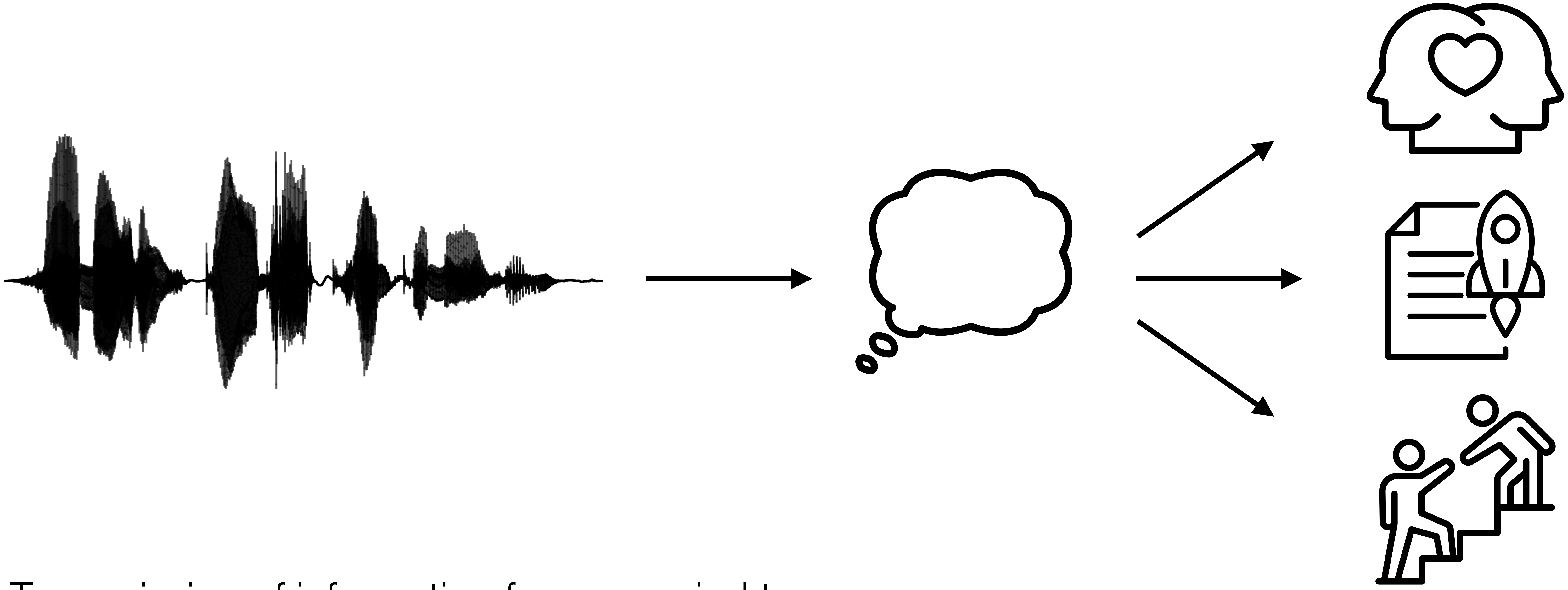
The Human Language System and Large Language Models (LLMs)

Deep Learning in Brains, Minds & Machines | Brown University

April 14, 2026

Greta Tuckute
gtuckute@fas.harvard.edu

Kempner INSTITUTE
For the Study of Natural
& Artificial Intelligence
at Harvard University



Transmission of information from my mind to yours

We share the code (*the language*) that maps between sounds and their meanings

Enabling downstream behaviors, such as social inference, planning, or executing actions

Language serves as a core component of intelligent behavior

The neural architecture of language: Integrative modeling converges on predictive processing

Martin Schrimpf[a,b,c,1], Idan Asher Blank[a,d,2], Greta Tuckute[a,b,2], Carina Kauf[a,b,2], Eghbal A. Hosseini[a,b], Nancy Kanwisher[a,b,c,1], Joshua B. Tenenbaum[a,c,3], and Evelina Fedorenko[a,b,1,3] 2021

Shared computational principles for language processing in humans and deep language models

Ariel Goldstein[1,2], Zaid Zada[1,8], Eliav Buchnik[2,8], Mariano Schain[2,8], Amy Price[1,8], Bobbi Aubrey[1,3,8], Samuel A. Nastase[1,8], Amir Feder[2,8], Dotan Emanuel[2,8], Alon Cohen[2,8], Aren Jansen[2,8], Harshvardhan Gazula[1], Gina Choe[1,3], Aditi Rao[1,3], Catherine Kim[1,3], Colton Casto[1],

Brains and algorithms partially converge in natural language processing

Charlotte Caucheteux[1,2] & Jean-Rémi King[1,3] 2022

Scaling laws for language encoding models in fMRI

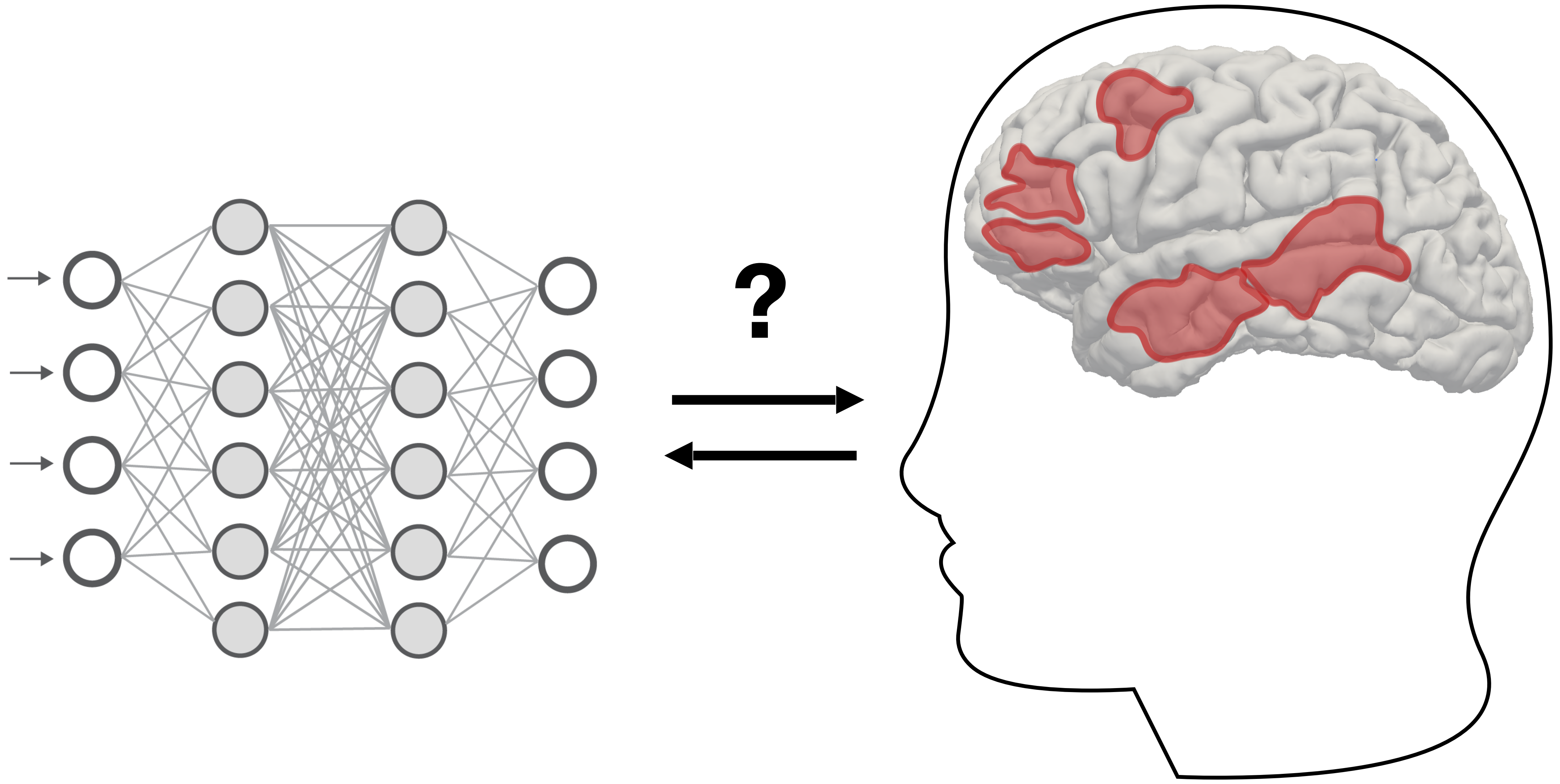
Aditya R. Vaidya
epartment of Computer Science
he University of Texas at Austin
avaidya@utexas.edu

ıth
: and Neuroscience
at Austin
edu 2023

Shared functional specialization in transformer-based language models and the human brain

Received: 21 July 2023
Accepted: 24 May 2024
Published online: 29 June 2024

Sreejan Kumar[1,6], Theodore R. Sumers[2,6], Takateru Yamakoshi[3], Ariel Goldstein[4], Uri Hasson[1,5], Kenneth A. Norman[1,5], Thomas L. Griffiths[2,5], Robert D. Hawkins[1,5] & Samuel A. Nastase[1] 2024



And others: Jain & Huth, 2018; Toneva & Wehbe 2019; Lopopolo et al., 2020; Pasquiou et al., 2022; Kauf, Tuckute et al., 2023; Caucheteux et al., 2023; Tucker & Tuckute, 2023; Tikochinski et al., 2024; Tuckute et al., 2024; Hosseini et al., 2024, Singh, Antonello et al., 2025 …

Why should we care?

- A(ny) model is a hypothesis of processes in the brain: language models are *computational* hypotheses.

Why should we care?

- A(ny) model is a hypothesis of processes in the brain: language models are *computational* hypotheses.
    - No animal models in language!
    - Models can be manipulated, observed, and lesioned.
- Language models perform *biologically-relevant tasks.*
- An accurate model of linguistic processes has *engineering/ therapeutic* applications.
- First time in history we have a system besides the human brain that generates language — do these two (seemingly) different systems share representational/computational principles?

Outline

- What is language?
  - How can we identify language in the brain?
  - 4 key properties of the language network
- Models of language processing
  - How it started
  - How it continued
  - How it's going
  - So what?

How can we identify language in the brain?

Sentence

"In the starlight he could see the palm trees"

>

Non-word string

"Ap shab succeams ob elt lant meapon"

Reading or listening to a sentence engages a lot of cognitive processes, for example:

- Visual processing of the letters
- Speech segmentation
- Articulation planning
- Accessing word meanings
- Combining words into sentence-level meanings

How can we identify language in the brain?

Speech

"In the starlight he could see the palm trees"

>

Muffled/foreign speech

"Žvaigždžių šviesoje jis galėjo matyti palmes"

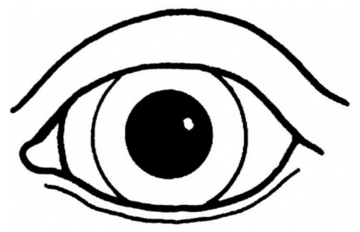
Sentence

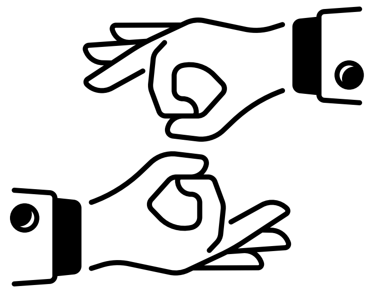
"In the starlight he could see the palm trees"

>

Non-word string

"Ap shab succeams ob elt lant meapon"





MacSweeney et al., 2018;
Karen Emmory in prep

How can we identify language in the brain?

Speech

“In the starlight he could see the palm trees”

Sentence

“In the starlight he could see the palm trees”

Linguistic input

\>

\>

\>

Muffled/foreign speech

“Žvaigždžių … galėjo matyti palmes”

Non-word string

“A… succeams ob … elt lant meapon”

Perceptually matched control

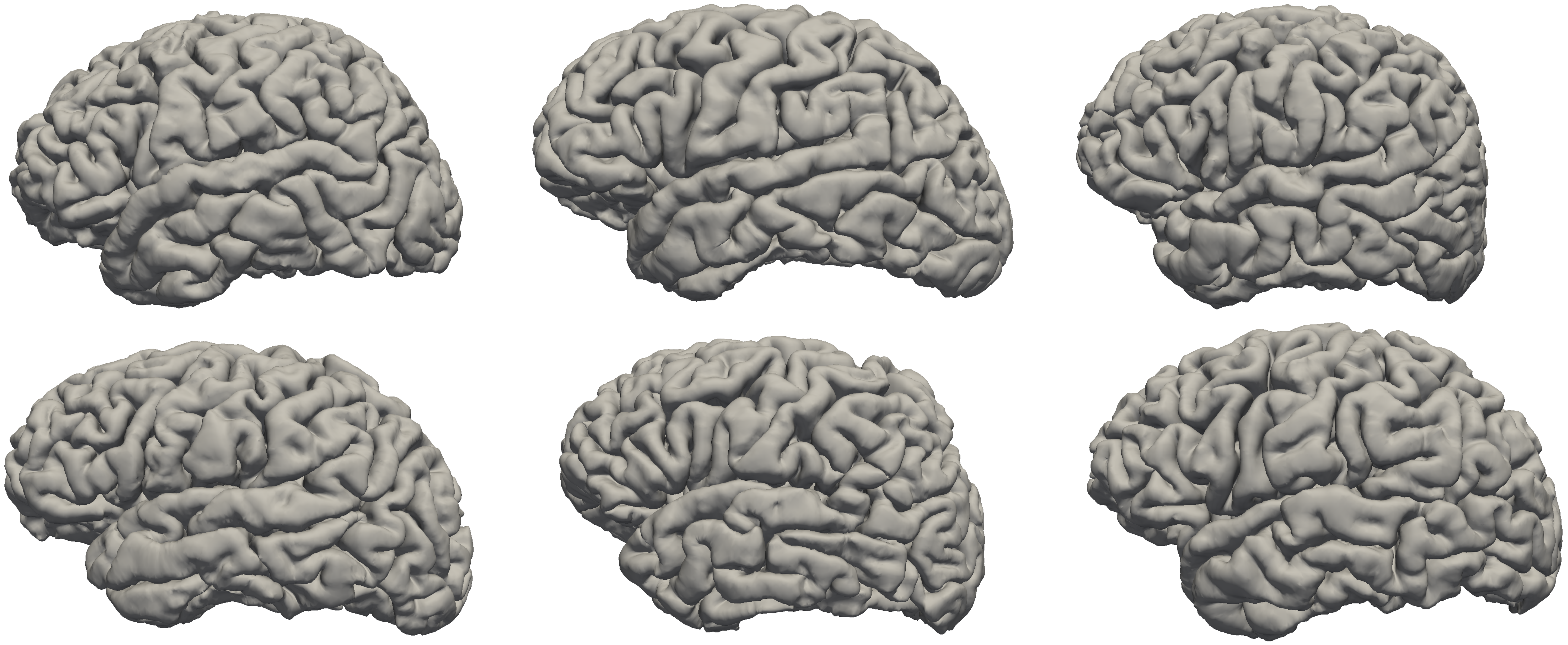
We aim to identify brain regions that are:

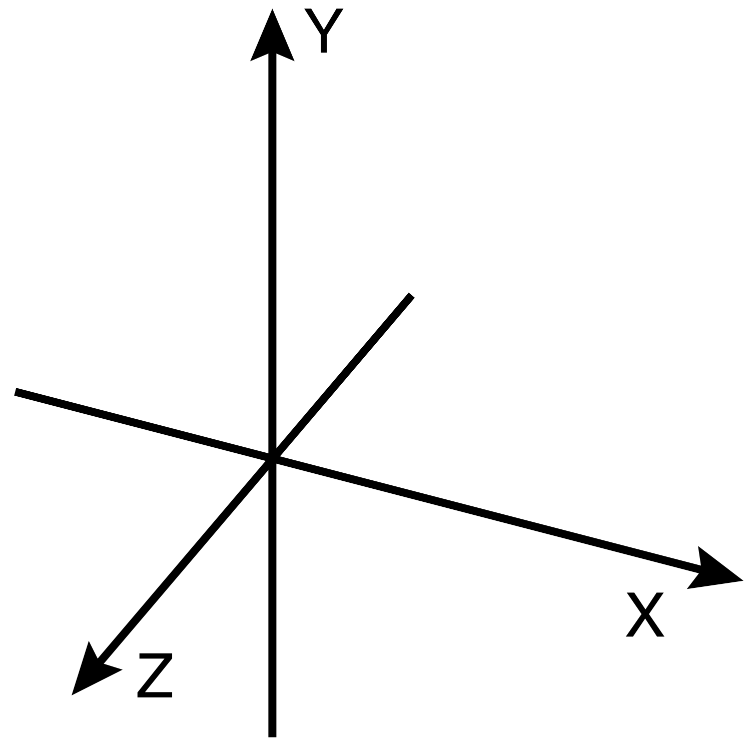
- Sensitive to word- and sentence-level meaning.
- ‘Higher-level’ (not perceptual or motor processing).

How can we identify language in the brain?

Practically (fMRI example)

Problem: Difficult to predict function from anatomical location alone





How can we identify language in the brain?

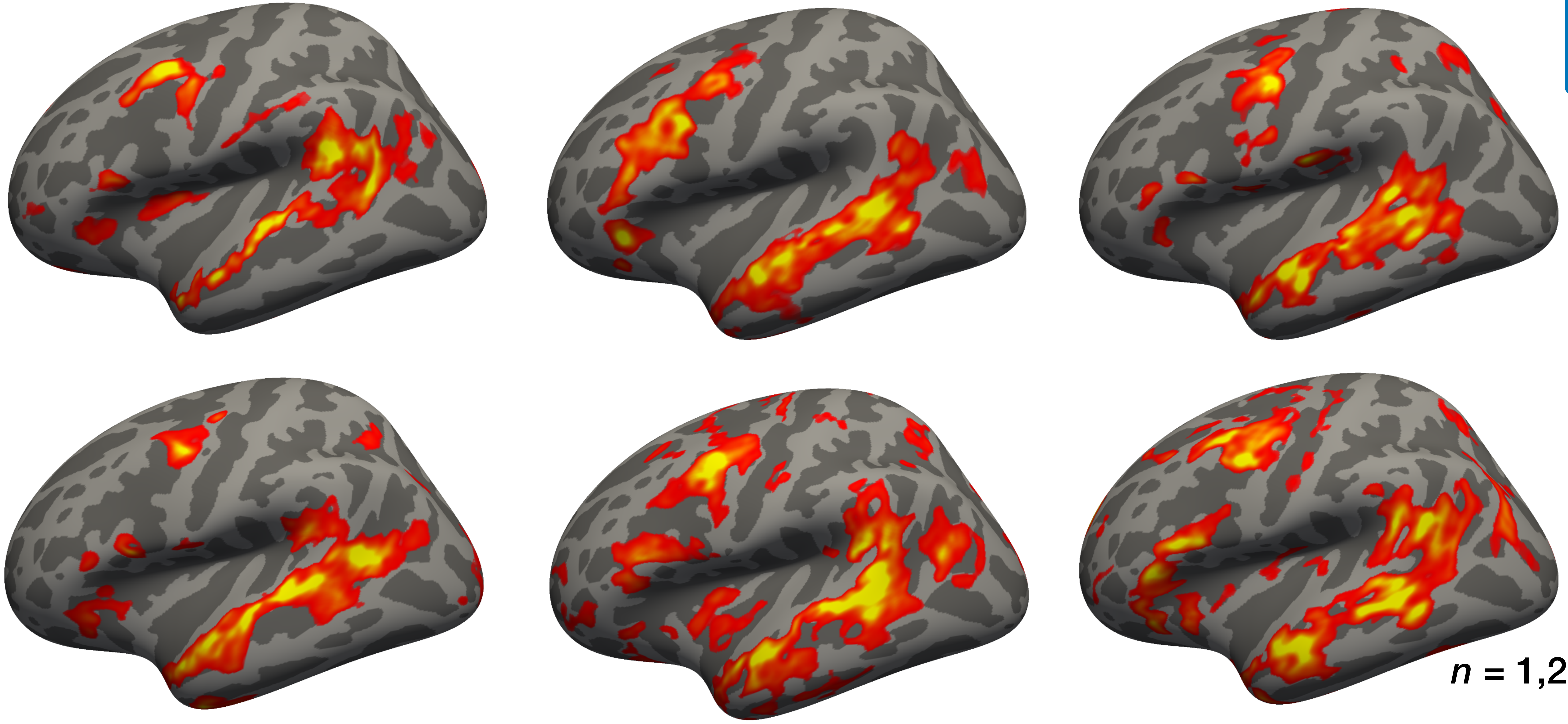
Practically (fMRI example)

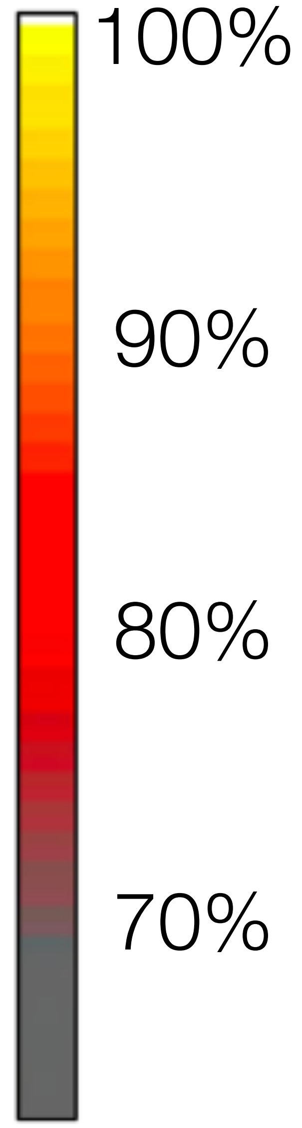
Linguistic input

>

Perceptually matched control

Solution: Localize the language network in each individual





n **= 1,2,3,4,5,6, … , 804**

Lipkin et al., 2022

How can we identify language in the brain?

Practically (fMRI example)

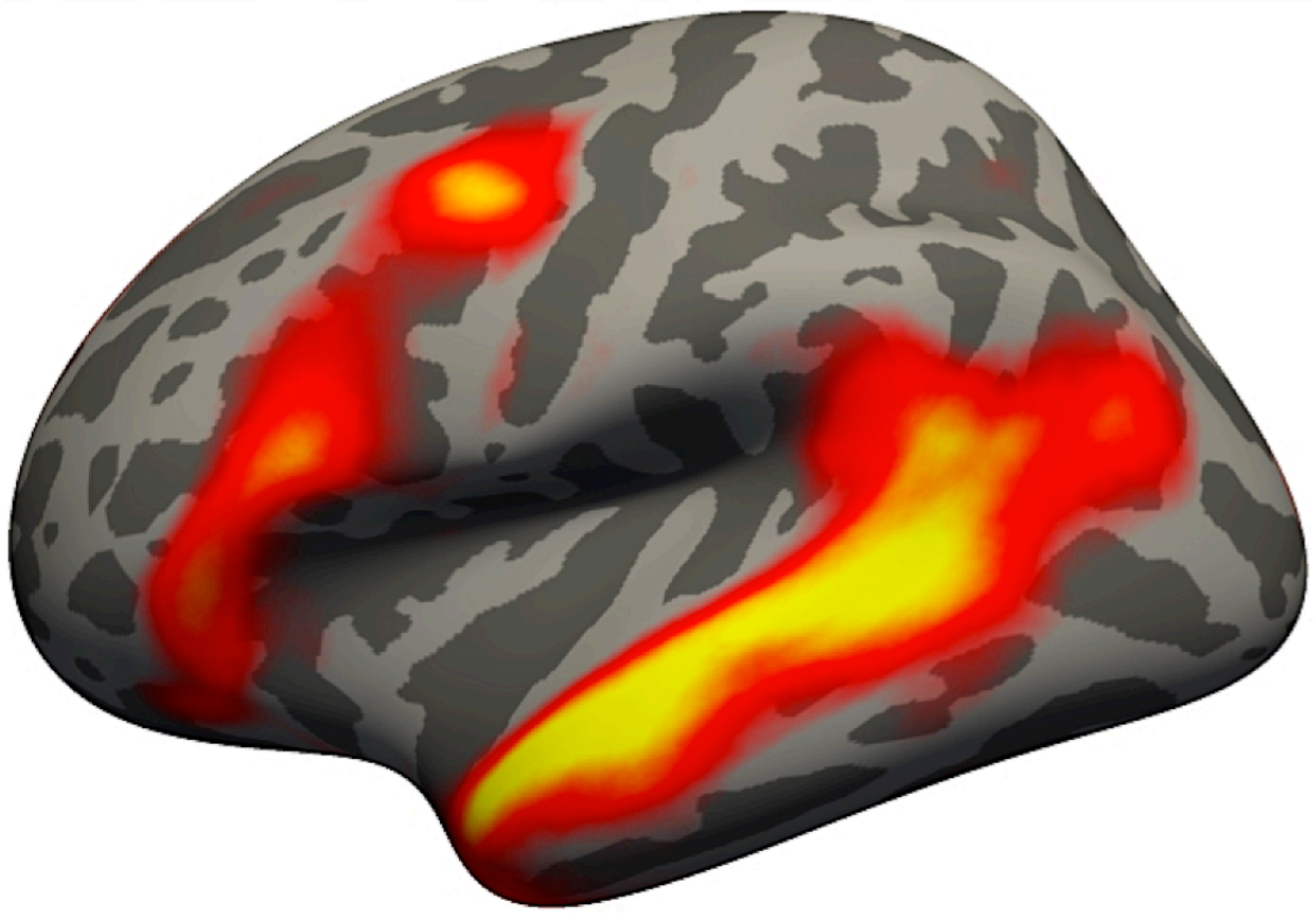
Solution: Localize the language network in each individual

Linguistic input

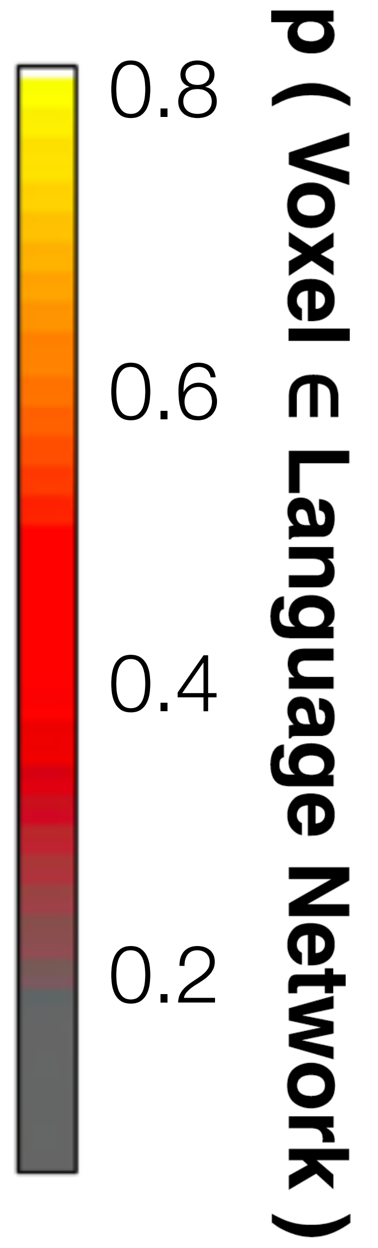
\>

Perceptually matched control

n = 804



Left hemisphere



Lipkin et al., 2022

How can we identify language in the brain?

