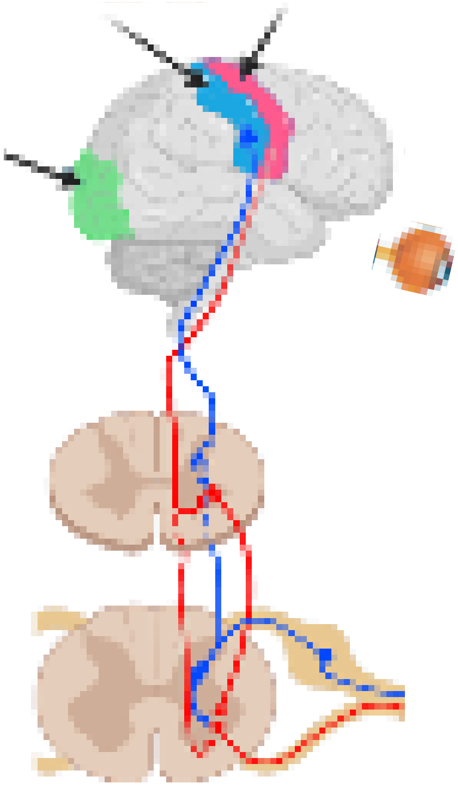
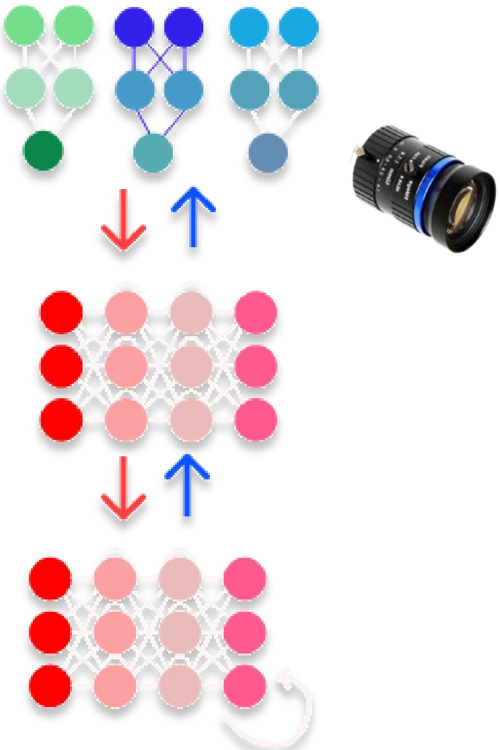
NX-414: Brain-like computation and intelligence

Martin Schrimpf

Lecture 9, 16 April 2025

Biological Intelligence ⟷ Artificial Intelligence

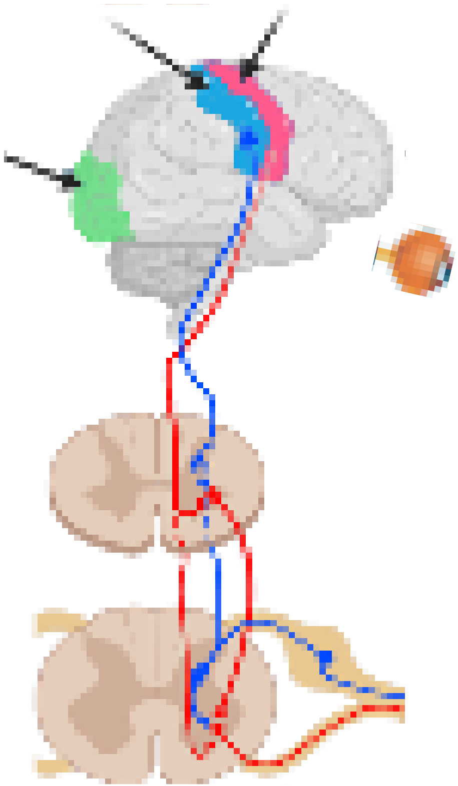




Hausmann & Marin-Vargas et al., 2021

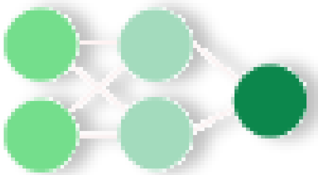
EPFL

Normative frameworks



Information theoretic

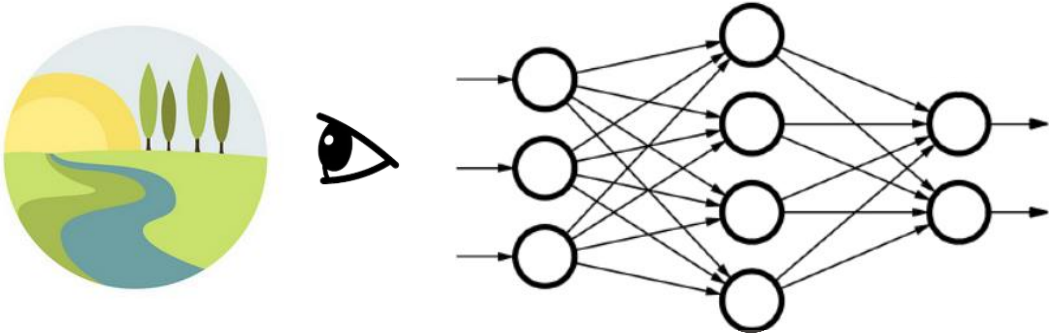
e.g. sparse coding,
redundancy reduction,
mutual information …

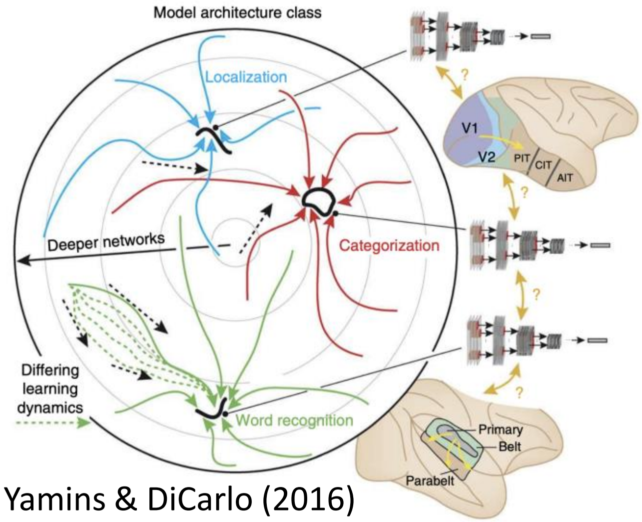


Utilitarian

e.g. **recognize objects**,
chase prey, navigate …

Using deep neural networks as goal-driven models of a system





Yamins & DiCarlo (2016)

Vision: object recognition. Yamins & Hong et al. (2014), Schrimpf & Kubilius et al. (2018)

Audition: speech recognition, speaker & sound identification. Kell et al. (2018)

Somatosentation: shape recognition. Zhuang et al. (2017)

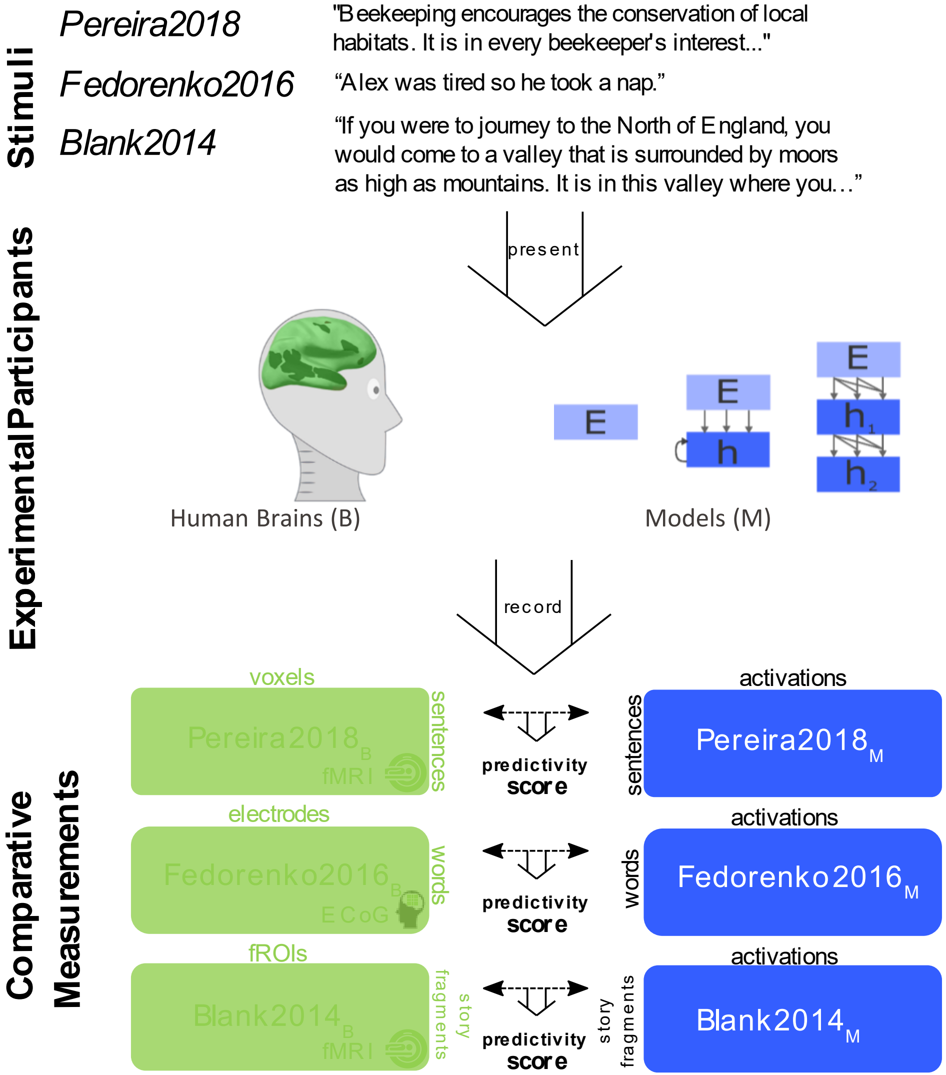
Language: next-word prediction. Schrimpf et al. (2021)

Decision making: context-dependent choice. Mante & Sussilo et al. (2013)

Proprioception: action recognition. Sandbrink et al. (2023)

