Brain-Inspired AI 2.0: Aligning Language Models Across Languages and Modalities

Subba Reddy Oota[1], Tanmoy Chakraborty[2], Manish Gupta[3,4], Raju S. Bapi[3]

[1]TU Berlin, Germany; [2]IIT Delhi, India; [3]IIIT Hyderabad, India; [4]Microsoft, India

subba.reddy.oota@tu-berlin.de, tanchak@iitd.ac.in, gmanish@microsoft.com, raju.bapi@iiit.ac.in



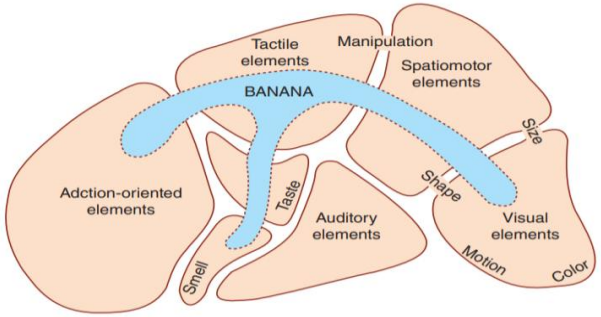

AAAI 2026: Brain-Inspired AI 2.0

Agenda

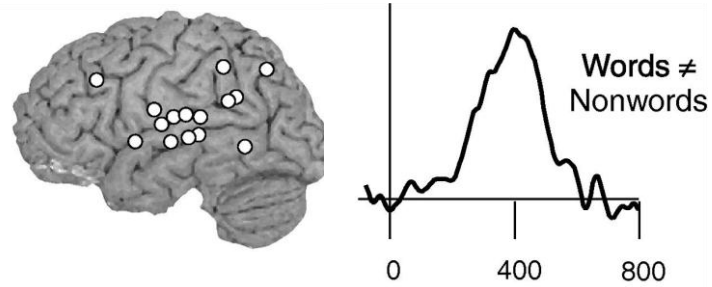
- **Introduction to the tutorial [10 min]**
- Introduction to Brain Encoding and Decoding [50 min]
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

Mechanistic understanding of language processing in the brain: four big questions

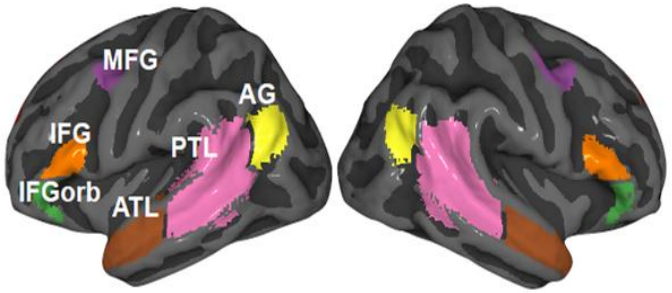
What



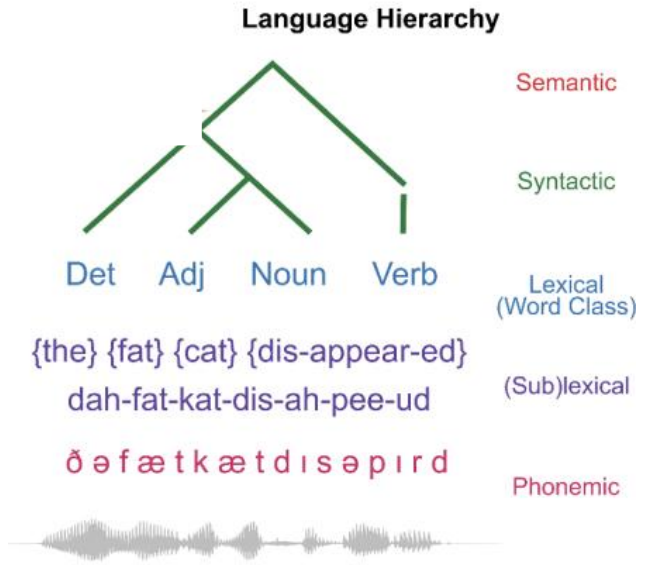
When



Where



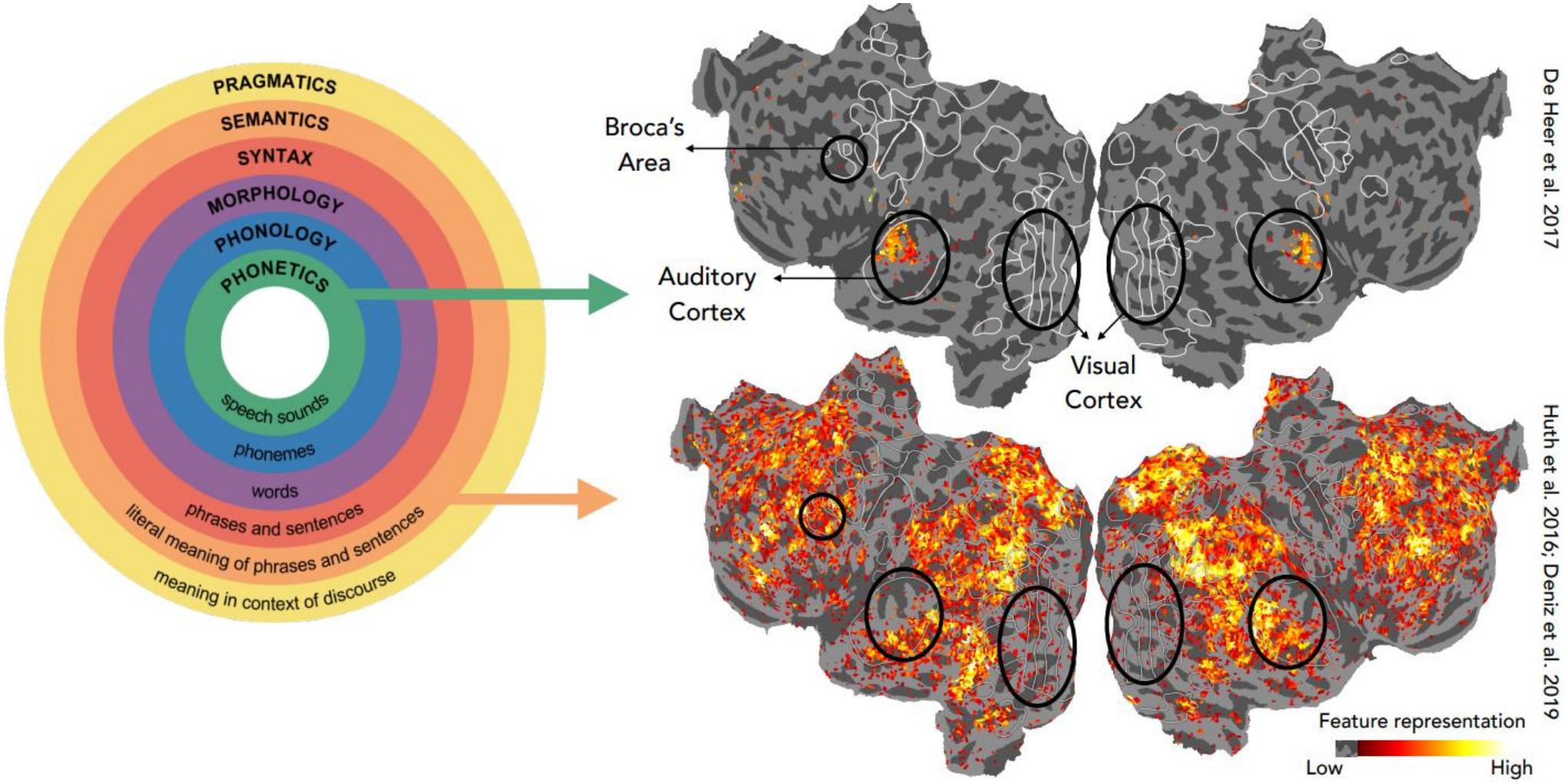
How



[Fedorenko et al. 2010, 2014]

Gwilliams et al. 2024

Natural language is composed of many different features

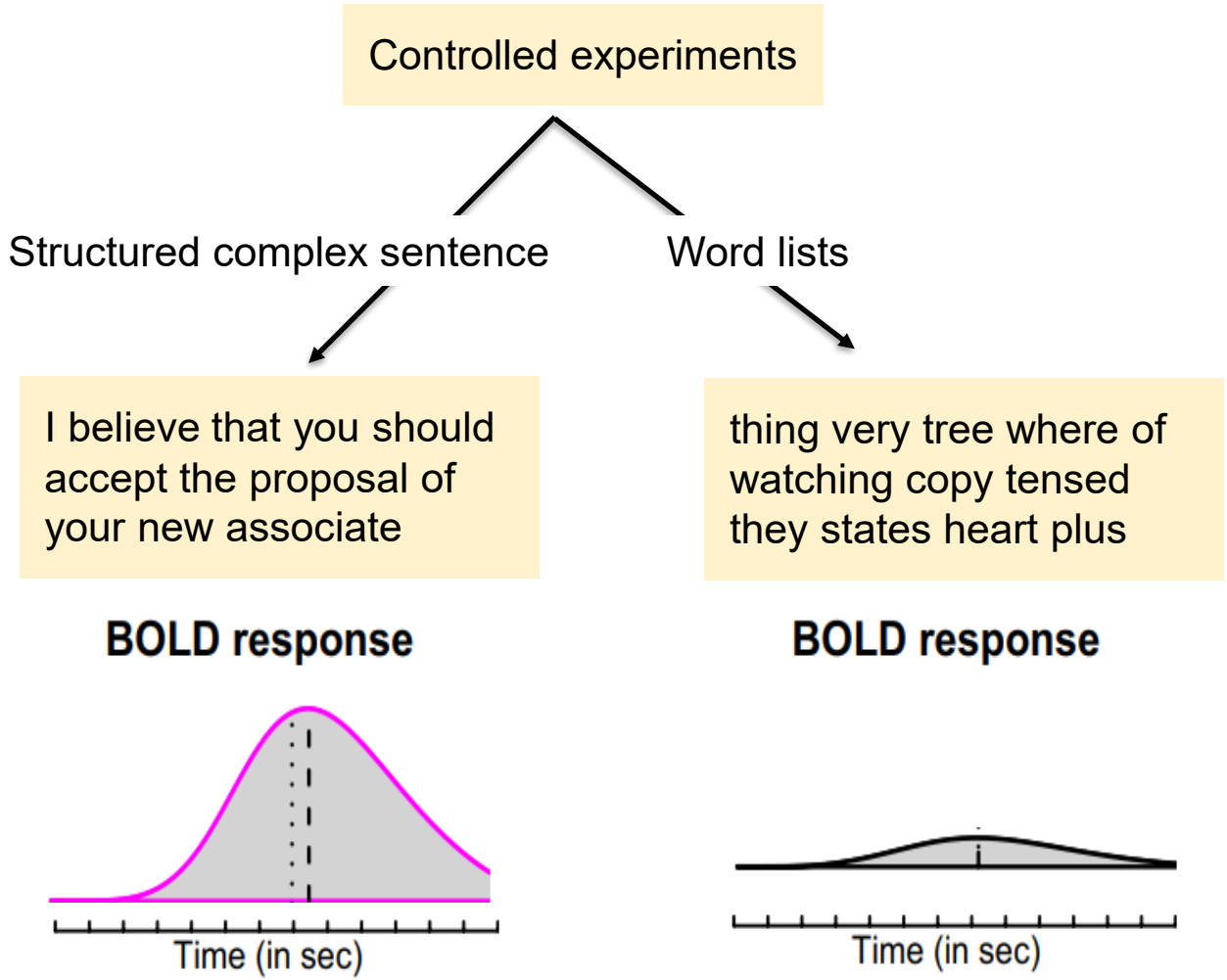


Source: Slide from Fatma Deniz's talk at NEAT-24 workshop

What features of the language stimulus drive the response in each brain area?

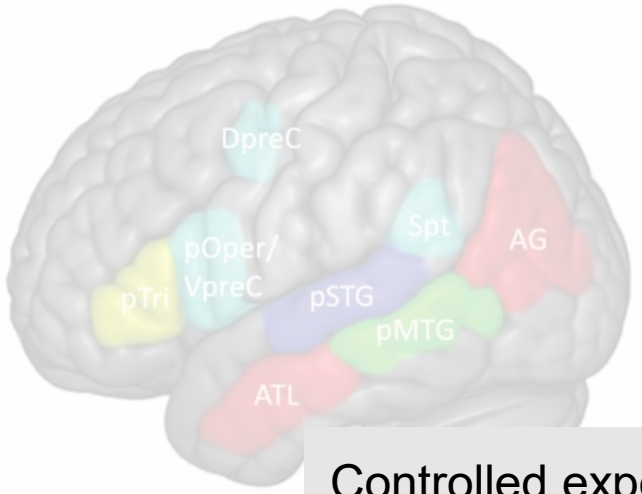
Typical studies of language processing with controlled experiments

- How the human brain computes and encodes syntactic structures?
  - **Syntax:** how do words structurally combine to form sentences and meaning?



Pallier et al. 2011

Language organization in the brain

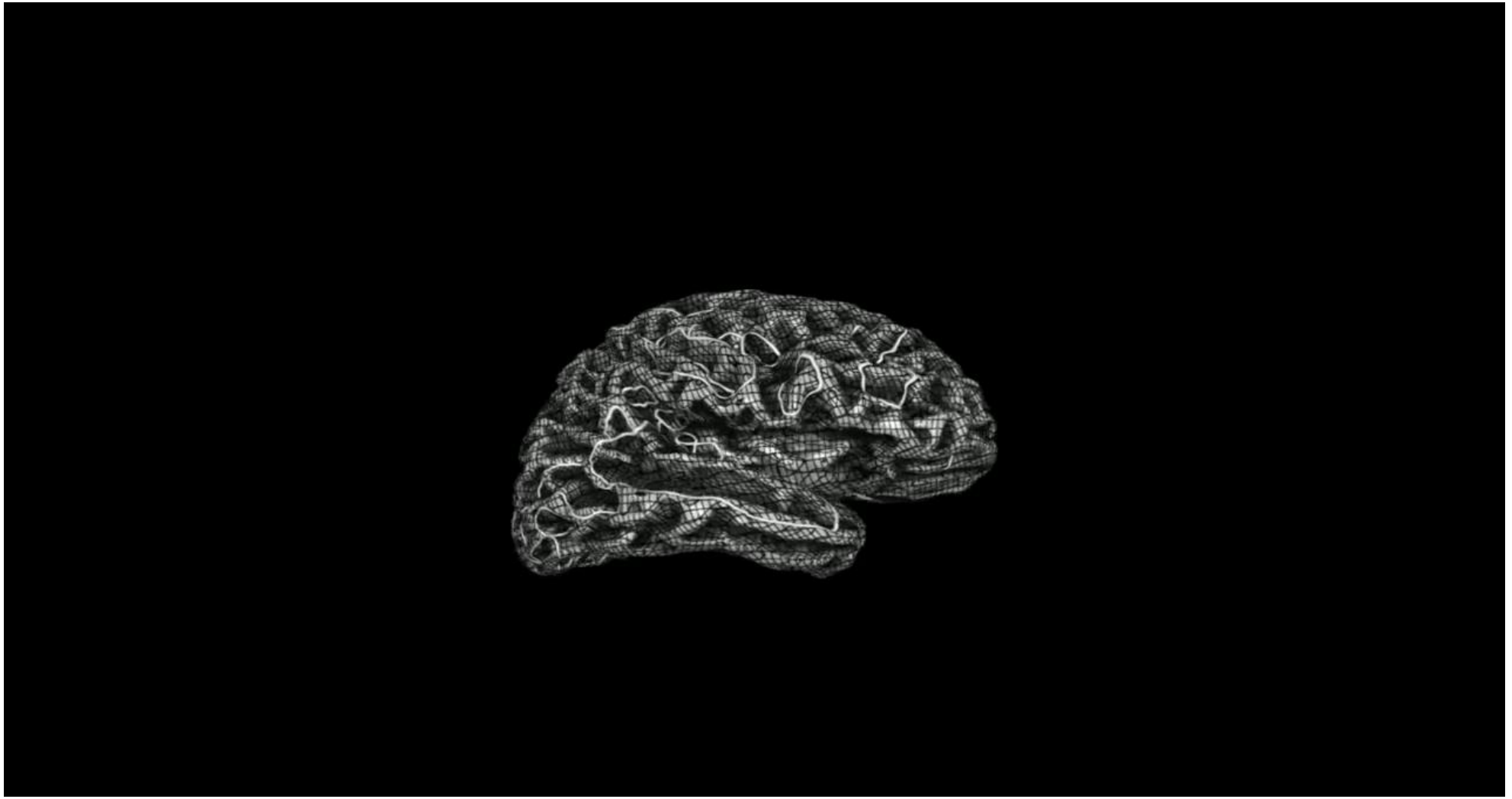


Controlled experiments are task-based and not ecological

ge at different features
Hierarchical syntactic information occurs in the cortical zone situated between auditory-phonological and semantic zones.

Matchin & Hickok et al. 2022

Designing a functional MRI experiment: watching movies



Source: Video from Gallant Lab

Increasingly available open source ecological stimuli datasets

With advancement of **ecological stimuli datasets** and **open source language models**, recent studies looked at interesting open questions?

Dataset	Modality	Subj	1-TR	# TRs
Full-Moth-Radio-Hour	Listening	8	2.0045s	9932
Subset-Moth-Radio-Hour	Reading	6	2.0045s	4028
Subset-Moth-Radio-Hour	Listening	6	2.0045s	4028
Narratives (21^{st}-Year)	Listening	18	1.5s	2250
Harry-Potter	Reading	8	2s	1211

Is the **"how" of the NLP system process language comprehension** the same as **"how" of the brain process language comprehension**?

How is information aggregated by the brain during language comprehension?

Deniz et al. 2019

Lebel et al. 2022

Nastase et al. 2021

Li et al. 2022

Zhang et al. 2021

Neuro-AI Pipeline

DNN Model Representations

Stimulus

Language | Audio | Visual | Multimodal

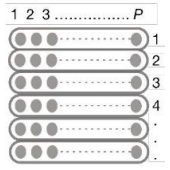
DNN Models

Language Model | Audio Model | Vision Model | Multimodal Model

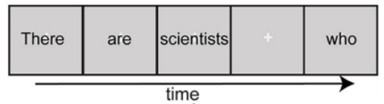
Stimulus Feature Space

Down sampling: Interpolation of the feature matrix

HRF estimation: use of FIR model, different delays



Human Brain Recordings



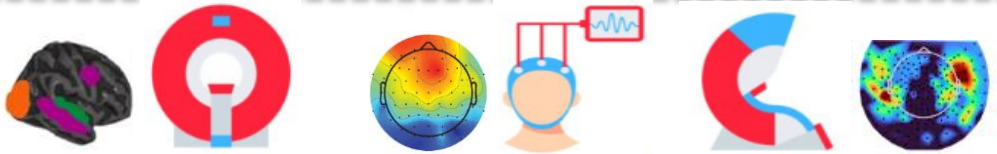
Reading

Listening

Static Image

Video clip (with or without audio)

Visual | Auditory



fMRI | EEG | MEG

Evaluation Metrics

PCC, R^2, 2V2 Accuracy, RDM, CKA, Noise Ceiling, Normalized brain alignment

fMRI: Whole brain, ROI level, Sub-ROI level, task-specific voxels

MEG: Sensor recordings over time points

EEG: Electrode signals recorded over time

Linear

Encoding Models →

Ridge | Bootstrap Ridge | Banded Ridge

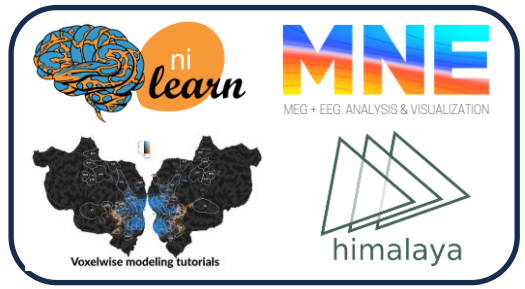
Lasso | PLS | Kernel Ridge

← **Decoding Models**

Non-Linear

Multi-Layer Perceptron | DNN Models

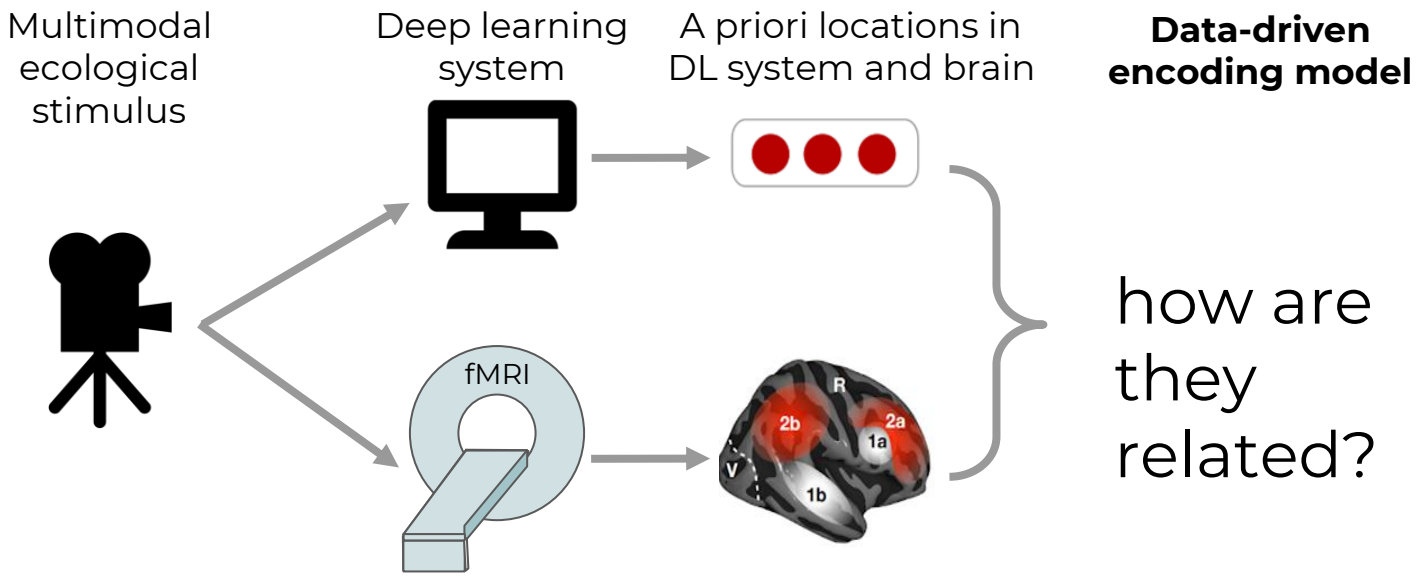
Visualization Tools



Oota et al (2025). Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding (Survey) [TMLR]

https://openreview.net/pdf?id=YxKJihRcby

Deep neural networks and brain alignment: brain encoding and decoding

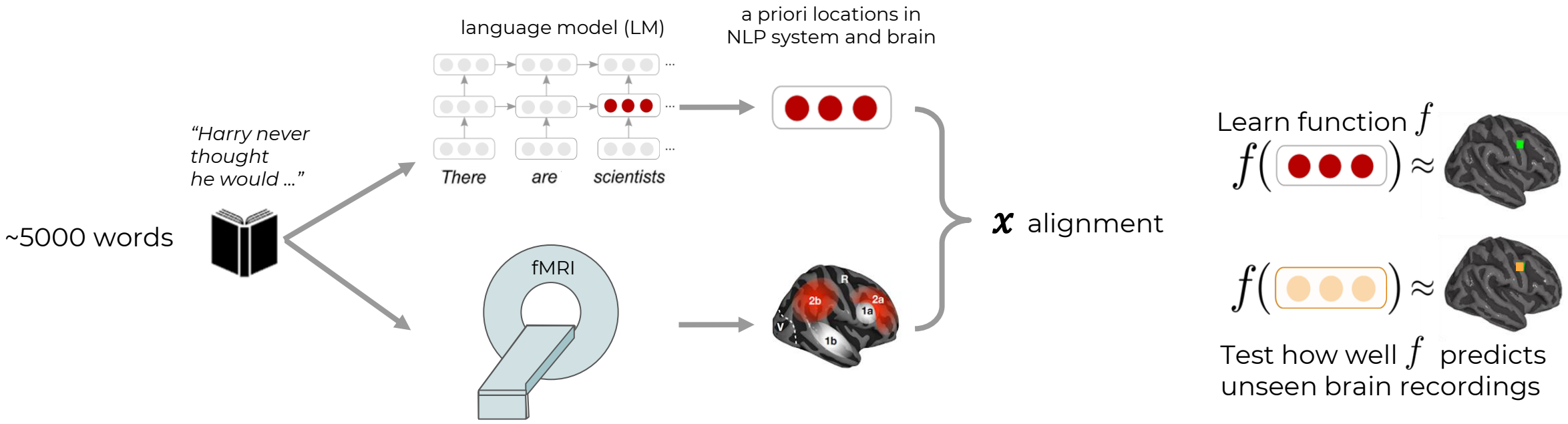


Wehbe et al. 2014,
Jain and Huth 2018,
Gauthier and Levy 2019

Toneva and Wehbe 2019,
Caucheteux et al. 2020,
Toneva et al. 2020

Jain et al. 2020,
Schrimpf et al. 2021,
Goldstein et al. 2022
...

General encoding pipeline to evaluate brain-LM alignment



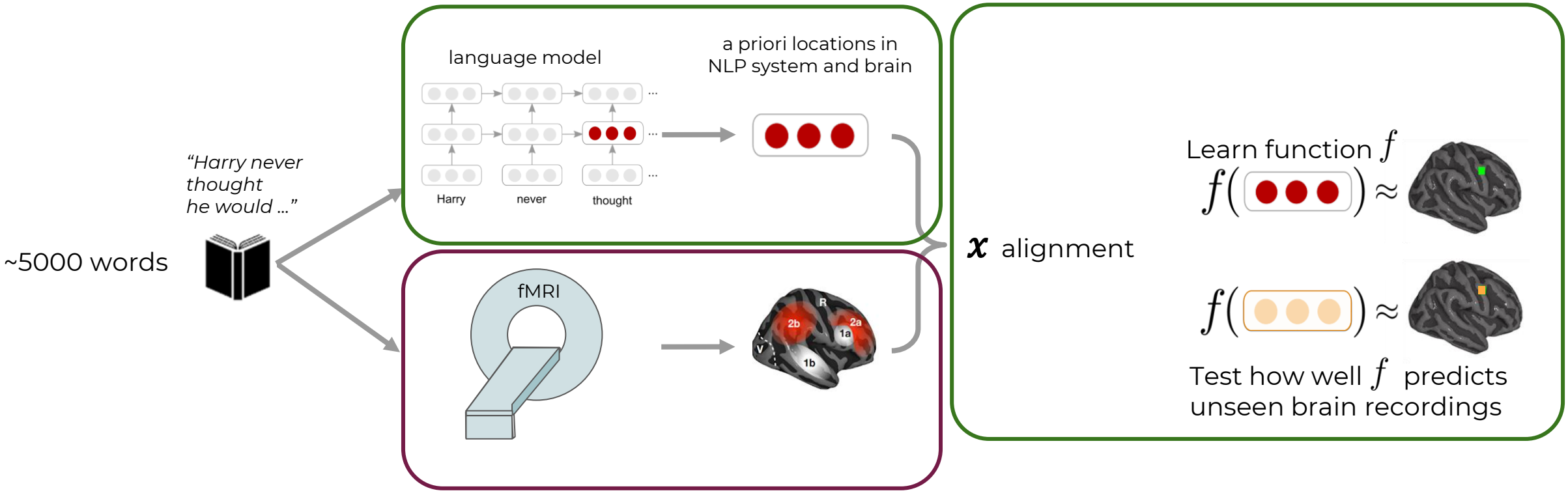
Brain alignment of a LM ⇒ how similar its representations are to a human brain's

Wehbe et al. 2014,
Jain and Huth 2018,
Gauthier and Levy 2019

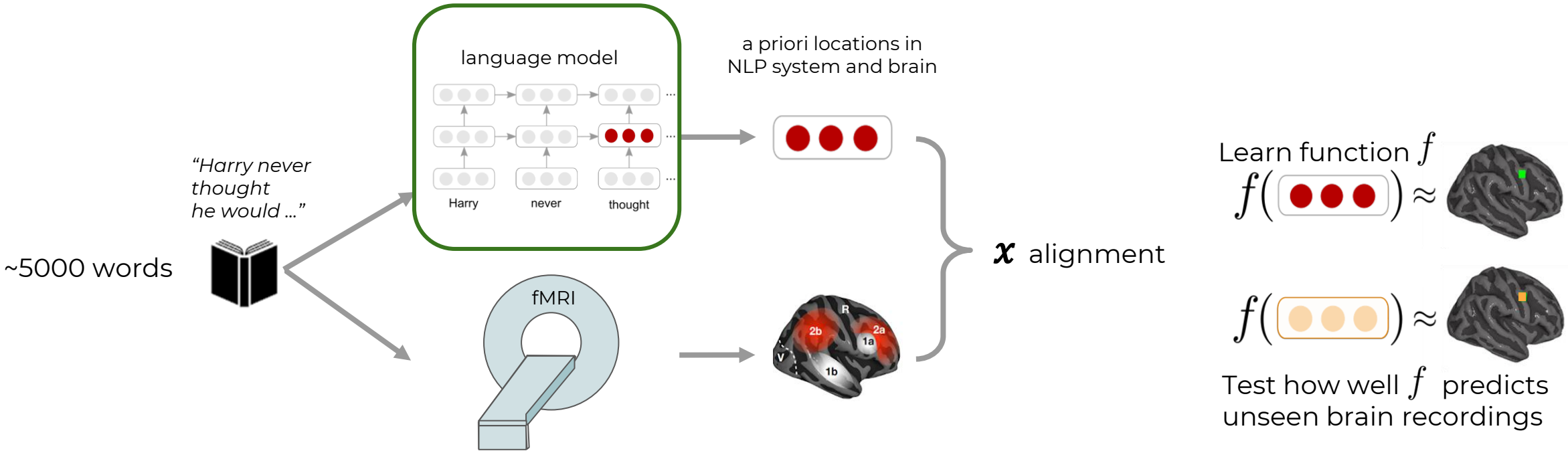
Toneva and Wehbe 2019,
Caucheteux et al. 2020,
Toneva et al. 2020

Jain et al. 2020,
Schrimpf et al. 2021,
Goldstein et al. 2022
...

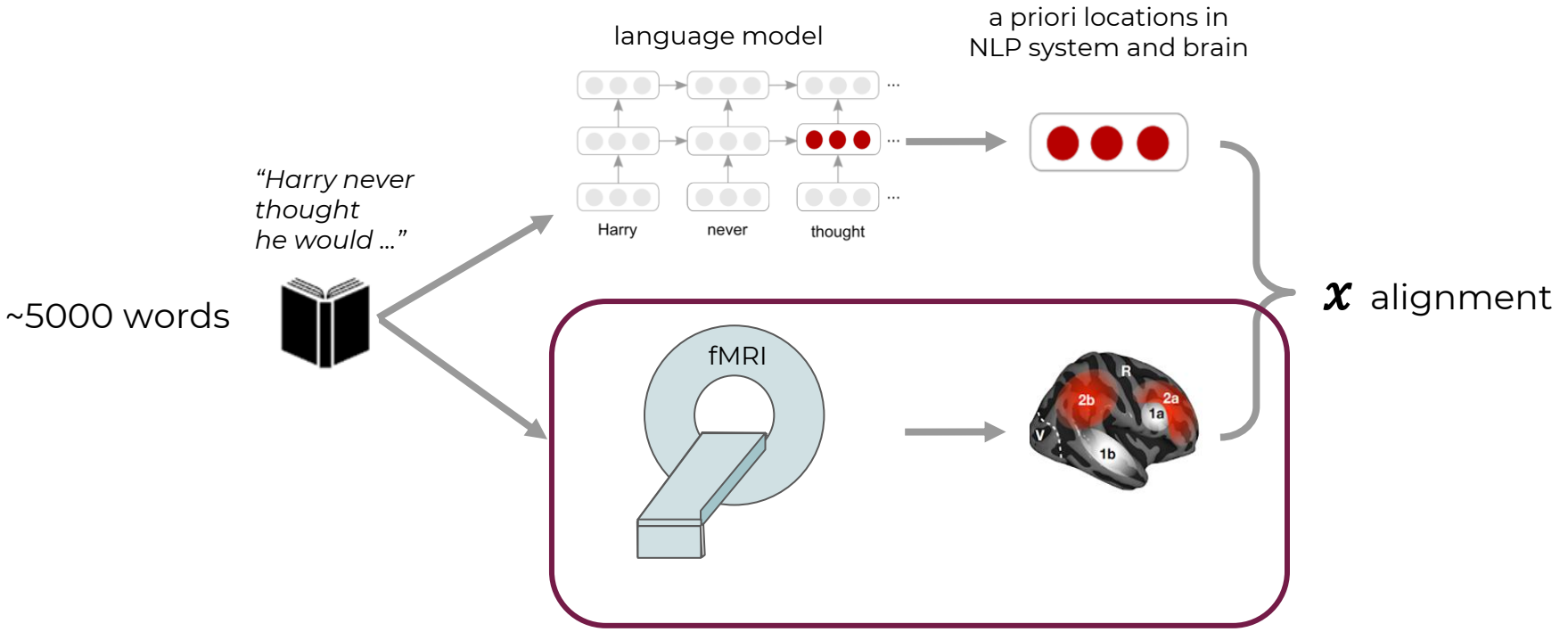
LLMs, estimating alignment, evaluation



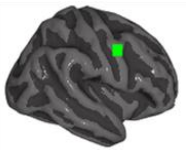
Part 1: LLMs + extracting representations

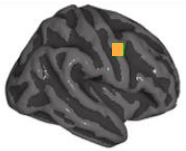


LLMs, estimating alignment, evaluation



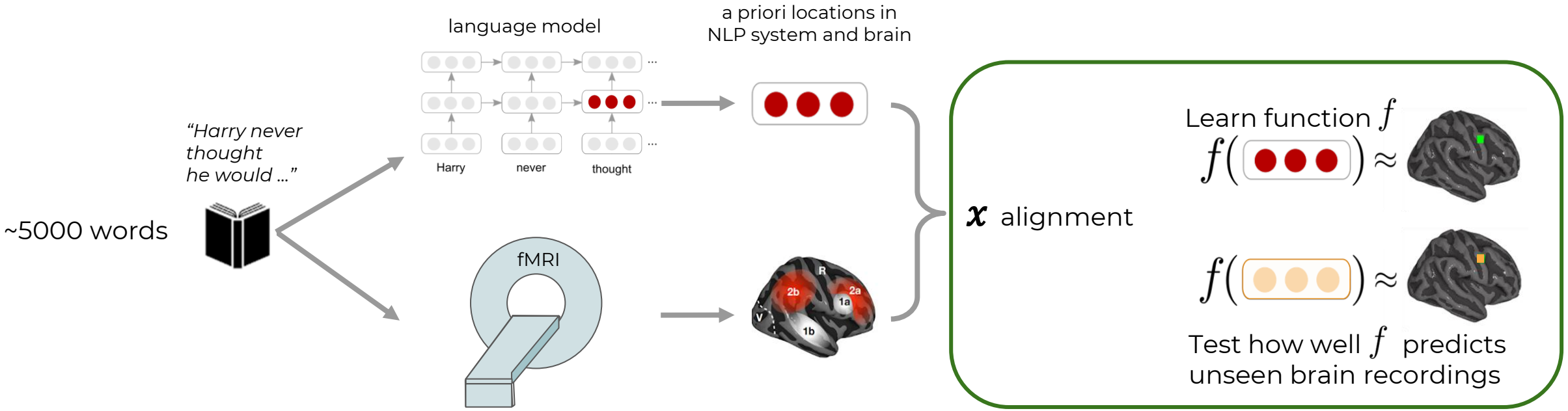
Learn function f

$f(\bullet\bullet\bullet) \approx$ 

$f(\bullet\bullet\bullet) \approx$ 

Test how well f predicts unseen brain recordings

Estimating brain-LM alignment + evaluation



Agenda

- Introduction to the tutorial [10 min]
- **Introduction to Brain Encoding and Decoding [50 min]**
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

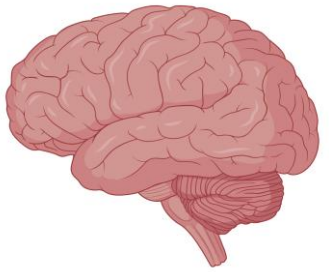
Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
  - Brain Encoding/Decoding and applications
  - Introduction to popular datasets
  - Text Stimulus Representations
  - Alignment Between AI Models and Human Brain Language Comprehension
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

Agenda

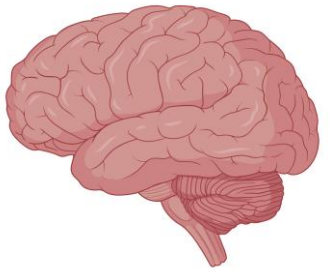
- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
  - **Brain Encoding/Decoding and applications**
  - Introduction to popular datasets
  - Text Stimulus Representations
  - Alignment Between AI Models and Human Brain Language Comprehension
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

Neuroscience



- Field of science that studies the structure and function of the nervous system of different species.
- How is information represented and processed in the brain?
  - How does the brain integrate text, vision, sound, and touch?
  - How do abstract concepts (like numbers or language) get encoded?
- How does the brain learn and change?
  - What are the neural mechanisms that allow learning, memory, and adaptation?
  - How are memories formed but not overwritten?
  - How does the brain learn from very few examples?
  - How does RL work biologically?
- How do large-scale brain circuits coordinate behavior?
  - How do distributed brain regions work together to produce coherent behavior?
  - How does the brain decide when to act or inhibit action?
  - How do sensory inputs get transformed into motor outputs?

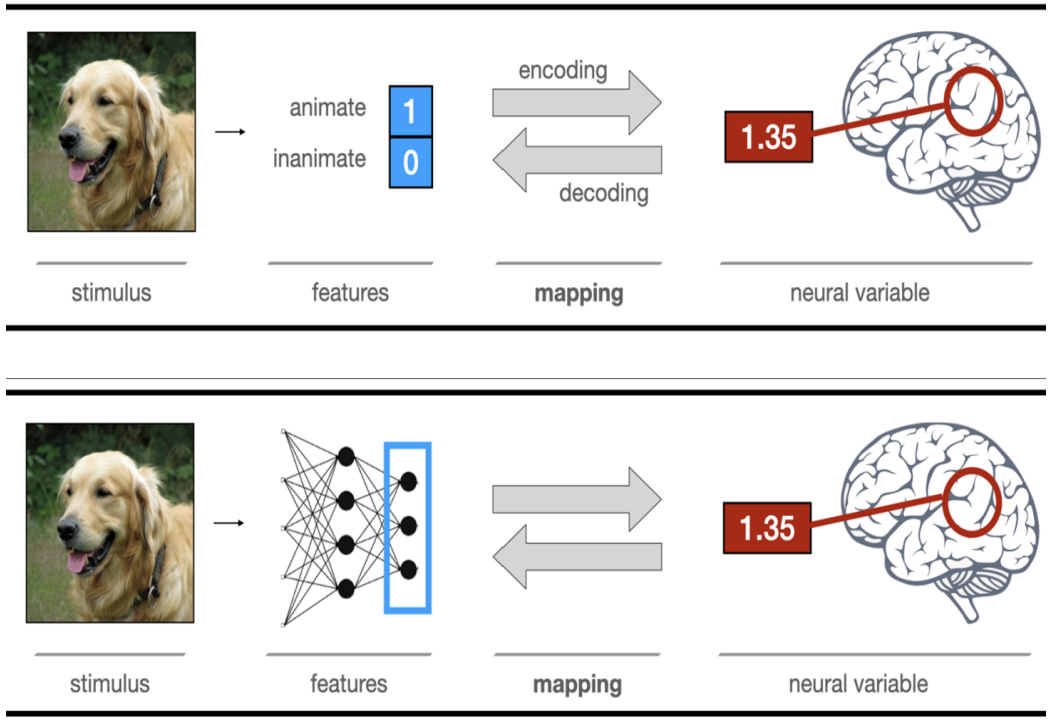
Neuroscience



- What goes wrong in neurological and psychiatric disorders, and how can we fix it?
  - How do disruptions in brain structure or function cause diseases (Alzheimer's, depression, Parkinson's, autism, and schizophrenia), and how can we treat them?
  - Precision psychiatry: Can we predict which treatment will work for whom, when, and why?
- How does the brain give rise to the mind and consciousness?
  - How do electrical and chemical signals in neurons produce subjective experiences like thoughts, emotions, awareness, and the sense of self?
  - Is consciousness localized or distributed?
  - Can machines ever be conscious?
  - What changes in the brain when consciousness is lost?
- Brain-computer interfaces

Brain encoding and decoding in cognitive neuroscience

- Encoding is the process of learning the mapping e from the stimuli S to the neural activation F.
  - Using feature engg or deep learning.
- Decoding constitutes learning mapping d, which predicts stimuli S back from the brain activation F.
  - Oftentimes, we predict a stimulus representation R rather than actually reconstructing S.

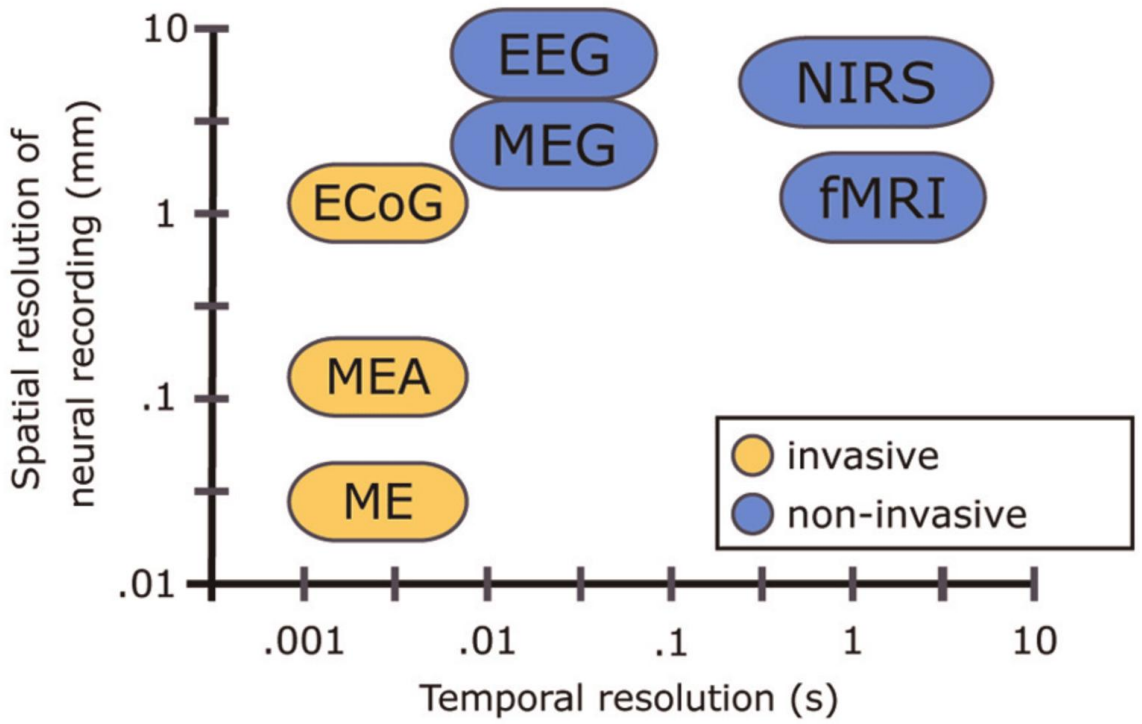


Ivanova, Anna A., Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. "Is it that simple? Linear mapping models in cognitive neuroscience." *bioRxiv* (2021).

Brain encoding and decoding

- For both encoding and decoding, the first step is to learn a stimulus representation R of the stimuli S at the train time.
- F is the brain response.
- Next
  - For encoding, a regression function $e: R \rightarrow F$ is trained.
  - For decoding, a function $d: F \rightarrow R$ is trained.
- These functions e and d can then be used at test time to process new stimuli and brain activations, respectively.