Practically (fMRI example)

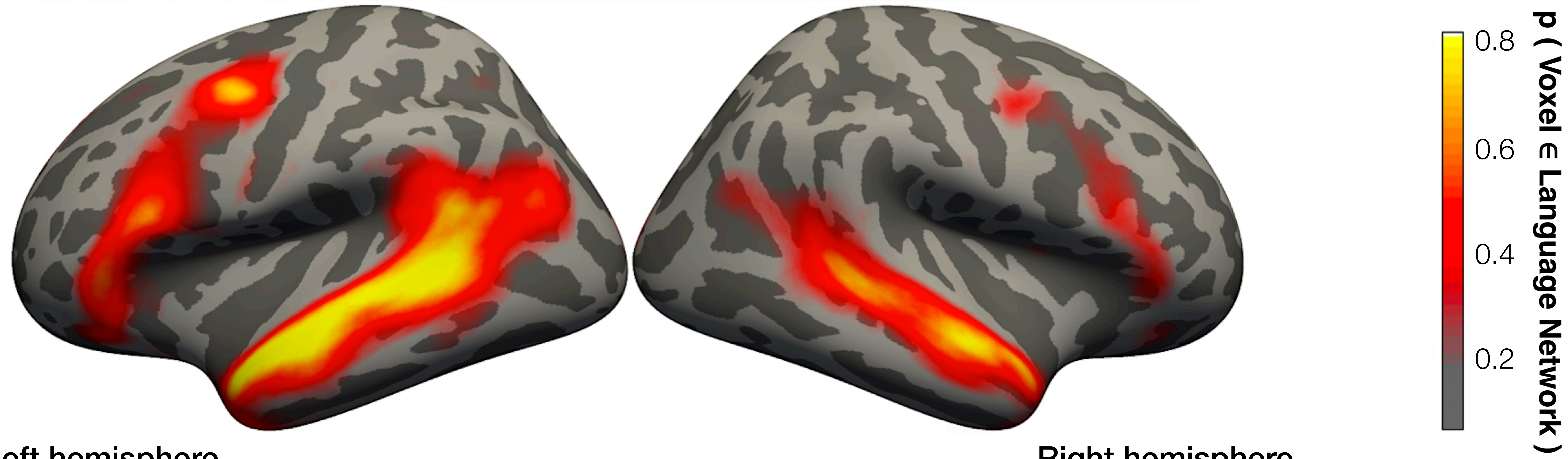
Solution: Localize the language network in each individual

Linguistic input

>

Perceptually matched control

n = 804



Left hemisphere

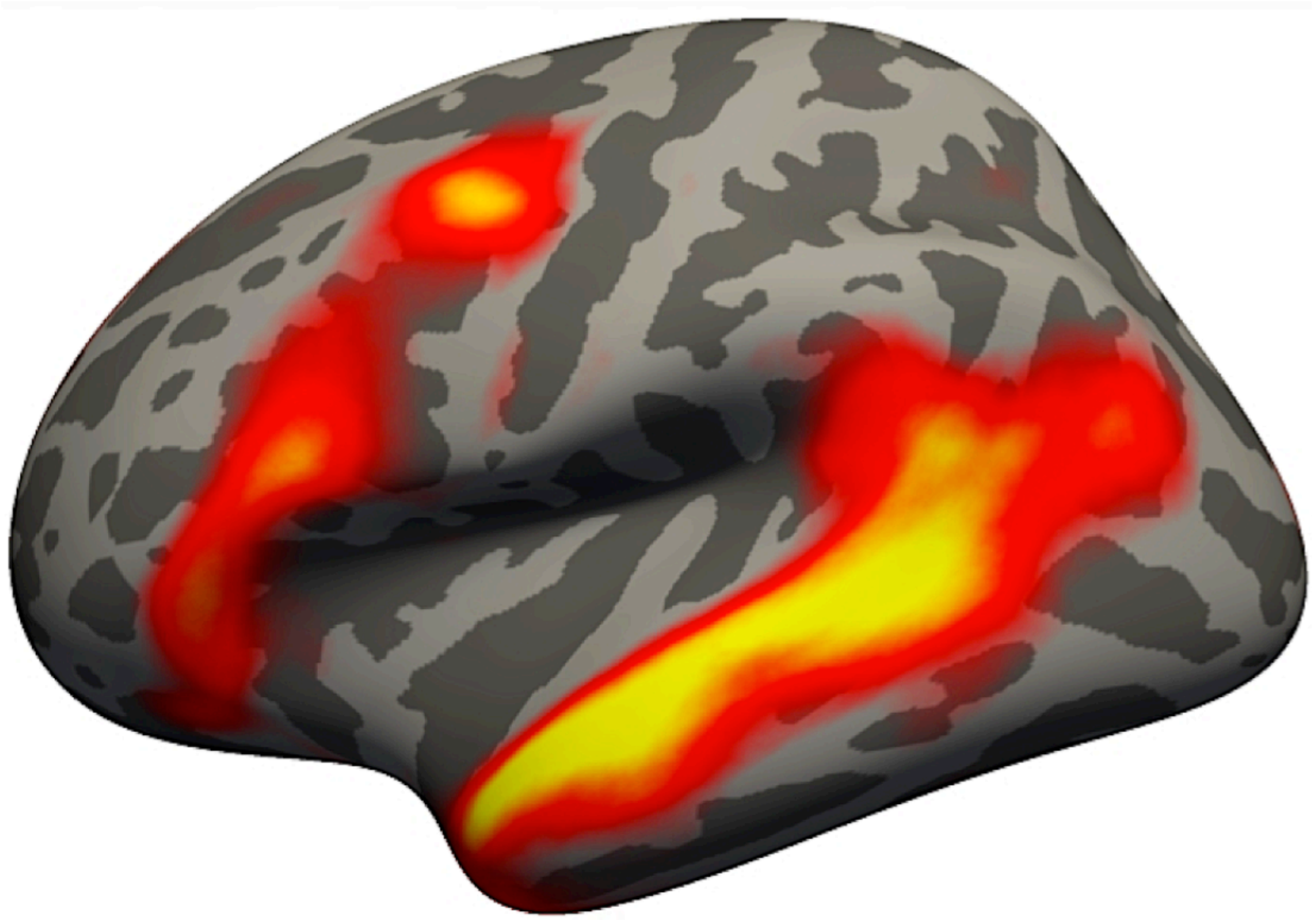
Right hemisphere

Lipkin et al., 2022

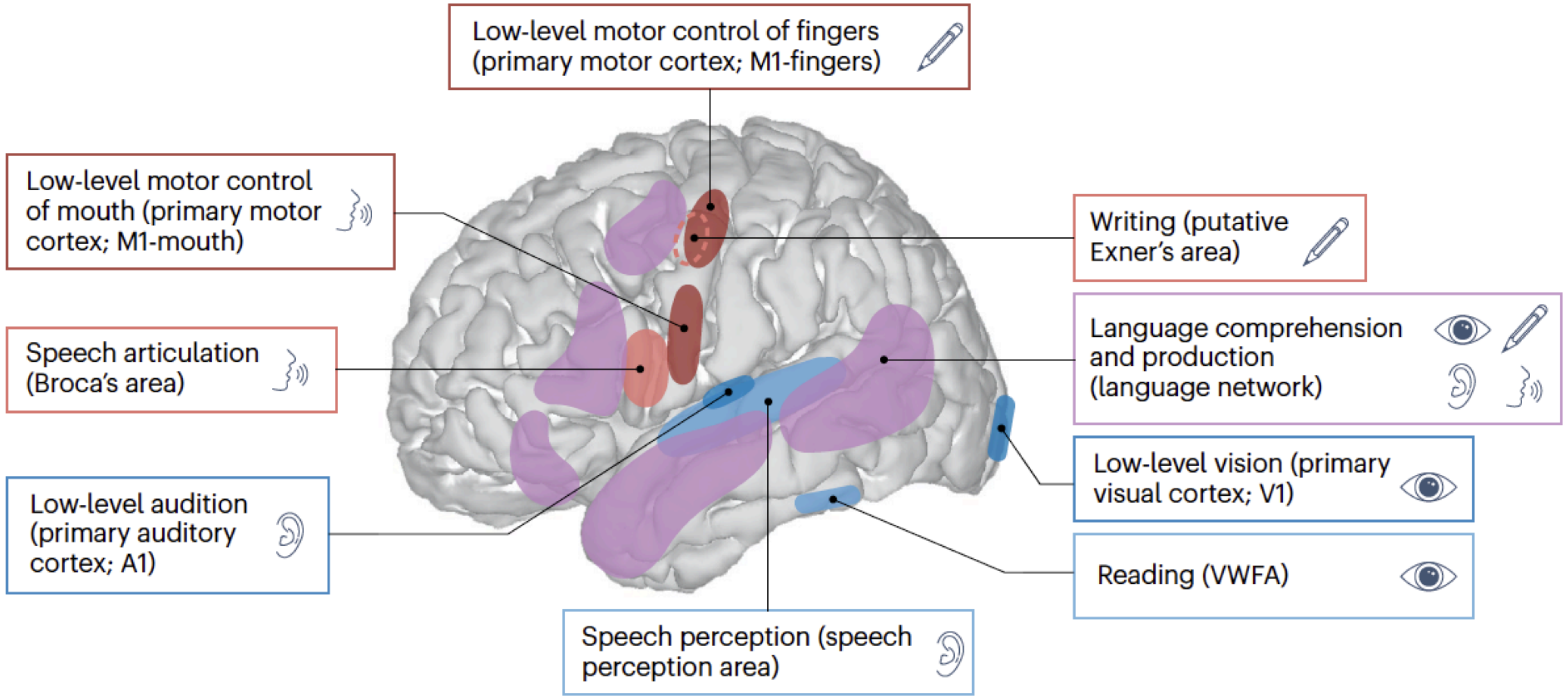
How can we identify language in the brain?

Practically (fMRI example)

Where are the perceptual and motor regions?



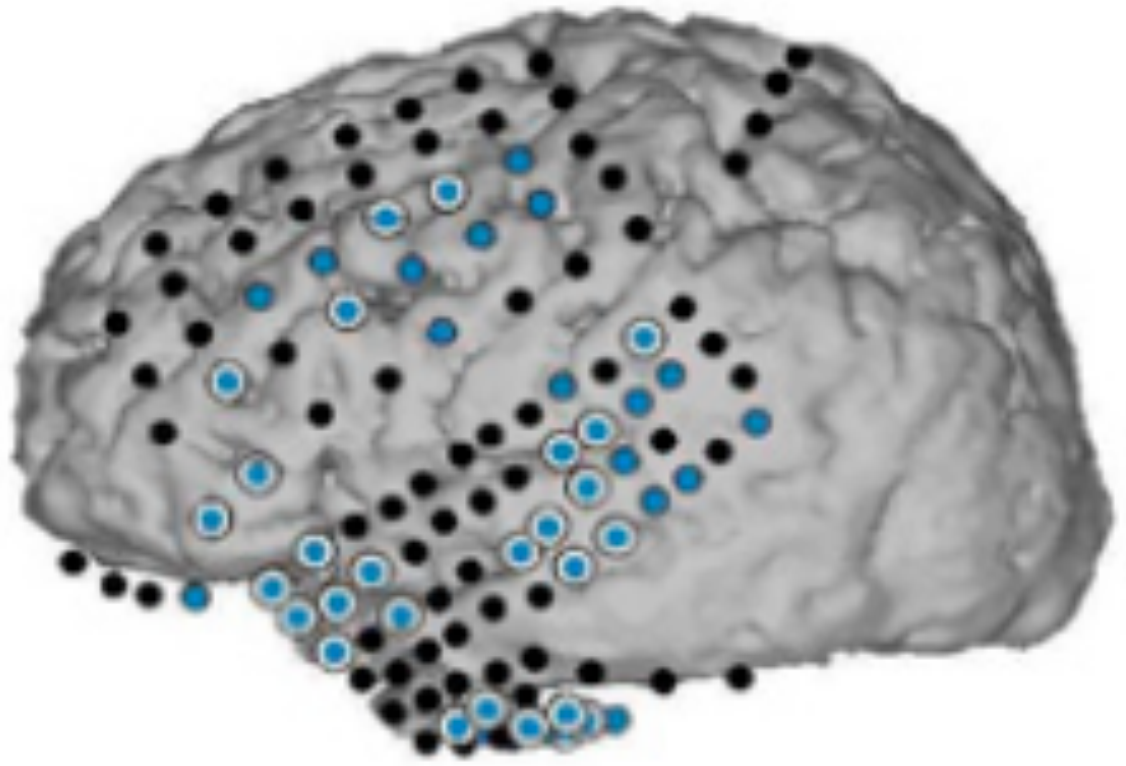
Left hemisphere

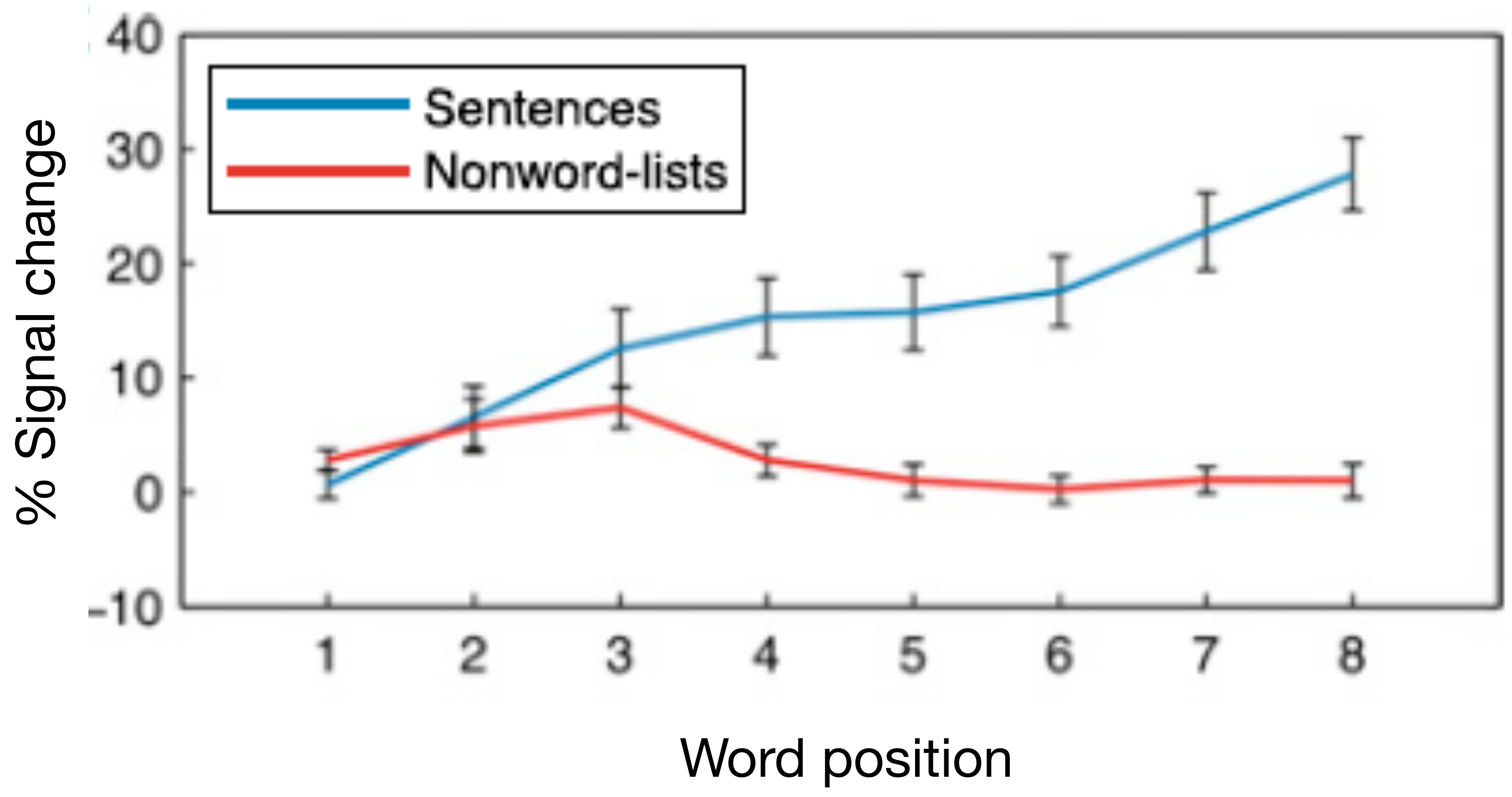


Fedorenko et al., 2024
Lipkin et al., 2022

How can we identify language in the brain?
Practically (intracranial example)

Electrocorticography (ECoG) grids
(example patient)





Fedorenko et al., 2016

Outline

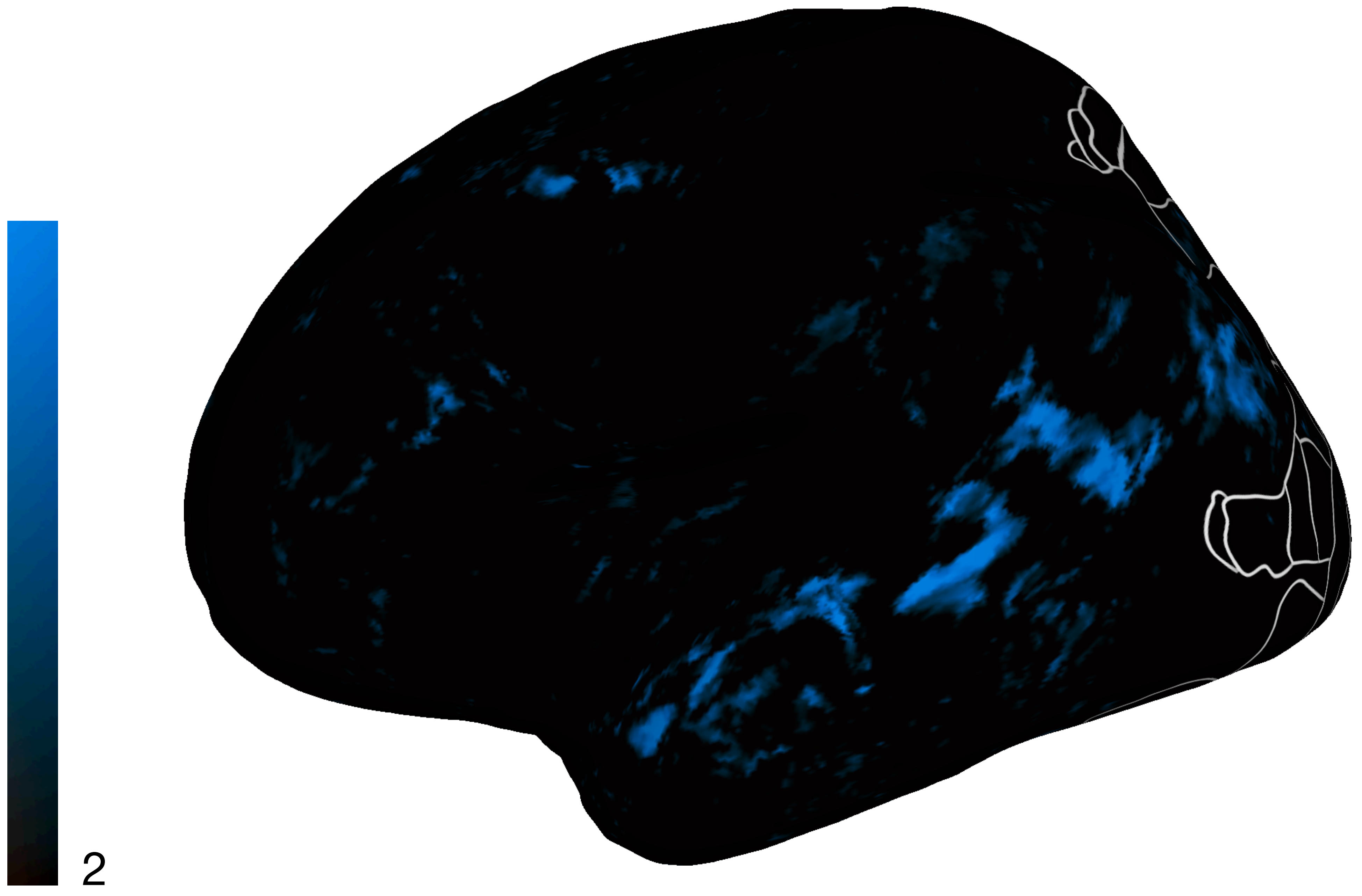
- What is language?
  - How can we identify language in the brain?
  - 4 key properties of the language network
- Models of language processing
  - How it started
  - How it continued
  - How it's going
  - So what?

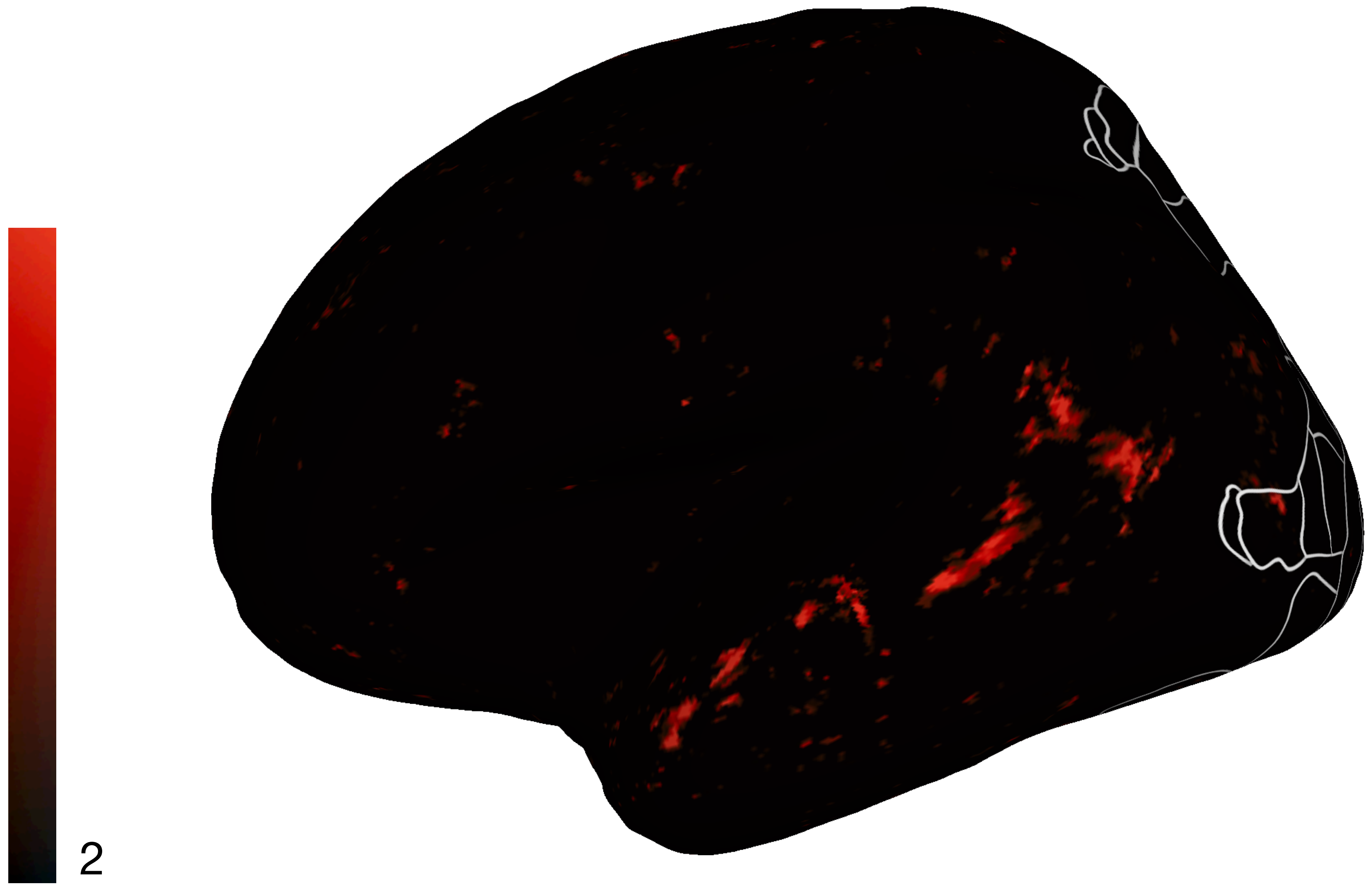
4 key properties of the language network

1. Modality independent (MacSweeney et al., 2002; Scott et al., 2017; Deniz et al., 2019; Trettenbrein et al., 2020)
2. Functionally integrated network (Blank et al., 2014; Braga et al., 2020; Paunov et al., 2019, Tuckute et al., 2024)
3. Dissociated from other cognitive abilities (Monti et al., 2009; 2011; Fedorenko et al., 2011; 2013; Ivanova et al., 2020; Chen, Affourtit et al., 2023, Tuckute, Lee et al., 2024)
4. Causally important for language (Mesulam et al., 2001; Dronkers, 2010; Ivanova et al., 2021, Wilson et al., 2023)