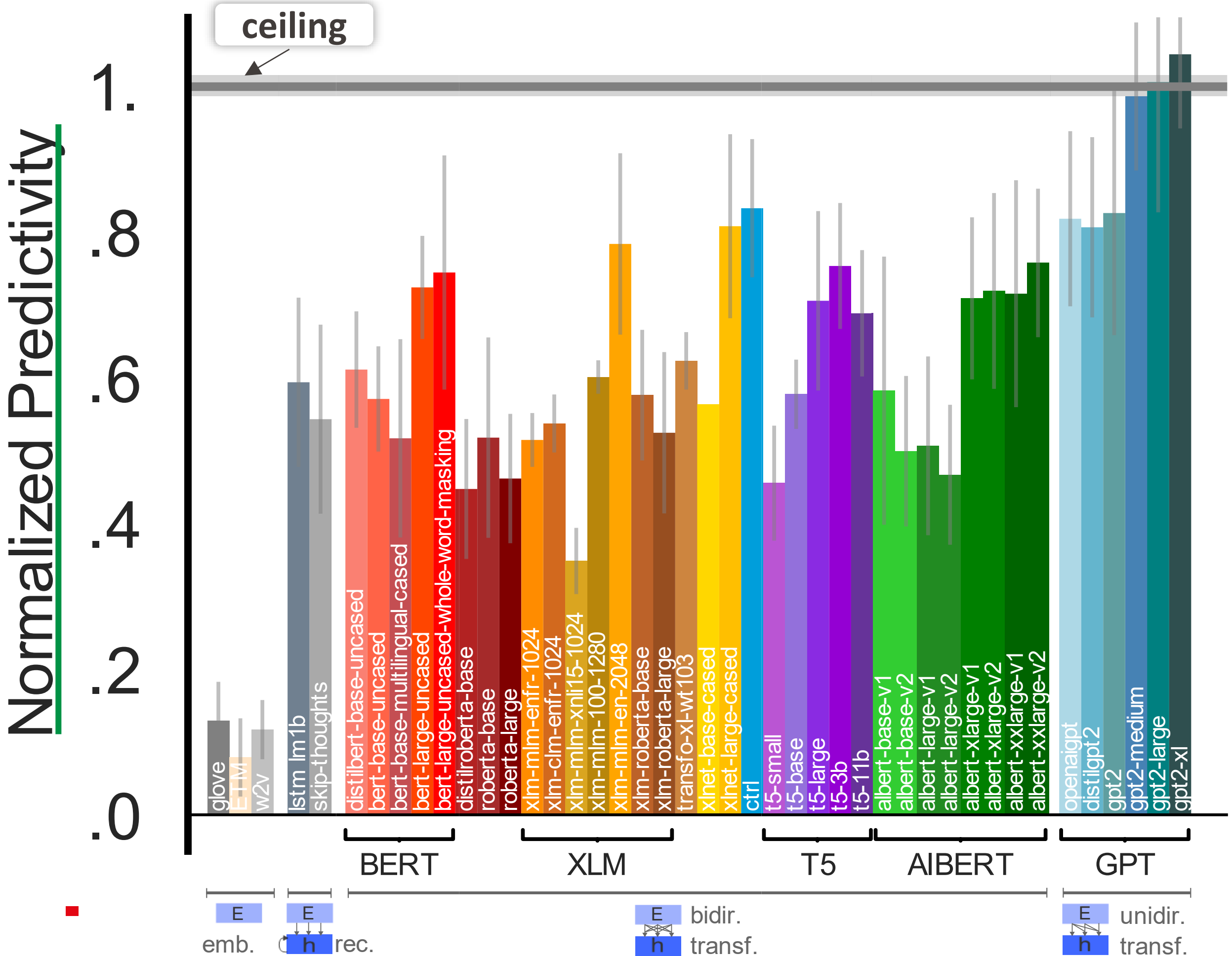
Recap from last time

- Language as a **bridge from perception to higher cognition**.
  Language is not thought.
- **Human language network**: functionally defined.
  Activation to sentences > lists of non-words
- Brain **recordings mostly fMRI**. Data limitations and noisiness,
  quantify via cross-subject consistency "ceiling"
- Model classes in **natural language processing**:
  embedding (e.g. GloVe), recurrent (e.g. LSTM), transformer (e.g. GPT)
- Evaluate model-to-brain similarity via **benchmarks.**
  Combine experimental paradigm, biological dataset, and similarity metric



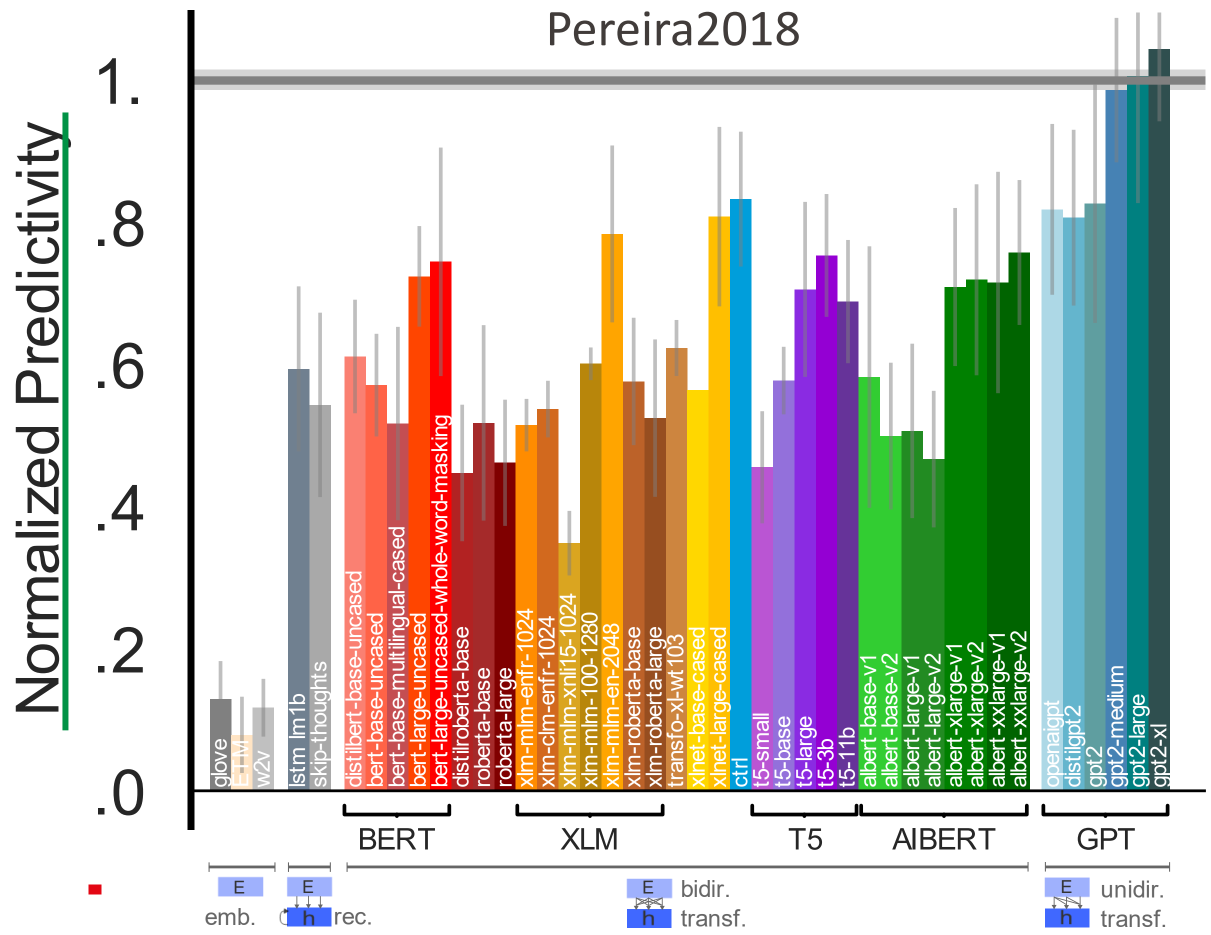
We want one model to predict *all* data

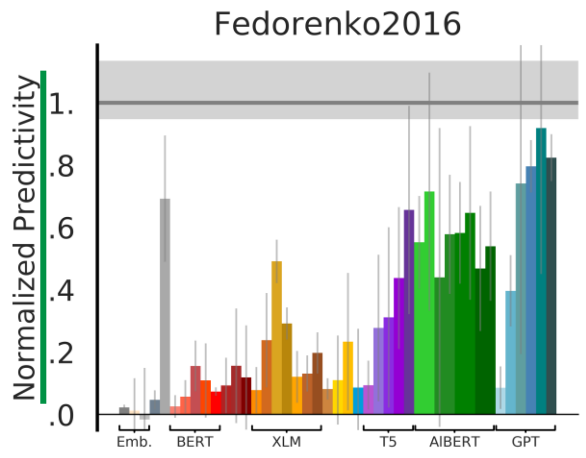
Certain language models predict human language recordings

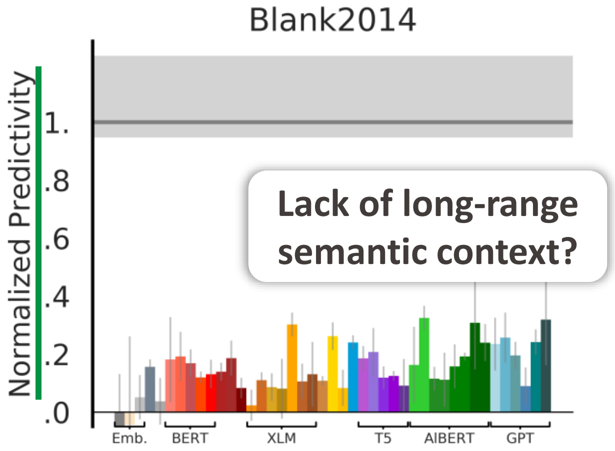


gpt2-xl hits our estimated ceiling for this benchmark

Language Models predict human language recordings





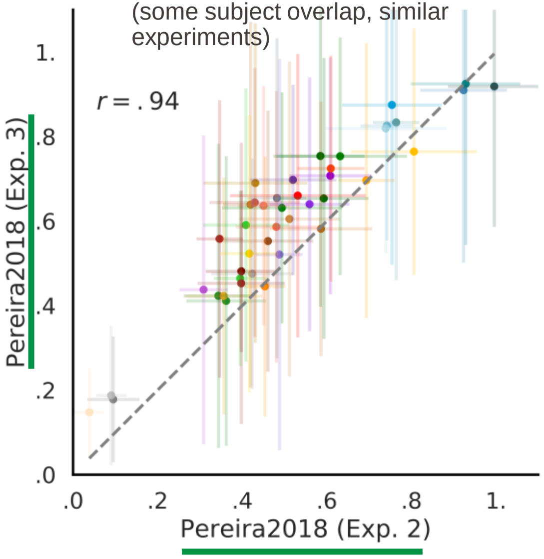


Control: model scores across benchmarks are correlated, although differences exist

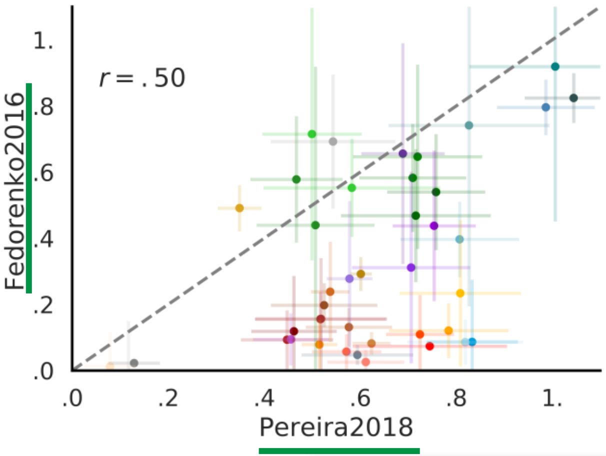
Scores generalize to a good extent

Are the discrepancies an issue? A plus?