Techniques for studying the brain function



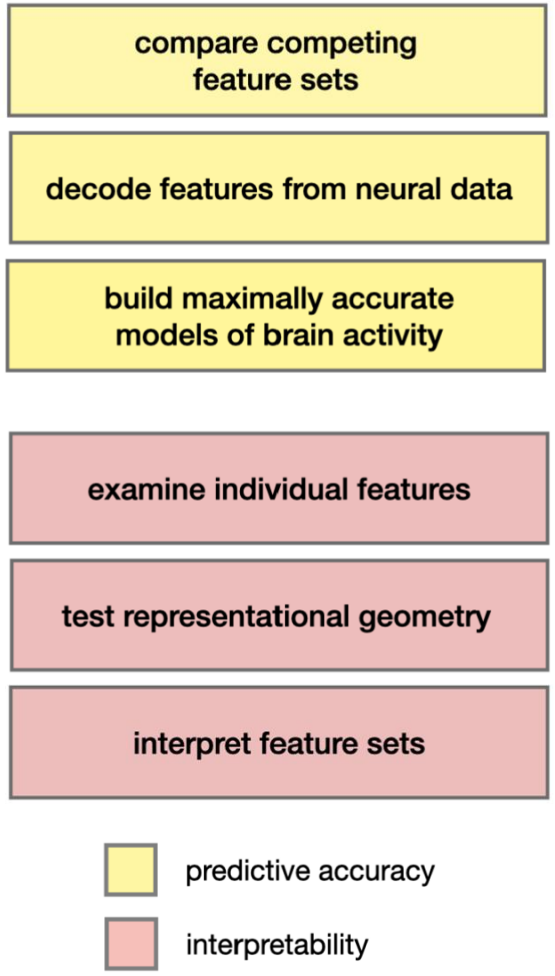
Single Micro-Electrode (ME), Micro-Electrode array (MEA), Electro-Cortico Graphy (ECoG), Positron emission tomography (PET), functional MRI (fMRI), Magneto-encephalography (MEG), Electro-encephalography (EEG), Near-Infrared Spectroscopy (NIRS)

- fMRI: high spatial but low time resolution.
  - Good to study a specific location in the brain
  - Unsuitable for sentence-level analysis. fMRI takes about two seconds to complete a scan. This is far lower than the speed at which humans can process language.
  - Cannot capture syntactic information (Gauthier and Levy, 2019)
- EEG: high time but low spatial resolution.
  - Can preserve rich syntactic information (Hale et al., 2018)
  - But cannot use for source analysis.
- fNIRS: compromise option
  - Time resolution better than fMRI
  - Spatial resolution better than EEG
  - Balance of spatial and temporal resolution may not be enough to compensate for the loss in both.

Vogel, Jörn, Sami Haddadin, Beata Jarosiewicz, John D. Simeral, Daniel Bacher, Leigh R. Hochberg, John P. Donoghue, and Patrick van der Smagt. "An assistive decision-and-control architecture for force-sensitive hand–arm systems driven by human–machine interfaces." *The International Journal of Robotics Research* 34, no. 6 (2015): 763-780.

Computational Cognitive Science Research goals

- Predictive Accuracy
  - Compare feature sets: Which feature set provides the most faithful reflection of the neural representational space?
  - Test feature decodability: “Does neural data Y contain information about features X?”
  - Build accurate models of brain data: Aim is to enable simulations of neuroscience experiments.
- Interpretability
  - Examine individual features: Which features contribute the most to neural activity?
  - Test correspondences between representational spaces
    - “CNNs vs ventral visual stream” or “Two text representations”
  - Interpret feature sets
    - Do features X, generated by a known process, accurately describe the space of neural responses Y?
    - Do voxels respond to a single feature or exhibit mixed selectivity?
  - How does the mapping relate to other models or theories of brain function?
- …



Ivanova, Anna A., Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. "Is it that simple? Linear mapping models in cognitive neuroscience." *bioRxiv* (2021).

Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
  - Brain Encoding/Decoding and applications
  - **Introduction to popular datasets**
  - Text Stimulus Representations
  - Alignment Between AI Models and Human Brain Language Comprehension
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

Types of stimuli and popular datasets

- Text (Words, Sentences, Paragraphs): Harry Potter Story, ZUCO EEG, Question-Answering MEG.
- Visual: Binary visual patterns, Natural Images (Vim-1), BOLD5000, Algonauts and SS-fMRI.
- Audio: Alice's Adventures in Wonderland, Narratives, The Moth Radio Hour, Audio stories.
- Videos: BBC's Doctor Who, Japanese Ads, Pippi Langkous, Algonauts.
- Other Multimodal Stimuli: Words + line drawing of concept named by each word, Pereira.

Forms of stimulus presentation and data collection

- Type: fMRI, EEG, MEG, …
- TR: Sampling time.
- Fixation points: location, color, shape.
- Form of stimuli presentation: text, video, audio, images.
- Task: question answering, property generation, understanding, …
- Time given to participants: 1 minute to list properties, …
- Type of participants: males/females, sighted/blind, …
- Number of times the response to stimuli was recorded.
- Stimulus Language

Text Stimulus Datasets

Dataset	Type	Language	Stimulus	#Subjects	Paradigm	Size	Task
Wehbe et al., 2014	fMRI	English	Chapter 9 of Harry Potter and the Sorcerer's Stone	9	Reading stories	5000 word chapter was presented in 45 minutes.	Story understanding
Handjaras et al., 2016	fMRI	Italian	Verbal, pictorial or auditory presentation of 40 concrete nouns	20	Reading, viewing or listening	40 nouns * 4 times.	Property Generation
Anderson et al., 2017	fMRI	Italian	70 concrete and abstract nouns from law/music.	7	Reading	70 nouns * 5 times.	Imagine a situation that they personally associate with the noun
Zurich Cognitive Language Processing Corpus (ZuCo): Hollenstein et al., 2018	EEG and eye-tracking	English	Sentences from movie reviews or Wikipedia	12	Reading natural sentences	21,629 words in 1107 sentences and 154,173 fixations	Rate movie quality, answer control questions, check for existence of a relation
Anderson et al., 2019	fMRI	English	240 active voice sentences describing everyday situations	14	Reading	240 sentences seen 12 times (by 10 subjects) and 6 times (by 4 subjects)	Passive reading
BCCWJ-EEG: Oseki and Asahara, 2020	EEG	Japanese	20 newspaper articles	40	Reading	1 time reading for ~30-40 minutes	Passive reading
Deniz et al., 2019	fMRI	English	Subset of Moth Radio Hour. 11 stories	9	Reading	11 10- to 15 min stories presented twice word by word	Passive reading and Listening

Visual Stimulus Datasets

Dataset	Type	Stimulus	#S	Paradigm	Size	Task
Thirion et al., 2006	fMRI	Rotating wedges, expanding/contracting rings, rotating Gabor filters, grid	9	Viewing visual patterns	Wedges/rings for 8 times, 36 Gabor filters for 4 times, grid 36 times	Passive viewing, imagine one of the 6 domino stimuli when prompted to.
Vim-1: Kay et al., 2008	fMRI	Sequences of natural photos	2	Viewing natural images	Each subject viewed 1750 (Stage 1)+ 120 (Stage 2) novel natural images	Passive viewing
Horikawa et al., 2017	fMRI	Object images	5	Viewing and Reading	Each subject: (1) Image presentation: 1,200 images from 150 object categories and 50 images from 50 object categories; (2) Imagery: 10 times.	One-back repetition detection task, imagine object images pertaining to the category
BOLD5000: Chang et al., 2019	fMRI	5254 images depicting real-world scenes	4	Viewing natural images	~20 hours of MRI scans per each of four participants	Passive viewing
Algonauts: Cichy et al., 2019	fMRI (EVC and IT)/MEG (early and late in time)	Object images	15	Viewing object images	92 silhouette object images and 118 images of objects on natural background	Passive viewing
Natural Scenes Dataset: Allen et al., 2022	fMRI	73000 natural scenes	8	Viewing natural scenes	~73000 distinct natural scene images from MSCOCO.	Passive viewing
THINGS: Hebart et al., 2023	fMRI/EEG	31188 natural images across 1,854 object concepts.	8	Viewing natural images	fMRI: 3 Participants. 8,740 unique images. 720 objects. MEG: 4 Participants. 22,448 unique images. 1,854 objects	oddball detection task (synthetic image).

Audio Stimulus Datasets

Dataset	Type	Language	Stimulus	#S	Paradigm	Size	Task
Handjaras et al., 2016	fMRI	Italian	Verbal, pictorial or auditory presentation of 40 concrete nouns	20	Reading, viewing or listening	40 nouns * 4 times.	Property Generation
Huth et al., 2016	fMRI	English	Eleven 10-minute stories	7	Listening	2 hours of stories from The Moth Radio Hour	Passive Listening
Brennan and Hale, 2019	EEG	English	Chapter one of Alice's Adventures in Wonderland as read by Kristen McQuillan	33	Listening	2,129 words in 84 sentences. The entire experimental session lasted 1–1.5 h (including QA).	8 MCQ Question answering concerning the contents of the story
Anderson et al., 2020	fMRI	English	One of 20 scenario names	26	Listening scenario name	20 scenario prompts displayed 5 times.	Imagine themselves personally experiencing common scenarios
Narratives: Nastase et al., 2021	fMRI	English	27 diverse naturalistic spoken stories	345	Listening	891 functional scans, totaling ~4.6 hours of unique stimuli (~43,000 words)	Passive Listening
Natural Stories: Zhang et al., 2020	fMRI	English	Moth-Radio-Hour naturalistic spoken stories	19	Listening	5 h 33 m (repeated twice). Each story is 6 m 48 s avg or 2492 words.	Passive Listening
The Little Prince: Li et al., 2021	fMRI	English, Chinese, French	Audiobook	112	Listening	English audiobook is 94 minutes long. Chinese: 99min. French: 97 min.	Passive Listening. 4 quiz questions.
MEG-MASC: Gwilliams et al., 2022	MEG	English	4 English fictional stories: Cable spool boy, LW1, Black willow, Easy money.	27	Listening	Two hours of naturalistic stories. 208 MEG sensors.	Passive Listening

Video Stimulus Datasets

Dataset	Type	Language	Stimulus	#Subjects	Paradigm	Size	Task
BBC's Doctor Who: Seeliger et al., 2019	fMRI	English	Spatiotemporal visual and auditory naturalistic stimuli (30 episodes of BBC's Doctor Who)	1	Viewing episode videos	120.830 whole-brain volumes (approx. 23 h) of single-presentation data, and 1.178 volumes (11 min) of repeated narrative short episodes (22 repetitions)	Passive viewing
Japanese Ads: Nishida et al., 2020	fMRI	Japanese	368 web and 2452 TV Japanese ad movies (15-30s)	40 and 28 for web and TV ads. 16 were overlapped	Viewing Ads	7200 train and 1200 test fMRIs for web; fMRIs from 420 ads.	Passive viewing
Pippi Langkous: Berezutskaya et al., 2020	ECoG	The movie was originally in Swedish but dubbed in Dutch	30 s excerpts of a feature film (in total, 6.5 min long), edited together for a coherent story	37 patients	Viewing	6.5 min movie.	Passive viewing
Algonauts: Cichy et al., 2021	fMRI	English	1000 short video clips	10	Viewing video clips	1000 short video clips (3 sec each)	Passive viewing
Natural Short Clips: Huth et al., 2022	fMRI	English	Natural short movie clips	5	Watching natural short movie clips	3870 responses per subject.	Passive viewing

Other Multimodal Stimulus Datasets

Dataset	Type	Language	Stimulus	#Subjects	Paradigm	Size	Task
Mitchell et al., 2008	fMRI	English	60 different word-picture pairs from 12 categories.	9	Viewing word-picture pairs	60 different word-picture pairs presented six times each	Passive viewing
Sudre et al., 2012	MEG	English	60 concrete nouns along with line drawings	9	Reading	60 stimuli × 20 questions = 1200 examples	Question answering
Zinszer et al., 2017	fNIRS	English	8 concrete nouns (audiovisual word and picture stimuli): bunny, bear, kitty, dog, mouth, foot, hand, and nose	24	Viewing and listening	12 blocks with the 8 stimuli per subject.	Passive viewing and listening
Pereira et al., 2018	fMRI	English	180 Words with Picture, Sentences, word clouds; 96 text passages; 72 passages	16	Viewing WP, sentences or word clouds	180 WP, S and WC per subject; 96+72 passages shown 3 times	Passive viewing
Cao et al., 2021	fNIRS	Chinese	50 concrete nouns from 10 semantic categories	7	Viewing and listening	Each stimulus is presented 7 times.	Passive viewing and listening
Courtois Neuromod	fMRI	full-length movies and TV show	6	Viewing and Listening	~100 hours of data per participant	Passive viewing	

Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
  - Brain Encoding/Decoding and applications
  - Introduction to popular datasets
  - **Text Stimulus Representations**
  - Alignment Between AI Models and Human Brain Language Comprehension
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

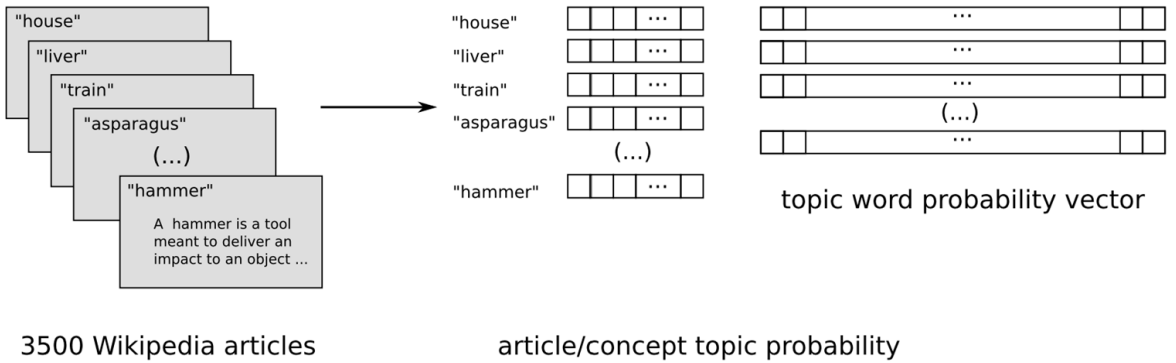
Text Stimulus Representations

- Basic NLP Representations
  - Corpus co-occurrence counts
  - Topic models
  - Linguistic: POS, dependencies, roles.
- Discourse
  - Characters, motion, speech, emotions, non-motion verbs
- Deep Learning based Representations
  - Embeddings
  - Longer context using LSTMs
  - Transformers
- Experiential attributes
  - Rated on 0-6 scale
  - Binary

Basic NLP Representations for Word Stimuli

- Corpus co-occurrence counts
  - 25 verbs (Mitchell et al., 2008; Pereira et al., 2013)
    - Basic sensory and motor activities, actions per formed on objects, and actions involving changes to spatial relationships.
    - Verbs: see, hear, listen, taste, smell, eat, touch, nib, lift, manipulate, run, push, fill, move, ride, say, fear, open, approach, near, enter, drive, wear, break, and clean.
    - For each (verb, stimulus word w), feature value = normalized co-occurrence count of w with any of three forms of the verb (e.g., taste, tastes, or tasted) over the text corpus.
  - 985 common English words (such as above, worry, and mother) in (Huth et al., 2016).

- Topic models (Pereira et al., 2013)
  - Get relevant Wiki pages (e.g., “airplane” is “Fixed-Wing Aircraft”) and other linked pages (e.g. “Aircraft cabin”)
  - LDA topic modelling on 3500 pages with #topics from 10 to 100, in increments of 5, setting the α parameter to 25/#topics.
  - LSA topic modelling (Wang et al., 2017)



Basic NLP Representations for Word Stimuli

- Word length
- Is the word related to one of the 28 unique parts of speech and 17 unique dependency relationships?
- Position of word in the sentence
- Roles
  - Main verb
  - Agent or experiencer
  - Patient or recipient
  - Predicate of a sentence (The window was dusty)
  - Modifier (The angry activist broke the chair)
  - Complement in adjunct and propositional phrase, including direction, location, and time (The restaurant was loud at night).

Wehbe, Leila, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. "Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses." *PloS one* 9, no. 11 (2014): e112575.

Wang, Jing, Vladimir L. Cherkassky, and Marcel Adam Just. "Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states." *Human brain mapping* 38, no. 10 (2017): 4865-4881.

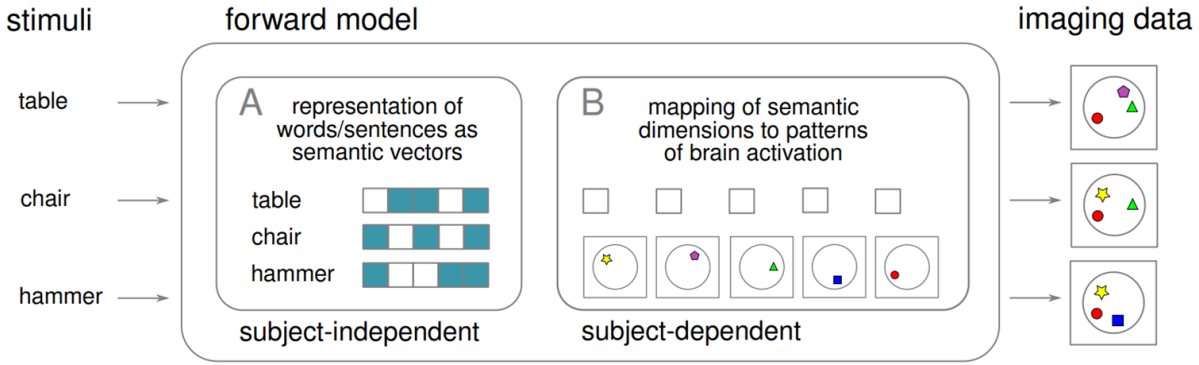
Discourse features (for Harry Potter dataset)

- Characters: Binary features to signal which of the 10 characters are mentioned.
- Motions: Identify a set of motions that occurred frequently in the chapter (e.g. fly, manipulate, collide physically, etc.).
- Speech: Indicate the parts of the story that correspond to direct speech between the characters. Used the presence of dialog as a feature.
- Emotions: Identified a set of emotions that were felt by the characters in the chapter (e.g. annoyance, nervousness, pride, etc.).
- Verbs: Identified a set of actions that occurred frequently in the chapter that were distinct from motion (e.g. hear, know, see, etc.).

Wehbe, Leila, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. "Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses." *PloS one* 9, no. 11 (2014): e112575.

Wang, Jing, Vladimir L. Cherkassky, and Marcel Adam Just. "Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states." *Human brain mapping* 38, no. 10 (2017): 4865-4881.

DL Representations: Using embeddings for word stimuli

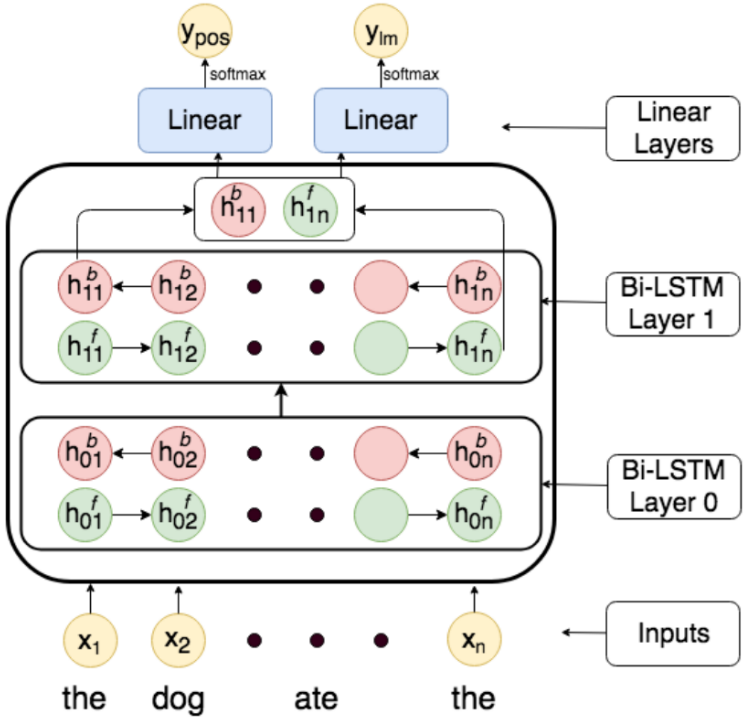


	Noun	Verb	Adjective
GloVe	**0.8768**(0.0792)	0.8544(0.0713)	0.8337(0.1081)
Word2Vec	**0.8386**(0.0942)	0.8309(0.0636)	0.8210(0.1028)
Fasttext	**0.8407**(0.0676)	0.8235(0.0766)	0.8077(0.0996)
RWSGwn	**0.8123**(0.0886)	0.7453(0.0771)	0.7425(0.1032)
ELMo	**0.9088**(0.0632)	0.8520(0.0797)	0.7993(0.1244)
ConceptNet	0.8646(0.0875)	**0.8702**(0.0695)	0.8249(0.0925)
Dependency	**0.8554**(0.0731)	0.8137(0.0755)	0.7891(0.0808)

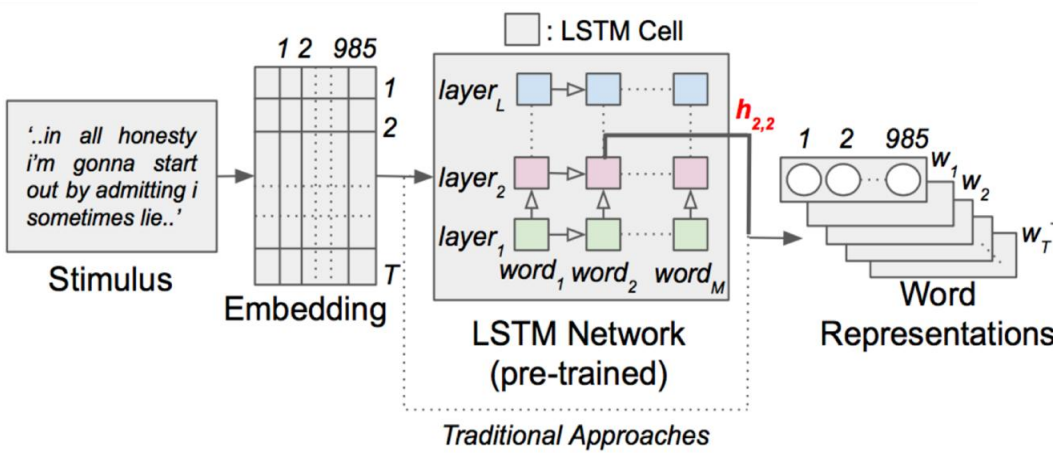
- GloVe 300D vectors (Pereira et al., 2016; Wang et al., 2017; Pereira et al., 2018; Anderson et al., 2019)
- 300D embeddings by training a skip-gram model using negative sampling (SGNS) on Italian and English Wikipedia dumps using Gensim. (Anderson et al., 2017a)
- FastText (Berezutskaya et al., 2020)
- Comparison across multiple embedding methods
  - GloVe, word2vec, WordNet2Vec, FastText, ELMo (Hollenstein et al., 2019)
  - word2Vec, fastText, GloVe, Dependency-based word2vec, RWSGwn, ConceptNet, ELMo, averaged and concatenated combinations (Wang et al., 2020)

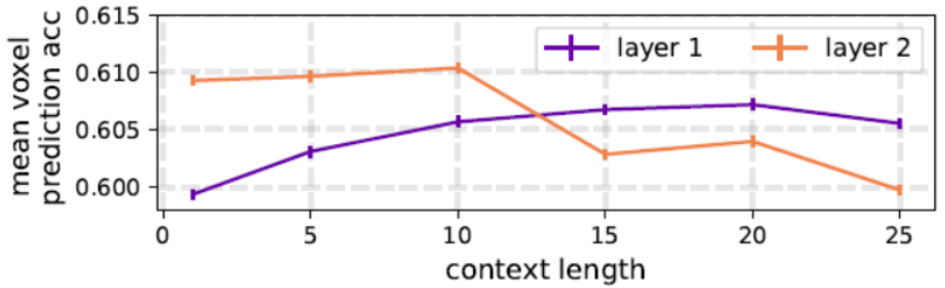
DL Representations: Using longer context for word stimuli

- Multi-task LSTMs
  - Predict next word and POS of next word.



- ELMo embeddings: LSTM based pretrained language model





(a) ELMo

Toneva, Mariya, and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." *Advances in Neural Information Processing Systems* 32 (2019).

Jain, Shailee, and Alexander Huth. "Incorporating context into language encoding models for fMRI." *Advances in neural information processing systems* 31 (2018).

Jat, Sharmistha, Hao Tang, Partha Talukdar, and Tom Mitchell. "Relating simple sentence representations in deep neural networks and the brain." *arXiv preprint arXiv:1906.11861* (2019).

DL Representations: Using sentence embeddings

- Unstructured Models: Ignore sentence structure
  - Simple Pooling Methods
    - Average/max/concat(max, avg) pooling over word embeddings.
  - Advanced Pooling Methods
    - FastSent (Hill, Cho, and Korhonen 2016)
    - SIF (Arora, Liang, and Ma 2016) adapts the naïve averaging of word embeddings to weighted averaging.
- Structured Models
  - Unsupervised Methods: Skip-thought, QuickThought.
  - Supervised Methods: InferSent, GenSen (Subramanian et al. 2018), Universal Sentence Encoder

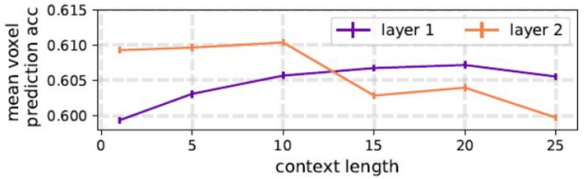
Topic	Passage	Sentence [b]
Musical Instruments	Piano	1. The piano is a popular musical instrument… 2. Pressing a piano key causes a felt-tipped hammer... 3. The piano has an enormous note range.
	Accordion	1. A clarinet is a woodwind musical instrument… 2. It is a long black tube with a flare at the bottom 3. The player chooses notes by pressing keys and holes.
	Clarinet	1. An accordion is a portable musical instrument. 2. One keyboard is used for individual notes 3. Accordions produce sound with bellow that blow air
…	…	…

[a]	Ridge			Lasso			MLP		
	topic	passa.	sente.	topic	passa.	sente.	topic	passa.	sente.
Max	0.88	0.76	0.65	0.88	0.75	0.70	0.83	0.70	0.63
Avg	0.90	0.83	0.73	0.92	0.81	0.78	0.89	0.78	0.67
Cat	0.92	0.83	0.74	0.90	0.81	0.80	0.86	0.74	0.66
Sif	0.89	0.84	0.69	0.91	0.77	0.72	0.84	0.73	0.65
Fast	0.92	0.81	0.74	0.90	0.79	0.77	0.88	0.76	0.67
Skip	0.90	0.82	0.75	0.91	0.80	0.79	0.86	0.81	0.73
Quik	0.91	0.84	0.75	0.91	0.81	0.79	0.90	0.82	0.77
Gen	0.91	0.84	0.78	0.92	0.84	**0.84**	0.91	**0.84**	**0.80**
Inf	**0.94**	**0.90**	**0.83**	**0.93**	**0.86**	**0.84**	**0.92**	**0.84**	0.79

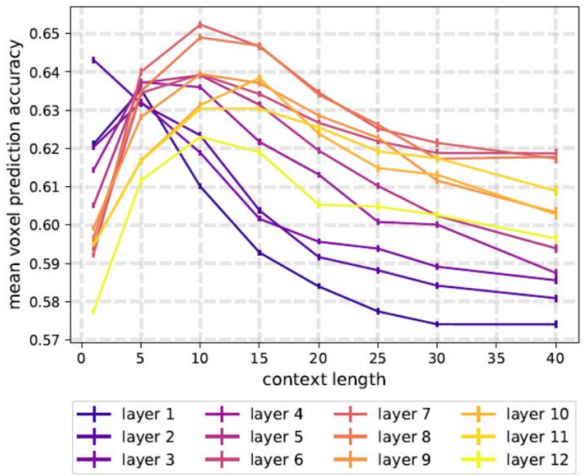
Toneva, Mariya, and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." *Advances in Neural Information Processing Systems* 32 (2019).

Sun, Jingyuan, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. "Towards sentence-level brain decoding with distributed representations." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 7047-7054. 2019.

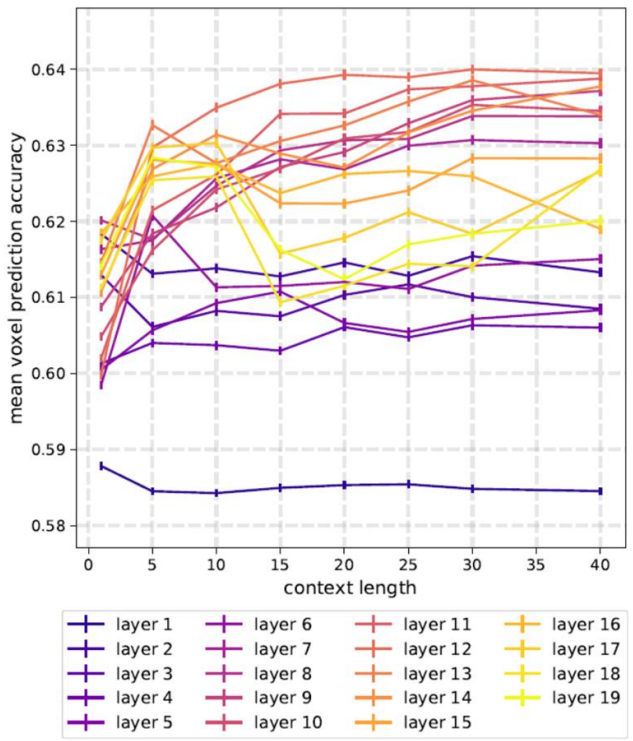
DL Representations: Transformer-based methods for text stimuli (Layer #, context length, architecture)



(a) ELMo

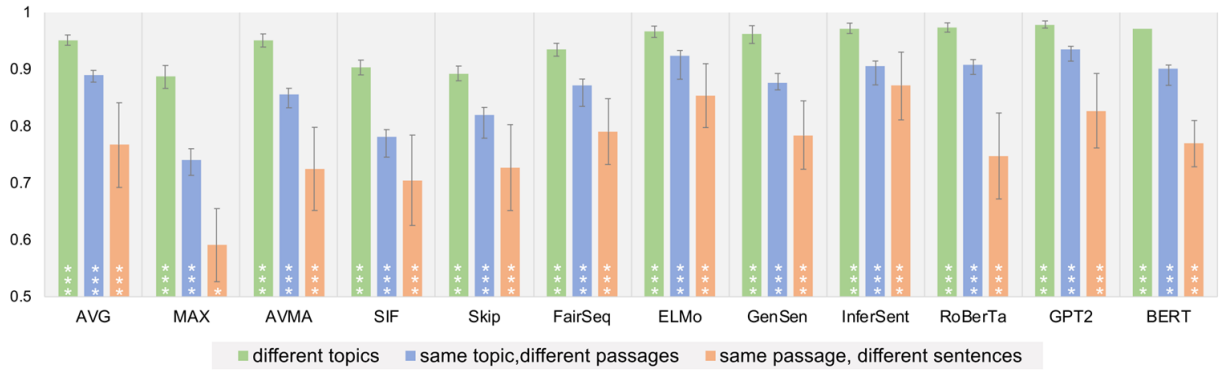


(b) BERT



(c) T-XL

DSM	Name	Structure and Training Task
Unstructured	AVG	Average Pooling
	MAX	Max Pooling
	AVMA	Concatenation of AVG and Max
	SIF	Weighted Average Pooling
Structured	FairSeq	CNN (language model)
	Skip	LSTM (language model)
	GenSen	BiLSTM (multi-task learning)
	InferSent	CNN-BiLSTM (natural language inference)
	ELMo	CNN-BiLSTM (language model)
	BERT	Transformer (language model)
	RoBerTa	
	GPT2	



Transformer-XL is the only model that continues to increase performance as the context length is increased. In all networks, the middle layers perform the best for contexts longer than 15 words. The deepest layers across all networks show a sharp increase in performance at short-range context (fewer than 10 words), followed by a decrease in performance. [Toneva and Wehbe, 2019]

Toneva, Mariya, and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." *Advances in Neural Information Processing Systems* 32 (2019).

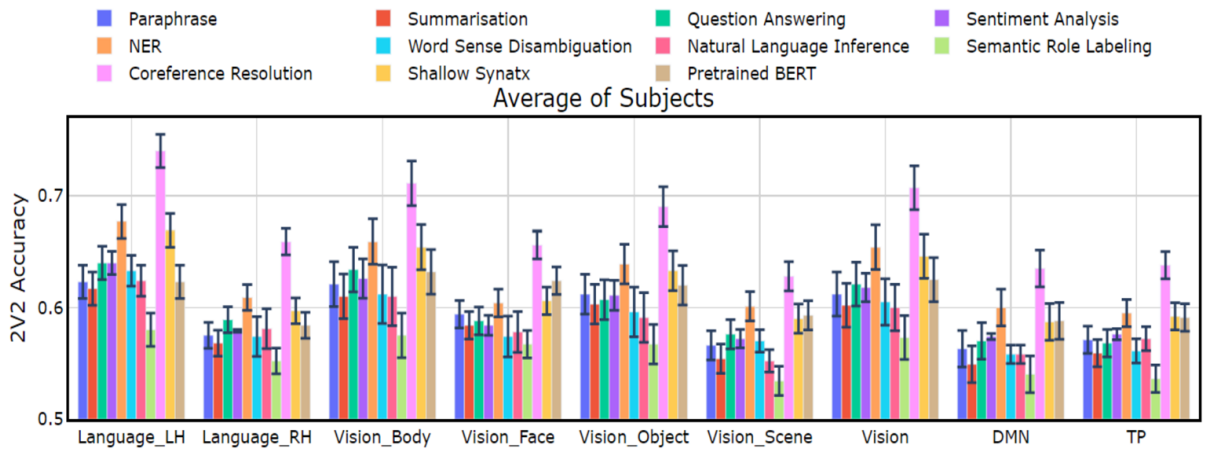
Sun, Jingyuan, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. "Neural encoding and decoding with distributed sentence representations." *IEEE Transactions on Neural Networks and Learning Systems* 32, no. 2 (2020): 589-603.

DL Representations: Transformer-based methods for text stimuli (NLP task finetuning)

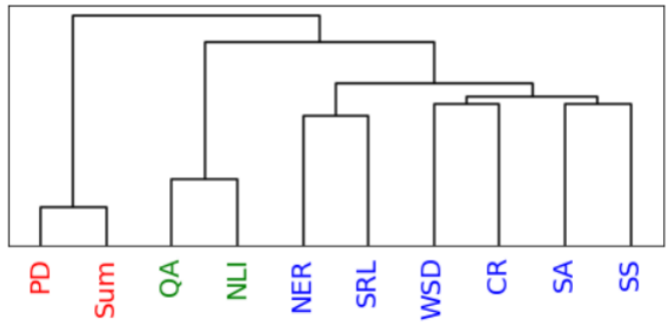
Task	HuggingFace Model Name	Dataset
NLI	bert-base-nli-mean-tokens	Stanford Natural Language Inference (SNLI), MultiNLI
PD	bert-base-cased-finetuned-mrpc	Microsoft Research Paraphrase Corpus (MRPC)
SS	bert-base-chunl	CoNLL-2003
Sum	bart-base-samsum	SAMSum
WSD	bert-base-baseline	English all-words
CR	bert_coreference_base	OntoNotes and GAP
NER	bert-base-NER	CoNLL-2003
QA	bert-base-qa	SQUAD
SA	bert-base-sst	Stanford Sentiment Treebank (SST)
SRL	bert-base-srl	English PropBank SRL

Tasks

Paraphrase, Summarization, Question Answering, Sentiment Analysis, NER, Word Sense Disambiguation, Natural Language Inference, Semantic Role Labeling, Coreference Resolution, Shallow Syntax Parsing



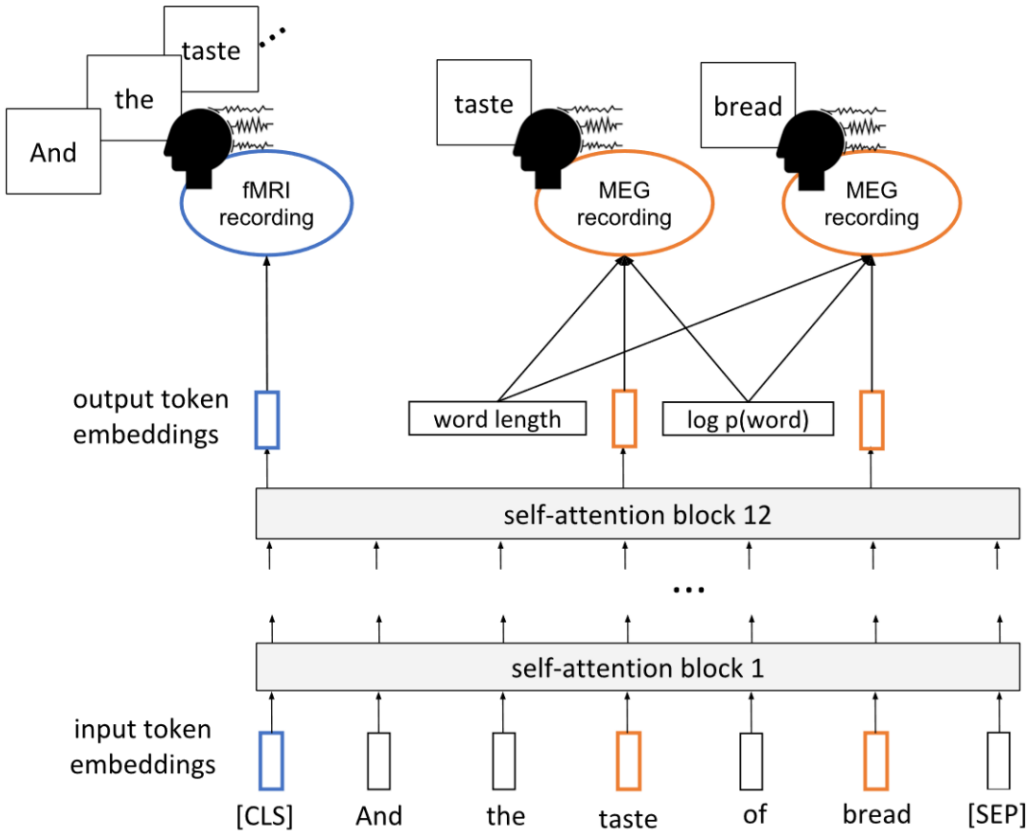
Pereira dataset: CR, NER, and SS perform the best.



Dendrogram constructed using similarity on representations from task-specific Transformer encoder models with stimuli from the dataset passed as input.

Oota, Subba Reddy, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampudi. "Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity?." *arXiv preprint arXiv:2205.01404* (2022).

DL Representations: Transformer-based methods for text stimuli (Multi-task setup)



- Settings
  - Finetune BERT vs not
  - Finetune BERT using one representative subject and train dense layer for each subject, vs finetune BERT for each subject.
  - Finetune BERT on MEG for all subjects, then finetune BERT on fMRI.
  - Multi-task finetune BERT for fMRI+MEG prediction task
- Results
  - Fine-tuned models predict fMRI data better than vanilla BERT
  - Relationships between text and brain activity generalize across experiment participants.
  - Using MEG data can improve fMRI predictions.
  - A single model can be used to predict fMRI activity across multiple experiment participants.

Schwartz, Dan, Mariya Toneva, and Leila Wehbe. "Inducing brain-relevant bias in natural language processing models." *Advances in neural information processing systems* 32 (2019).

Experiential attributes model for text stimuli

- Represents words in terms of human ratings of their degree of association with different attributes of experience
  - "On a scale of 0 to 6, to what degree do you think of a banana as having a characteristic or defining color?"
  - Anderson et al., 2019: 65 attributes spanning sensory, motor, affective, spatial, temporal, causal, social, and abstract cognitive experiences.
- Value-add on top of text models: a lot of experiential information goes unstated in natural verbal communication.
  - E.g., it is rarely useful to communicate the color of bananas because it is obvious to all those with experience of bananas.
  - E.g., it would be unusual to specify that dropping things involves movement.
- Nishida et al., 2020 use a subset of 20 attributes.

Table 1 List of attributes first arranged by modality, and then subdivided into individual attributes

Dominant modality	Attribute
Vision	vision, bright, dark, color, pattern, large, small, motion, biomotion, fast, slow, shape, complexity, face, body.
Auditory	audition, loud, low, high, sound, music, speech.
Somatosensory	touch, temperature, texture, weight, pain.
Gustatory +Smell	taste, smell.
Motor	head, upper limb, lower limb, practice.
Attention	attention, arousal.
Event	duration, long, short, caused, consequential, social, time.
Evaluation	benefit, harm, pleasant, unpleasant.
Cognition	human, communication, self, cognition, number.
Emotion	happy, sad, angry, disgusted, fearful, surprised.
Drive	drive, needs.
Spatial	landmark, path, scene, near, toward, away.

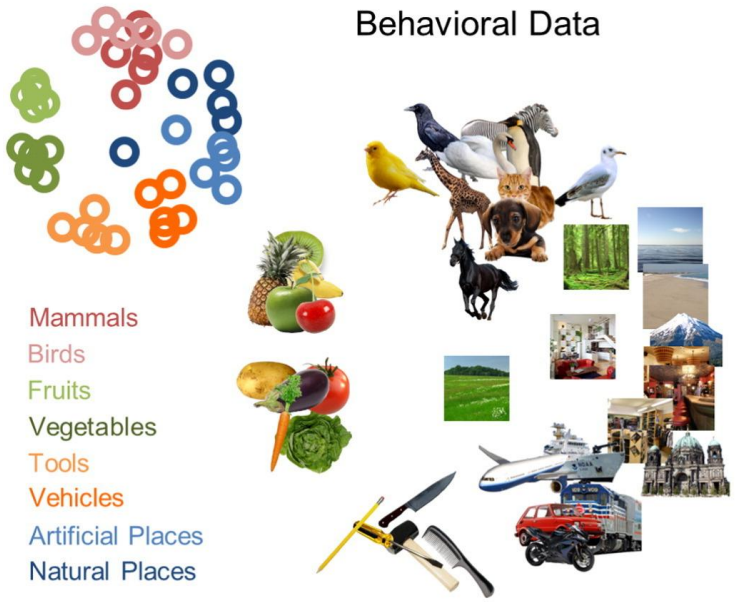
Anderson, Andrew James, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev DS Raizada, Feng Lin, and Edmund C. Lalor. "An integrated neural decoder of linguistic and experiential meaning." *Journal of Neuroscience* 39, no. 45 (2019): 8969-8987.

Anderson, Andrew James, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev DS Raizada. "Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation." *Cerebral Cortex* 27, no. 9 (2017): 4379-4395.

Anderson, Andrew James, Kelsey McDermott, Brian Rooks, Kathi L. Heffner, David Dodell-Feder, and Feng V. Lin. "Decoding individual identity from brain activity elicited in imagining common experiences." *Nature communications* 11, no. 1 (2020): 1-14.

Binary attribute representations

- Each stimulus is represented using a binary vector capturing membership to one of the eight semantic categories.



- 42 neurally plausible semantic features (NPSFs)
  - Perceptual and affective characteristics of an entity (10 NPSFs coded such features, such as man-made, size, color, temperature, positive affective valence, high affective arousal), animate beings (person, human-group, animal), and time and space properties (e.g. unenclosed setting, change of location)

Word	NPSF features
Interview	Social, Mental action, Knowledge, Communication, Abstraction
Walk	Physical action, Change of location
Hurricane	Event, Change of physical state, Health, Natural, Negative affective valence, High affective arousal
Cellphone	Social action, Communication, Man-made, Inanimate
Judge	Social norms, Knowledge, Communication, Person
Clever	Attribute, Mental action, Knowledge, Positive affective valence, Abstraction

Handjaras, Giacomo, Emiliano Ricciardi, Andrea Leo, Alessandro Lenci, Luca Cecchetti, Mirco Cosottini, Giovanna Marotta, and Pietro Pietrini. "How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge." *Neuroimage* 135 (2016): 232-242.

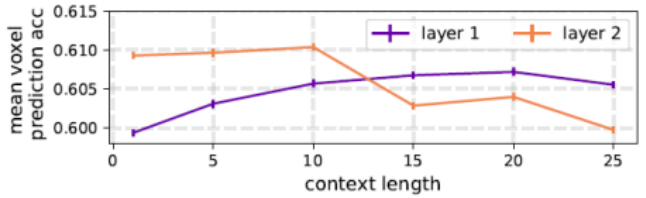
Wang, Jing, Vladimir L. Cherkassky, and Marcel Adam Just. "Predicting the brain activation pattern associated with the propositional content of a sentence: modeling neural representations of events and states." *Human brain mapping* 38, no. 10 (2017): 4865-4881.