1. Modality independent

We can record brain responses to visual and auditory language localizers:
(some of my pilot data using ultra-high-field fMRI)





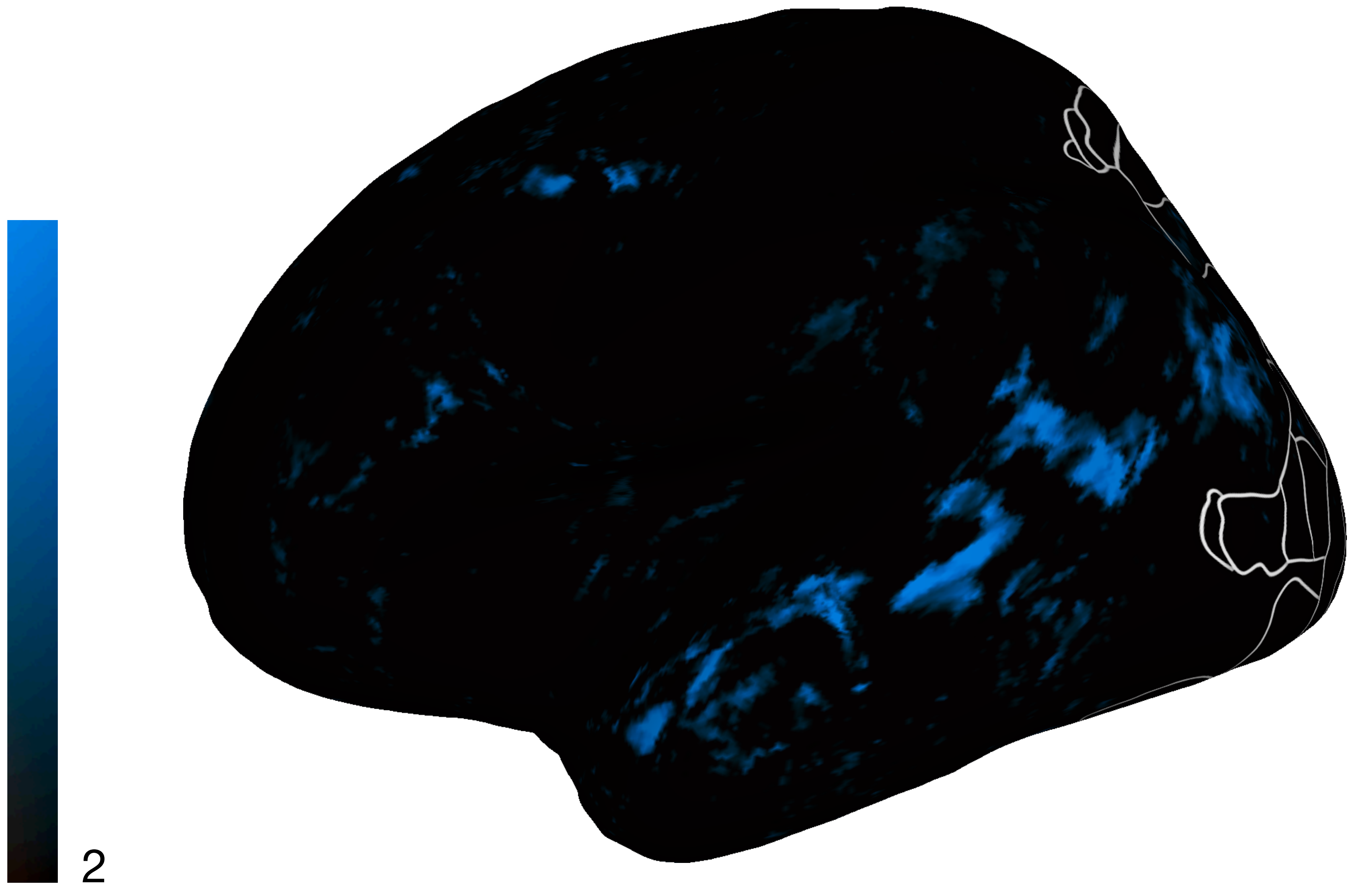
Visual
Sentences > non-words *t*-statistic

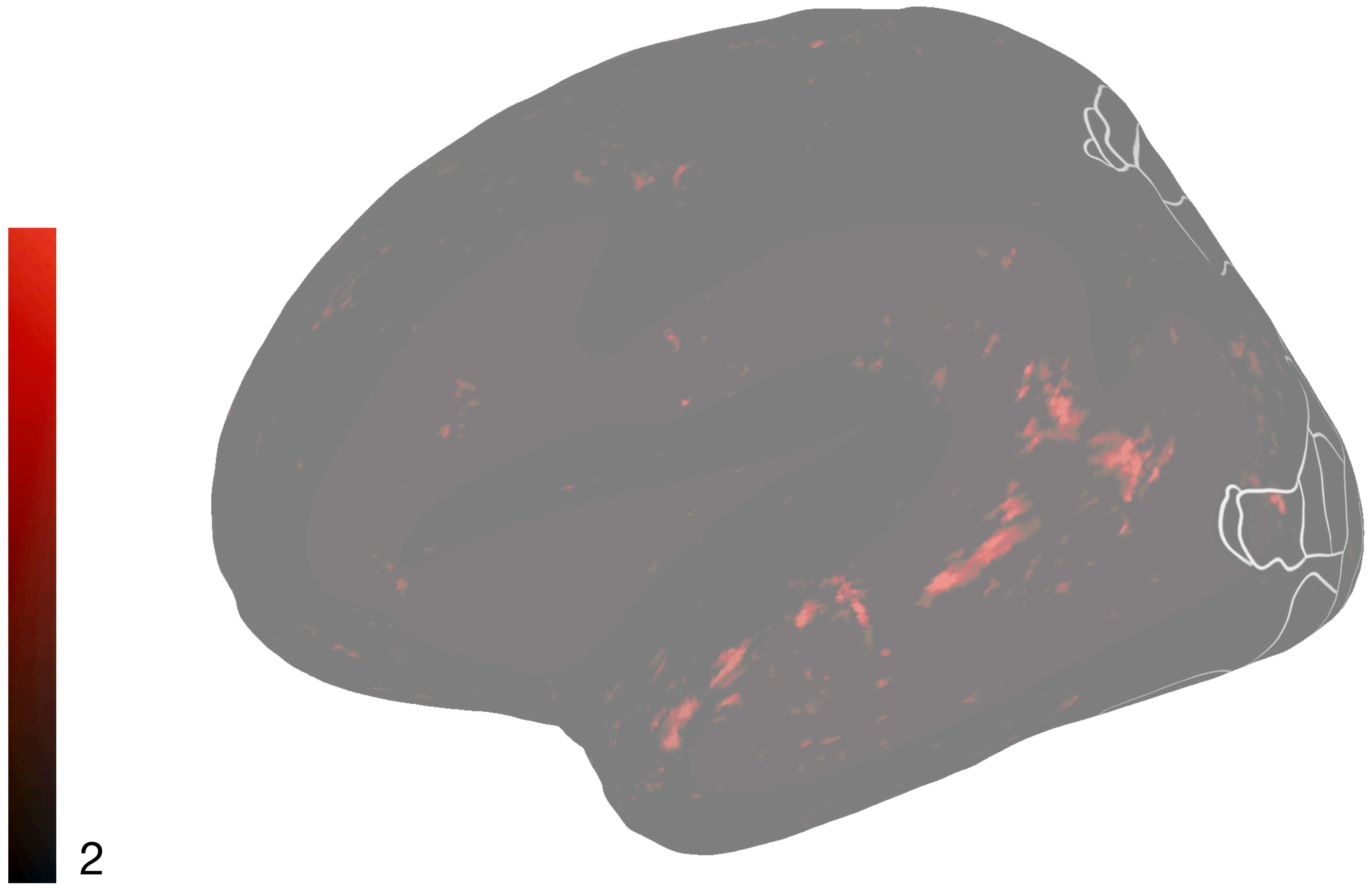
Auditory
Speech > degraded speech *t*-statistic

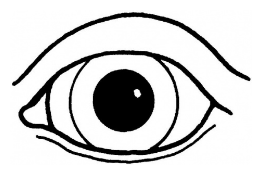
In prep w. Kendrick Kay; also see Scott et al., 2017; Deniz et al., 2019

1. Modality independent

We can record brain responses to visual and auditory language localizers:
(some of my pilot data using ultra-high-field fMRI)





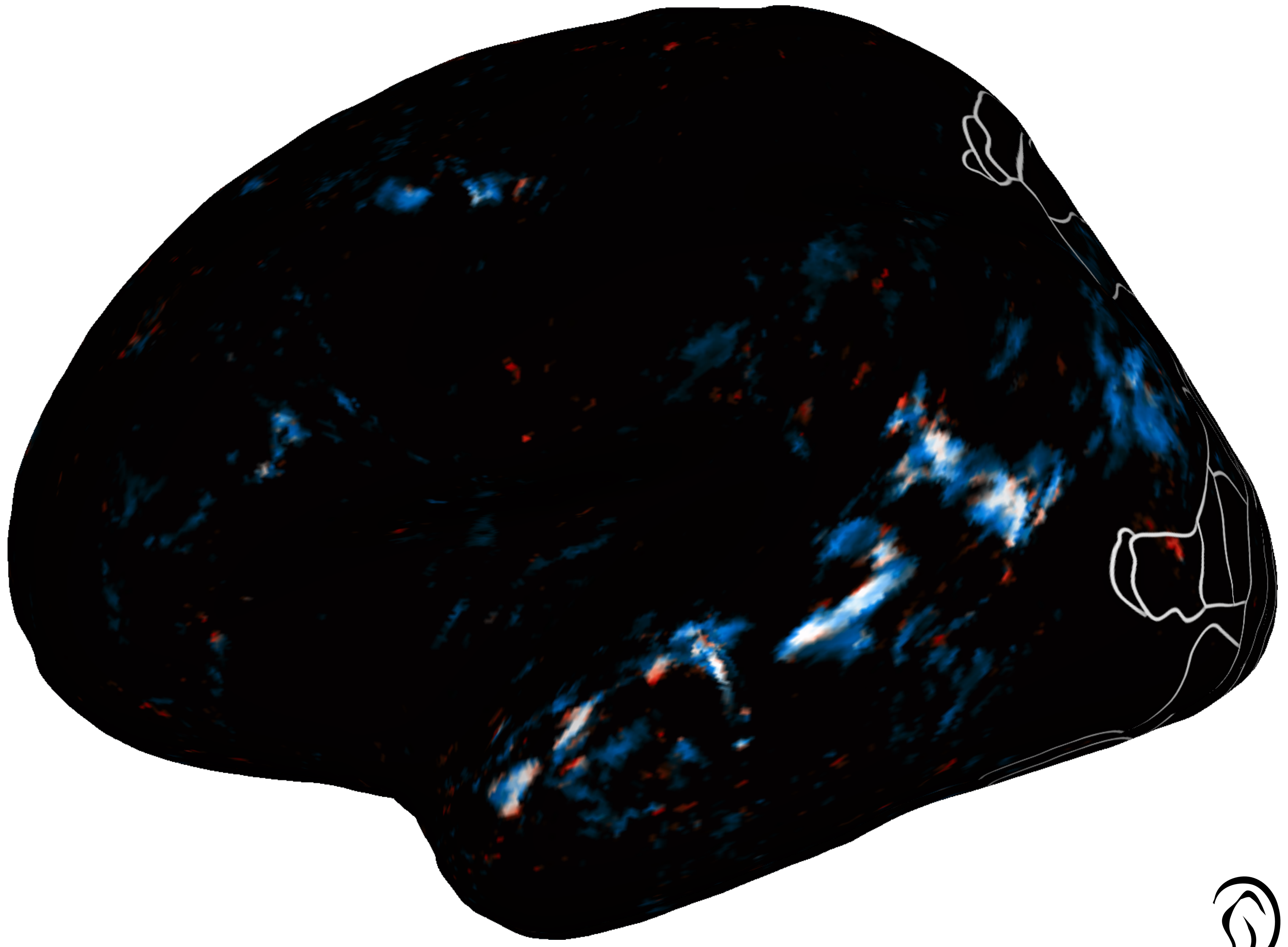
Visual
Sentences > non-words *t*-statistic 

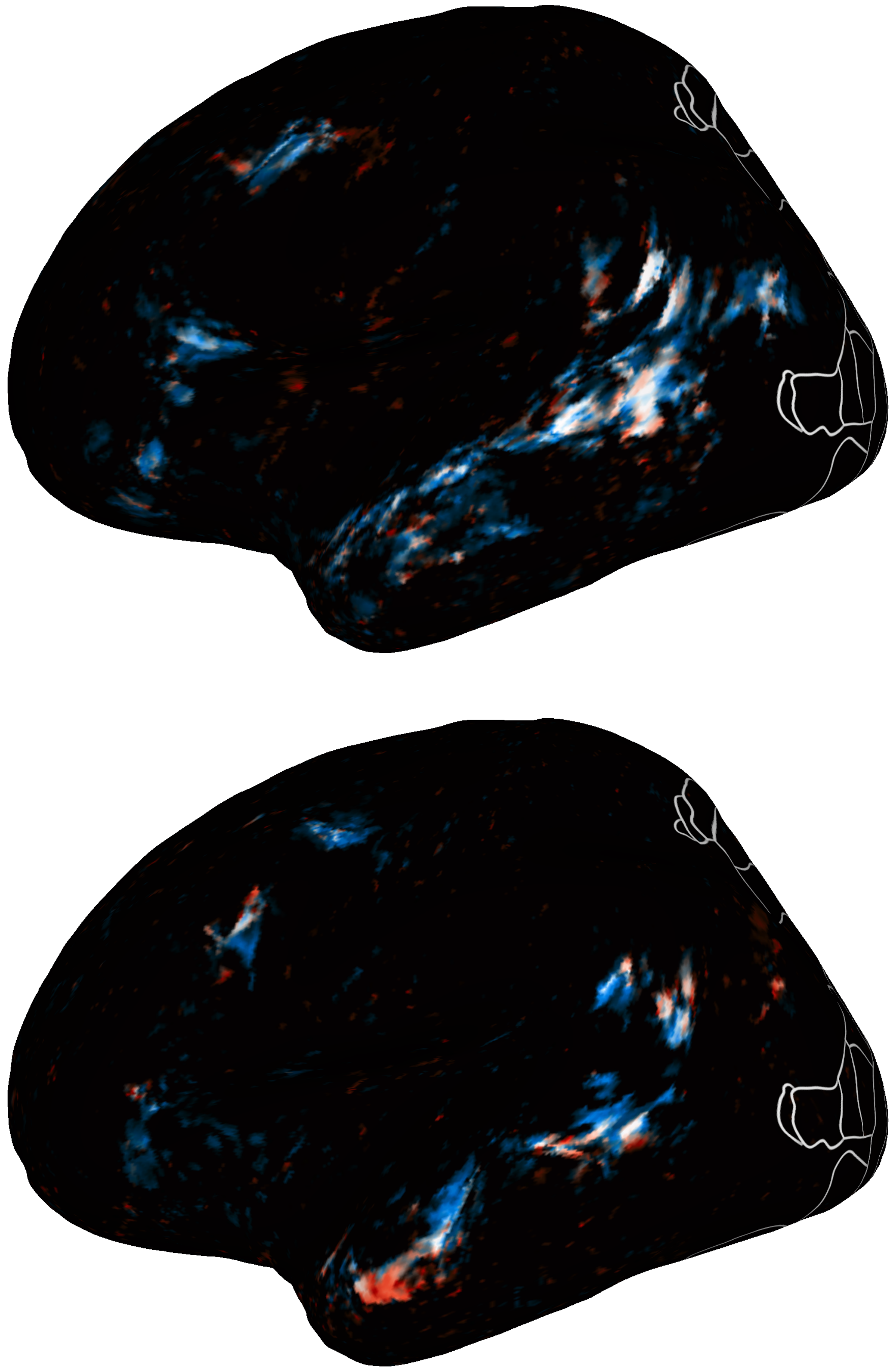
Auditory
Speech > degraded speech *t*-statistic 

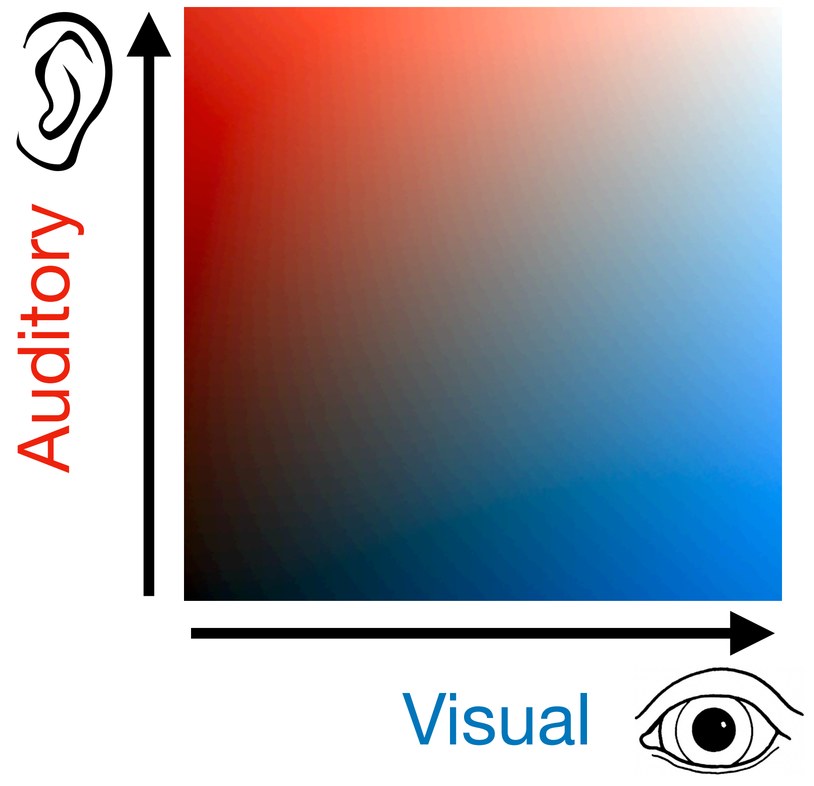
In prep w. Kendrick Kay; also see Scott et al., 2017; Deniz et al., 2019

1. Modality independent

We can record brain responses to visual and auditory language localizers:
(some of my pilot data using ultra-high-field fMRI)

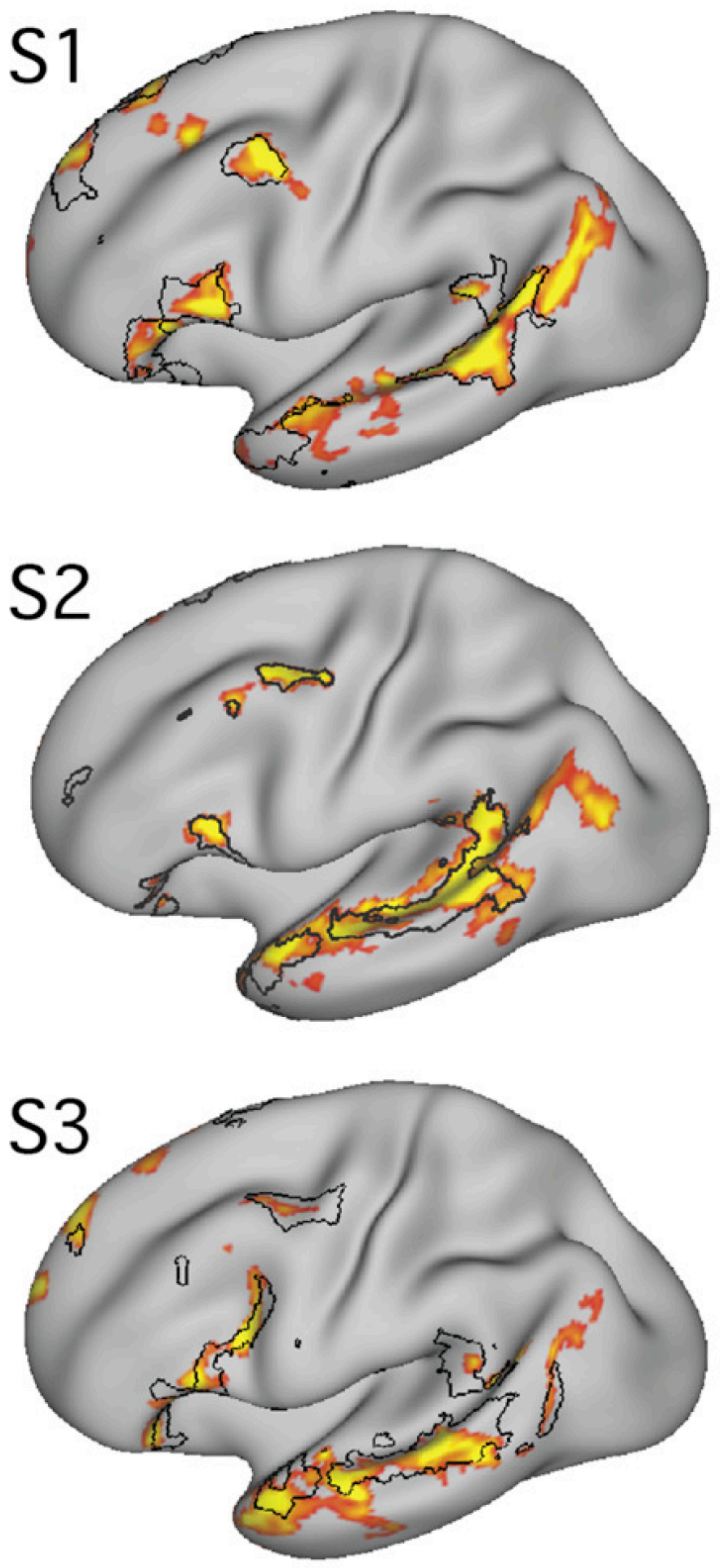


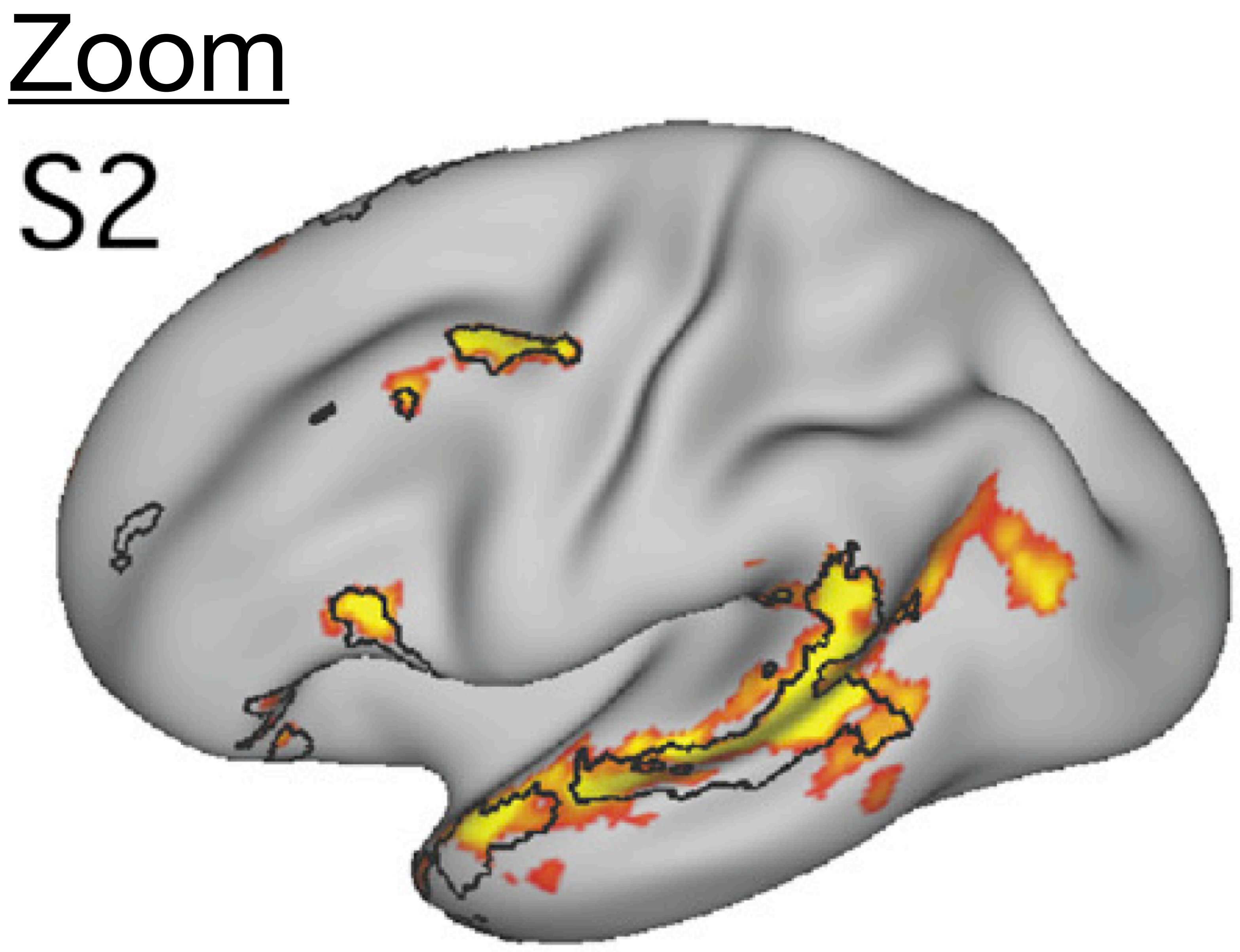




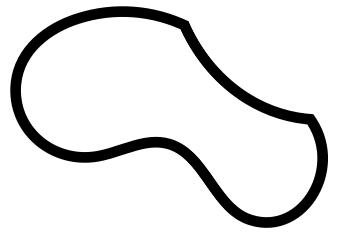
In prep w. Kendrick Kay; also see Scott et al., 2017; Deniz et al., 2019

