Large Language Models Are Aligned With The Human Language System

Martin Schrimpf

PATT, Schools of Life Sciences, and of
Computer and Communication Sciences

Neuro X Institute

EPFL

martin.schrimpf@epfl.ch @martin_schrimpf mastodon.social/@mschrimpf

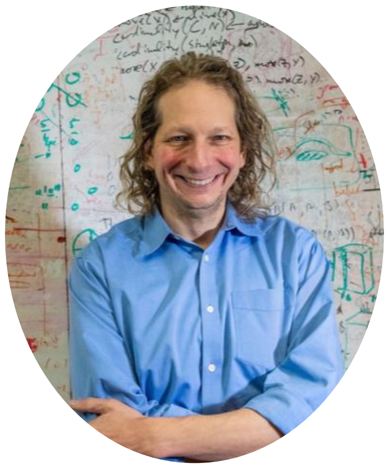

Idan Blank


Greta Tuckute


Carina Kauf

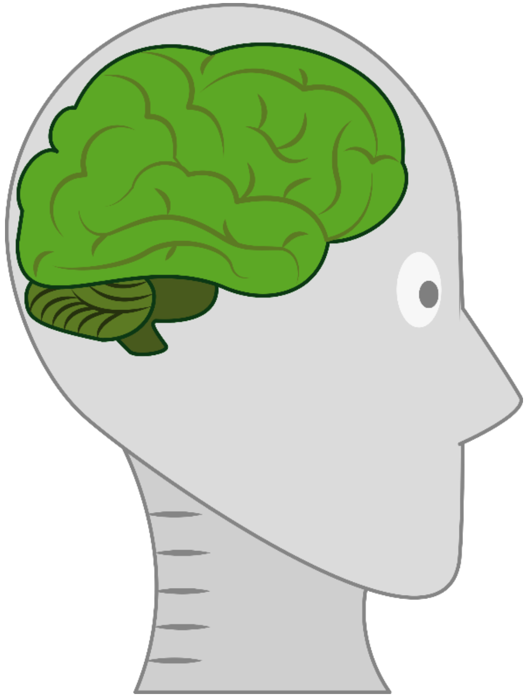

Eghbal Hosseini


Nancy Kanwisher

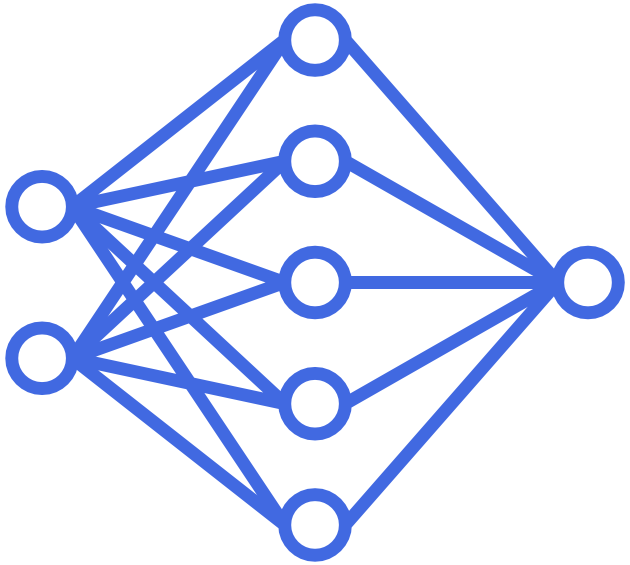

Josh Tenenbaum


Ev Fedorenko

<u>Goal</u> (broadly): Model Natural (Human) Intelligence and the Underlying Neural Mechanisms



$\approx$

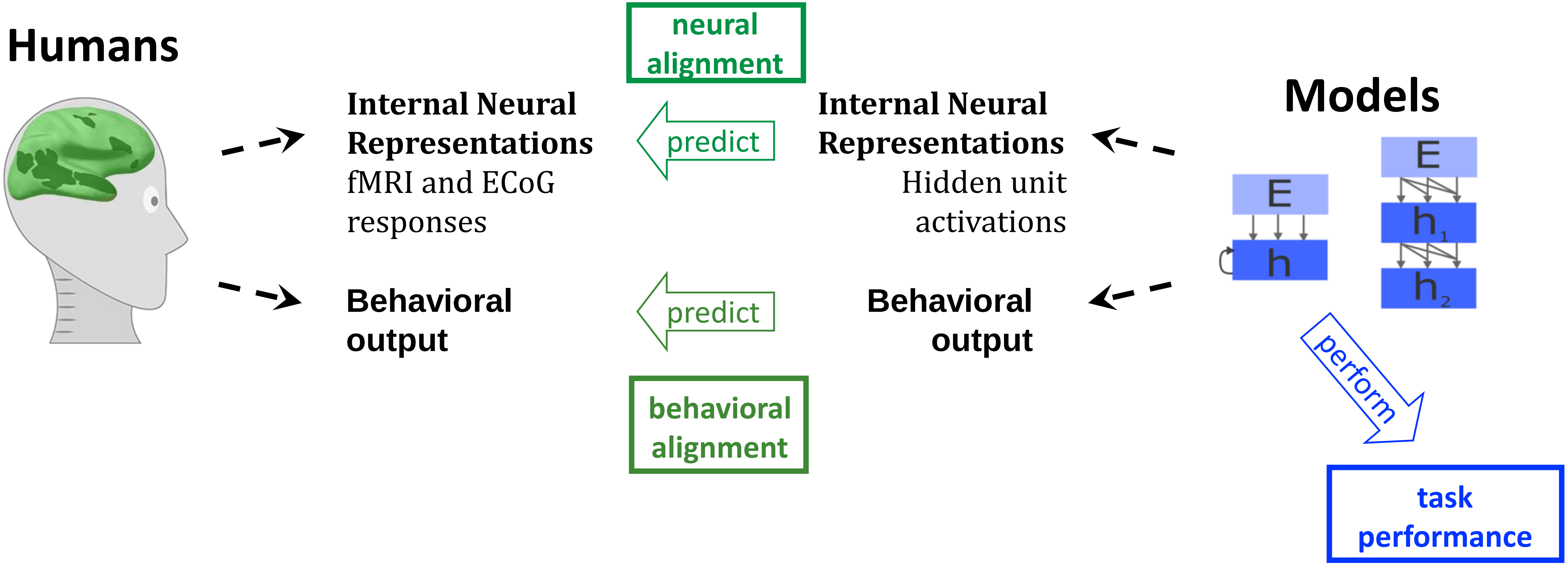


Mechanistic understanding of human intelligence

Next-generation intelligence algorithms

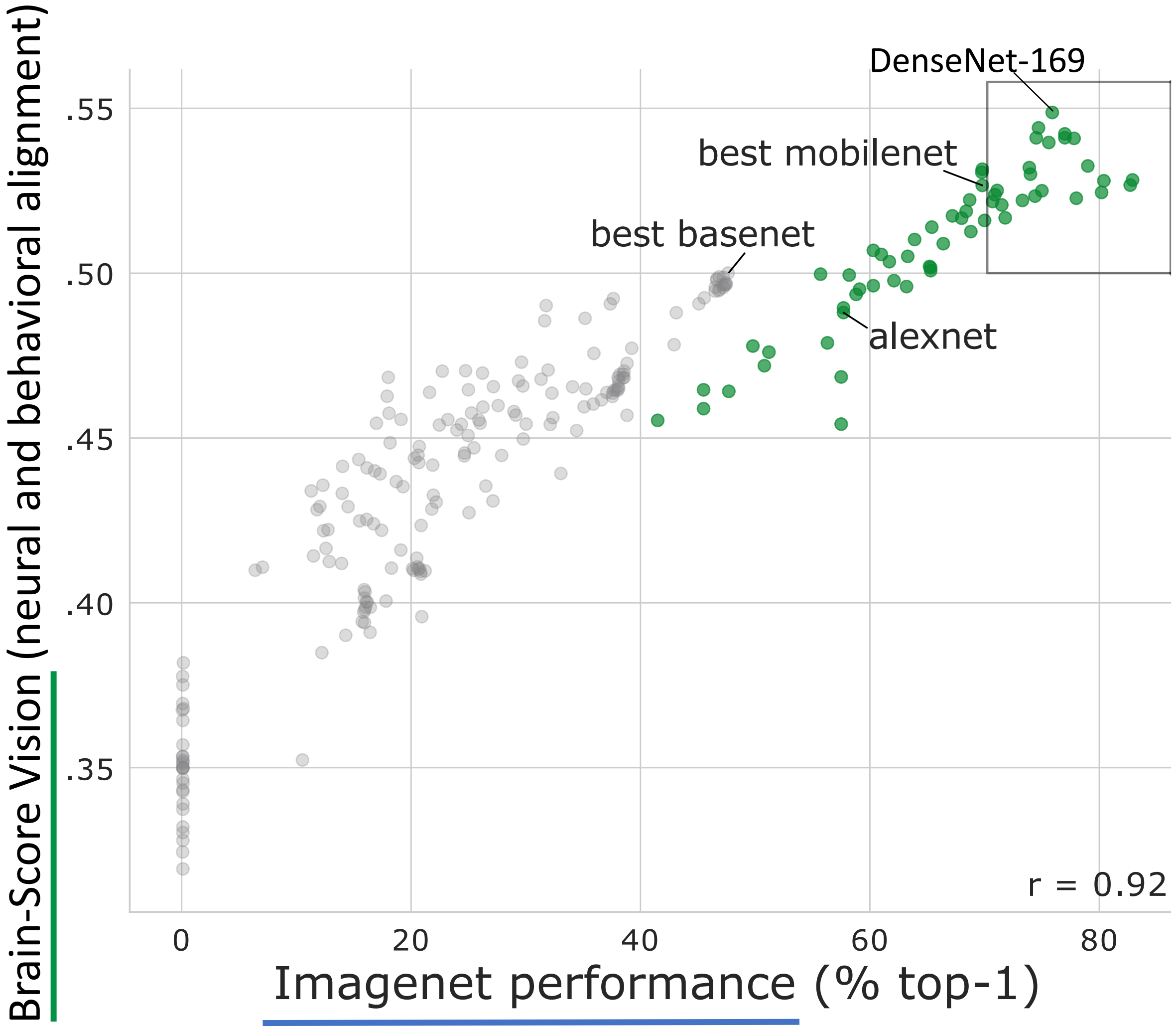
Future clinical applications

Goal (today): Model the Human Language System



Schrimpf et al. 2020, 2021

What kinds of models could align with the human language system?



In sensory cortex:

- Artificial Neural Networks (ANNs) are the leading class of models for explaining brain and behavior
- ANNs make predictions for any visual input and work well for real-world stimuli
- ANNs with higher task performance generally are more aligned to brain and behavior

Schrimpf, Kubilius*, et al. 2018 | Kubilius*, Schrimpf*, et al. 2019*

see also www.brain-score.org | Yamins, Hong*, et al. 2013, 2014 | Khaligh-Razavi & Kriegeskorte 2014 | Zhuang et al. 2017 | Kell et al. 2018*

Modeling **higher cognition**

Perception

Language

High-level reasoning

Artificial neural networks have worked well in modeling sensory cortex – could they also predict higher cognition?

The human language network

<u>working definition:</u>
a set of **left-lateralized** regions on the lateral surfaces of **frontal** and **temporal** cortex that support **high-level** language processing.

Language	>	Perceptually matched control
Sentences	>	Lists of nonwords

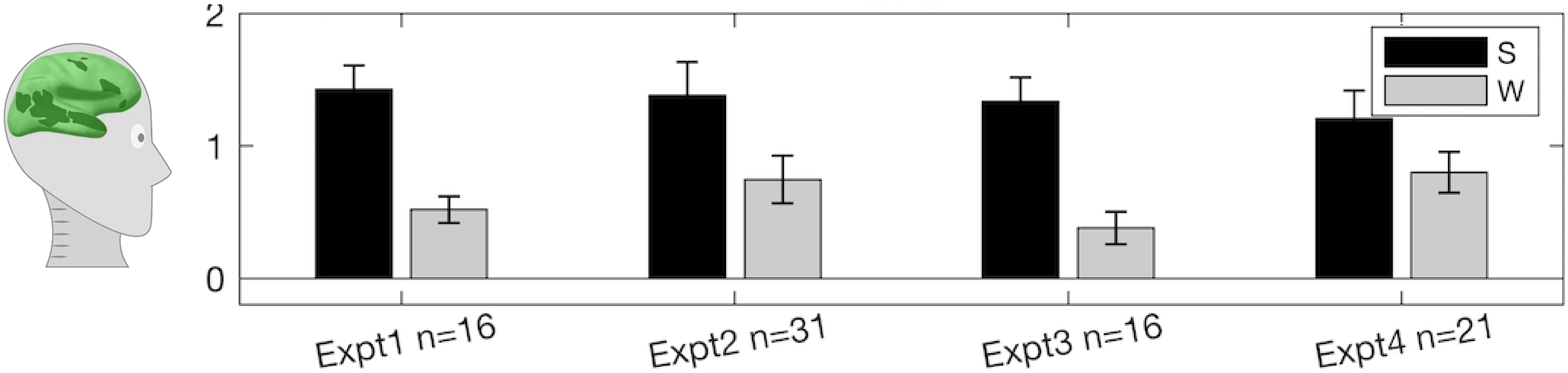
Fedorenko and Thompson-Schill 2014 *Braga, DiNicola and Buckner 2019*

The human language network

the dog is taking
a bath

>

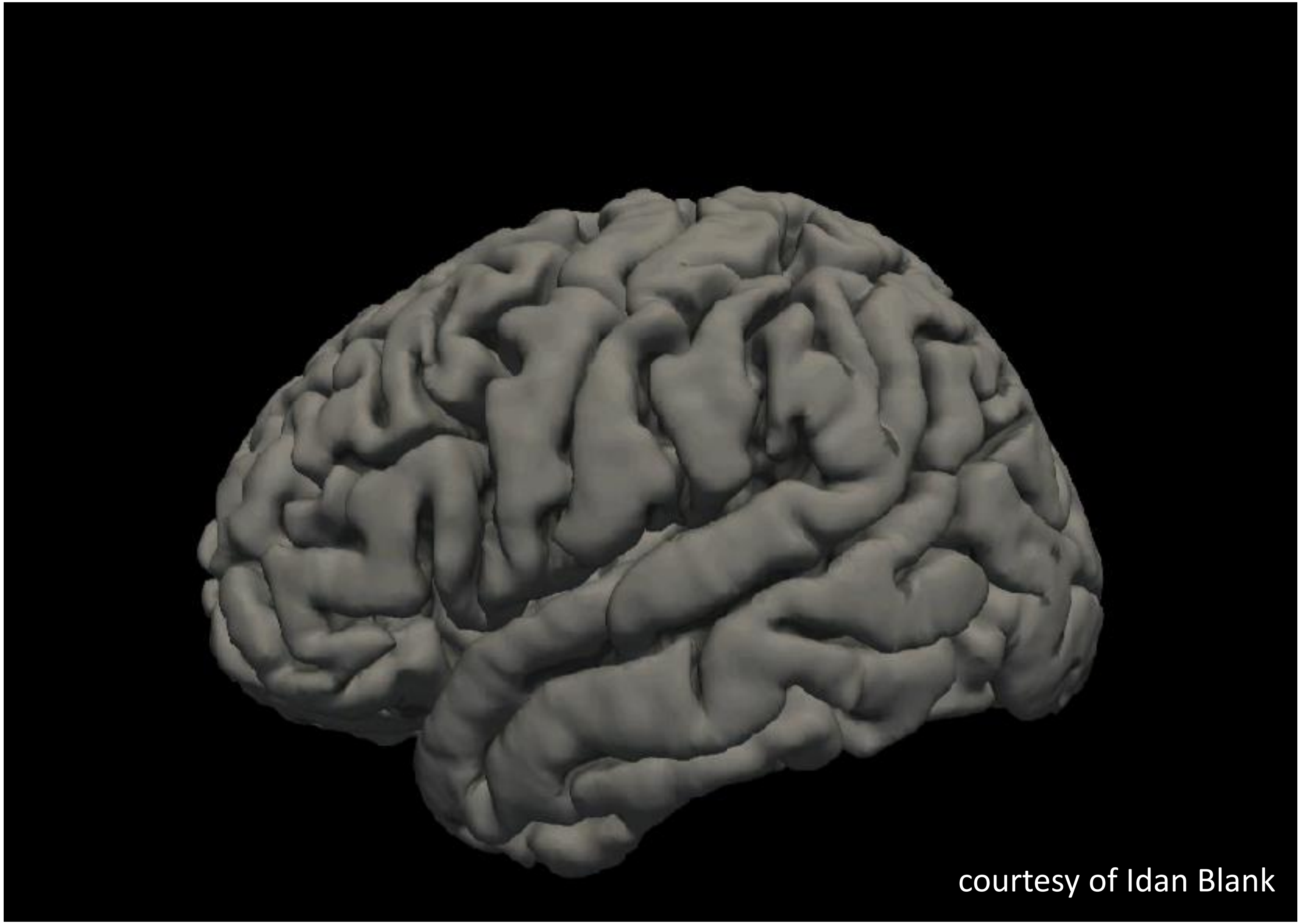
dap drello smop ub
plid kav



Key signature: stronger response to sentences than lists of unconnected words

Fedorenko, Behr and Kanwisher 2011 | Fedorenko et al. 2020

The human language network



What are the mechanisms underlying human language comprehension?

```
the dog is taking
a bath
```

"meaning"

What are the mechanisms underlying human language comprehension?