Agenda

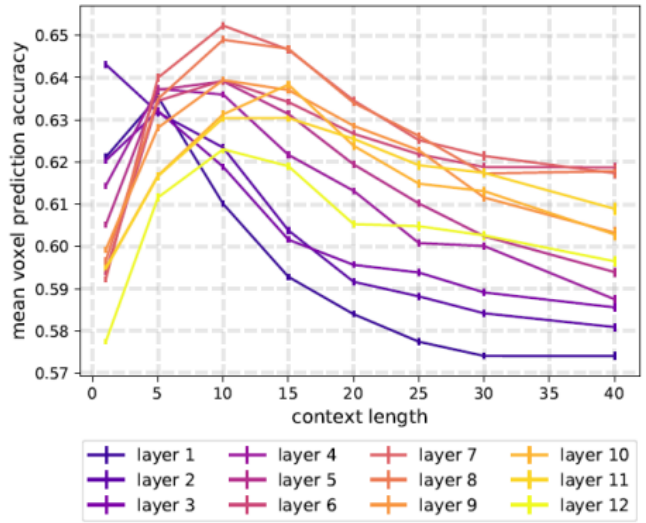
- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
  - Brain Encoding/Decoding and applications
  - Introduction to popular datasets
  - Text Stimulus Representations
  - **Alignment Between AI Models and Human Brain Language Comprehension**
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

Does context size impact alignment?

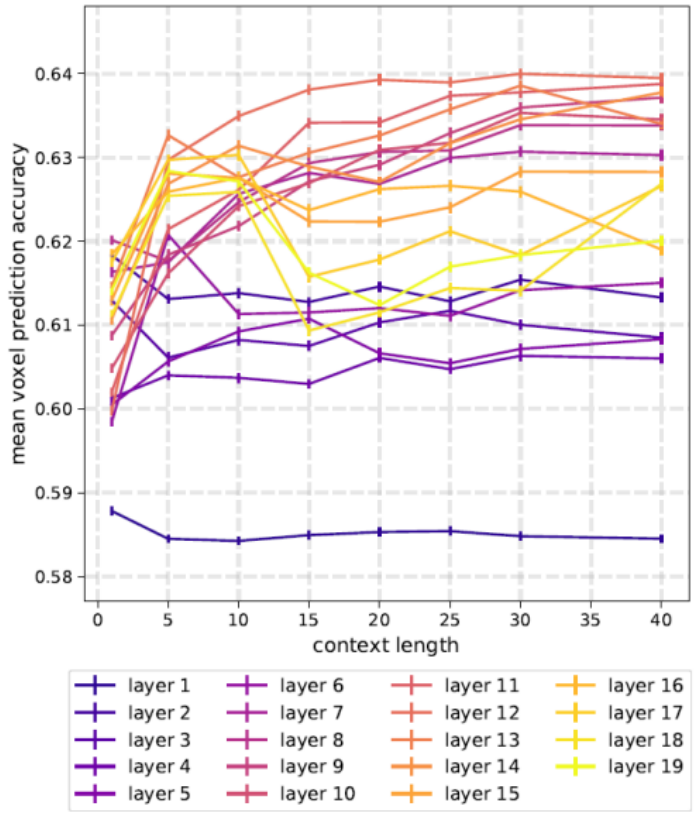
- ELMO, BERT, USE (Universal Sentence Encoder) and T-XL.
- Transformer-XL continues to increase perf as the context length is increased.
- Middle layers perform the best for contexts of 15+ words.
- Deepest layers show a sharp increase in perf at short-range context (<10 words).



(a) ELMo



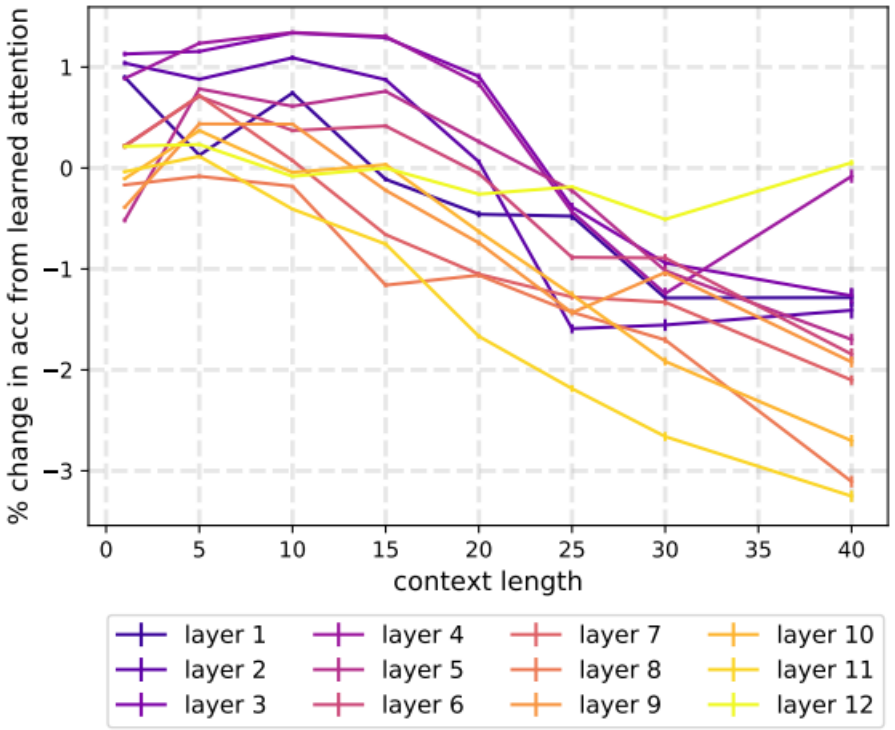
(b) BERT



(c) T-XL

Toneva, Mariya, and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." *Advances in neural information processing systems* 32 (2019).

What if we use uniform attention in first few layers?



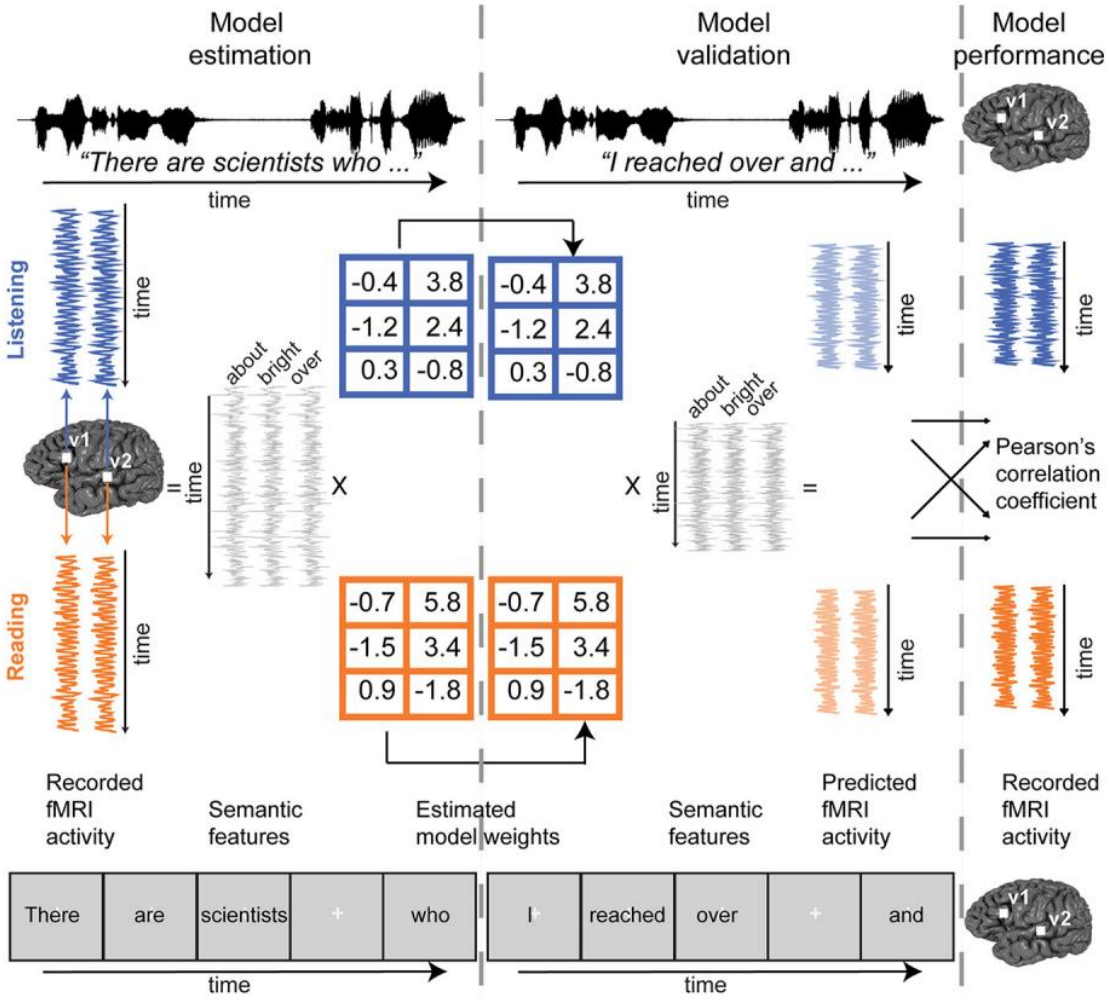
condition	uni L1	uni L2	uni L6	uni L11	base
simple	**1.00**	**1.00**	**1.00**	0.98	**1.00**
in a sentential complement	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**
short VP coordination	0.88	0.90	**0.91**	0.88	0.89
long VP coordination	0.96	0.97	**1.00****	0.96	0.98
across a prepositional phrase	0.86	**0.93****	0.88	0.82	0.85
across a subject relative clause	0.83	0.83	**0.85****	0.83	0.84
across an object relative clause	0.87	0.91	**0.92****	0.86	0.89
across an object relative clause (no that)	**0.87**	0.80	**0.87**	0.84	0.86
in an object relative clause	**0.97****	0.95	0.91	0.93	0.95
in an object relative clause (no that)	**0.83****	0.72	0.74	0.72	0.79
reflexive anaphora: simple	0.91	0.94	**0.99****	0.95	0.94
reflexive anaphora: in a sent. complem.	0.88	0.85	0.86	0.85	**0.89**
reflexive anaphora: across rel. clause	0.79	**0.84****	0.79	0.76	0.80

- Shallow layers benefit from the uniform attention for context lengths up to 25 words.

- Make attention in layers 1-6 in base BERT uniform, a single layer at a time
- Altered models > pretrained model (‘base’) in 8/13 tasks and ≡ in 4/5 tasks.

Toneva, Mariya, and Leila Wehbe. "Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)." *Advances in neural information processing systems* 32 (2019).

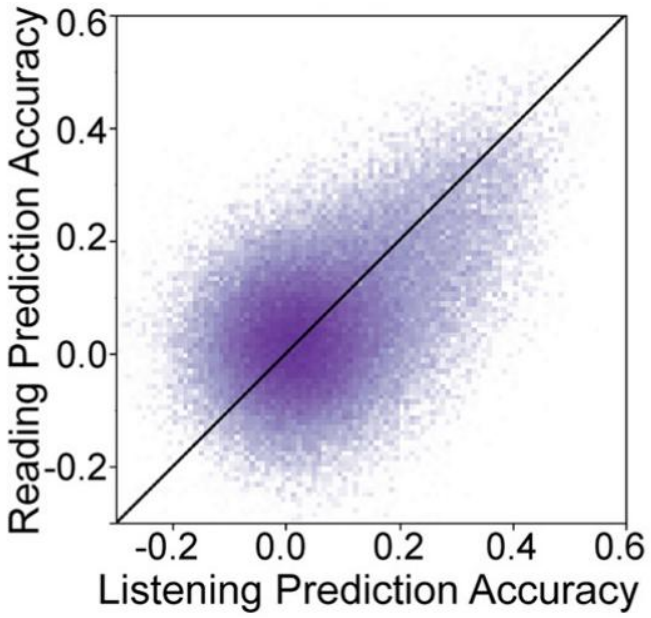
Is the representation of semantic information different when reading vs listening?



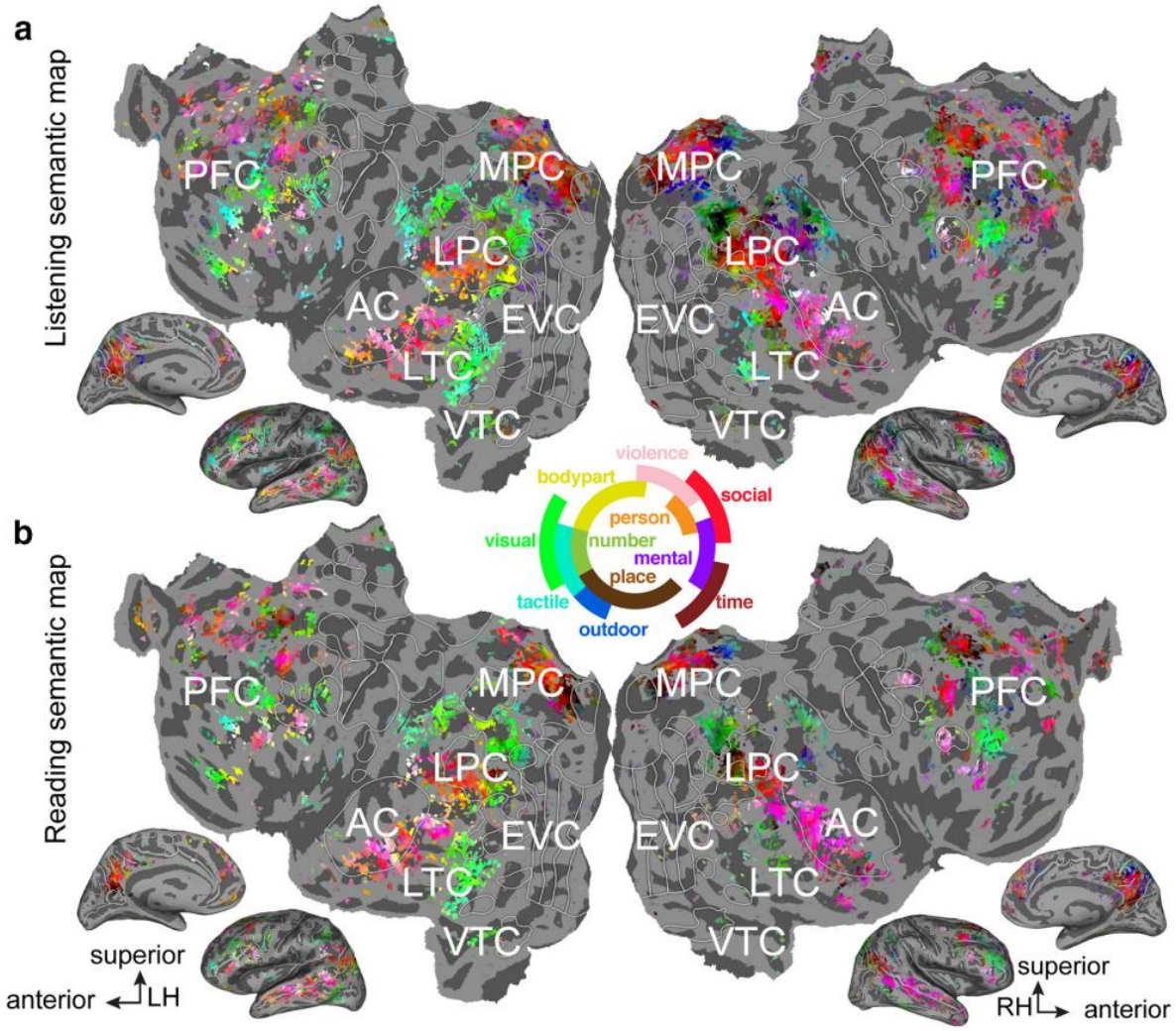
- Train on fMRI for 10 narrative stories (from The Moth Radio Hour) while participants listened to or read several hours
- Voxel-wise modeling with banded ridge regression
- During reading, each word was presented for a duration=duration of that word in the spoken story.
- Features
  - 39 Motion-energy features using spatiotemporal Gabor pyramid
  - 80 spectral audio features based on cochleogram.
  - Word rate, phoneme rate, letter rate, word length variation per TR
  - 39 phoneme frequency
  - 26 letter frequency
  - 56 syntactic binary features: 12 POS and 44 dependency
  - 985 co-occurrence semantics

Deniz, Fatma, Anwar O. Nunez-Elizalde, Alexander G. Huth, and Jack L. Gallant. "The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality." *Journal of Neuroscience* 39, no. 39 (2019): 7722-7736.

Is the representation of semantic information different when reading vs listening?

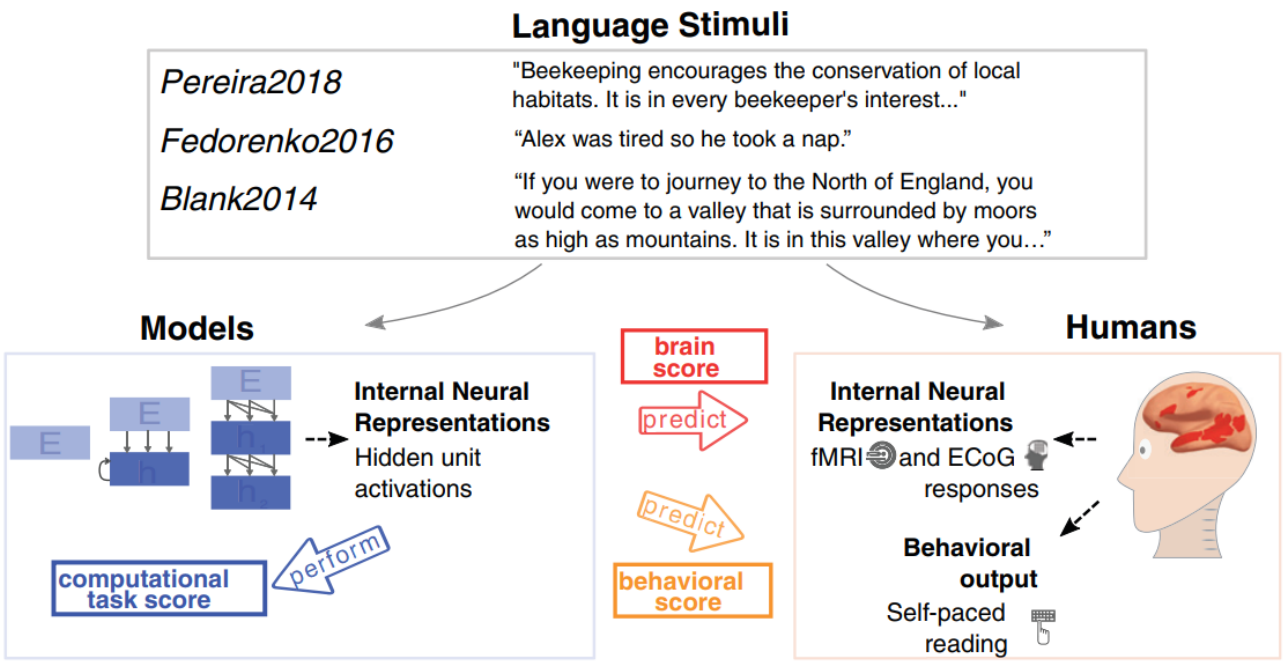


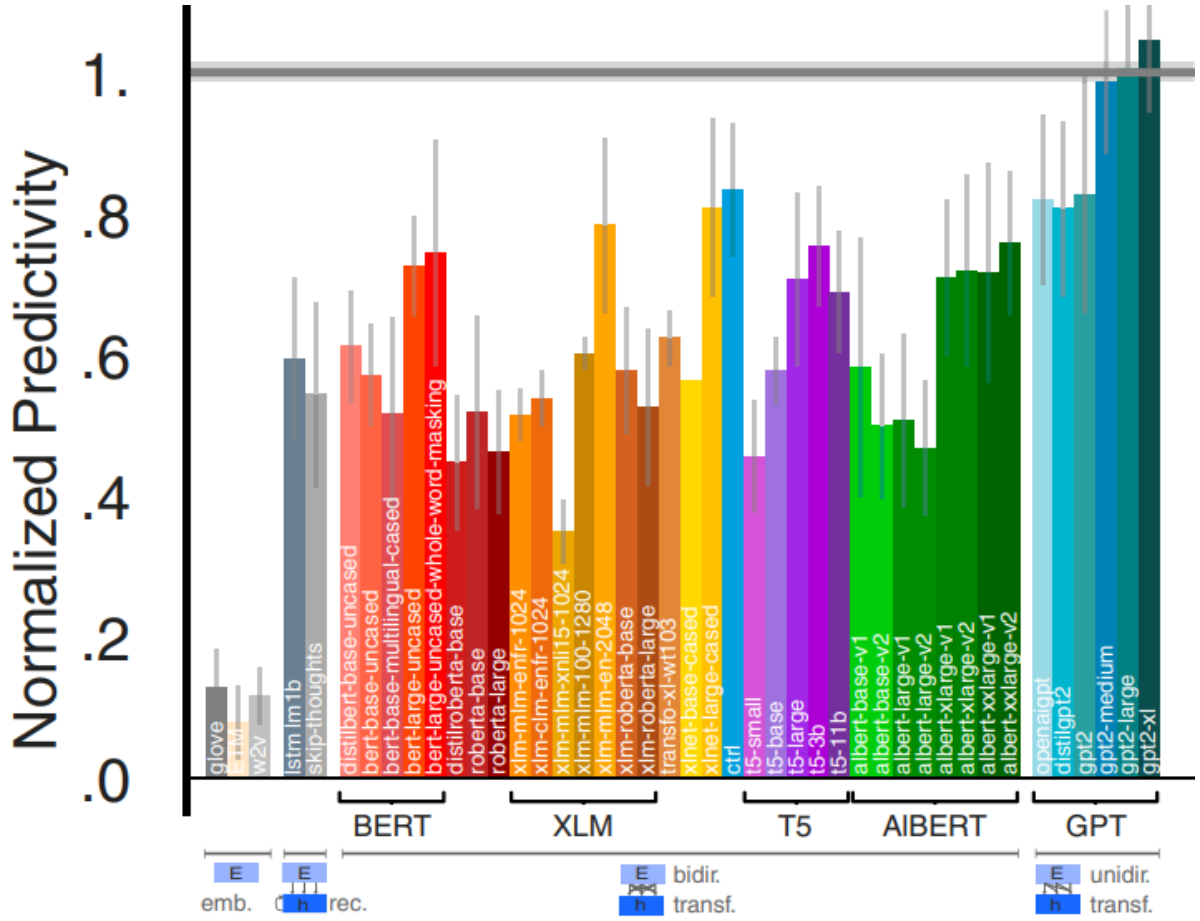
- Semantic tuning during listening and reading are highly correlated in most semantically selective regions of cortex
- Models estimated using one modality accurately predict voxel responses in the other modality.



Deniz, Fatma, Anwar O. Nunez-Elizalde, Alexander G. Huth, and Jack L. Gallant. "The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality." *Journal of Neuroscience* 39, no. 39 (2019): 7722-7736.

Do larger Transformer models lead to better brain-encoding accuracy?

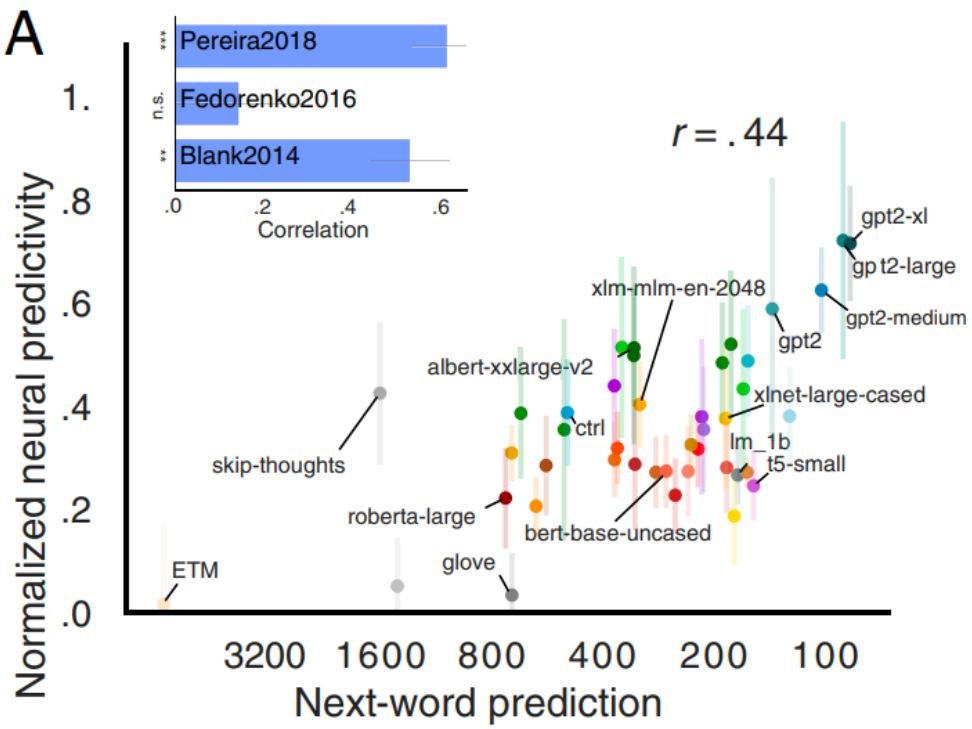


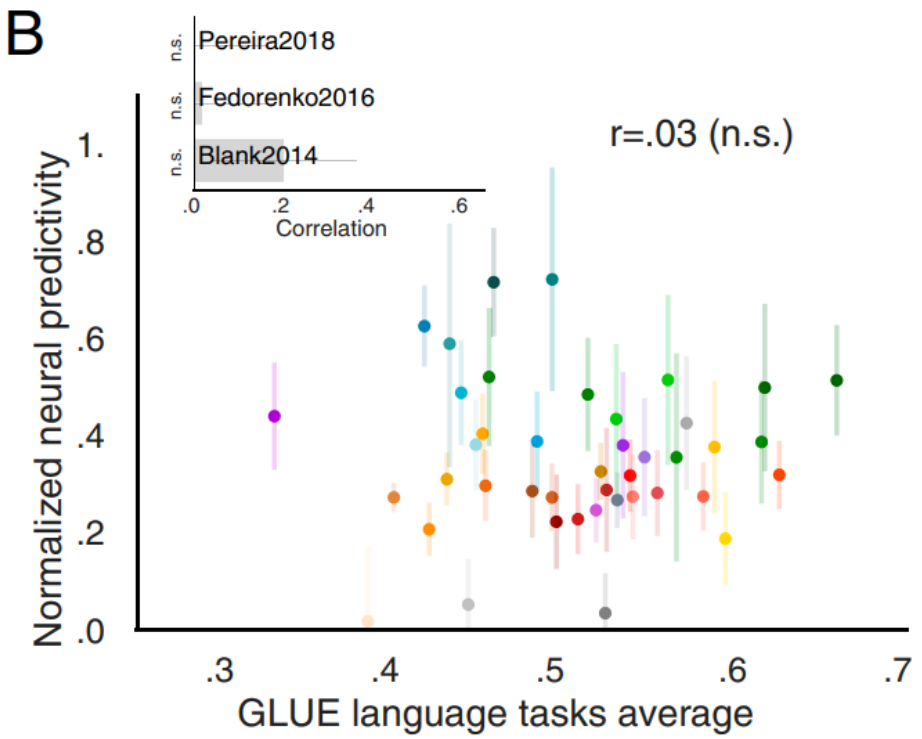


- Transformer models predict nearly 100% of explainable variance in neural responses to sentences.
  - Larger models are better. GPT2-XL is the best.

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. "The neural architecture of language: Integrative modeling converges on predictive processing." *Proceedings of the National Academy of Sciences* 118, no. 45 (2021): e2105646118.

Does improved perf of Transformer models on NLP benchmarks translate to better brain-encoding accuracy?

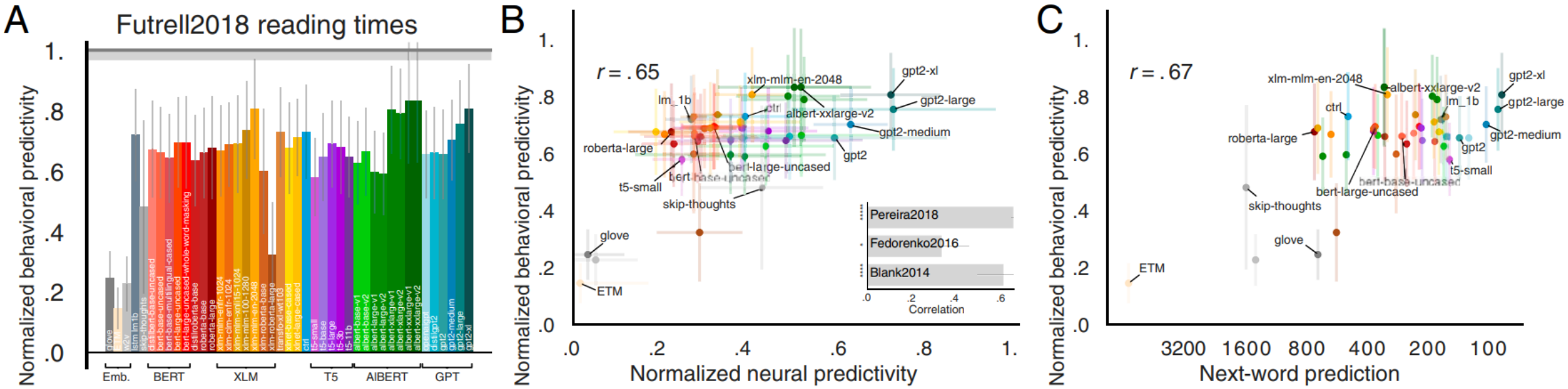




- Next-Word-Prediction Task Performance Selectively Predicts Brain Scores.
- Performance on GLUE tasks does not predict brain scores.

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. "The neural architecture of language: Integrative modeling converges on predictive processing." *Proceedings of the National Academy of Sciences* 118, no. 45 (2021): e2105646118.

Does improved perf of Transformer models on NLP benchmarks translate to better brain-encoding accuracy?

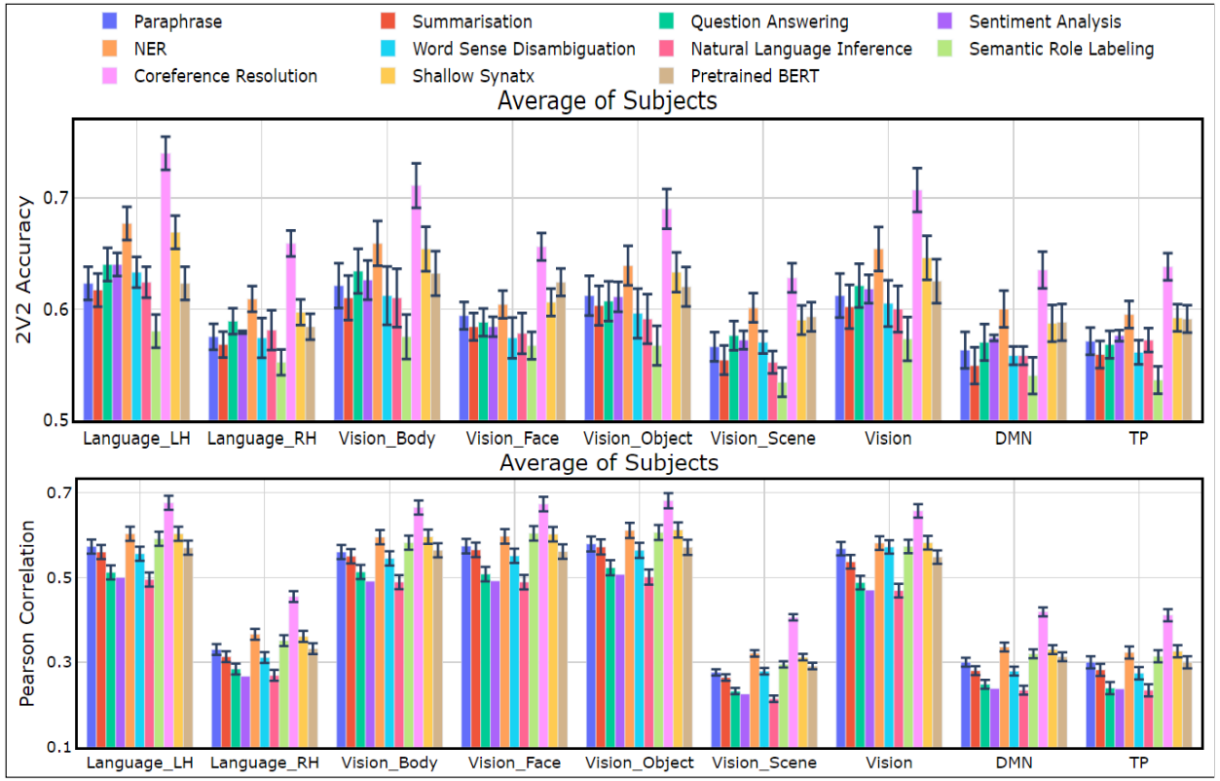


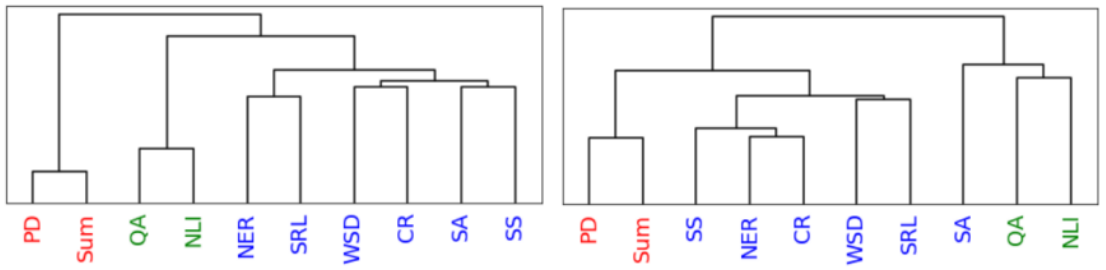
- Behavioral scores (self-paced reading times), brain scores, and next-word-prediction task perf are pairwise correlated.
- Perf on other GLUE tasks does not correlate with behavioral scores

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. "The neural architecture of language: Integrative modeling converges on predictive processing." *Proceedings of the National Academy of Sciences* 118, no. 45 (2021): e2105646118.

Which NLP Tasks are the most Predictive of fMRI Brain Activity?

- Coreference resolution, NER, and shallow syntax parsing explain greater variance for the reading activity.
- For the listening activity, paraphrase generation, summarization, and natural language inference show better encoding performance.
- Tree derived from brain representation vs tree based on Transformer encoder embeddings.

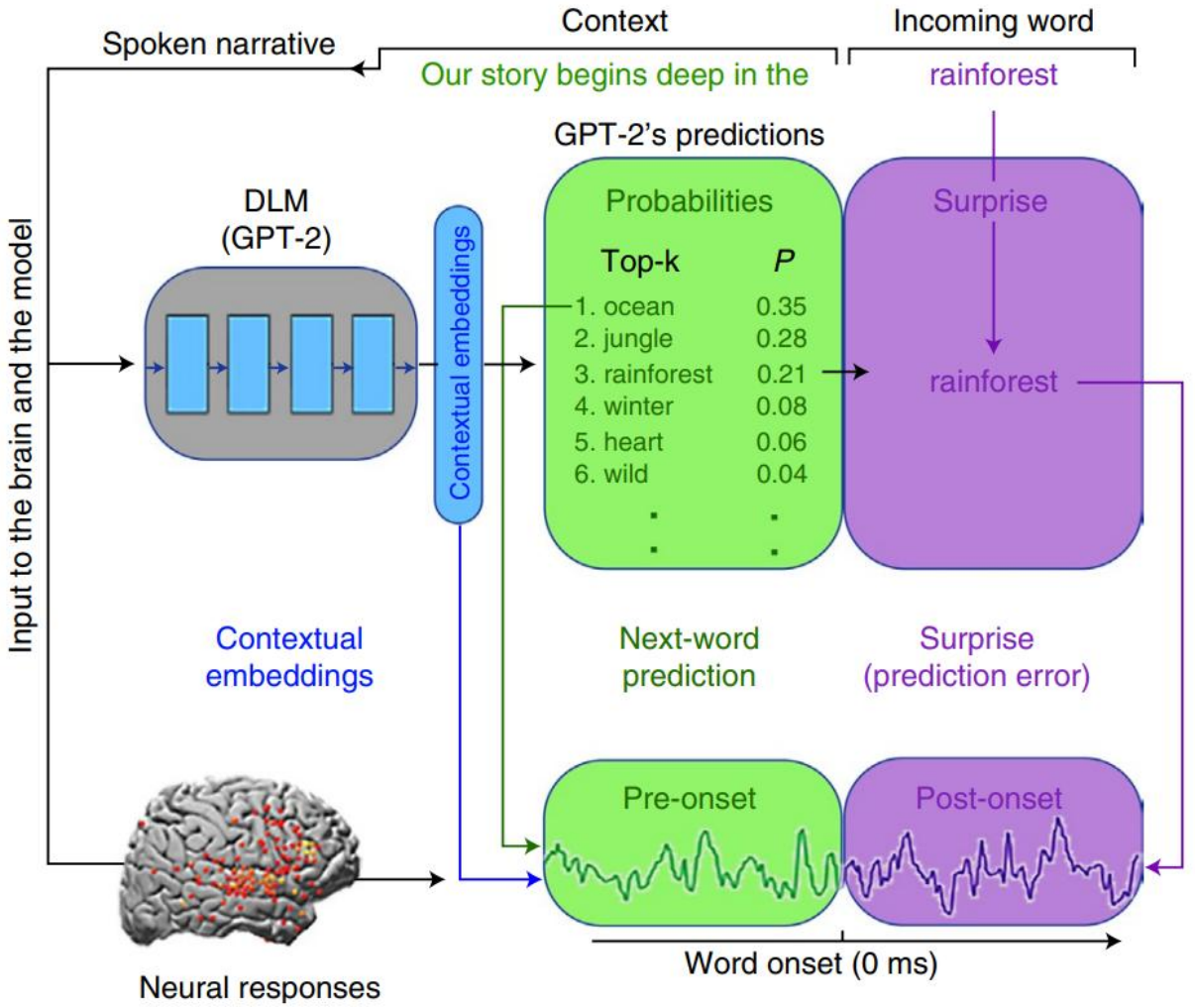




Oota, Subba Reddy, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. "Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity?." In *NAACL-HLT*, pp. 3220-3237. 2022.

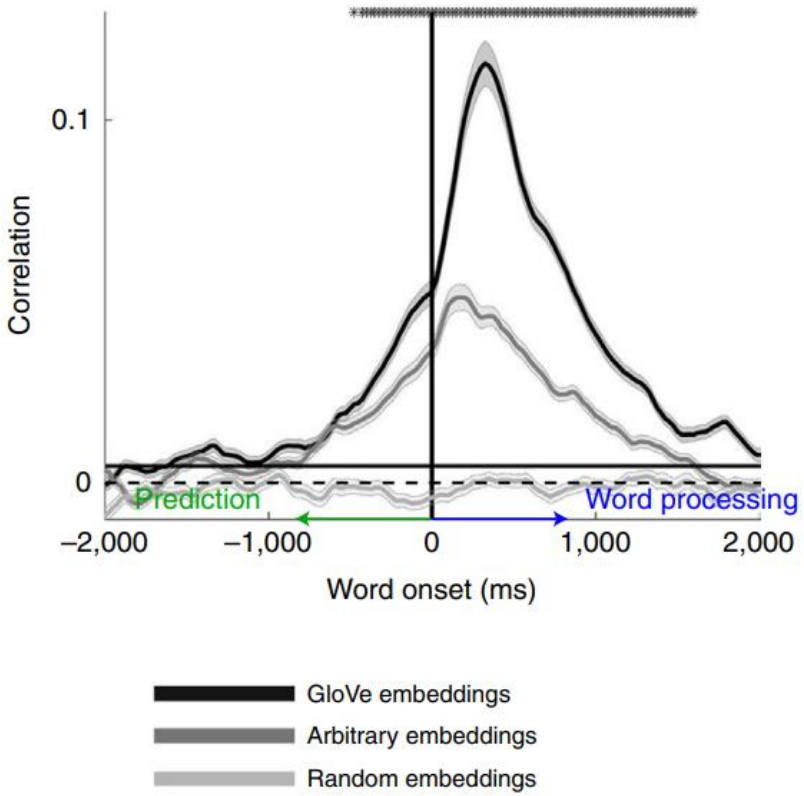
Does the brain also perform next word prediction and is surprised when next word does not match the prediction?

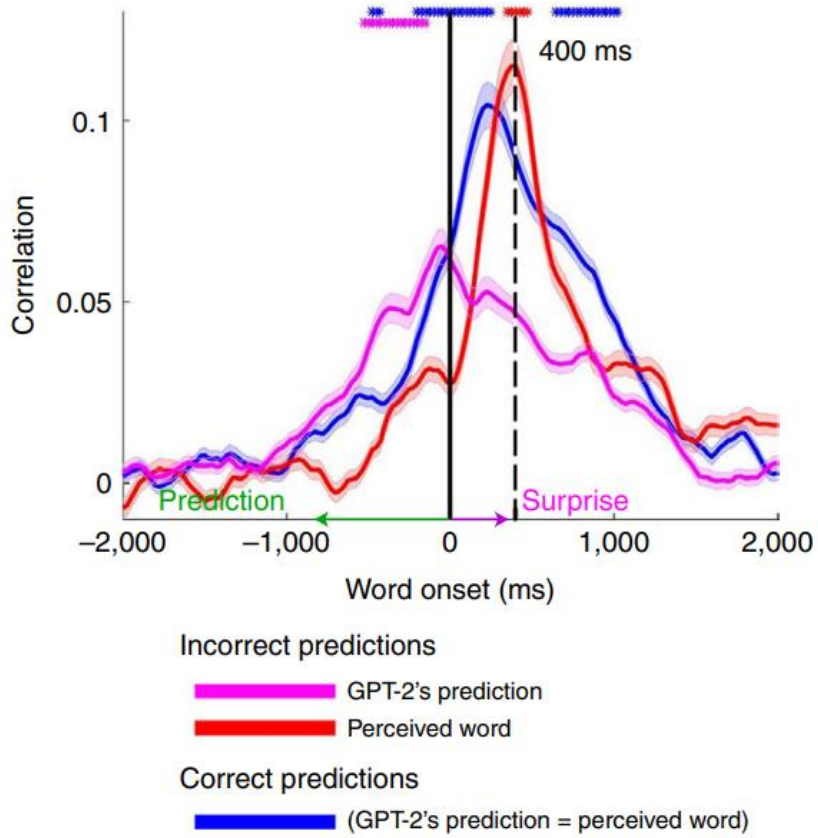
- ECoG while listening to a 30-min podcast.
- Human brains and DLMs
  - Both are engaged in continuous next-word prediction before word onset
  - Both match their pre-onset predictions to the incoming word to calculate post-onset surprise
  - Both rely on contextual embeddings to represent words in natural contexts.



Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase et al. "Shared computational principles for language processing in humans and deep language models." *Nature neuroscience* 25, no. 3 (2022): 369-380.

Does the brain also perform next word prediction and is surprised when next word does not match the prediction?