2. Functionally integrated network






Sentences > non-words response


Resting state correlations

Linguistic input

>

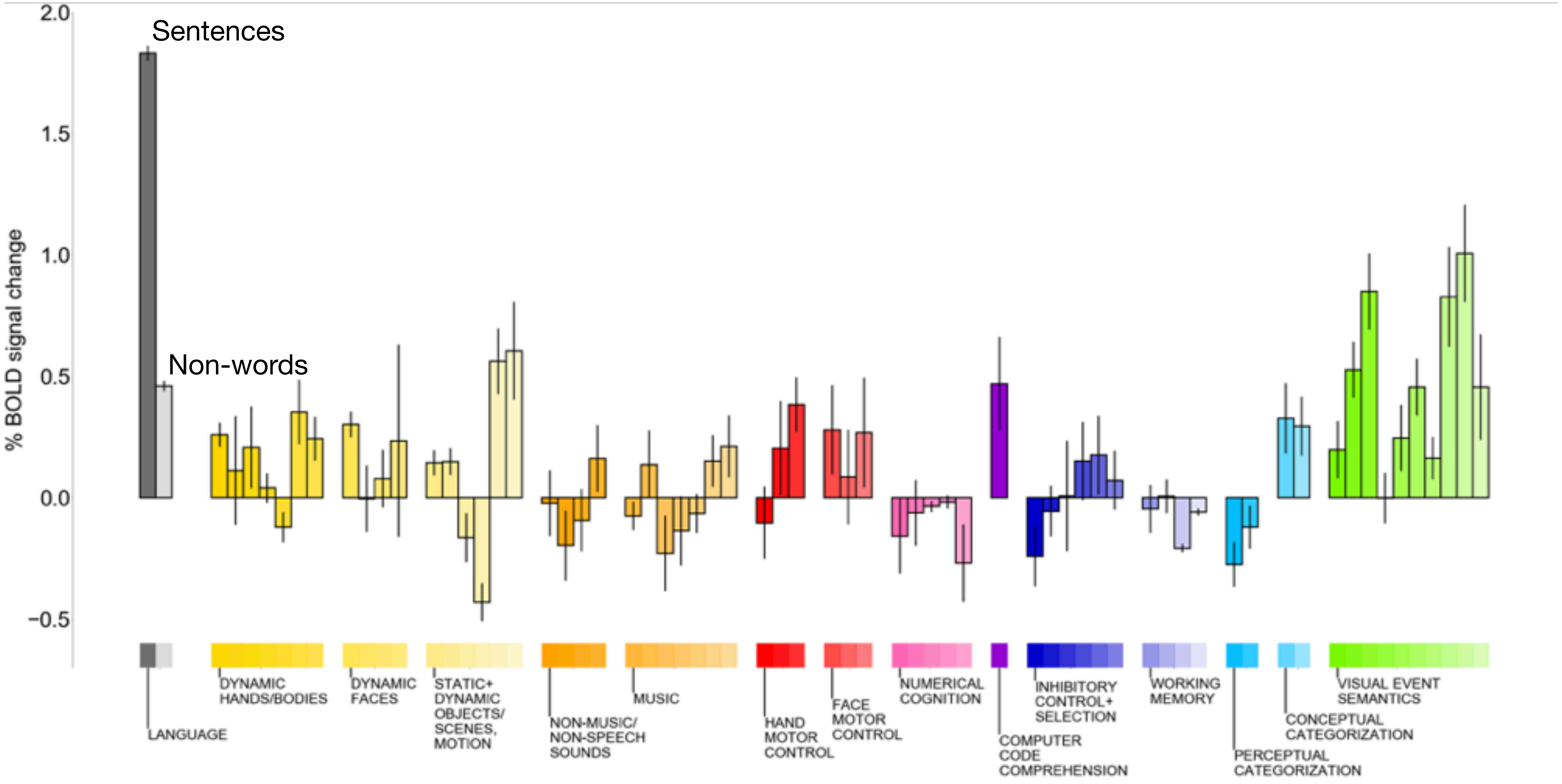
Perceptually matched control

Braga et al., 2020; also: Blank et al., 2014; Shain & Fedorenko, 2025

3. Dissociated from other cognitive abilities

3. Dissociated from other cognitive abilities

32 experiments, 64 conditions; 761 participants across 1,007 scanning sessions



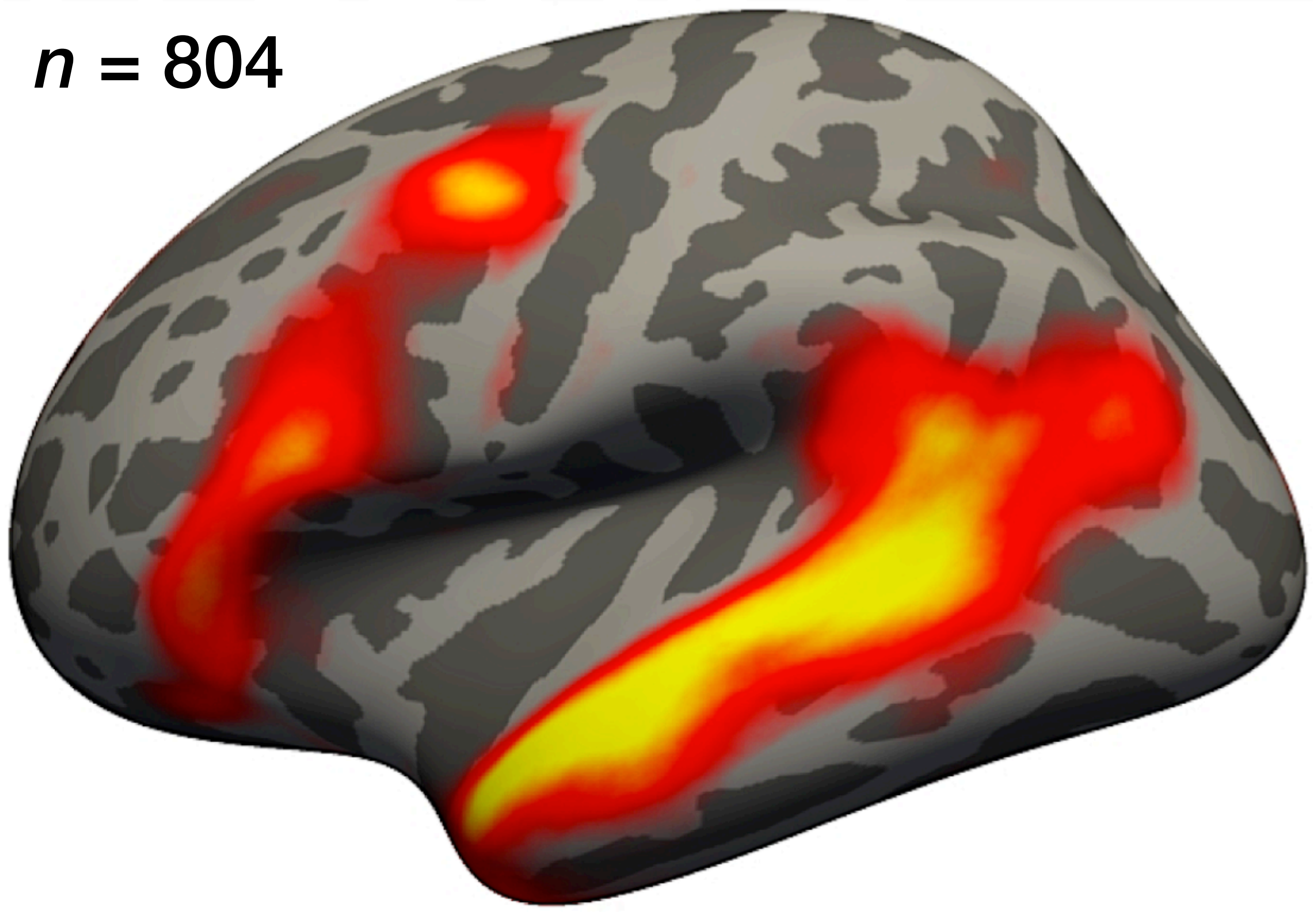
Courtesy of EvLab. Related papers: Monti et al., 2009; 2011; Fedorenko et al., 2011; 2013; Ivanova et al., 2020; Chen, Affourtit et al., 2023; see Fedorenko et al., 2024 for a recent review

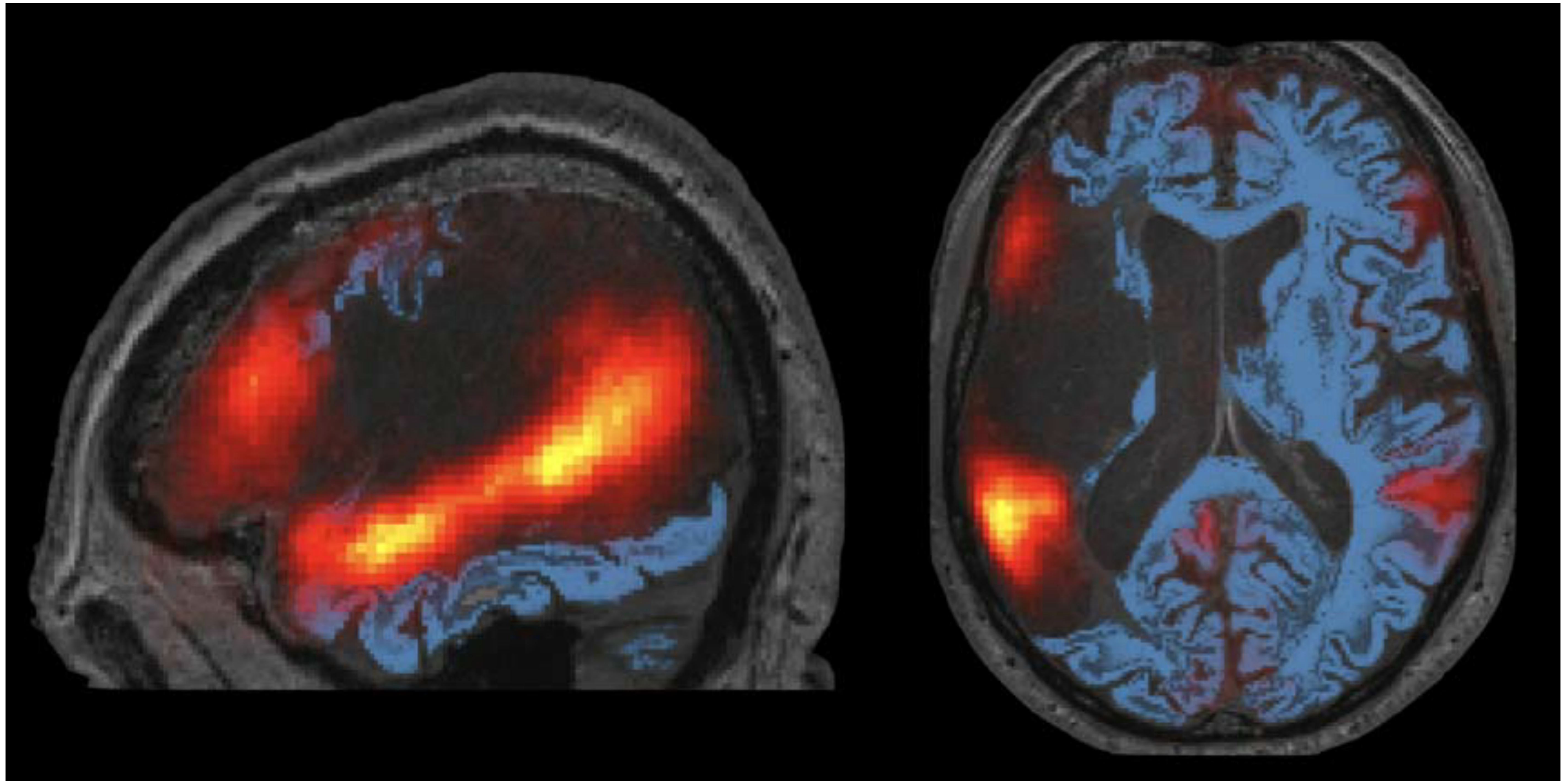
4. Causally important for language

Individuals with damage to the frontal and temporal lobes have deficits in understanding language and speaking (Broca, 1861; Wernicke, 1874; Luria, 1970; Mesulam et al., 2014).

Remarks on the Seat of the Faculty of Articulated Language, Following an Observation of Aphemia (Loss of Speech)

by Mr. Paul Broca (1861)
Surgeon of the Bicêtre Hosptial

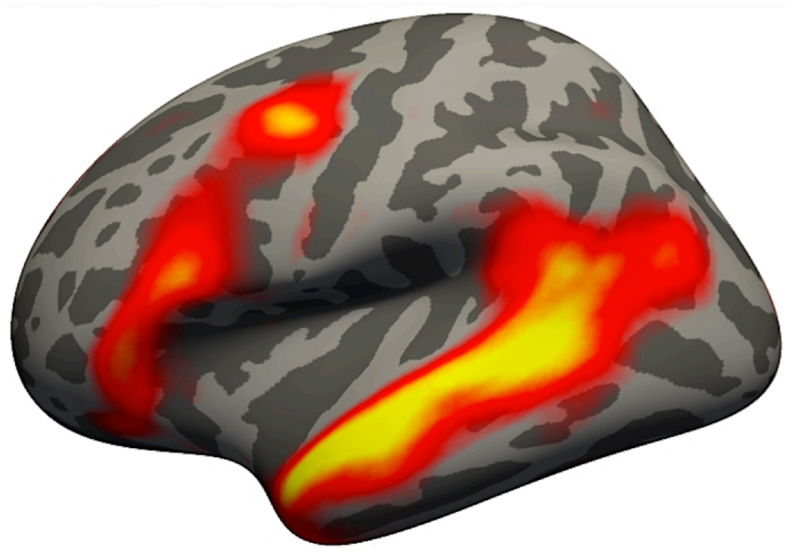


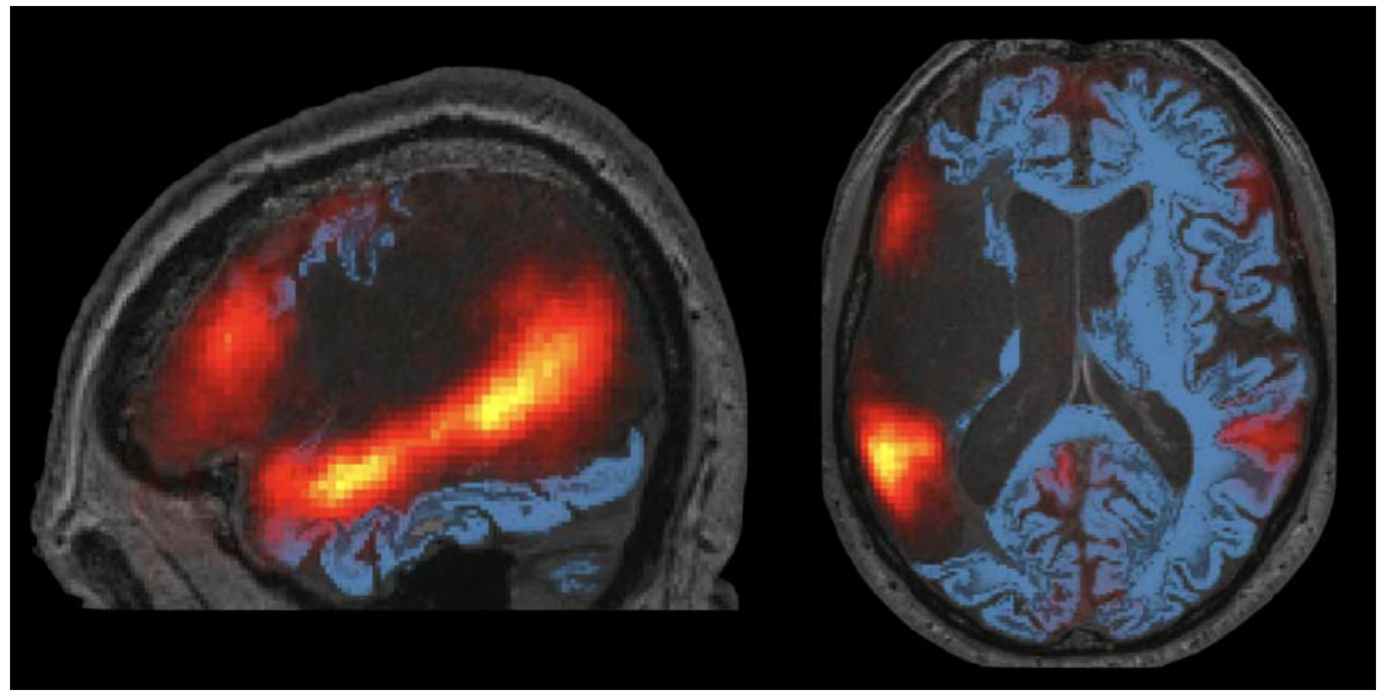


Ivanova et al., 2021

4. Causally important for language

We usually localize language “units” in humans:





Sentence

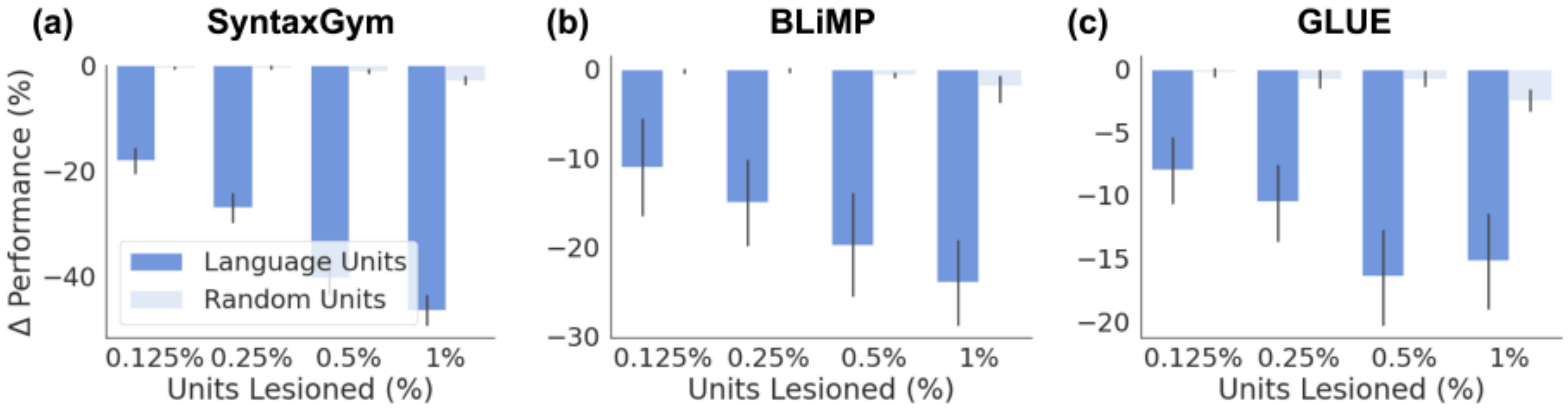
“In the starlight he could see the palm trees”

>

Non-word string

“Ap shab succeams ob elt lant meapon”

We can do it in LLMs too, and identify “LLM language units”. What happens if we ablate them?



AlKhamissi, Tuckute, Bosselut, Schrimpf, 2024

Outline

- What is language?
  - How can we identify language in the brain?
  - 4 key properties of the language network
- Models of language processing
  - How it started
  - How it continued
  - How it's going
  - So what?

Models of language processing

How it started

How it continued..

How it's going

Models of language processing

How it started *Descriptive models*
’Boxes and arrows’ (e.g., Lichtheim, 1885; Geschwind, 1965; Hickok & Poeppel, 2007)

How it continued..

Review

Broca and Wernicke are dead, or moving past the classic model of language neurobiology

Pascale Tremblay [a,b], Anthony Steven Dick [c,*] 2016

How it’s going

Classic “Wernicke-Lichtheim-Geschwind” model

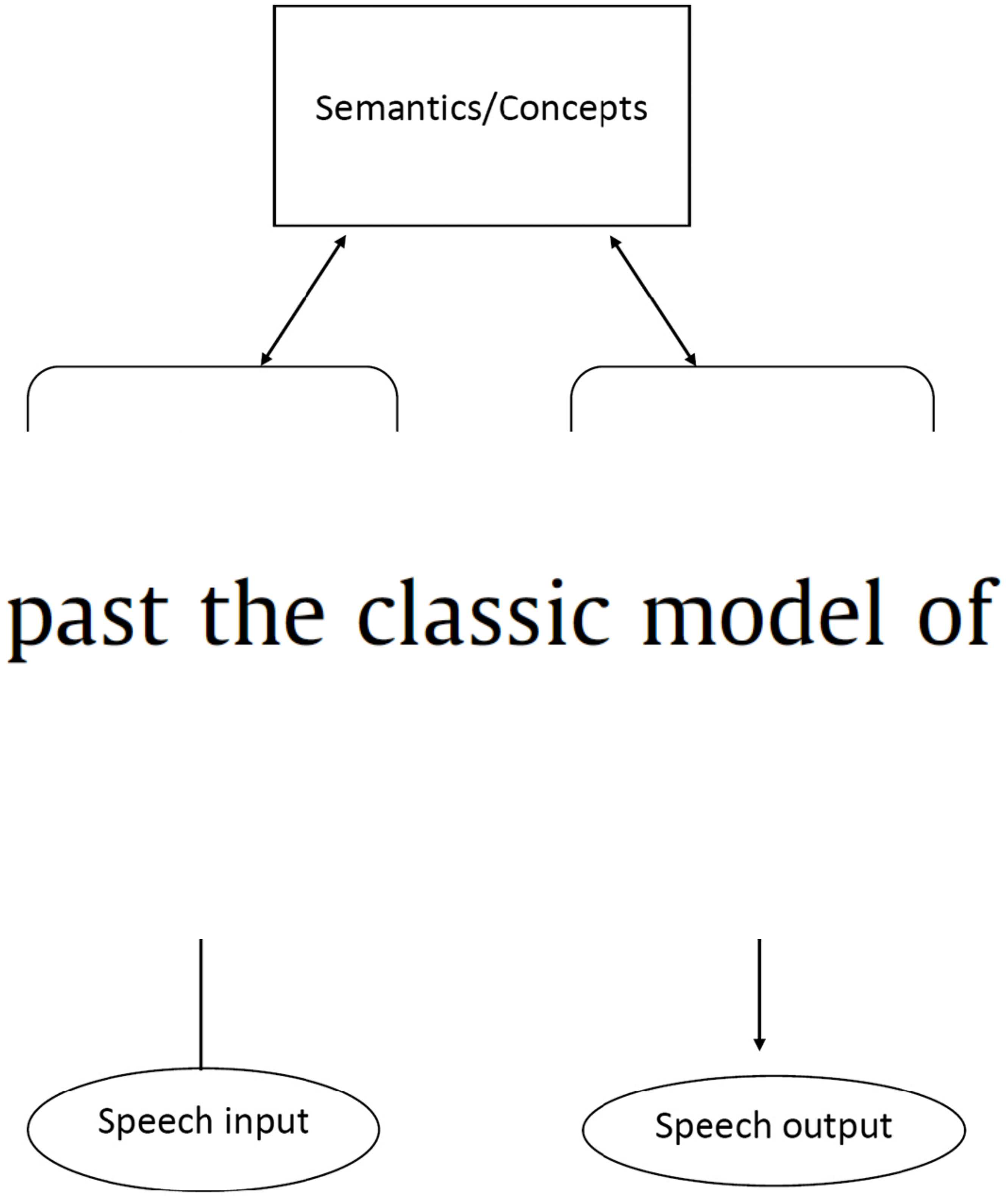


Illustration from Surbeck et al., 2020

Models of language processing

How it started *Descriptive models*
'Boxes and arrows' (e.g., Lichtheim, 1885; Geschwind, 1965; Hickok & Poeppel, 2007)

How it continued..
Co-occurrence word vectors

How it's going

Models of language processing

How it started *Descriptive models*
'Boxes and arrows' (e.g., Lichtheim, 1885; Geschwind, 1965; Hickok & Poeppel, 2007)

How it continued.. *Quantitative models*
Co-occurrence word vectors (e.g., Mitchell et al., 2008; Devereux et al., 2010; Pereira et al., 2011)

How it's going

Predicting Human Brain Activity Associated with the Meanings of Nouns 2008

Tom M. Mitchell,[1*] Svetlana V. Shinkareva,[2] Andrew Carlson,[1] Kai-Min Chang,[3,4] Vicente L. Malave,[5] Robert A. Mason,[3] Marcel Adam Just[3]

Predictive model

"... they [older theories] are primarily descriptive theories that make no attempt to predict the specific brain activation that will be produced when a human subject reads a particular word or views a drawing of a particular object."

A word's co-occurrence with 25 verbs extracted from a trillion-word text corpus

Mapping learned from fMRI training data

Models of language processing

How it started *Descriptive models*
'Boxes and arrows' (e.g., Lichtheim, 1885; Geschwind, 1965; Hickok & Poeppel, 2007)

How it continued.. *Quantitative models*
Co-occurrence word vectors (e.g., Mitchell et al., 2008; Devereux et al., 2010; Pereira et al., 2011)

Recurrent neural networks

How it's going

Models of language processing

How it started *Descriptive models*
'Boxes and arrows' (e.g., Lichtheim, 1885; Geschwind, 1965; Hickok & Poeppel, 2007)

How it continued.. *Quantitative models*
Co-occurrence word vectors (e.g., Mitchell et al., 2008; Devereux et al., 2010; Pereira et al., 2011)

Recurrent neural networks (e.g., Wehbe et al., 2014; Qian et al., 2016; Jain & Huth, 2018)

How it's going
LLMs (Transformers) (e.g., Schrimpf et al., 2021; Caucheteux & King, 2022; Antonello et al., 2024; *inter alia*)

Models of language processing

How it started *Descriptive models*
'Boxes and arrows' (e.g., Lichtheim, 1885; Geschwind, 1965; Hickok & Poeppel, 2007)

How it continued.. *Quantitative models*
Co-occurrence word vectors (e.g., Mitchell et al., 2008; Devereux et al., 2010; Pereira et al., 2011)

Recurrent neural networks (e.g., Wehbe et al., 2014; Qian et al., 2016; Jain & Huth, 2018)

How it's going
LLMs (Transformers) (e.g., Schrimpf et al., 2021; Caucheteux & King, 2022; Antonello et al., 2024; *inter alia*)

A priori, why might we expect LLMs to capture something about human language processing?