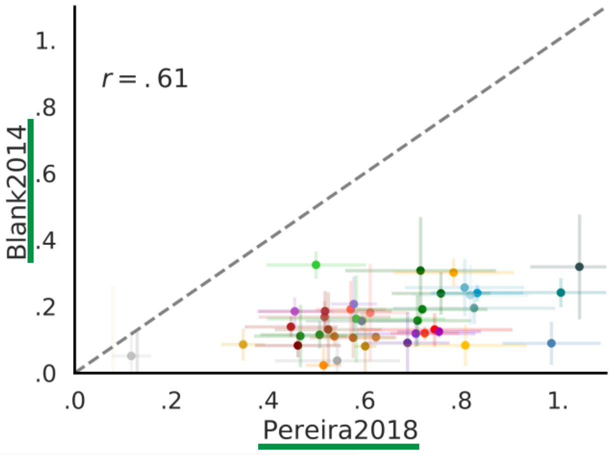
within Pereira2018
(some subject overlap, similar experiments)



Pereira2018 → Fedorenko2016

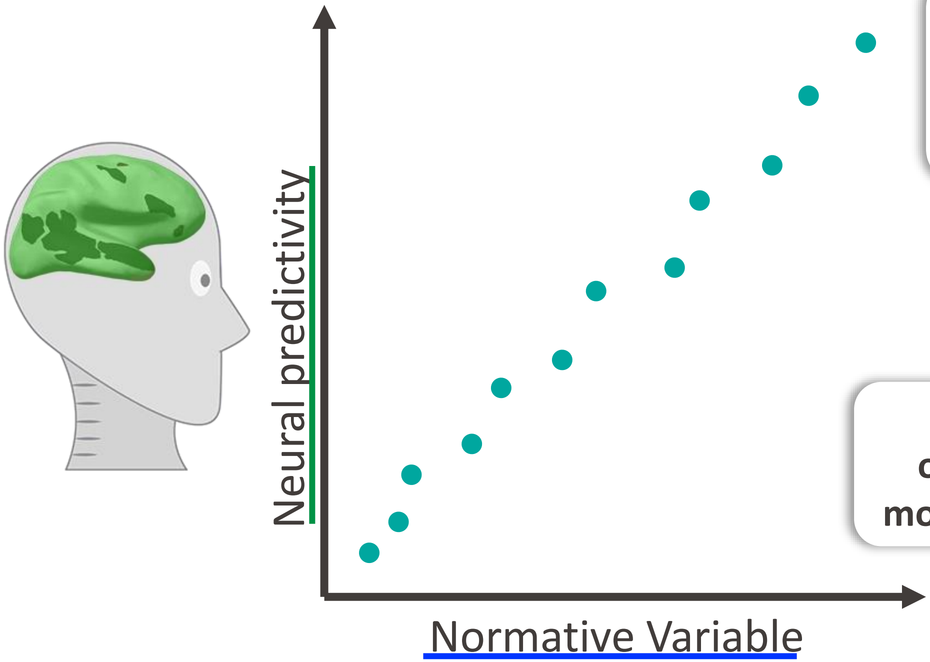


Pereira2018 → Blank2014



But there are also differences, making each individual benchmark valuable

What explains the model differences?



Goal 1: possible explanation *why* some models are better than others (hinting at optimization in the brain)

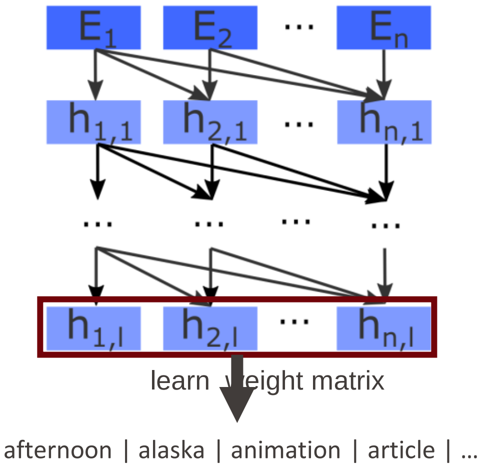
Goal 2: if x-axis is easier to optimize than y-axis, we can more *efficiently* improve models

Next-Word Prediction on WikiText-2

= Gold dollar =
The gold dollar or gold one @-@ dollar piece was a coin struck as a regular issue by the United States Bureau of the Mint from 1849 to 1889 . The coin had three types over its lifetime , all designed by Mint Chief Engraver James B. Longacre . The Type 1 issue had …

	WikiText-2		
	Train	**Valid**	**Test**
Articles	600	60	60
Tokens	2,088,628	217,646	245,569
Vocab	33,278		
OoV	2.6%		

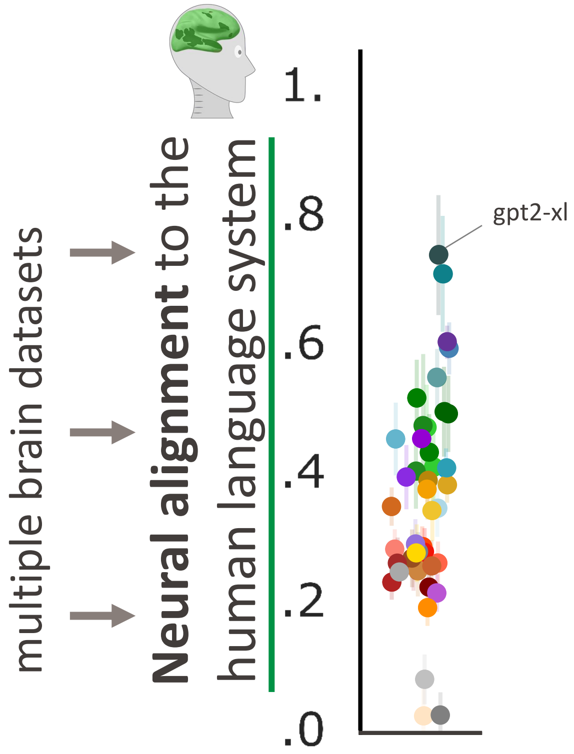
Alaska
Alaska is
Alaska is about
Alaska is about twelve
Alaska is about twelve times
Alaska is about twelve times larger
Alaska is about twelve times larger than
Alaska is about twelve times larger than New
Alaska is about twelve times larger than New York



Surprisal of seeing actual next word: **perplexity** = exp(NLL Loss)

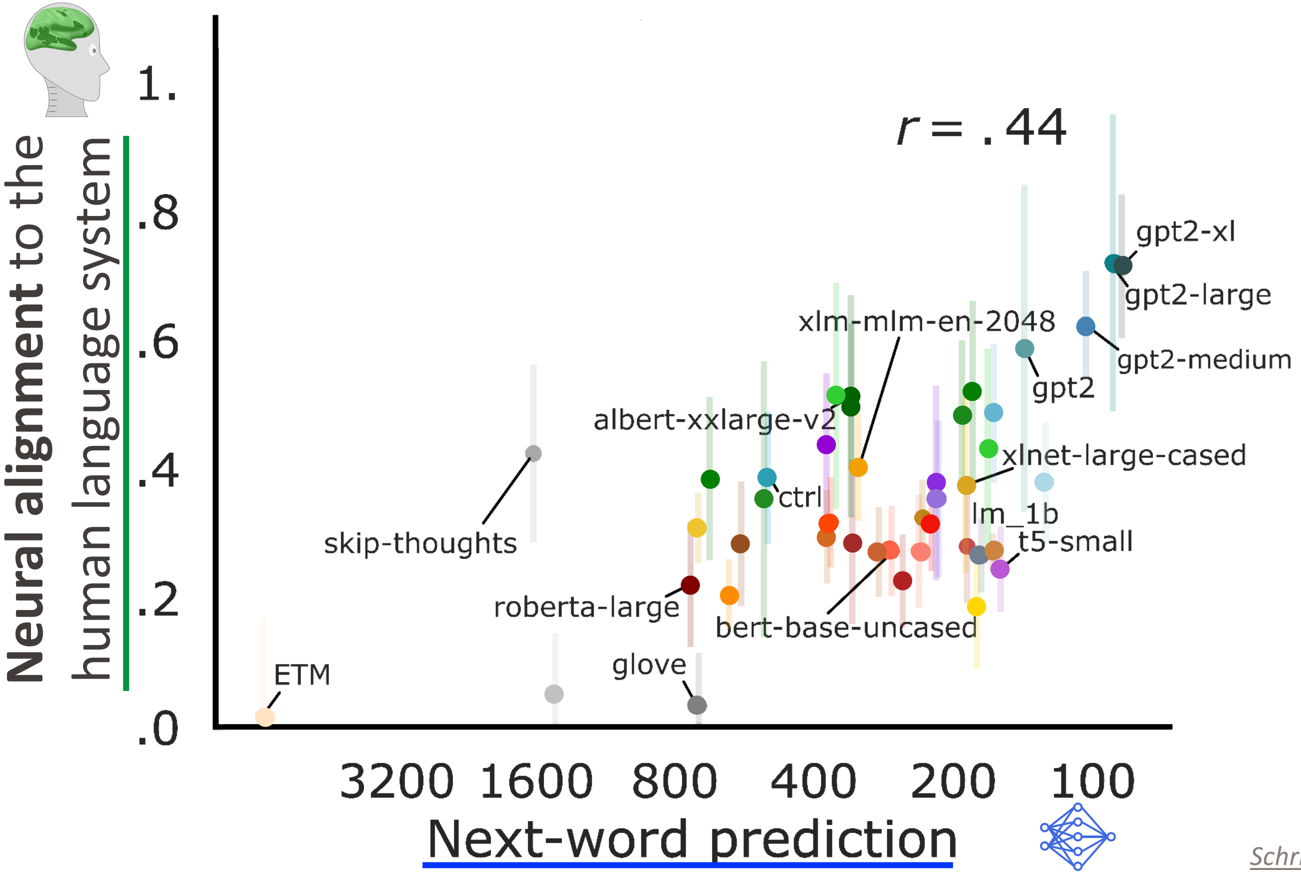
Merity et al. 2016

The better models can predict the next word, the more brain-like they are



Schrimpf et al. (PNAS 2021)

The better models can predict the next word, the more brain-like they are



Schrimpf et al. (PNAS 2021)

EPFL

What about other language tasks?

GLUE

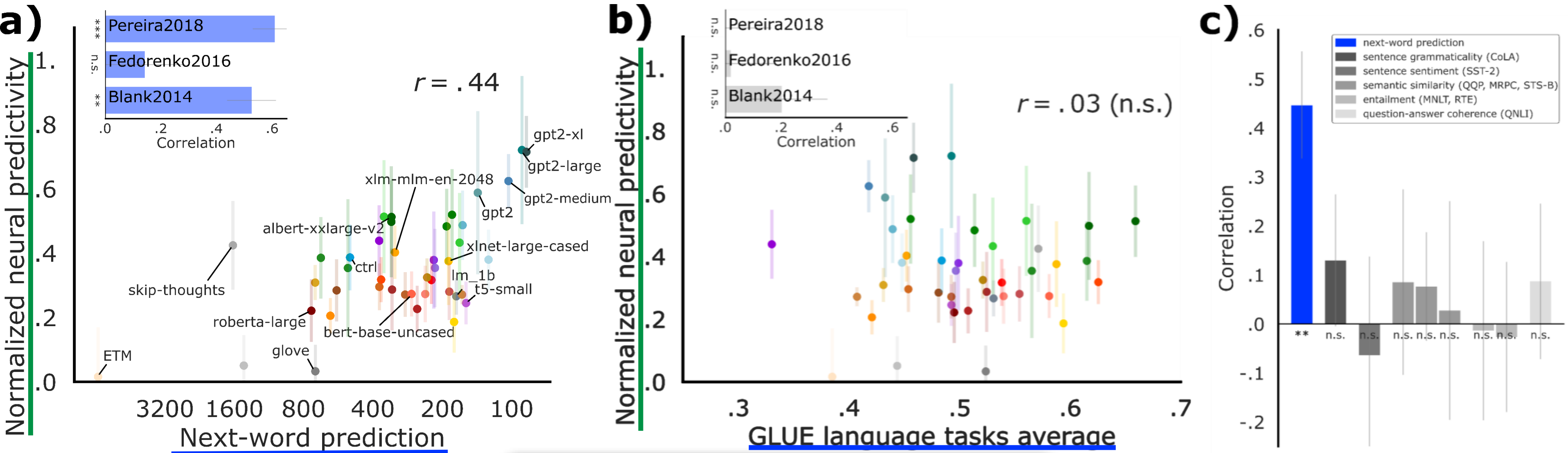
9 "General Language Understanding Evaluation" tasks:

Sentence grammaticality (CoLa)
Sentence sentiment (SST-2)
Semantic similarity (QQP, MRPC, STS-B)
Entailment (MNLT, RTE)
Question-answer coherence (QNLI)
Winograd (WNLI; ignored due to known issues)

Which of these model task performances will correlate with brain alignment?