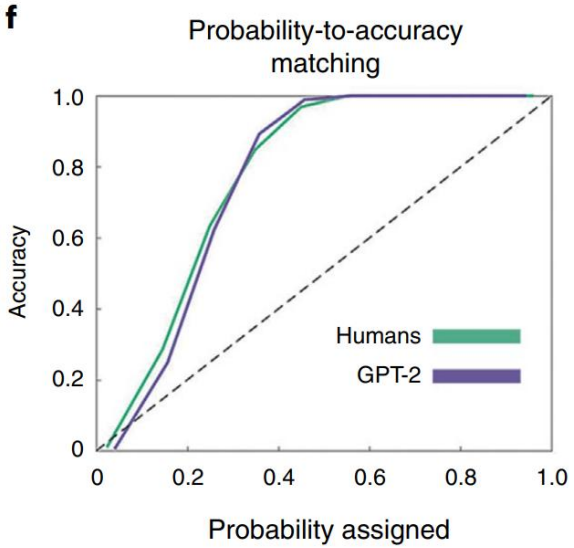


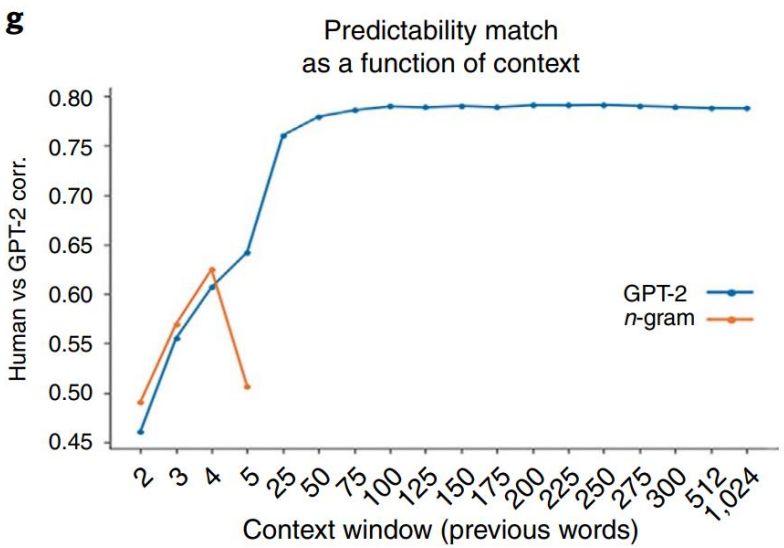


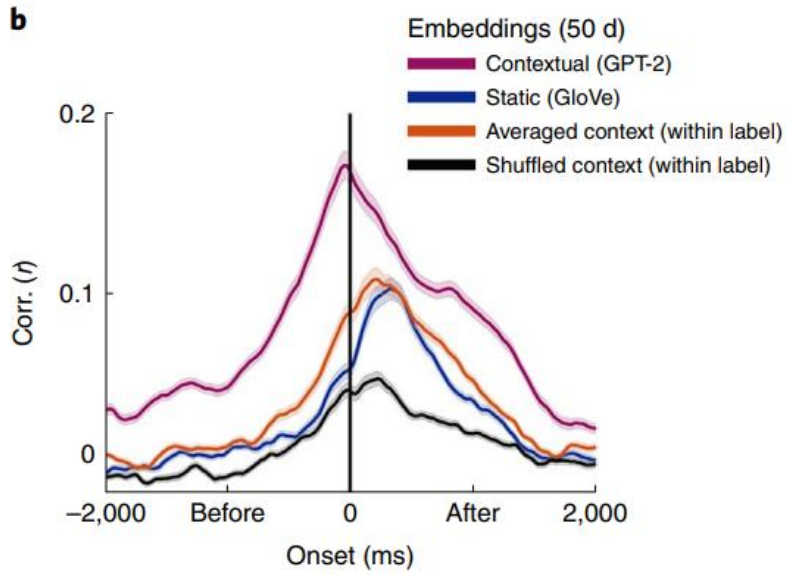
- GloVe yielded significant correlations with predicted neural responses to upcoming words up to 800 ms before word onset.
- Pre-onset activity associated with next-word prediction matches prediction content even when the prediction was incorrect.
- Post-onset activity matches content of the incoming word, even if it was unpredicted.
- Increase in encoding perf for surprising words compared to predicted words 400 ms after word onset.

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase et al. "Shared computational principles for language processing in humans and deep language models." *Nature neuroscience* 25, no. 3 (2022): 369-380.

Does the brain depend on context for processing text?







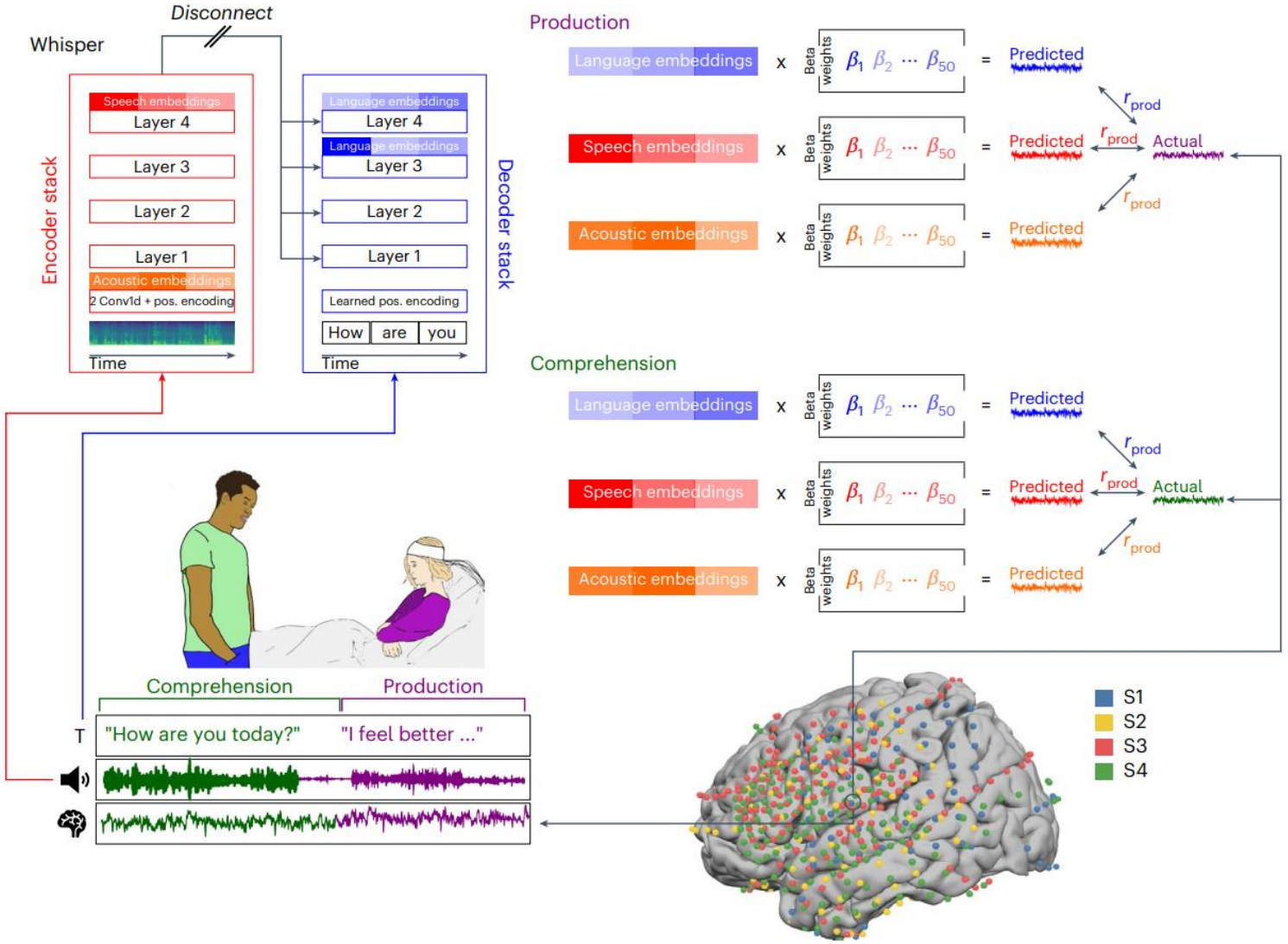
- Match between confidence level and the accuracy level of GPT-2 and human predictions
  - Under-confidence in their predictions and were above 95% correct when the probabilities were higher than 40%.
- Correlation between human and GPT2 word predictions improved as the contextual window increased.

- Contextual embeddings (GPT-2) > static embeddings (GloVe)
- Contextual embeddings (GPT-2) > concatenated GloVe embeddings for 10 prev words
- Avg context: Averaging all embeddings for each unique word (all occurrences of ‘monkey’) into 1 vector
- Scramble the embeddings across different occurrences of the same word in the story (switch embedding of ‘monkey’ in sentence 5 with that in sentence 50).

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase et al. "Shared computational principles for language processing in humans and deep language models." *Nature neuroscience* 25, no. 3 (2022): 369-380.

Do speech models align with language and speech ROIs during speech production and comprehension?

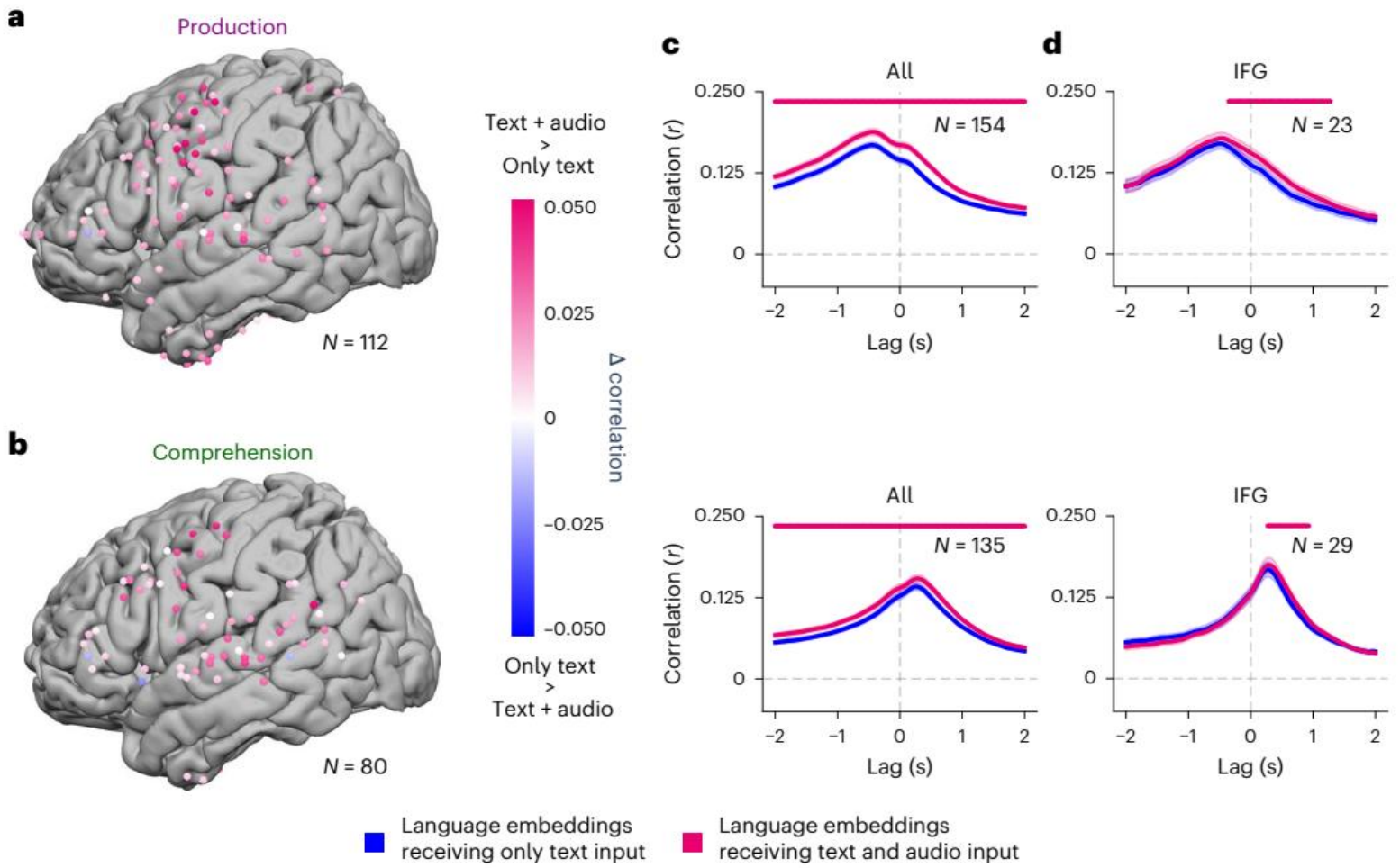
- Extracted low-level acoustic, mid-level speech and contextual word embeddings from a multimodal speech-to-text model (Whisper).
- Sensory and motor regions better align with the model's speech embeddings
- Higher-level language areas better align with the model's language embeddings.



Goldstein, Ariel, Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer et al. "A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations." *Nature human behaviour* (2025): 1-15.

Are both text and audio needed for speech alignment?

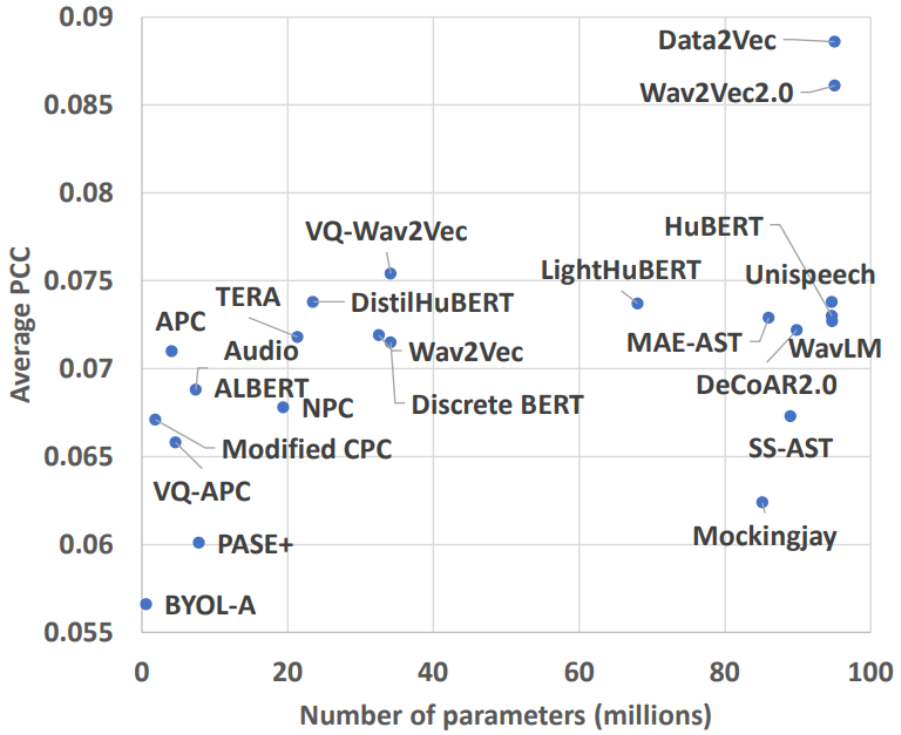
- Only text input (conversation transcripts) vs language receiving audio and text inputs (speech recordings and conversation transcripts)
- Language embeddings fused with auditory features outperform text-only language embeddings in predicting neural activity across multiple electrodes.
- Even though IFG is associated with linguistic processing, across multiple lags, the audio-fused language embeddings yield higher encoding performance during both production and comprehension.



Goldstein, Ariel, Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer et al. "A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations." *Nature human behaviour* (2025): 1-15.

Which speech model aligns the best?

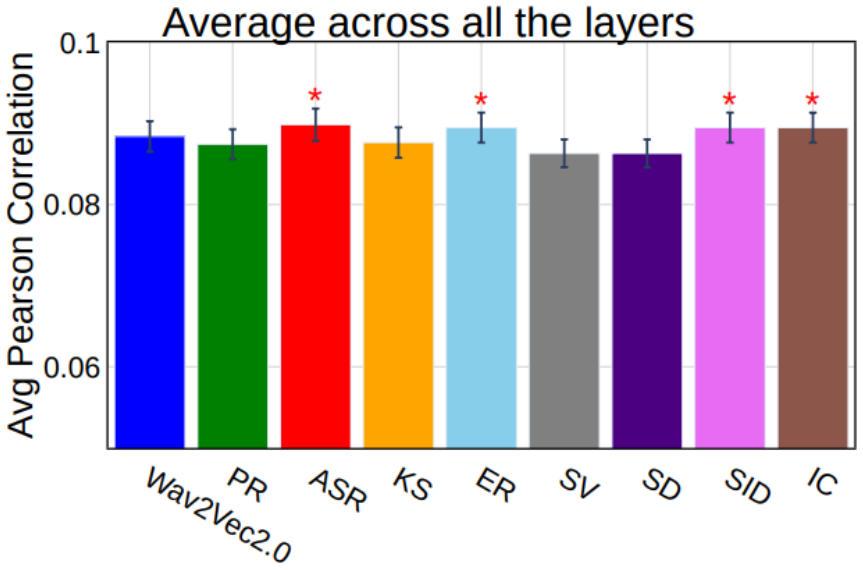
- Wav2Vec2.0, HuBERT, Data2Vec align well.
- Data2Vec aligns best with both language and auditory brain regions

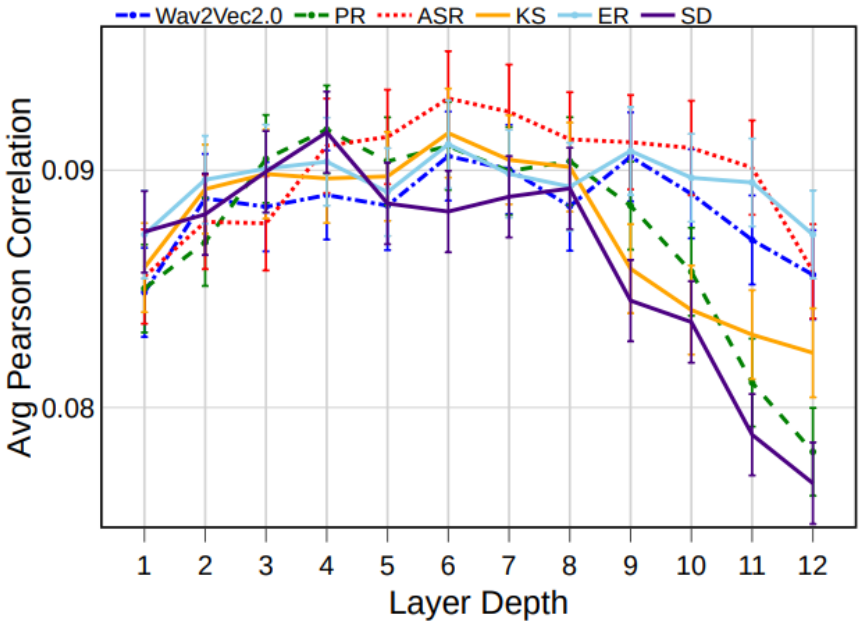


Category	Model	AC	Broca	Whole Brain
Traditional non-DL & non-SS DL Methods	Spectrogram	0.0545	0.0511	0.0495
	Filter bank	0.0477	0.0450	0.0498
	Mel	0.0489	0.0515	0.0511
	MFCC	0.0495	0.0520	0.0517
	VGGish	**0.1612**	**0.0785**	**0.0605**
Generative Self-Supervised Methods	PASE+	0.1272	0.0719	0.0601
	DeCoAR	0.2332	0.1017	0.0695
	DeCoAR2.0	0.2293	**0.1142**	0.0722
	NPC	0.2123	0.0995	0.0678
	TERA	0.2332	0.1052	0.0718
	Mockingjay	0.1812	0.0946	0.0624
	APC	**0.2382**	0.0991	0.0710
	VQ-APC	0.2085	0.0891	0.0658
	Audio ALBERT	0.2184	0.0992	0.0688
	MAE-AST	0.2355	0.1132	**0.0729**
	SS-AST	0.2193	0.1023	0.0673
Contrastive Self-Supervised Methods	Modified CPC	0.2128	0.1019	0.0671
	Wav2Vec	0.2209	0.1044	0.0719
	VQ-Wav2Vec2.0	0.2307	0.1167	0.0754
	Wav2Vec2.0	0.2662	0.1741	0.0861
	Wav2Vec2.0-Large	**0.2676**	**0.1750**	**0.0882**
	Wav2Vec2.0-C	0.2655	0.1740	0.0860
	Discrete BERT	0.2277	0.1065	0.0715
	BYOL-A	0.1302	0.0784	0.0566
	Unispeech	0.2378	0.1356	0.0738
Predictive Self-Supervised Methods	WavLM	0.2356	0.1116	0.0727
	HuBERT	0.2298	0.1088	0.0730
	Data2Vec	**0.2683**	**0.1756**	**0.0886**
	DistilHuBERT	0.2323	0.1101	0.0738
	LightHuBERT	0.2328	0.1102	0.0737

Oota, Subba Reddy, Khushbu Pahwa, Mounika Marreddy, Manish Gupta, and Bapi S. Raju. "Neural architecture of speech." In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5. IEEE, 2023.

Which Speech Tasks are the most Predictive of fMRI Brain Activity?





- 8 tasks from Speech processing Universal PERformance Benchmark (SUPERB): Phoneme Recognition (PR), Automatic Speech Recognition (ASR), Keyword Spotting (KS), Intent Classification (IC), Speaker Diarization (SD), Speaker Verification (SV), Speaker Identification (SID), and Emotion Recognition (ER)

- ASR finetuning (middle layers) yields the best encoding performance for the whole brain, language and auditory regions.
- Finetuning on ER, SID and IC leads to the best alignment for the early auditory cortex.

Oota, Subba Reddy, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Raju Surampudi Bapi. "Speech taskonomy: Which speech tasks are the most predictive of fmri brain activity?." In *INTERSPEECH 2023-24th INTERSPEECH Conference*, pp. 5167-5171. 2023.

Agenda

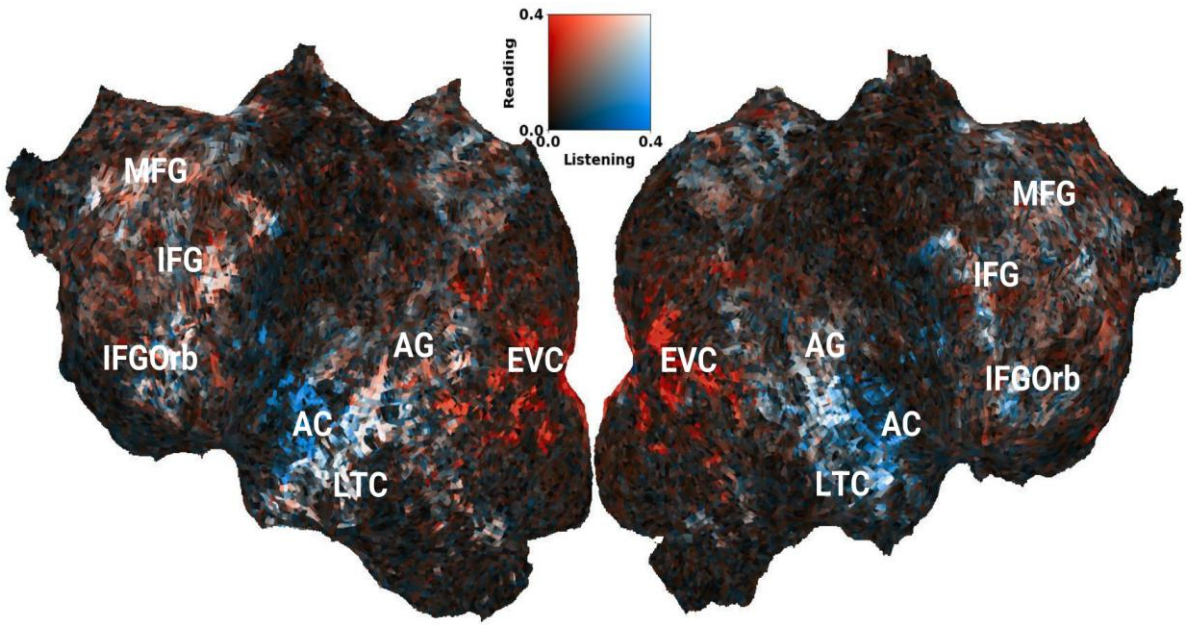
- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- **Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]**
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

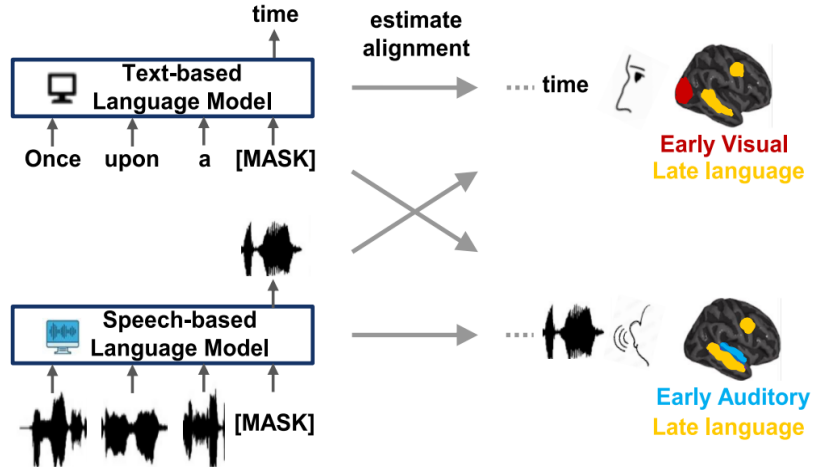
Agenda

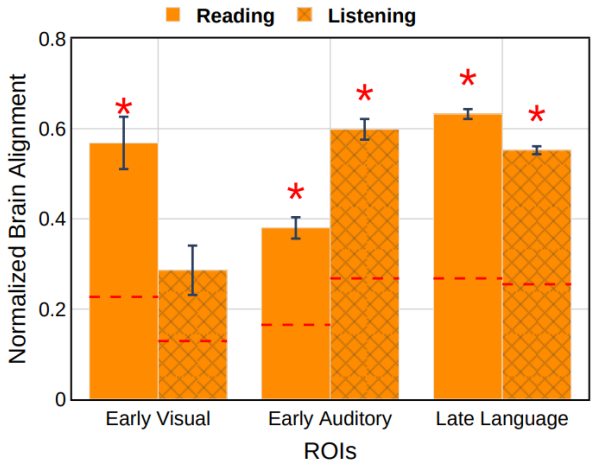
- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- **Brain Encoding [60 min] :**
  - **Text vs. Speech Models**
  - Scaling Laws,
  - Multilinguality,
  - Multimodal and Instruction-tuned Models
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

Text- vs. Speech-based language models : brain alignment

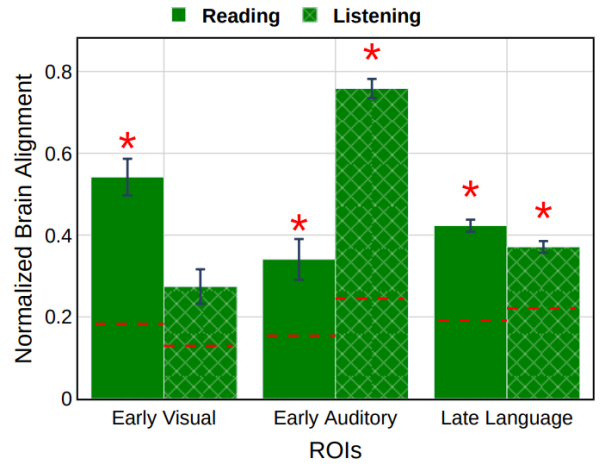
- Stimuli: Subset-Moth-Radio-Hour
- Stimulus representation: pretrained NLP models and speech models
- Brain recording & modality: fMRI, Reading, Listening







(a) Text Models



(b) Speech Models

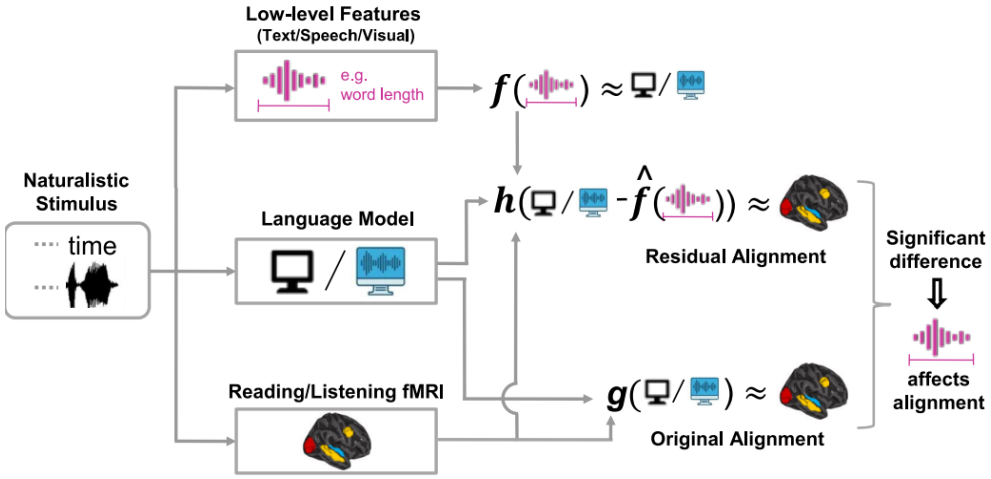
- **Late language regions**: Both types of models show high brain alignment with **late language regions**, but **speech models** trails behind **text models**
- Highly predict early visual and auditory areas.

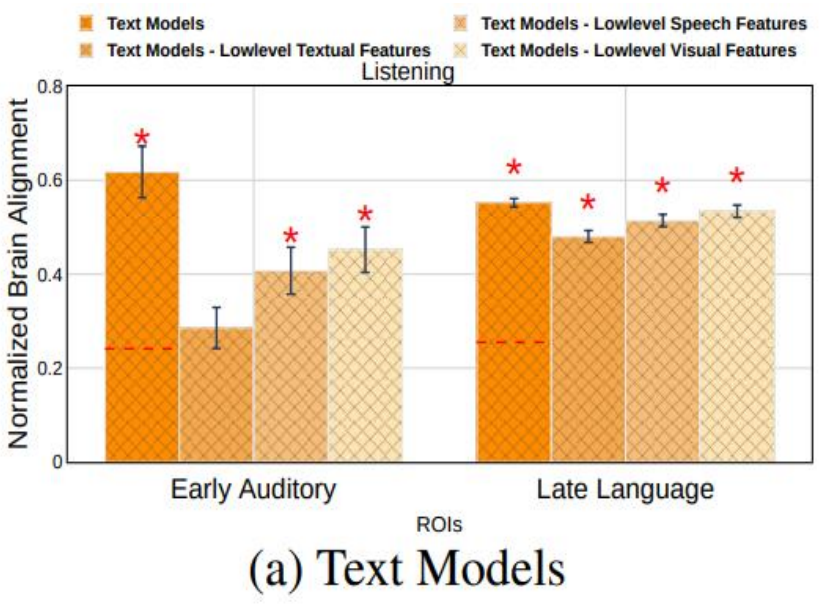
Subba Reddy Oota, Emin Çelik, Fatma Deniz, Mariya Toneva. Speech language models lack important brain-relevant semantics. ACL 2024.

Speech language models lack important brain relevant semantics

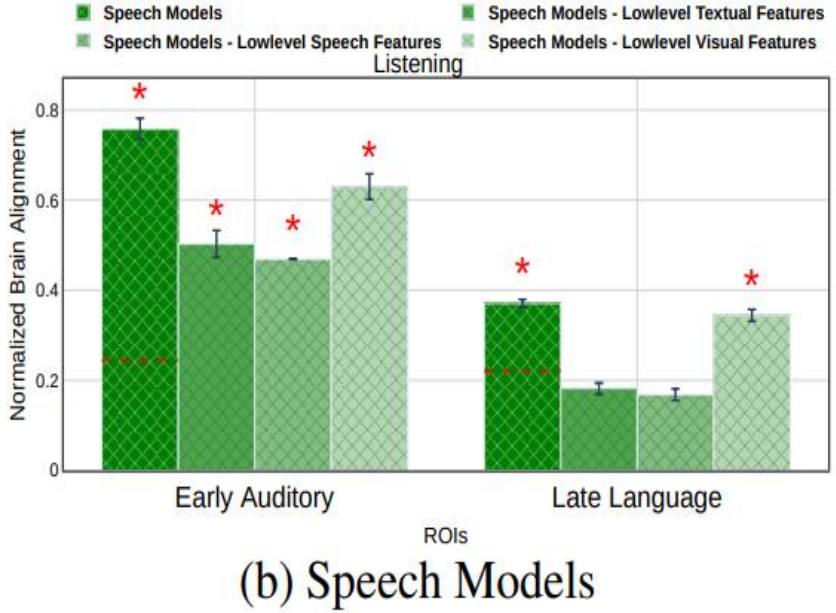
- **Questions:** Why do text-based language models predict early auditory cortices to an impressive degree?
- What types of information do language models truly predict in the Brain?
- How does the type of model (text vs. speech) affect the resulting alignment?

Investigate via a perturbation approach





(a) Text Models

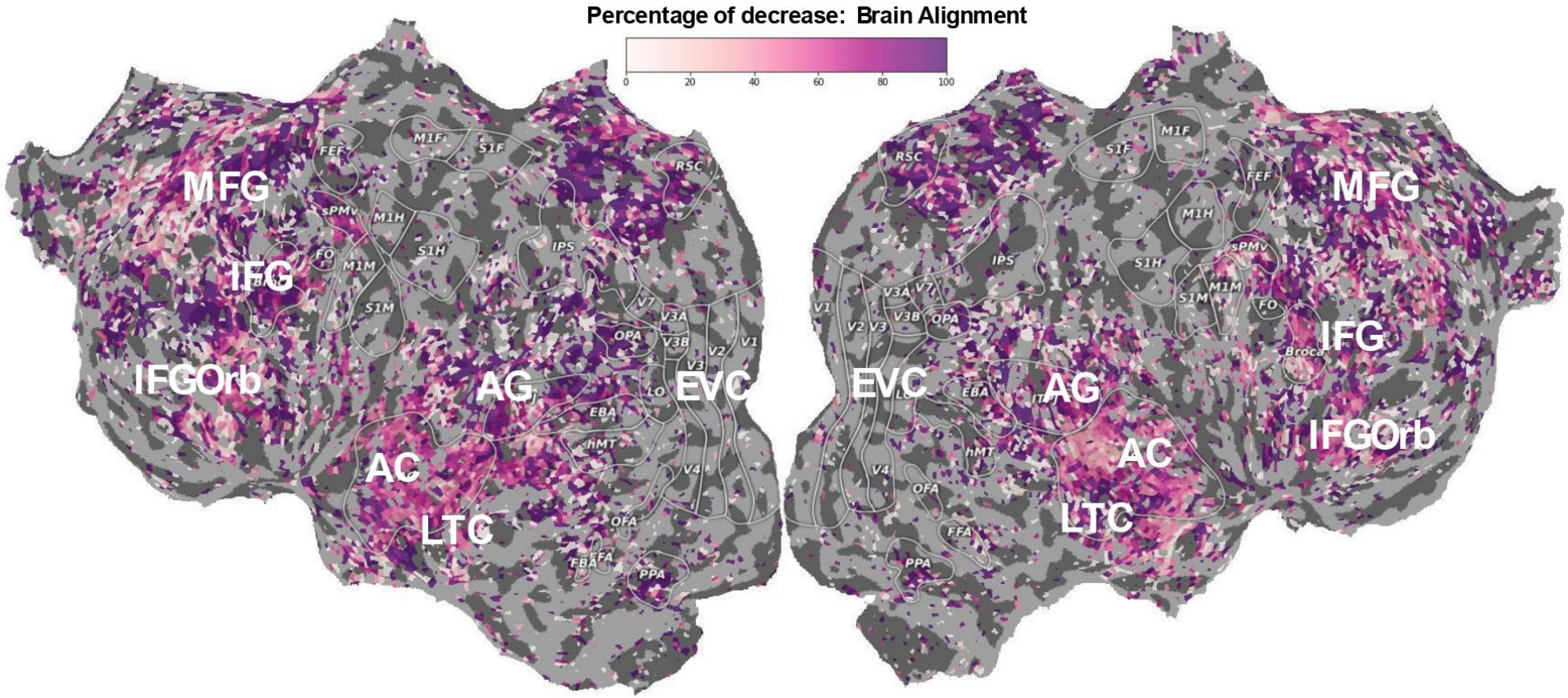


(b) Speech Models

- **Text models:**
  - high alignment in late language regions is not due to **low-level features**
- **Speech models:**
  - alignment in late language regions entirely due to **low-level stimulus features**

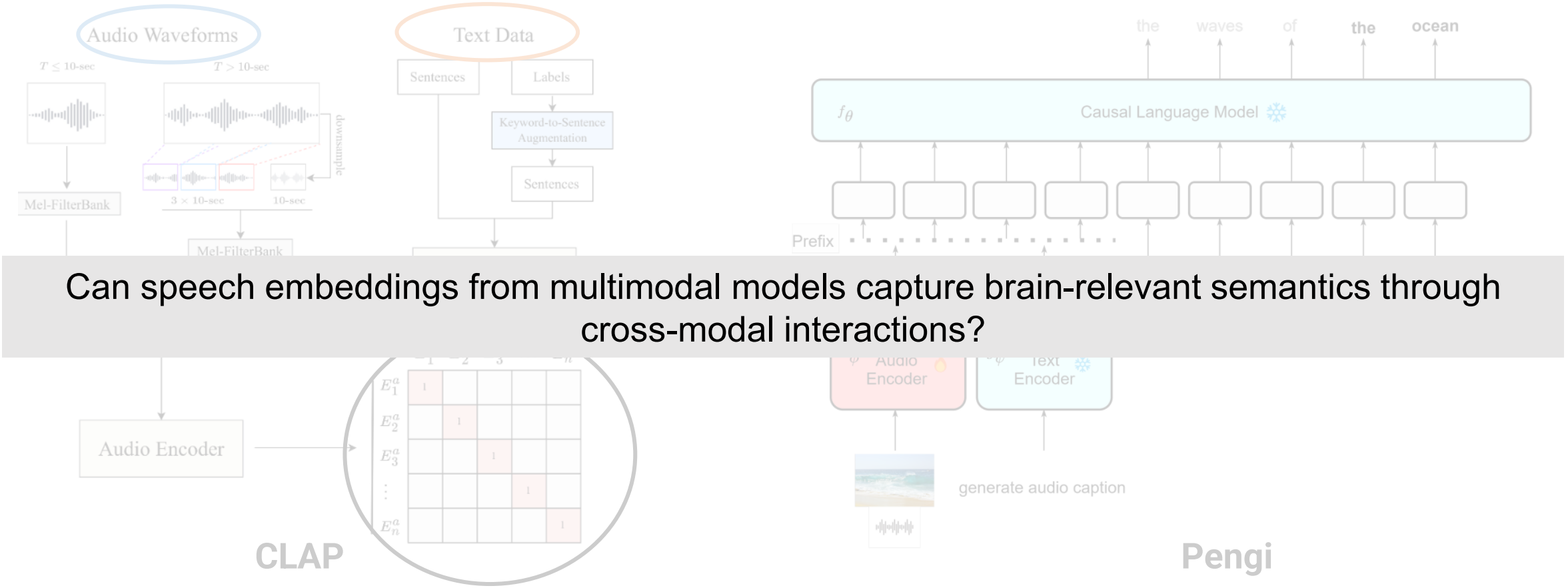
Subba Reddy Oota, Emin Çelik, Fatma Deniz, Mariya Toneva. "Speech language models lack important brain-relevant semantics." ACL 2024.

Phonological properties account for most of the alignment between speech models and the human brain



Subba Reddy Oota, Emin Çelik, Fatma Deniz, Mariya Toneva. "Speech language models lack important brain-relevant semantics." ACL 2024.

Multimodal (Text+Speech) models enables learning audio concepts from natural language supervision

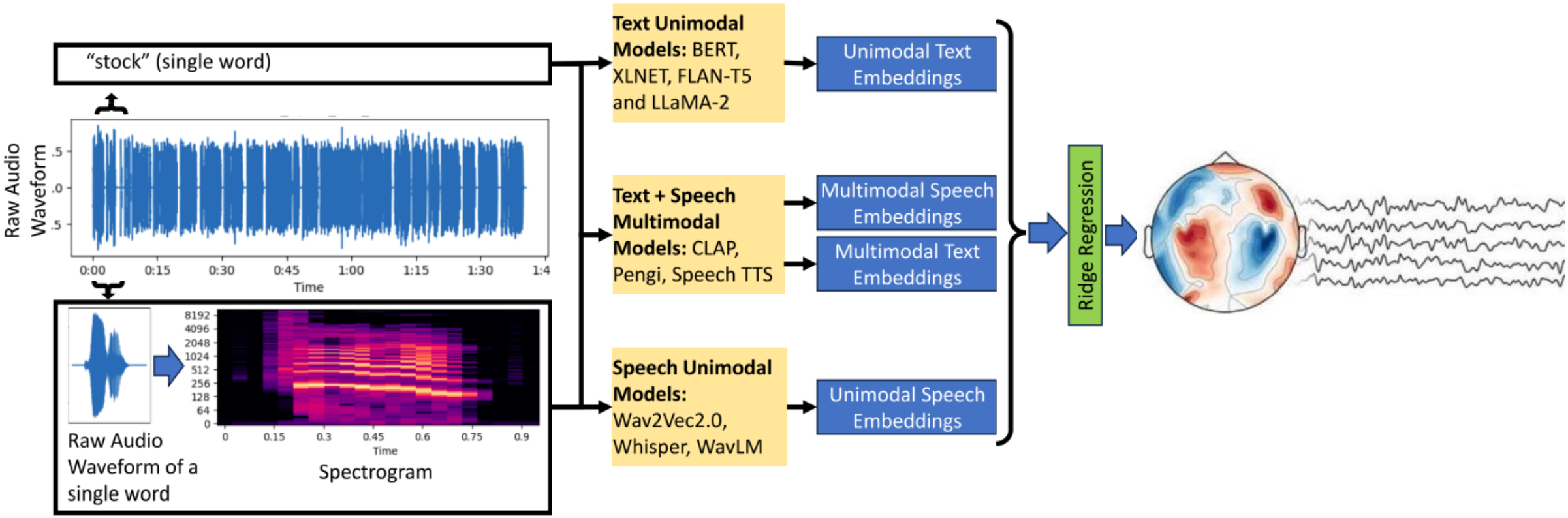


Can speech embeddings from multimodal models capture brain-relevant semantics through cross-modal interactions?

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, Huaming Wang. "CLAP : LEARNING AUDIO CONCEPTS FROM NATURAL LANGUAGE SUPERVISION." Arxiv 2022.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, Huaming Wang. "Pengi: An Audio Language Model for Audio Tasks." NeurIPS 2023.

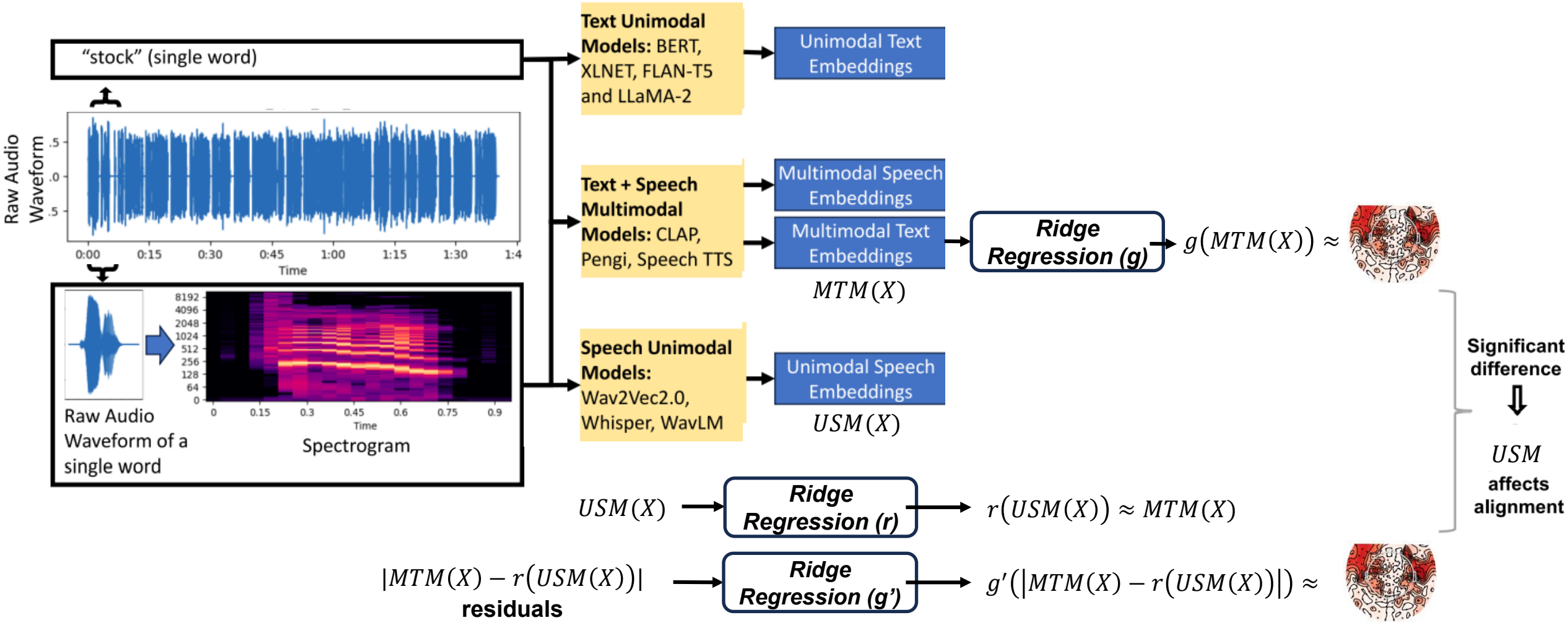
Multi-modal vs. Unimodal models: Brain alignment



- How well do text/speech embeddings from multi-modal models predict speech-evoked brain activity over unimodal models?
- Is there asymmetric knowledge transfer across modalities in multimodal models, or do multimodal-text and multimodal-speech perform equally well?