- Formal linguistic competence.
- Sensitivity to multiple levels of linguistic structure.
- Acquiring language does not necessarily require learning formal reasoning.

Models of language processing

How it started *Descriptive models*
'Boxes and arrows' (e.g., Lichtheim, 1885; Geschwind, 1965; Hickok & Poeppel, 2007)

How it continued.. *Quantitative models*
Co-occurrence word vectors (e.g., Mitchell et al., 2008; Devereux et al., 2010; Pereira et al., 2011)

Recurrent neural networks (e.g., Wehbe et al., 2014; Qian et al., 2016; Jain & Huth, 2018)

How it's going
LLMs (Transformers) (e.g., Schrimpf et al., 2021; Caucheteux & King, 2022; Antonello et al., 2024; *inter alia*)

A priori, why might we expect LLMs to capture something about human language processing?

- Formal linguistic competence.
- Sensitivity to multiple levels of linguistic structure.
- Acquiring language does not necessarily require learning formal reasoning.

And why might we not?

- Language learning: Amount/kinds of training data and interaction with the world more broadly.
- Operate on different input signals (tokens).
- Memory of preceding linguistic input.
- Hardware differences.

LLMs as models of language processing

First large-scale claims

New Results

View current version of this article

Follow this preprint

Previous

Posted June 27, 2020.

Artificial Neural Networks Accurately Predict Language Processing in the Brain

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, Evelina Fedorenko

Download PDF

Print/Save Options

New Results

View current version of this article

Follow this preprint

Previous

Posted July 04, 2020.

Language processing in brains and deep neural networks: computational convergence and its limits

Charlotte Caucheteux, Jean-Rémi King

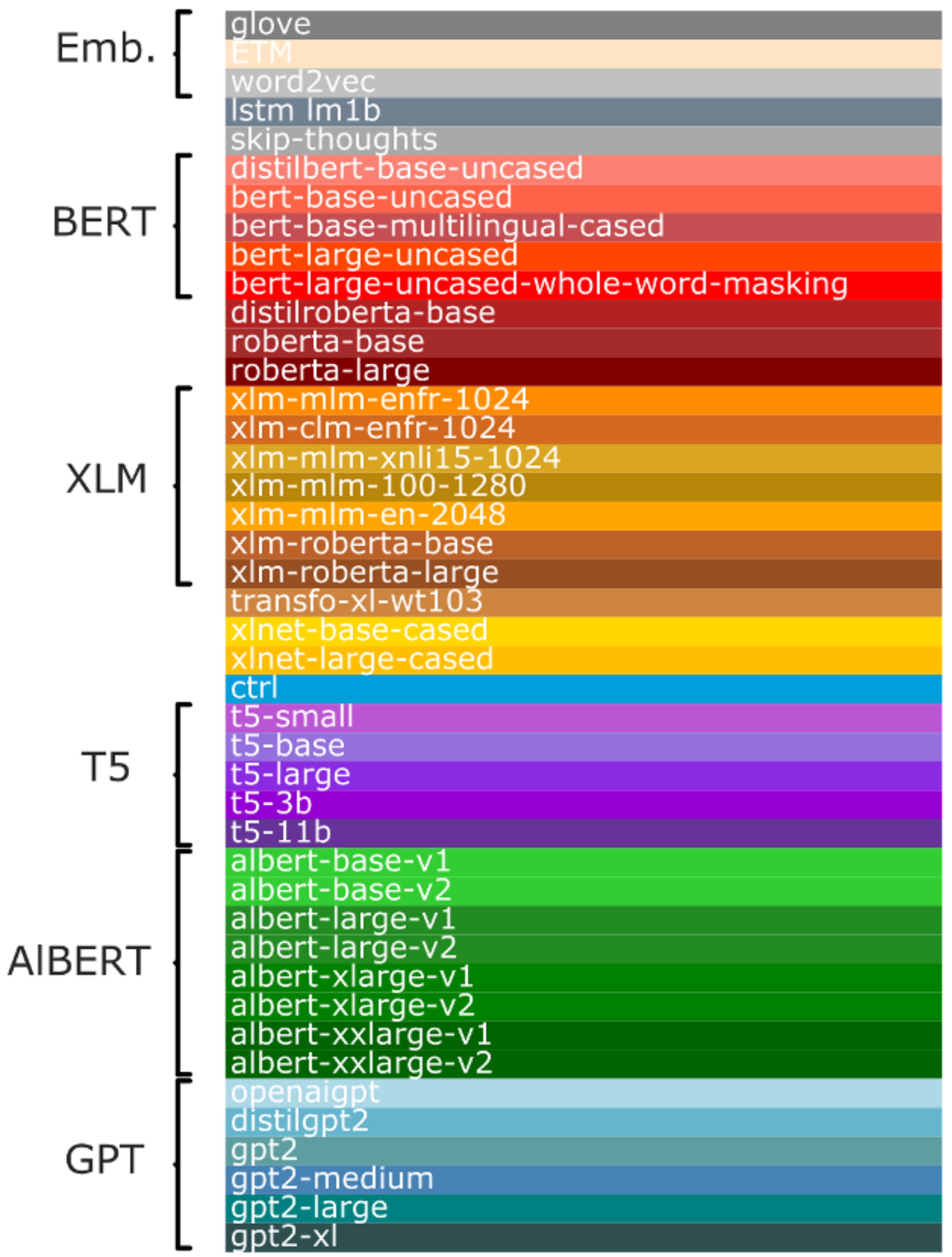
Download PDF

Print/Save Options

LLMs as models of language processing

First large-scale claims Schrimpf et al., 2021

Models



n = 43

Brain/behavior

fMRI — **Pereira et al., 2018, *n* = 10**

fMRI — **Blank et al., 2014, *n* = 5**

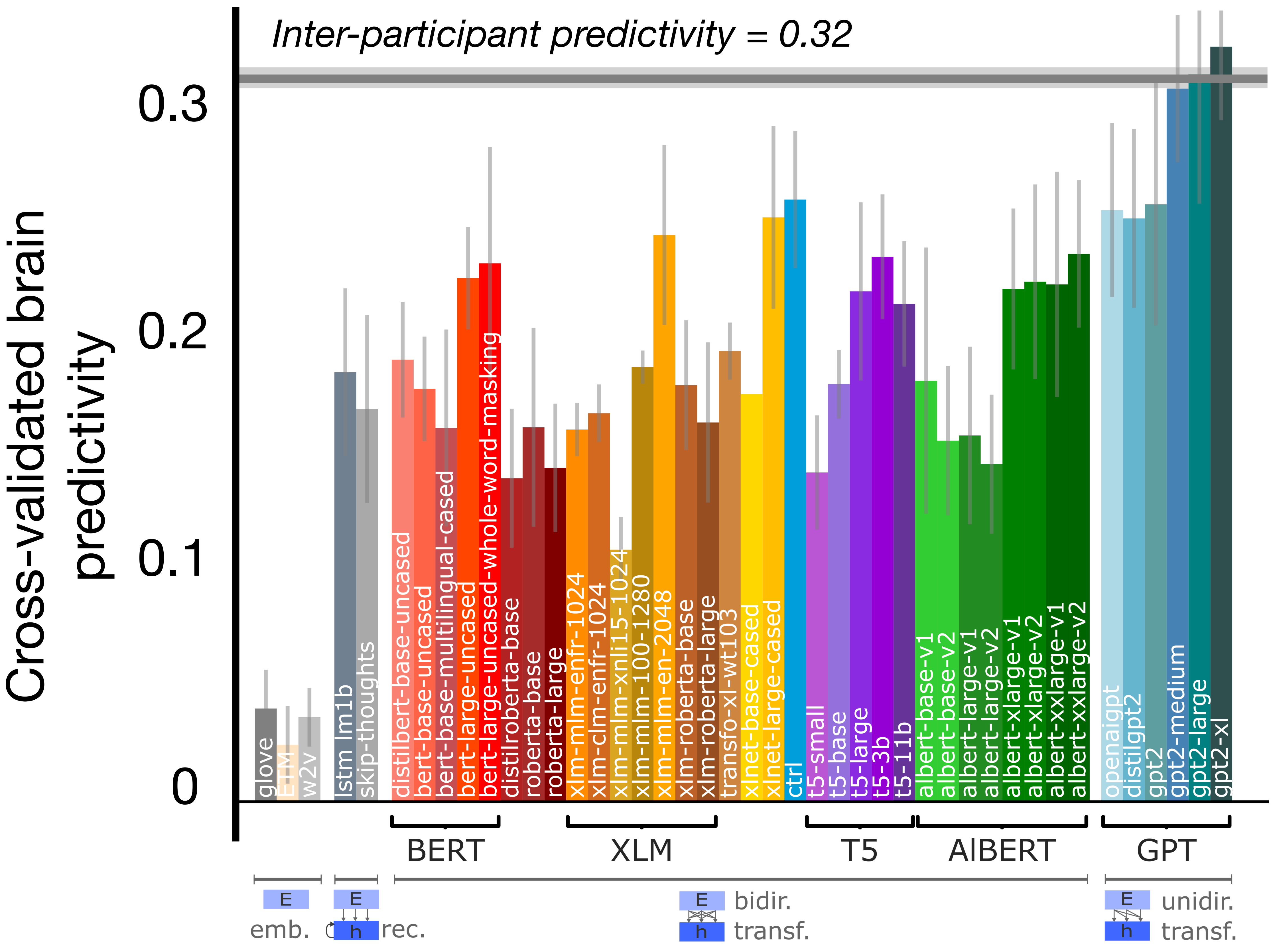
ECoG — **Fedorenko et al., 2016, *n* = 5**

Behavior — **Futrell et al., 2018, *n* = 179**

LLMs as models of language processing

First large-scale claims Schrimpf et al., 2021

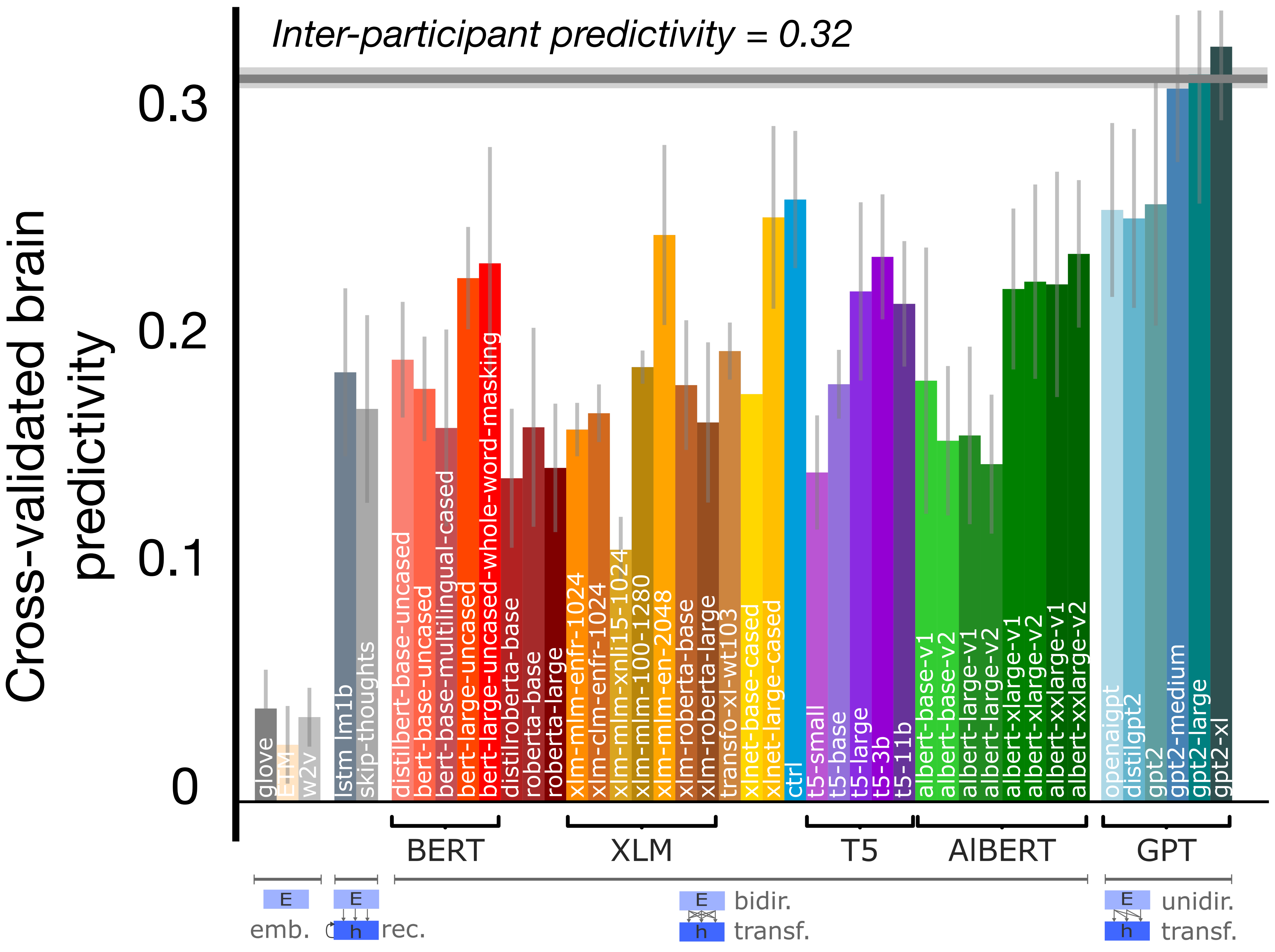
Pereira et al., 2018



LLMs as models of language processing

First large-scale claims Schrimpf et al., 2021

Pereira et al., 2018

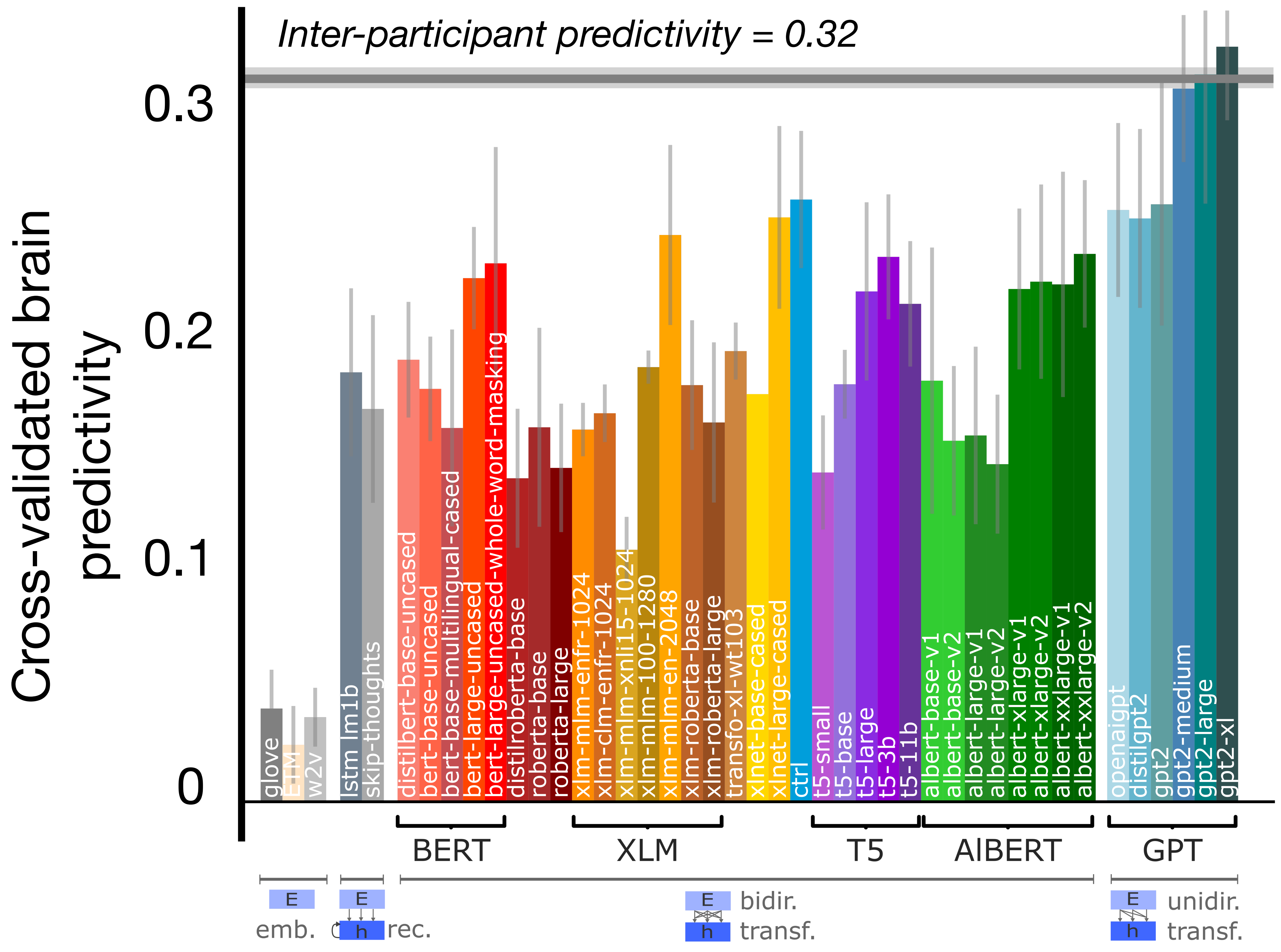


- *Representational* similarity between models and brain responses: Transformer models outperform word embedding models, and there is variability across models.
- Why are certain models better than others? *We will get to that in a bit*

LLMs as models of language processing

First large-scale claims Schrimpf et al., 2021

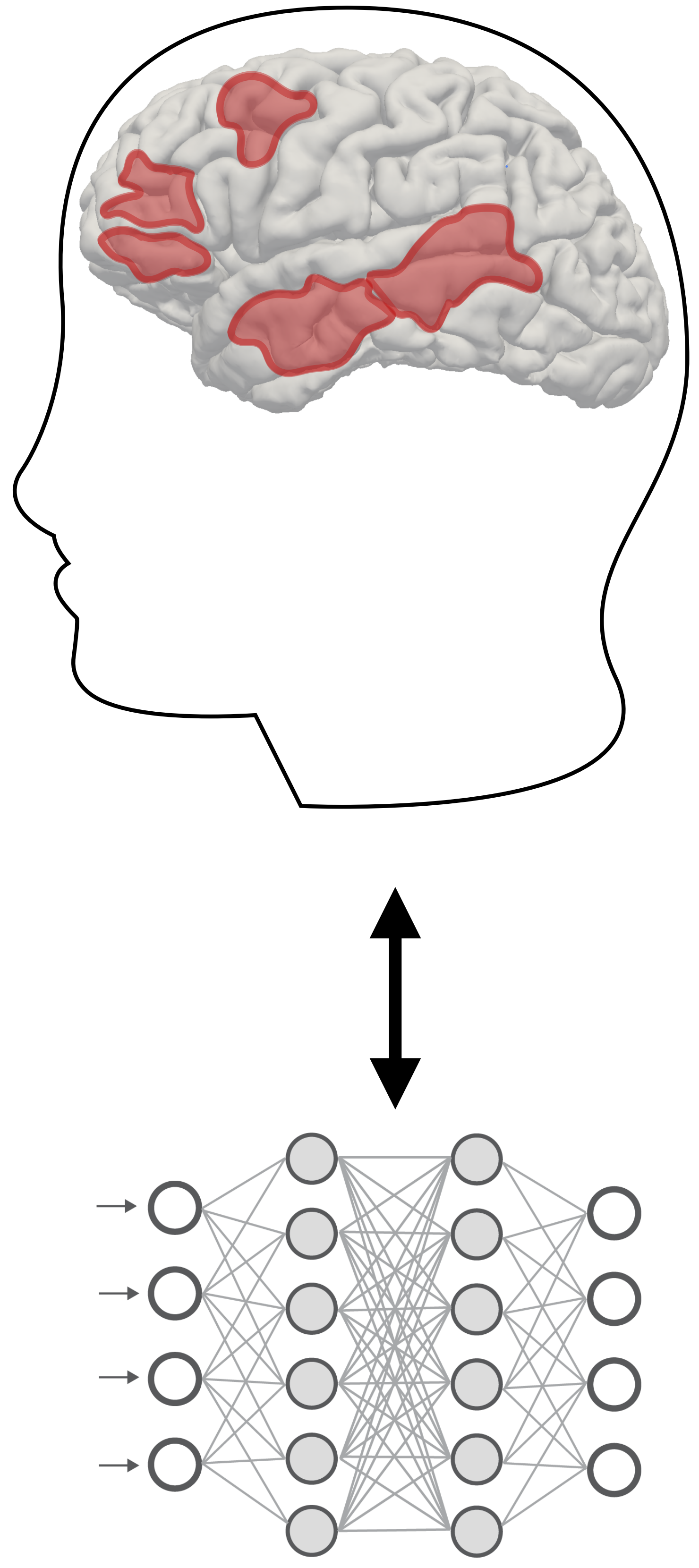
Pereira et al., 2018



This apparent similarity is because LLMs capture something idiosyncratic about brains… or something idiosyncratic about LLM model representations...

Driving and suppressing brain responses

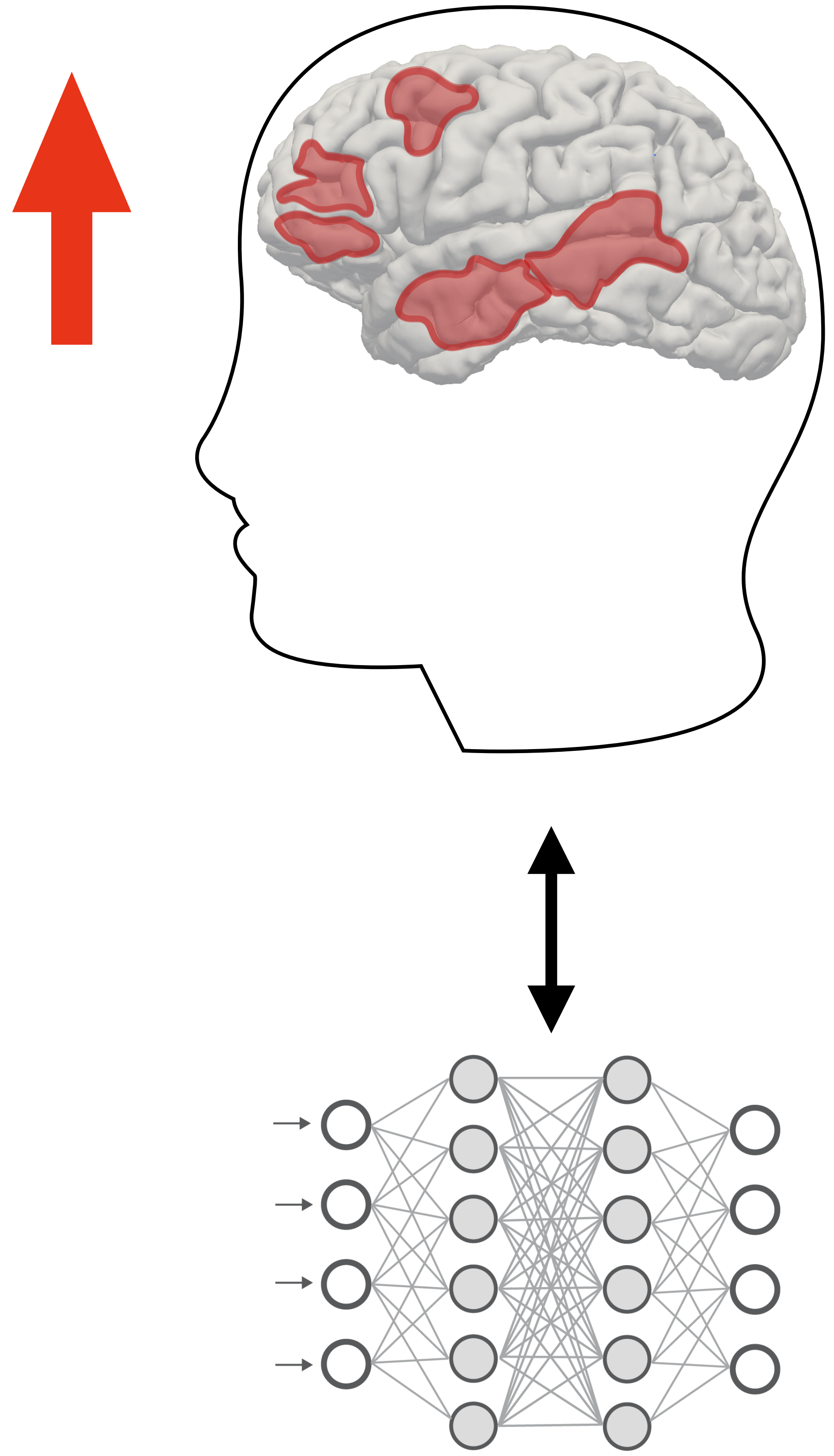
LLMs are predictive of brain responses during language processing.



Driving and suppressing brain responses

LLMs are predictive of brain responses during language processing.