1. *None*
2. *Some*
3. *All*

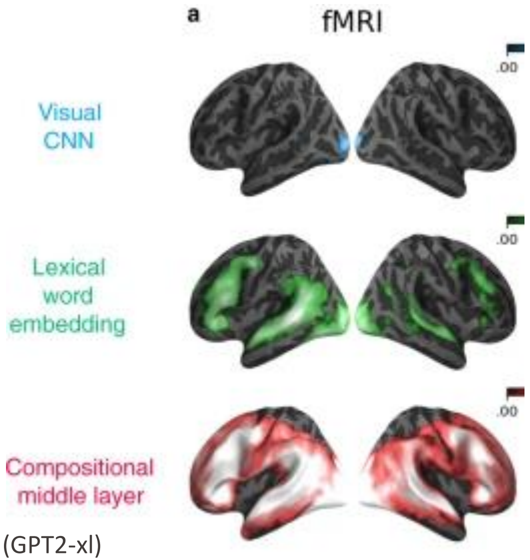
Next-Word Prediction performance selectively correlates with neural predictivity



Online prediction may fundamentally shape language processing in the brain

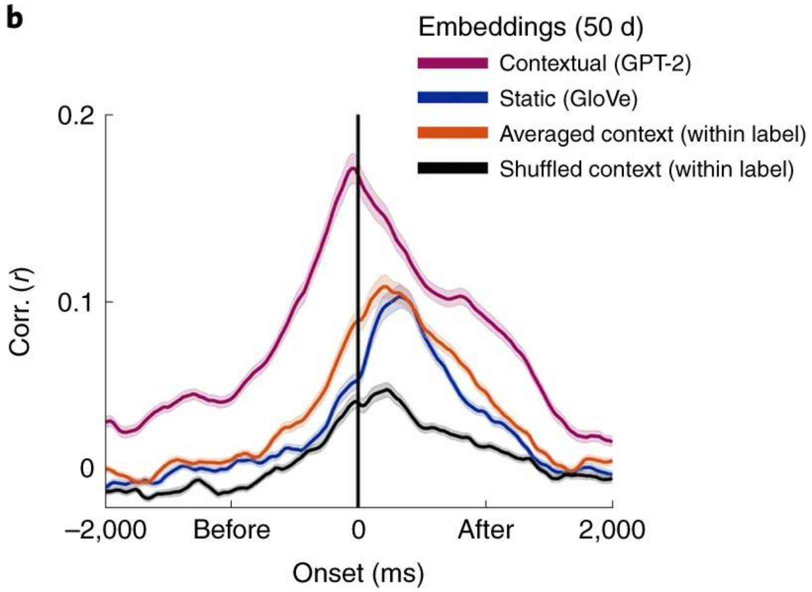
Related work

[Caucheteux et al. 2021]



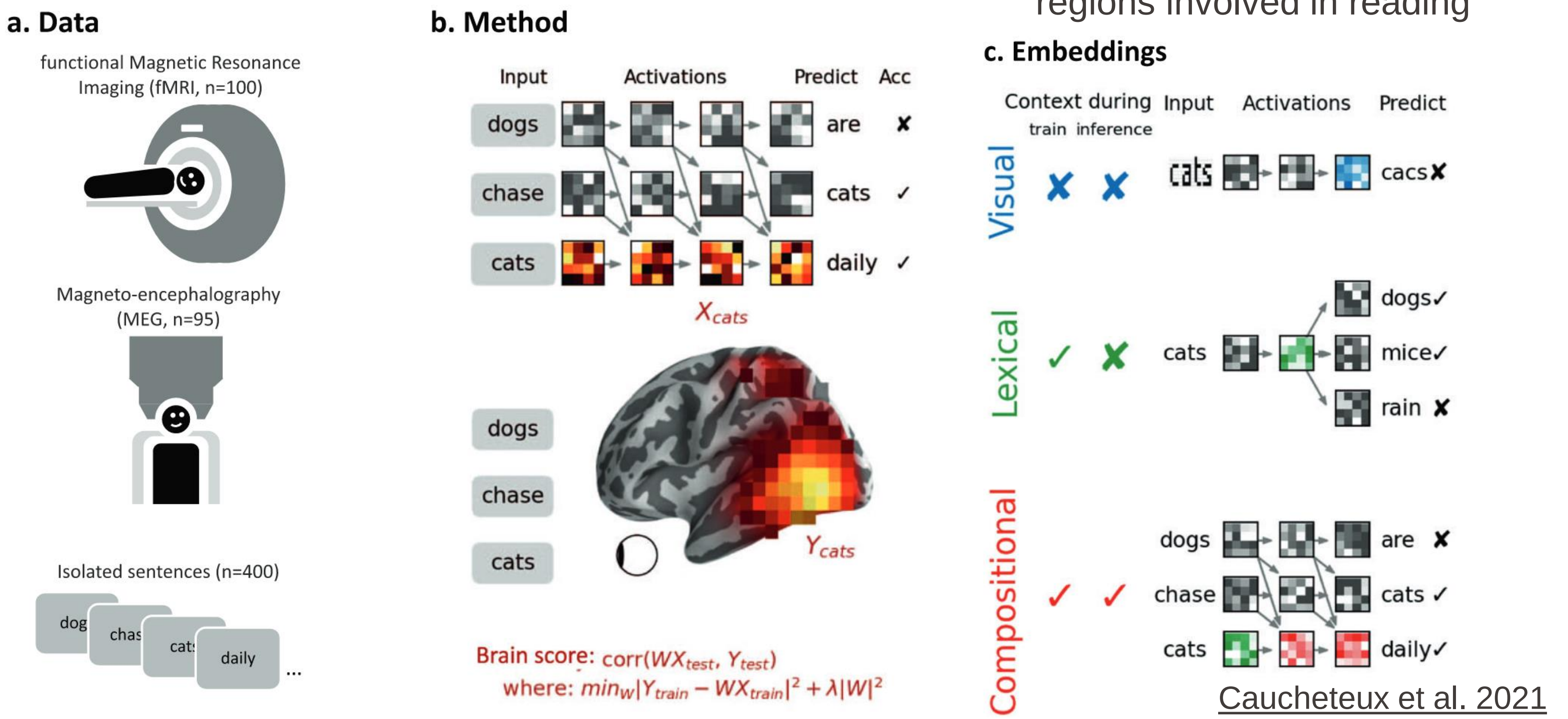
(GPT2-xl)

[Goldstein et al. 2022]



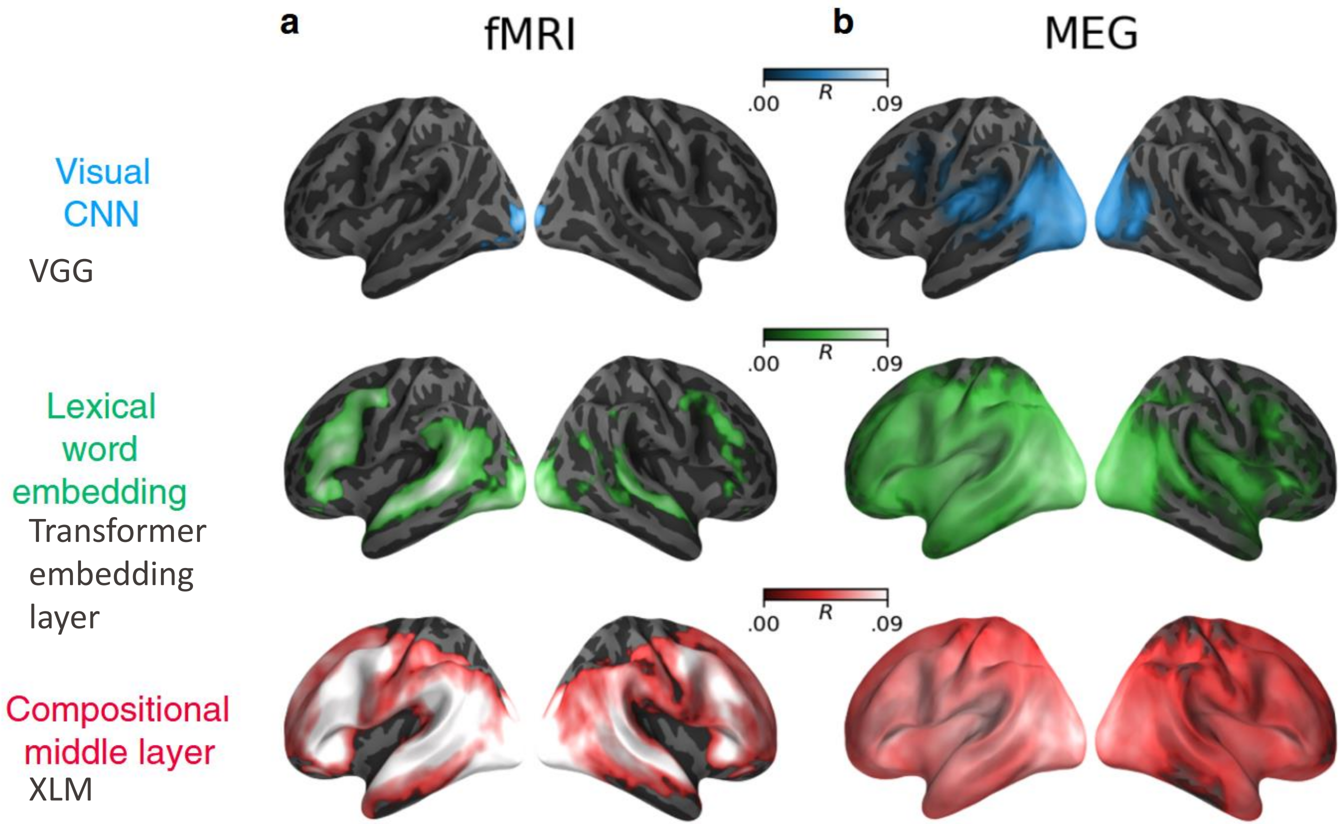
Separating different brain regions with different model types

- Idea: use a visual model, a non-contextual language model, and a contextual language model to identify a hierarchy of brain regions involved in reading