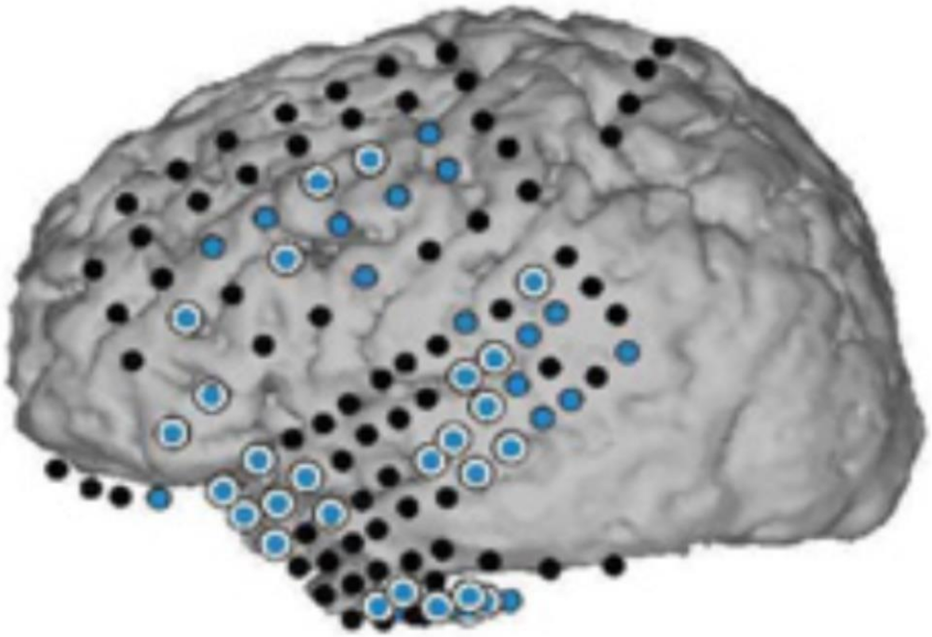
stimulus

observed data

Want to find the neural mechanisms that predict our data

Data target: human neural recordings

fMRI



ECoG

Data target: human neural recordings

Pereira et al. 2018 fMRI

627 sentences x 13,517 voxels in 10 subjects

Beekeeping encourages the conservation of local habitats. | It is in every beekeeper's interest …

Fedorenko et al. 2016 ECoG

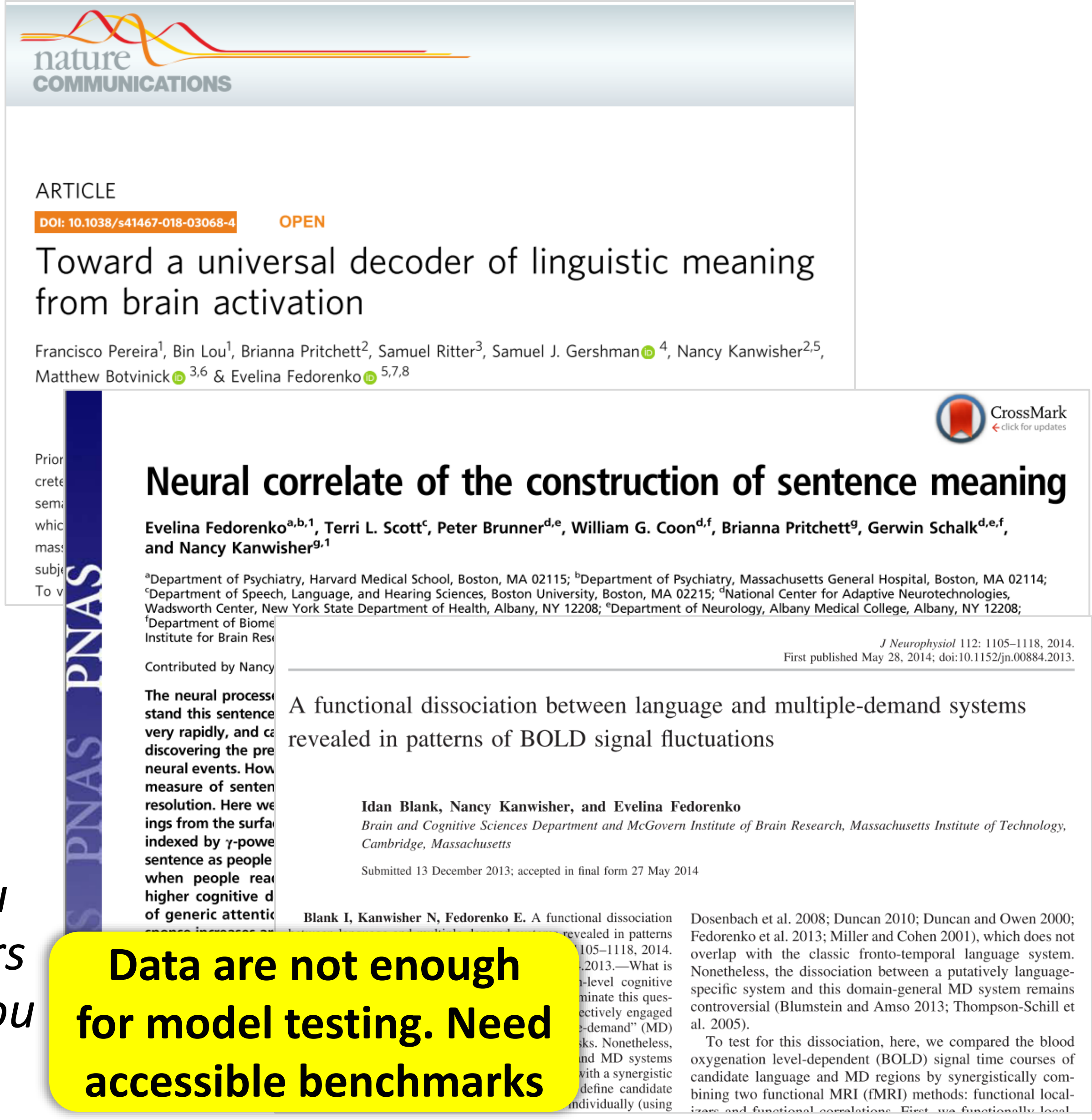
416 words x 97 electrodes in 5 subjects

ALEX | WAS | TIRED | SO | HE | TOOK | A | NAP

Blank et al. 2014 fMRI

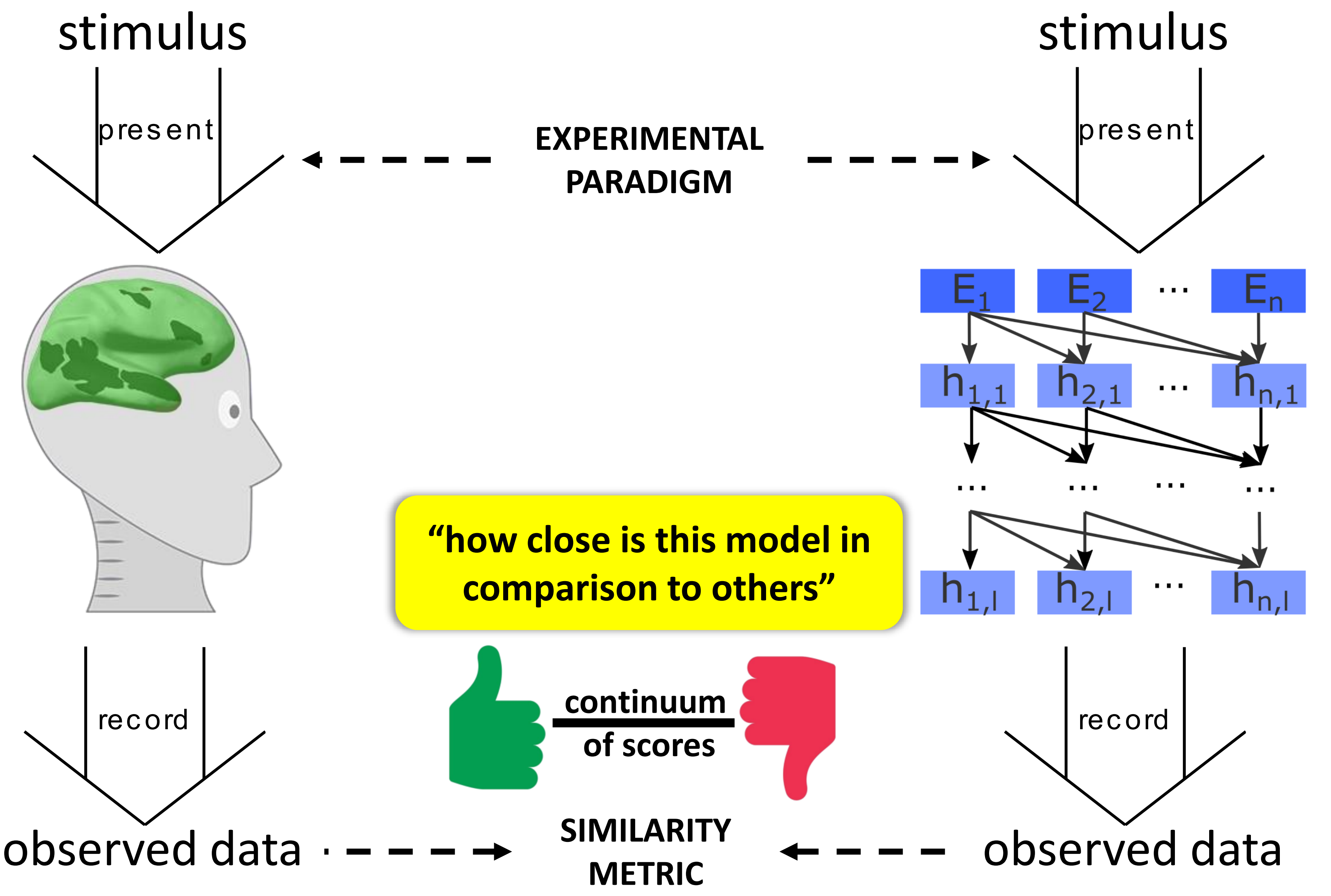
1,317 story fragments x 60 fROIs in 5 subjects

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, | …



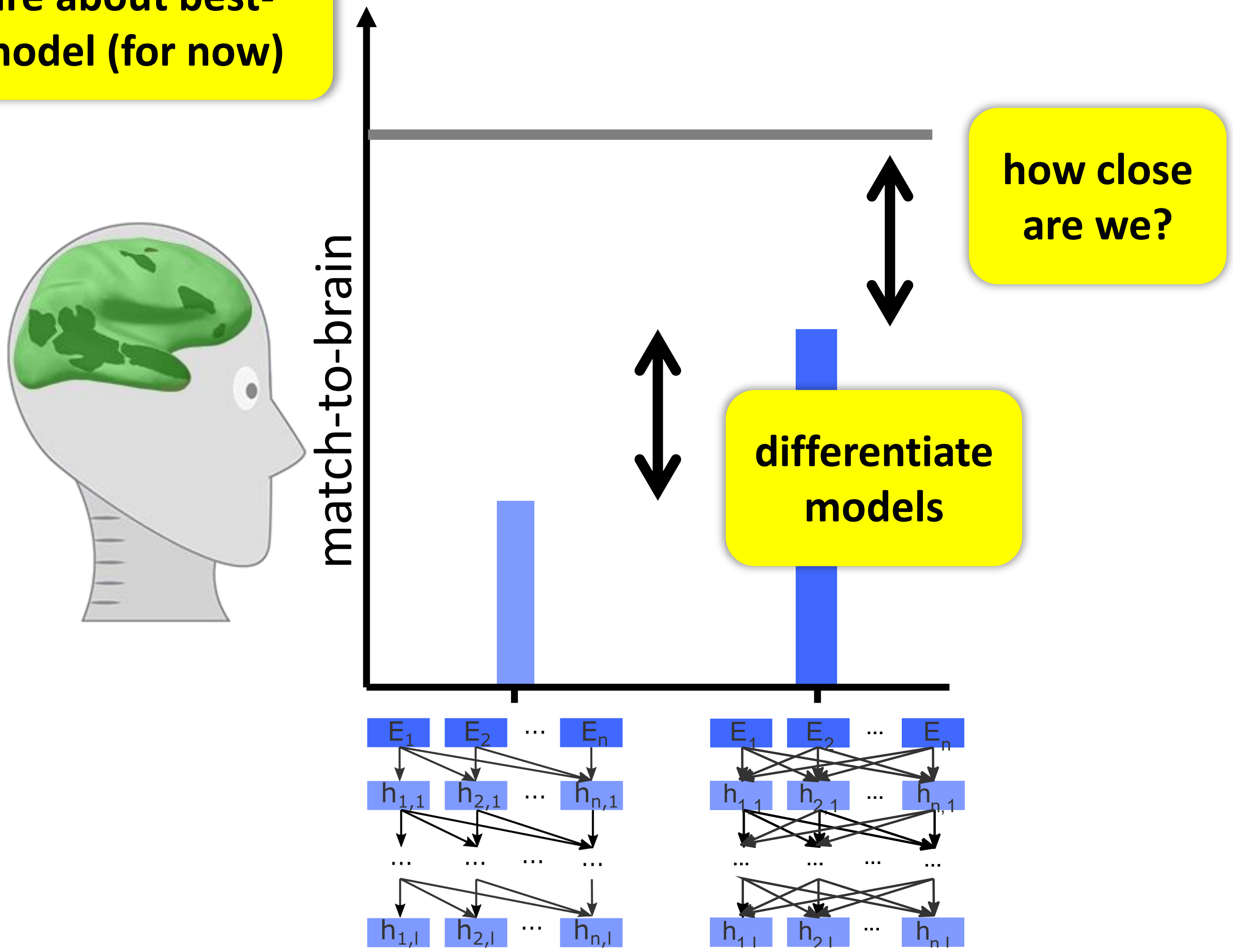
Data are not enough for model testing. Need accessible benchmarks

Quantifying match-to-brain: Benchmarking



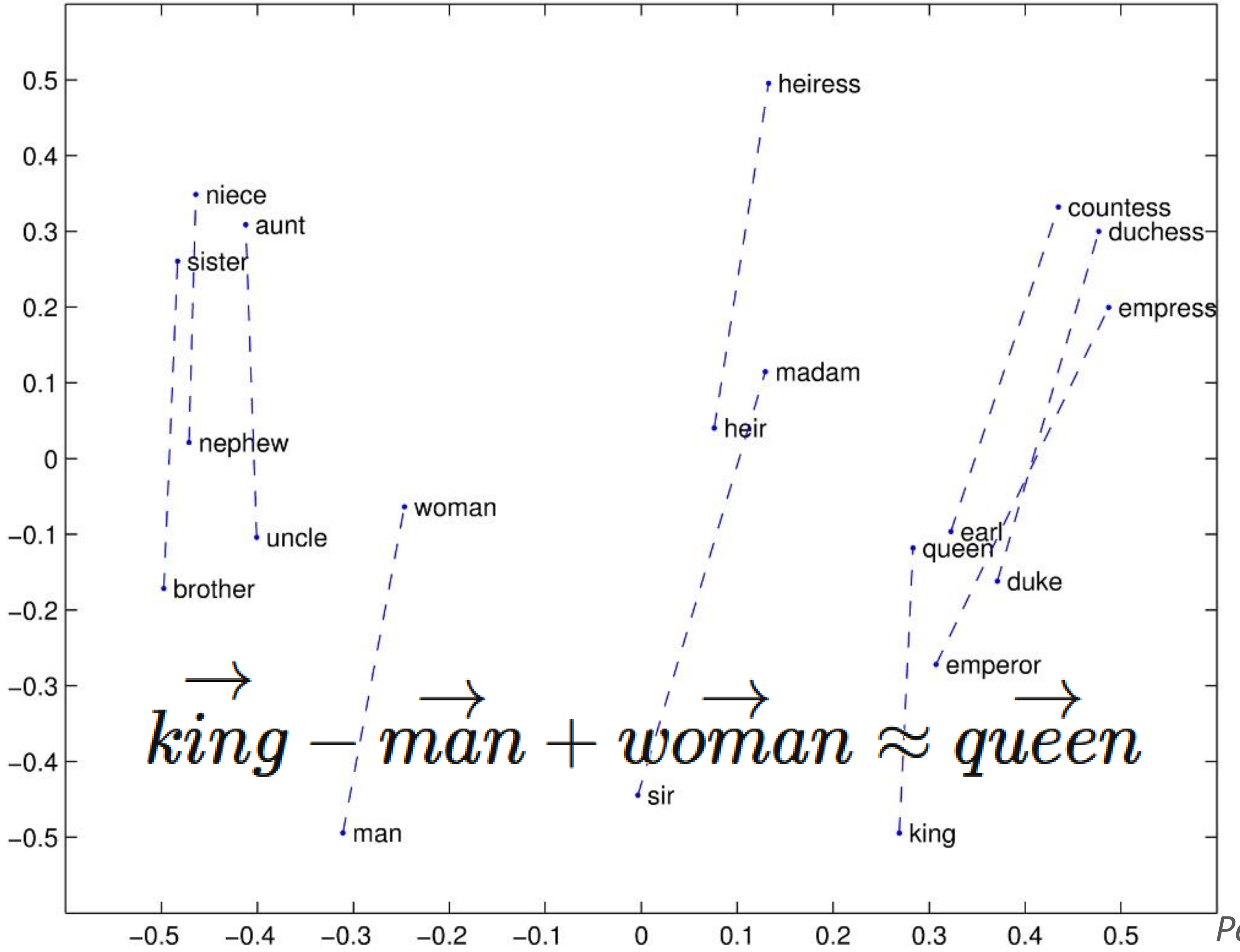
Quantifying match-to-brain: Benchmarking