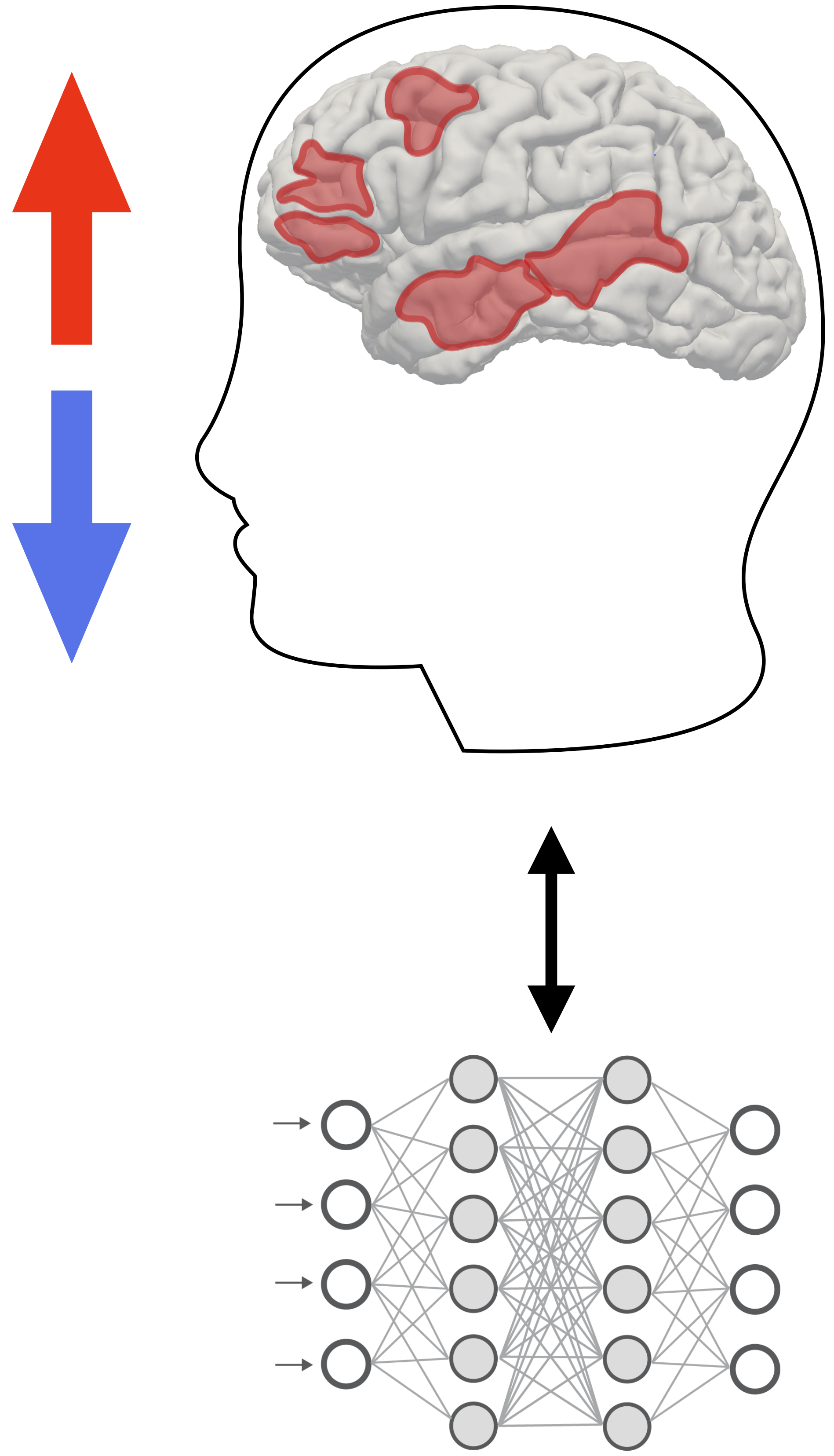
Can we leverage the predictive power of LLMs to identify new stimuli to maximally **drive** or suppress brain responses in the language network of new individuals?



Driving and suppressing brain responses

LLMs are predictive of brain responses during language processing.

Can we leverage the predictive power of LLMs to identify new stimuli to maximally **drive** or **suppress** brain responses in the language network of new individuals?

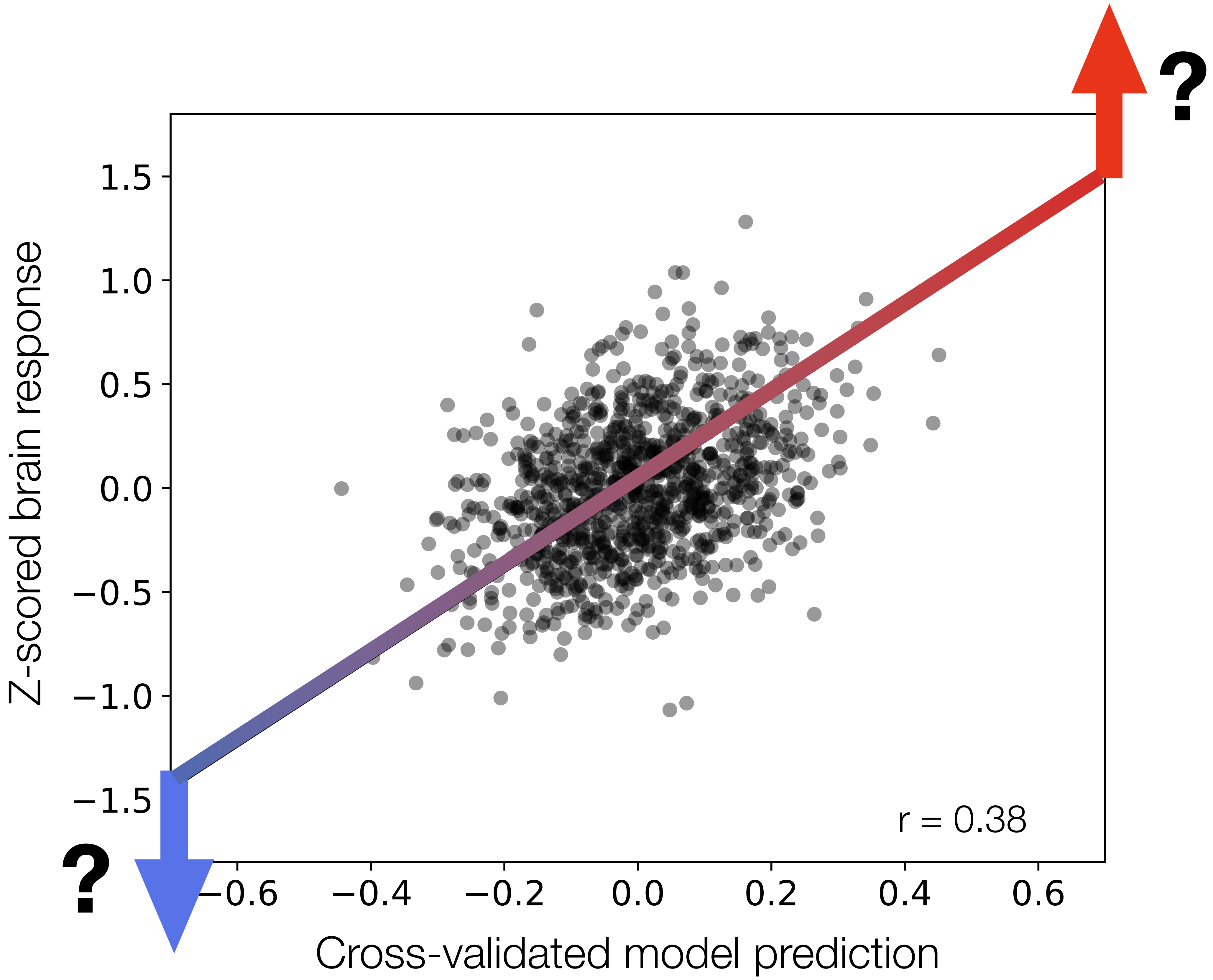


Driving and suppressing brain responses

Cross-validated model performance: 5 participants, 1,000 sentences each (noise ceiling: r = 0.56)

Use LLM-based encoding model to identify sentences that elicit maximal activity (drive sentences) or minimal activity (suppress sentences)

Search across 1.8M sentences from existing corpora

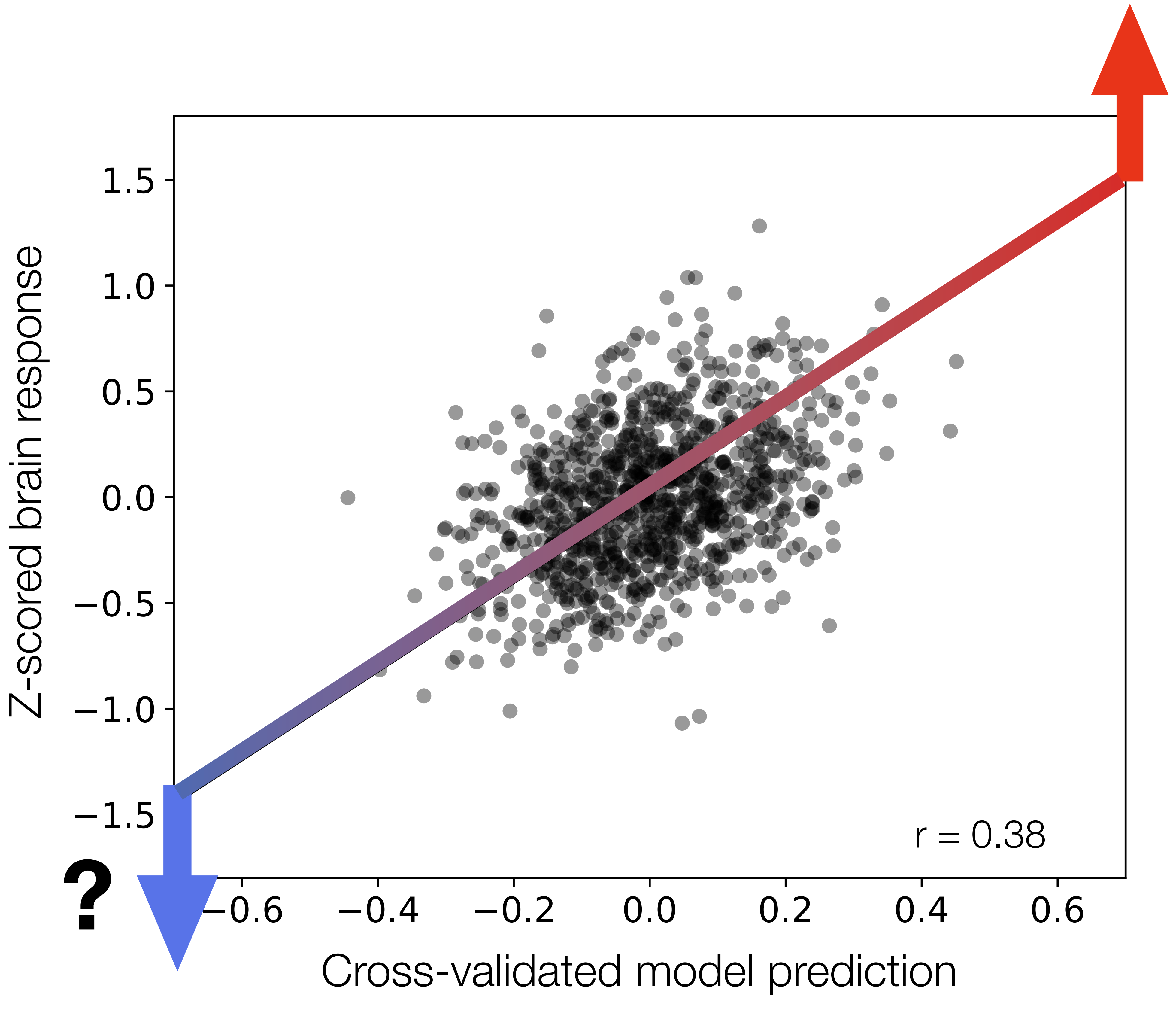


Tuckute et al., 2024

Driving and suppressing brain responses

Cross-validated model performance: 5 participants, 1,000 sentences each (noise ceiling: r = 0.56)

Use LLM-based encoding model to identify sentences that elicit maximal activity (drive sentences) or minimal activity (suppress sentences)



Drive (n=250)

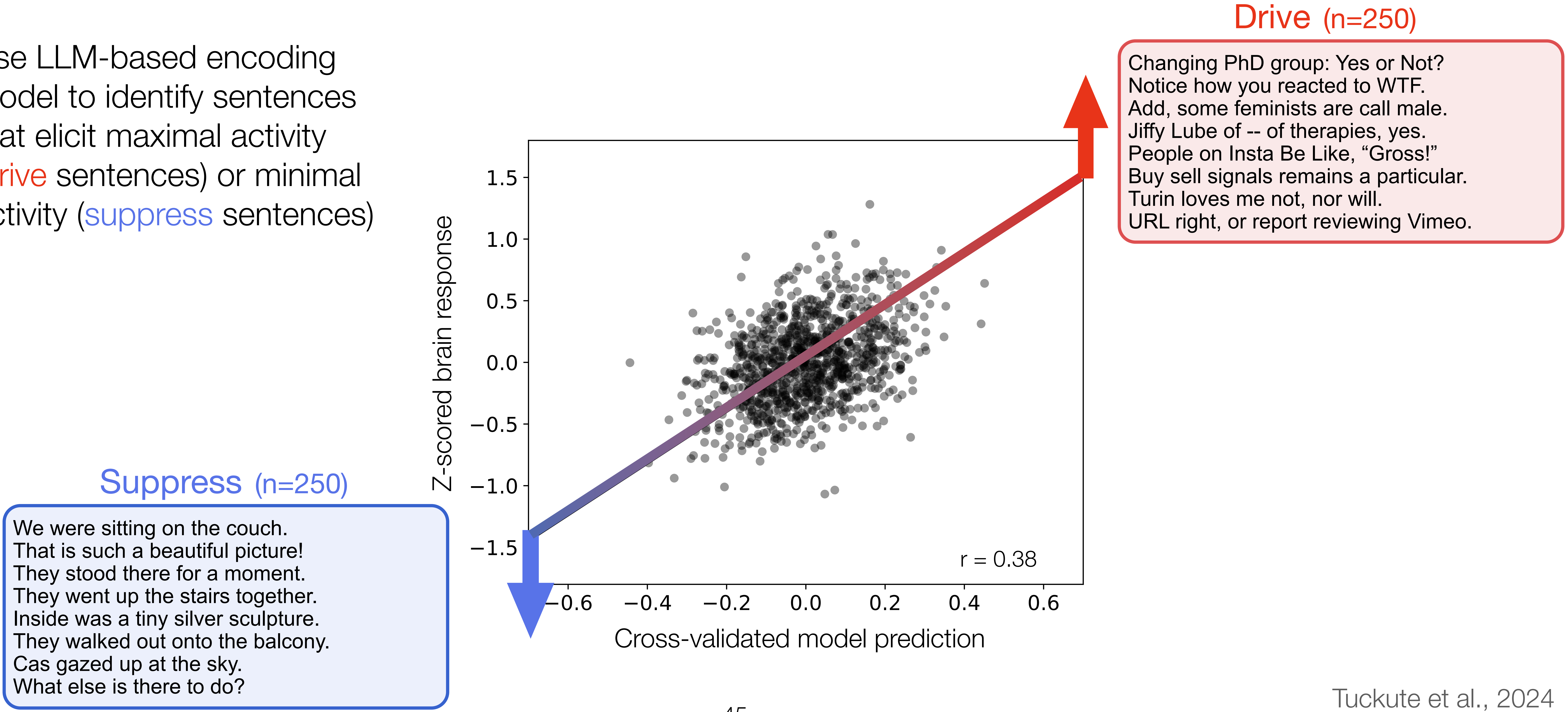
Changing PhD group: Yes or Not?
Notice how you reacted to WTF.
Add, some feminists are call male.
Jiffy Lube of -- of therapies, yes.
People on Insta Be Like, “Gross!”
Buy sell signals remains a particular.
Turin loves me not, nor will.
URL right, or report reviewing Vimeo.

Tuckute et al., 2024

Driving and suppressing brain responses

Cross-validated model performance: 5 participants, 1,000 sentences each (noise ceiling: r = 0.56)

Use LLM-based encoding model to identify sentences that elicit maximal activity (drive sentences) or minimal activity (suppress sentences)



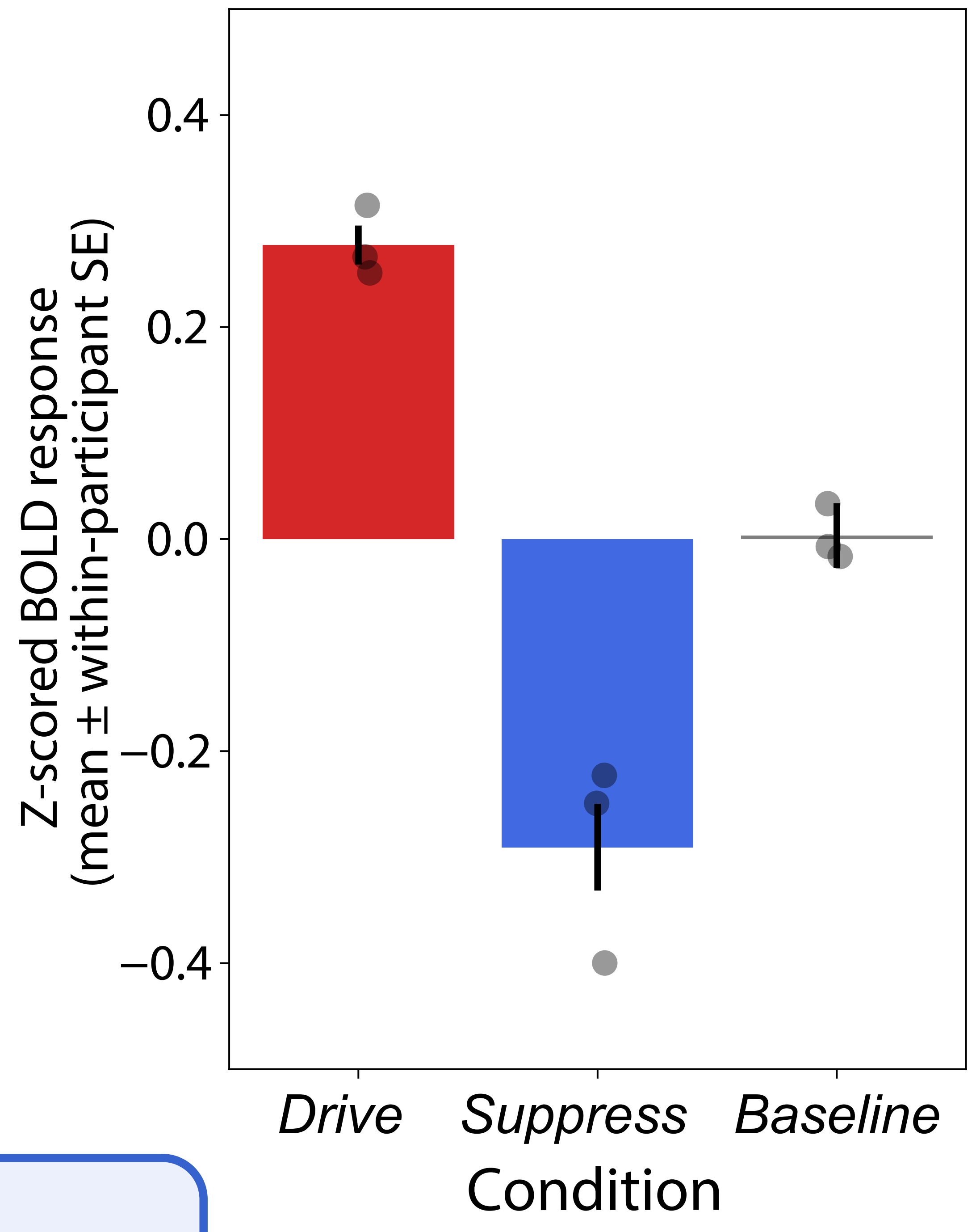
Tuckute et al., 2024

Driving and suppressing brain responses

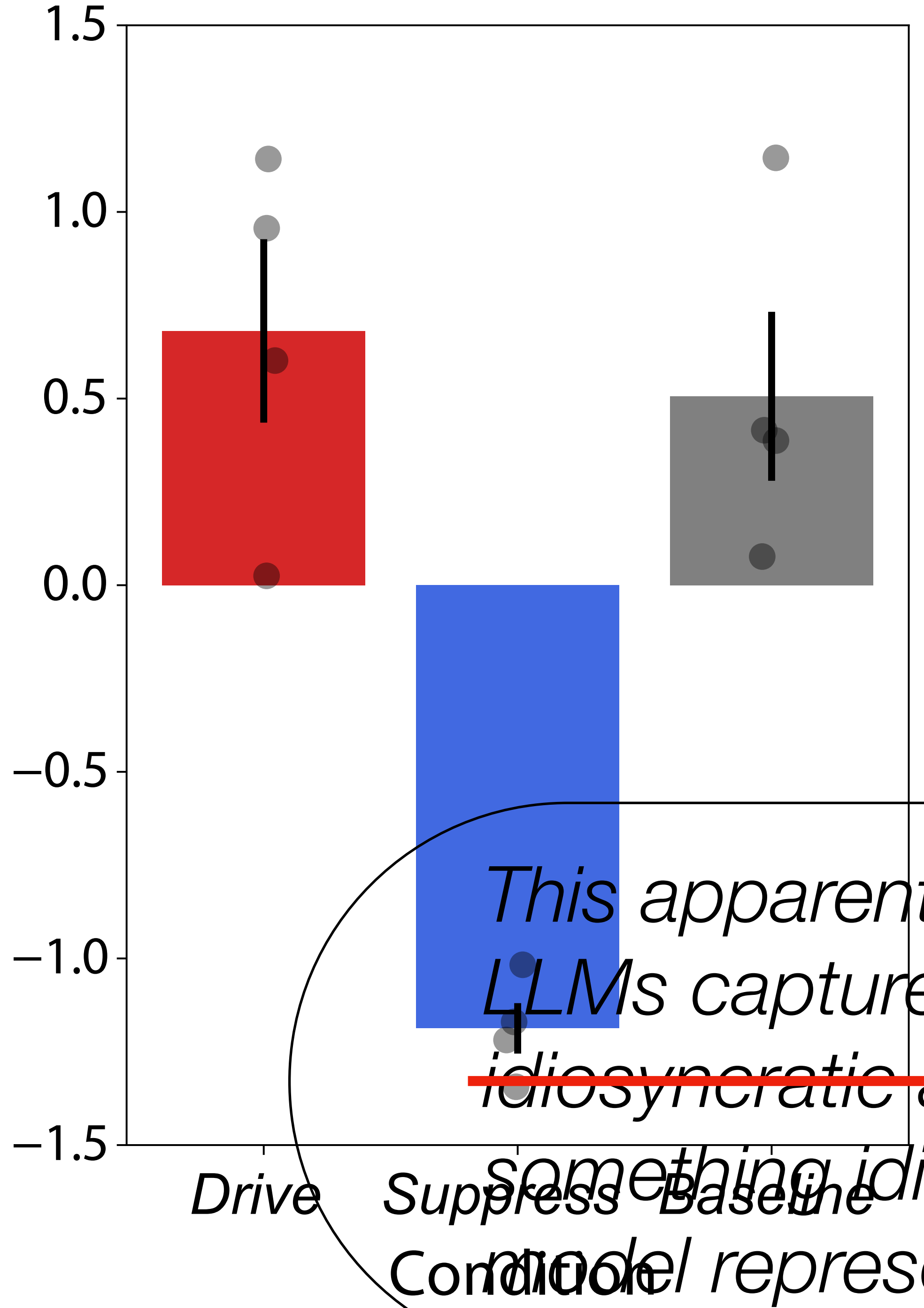
Modulation of brain responses

Collect brain responses to these new sentences in new individuals.

Successful modulation of brain responses to language in a closed-loop manner *across individuals*.



n=3 participants



n=4 participants

Drive (n=250)

Changing PhD group: Yes or Not?
Notice how you reacted to WTF.
Add, some feminists are call male.
Jiffy Lube of -- of therapies, yes.
People on Insta Be Like, "Gross!"
Buy sell signals remains a particular.
Turin loves me not, nor will.
URL right, or report reviewing Vimeo.

This apparent similarity is because LLMs capture something ~~idiosyncratic about brains~~… or something idiosyncratic about LLM model representations...

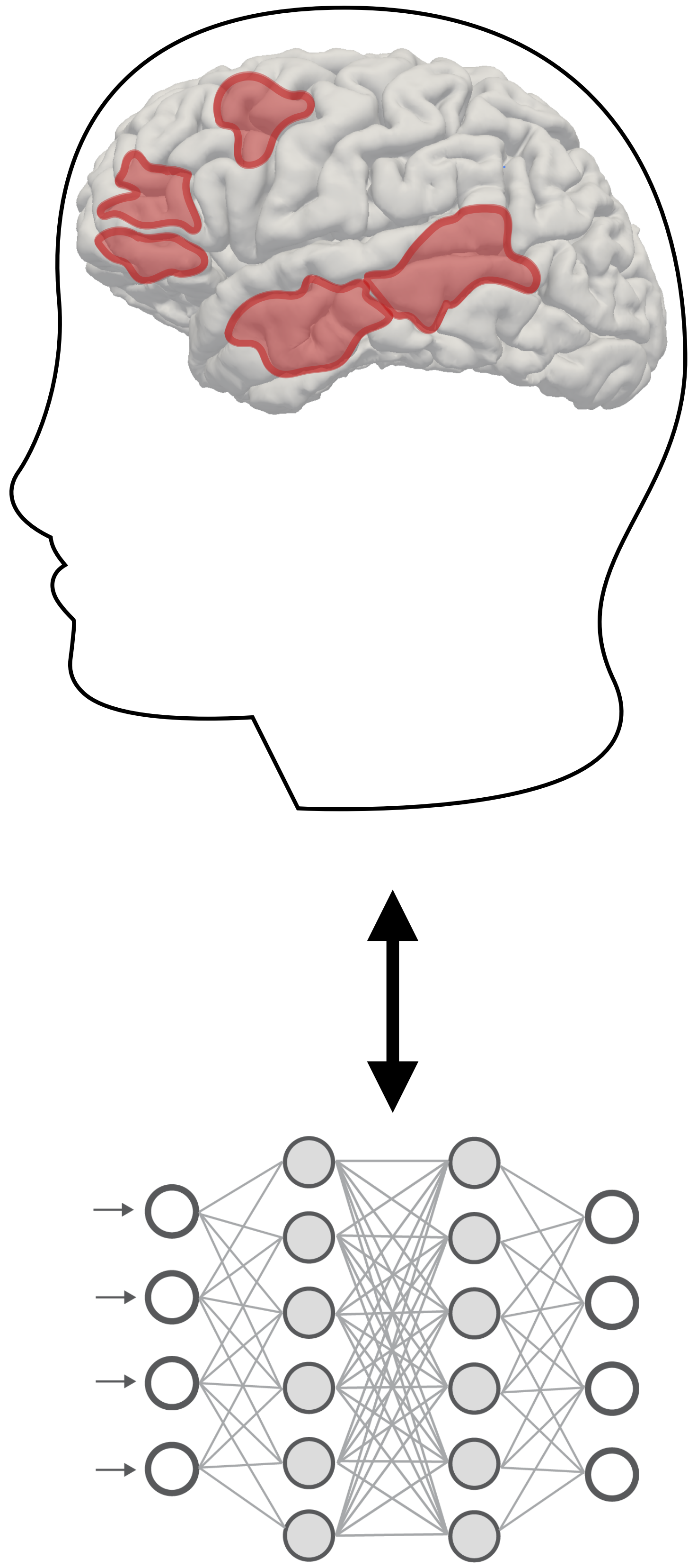
Suppress (n=250)

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?