Which modality of representations in multi-modal models leads to high brain alignment?

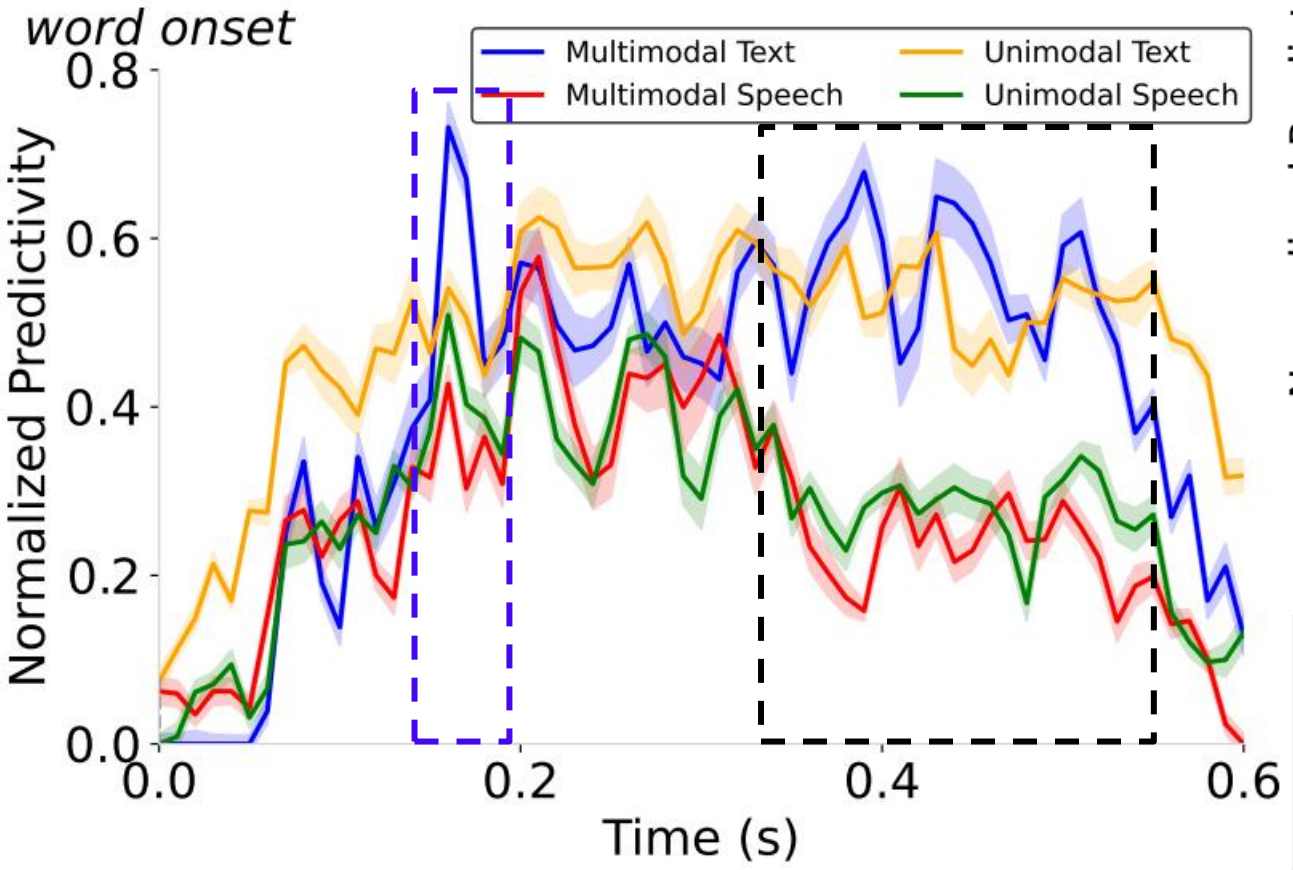
Investigate via a residual approach

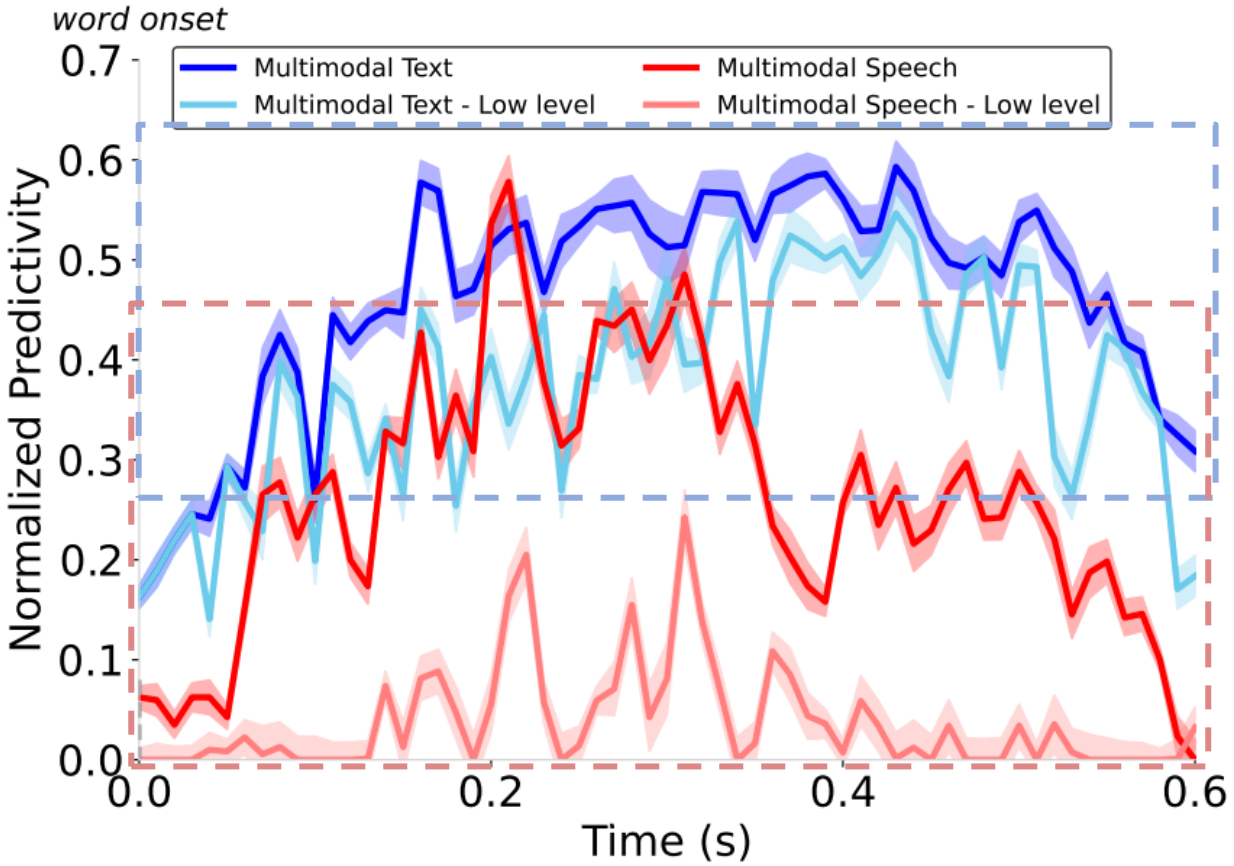


Toneva et al. 2022 Nature Computational Science, Oota et al. 2023 NeurIPS

Multimodal speech embeddings still lack important brain relevant semantics

- Stimuli: Naturalistic stories
- Stimulus representation: pretrained NLP models, speech and multimodal text-speech models
- Brain recording & modality: MEG, Listening





- **Multimodal Text Embeddings**:
  - high brain alignment due to **brain-relevant semantics**
- **Multimodal Speech Embeddings:**
  - brain alignment mostly due to **low-level speech features**

Padakanti Srijith, Khushbu Pahwa, Radhika Mamidi, Bapi Raju Surampudi, Manish Gupta, Subba Reddy Oota. "Aligning Text/Speech Representations from Multimodal Models with MEG Brain Activity During Listening." EMNLP 2025.

Conclusions for neuro-AI research field

1. **Speech models (🔊)** useful for modeling **early listening (🎧)**: investigate speech models to learn more about AC

2. **Text models (🔤)** useful for modeling **language processing** in both 🎧 and 📖

3. **Multimodal Speech models** are useful for modeling **early auditory** and **lexical word processing.** Need to investigate speech models to learn more about word processing

4. **Multimodal Text models** are useful for modeling both **early auditory** and **high-level semantic information processing**

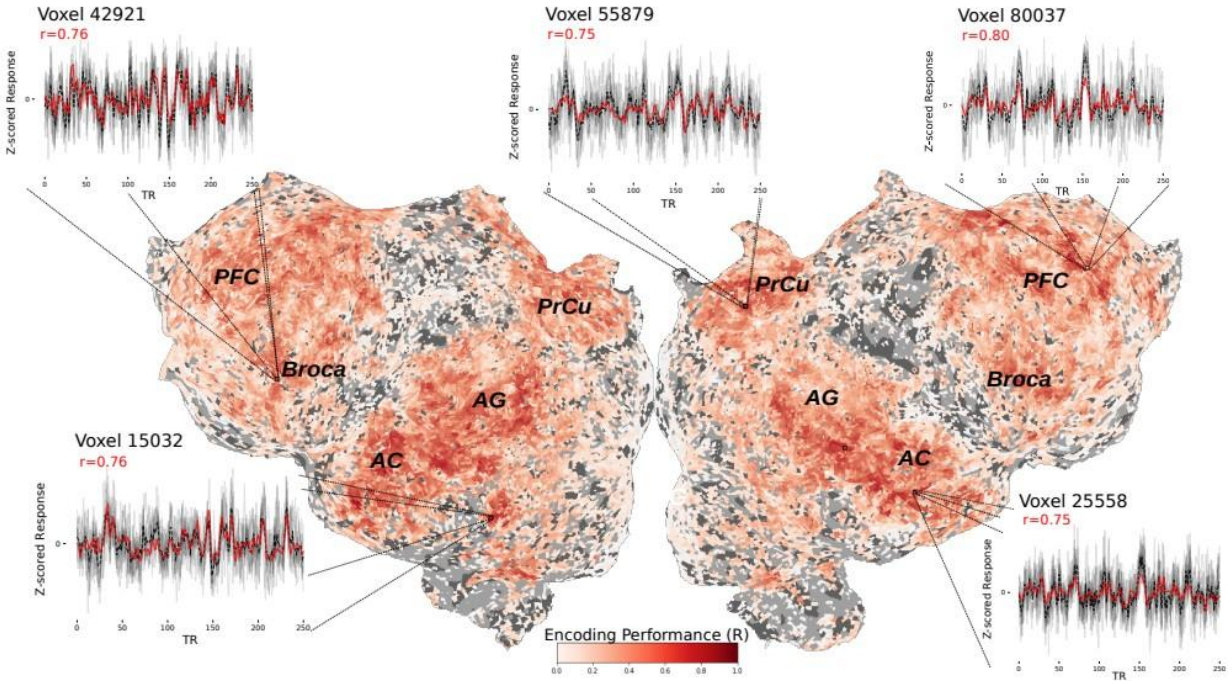
5. More work to do for a complete **end-to-end multimodal model** capable of bi-directional cross-knowledge transfer between text and speech

Agenda

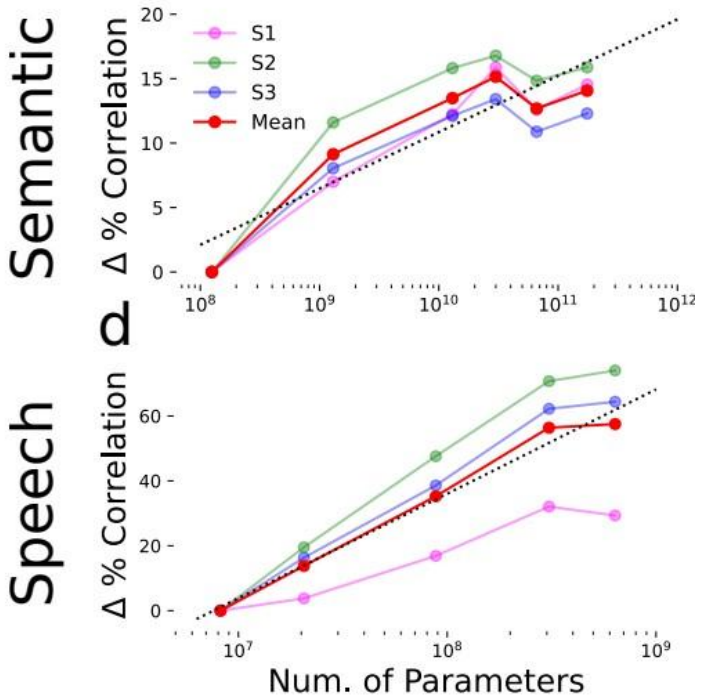
- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- **Brain Encoding [60 min] :**
  - Text vs. Speech Models
  - **Scaling Laws,**
  - Multilinguality,
  - Multimodal and Instruction-tuned Models
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

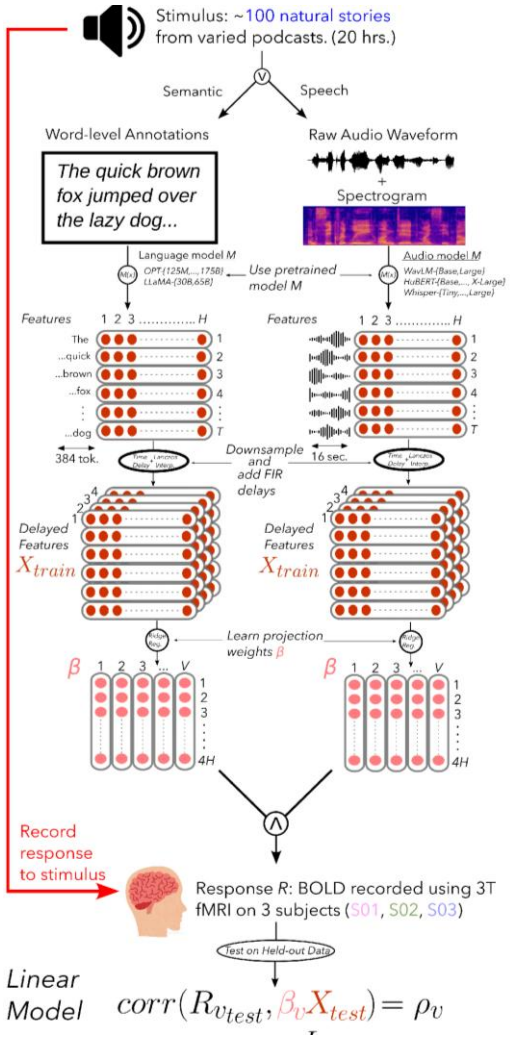
Brain prediction performance scales logarithmically with model size

- Stimuli: Moth-Radio-Hour (English)
- Stimulus representation: OPT (125M to 175B) and LLaMA (33B, 66B)
- Audio models: HuBERT (95M to 964M), Whisper (8M to 637M), WavLM (95M, 317M)
- Brain recording & modality: fMRI, Listening



OPT Model



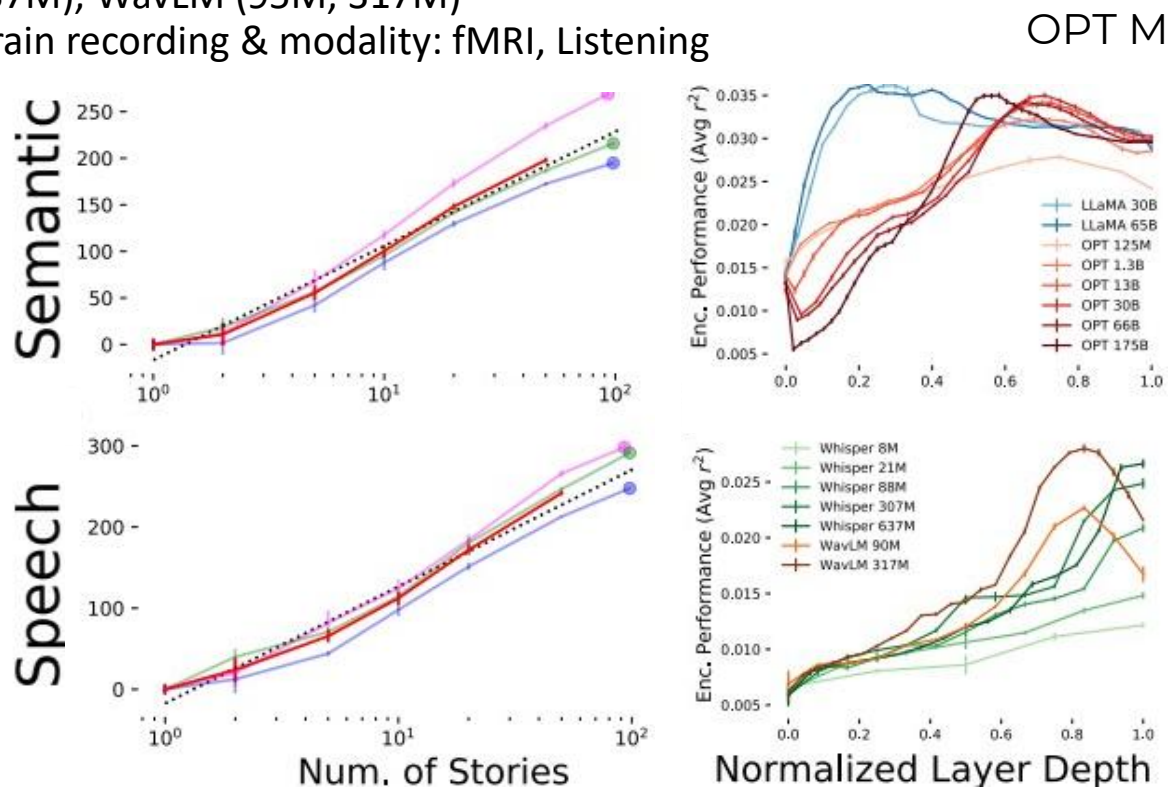


- For each order of magnitude increase in the number of parameters in the language, the encoding performance of the average subject increases by roughly 4.4%

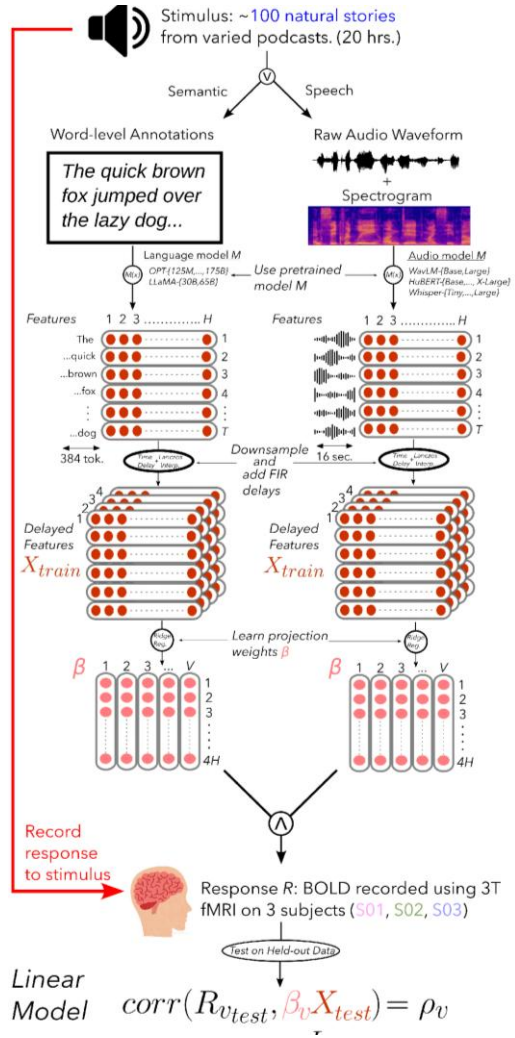
Richard Antonello, Aditya R. Vaidya, Alexander Huth. "Scaling laws for language encoding models in fMRI" 2023 NeurIPS.

Brain prediction performance scales logarithmically with the size of the fMRI training set

- Stimuli: Moth-Radio-Hour (English)
- Stimulus representation: OPT (125M to 175B) and LLaMA (33B, 66B)
- Audio models: HuBERT (95M to 964M), Whisper (8M to 637M), WavLM (95M, 317M)
- Brain recording & modality: fMRI, Listening



- 10× more stories -> +122% encoding (avg subject)
- LLaMA slightly > OPT (encoding)
- LLaMA peaks early, then slowly declines
- OPT peaks late (~75% depth)



- Higher encoding performance leads to more trustworthy model predictions and more accurate conclusions

Richard Antonello, Aditya R. Vaidya, Alexander Huth. "Scaling laws for language encoding models in fMRI" 2023 NeurIPS.

Larger LLMs better capture the structure of natural language and better predict neural activity

- Stimuli: Audio podcast (English)
- Stimulus representation: GPT2 (82M to 1.5B), GPT-Neo (125M to 20B), OPT (125M to 66B), and LLaMA-2 (7B to 70B)
- Brain recording & modality: ECoG, Listening

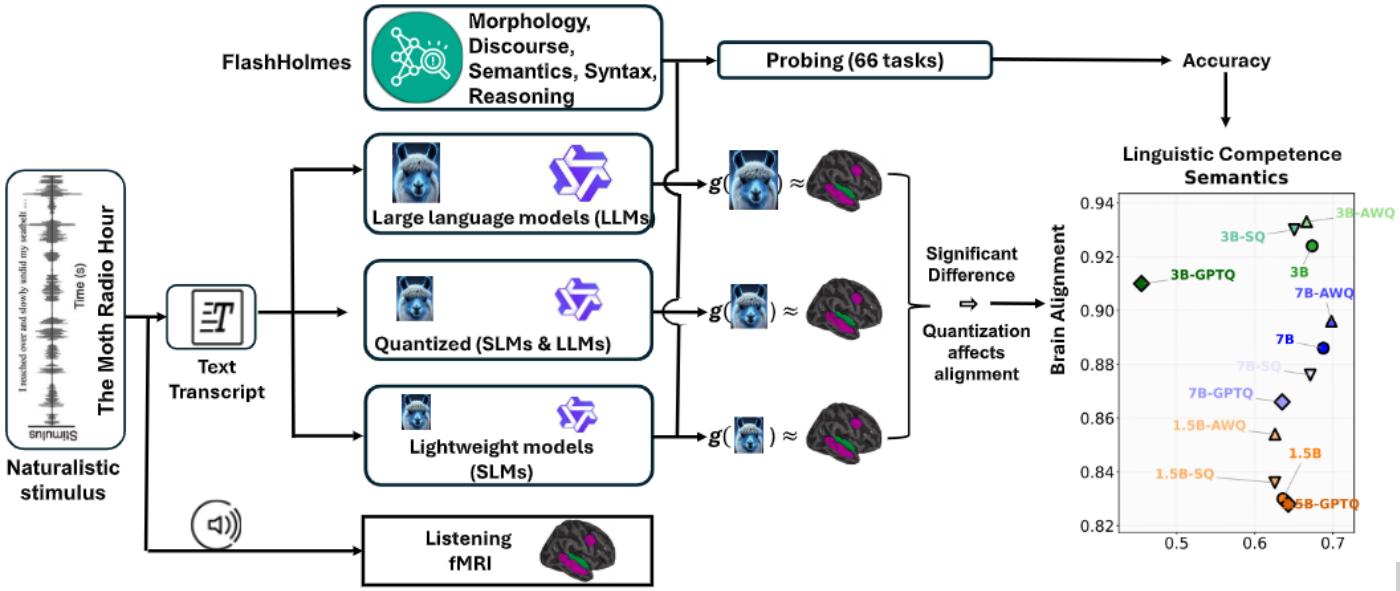
If encoding scales with size/data, does interpretability scale too?

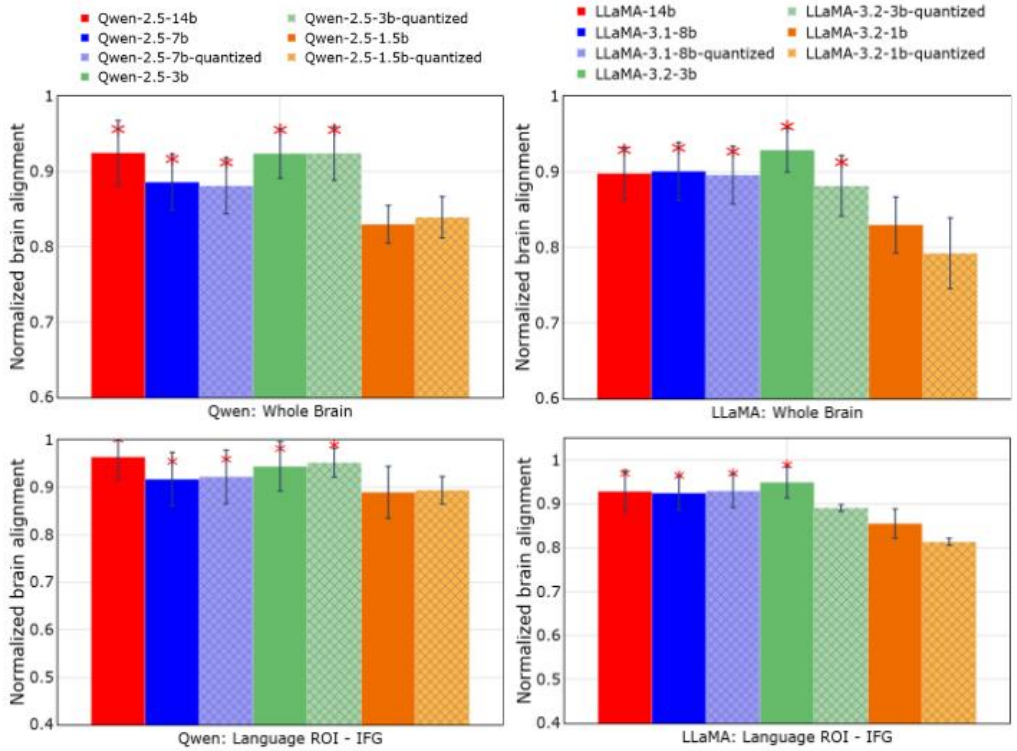
- A log-linear relationship where the encoding performance peaks in relatively earlier layers as model size increases

Zhuoqiao Hong, Haocheng Wang, Zaid Zada, Harshvardhan Gazula, David Turner, Bobbi Aubrey, Leonard Niekerken, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Orrin Devinsky, Adeen Flinker, Uri Hasson, Samuel A Nastase, Ariel Goldstein. "Scale matters: Large language models with billions (rather than millions) of parameters" 2025 BioArxiv.

SLMs achieve brain alignment comparable to their larger counterparts

- Stimuli: Subset-Moth-Radio-Hour
- Stimulus representation: Qwen2.5 (1.5B to 14B), LLaMA-3 (1B to 14B), DeepSeek-R1 (1B to 14B)
- Brain recording & modality: fMRI, Listening





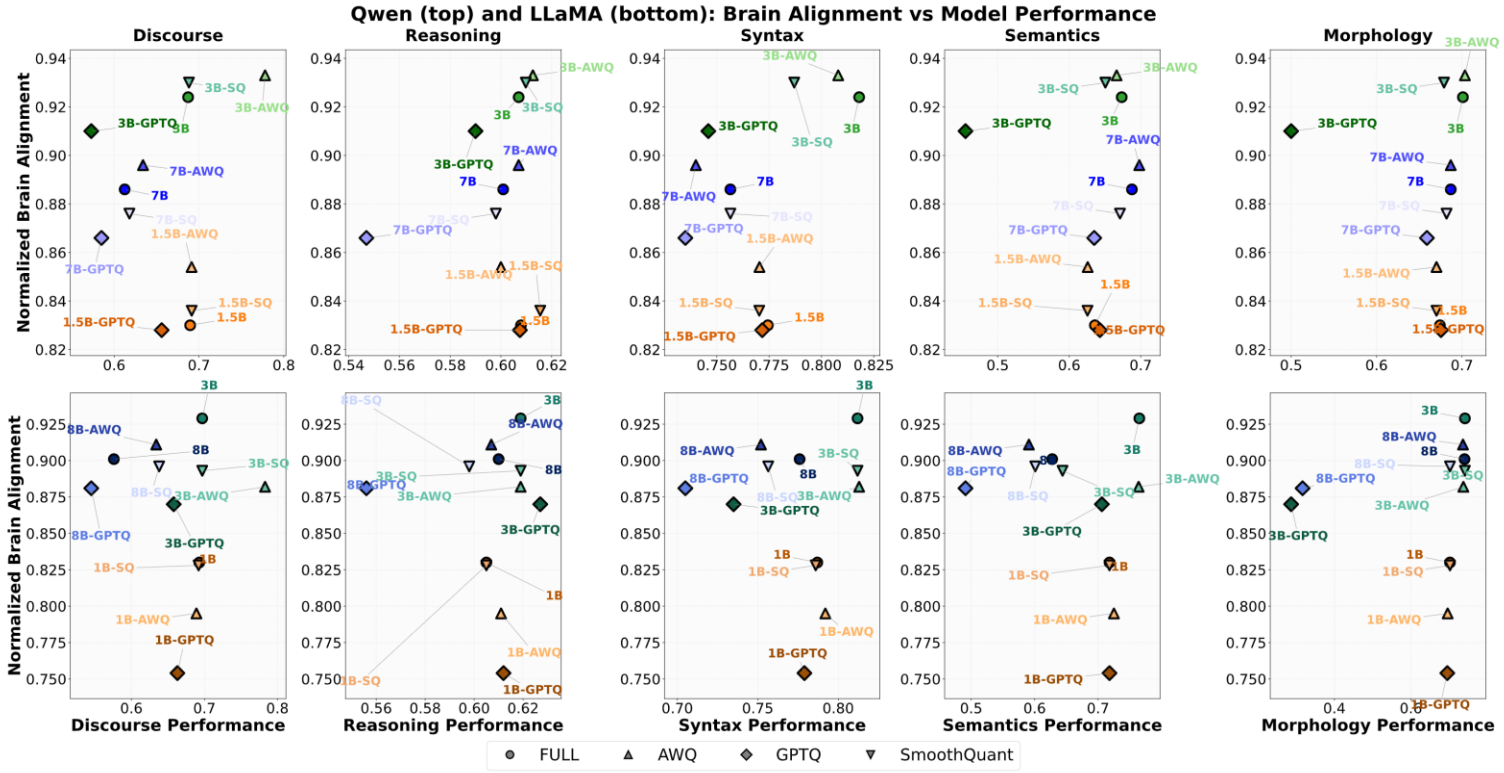
- 3B SLMs preserve brain alignment comparable to 7B and 14B LLMs, while 1B-1.5B models consistently show lower brain alignment

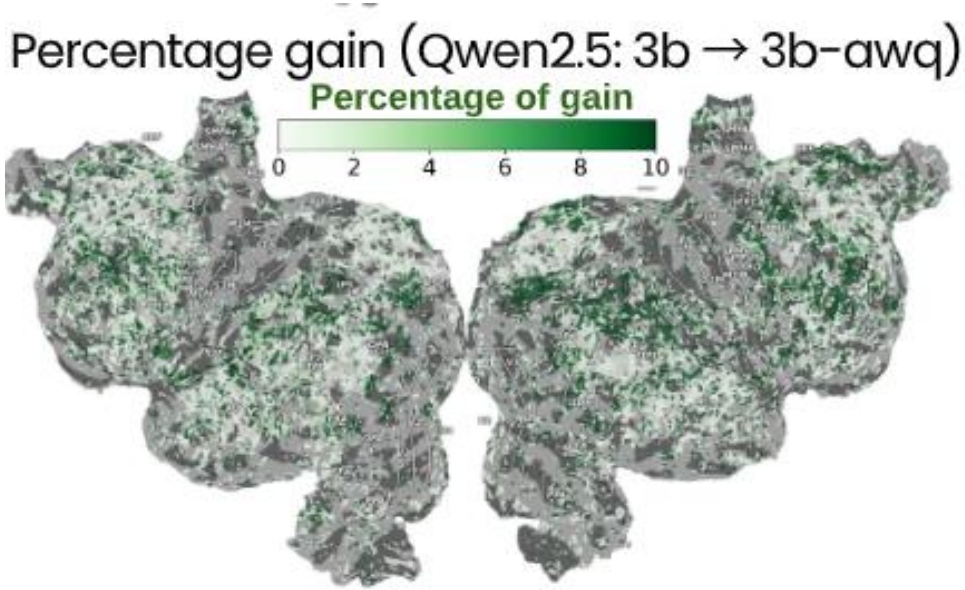
- Which linguistic properties are maintained or lost across small, large, and compressed models

Oota et al. 2025. "Linguistic Properties and Model Scale in Brain Encoding: From Small to Compressed Language Models" 2025 OpenReview.

Behavioral competence alone does not ensure brain-like representations

- Stimuli: Subset-Moth-Radio-Hour
- Stimulus representation: Qwen2.5 (1.5B to 14B), LLaMA-3 (1B to 14B), DeepSeek-R1 (1B to 14B)
- Brain recording & modality: fMRI, Listening
- Linguistic competence: FlashHolmes benchmark (200+ tasks)





- 1B SLMs maintain FlashHolmes performance yet show significant drop in brain alignment
- GPTQ selectively degrades discourse, syntax, and reasoning
- AWQ preserves ~98% of alignment while maintaining or improving linguistic competence

3B models match similar brain alignment as 7B–14B models, while 1B–1.5B models show a significant drop in brain alignment, supporting a local capacity threshold around ~3B.

Oota et al. 2025. "Linguistic Properties and Model Scale in Brain Encoding: From Small to Compressed Language Models" 2025 OpenReview.

Scaling helps, but NeuroAI needs the right capacity, compression, and layers

1. **Practical modeling:** for brain-encoding pipelines, prioritize mid-size (~3B, 7B) + right layer selection over blindly scaling parameters

2. **Where the brain signal lives differs by family:** different LLM families peak at different layer depths, implying architecture/training choices shape brain similarity.

3. **Brain encoding scales predictably with model size** (roughly log-linear / diminishing returns).

4. **Behavioral competence ≠ brain-likeness:** high benchmark performance alone does not guarantee brain-like representations.

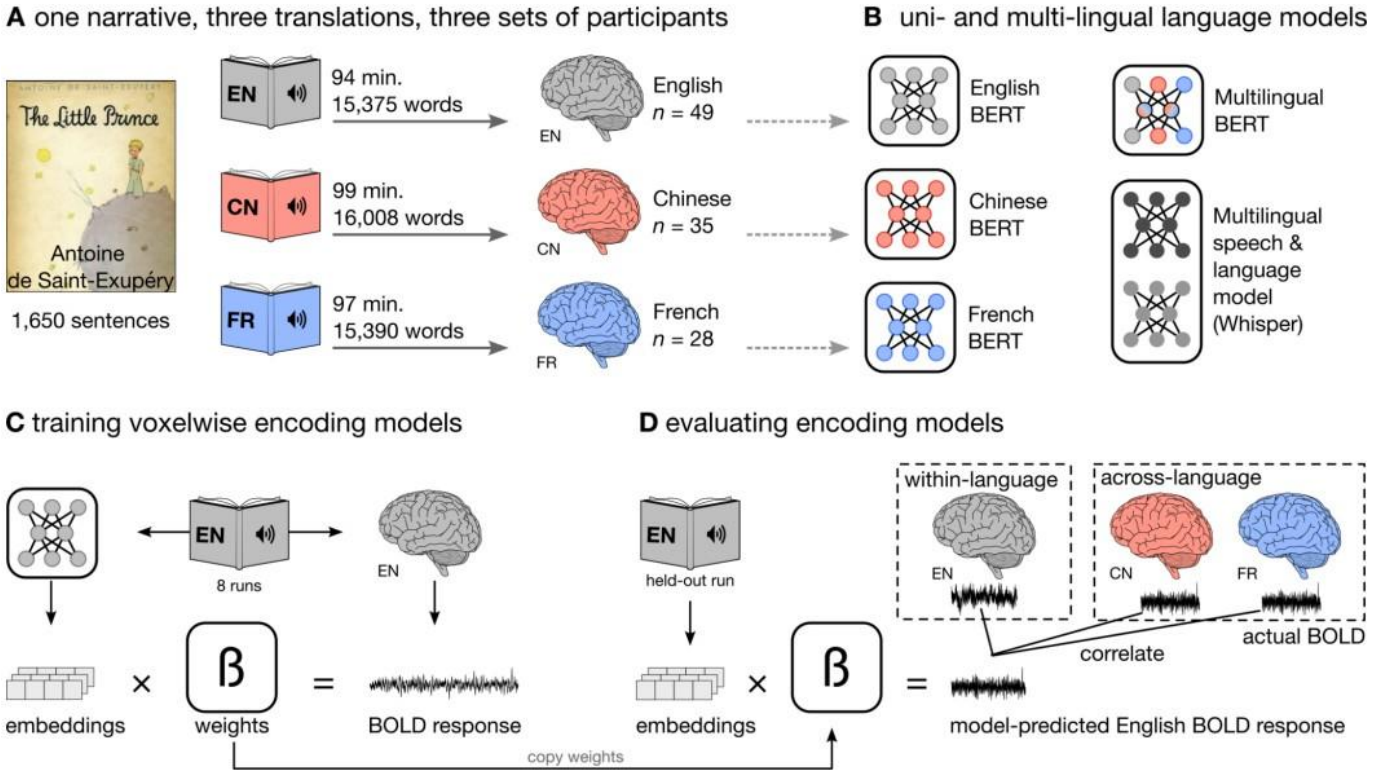
5. **Model selection principle:** still open questions on models like diffusion LLMs, and Mamba-like architectures.

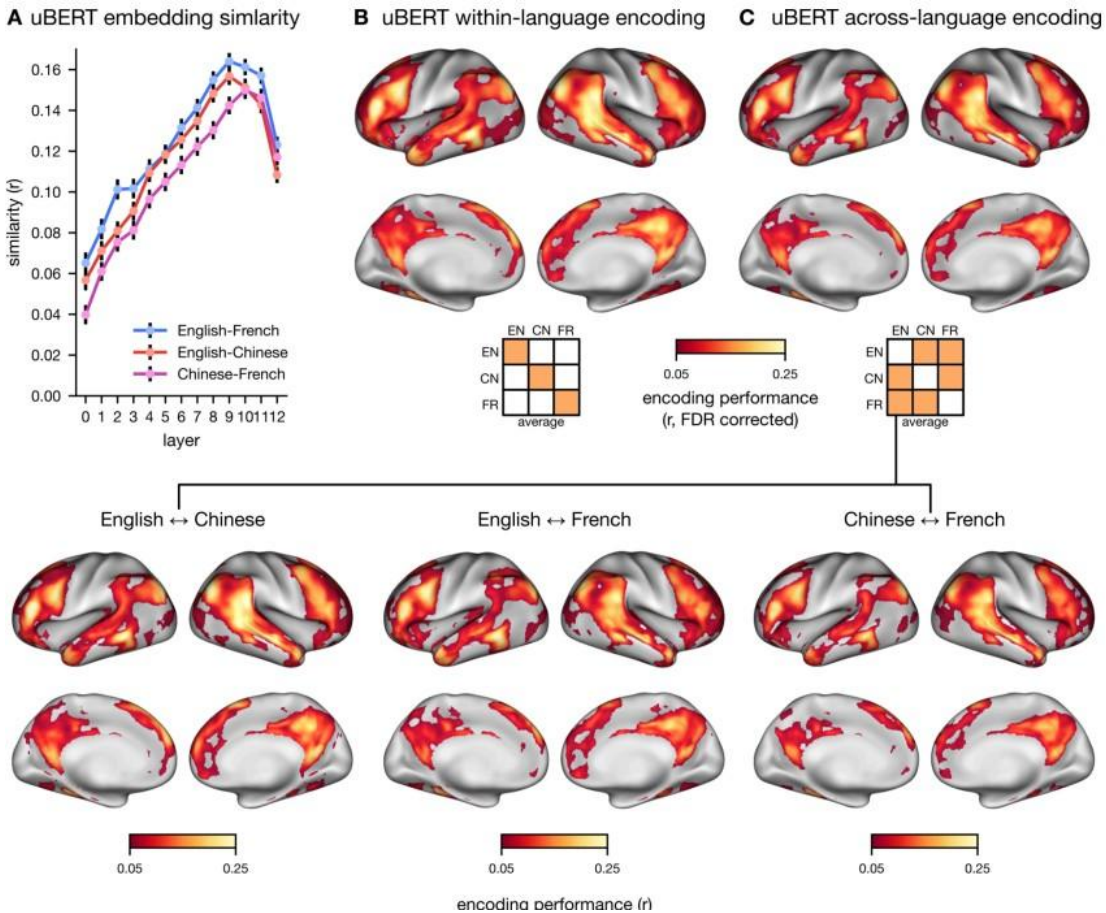
Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- **Brain Encoding [60 min] :**
  - Scaling Laws,
  - **Multilinguality,**
  - Multimodal and Instruction-tuned Models
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

Unilingual models and brains converge on shared linguistic representations

- Stimuli: Little Prince Narrative (3 languages)
- Stimulus representation: multilingual language models
- Brain recording & modality: fMRI, Listening



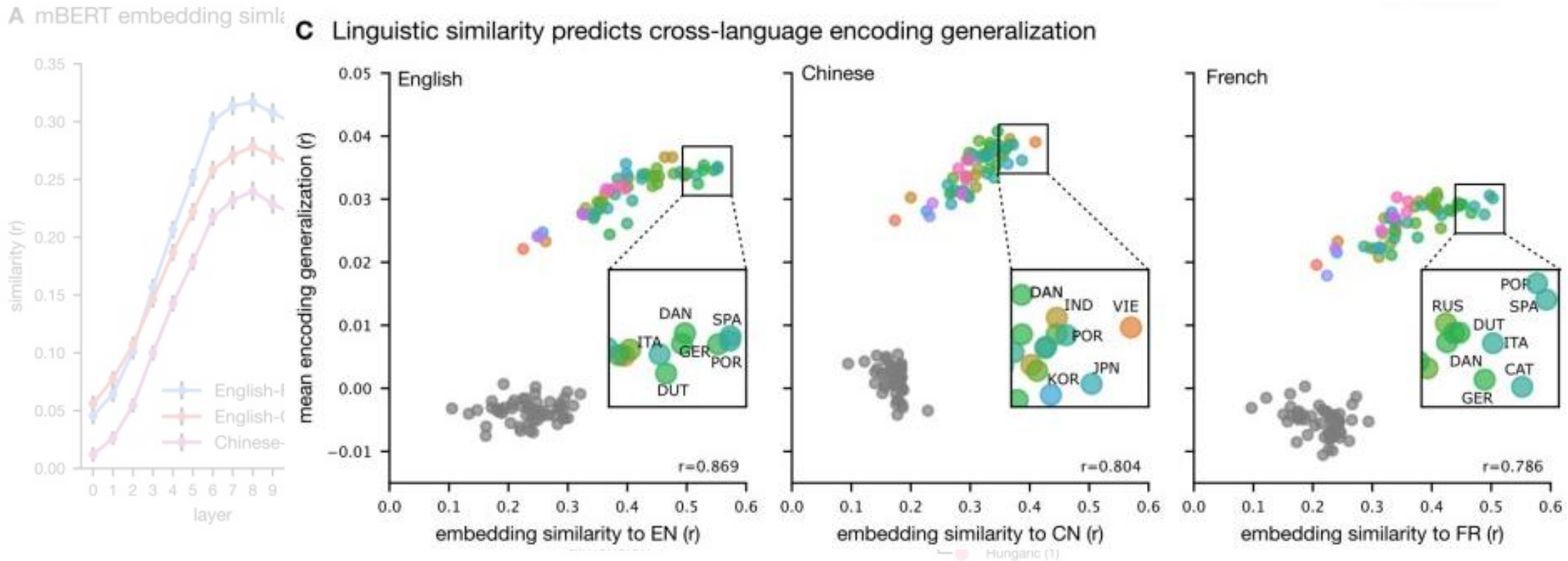


- Shared content across different languages drives similarity between models and brains

Zaid Zada, Samuel A. Nastase, Jixing Li, Uri Hasson. "Brains and language models converge on a shared conceptual space across different languages" 2025 BioarXiv.