Caucheteux et al. 2021

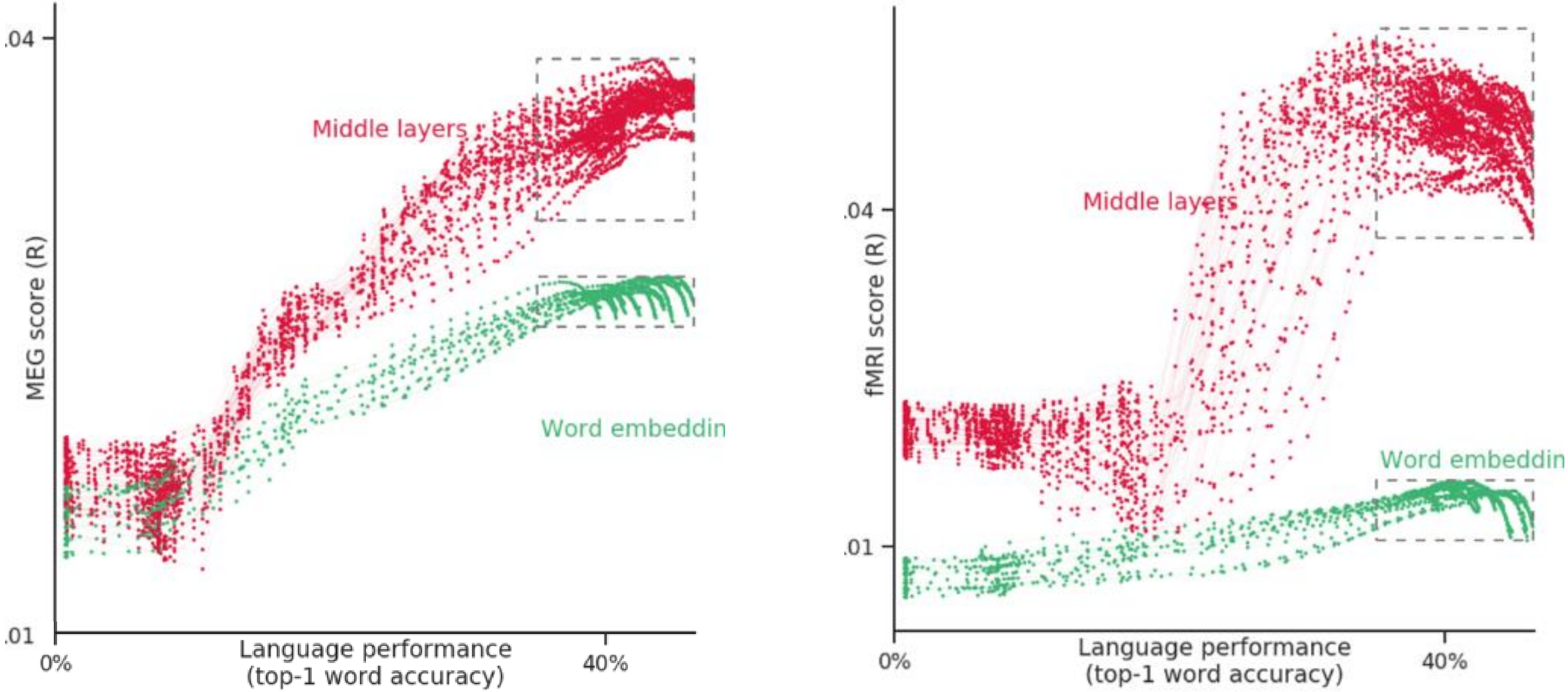
Separating different brain regions with different model types



- Different model types best explain different brain regions
- Visual model best explains early visual cortex
- Contextual language model explains downstream regions

Caucheteux et al. 2021

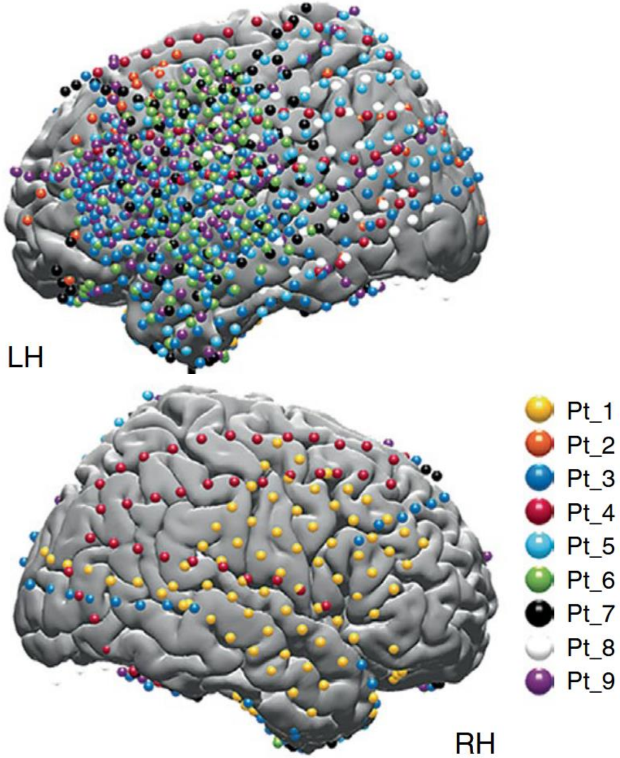
Same observation as we saw before: next-word prediction performance correlates with brain alignment



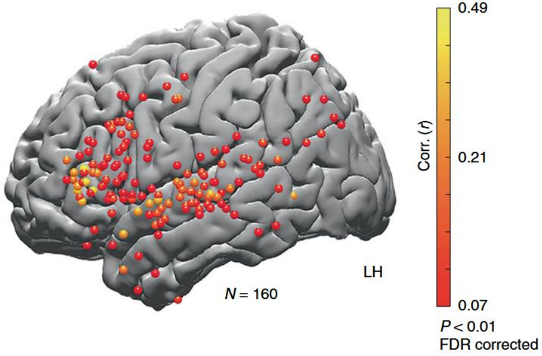
Caucheteux et al. 2021

The brain's language system might itself engage in next-word prediction

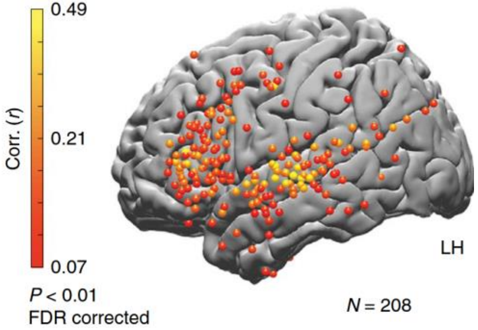
Electrode coverage



glove



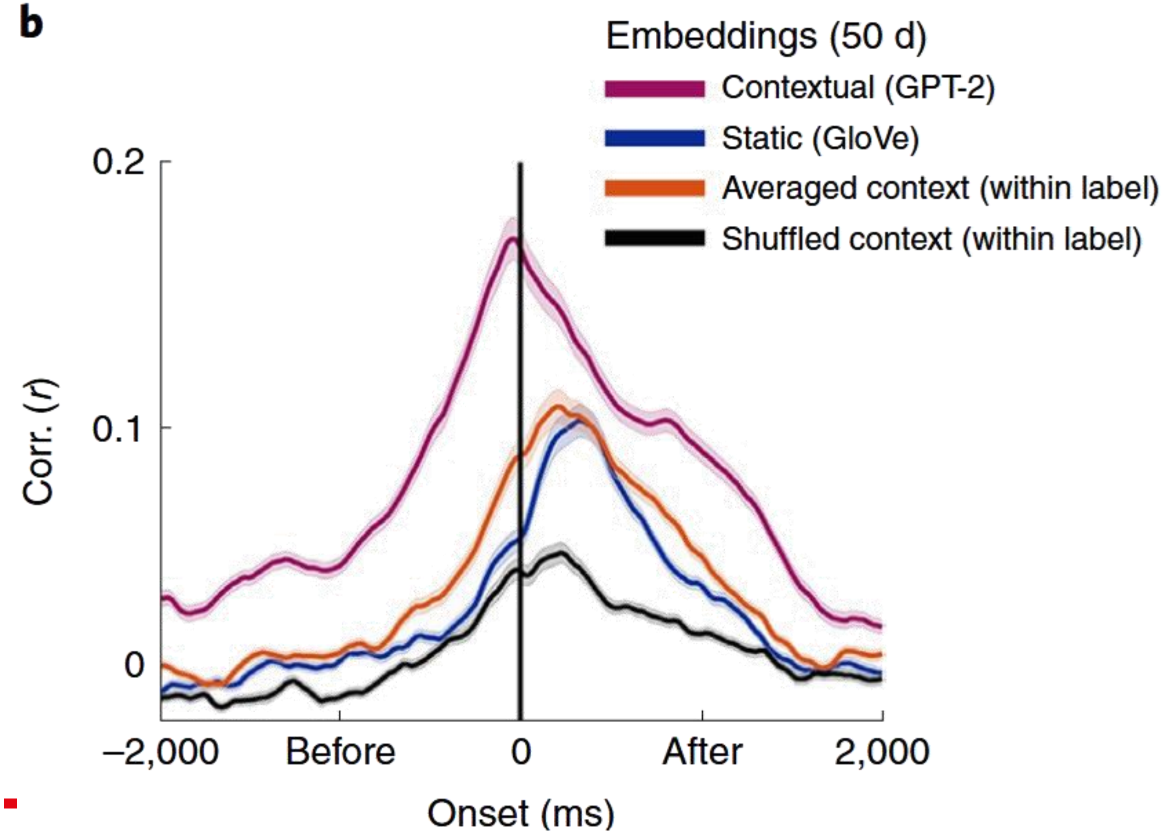
gpt2-xl



- Also on human electrode recordings, GPT2 outperforms GloVe

Goldstein et al. 2022

The brain's language system might itself engage in next-word prediction



- Contextual embeddings in GPT2 outperform non-specific context and non-contextual embeddings
- Contextual embeddings predict brain activity even before the next word occurs. Since GPT2 predicts the next token, its representations should be focused on the future
- The authors infer that the brain therefore also performs next-word prediction

Goldstein et al. 2022

Why build models in the first place?