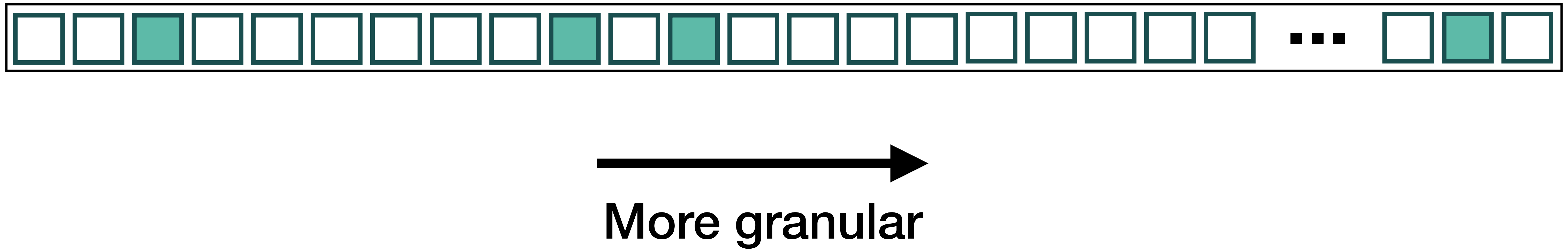
Feature importance in LLM-brain alignment

Mechanistic interpretability gives us tools to decompose what LLMs have learned into interpretable, hierarchically organized features (Bussmann et al., 2025).

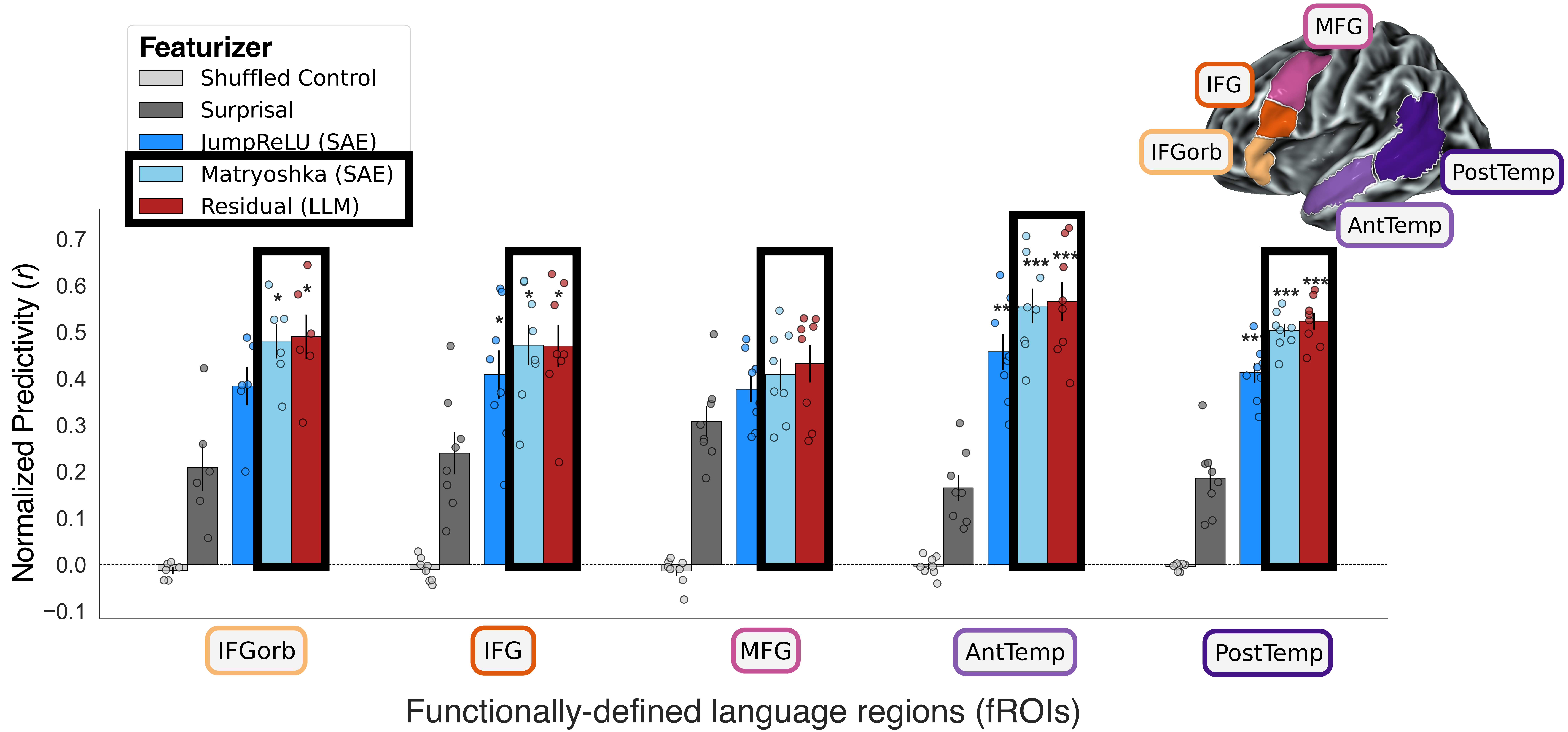
With Mikey Lepori!



SAE features are organized at varying levels of granularity, from coarse (= **widely applicable**) to more granular (= **residuals**).



Feature importance in LLM-brain alignment



Lepori, Kay, Tuckute, *in prep*

Feature importance in LLM-brain alignment

Brain responses in the language network are largely explained by the coarse, widely-applicable features of LLM representations—not by idiosyncratic residual variation.

This apparent similarity is because LLMs capture something ~~idiosyncratic about brains~~… or ~~something idiosyncratic about LLM model representations…~~

Lepori, Kay, Tuckute, *in prep*

Outline

- What is language?
  - How can we identify language in the brain?
  - 4 key properties of the language network
- Models of language processing
  - How it started
  - How it continued
  - How it's going
  - So what?

So what?

- **Representational content of the language network:**
  Distilling the properties of LLMs that enable them to capture human responses to language.
- **Engineering applications & model-guided experiments:**
  Leveraging LLMs for decoding, simulation, and neural control.

Distilling LLM properties

A. Model architecture

- General architecture class (e.g., Transformer, RNN)
- Number of layers (e.g., Transformer blocks)
- Number of parameters (number of learnable parameters in the model)
- Embedding dimensionality (the vector size for each token)

B. Model behavior

- Performance on the main training objective (typically next-word prediction)
- Performance on other tasks (e.g., sentiment analysis, summarization)
- Performance on experimentally altered inguistic input

C. Model training

- Training objective
  - Main training objective
    - Causal/masked word prediction
    - Other (less common) objectives
  - Fine-tuning training objective
- Training data
  - Data amount (number of tokens)
  - Data type (e.g., different modalities, content types, languages)

Distilling LLM properties

A. Model architecture

- General architecture class (e.g., Transformer, RNN)
- Number of layers (e.g., Transformer blocks)
- Number of parameters (number of learnable parameters in the model)
- Embedding dimensionality (the vector size for each token)

B. Model behavior

- Performance on the main training objective (typically next-word prediction)
- Performance on other tasks (e.g., sentiment analysis, summarization)
- Performance on experimentally altered inguistic input

C. Model training

- Training objective
  - Main training objective
    - Causal/masked word prediction
    - Other (less common) objectives
  - Fine-tuning training objective
- Training data
  - Data amount (number of tokens)
  - Data type (e.g., different modalities, content types, languages)

Distilling LLM properties

A. Model architecture

- General architecture class (e.g., Transformer, RNN)
- Number of layers (e.g., Transformer blocks)
- Number of parameters (number of learnable parameters in the model)
- Embedding dimensionality (the vector size for each token)

B. Model behavior

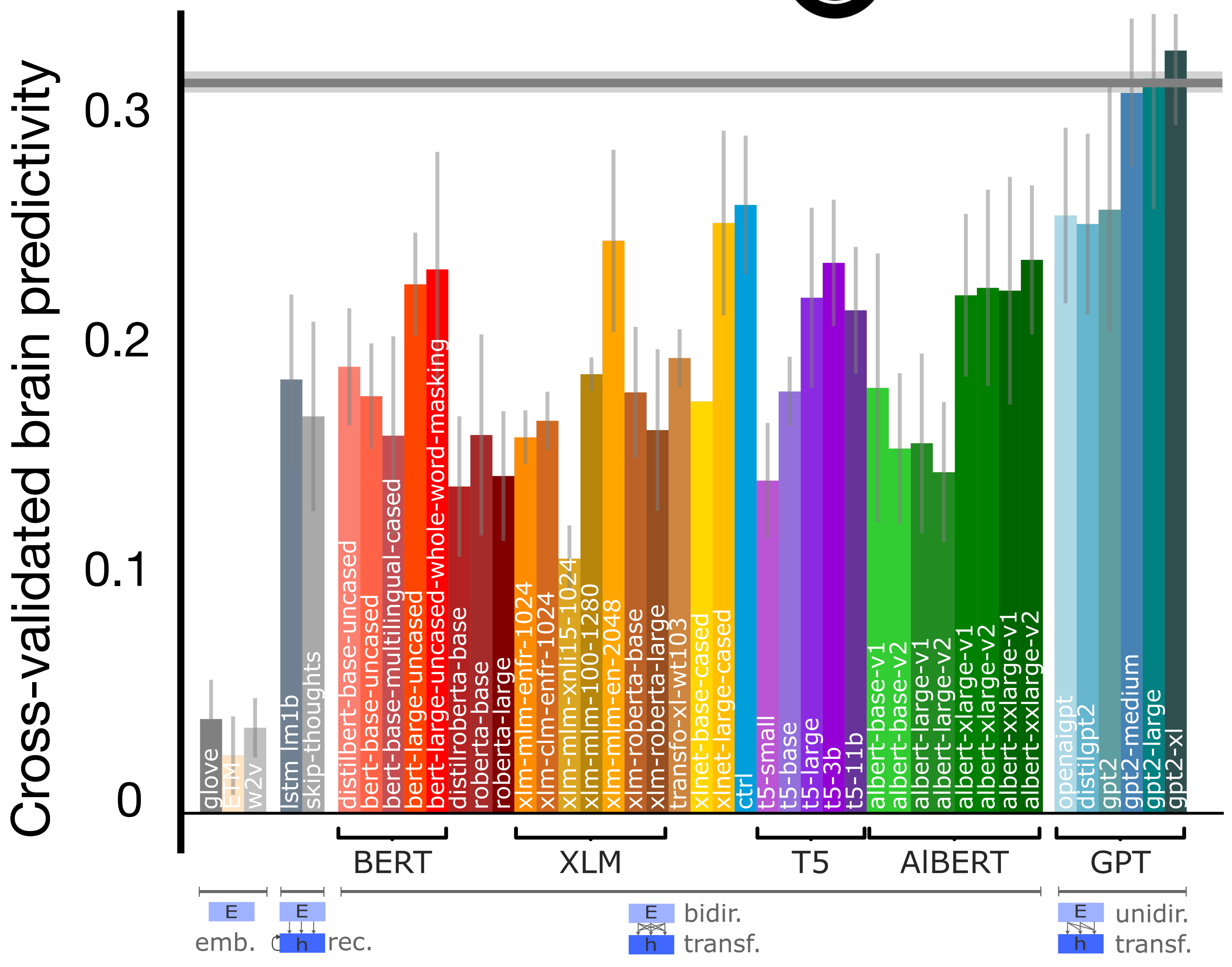
- Performance on the main training objective (typically next-word prediction)
- Performance on other tasks (e.g., sentiment analysis, summarization)
- Performance on experimentally altered inguistic input

C. Model training

- Training objective
  - Main training objective
    - Causal/masked word prediction
    - Other (less common) objectives
  - Fine-tuning training objective
- Training data
  - Data amount (number of tokens)
  - Data type (e.g., different modalities, content types, languages)

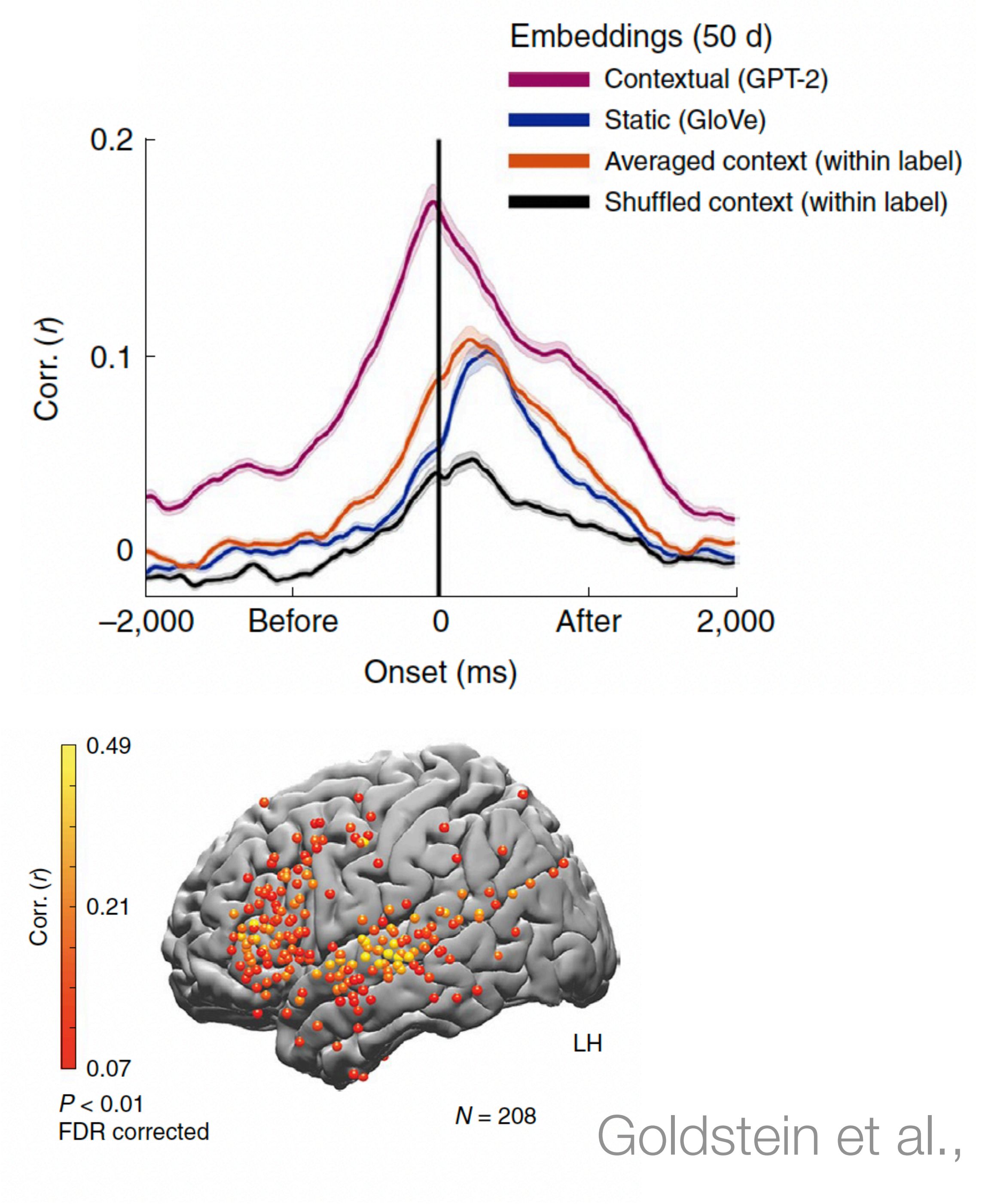
- Many LLM architecture instantiations fit brain data well; no particular architectural motif critically modulates model-brain alignment.
  - Decontextualized models align poorly to the brain.
  - If anything, the clearest pattern is causal > masked.

General architecture class



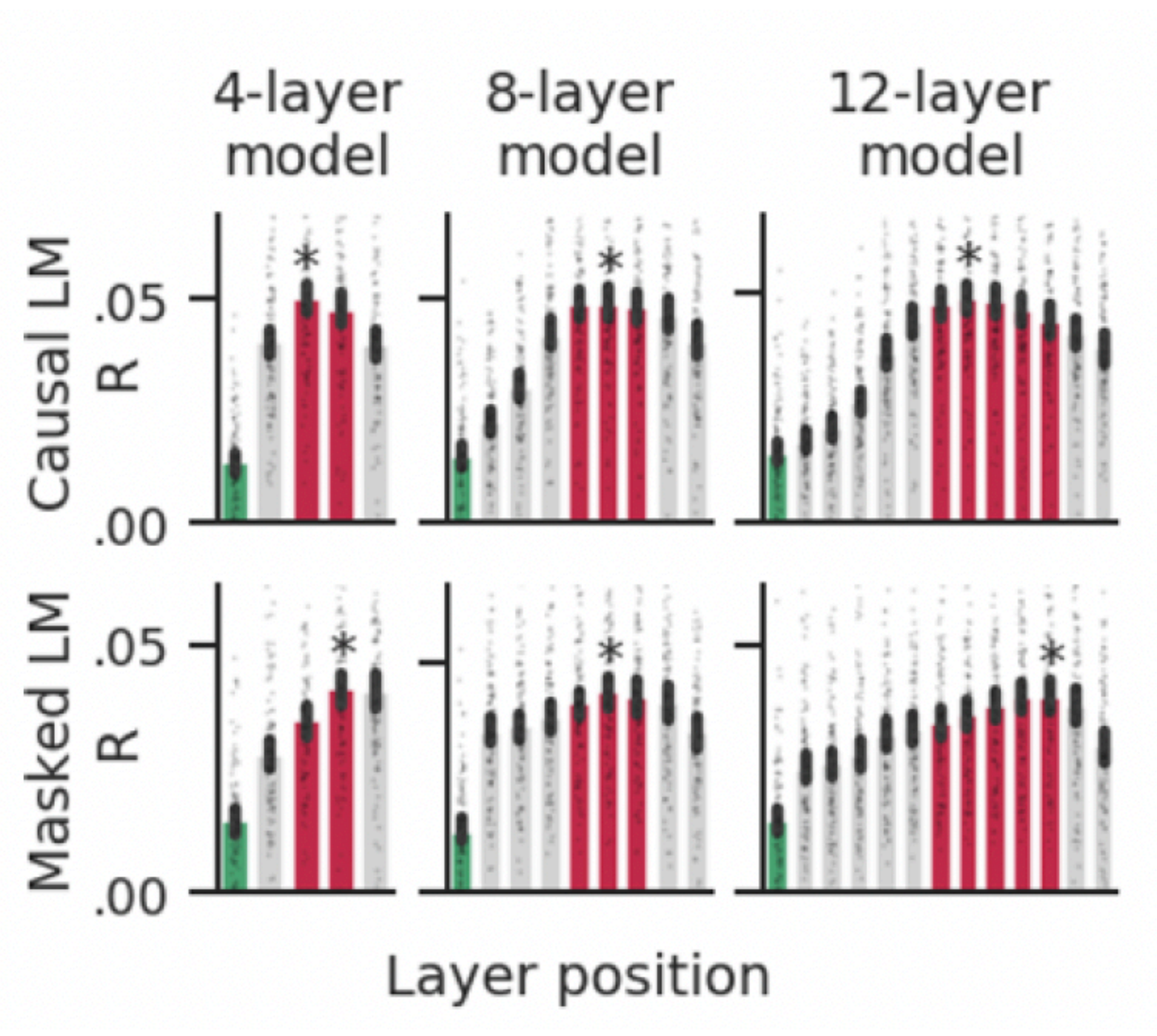
Schrimpf et al., 2021

Contextual vs. decontextualized



Goldstein et al., 2022

Causal vs. masked

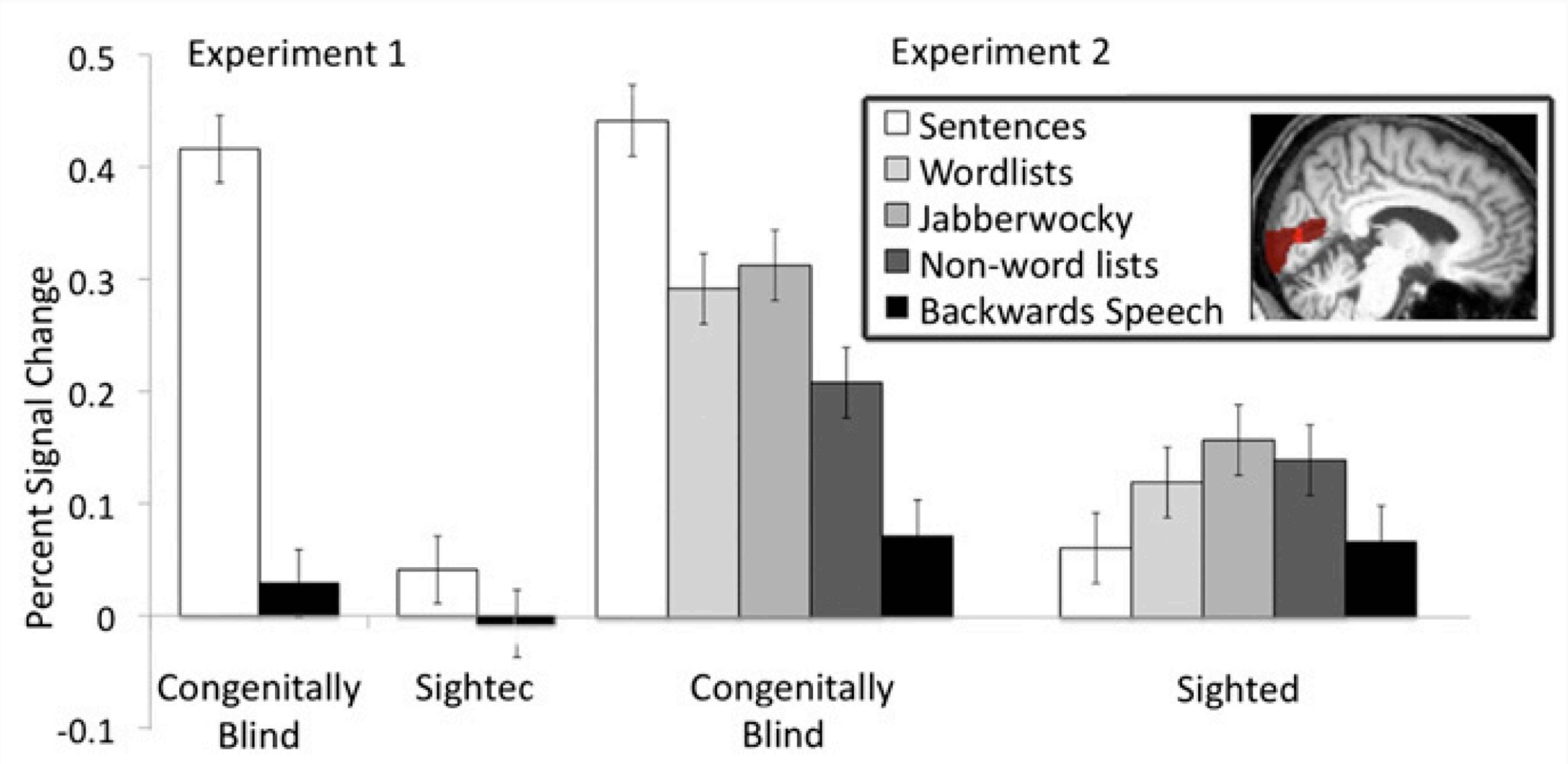


Caucheteux & King, 2022

Distilling LLM properties

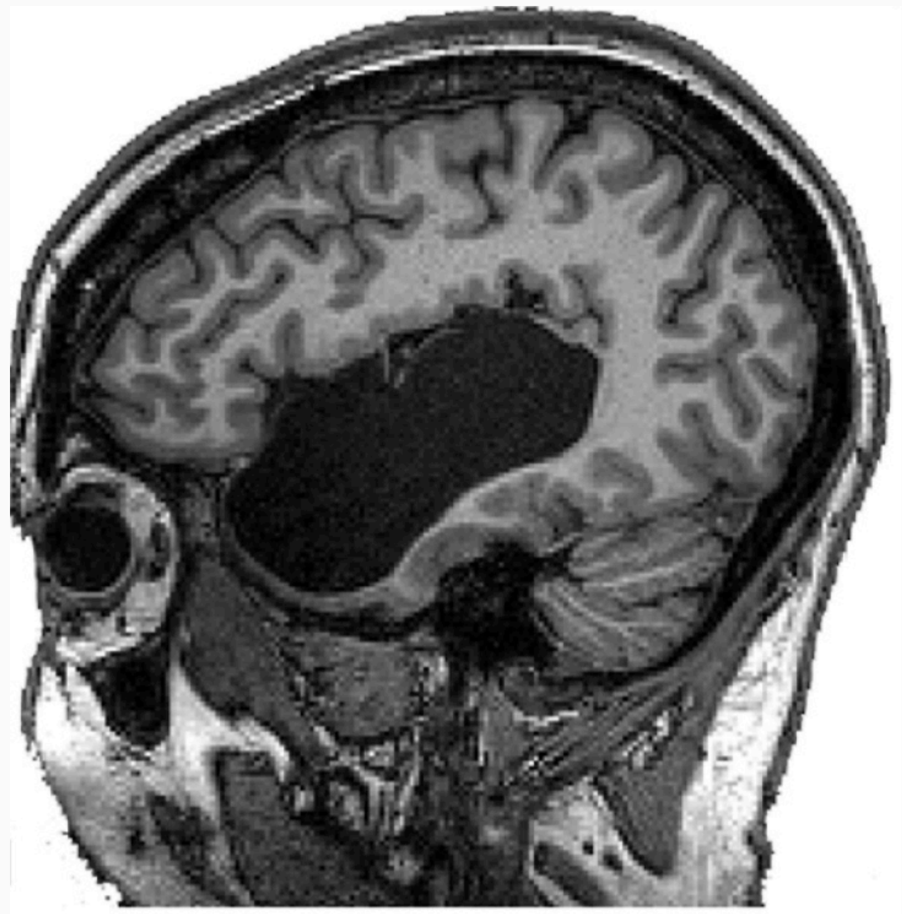
Humans can also learn language under very different neural "architectures"…

The case of blind individuals



Bedny et al., 2011

or large (early) lesions



Tuckute et al., 2022

- Many LLM architecture instantiations fit brain data well; **no particular architectural motif** critically modulates model-brain alignment.
  - Decontextualized models align poorly to the brain.
  - If anything, the clearest pattern is causal > masked.