We only care about best-matching model (for now)



Models tested (n=43)

Embedding type models: GloVe, word2vec, topicETM

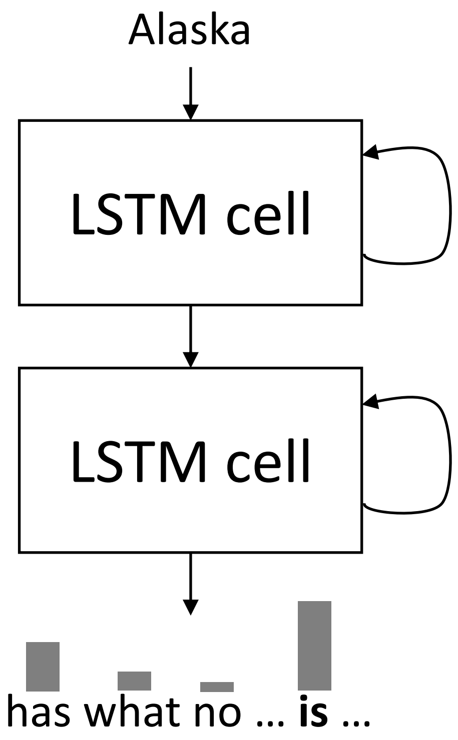


Pennington, Socher, Manning 2014

Models tested (n=43)

Embedding type models: GloVe, word2vec, topicETM

Recurrent networks: skip-thoughts, LSTM lm_1b



Language Modeling

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

Image from https://www.quora.com/What-is-a-masked-language-model-and-how-is-it-related-to-BERT

Jozefowicz, vinyals, Schuster, Shazeer, Wu 2016

Models tested (n=43)

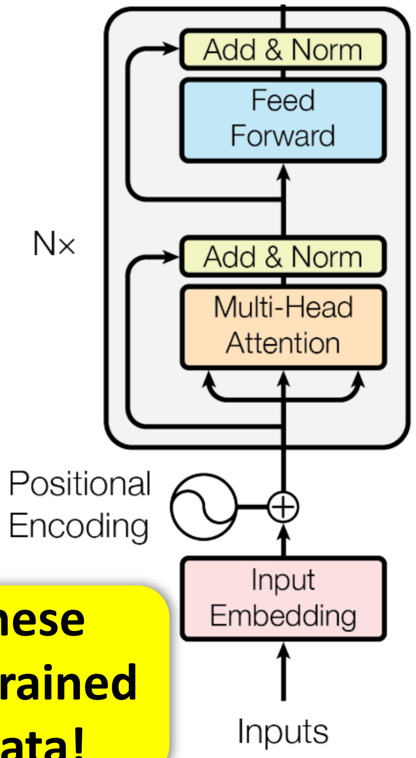
Embedding type models: GloVe, word2vec, topicETM

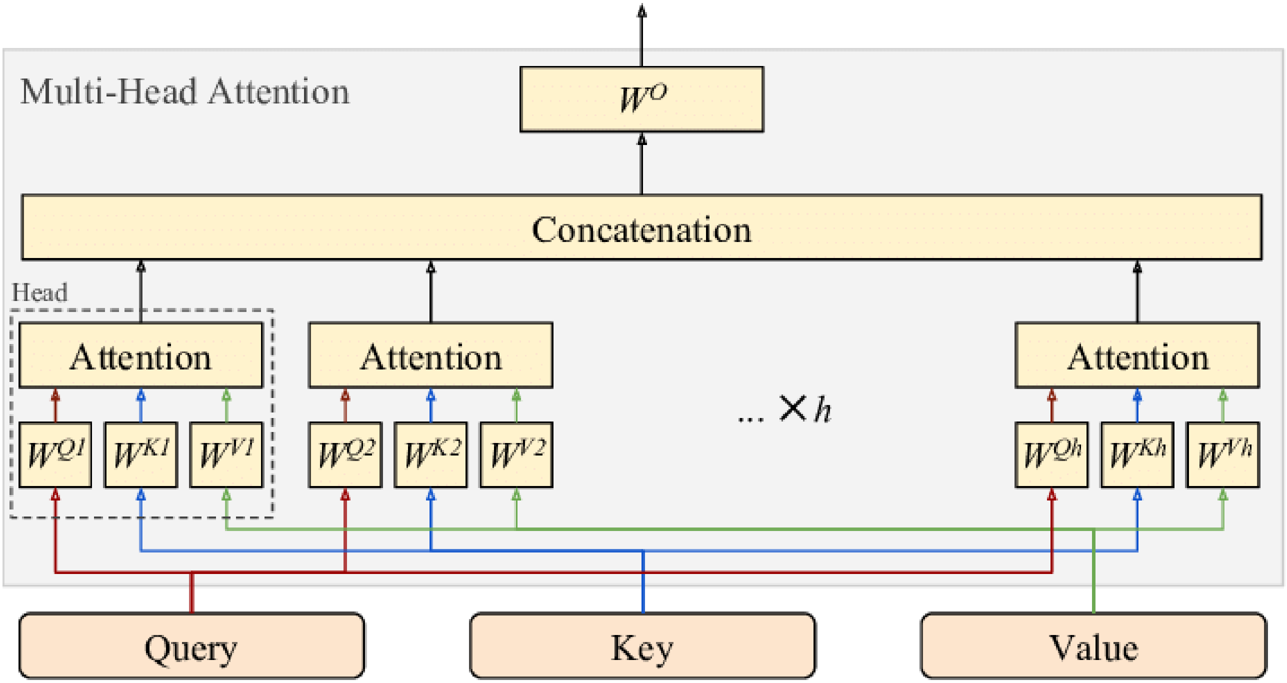
Recurrent networks: skip-thoughts, LSTM lm_1b

Transformers

- BERTs
- RoBERTas
- XLMs
- Transformer-XLs
- XLNets
- CTRL
- T5s
- AlBERTs
- GPTs

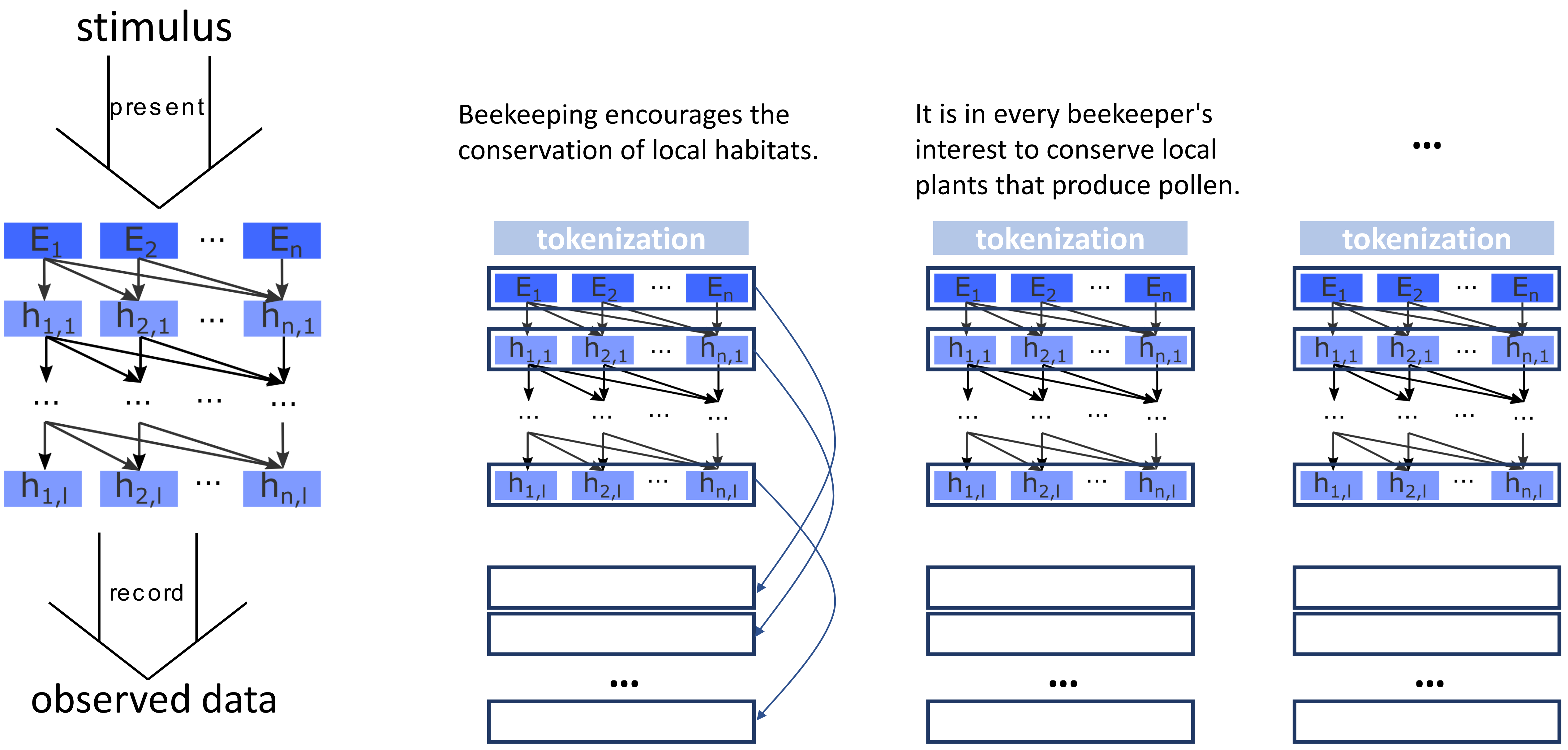
None of these models are trained on brain data!



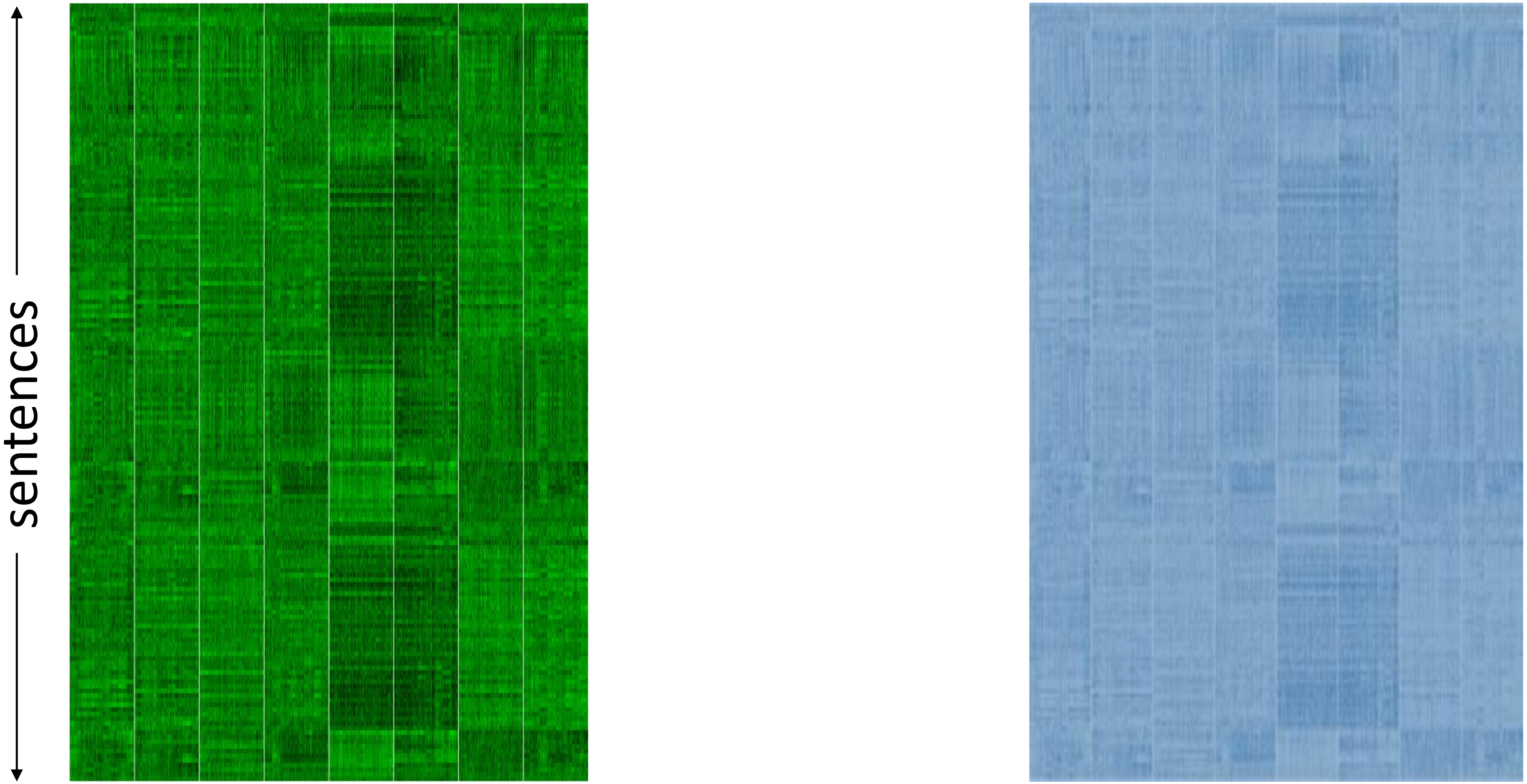


e.g. Pennington et al. 2014 | Jozefowicz et al. 2016 | Vaswani, Shazeer*, Parmar*, Uszkoreit*, Jones*, Gomez*, Kaiser*, Polosukhin* 2017 | Devlin et al. 2018*