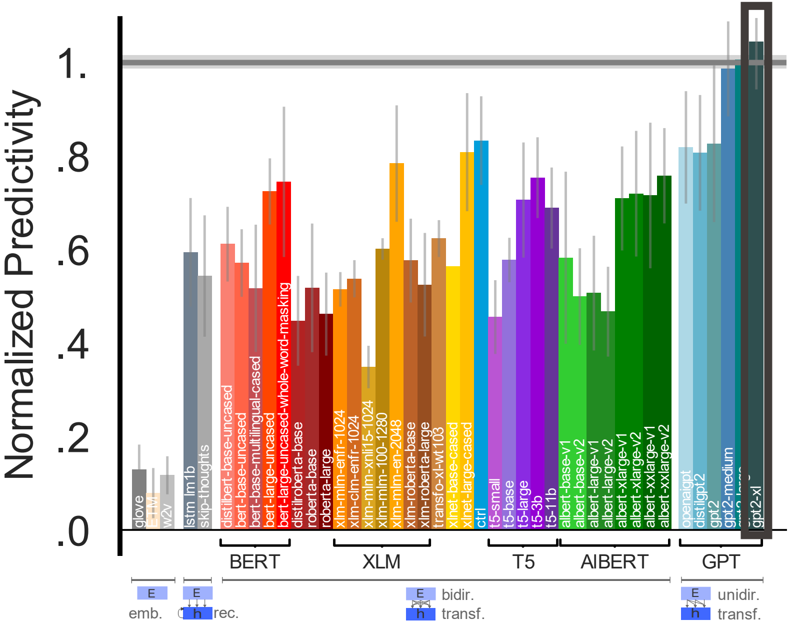
Efficient science

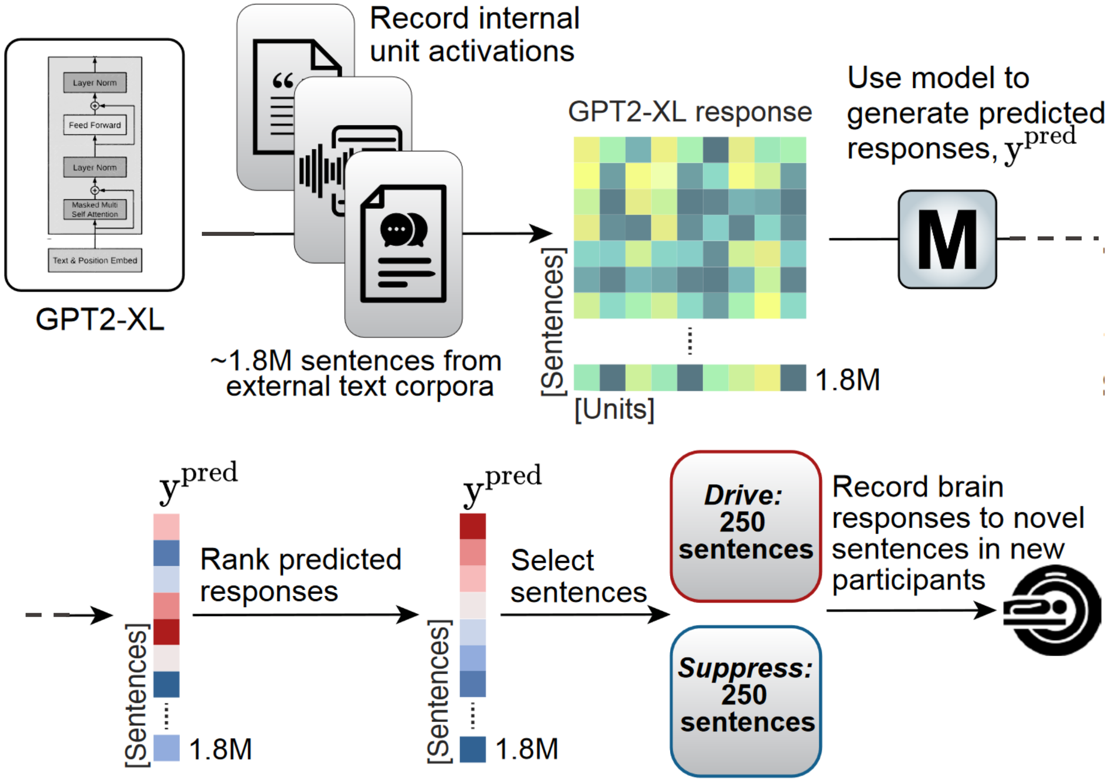
- Reproducible and uniquely specified (machine-executable)
- Integrative codification of state-of-the-art hypotheses across many pieces of evidence (potentially beyond the mind of any one individual)
- Quick prototyping of new experiments

Long-term benefits

- Better AI (personally I’m not holding my breath on this one)
- Computational understanding of human behavior and underlying neural mechanisms
- Clinical applications

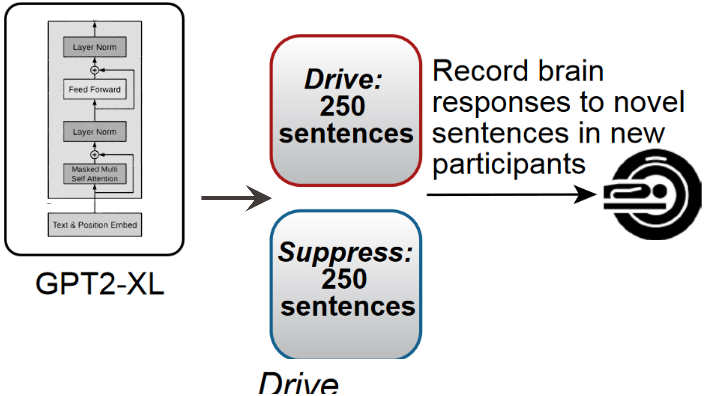
We can use brain-aligned LLMs to noninvasively control neural activity





Tuckute et al. NatHumBeh 2024

We can use brain-aligned LLMs to noninvasively control neural activity



Drive

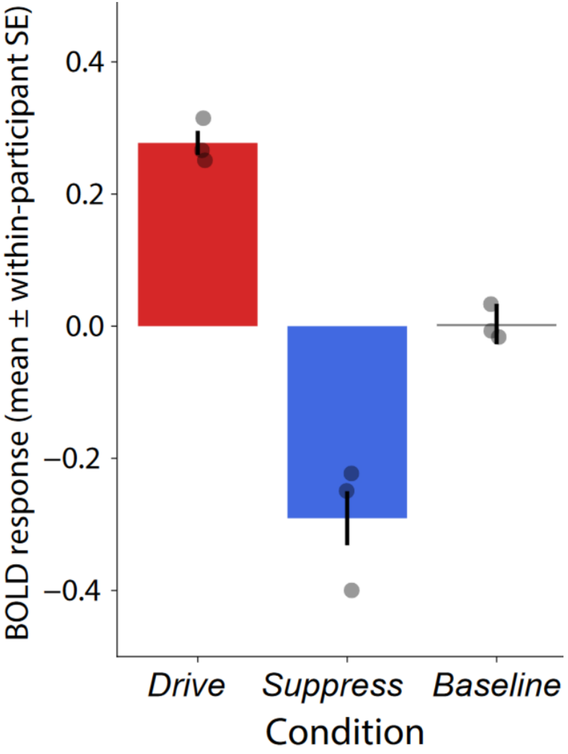
Sent
res

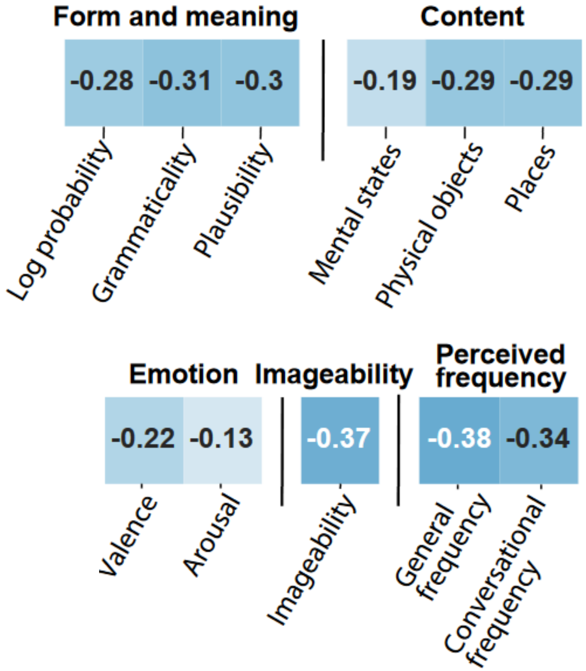
Changir
Notice h
Add, so
Jiffy Lub
People
Buy sell
Turin lo
URL rig

Suppress

Sentences identified to elicit minimal response in the language network

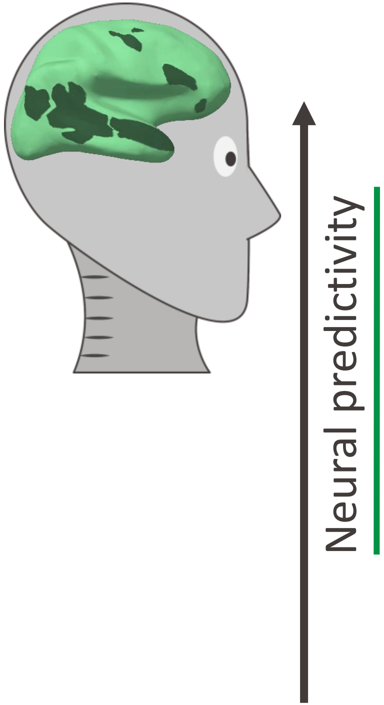
We were sitting on the couch.
That is such a beautiful picture!
They stood there for a moment.
They went up the stairs together.
Inside was a tiny silver sculpture.
They walked out onto the balcony.
Cas gazed up at the sky.
What else is there to do?



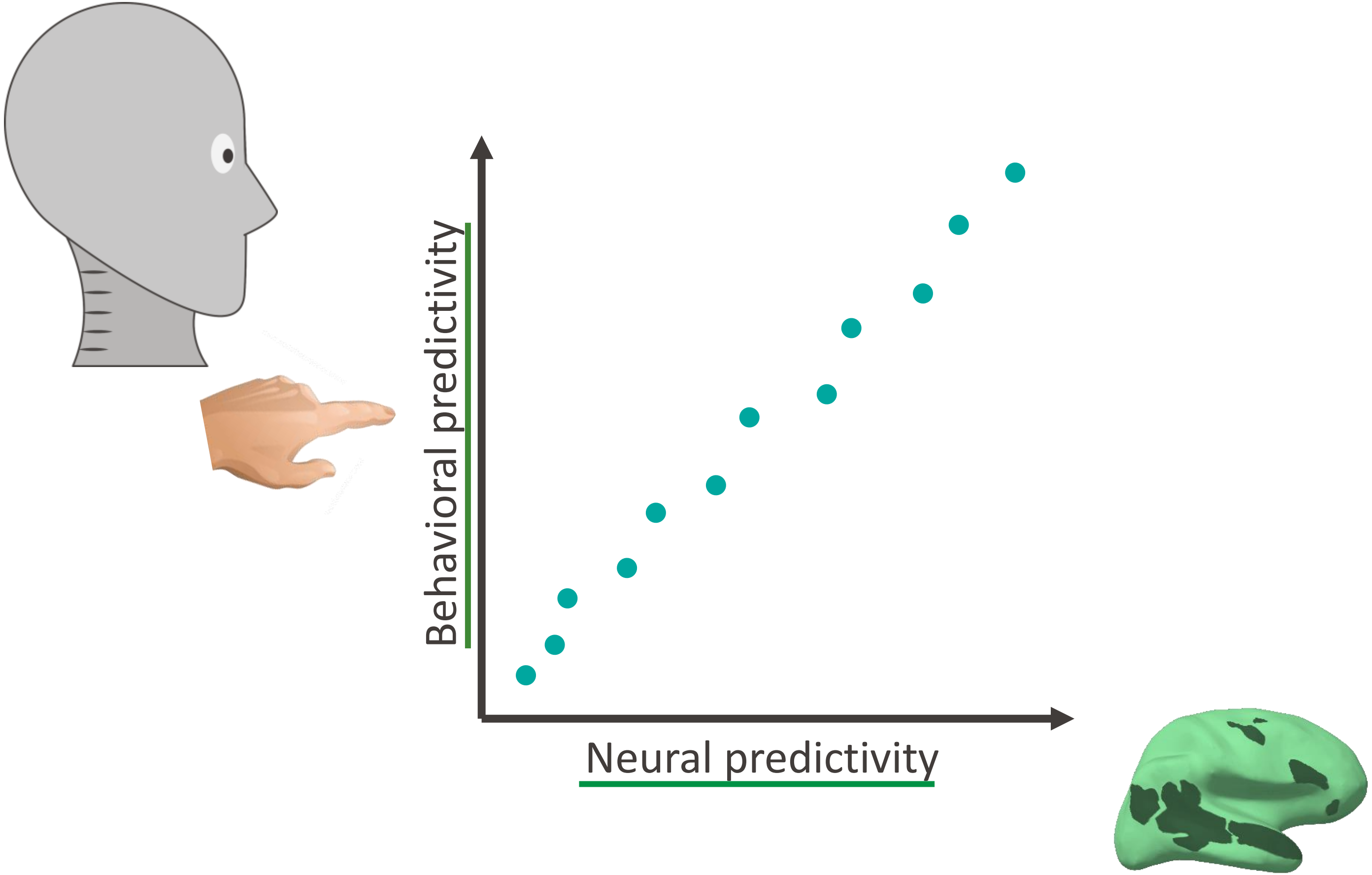


Tuckute et al. NatHumBeh 2024

Is any of this behaviorally relevant?



Is any of this behaviorally relevant?