Language family structure reflected in multilingual language model embeddings

- Stimuli: Little Prince Narrative (3 languages)
- Stimulus representation: multilingual language models
- Brain recording & modality: fMRI, Listening

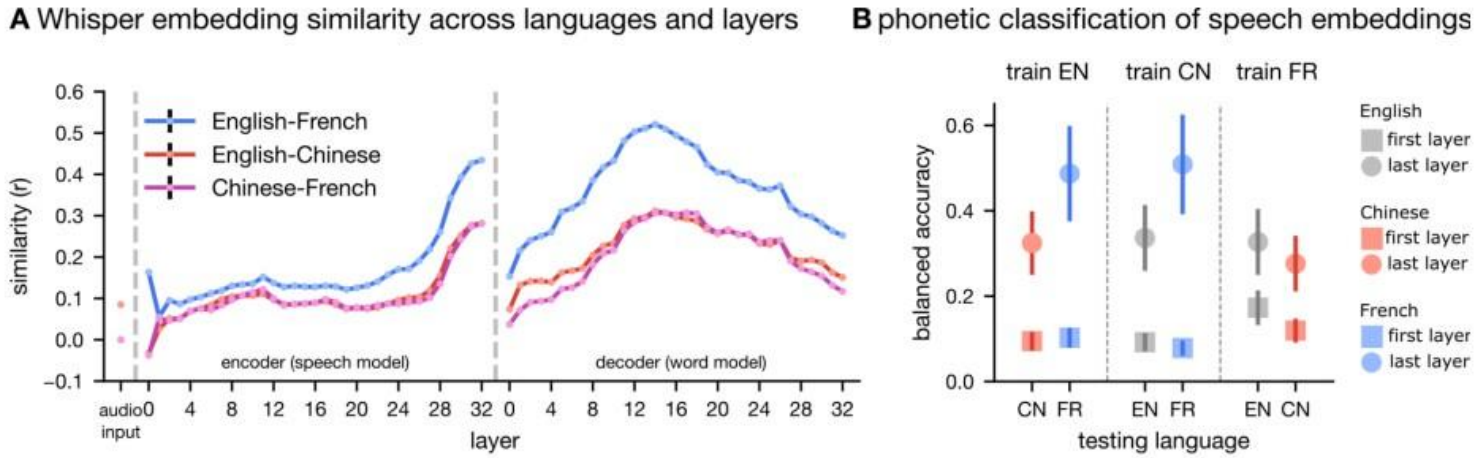


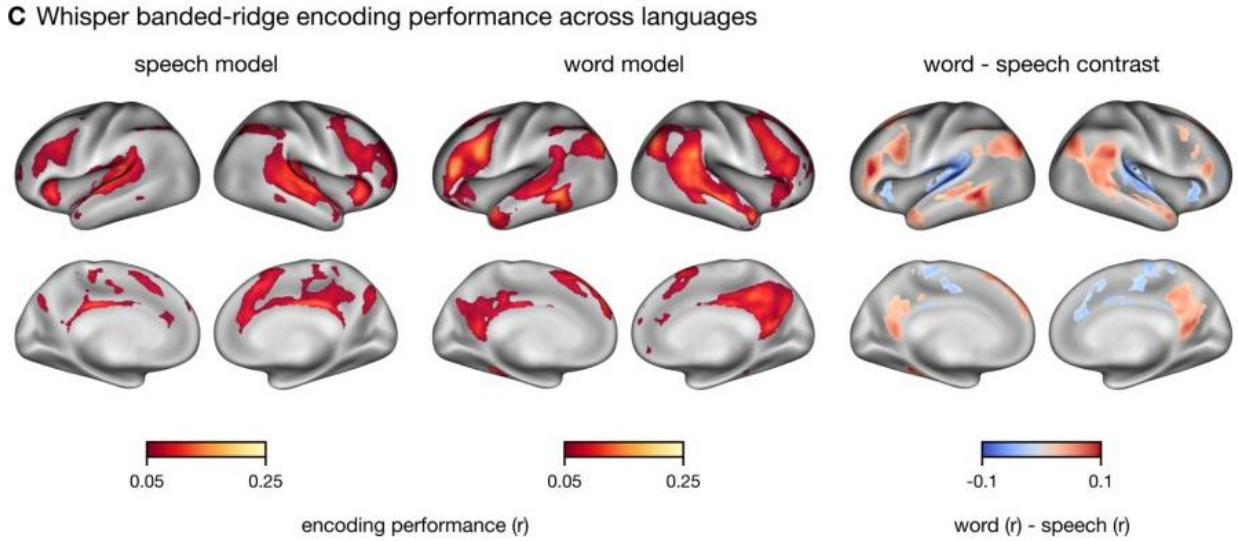
- Language models trained on multiple languages leverage conceptual similarities across languages and learn features that are shared across multiple languages

Zaid Zada, Samuel A. Nastase, Jixing Li, Uri Hasson. "Brains and language models converge on a shared conceptual space across different languages" 2025 BioarXiv.

Shared multilingual speech features in Whisper and across brains

- Stimuli: Little Prince Narrative (3 languages)
- Stimulus representation: multilingual language models
- Brain recording & modality: fMRI, Listening



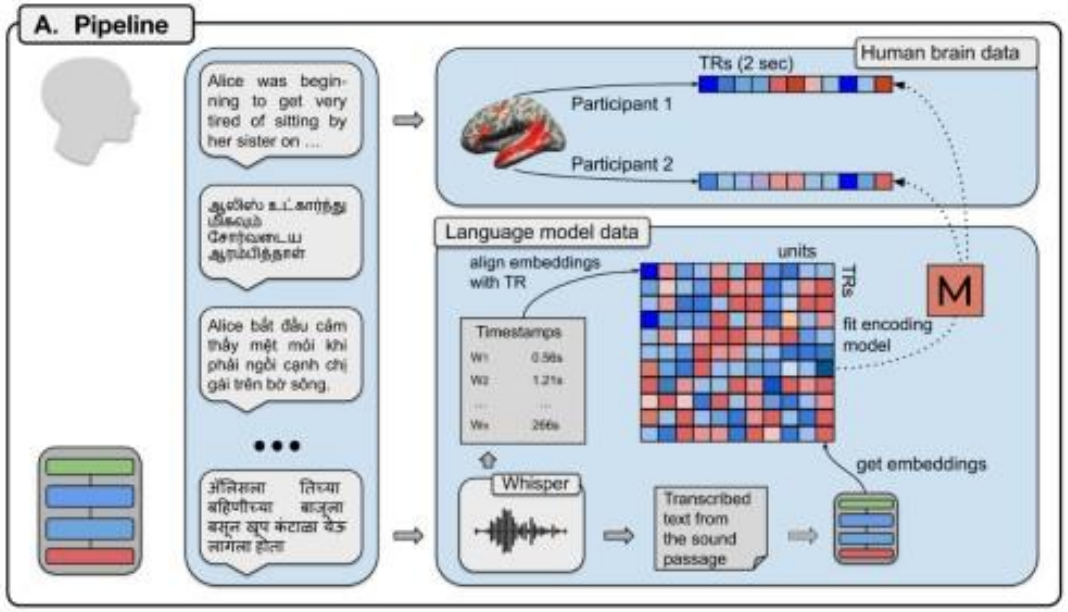


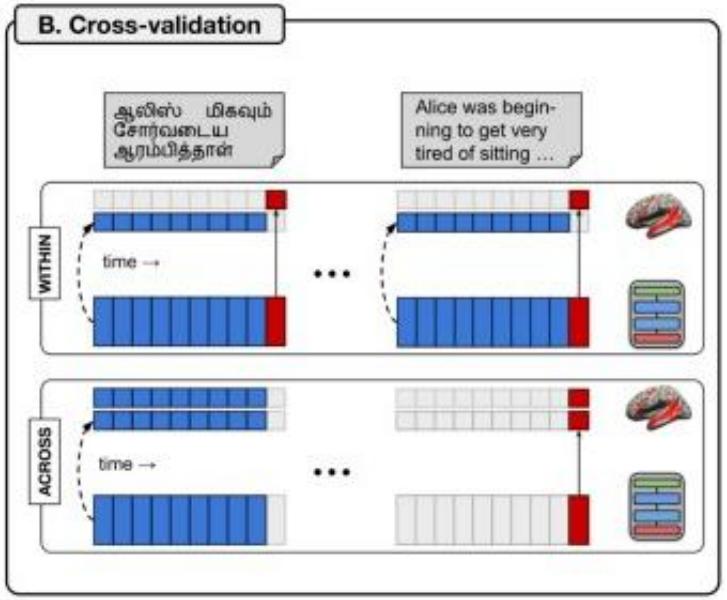
- Shared meaning across languages is not limited to abstract conceptual representations but that lower-level speech features are also shared

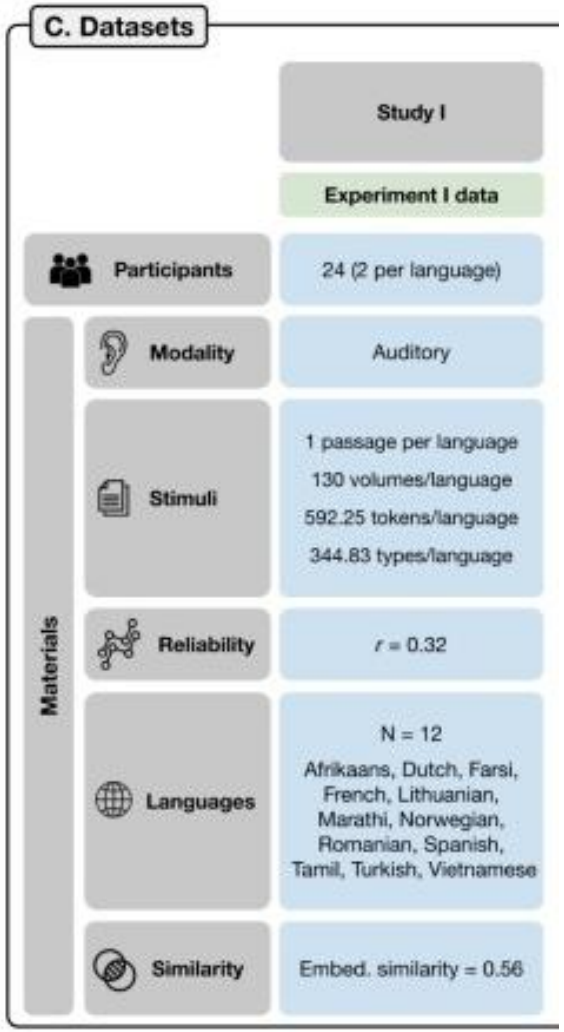
Zaid Zada, Samuel A. Nastase, Jixing Li, Uri Hasson. "Brains and language models converge on a shared conceptual space across different languages" 2025 BioarXiv.

Encoding models can be transferred zero-shot across languages

- Stimuli: Passage from book (21 languages: 12 for study-1 and 9 for study-2)
- Stimulus representation: multilingual language models
- Brain recording & modality: fMRI, Listening





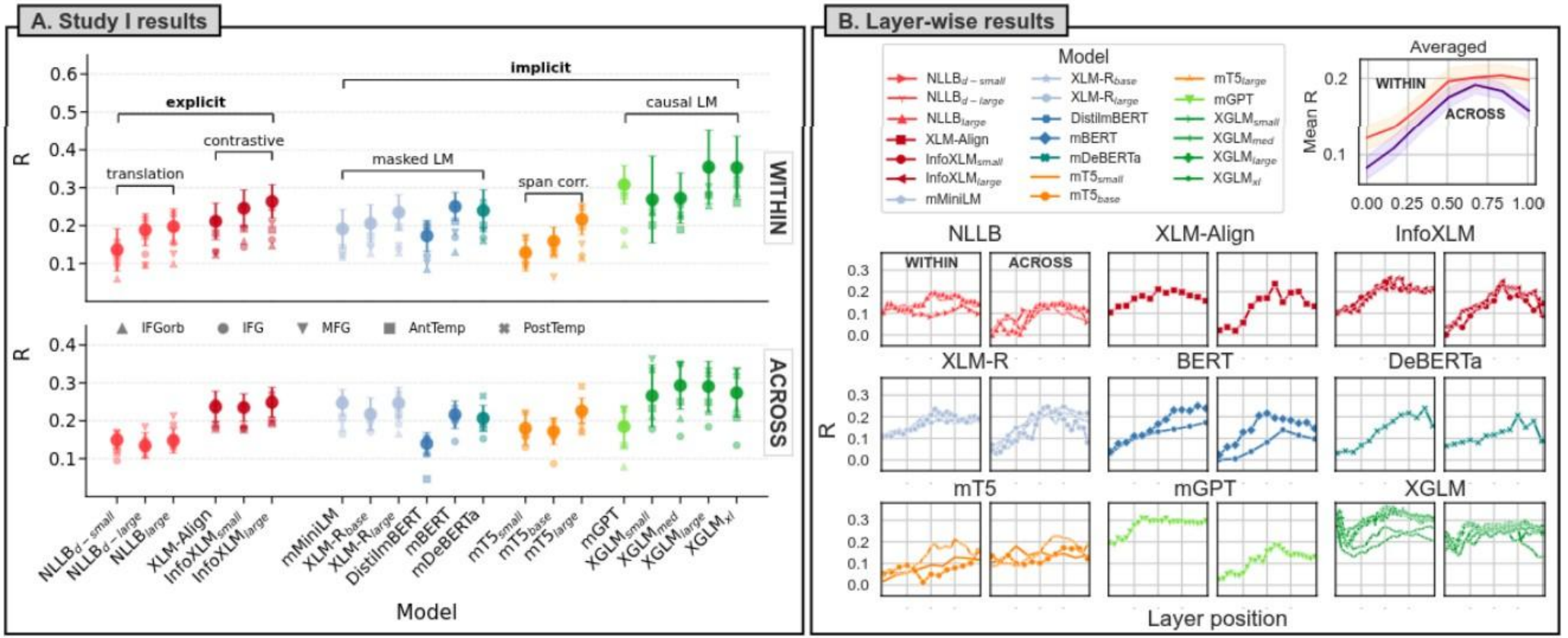


- Semantic representations are largely shared across languages

Andrea Gregor de Varda1,2, Saima Malik-Moraleda2Andrea Gregor de Varda, Saima Malik-Moraleda, Greta Tuckute, Evelina Fedorenko. "Multilingual Computational Models Capture a Shared Meaning Component in Brain Responses across 21 Languages" 2025 BioarXiv.

Intermediate-to-deep layers show higher levels of predictivity

- Stimuli: Passage from book (21 languages)
- Stimulus representation: multilingual language models
- Brain recording & modality: fMRI, Listening

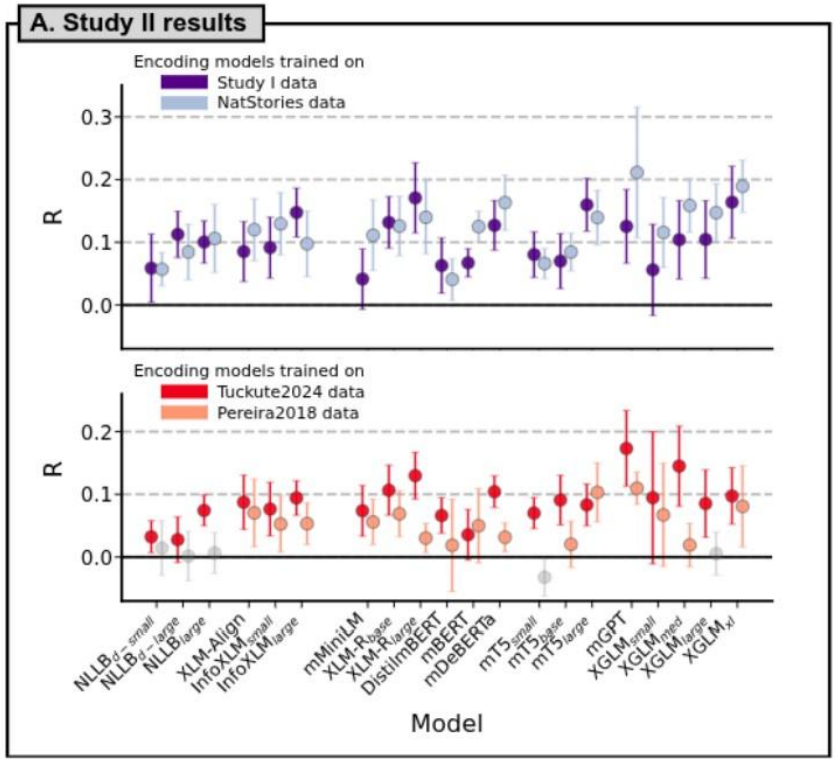


- Encoding models can be successfully transferred zero-shot to unseen languages

Andrea Gregor de Varda1,2, Saima Malik-Moraleda2Andrea Gregor de Varda, Saima Malik-Moraleda, Greta Tuckute, Evelina Fedorenko. "Multilingual Computational Models Capture a Shared Meaning Component in Brain Responses across 21 Languages" 2025 BioarXiv.

Encoding models can be transferred to new languages across types of language materials and presentation modalities

- Stimuli: Passage from book (21 languages)
- Stimulus representation: multilingual language models
- Brain recording & modality: fMRI, Listening



C. Datasets

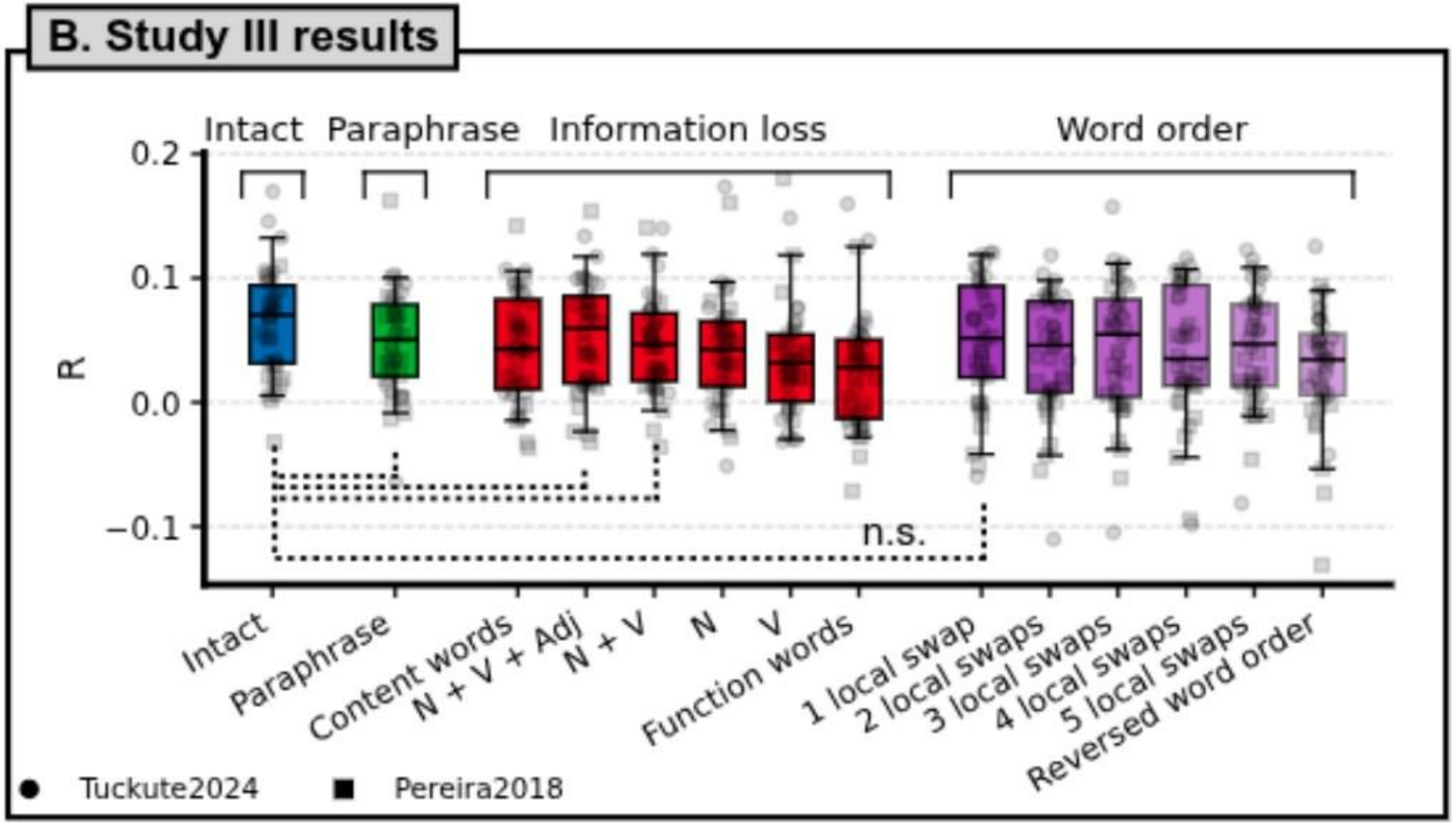
		Study II				
		Train				Test
		Experiment I data	NatStories	Pereira2018	Tuckute2024	Experiment II data
	Participants	24 (2 per language)	97 (28.44 per story)	10 (7.84 per sentence)	10 (10 per sentence)	27 (3 per language)
Materials	Modality	Auditory	Auditory	Visual	Visual	Auditory
	Stimuli	1 passage per language 130 volumes/language 592.25 tokens/language 344.83 types/language	9 passages 1,516 volumes 10,206 tokens 3,037 types	627 sentences 627 volumes 7,711 tokens 2,968 types	1,000 sentences 1,000 volumes 6,000 tokens 2,894 types	3 passages per language 390 volumes/language 1,751.11 tokens/lang. 851.78 types/language
	Reliability	$r = 0.32$	$r = 0.59$	$r = 0.35$	$r = 0.30$	$r = 0.30$
	Languages	N = 12 Afrikaans, Dutch, Farsi, French, Lithuanian, Marathi, Norwegian, Romanian, Spanish, Tamil, Turkish, Vietnamese	English	English	English	N = 9 Italian, Portuguese, German, Russian, Polish, Hindi, Arabic, Mandarin, Korean
	Similarity	Embed. similarity = 0.55 Word overlap = 0.49	Embed. similarity = 0.53 Word overlap = 0.09	Embed. similarity = 0.48 Word overlap = 0.06	Embed. similarity = 0.49 Word overlap = 0.05	N.A.

- Multilingual language models capture properties of language processing that remain consistent across a range of linguistic and experimental contexts

Andrea Gregor de Varda1,2, Saima Malik-Moraleda2Andrea Gregor de Varda, Saima Malik-Moraleda, Greta Tuckute, Evelina Fedorenko. "Multilingual Computational Models Capture a Shared Meaning Component in Brain Responses across 21 Languages" 2025 BioarXiv.

Cross-lingual transfer is primarily driven by compositional semantic content

- Stimuli: Passage from book (21 languages)
- Stimulus representation: multilingual language models
- Brain recording & modality: fMRI, Listening
- Encoding models training: “perturbed” versions of the **Tuckute2024** and **Pereira2018**

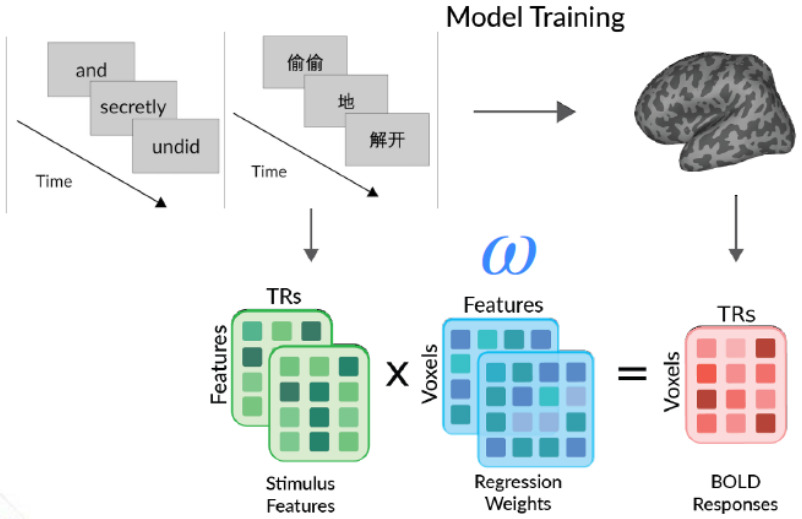


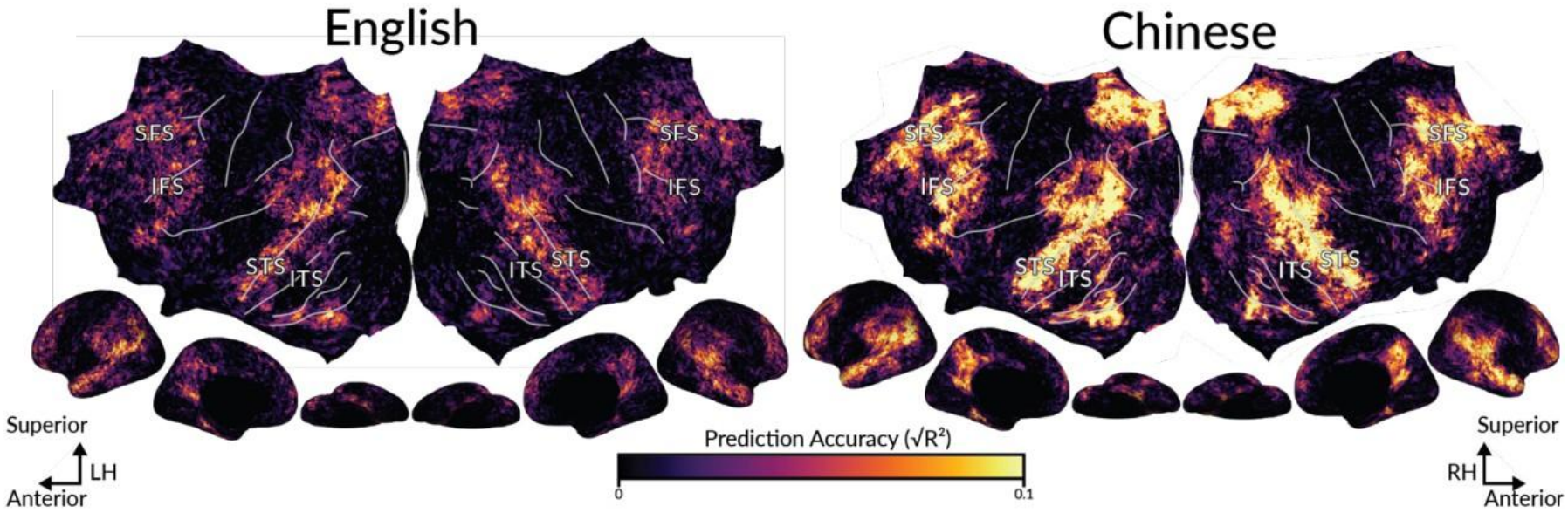
- All “perturbed” encoding models performed significantly worse than the intact encoding model apart from those marked with the dotted lines (n.s.; N + V + Adj, N + V, Paraphrase, 1 local swap)

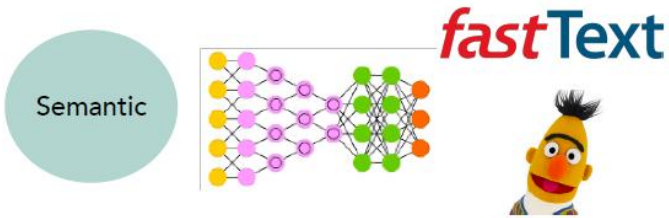
Andrea Gregor de Varda1,2, Saima Malik-Moraleda2Andrea Gregor de Varda, Saima Malik-Moraleda, Greta Tuckute, Evelina Fedorenko. "Multilingual Computational Models Capture a Shared Meaning Component in Brain Responses across 21 Languages" 2025 BioarXiv.

English- vs. Chinese: Bilingual language processing

- Stimuli: Bilingual-Moth-Radio-Hour (Chinese and English)
- Stimulus representation: facebook FastText model
- Brain recording & modality: fMRI, Reading





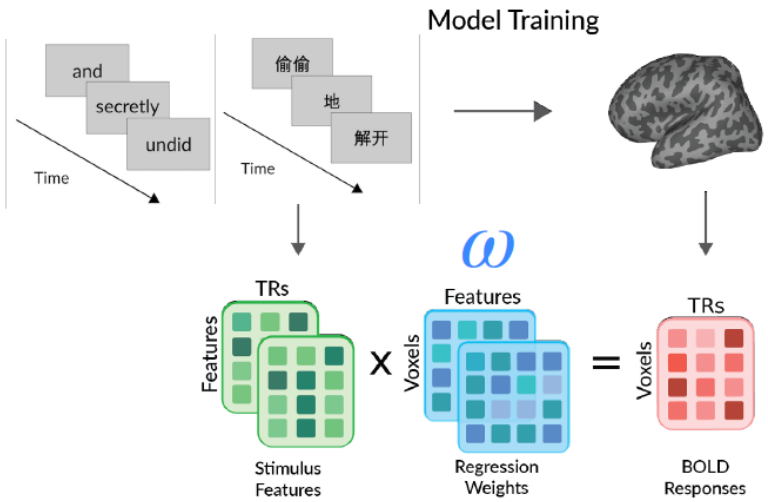


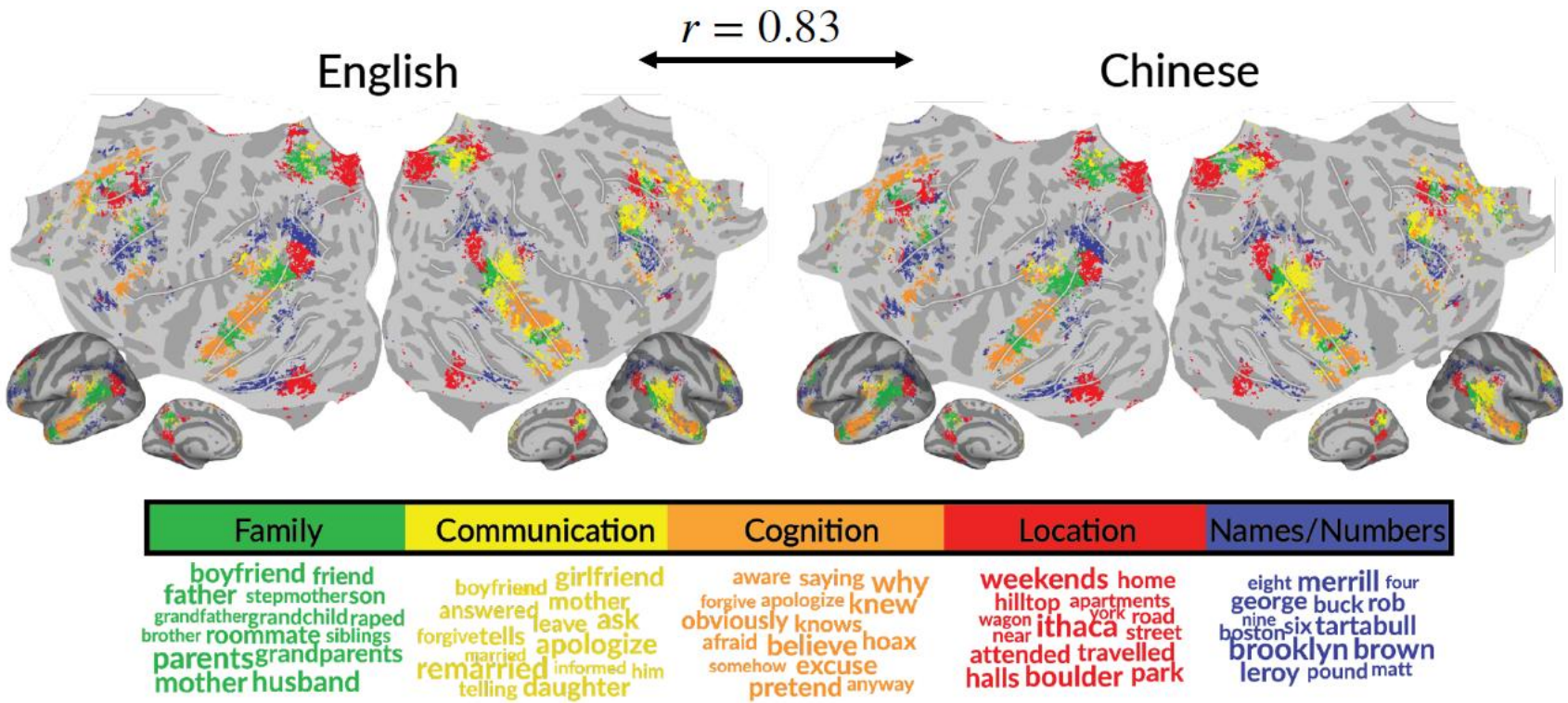
- Semantic representations are largely shared across languages in bilingual individuals

Catherine Chen, Xue L. Gong, Christine Tseng, Daniel L. Klein, Jack L. Gallant, Fatma Deniz. "Bilingual Language Processing Relies on Shared Semantic Representations that are Modulated by Each Language" 2024 arXiv.

Semantic representations are largely shared between languages in individual participant

- Stimuli: Bilingual-Moth-Radio-Hour (Chinese and English)
- Stimulus representation: facebook FastText model
- Brain recording & modality: fMRI, Reading





- Each voxel should be tuned towards similar word meanings between languages

Catherine Chen, Xue L. Gong, Christine Tseng, Daniel L. Klein, Jack L. Gallant, Fatma Deniz. "Bilingual Language Processing Relies on Shared Semantic Representations that are Modulated by Each Language" 2024 arXiv.

Conclusions for neuro-AI research field

1. **Build "universal meaning" representations:** train/align models to separate language-invariant vs language-specific components.

2. **Shared representations exist at multiple levels:** not only semantics, but also speech/phonetic features (e.g., Whisper-like features).

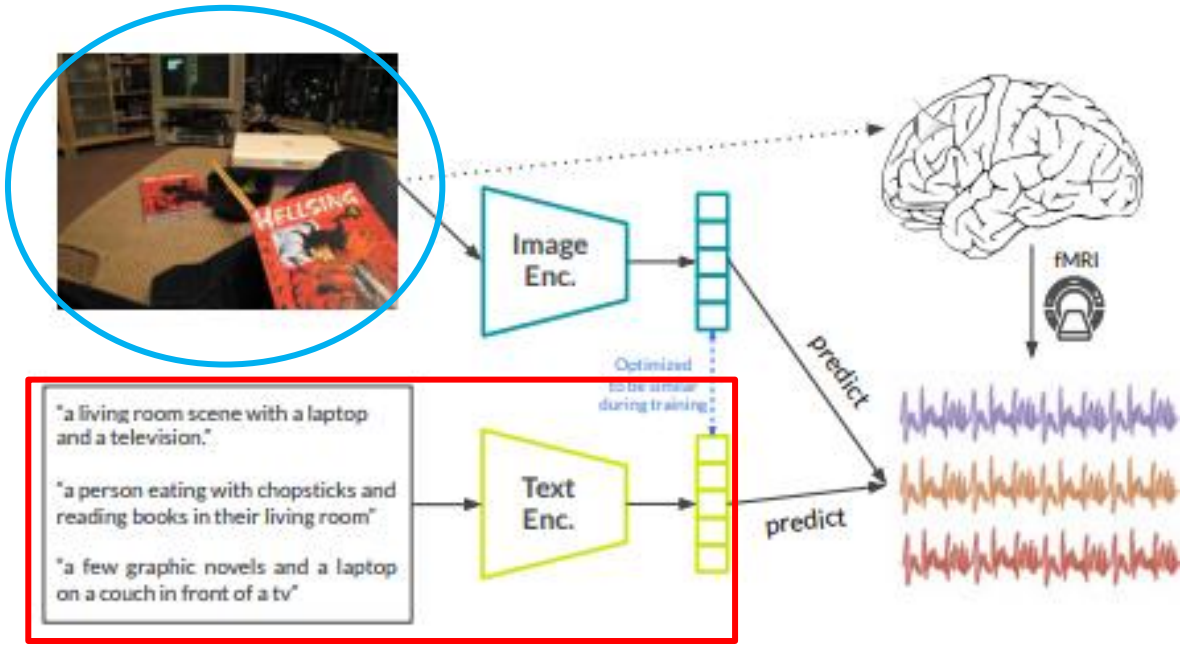
3. **Cross-lingual transfer reduces data needs:** enable low-resource languages and fewer subject-language sessions via zero-shot encoding.

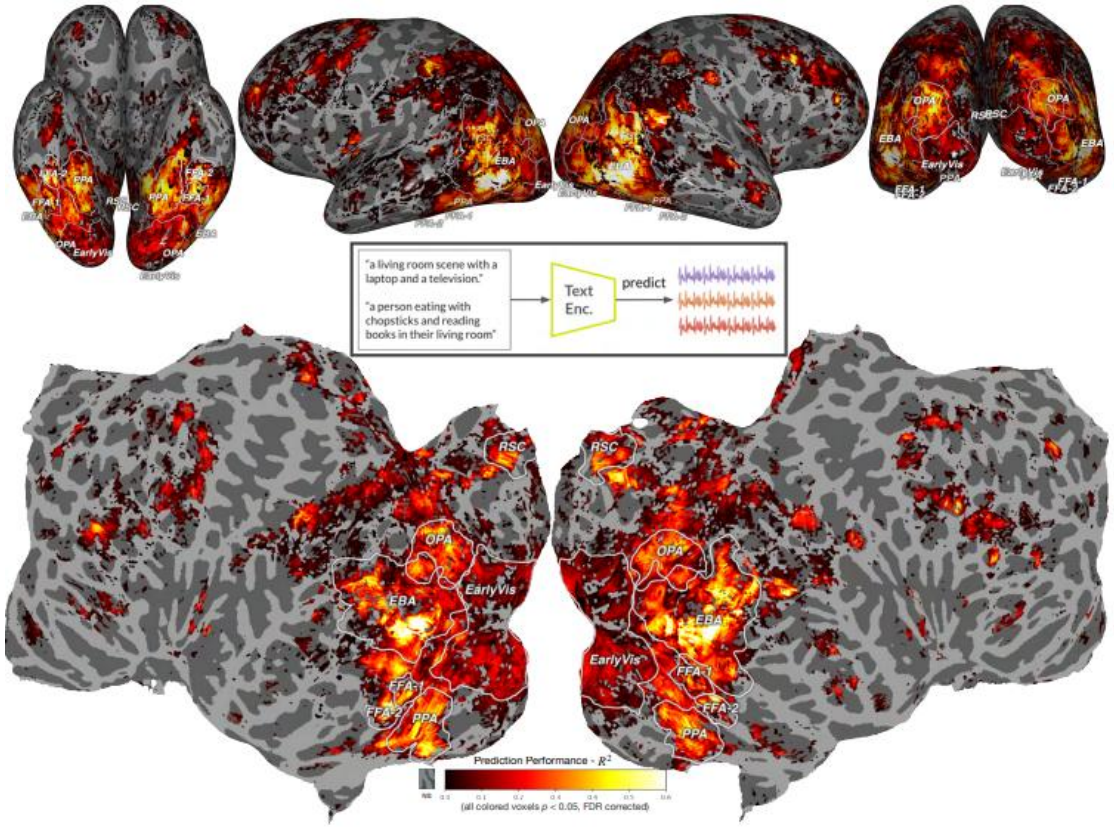
4. **Bilingual brains show shared semantics with language-specific modulation** at voxel/region level.

Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- **Brain Encoding [60 min] :**
  - Scaling Laws,
  - Multilinguality,
  - **Multimodal** and Instruction-tuned Models
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

Multimodal Transformer models can predict visual brain activity impressively well, even with text modality representations

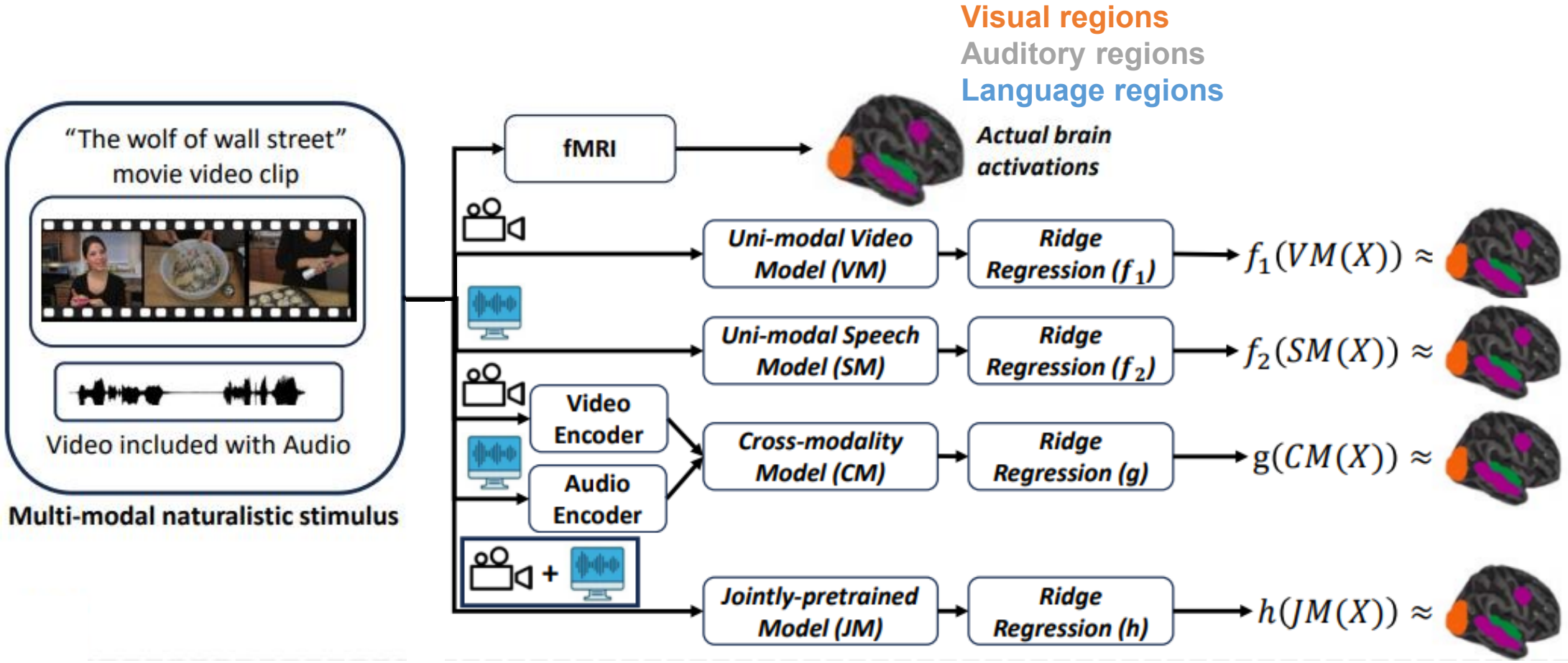




How accurately multi-modal models predict brain activity evoked by multi-modal stimuli?

Aria Y. Wang, Kendrick Kay. "Incorporating natural language into vision models." Bioarxiv 2023.