Treating models as experimental subjects



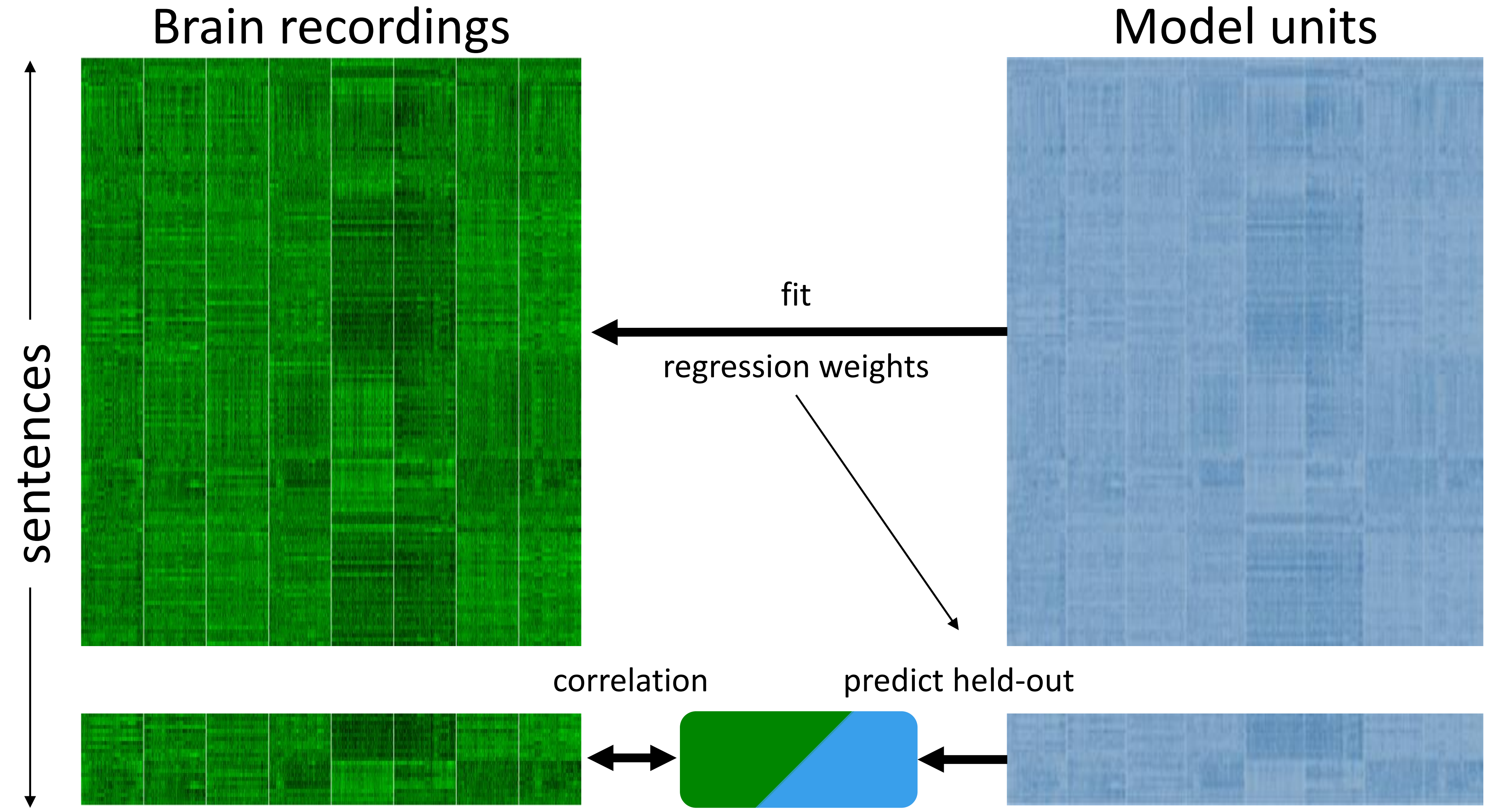
Neural benchmarks



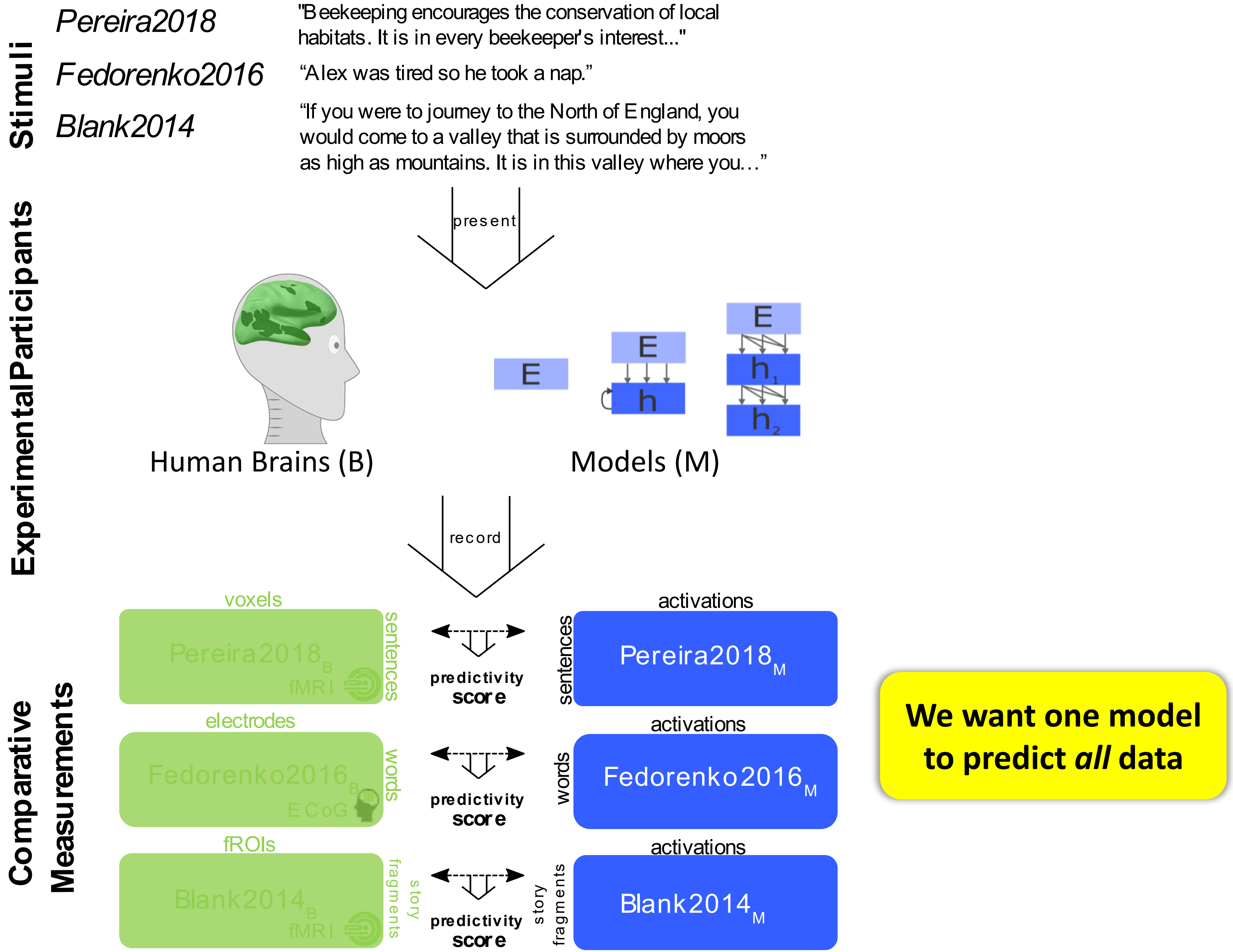
Neural benchmarks

similarity metric

neural predictivity

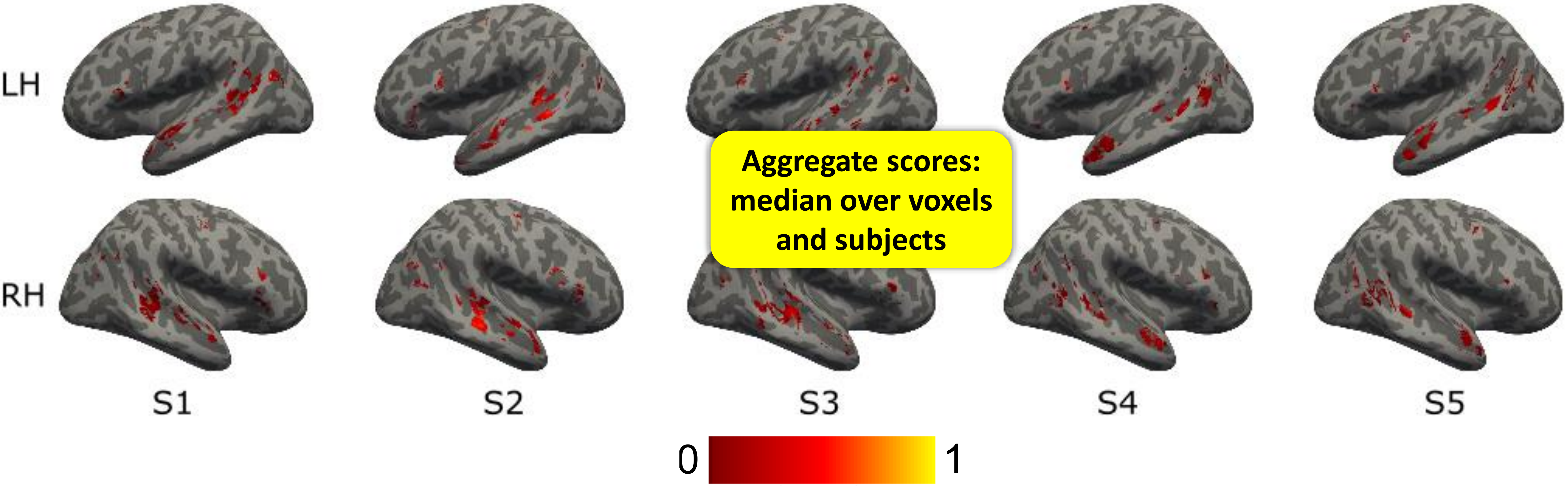


Yamins*, Hong*, et al. (PNAS 2014)   Schrimpf*, Kubilius*, et al. (bioRxiv 2018)


Stimuli
Pereira2018
"Beekeeping encourages the conservation of local habitats. It is in every beekeeper's interest..."
Fedorenko2016
"Alex was tired so he took a nap."
Blank2014
"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…"
present
Experimental Participants
Human Brains (B)
Models (M)
E
h
h1
h2
record
Comparative Measurements
voxels
Pereira2018B
fMRI
sentences
predictivity score
activations
Pereira2018M
electrodes
Fedorenko2016B
ECoG
words
Fedorenko2016M
fROIs
Blank2014B
story fragments
Blank2014M
We want one model to predict all data

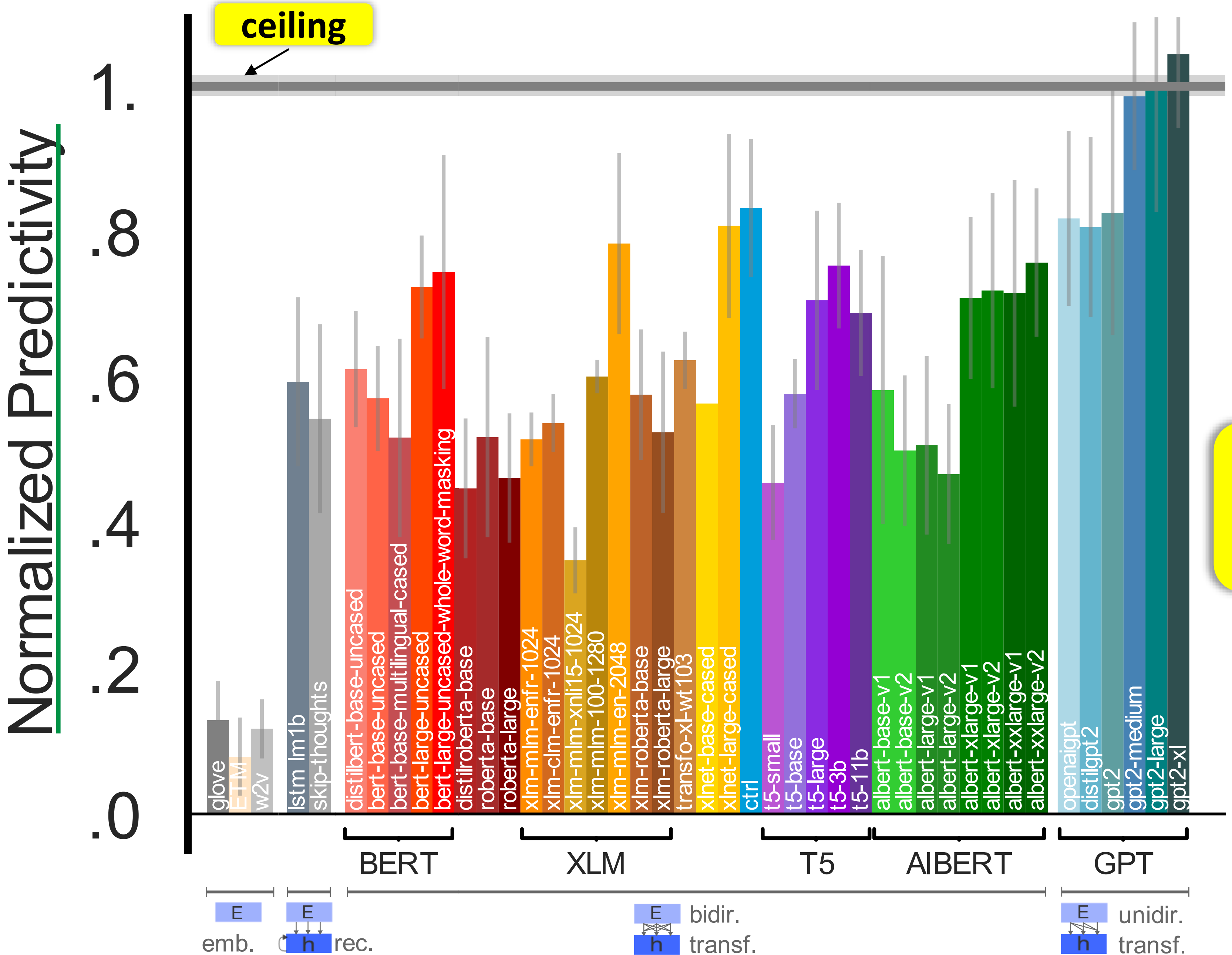
GloVe voxel-wise predictivity scores

← GloVe



Aggregate scores: median over voxels and subjects

Certain language models predict human language recordings

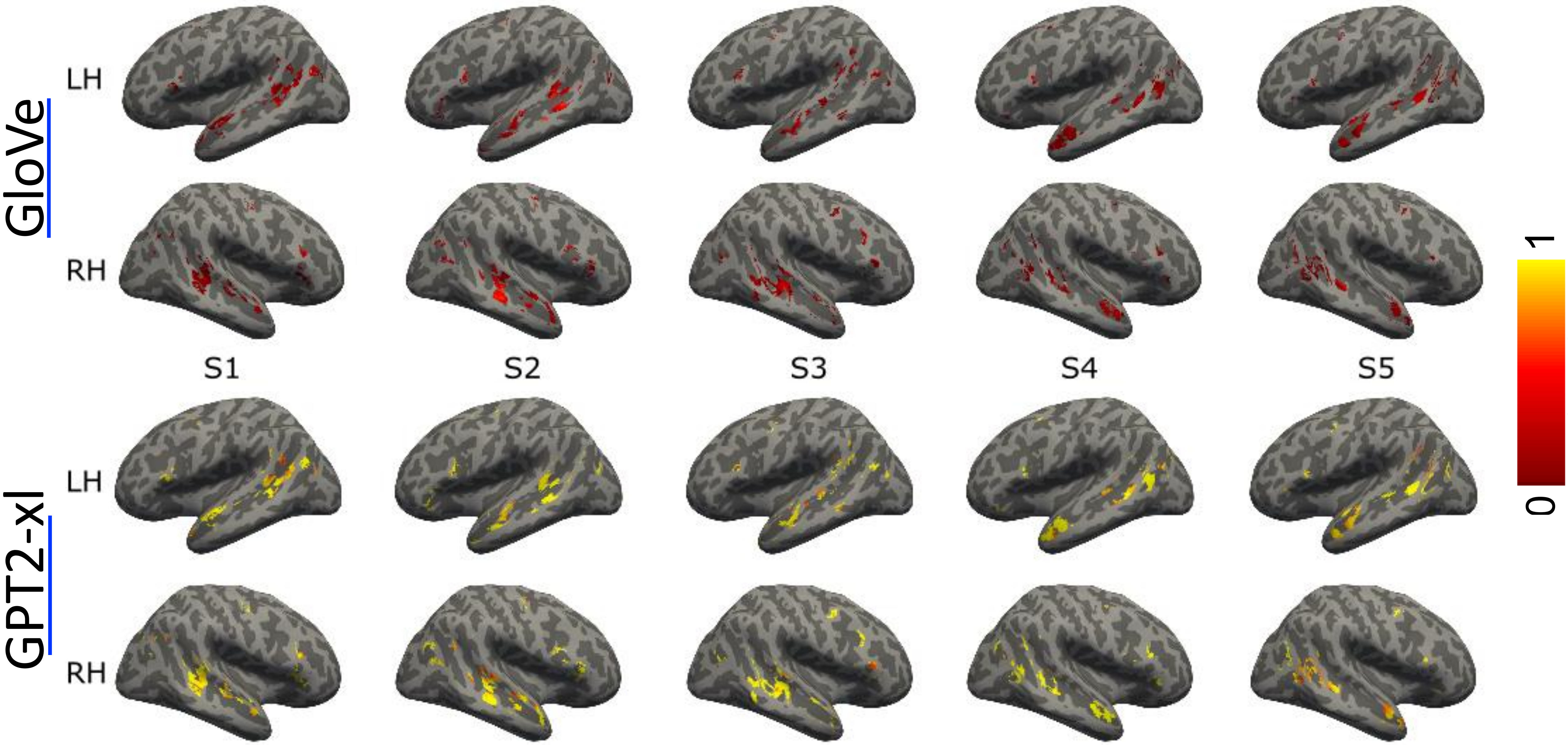


gpt2-xl hits our estimated ceiling for this benchmark

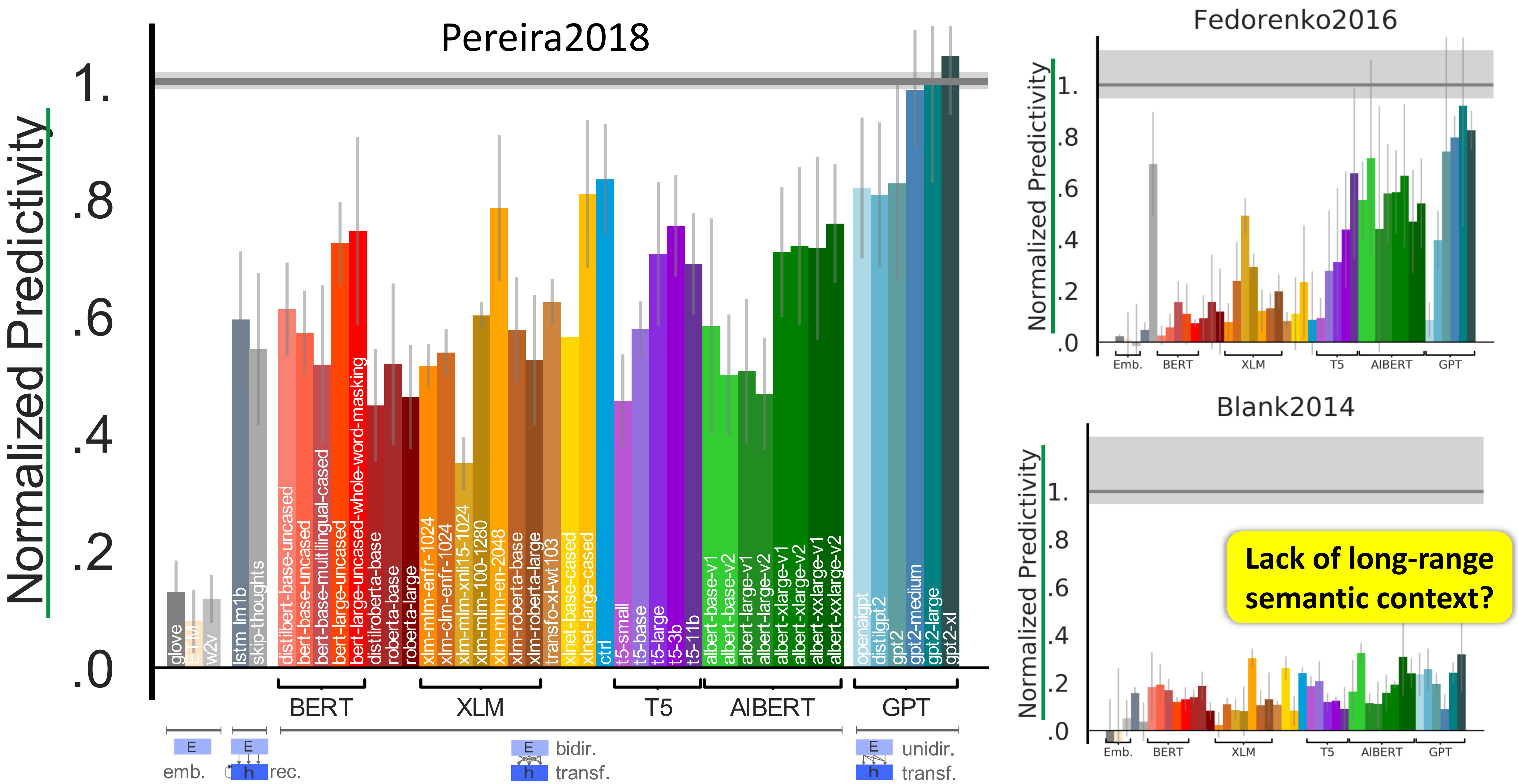
Small differences can lead to very different brain predictivities, warranting a full survey