~Illustration of behavioral setup: speed reading

https://youtu.be/SPPe7-tusOg?t=15

Behavioral target: human reading times

Futrell et al. 2018

- 10256 words x 179 subjects
- *If | you | were | to | journey | to | the | North | of | England, | you | would | come | to | a | valley | that | is | surrounded | by | moors | as | high | as | mountains. | It | is | in | this | valley | where | you | would | find | the | city | of | Bradford, | where | once | a | thousand | spinning | …*

Treat reading times as representation target

The Natural Stories Corpus

Richard Futrell[1], Edward Gibson[1], Harry J. Tily[2], Idan Blank[1], Anastasia Vishnevetsky[1], Steven T. Piantadosi[3], and Evelina Fedorenko[4,5]

[1]MIT Department of Brain and Cognitive Sciences [2]Netflix, Inc.

[3]University of Rochester Department of Brain and Cognitive Sciences

[4]Massachusetts General Hospital Department of Psychiatry

[5]Harvard Medical School Department of Psychiatry

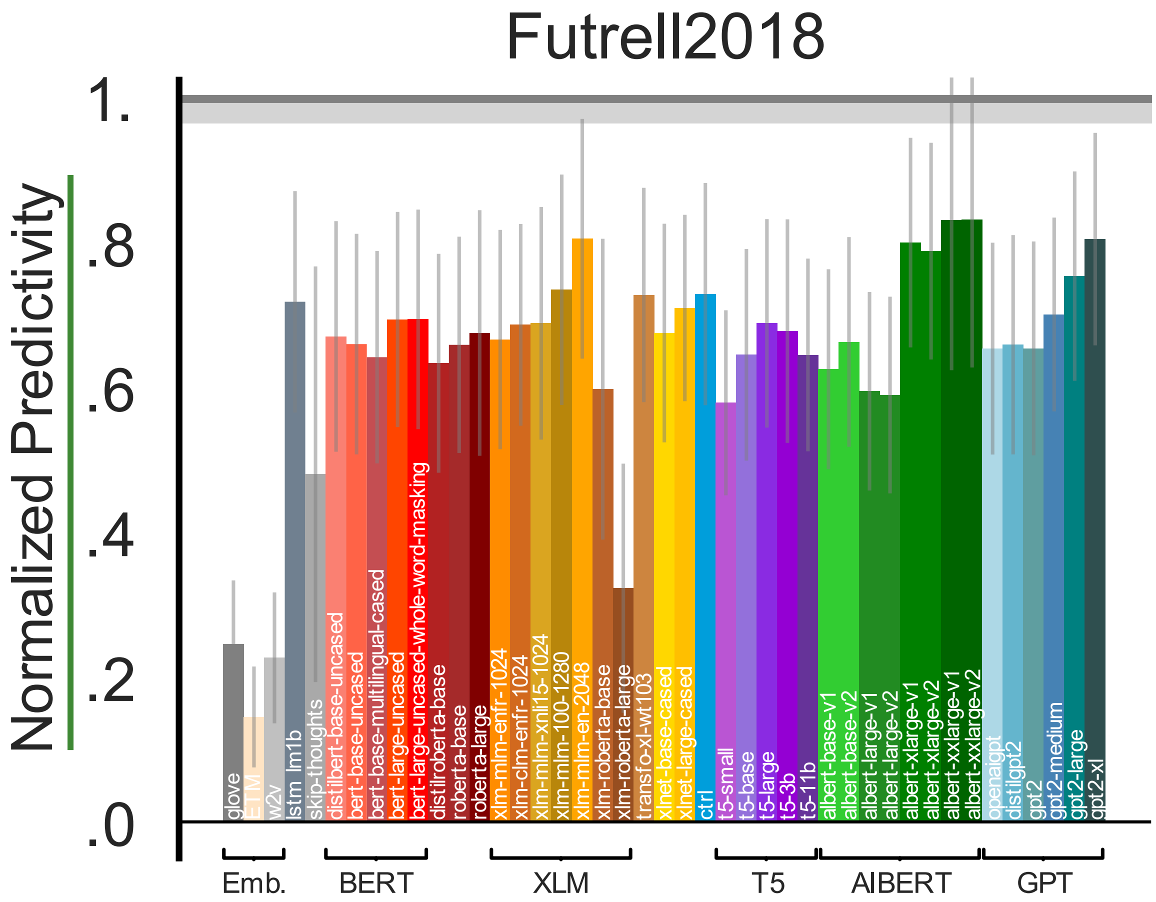
{futrell, egibson, iblank, evelina9}@mit.edu, hal.tily@gmail.com, staseyvi@mail.med.upenn.edu

Abstract

It is now a common practice to compare models of human language processing by comparing how well they predict behavioral and neural measures of processing difficulty, such as reading times, on corpora of rich naturalistic linguistic materials. However, many of these corpora, which are based on naturally-occurring text, do not contain many of the low-frequency syntactic constructions that are often required to distinguish between processing theories. Here we describe a new corpus consisting of English texts edited to contain many low-frequency syntactic constructions while still sounding fluent to native speakers. The corpus is annotated with hand-corrected Penn Treebank-style parse trees and includes self-paced reading time data and aligned audio recordings. Here we give an overview of the content of the corpus and release the data.

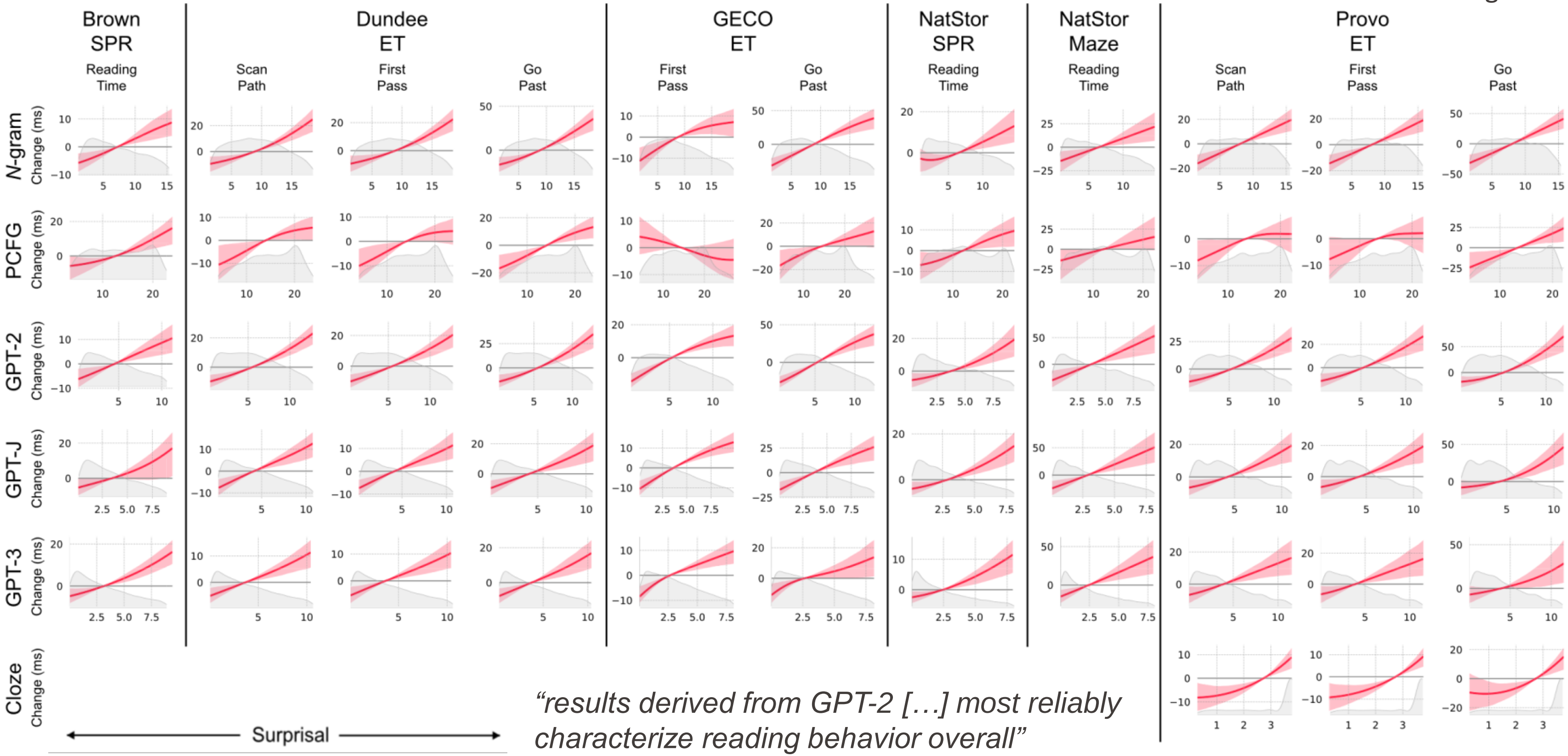
Keywords: Cognitive modeling, reading time, psycholinguistics

Behavioral scores



GPT-2 continues to shine in predicting human reading times

See also Smith & Levy 2013: effect of word probability on reading time is logarithmic

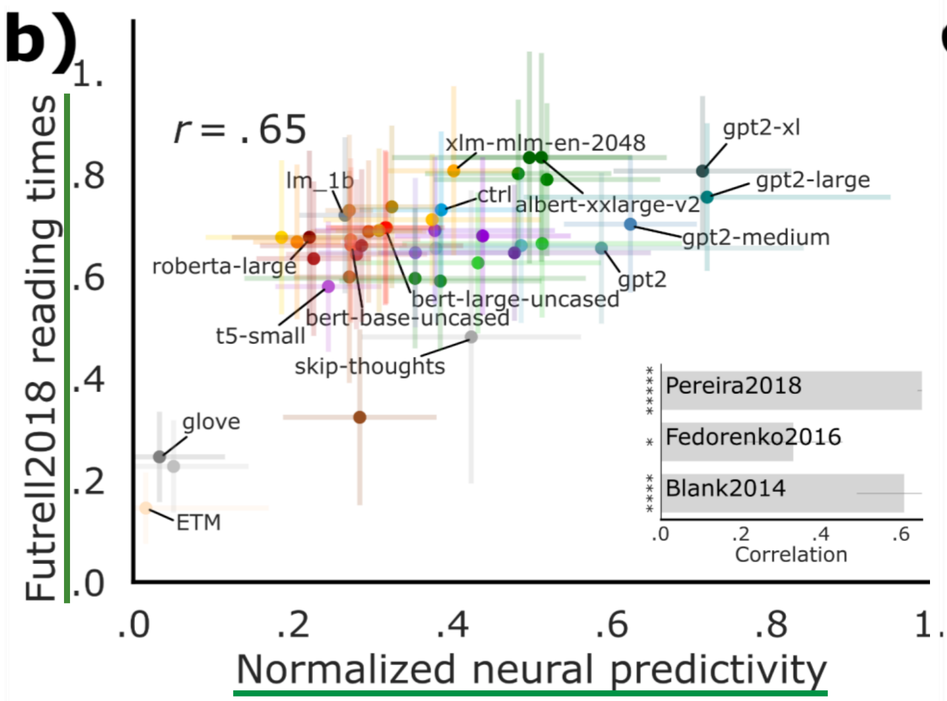


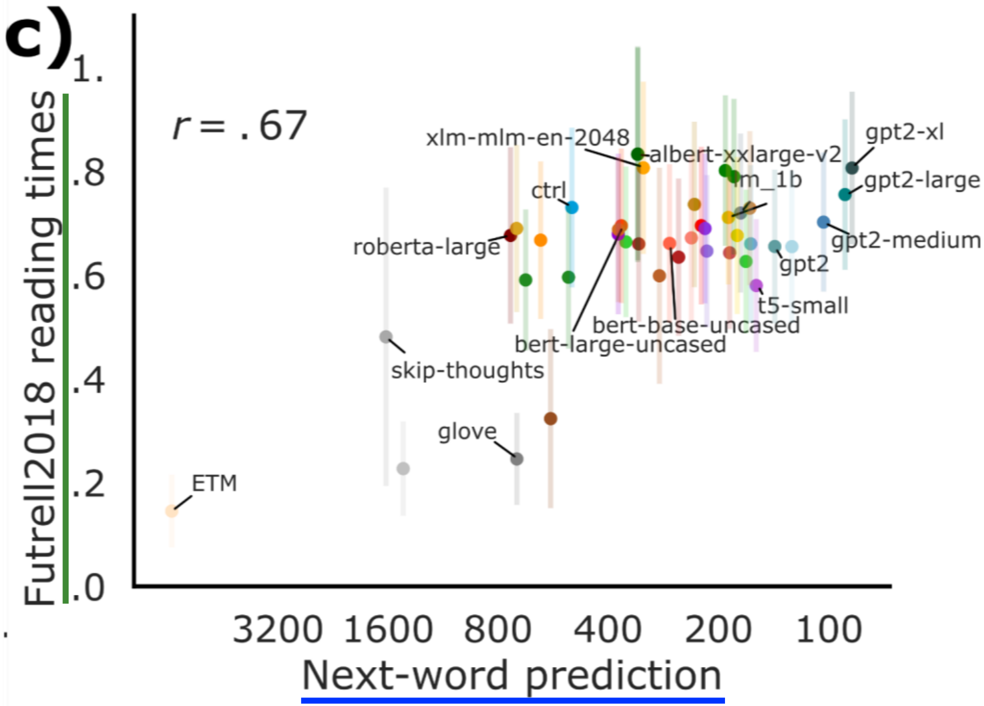
"results derived from GPT-2 […] most reliably characterize reading behavior overall"