Schrimpf et al., 2021

Goldstein et al., 2022

Causal vs. masked

Caucheteux & King, 2022

Distilling LLM properties

A. Model architecture

- General architecture class (e.g., Transformer, RNN)
- Number of layers (e.g., Transformer blocks)
- Number of parameters (number of learnable parameters in the model)
- Embedding dimensionality (the vector size for each token)

B. Model behavior

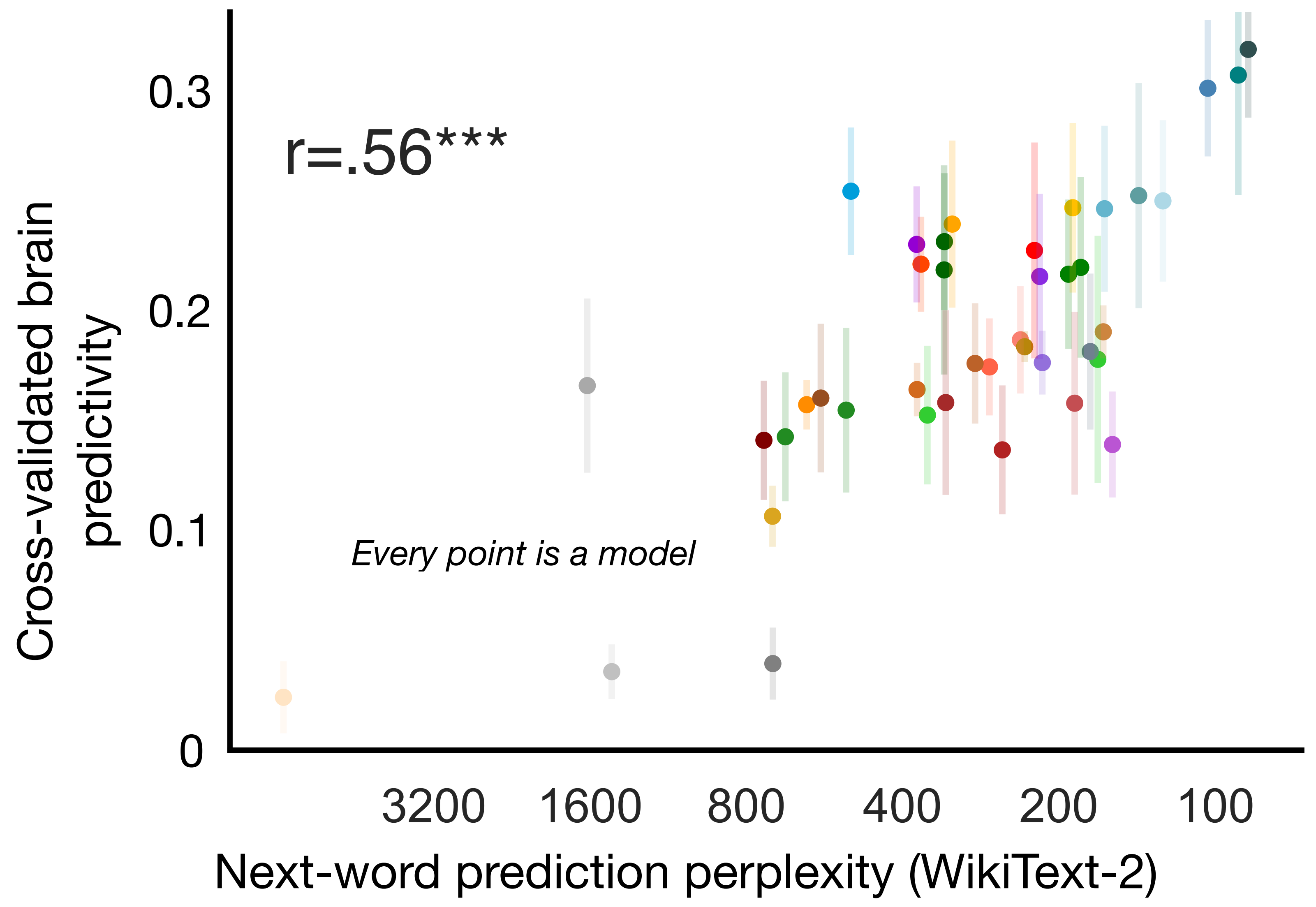
- Performance on the main training objective (typically next-word prediction)
- Performance on other tasks (e.g., sentiment analysis, summarization)
- Performance on experimentally altered inguistic input

C. Model training

- Training objective
  - Main training objective
    - Causal/masked word prediction
    - Other (less common) objectives
  - Fine-tuning training objective
- Training data
  - Data amount (number of tokens)
  - Data type (e.g., different modalities, content types, languages)

- Next-word prediction performance is positively correlated with brain alignment, but representational generality is too.
  - No other objective besides next-word prediction has led to similarly strong features.

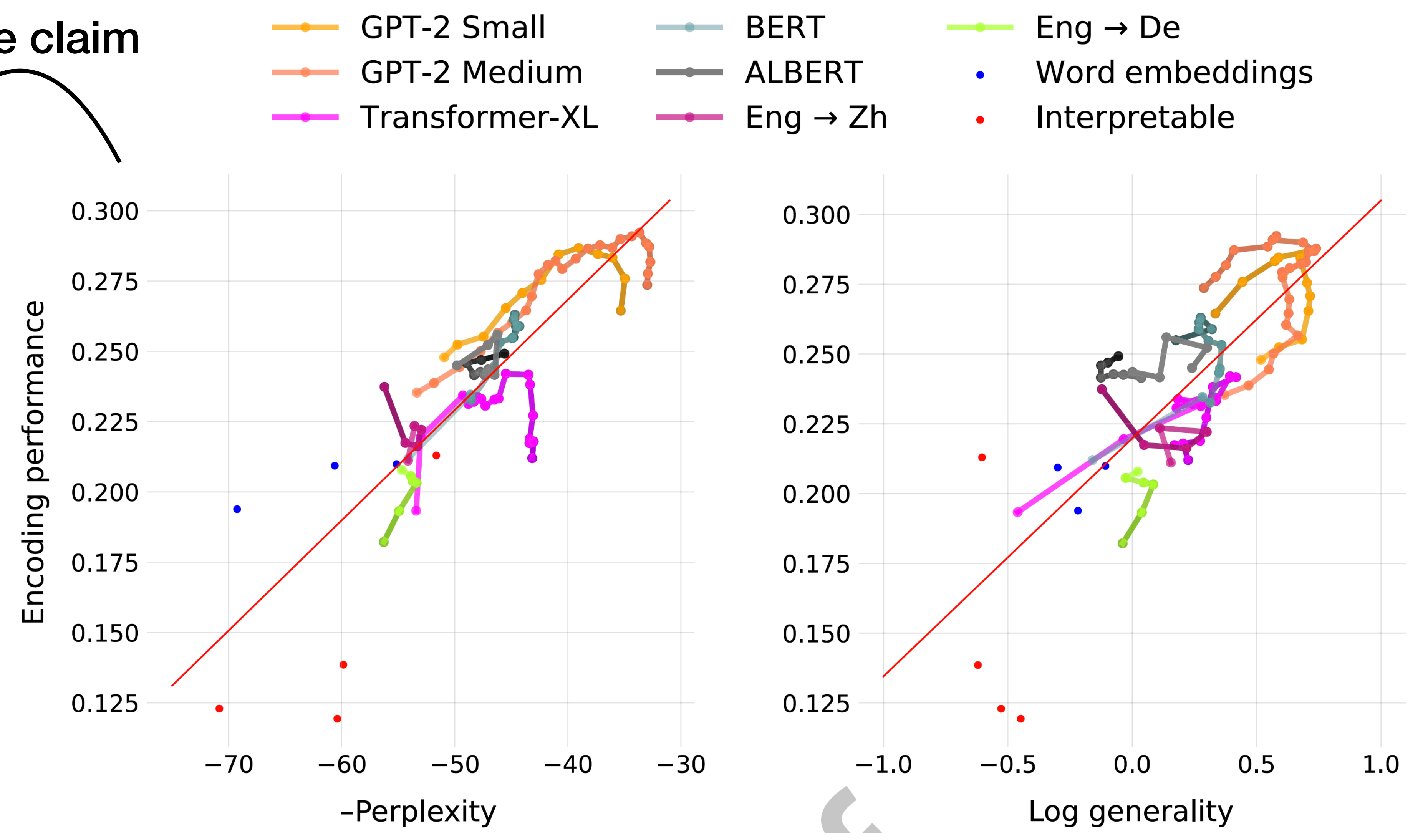
Next-word prediction performance



Schrimpf et al., 2021

"Representational generality"

Same claim



Antonello & Huth, 2022

Distilling LLM properties

A. Model architecture

- General architecture class (e.g., Transformer, RNN)
- Number of layers (e.g., Transformer blocks)
- Number of parameters (number of learnable parameters in the model)
- Embedding dimensionality (the vector size for each token)

B. Model behavior

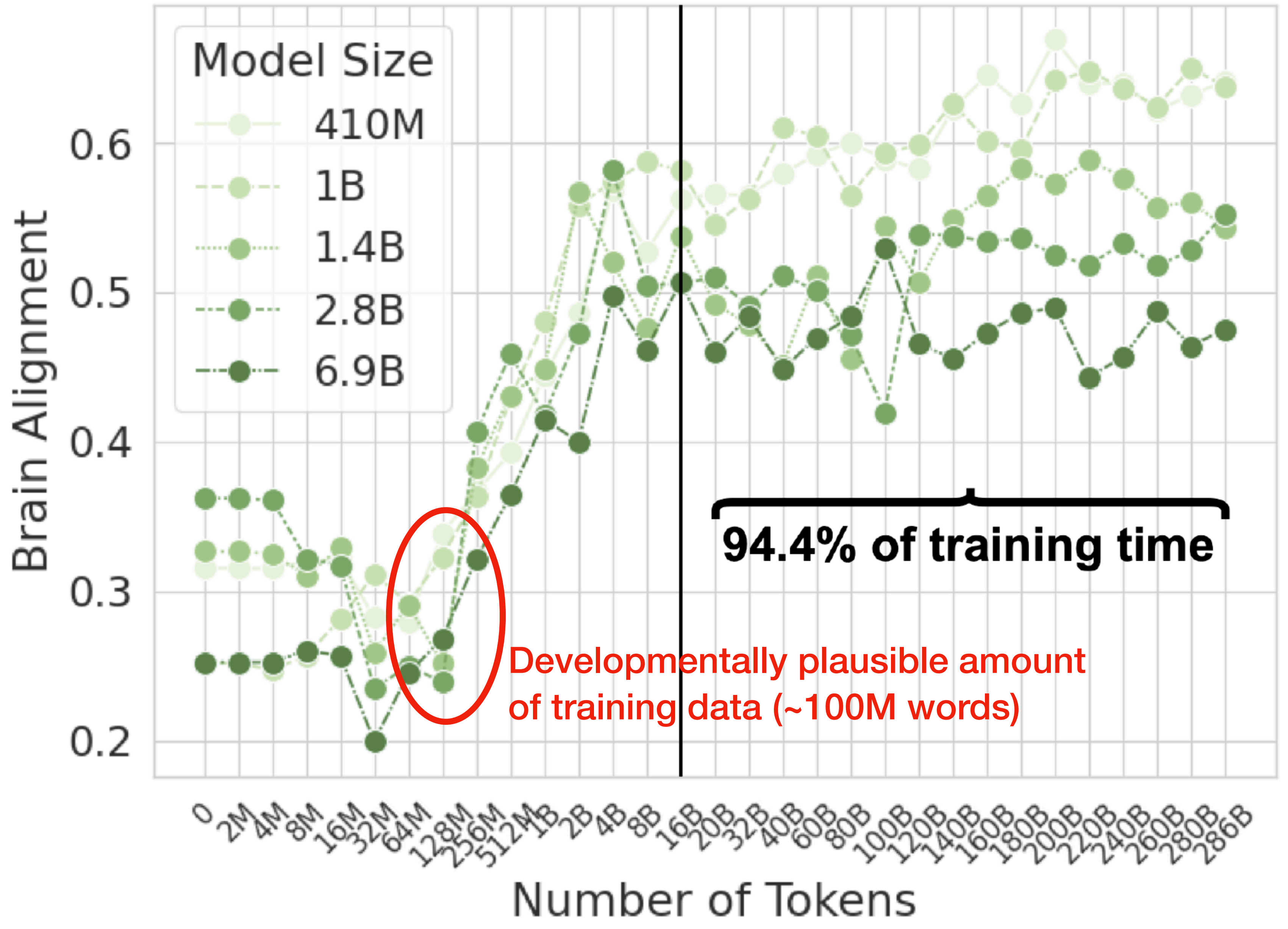
- Performance on the main training objective (typically next-word prediction)
- Performance on other tasks (e.g., sentiment analysis, summarization)
- Performance on experimentally altered inguistic input

C. Model training

- Training objective
  - Main training objective
    - Causal/masked word prediction
    - Other (less common) objectives
  - Fine-tuning training objective
- Training data
  - Data amount (number of tokens)
  - Data type (e.g., different modalities, content types, languages)

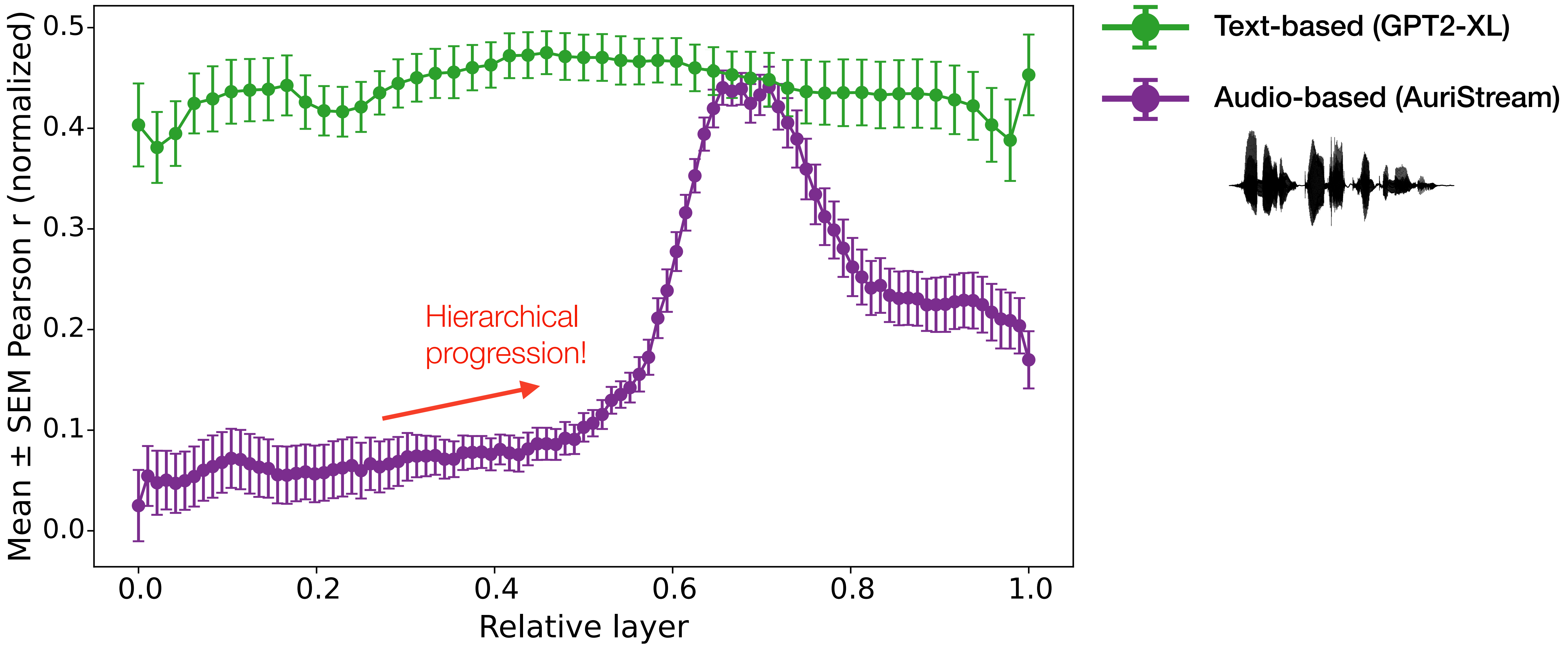
- LLMs do not need excessive amounts of data to align with brain responses.
  - But text tokens are the 'wrong' starting point…

Training data amount



AlKhamissi, Tuckute, .., et al., 2024
(also see Hosseini et al., 2024)

Training data input



Tuckute, Kotar et al., 2025 (+ in prep)

So what?

- **Representational content of the language network:**
  Distilling the properties of LLMs that enable them to capture human responses to language.

- **Engineering applications & model-guided experiments:**
  Leveraging LLMs for decoding, simulation, and neural control.

So what?

- **Representational content of the language network:**
  Distilling the properties of LLMs that enable them to capture human responses to language.

- **Engineering applications & model-guided experiments:**
  Leveraging LLMs for decoding, simulation, and neural control.

Article | Published: 01 May 2023

Semantic reconstruction of continuous language from non-invasive brain recordings

Jerry Tang, Amanda LeBel, Shailee Jain & Alexander G. Huth

Nature Neuroscience **26**, 858–866 (2023) | Cite this article

Actual stimulus

i got up from the air mattress and pressed my face against the glass of the bedroom window expecting to see eyes staring back at me but instead finding only darkness

Decoded stimulus

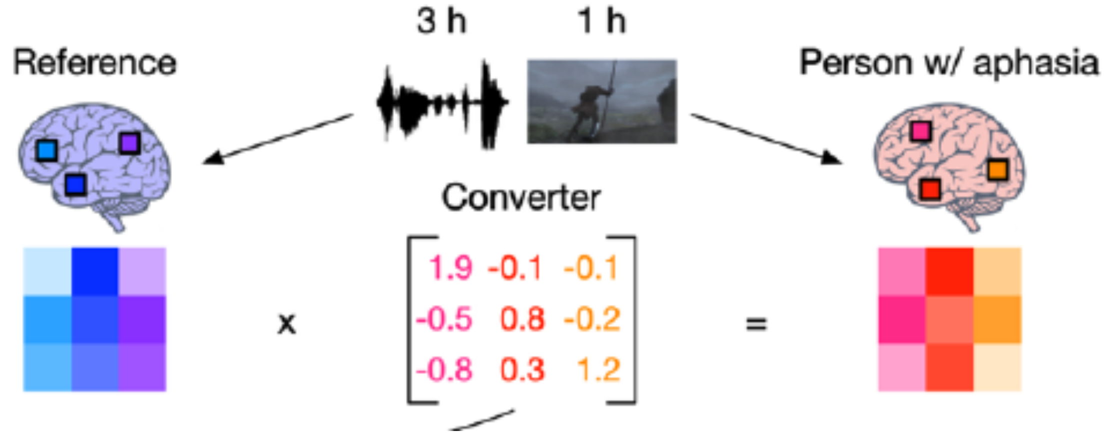
i just continued to walk up to the window and open the glass i stood on my toes and peered out i didn't see anything and looked up again i saw nothing

New Results

Follow this preprint

Decoding concept representations in aphasia after stroke

Jerry Tang, Carly Millanski, Allison Chen, Lisa D. Wauters, Jordyn Anders, Shilpa Shamapant, Stephen M. Wilson, Alexander G. Huth, Maya L. Henry



So what?

- **Representational content of the language network:**
  Distilling the properties of LLMs that enable them to capture human responses to language.

- **Engineering applications & model-guided experiments:**
  Leveraging LLMs for decoding, simulation, and neural control.

SPECIAL ISSUE:
Cognitive Computational Neuroscience of Language

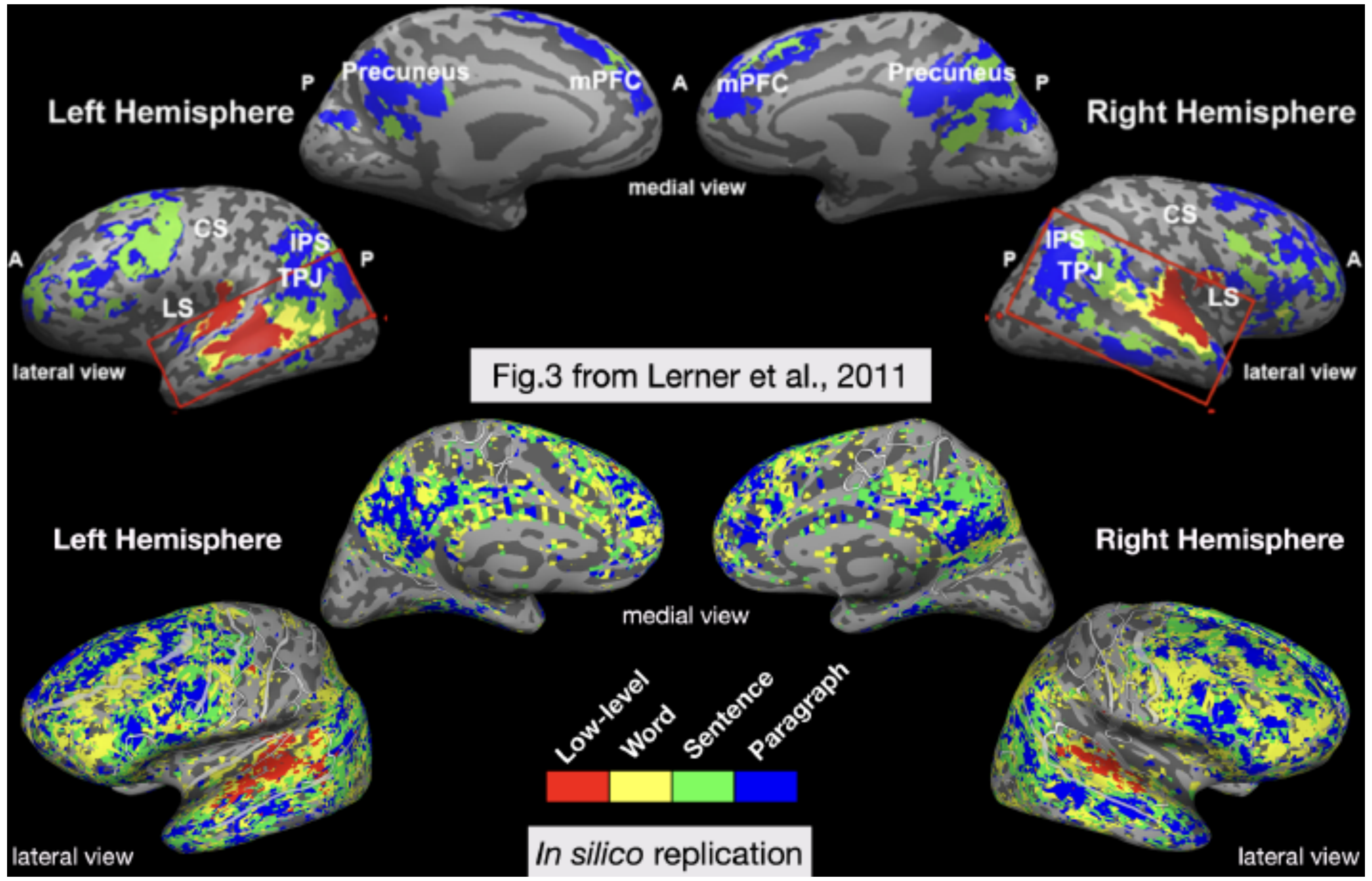
Computational Language Modeling and the Promise of in Silico Experimentation

Shailee Jain[1], Vy A. Vo[3], Leila Wehbe[4,5], and Alexander G. Huth[1,2]

Poster **B183** in Poster Session B: *Wednesday, August 13, 1:00 – 4:00 pm, de Brug & E-Hall*

Catalyzing in silico neuroscience with a toolkit of accurate encoding models of the brain

Alessandro Thomas Gifford[1], Domenic Bersch[2], Daniel Janini[1], Gemma Roig[2], Radoslaw Martin Cichy[1]; [1]Freie Universität Berlin, [2]Johann Wolfgang Goethe Universität Frankfurt am Main



So what?

- **Representational content of the language network:**
  Distilling the properties of LLMs that enable them to capture human responses to language.

- **Engineering applications & model-guided experiments:**
  Leveraging LLMs for decoding, simulation, and neural control.

Neural control: using models to make quantitative predictions about a neural target and subsequently using those predictions to successfully modulate neural activity in the target in a closed-loop manner.

nature human behaviour

Driving and suppressing the human language network using large language models

Received: 6 May 2023

Accepted: 10 November 2023

Published online: 03 January 2024

Greta Tuckute[1,2], Aalok Sathe[1,2], Shashank Srikant[3,4], Maya Taliaferro[1,2], Mingye Wang[1,2], Martin Schrimpf[2,5,6], Kendrick Kay[7] & Evelina Fedorenko[1,2,8]

Thanks for listening and thanks to all my brilliant collaborators!

Questions / thoughts

gtuckute@fas.harvard.edu