Jain & Huth 2018
Gauthier & Ivanova 2018
Jat et al. 2019
Toneva & Wehbe 2019
Gauthier & Levy 2020
Wang et al. 2020

GPT2-xl accurately predicts a large portion of voxels

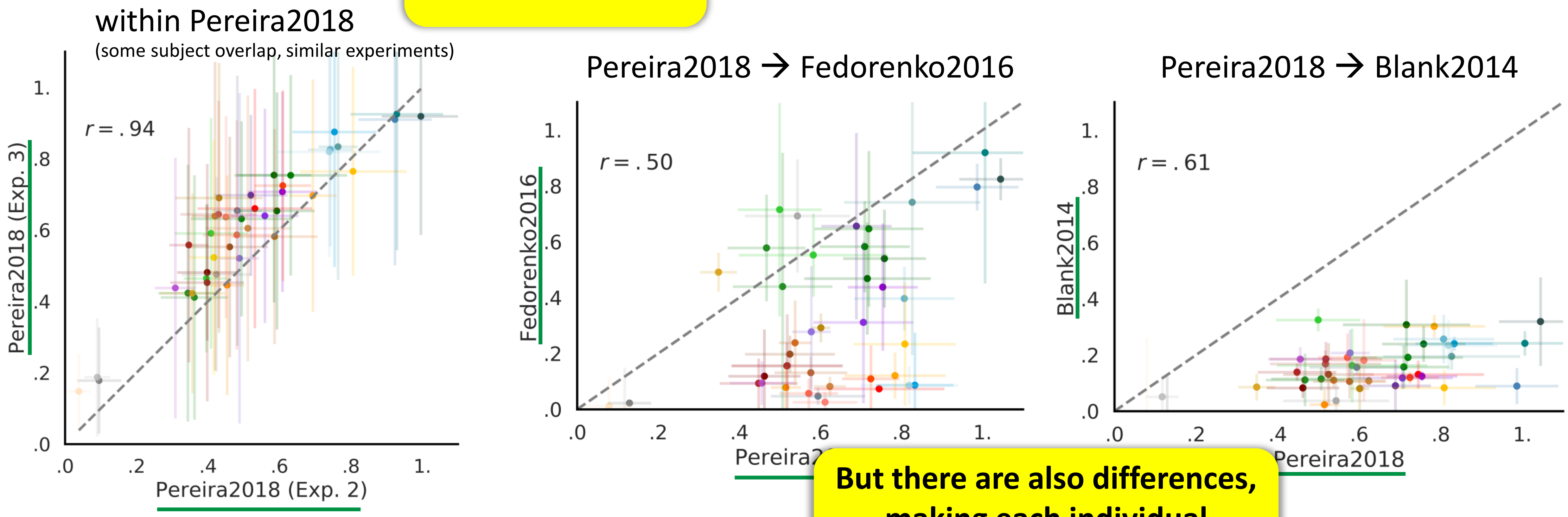


Language Models predict human language recordings



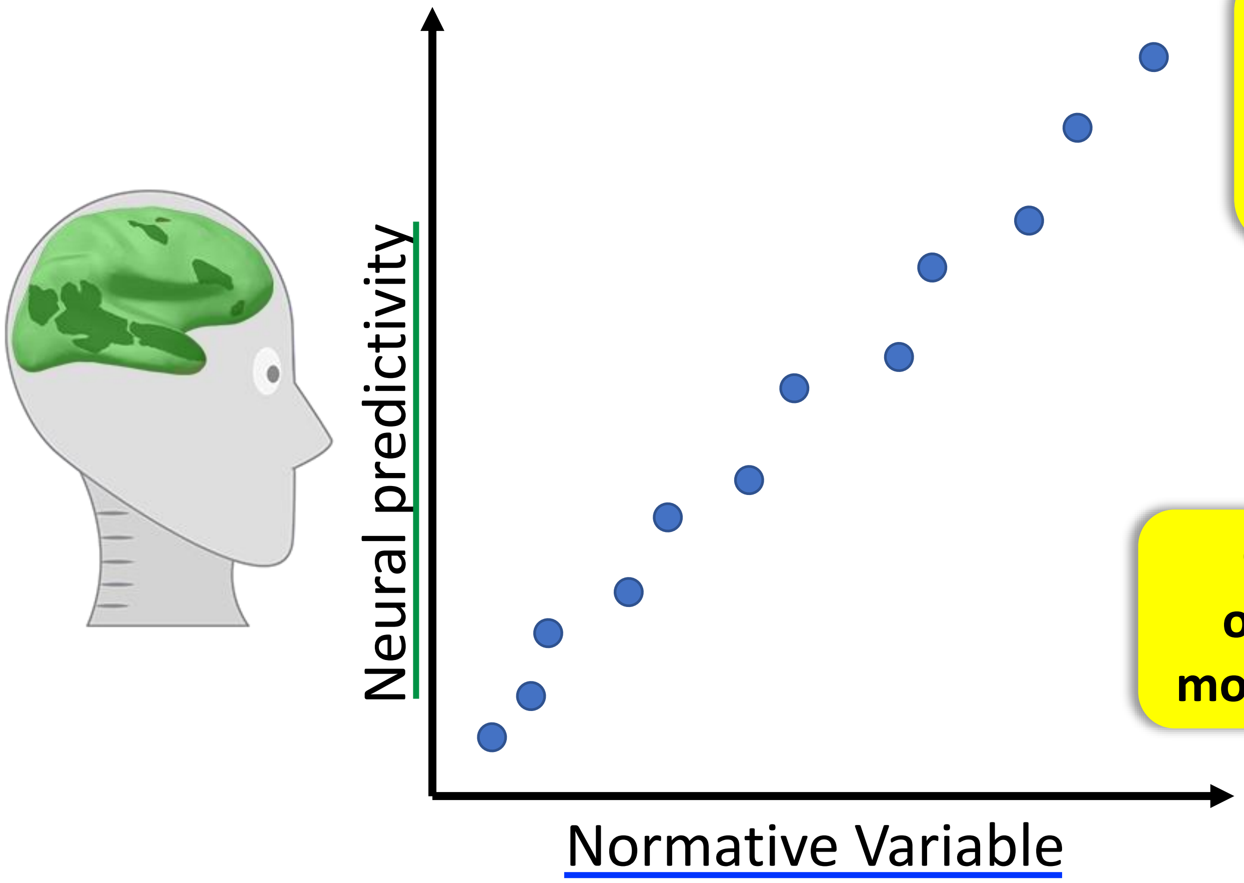
Control: model scores across benchmarks are correlated, although differences exist

Scores generalize to a good extent



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 *efficient*ly 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

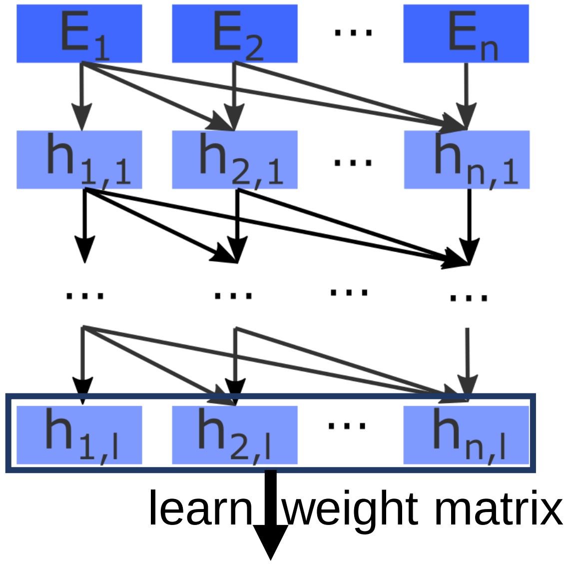
Alaska is about twelve times

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

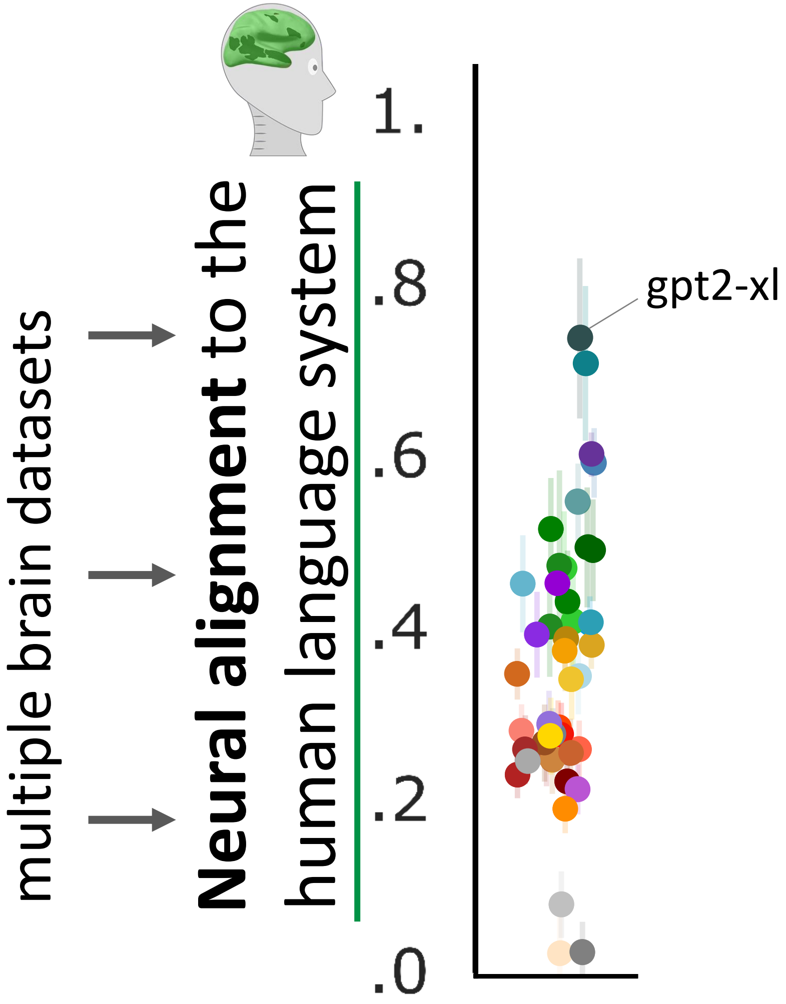


afternoon | alaska | animation | article | …

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

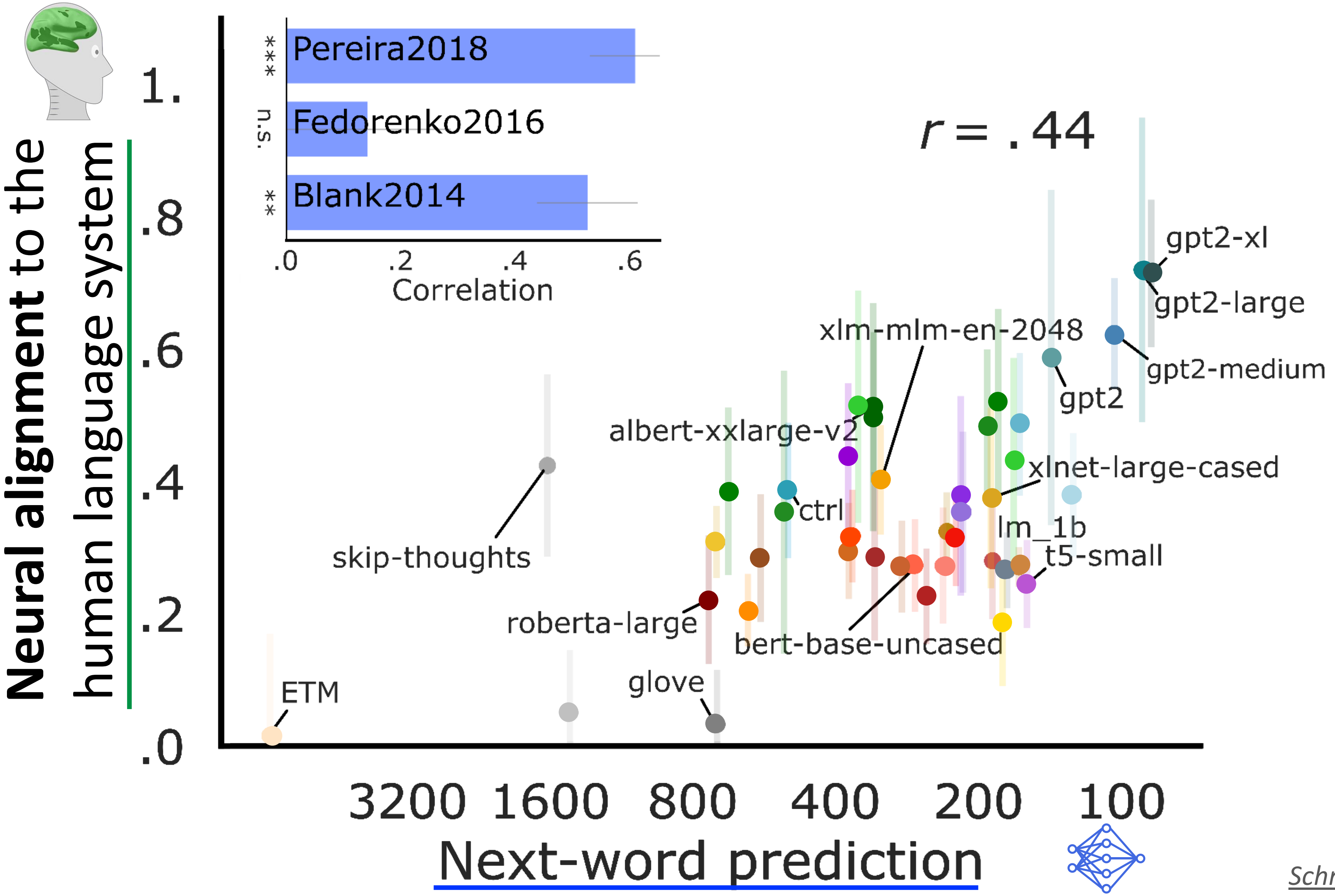
Merity et al. 2016

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)

What about other language tasks?