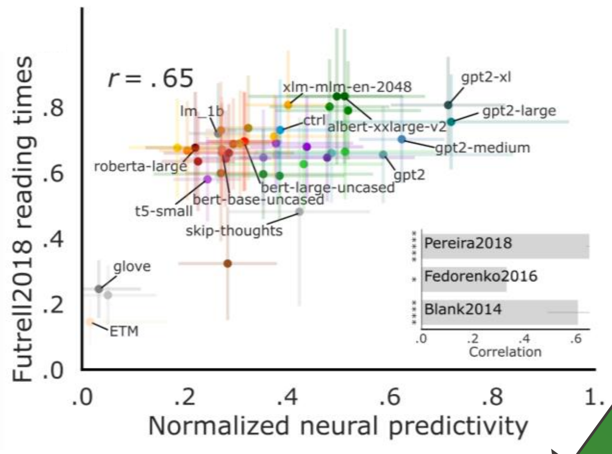
Shain et al. PNAS 2024

Neural scores correlate with Behavioral scores

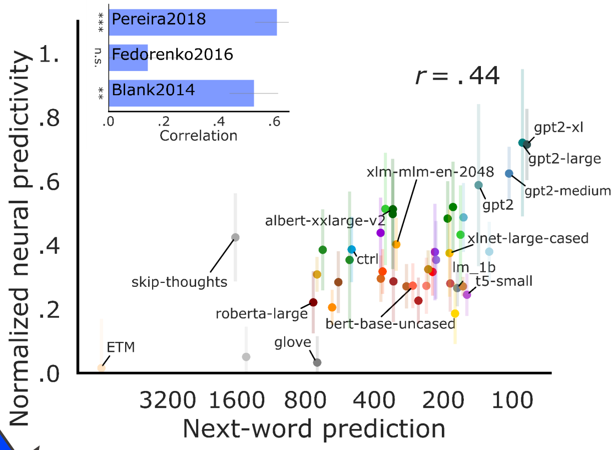
Task scores correlate with Behavioral scores







Neural

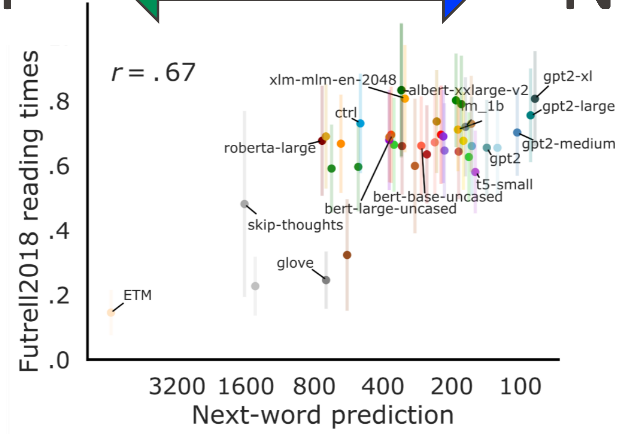


Integrative Modeling:
link neural mechanisms,
behavior, and computation

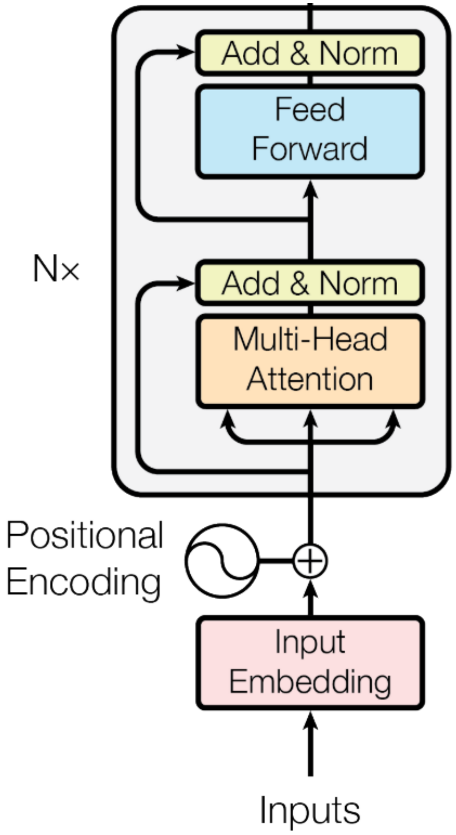
Schrimpf et al. Neuron 2020

Behavioral

Normative Task



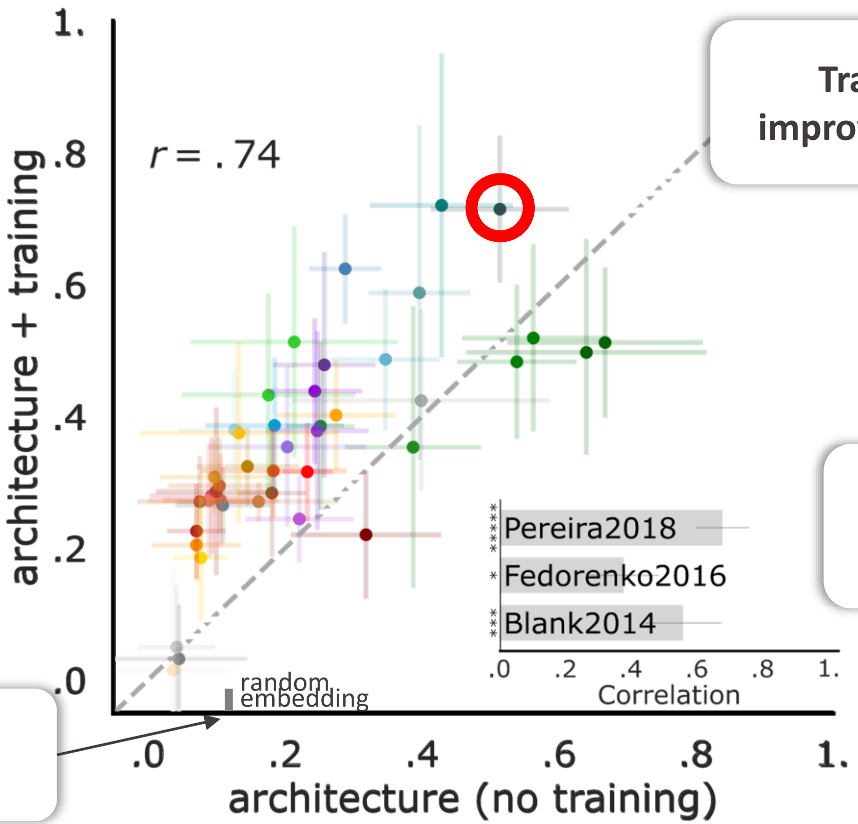
What is the relative importance of evolutionary and learning-based optimization?



Evolution ≃ community optimization over architectural properties

Experience-dependent learning ≃ updating of weights over training

Architecture substantially contributes to models' brain predictivity

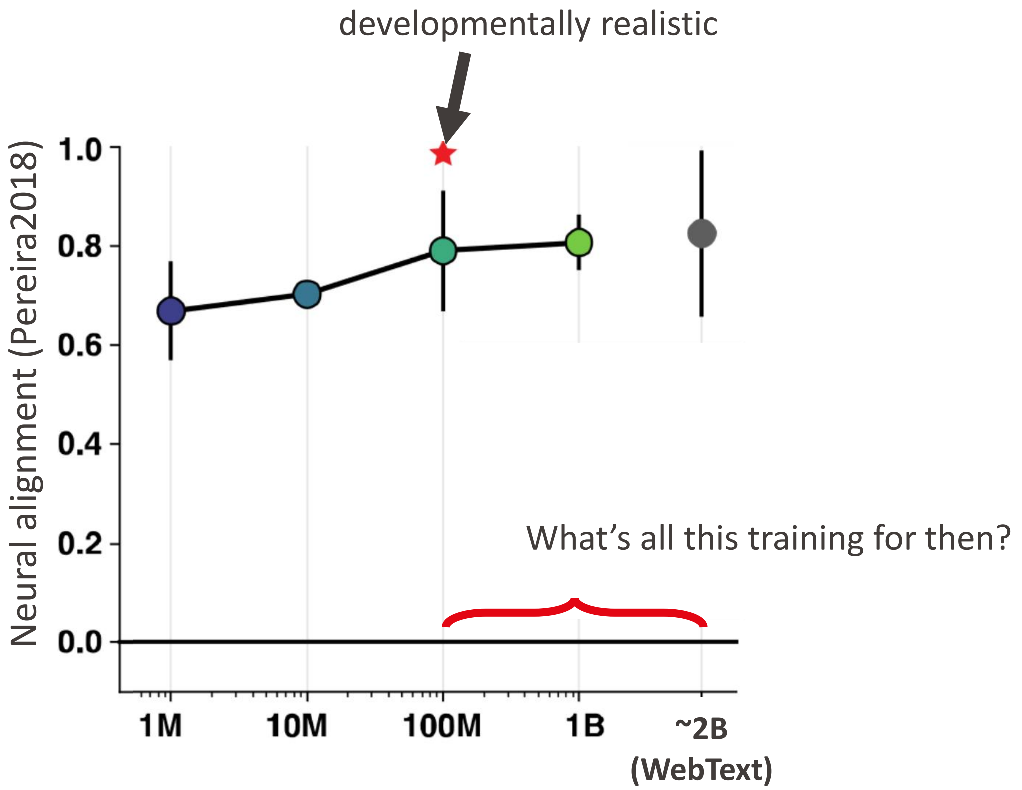


Training generally improves scores by ~53%

Inherent structure might be a key driver of brain-like language representations

Large feature size without structure is insufficient

LLMs align to the brain's language system after developmentally realistic amounts of training



Hosseini et al. 2022

Take-home messages

- Particular language models predict the human language system and behaviors
- Model-to-brain alignment is explained by next-word-prediction performance
- Model-to-behavior alignment correlates with brain, and task performance
- The best models can be used to noninvasively control brain activity
- Architecture and training both contribute to the brain-likeness of model representations