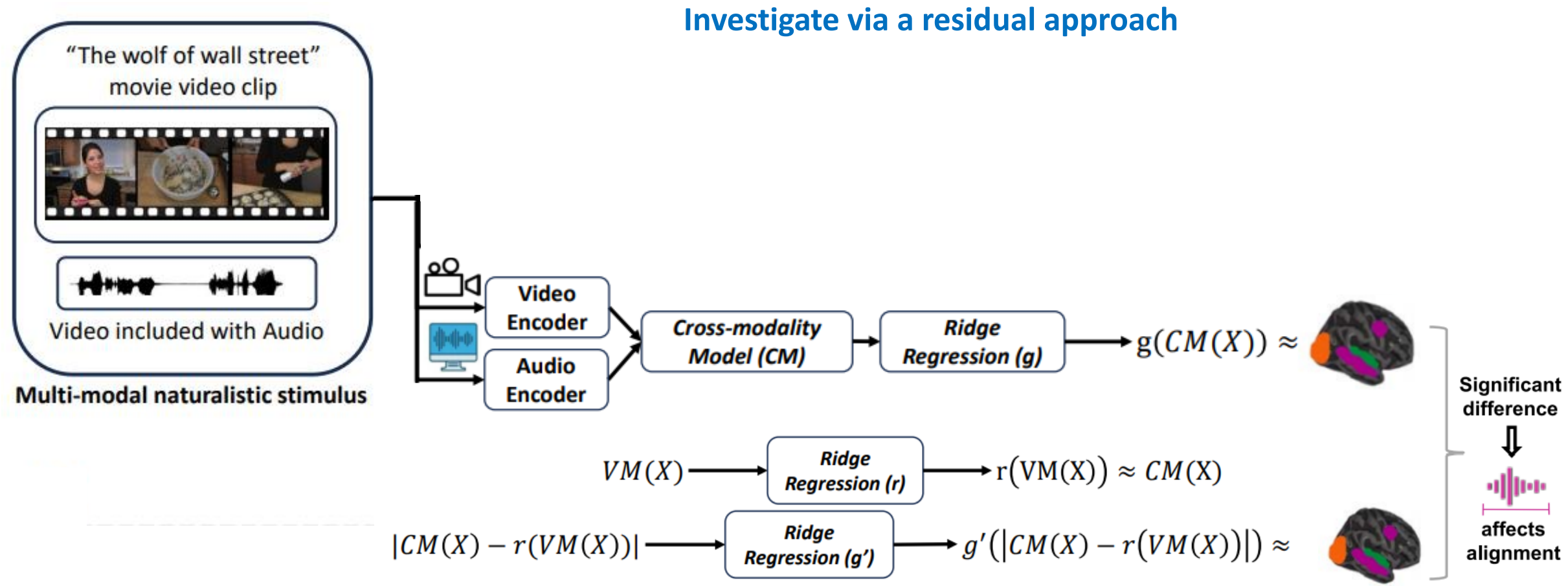
Multimodal vs. Unimodal models : brain alignment



- How well do multimodal models predict multimodal stimulus-evoked brain activity over unimodal models?
- How our brains separates and integrates information across modalities through a hierarchy of early sensory regions to higher cognition (language regions)?

Subba Reddy Oota, Khushbu Pahwa, mounika marreddy, Maneesh Kumar Singh, Manish Gupta, Bapi Raju Surampudi. "Multi-modal brain encoding models for multi-modal stimuli" ICLR-2025.

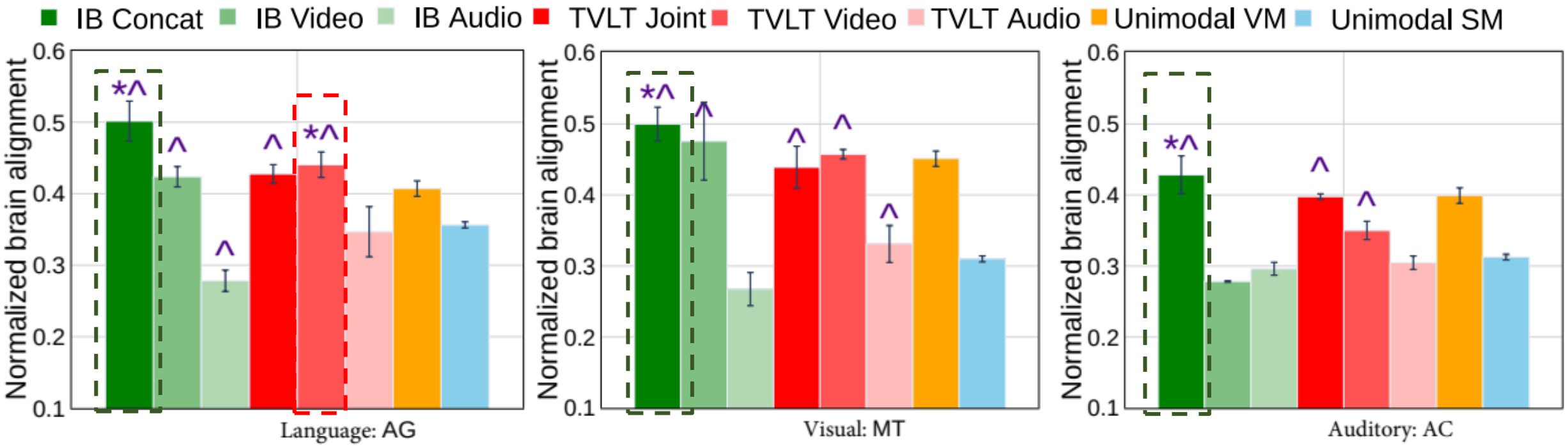
Which modality of representations in multimodal models lead to high brain alignment?



Subba Reddy Oota, Khushbu Pahwa, mounika marreddy, Maneesh Kumar Singh, Manish Gupta, Bapi Raju Surampudi. "Multi-modal brain encoding models for multi-modal stimuli" ICLR-2025.

Surprising Trends in Brain Alignment: Unimodal vs. Multimodal Models

- Stimuli: Movie videos
- Stimulus representation: ImageBind, TVLT, VideoMAE, AST
- Brain recording & modality: fMRI, Watching movies

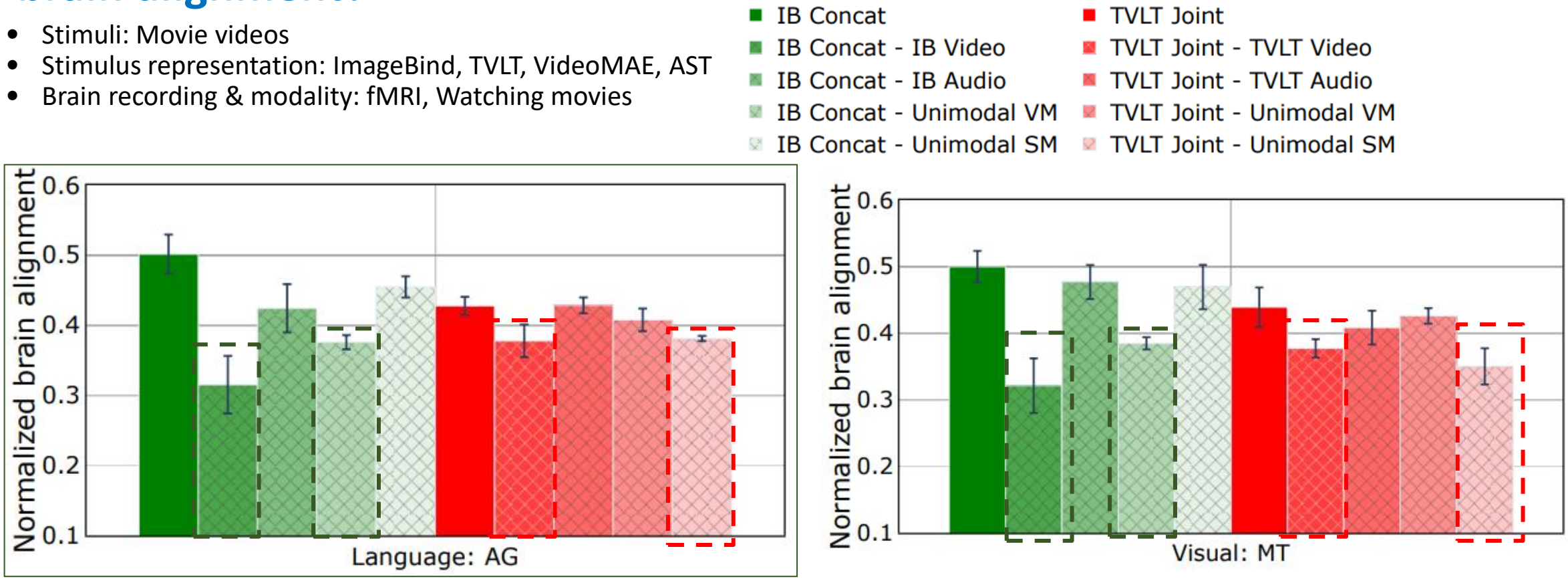


- **Language region (AG)**:
  - Both types of multi-modal models show high brain alignment than unimodal video and speech models with **language regions**, but **audio models** trails behind **video models**
- **Higher-visual (MT) and Early-sensory (AC):**
  - **Cross-Modal Models**: concat embeddings improve alignment, while **jointly-pretrained models** perform similarly to unimodal video models.

Subba Reddy Oota, Khushbu Pahwa, mounika marreddy, Maneesh Kumar Singh, Manish Gupta, Bapi Raju Surampudi. "Multi-modal brain encoding models for multi-modal stimuli" ICLR-2025.

Which modality of representations in multi-modal models lead to high brain alignment?

- Stimuli: Movie videos
- Stimulus representation: ImageBind, TVLT, VideoMAE, AST
- Brain recording & modality: fMRI, Watching movies



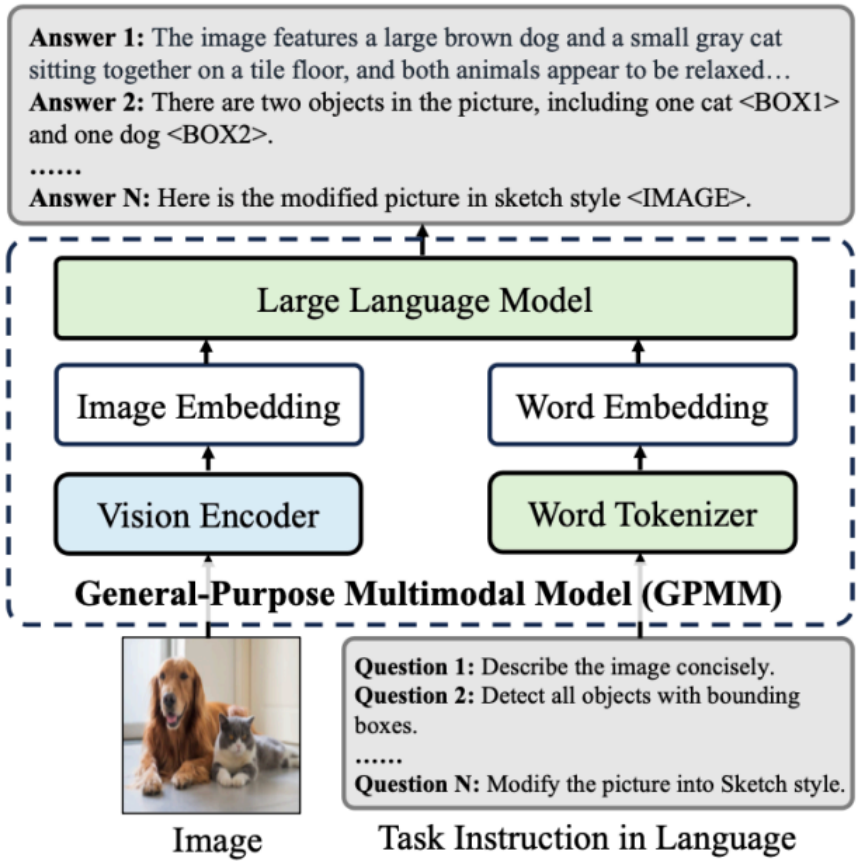
- **Cross-modal models**:
  - brain alignment **partially explained** by video features, but removal of speech features does not lead to drop in brain alignment.
- **Jointly-pretrained models**:
  - brain alignment **partially explained** by both video and audio features

Subba Reddy Oota, Khushbu Pahwa, mounika marreddy, Maneesh Kumar Singh, Manish Gupta, Bapi Raju Surampudi. "Multi-modal brain encoding models for multi-modal stimuli" ICLR-2025.

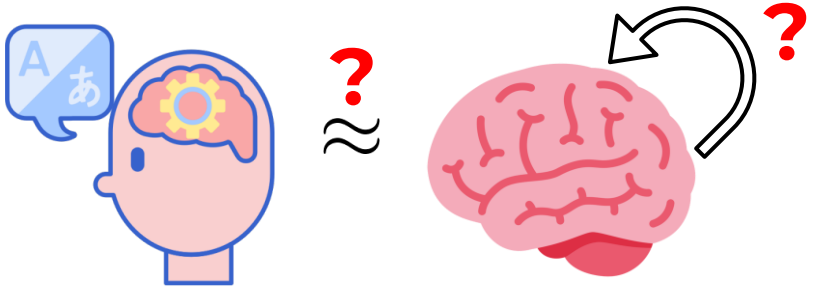
Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- **Brain Encoding [60 min] :**
  - Scaling Laws,
  - Multilinguality,
  - Multimodal and **Instruction-tuned Models**
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

Multimodal instruction tuning enables models to generalize to new tasks by following unseen instructions



INPUT: <image>Describe this image in detail.
OUTPUT: <long descriptions>



How do multimodal instruction-tuned LLMs process visual images when guided by natural language task instructions?

How does the brain integrate information during the processing of visual images?

Do multimodal instruction-tuned models prompted with natural language improve brain alignment and capture instruction-specific representations?

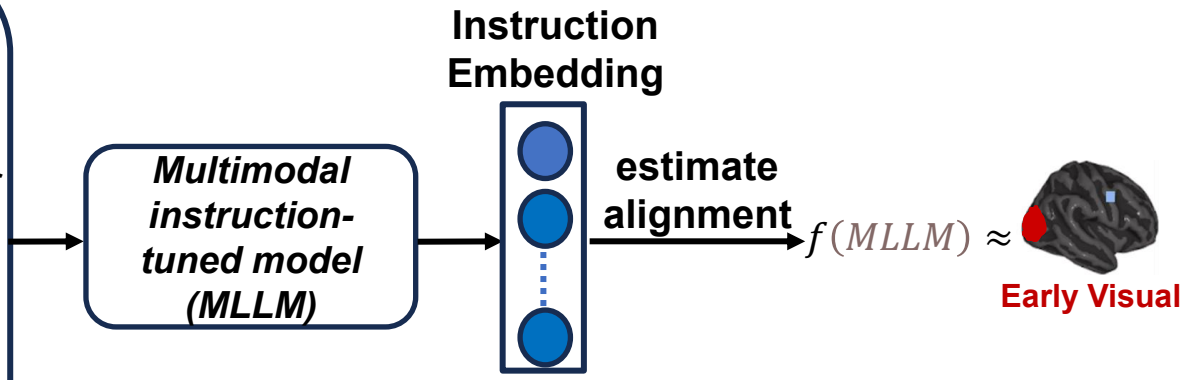
Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, Enhong Chen. "A Survey on Multimodal Large Language Models." Arxiv 2024.

Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain)

NSD dataset naturalistic Image stimulus

Image Captioning:
What is the caption of the image?
Image Understanding:
Describe the most dominant color in the image.
Visual Relationship:
What objects are being used by the largest animal in this image?

Task-specific instructions



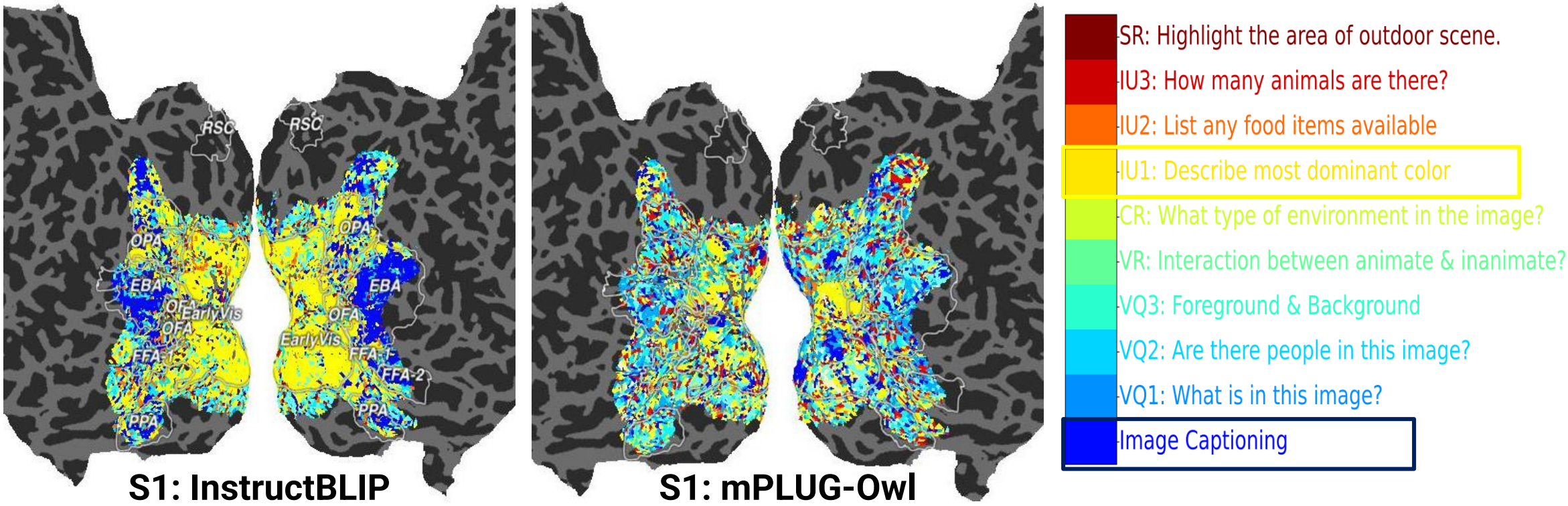
Task	Description
Image Understanding	IU1: Describe the most dominant color in the image IU2: List any food items visible. IU3: How many animals are there in the image?
Visual Question Answering	VQ1: What is in this image? VQ2: Are there any people in this image? If yes, describe them. VQ3: What is the foreground of the image? What is in the background?
Image Captioning	IC: Generate some text to describe the image
Scene Recognition	SR: Highlight the area that shows a natural outdoor scene.
Commonsense Reasoning	CR: What type of environment is shown in the image?
Visual Relationship	VR: What kind of interaction is happening between the animate and inanimate objects here?

- How well do MLLMs predict brain activity evoked by visual stimuli under task-specific instructions compared to unimodal and multimodal models?
- Do instruction-specific representations in MLLMs differentiate visual brain regions involved in processing, thereby aligning with the mechanisms of human visual cognition?

Subba Reddy Oota, Akshett Rai Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi GNVV, Manish Shrivastava, Maneesh Kumar Singh, Bapi Raju Surampudi, Manish Gupta. "Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain)" ICLR-2025.

Which task-specific instructions are highly correlated to visual function localizers?

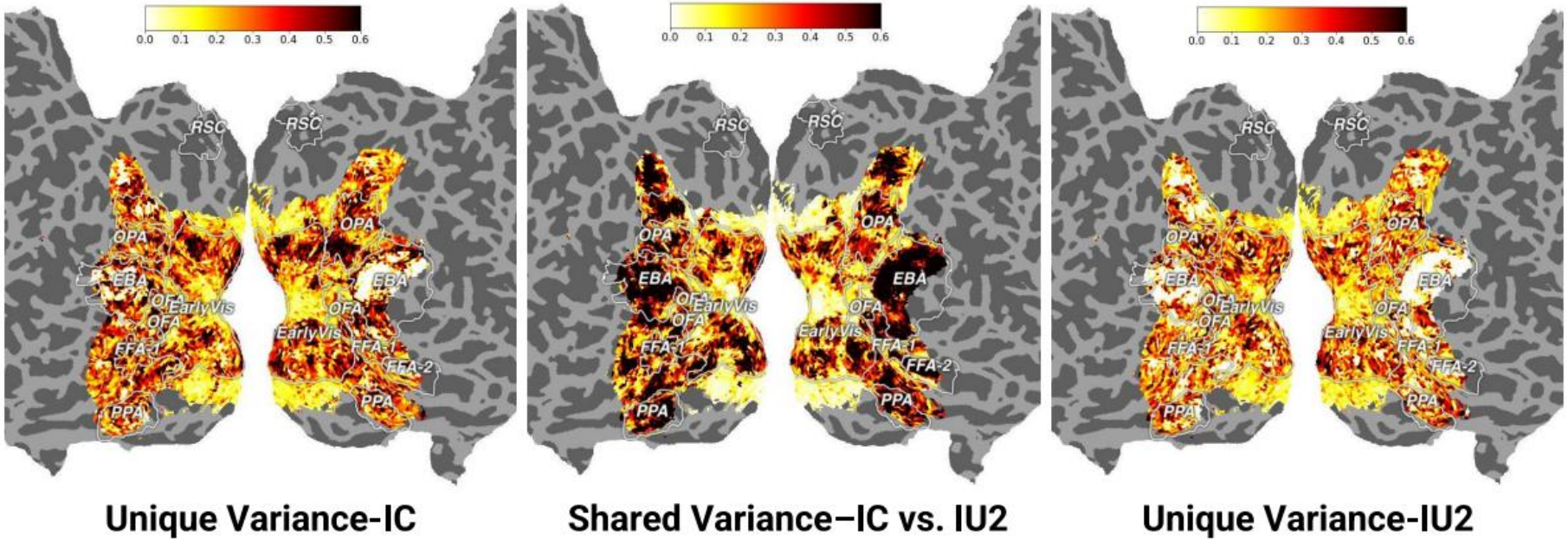
- Stimuli: Naturalistic images
- Stimulus representation: InstructBLIP, mPLUG-Owl
- Brain recording & modality: fMRI, Watching images



- Not all instructions lead to increased brain alignment across all regions
- Certain instructions (IC, VQ2, and IU1) are more effective than others in encoding brain activity.

Subba Reddy Oota, Akshett Rai Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi GNVV, Manish Shrivastava, Maneesh Kumar Singh, Bapi Raju Surampudi, Manish Gupta. "Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain)" ICLR-2025.

Partitioning explained variance between task-specific instructions



- Between **Image Captioning (IC)** and **Image Understanding (IU2):** there is no unique variance for **IU2** in the **EBA region (higher-visual)**, while **IC** retains some unique variance.
- Task-specific instructions exhibit **moderate shared variance** in the **early visual cortex**, while **shared variance is significantly higher** in **higher visual ROIs**

Subba Reddy Oota, Akshett Rai Jindal, Ishani Mondal, Khushbu Pahwa, Satya Sai Srinath Namburi GNVV, Manish Shrivastava, Maneesh Kumar Singh, Bapi Raju Surampudi, Manish Gupta. "Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain)" ICLR-2025.

Overview of multimodal model evaluation settings in brain encoding studies

Study	Model Type	Stimulus Modality	Brain Data	Dataset	Instruction-Tuned
Doerig et al. (2022)	Vision-Language (CLIP)	Unimodal (Images)	fMRI	NSD	✗
Wang et al. (2023)	Vision-Language (CLIP)	Unimodal (Images)	fMRI	NSD	✗
Oota et al. (2022b)	Vision-Language (CLIP, VisualBERT, LXMERT)	Unimodal (Images)	fMRI	BOLD5000	✗
Popham et al. (2021)	Vision-Only CNNs vs. Vision-Language	Unimodal (Silent Videos)	fMRI	Gallant lab short video clips	✗
Tang et al. (2022)	non-instruction-tuned multimodal model (BridgeTower)	Unimodal (Silent Videos), Unimodal (listening stories)	fMRI	Gallant lab short video clips	✗
Oota et al. (2025a)	Instruction-tuned Image+Text MLLMs	Unimodal (Images)	fMRI	NSD	✓
Sartzetaki et al. (2025)	Image Recognition models, Action recognition models	Unimodal (Visual)	fMRI	Bold Moments Dataset	✗
Nakagi et al. (2024)	Language models (BERT, GPT-2, Lllama2, OPT)	Multimodal (Videos with audio)	fMRI	8.3 hours of video dataset	✗
Subramaniam et al. (2024)	non-instruction-tuned multimodal models (SLIP-CLIP, SimCLR, BLIP, BEIT)	Image frame-text pairs (Movies)	SEEG	AMMT	✗
Dong & Toneva (2023a)	non-instruction-tuned multimodal models (Merlose-serve)	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Friends dataset	✗
Oota et al. (2025b)	non-instruction-tuned multimodal models (TVLT and ImageBind)	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Movie10	✗
Our study	instruction-tuned video and audio MLLMs	Multimodal (Movies: Videos with audio)	fMRI	Neuromod Movie10	✓

- Stimuli: Movie videos
- Stimulus representation: IT MLLMs, ICL MLLMs, Multimodal and Unimodal models
- Brain recording & modality: fMRI, Watching movies

Model Name	IT	#Layers	Modality
InstructBLIPVideo	✓	33	Video+Text
Video-LLaVA	✓	33	Video+Text
LLaVa-NeXT-Video	✓	33	Video+Text
Qwen-2.5-VL	✓	29	Video+Text
Videochat-R1	✓	29	Video+Text
LLaVA-OneVision	✓	28	Video+Text
Qwen-2.5-Audio	✓	29	Audio+Text
Kimi-Audio	✓	29	Audio+Text
TVLT	×	12	Video+Audio
VideoMAE	×	24	Video
AST	×	24	Audio

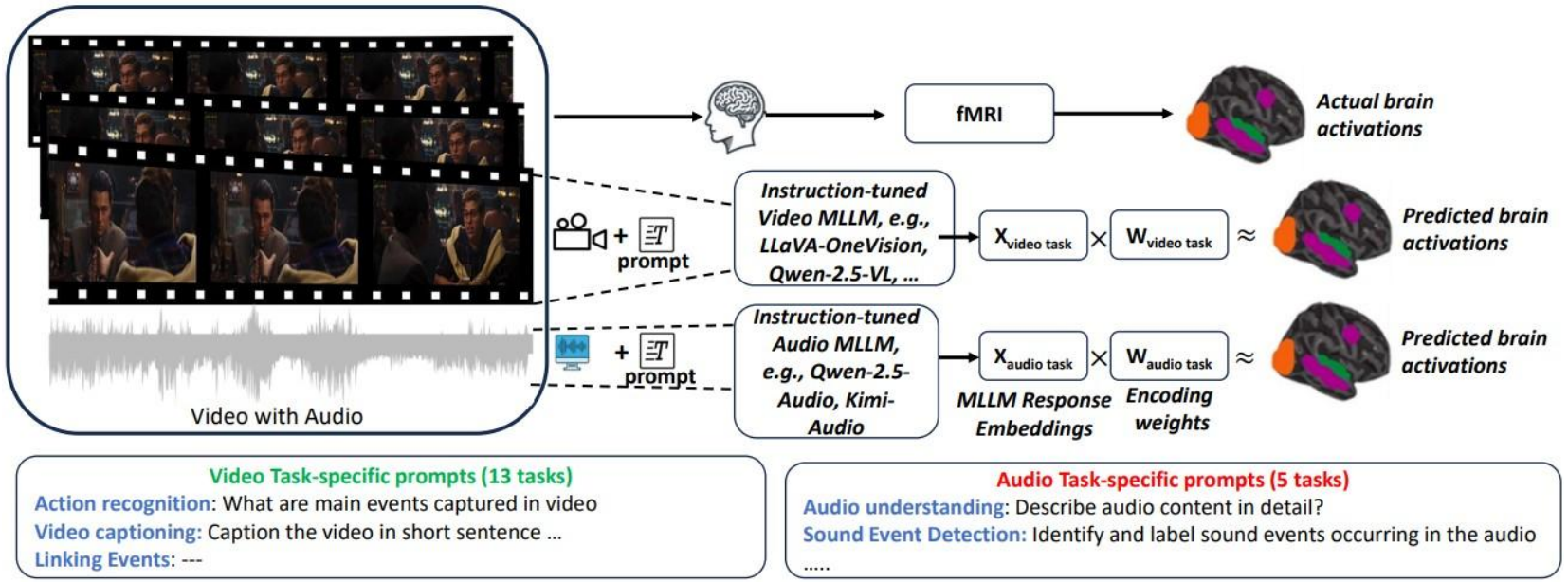
- Do instruction-tuned video/audio MLLMs prompted with natural language yield better brain alignment than their non-instruction-tuned counterparts and distinguish task-specific representations?

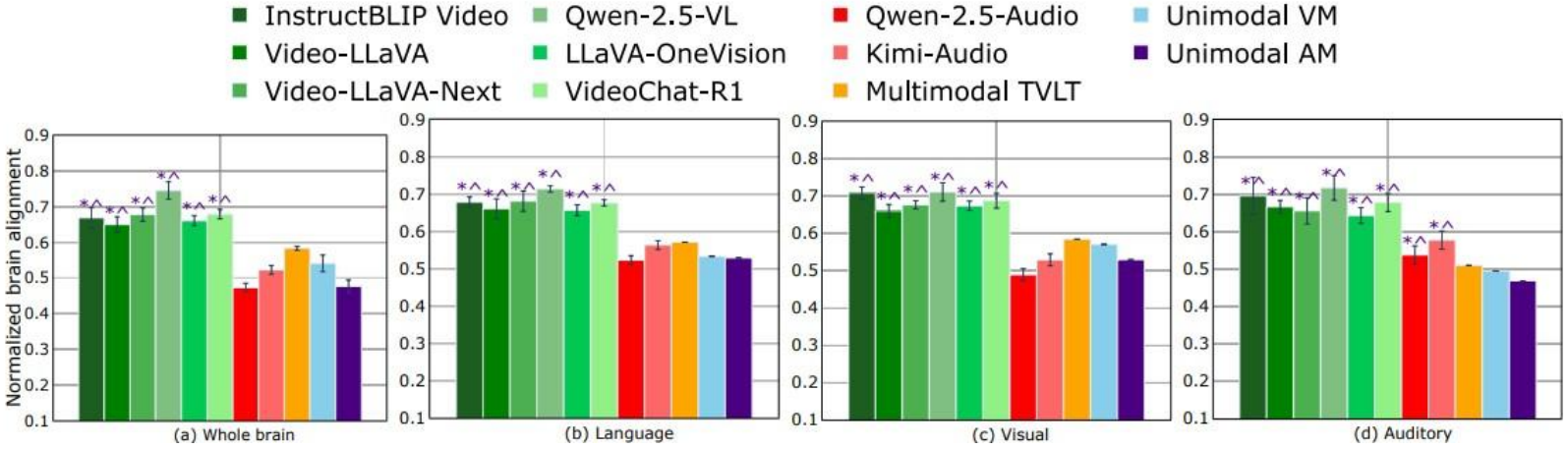
Subba Reddy Oota, Khushbu Pahwa, Satya Sai Srinath Namburi GNVV, Maneesh Kumar Singh, Bapi Raju Surampudi, Manish Gupta. "INSTRUCTION-TUNED VIDEO-AUDIO MODELS ELUCIDATE FUNCTIONAL SPECIALIZATION IN THE BRAIN" Arxiv 2025.

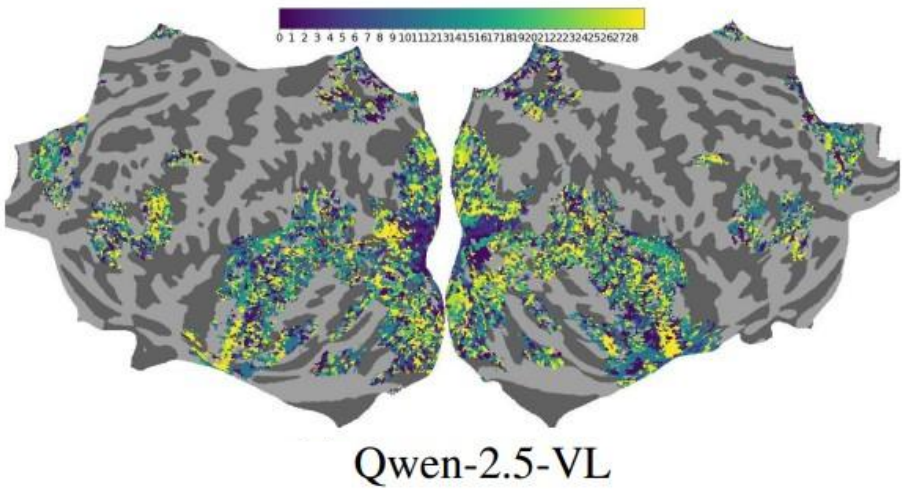
Instruction-tuned video and audio MLLMs successfully differentiate task-specific instructions

- Stimuli: Movie videos
- Stimulus representation: IT MLLMs, ICL MLLMs, Multimodal and Unimodal models
- Brain recording & modality: fMRI, Watching movies

- MLLM layers align hierarchically with the brain
- Language-guided instructions shows clear disentanglement in task-specific representations from MLLMs







Subba Reddy Oota, Khushbu Pahwa, Satya Sai Srinath Namburi GNVV, Maneesh Kumar Singh, Bapi Raju Surampudi, Manish Gupta. "INSTRUCTION-TUNED VIDEO-AUDIO MODELS ELUCIDATE FUNCTIONAL SPECIALIZATION IN THE BRAIN" Arxiv 2025.

Conclusions for neuro-AI research field

1. 🗣️ **MLLMs** generate task-specific output tokens based on instructions, but **not all instructions lead to better brain alignment.**

2. Use **instructions** as **experimental probes**: a new “task-specific prompts” for vision/speech can map functional specialization and dissociate ROIs.

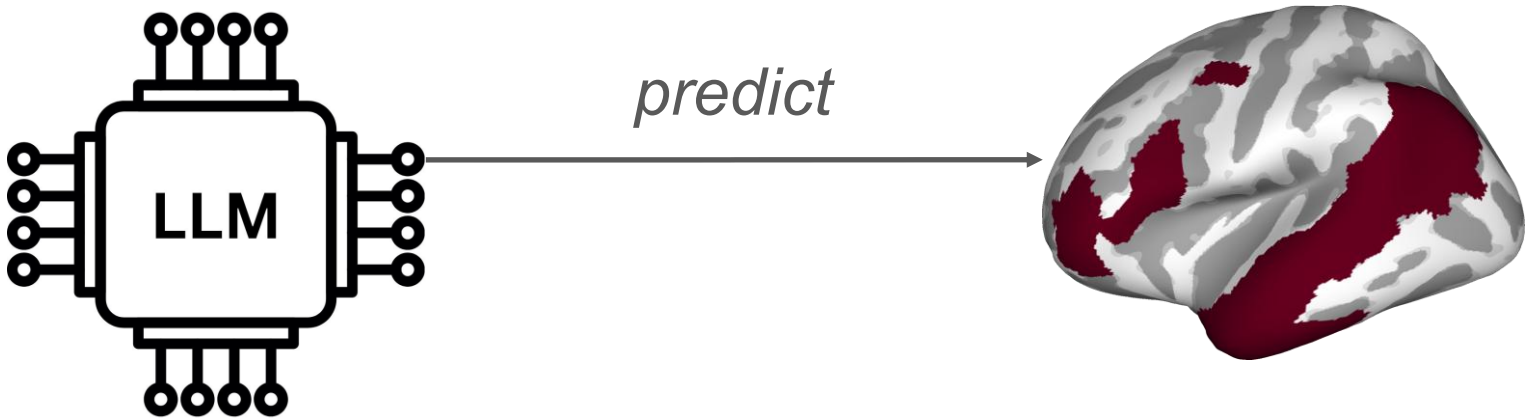
3. The **variance** in brain alignment is shared across task-specific instructions:
   - Moderate in 🧠 *early visual areas*
   - Higher in 🧠 *high-level visual regions*

4. **Better benchmarks:** evaluate models by instruction-specific brain predictivity + variance partitioning, not just average alignment.

Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- Brain Encoding [60 min] :
  - Scaling Laws,
  - Multilinguality,
  - Multimodal and Instruction-tuned Models
- Coffee Break & Networking [30 min]
- **Brain-informed Fine-tuning of Language Models [30 min]**
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

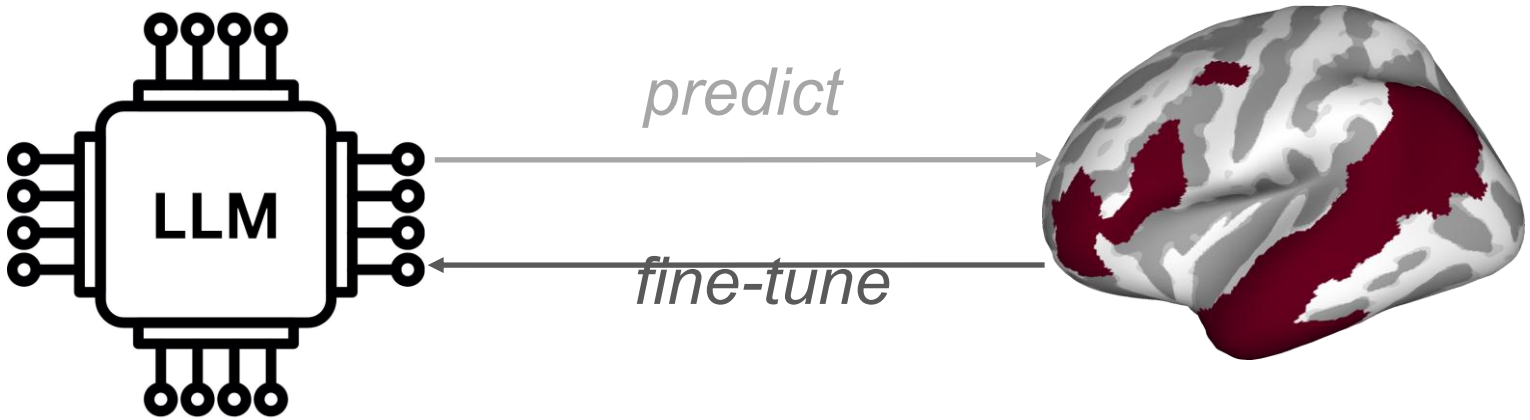
Language models accurately predict brain activity during language processing



Wehbe et al., 2014b; Jain & Huth, 2018; Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux & King, 2022; Goldstein et al., 2022; Karamolegkou et al., 2023; Oota et al., 2025

*LLM: Large Language Model

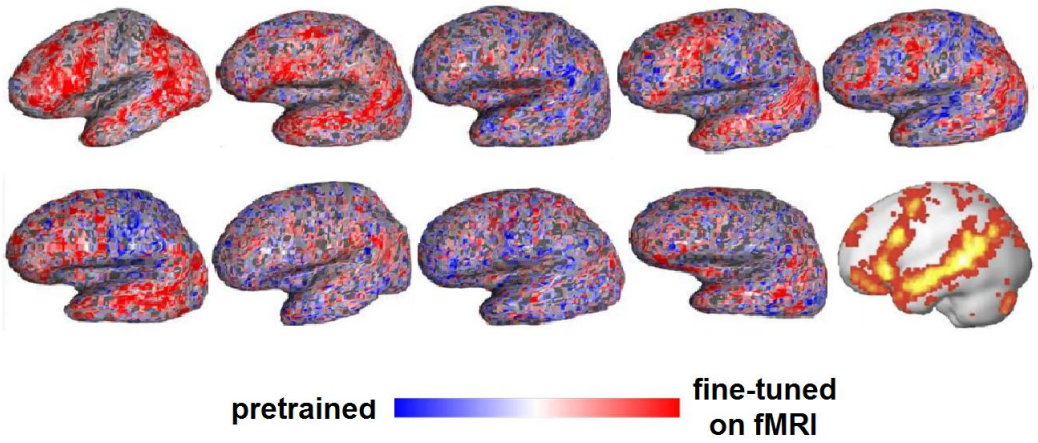
Fine-tuning language models with brain data

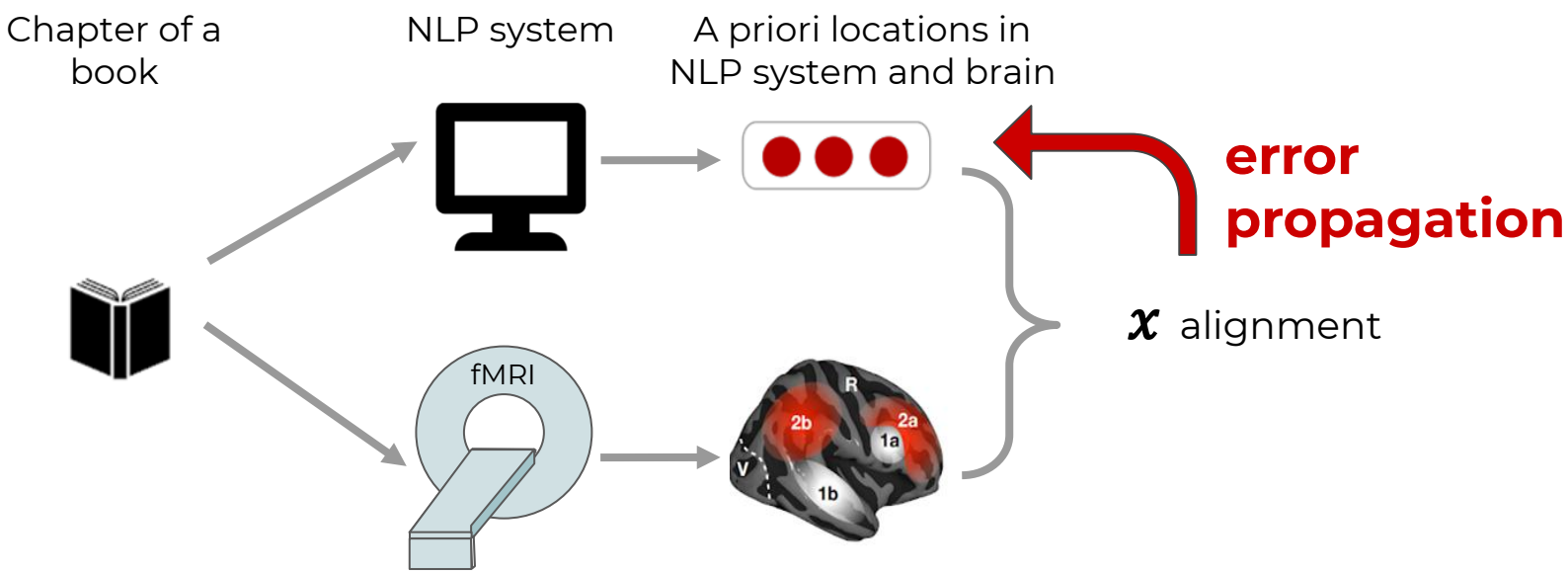


Schwartz et al., 2019

Training DL models using brain recordings

- Stimuli: one chapter of Harry Potter
- Stimulus representation: brain-optimized NLP model
- Brain recording & modality: fMRI & MEG, reading

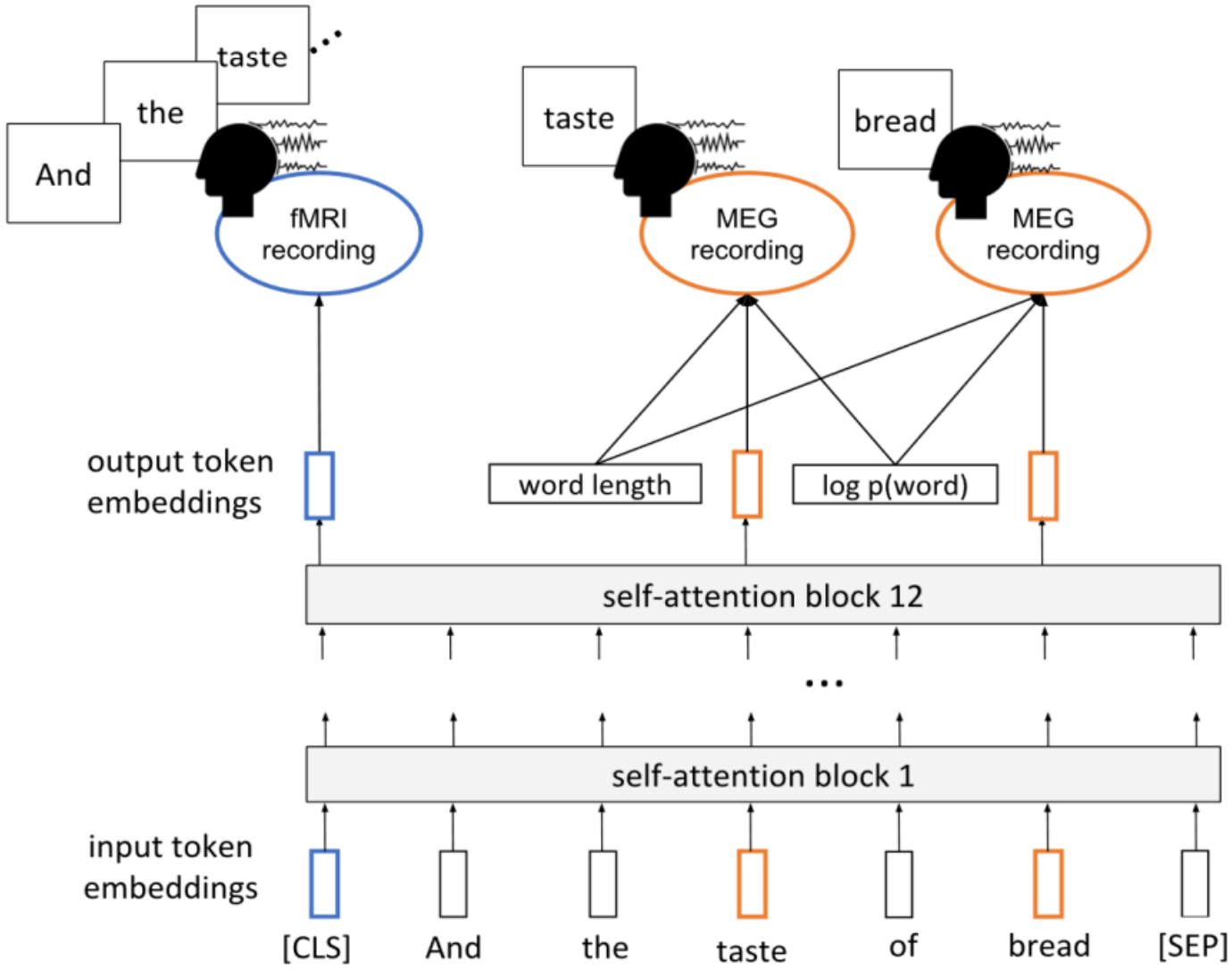




Brain-optimized NLP model predicts unseen fMRI recordings better, especially in canonical language regions

Schwartz, Dan, Mariya Toneva, and Leila Wehbe. "Inducing brain-relevant bias in natural language processing models." Advances in neural information processing systems 32 (2019).