GLUE

9 “General Language Understanding Evaluation” tasks:

Sentence grammaticality (CoLa)

Sentence sentiment (SST-2)

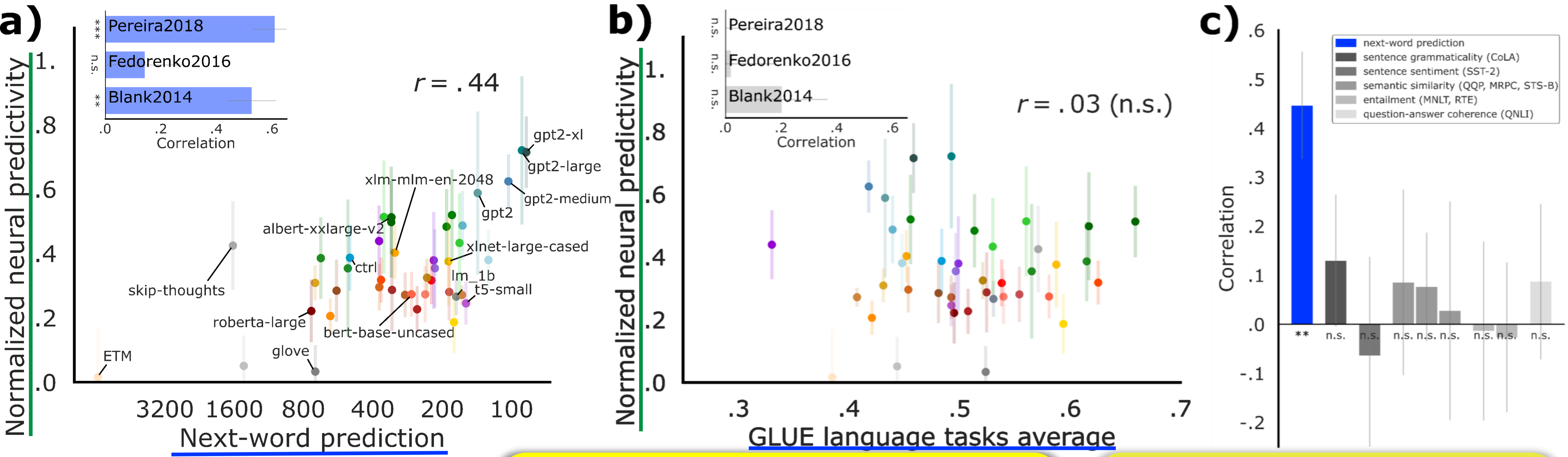
Semantic similarity (QQP, MRPC, STS-B)

Entailment (MNLT, RTE)

Question-answer coherence (QNLI)

Winograd (WNLI; ignored due to known issues)

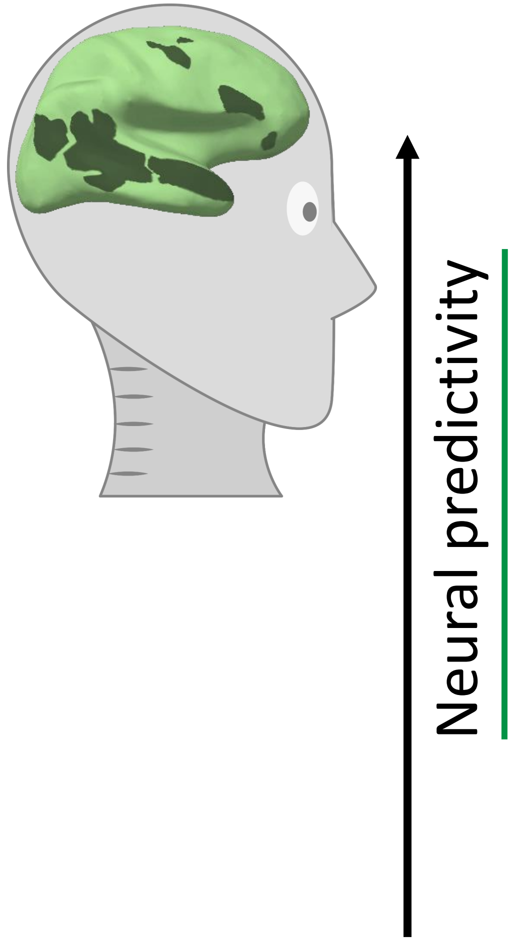
Next-Word Prediction performance **selectively** correlates with neural predictivity



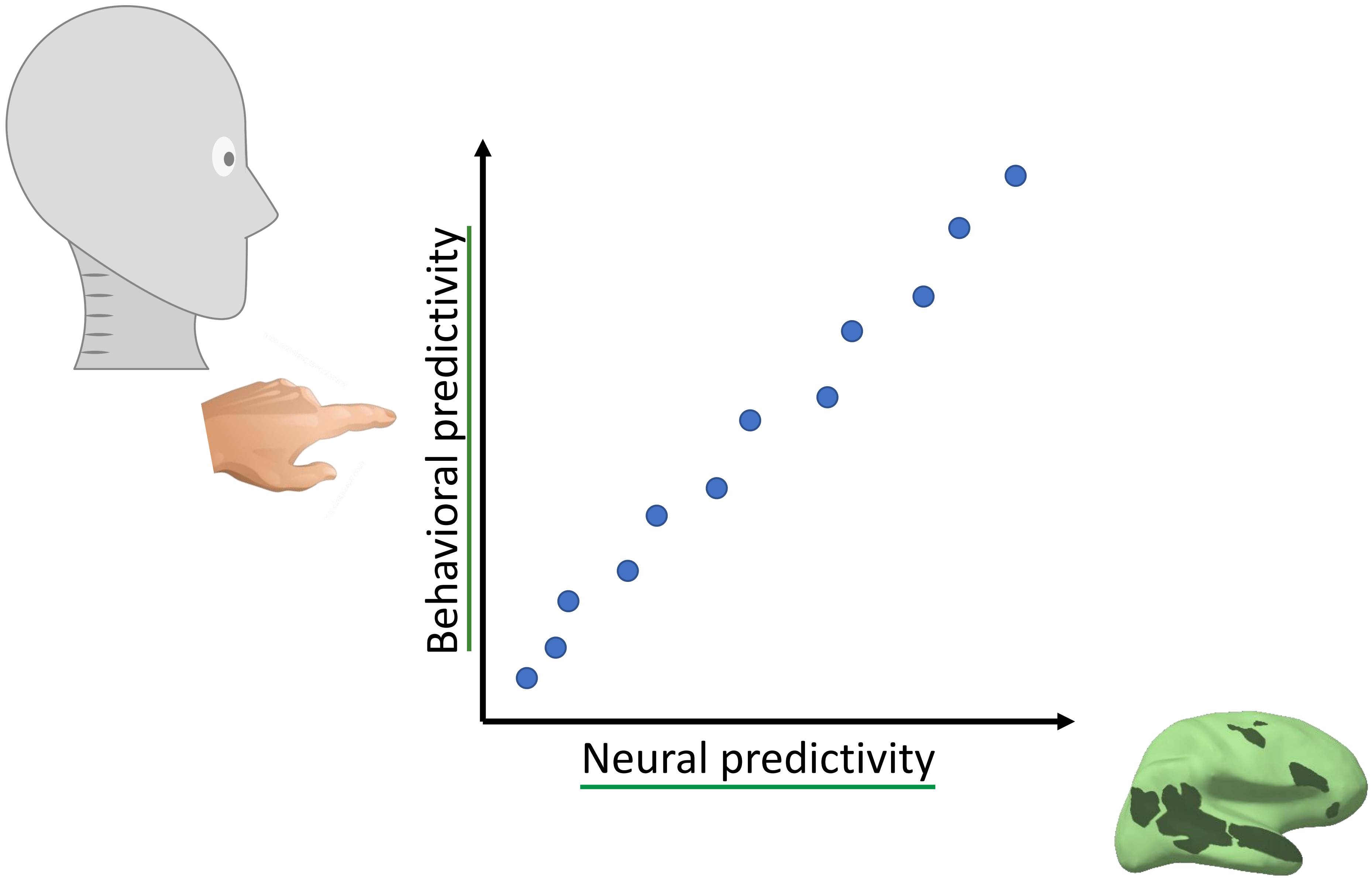
Online prediction may fundamentally shape language processing in the brain

Ongoing work: for high-performing models, reasoning capabilities (MMLU/BBH) also drive brain alignment *(Aw et al. in prep)*

Is any of this behaviorally relevant?



Is any of this behaviorally relevant?



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

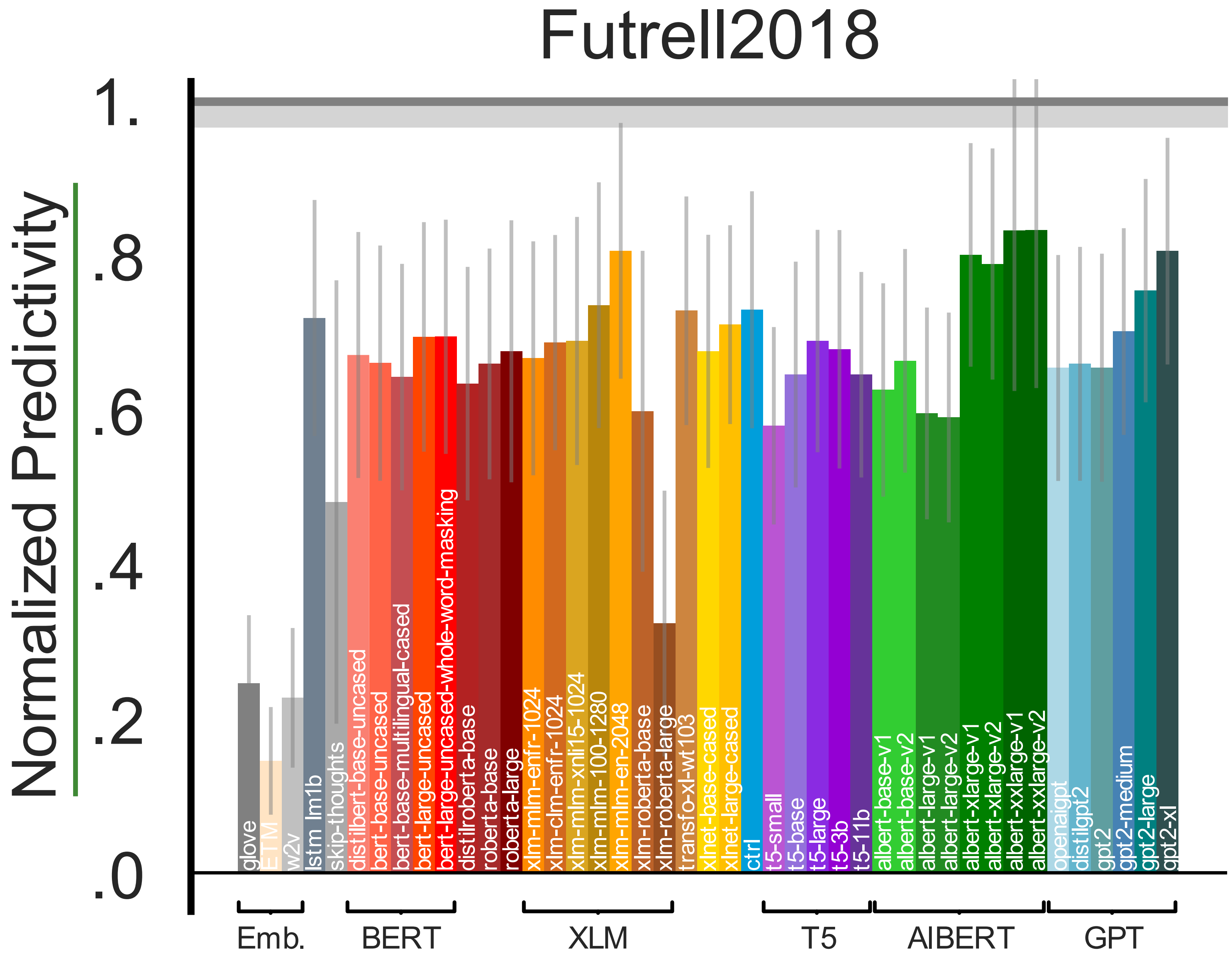
Richard Futrell[1], Edward Gibson[1], Harry J. Tily[2], Idan Blank[1], Anastasia Vishnevetsky[1], Steven T. Piantadosi[3], and Evelina Fedorenko[4,5]
[1]MIT Department of Brain and Cognitive Sciences [2]Netflix, Inc.
[3]University of Rochester Department of Brain and Cognitive Sciences
[4]Massachusetts 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.

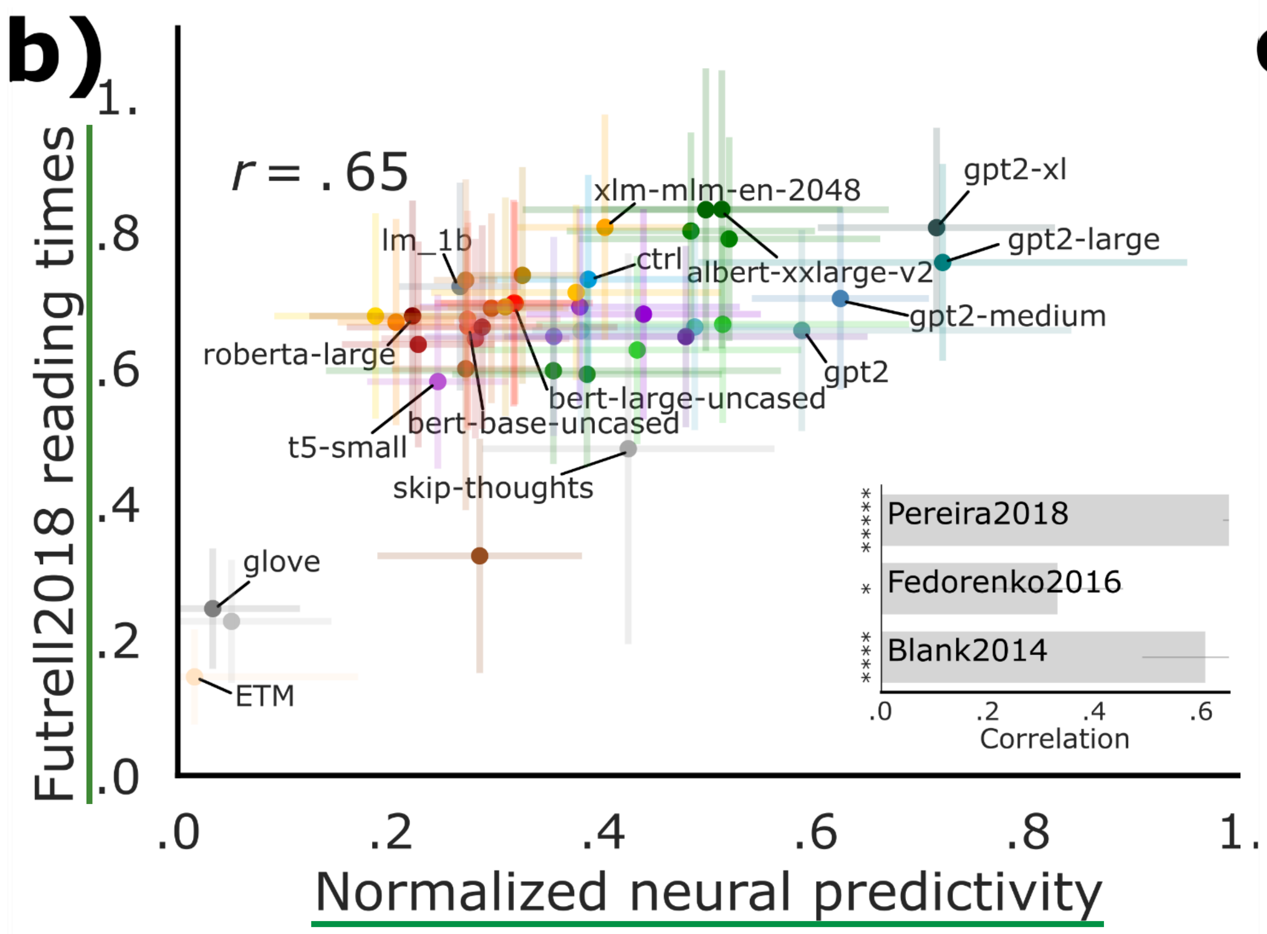
Keywords: Cognitive modeling, reading time, psycholinguistics

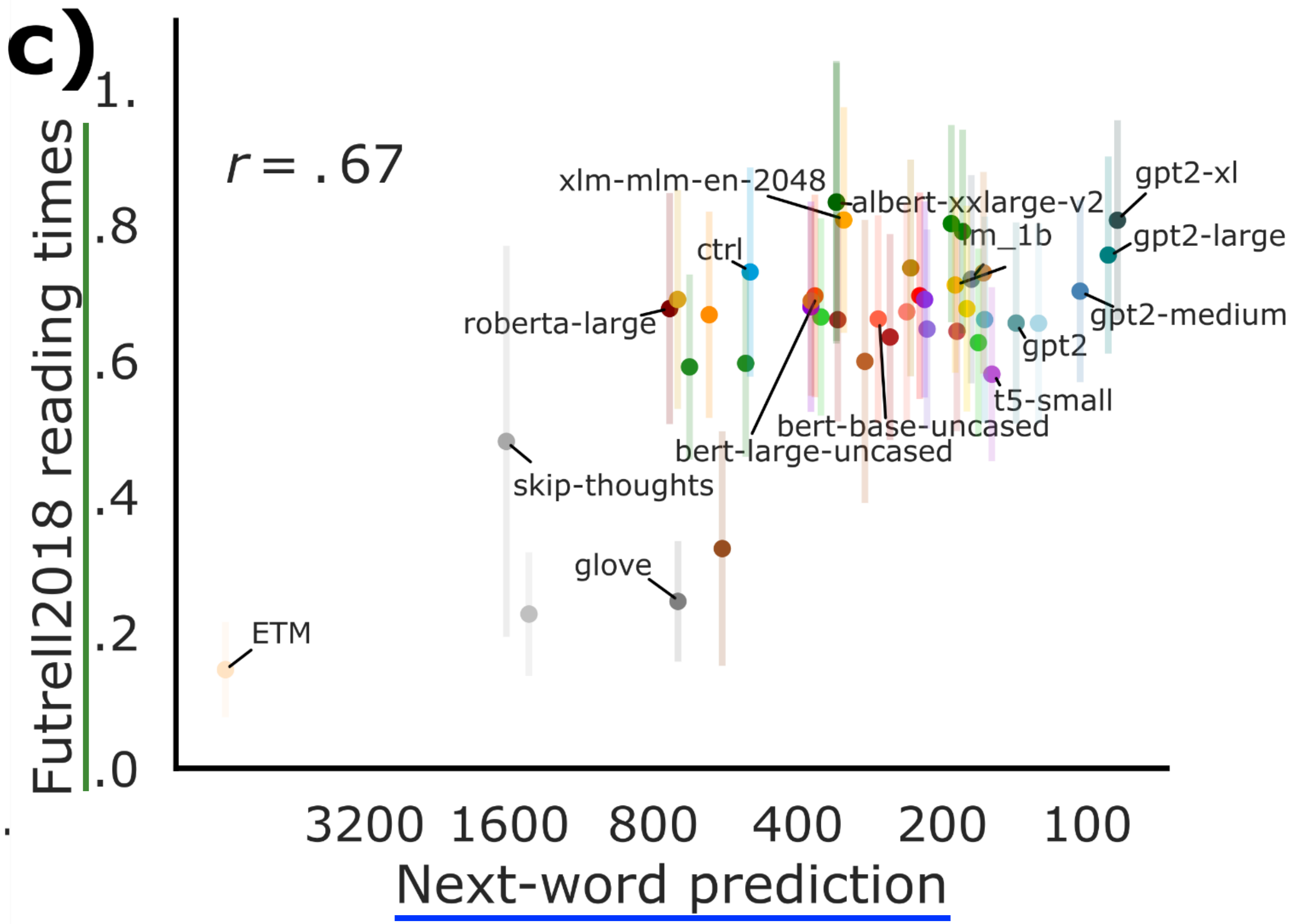
Behavioral scores

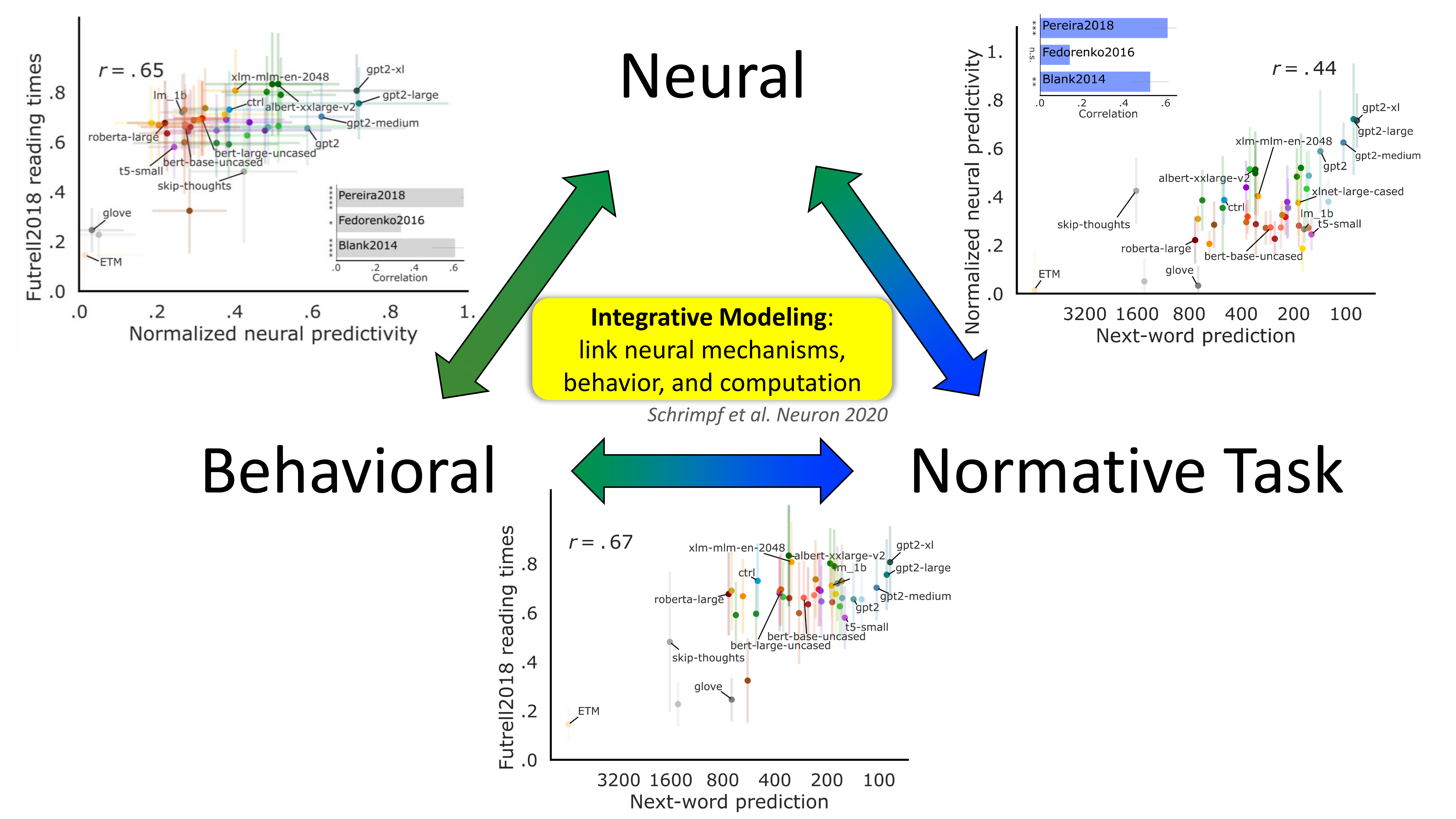


Neural scores correlate with Behavioral scores

Task scores correlate with Behavioral scores

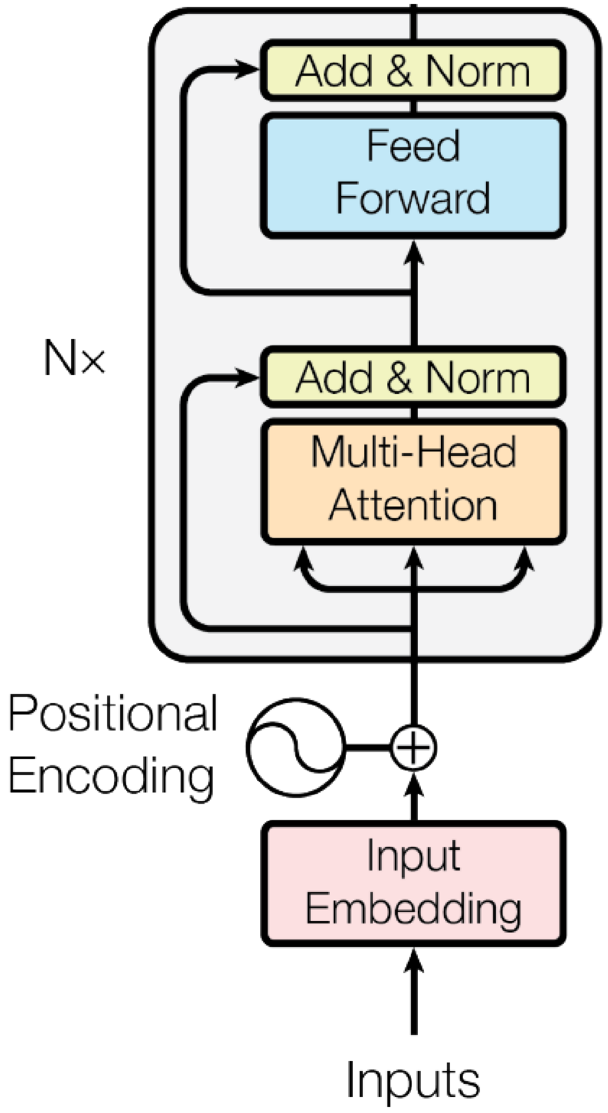






Neural
Behavioral
Normative Task
Integrative Modeling:
link neural mechanisms,
behavior, and computation
Schrimpf et al. Neuron 2020
r = .65
Futrell2018 reading times
Normalized neural predictivity
Pereira2018
Fedorenko2016
Blank2014
Correlation
r = .44
Normalized neural predictivity
Next-word prediction
r = .67
Futrell2018 reading times
Next-word prediction
gpt2-xl
gpt2-large
gpt2-medium
gpt2
xlm-mlm-en-2048
albert-xxlarge-v2
xlnet-large-cased
ctrl
lm_1b
t5-small
roberta-large
bert-large-uncased
bert-base-uncased
skip-thoughts
glove
ETM

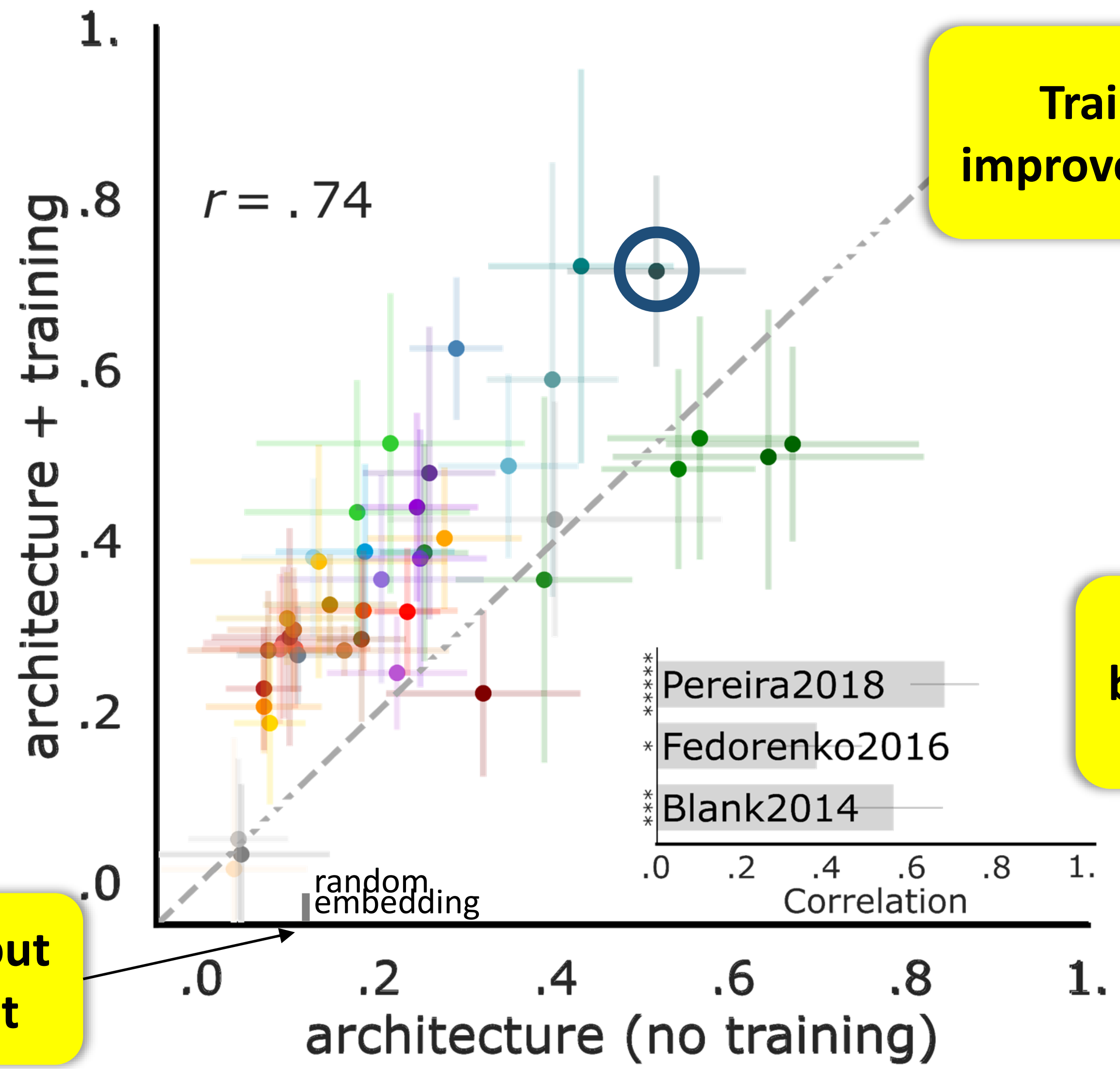
What is the relative importance of evolutionary and learning-based optimization?



Evolution $\simeq$ community optimization over architectural properties

Experience-dependent learning $\simeq$ updating of weights over training

Architecture substantially contributes to models' brain predictivity

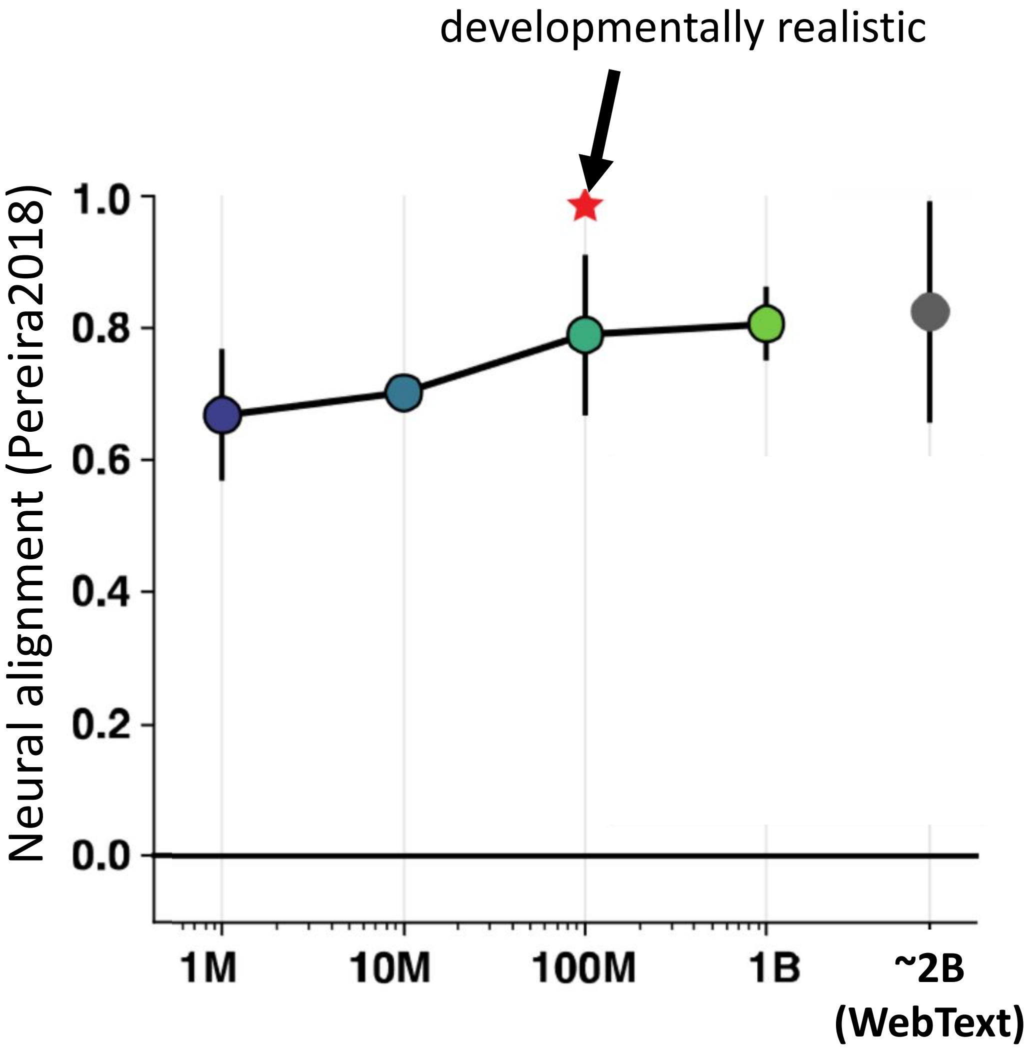


Training generally improves scores by ~53%

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

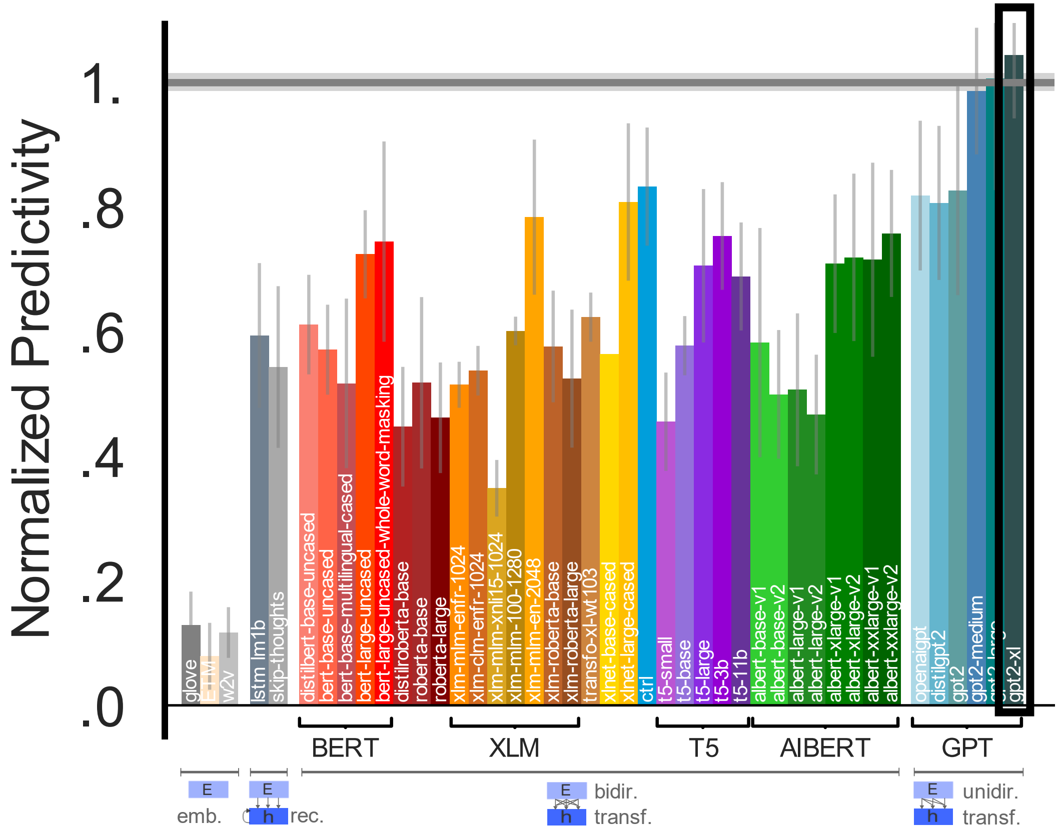
Large feature size without structure is insufficient

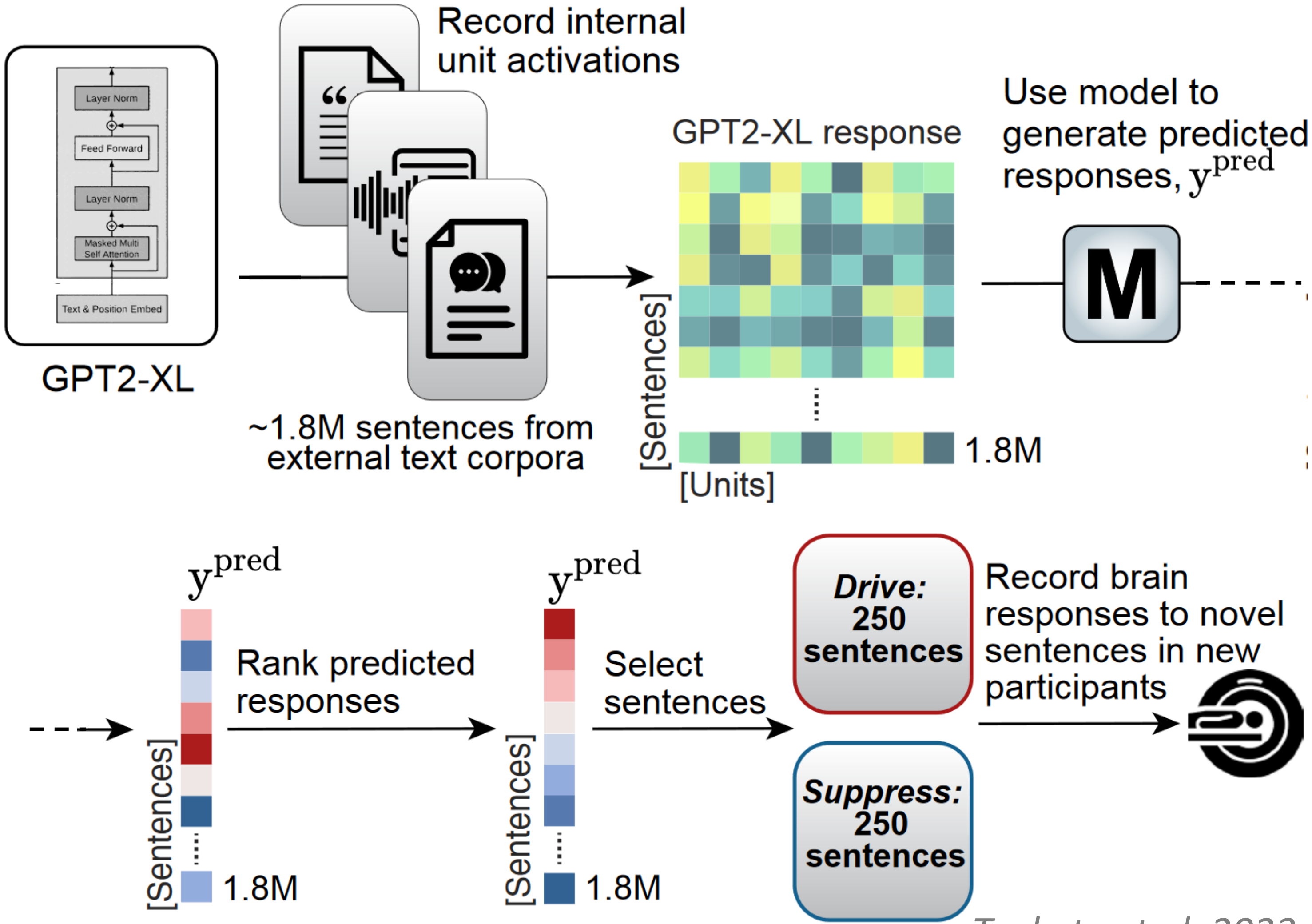
LLMs align to the brain's language system after developmentally realistic amounts of training



Hosseini et al. 2022

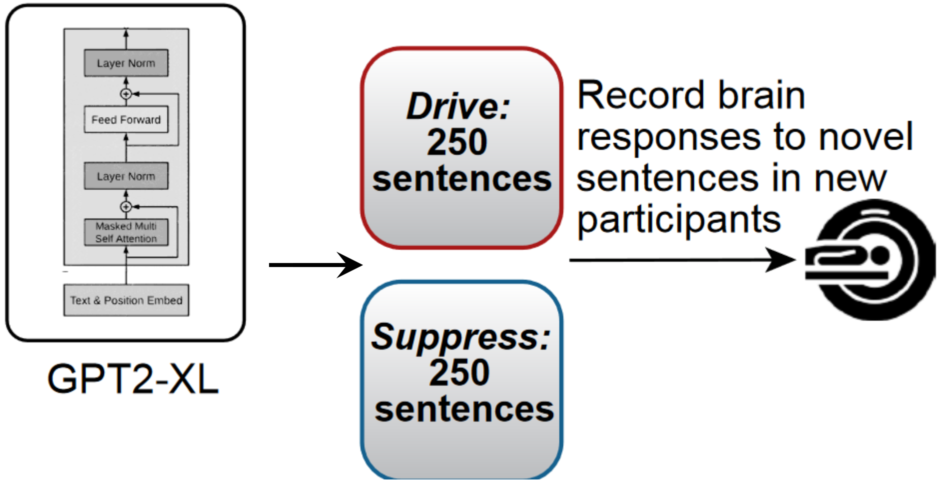
We can use brain-aligned LLMs to noninvasively control neural activity

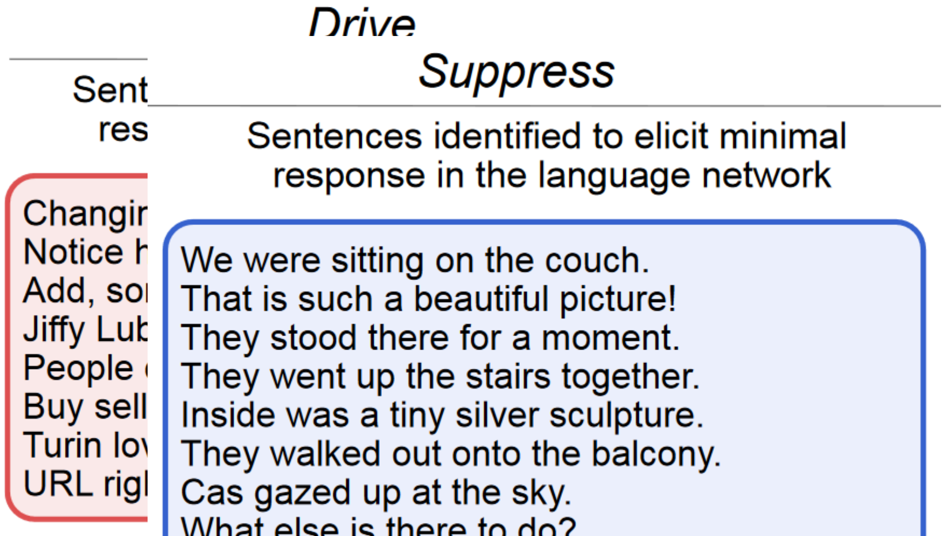


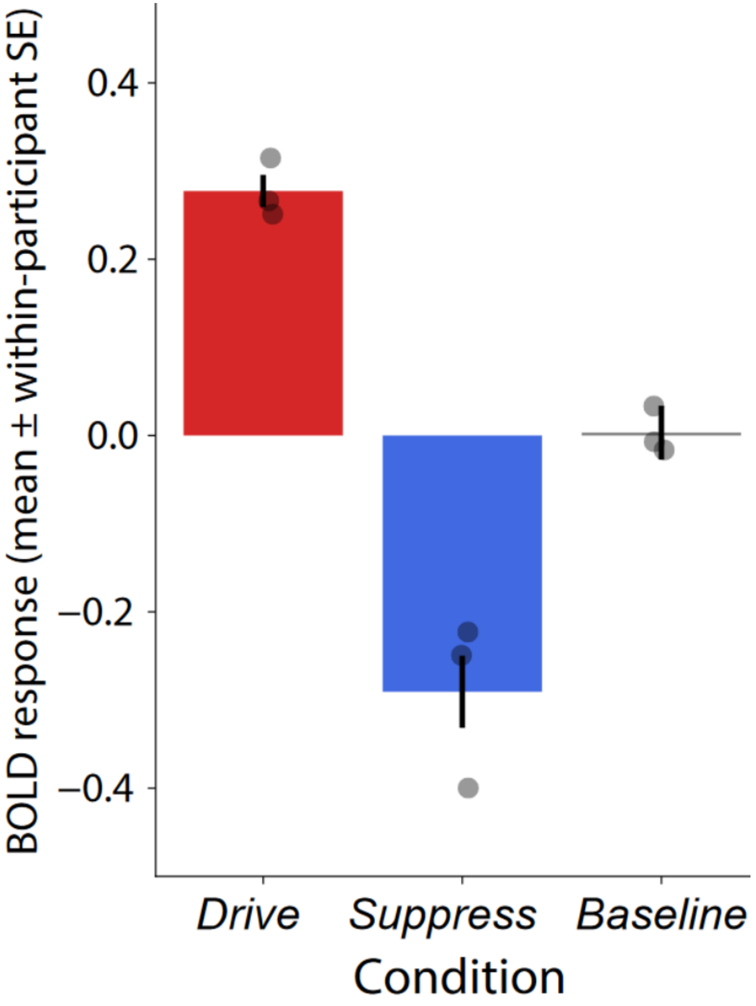


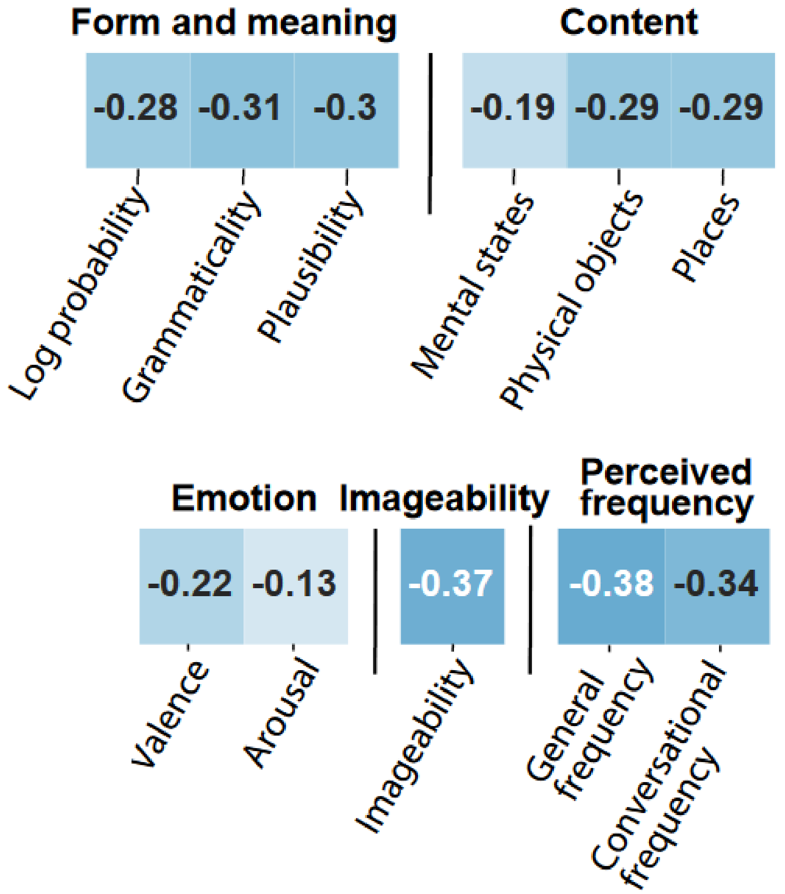
Tuckute et al. 2023

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





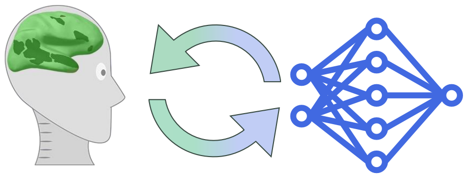


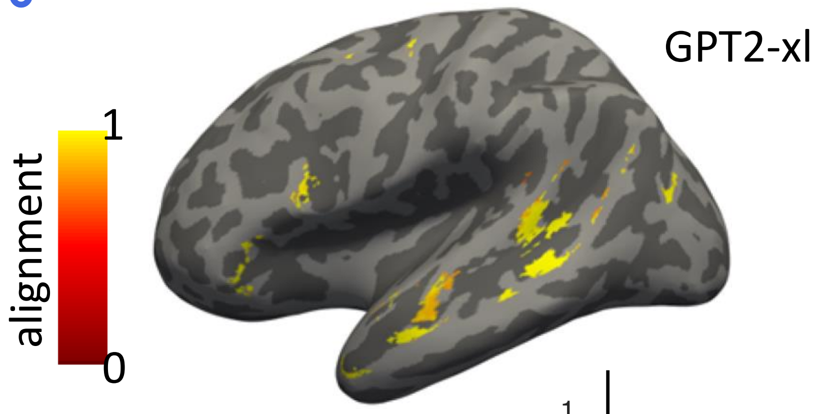


Tuckute et al. 2023

martin.schrimpf@epfl.ch
@martin_schrimpf
mastodon.social/
@mschrimpf

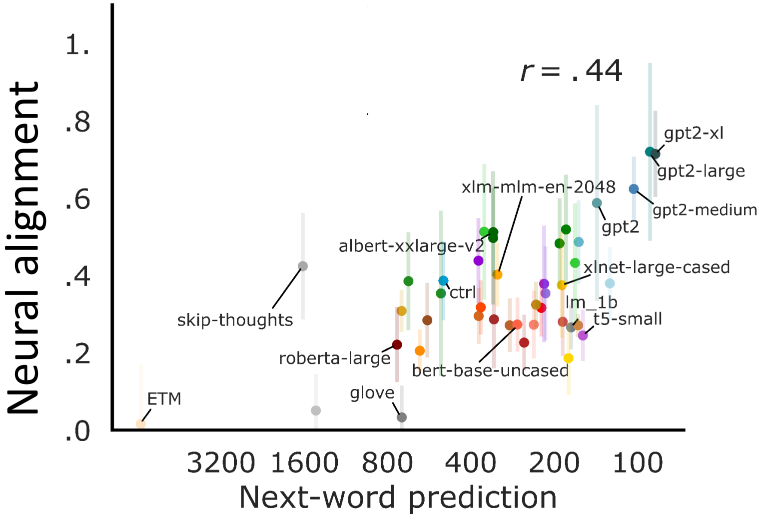
Contributions



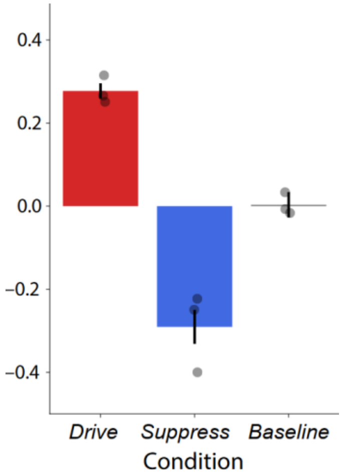


1 Particular LLMs are strong models of the human language system

2 Next-word prediction performance relates to brain and behavioral alignment



3 The best models can be used to noninvasively control neural activity



www. Brain-Score .org/language

Schrimpf et al. 2021
Hosseini et al. 2022 | Tuckute et al. 2023