Inducing Brain Relevant Bias

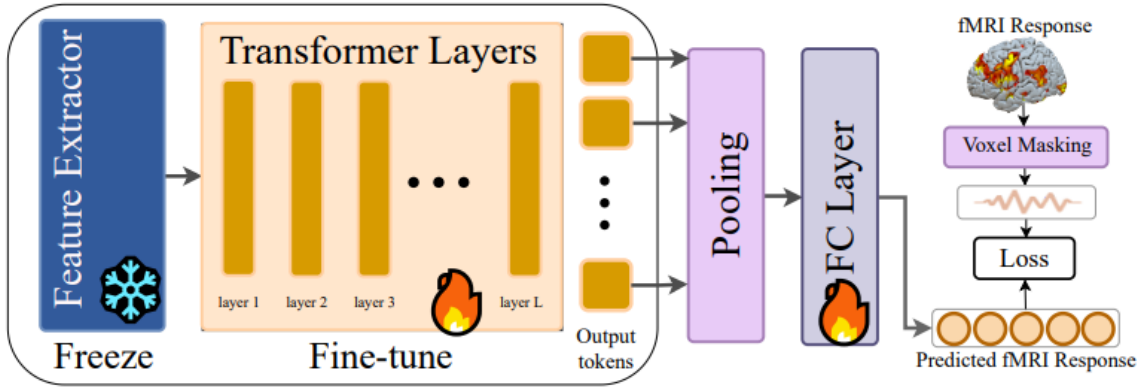


Metric	Vanilla	MEG	Joint
CoLA	57.29	57.63	**57.97**
SST-2	93.00	**93.23**	91.62
MRPC (Acc.)	83.82	83.97	**84.04**
MRPC (F1)	88.85	**88.93**	88.91
STS-B (Pears.)	**89.70**	89.32	88.60
STS-B (Spear.)	**89.37**	88.87	88.23
QQP (Acc.)	90.72	**91.06**	90.87
QQP (F1)	87.41	**87.91**	87.69
MNLI-m	83.95	**84.26**	84.08
MNLI-mm	84.39	84.65	**85.15**
QNLI	89.04	**91.73**	91.49
RTE	61.01	**65.42**	62.02
WNLI	53.52	**53.80**	51.97

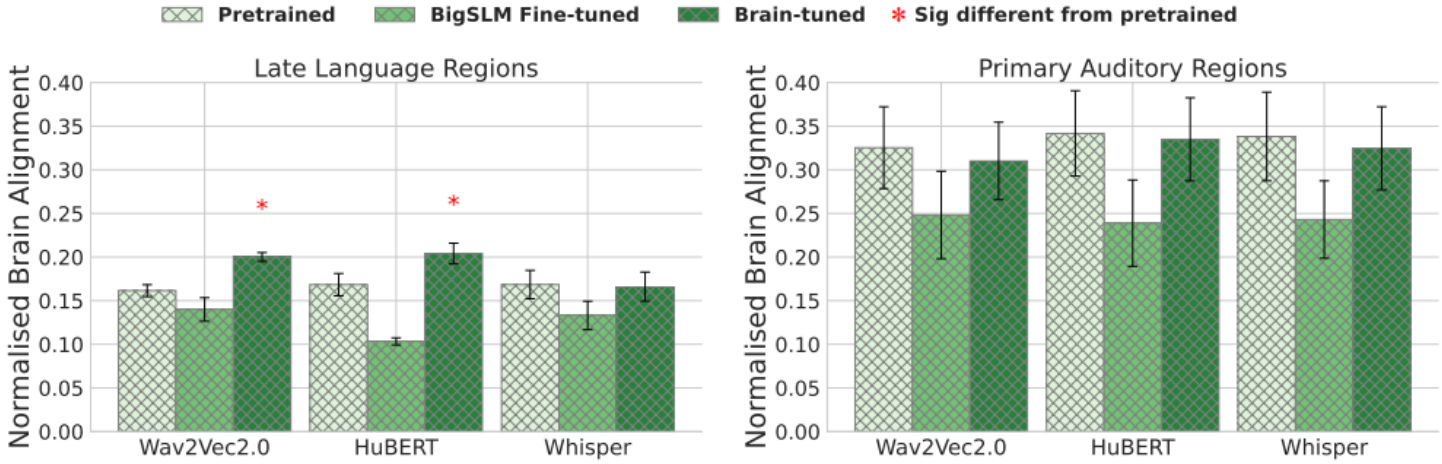
Schwartz, Dan, Mariya Toneva, and Leila Wehbe. "Inducing brain-relevant bias in natural language processing models." Advances in neural information processing systems 32 (2019).

Training Speech models using brain recordings

- Stimuli: Moth-Radio-Hour
- Stimulus representation: brain-optimized speech model
- Brain recording & modality: fMRI, listening

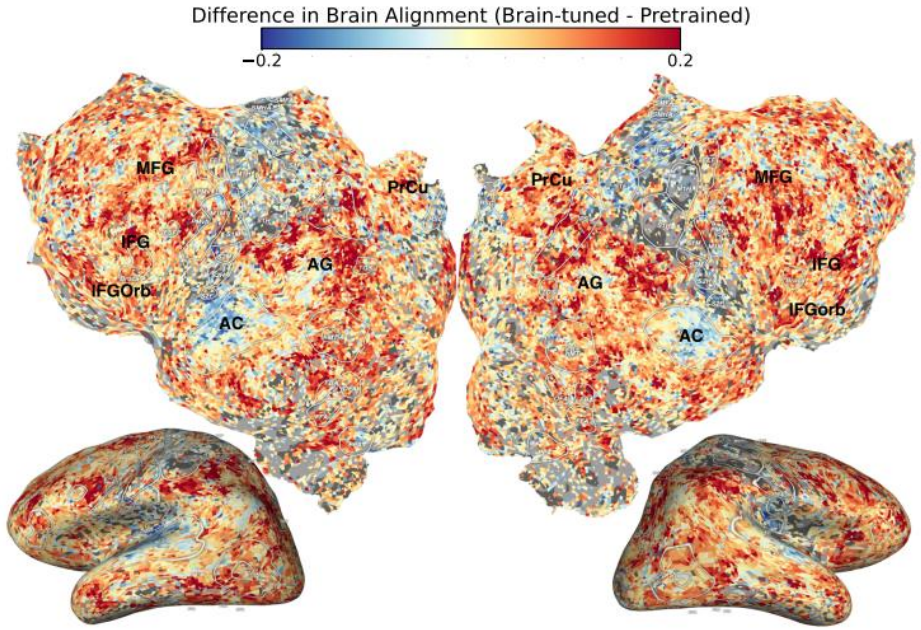


(a) Proposed brain-tuning approach



(a) Normalized alignment for late language regions

(b) Normalized alignment for primary auditory

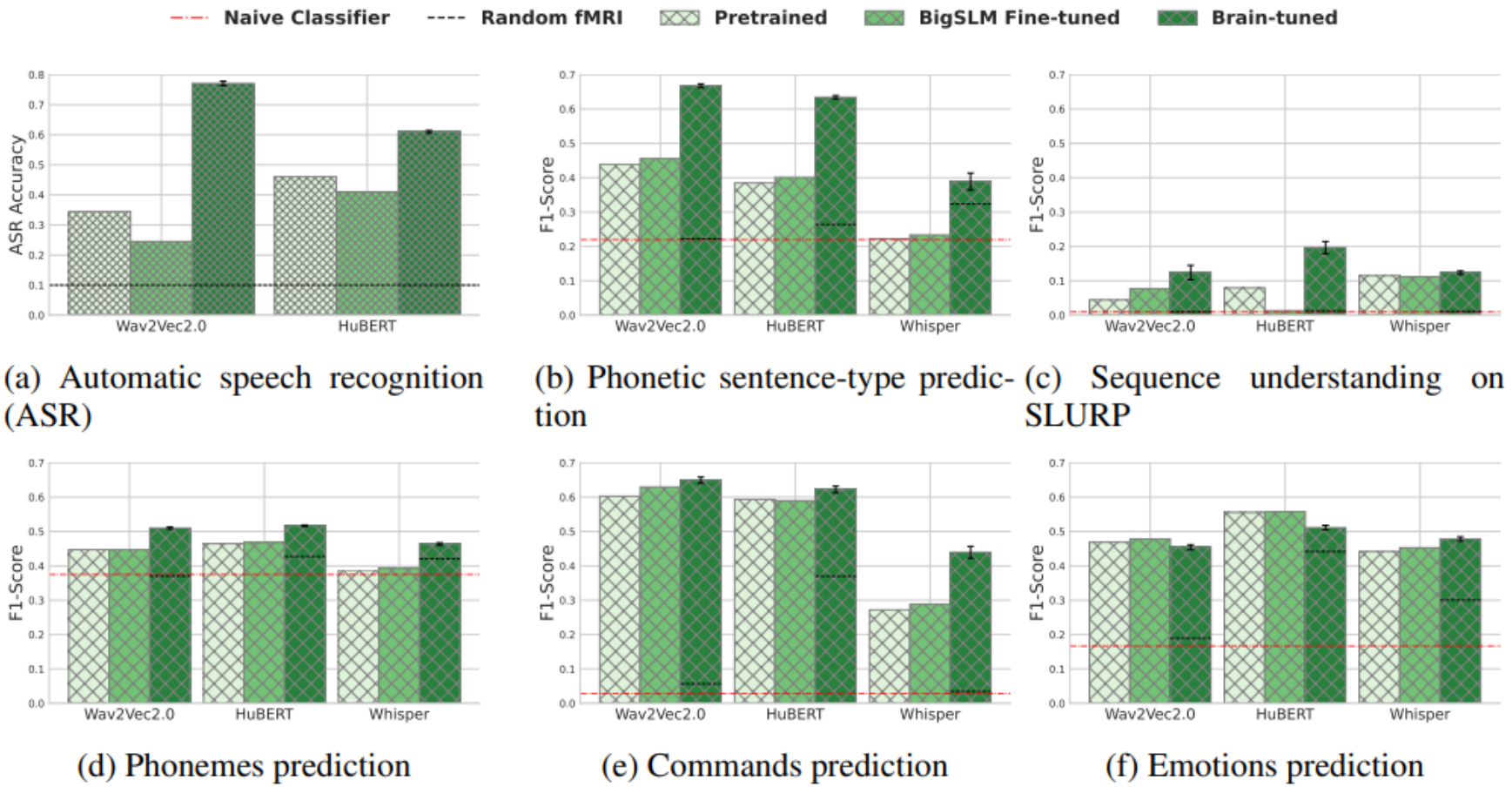


(c) Difference in brain alignment due to brain-tuning of Wav2vec2.0

- Brain-tuning may improve the brain-relevant semantics in at least some speech language models

Omer Moussa, Dietrich Klakow, Mariya Toneva. "Improving semantic understanding in speech language models via brain-tuning" Arxiv 2024.

Downstream performance



(a) Automatic speech recognition (ASR)

(b) Phonetic sentence-type prediction

(c) Sequence understanding on SLURP

(d) Phonemes prediction

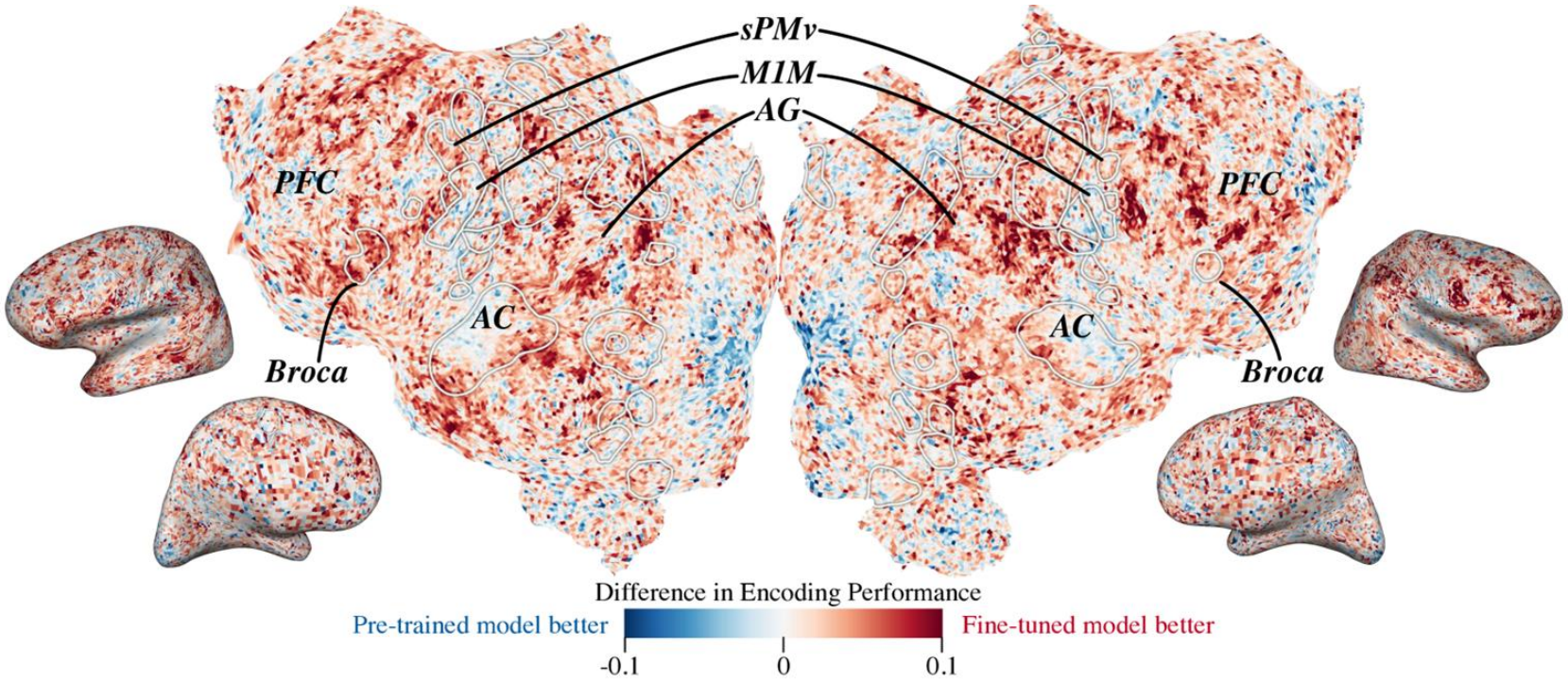
(e) Commands prediction

(f) Emotions prediction

- Brain-tuned models show consistent improvement over the baselines, with biggest gains in more semantic tasks (ASR and phonetic sentence-type prediction)

Omer Moussa, Dietrich Klakow, Mariya Toneva. "Improving semantic understanding in speech language models via brain-tuning" Arxiv 2024.

Fine-tuning language models with brain data



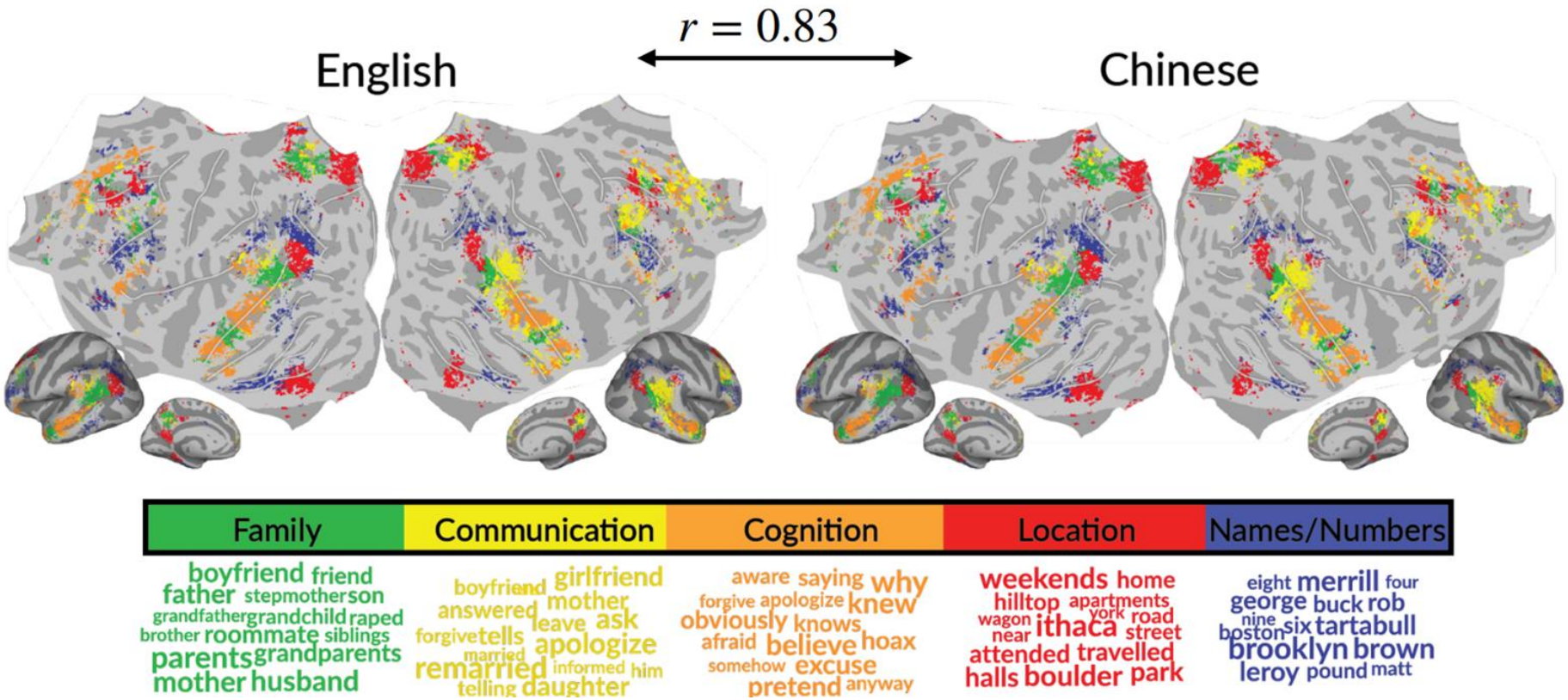
- ➔ improves alignment with the brain
- ➔ improves their semantic downstream task performance

! fine-tuning with monolingual brain data (English)

! only monolingual models were evaluated

Schwartz et al., 2019; Vattikonda, Vaidya, Antonello, Huth 2025

Shared semantic representations in bilinguals accurately predict

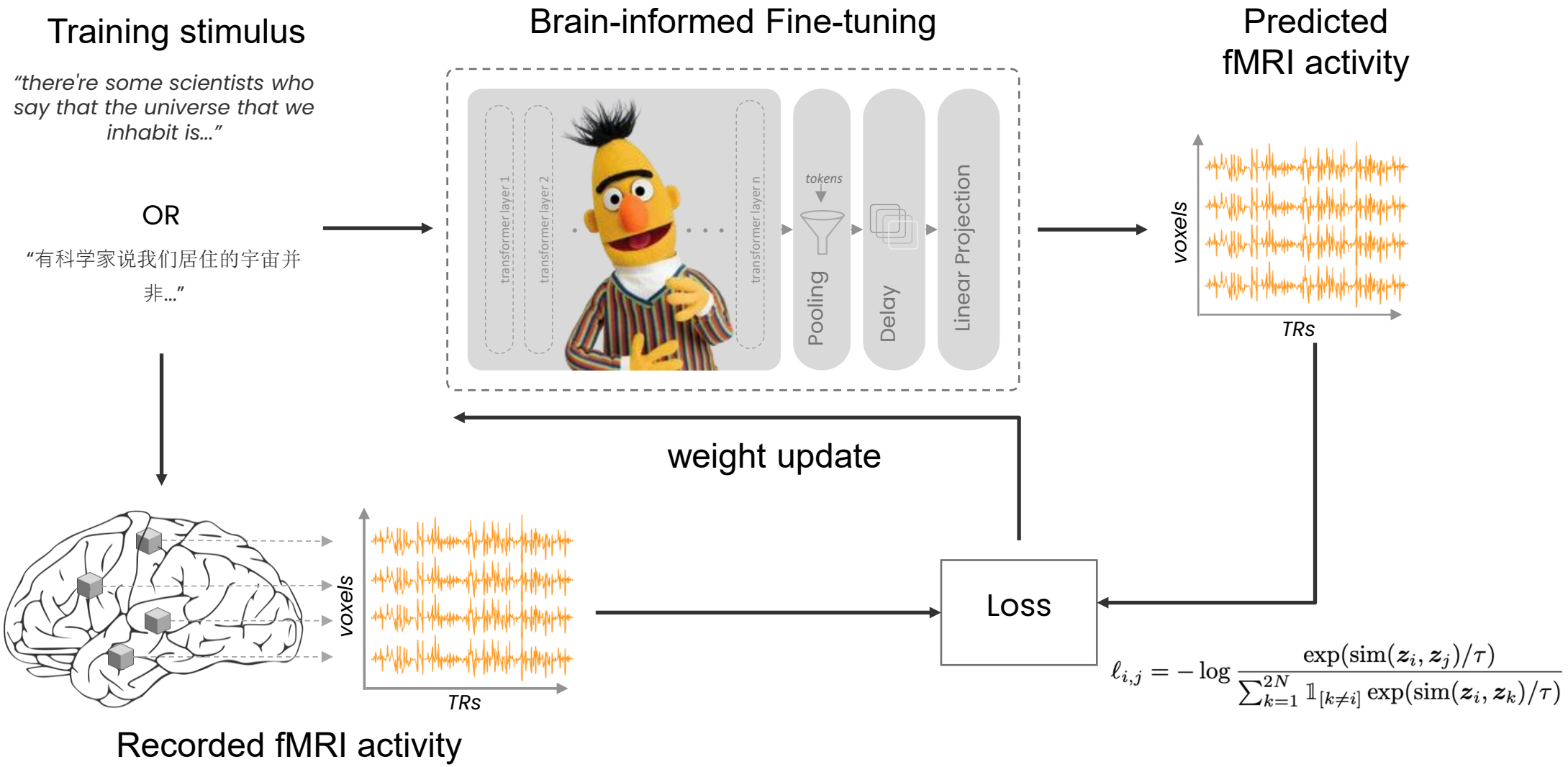


➔ Bilingual language processing relies on shared semantic representations

Can **fine-tuning** language models with **bilingual brain data** elicit multilingual capabilities in them?

Chen, Gong, Tseng, Klein, Gallant, Deniz 2024

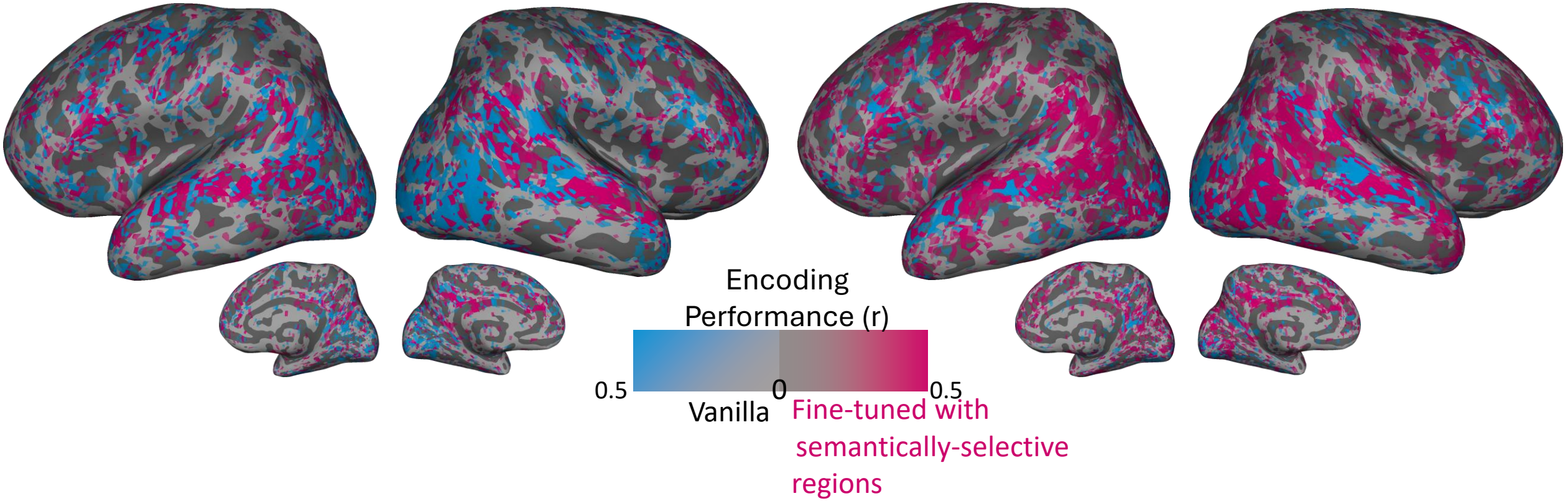
Brain-informed fine-tuning with bilingual brain data



Brain-informed fine-tuning improves brain alignment

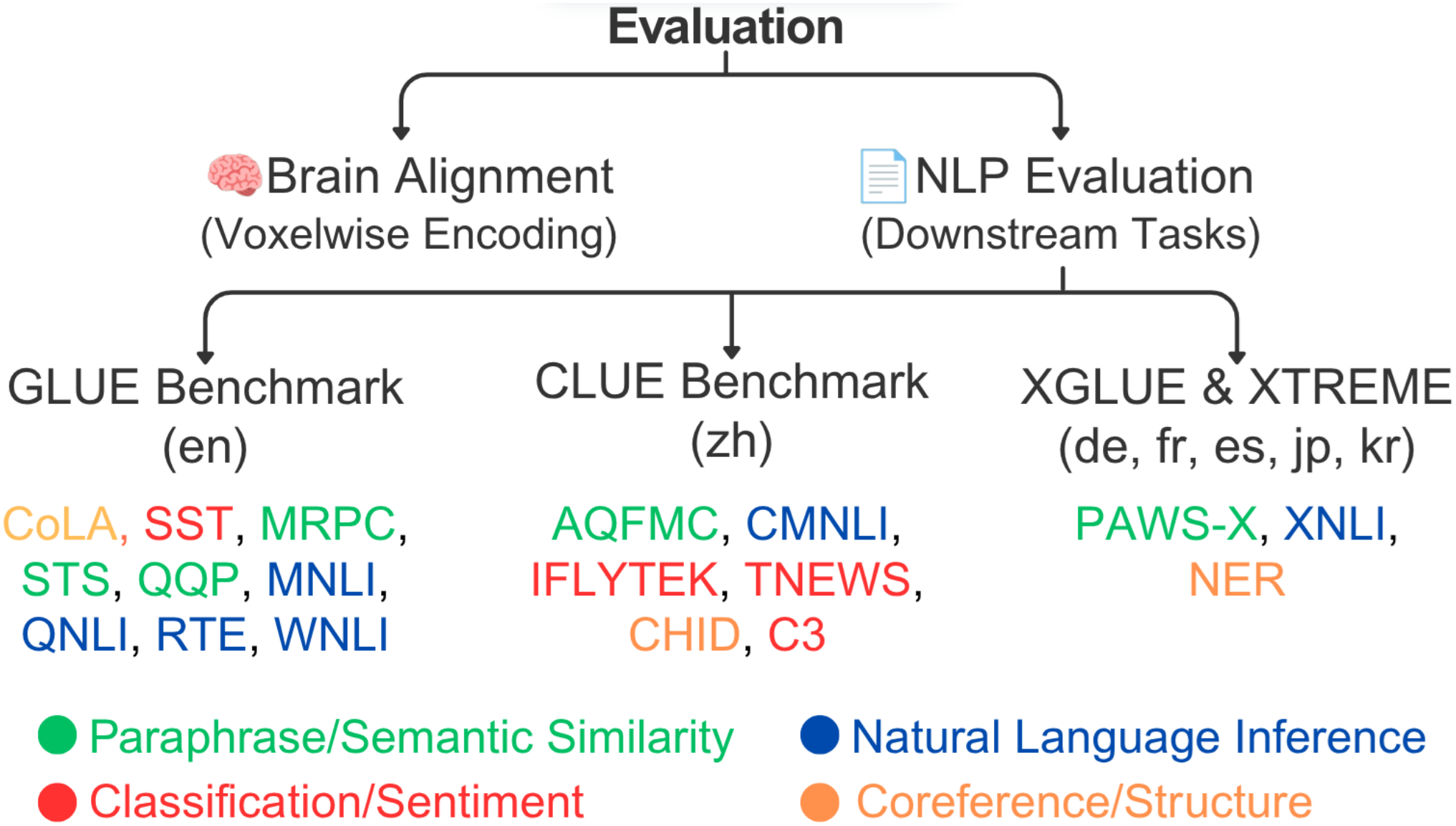
BERT-en fine-tuned with English brain data

BERT-zh fine-tuned with Chinese brain data

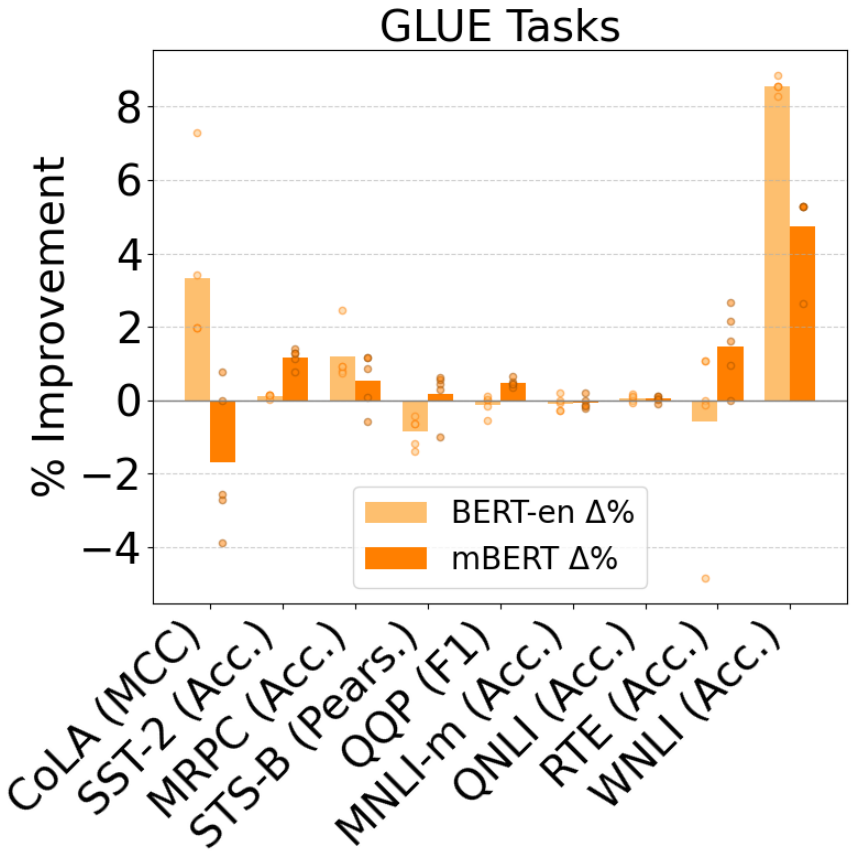


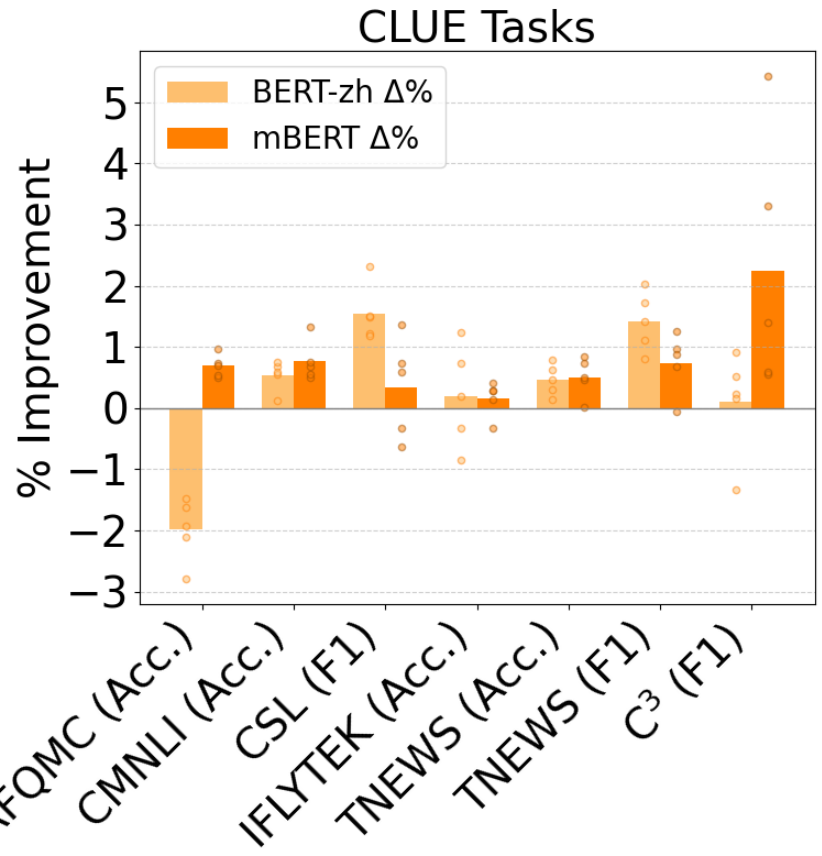
*~**70%** semantic voxels prefer a fine-tuned model over vanilla model*

Evaluating brain-informed fine-tuned models



Fine-tuning and Evaluation in the Same Language





Monolingual model

⬆️ 7/9 on English benchmark (GLUE)

⬆️ 6/7 on Chinese benchmark (CLUE)

Multilingual model

⬆️ 8/9 on English benchmark (GLUE)

⬆️ 7/7 on Chinese benchmark (CLUE)

Conclusions

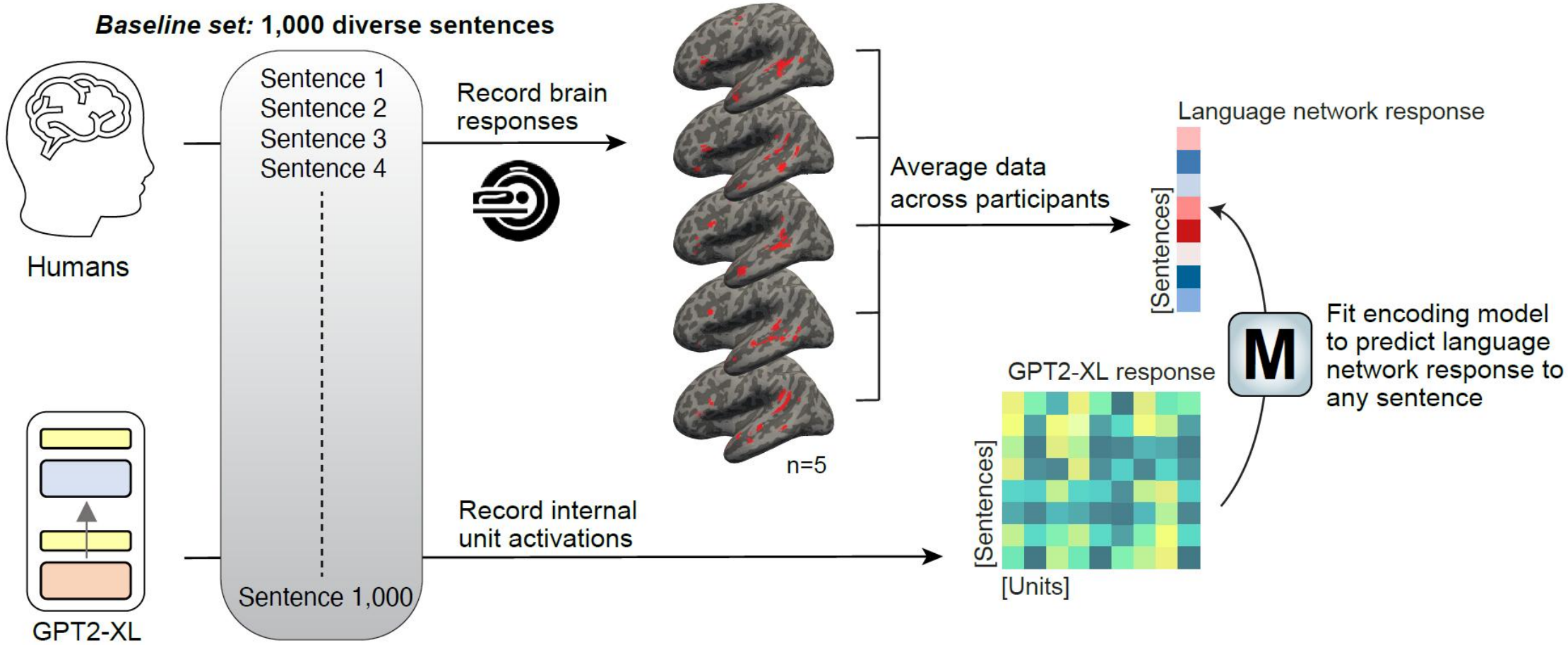
- ➔ First study to perform brain-informed fine-tuning using bilingual brain data.
- ➔ Brain-informed fine-tuning improves
  - brain alignment
  - downstream task performance across within-, cross-, and unseen language settings.
- ➔ Improvements are driven specifically by fine-tuning with bilingual brain data, not brain data in general.
- ➔ Potential of leveraging bilingual brain representations for developing language-agnostic models.

Future Work: Explore which linguistic properties the model captures (e.g., syntax, morphology, discourse) to improve model training and evaluation.

Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- **Brain-based Interpretability and Causal Testing of AI Models** [20 min]
- Brain Decoding [30 min]
- Summary and Future Trends [10 min]

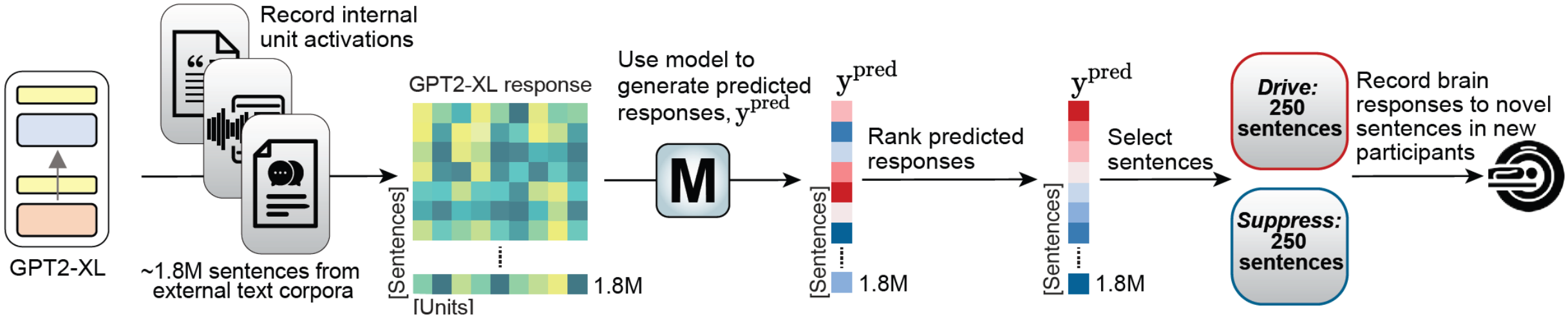
Model-selected sentences control language network responses



Encoding model development: Fitting a predictive model of the language network

Tuckute et al. Nat. Hum. Behav. 2024

Model-selected sentences control language network responses



Encoding model evaluation: Selection of novel drive and suppress sentences

Tuckute et al. Nat. Hum. Behav. 2024

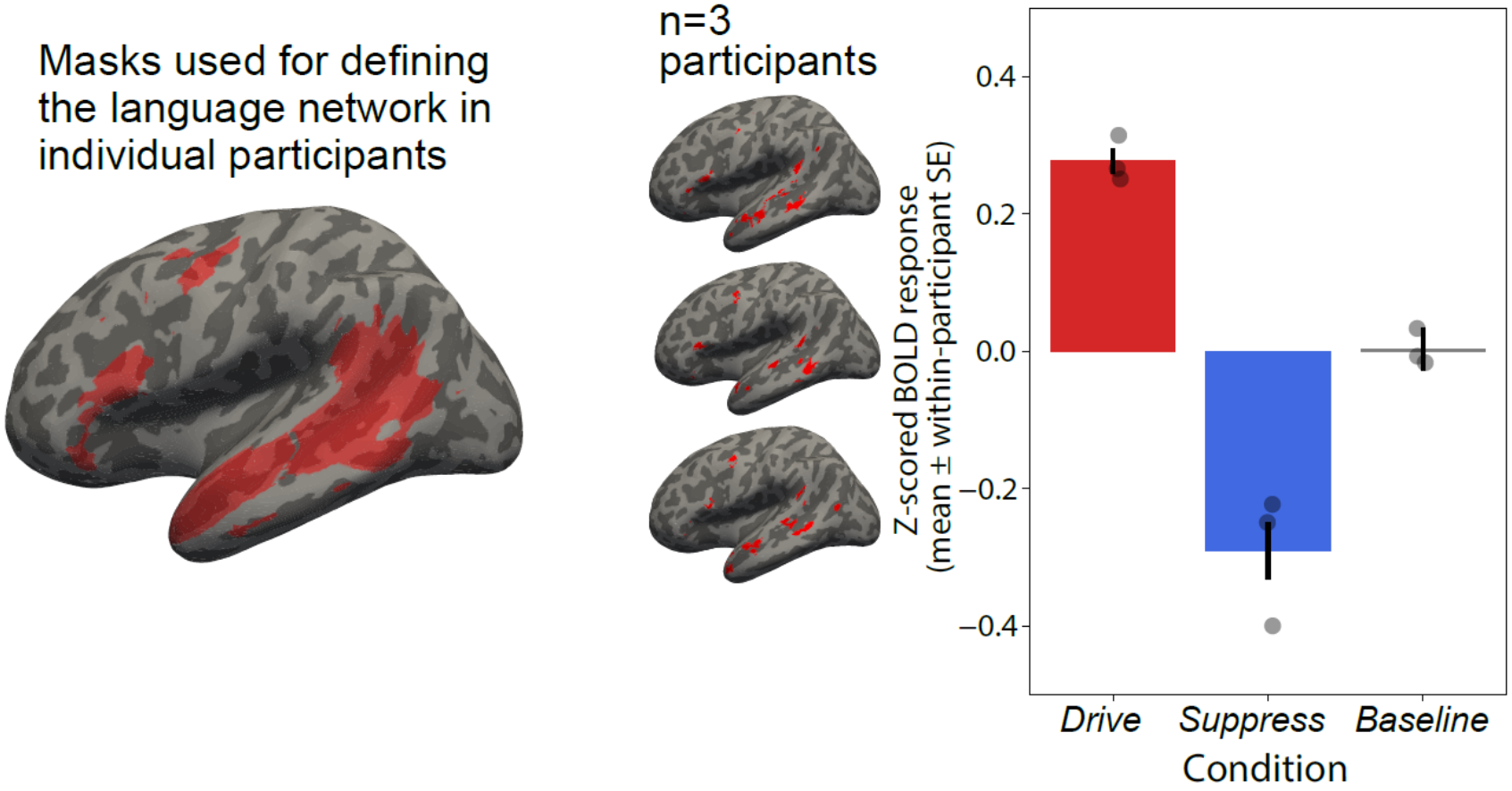
Model-selected sentences control language network responses

Drive	*Suppress*	*Baseline*
Sentences identified to elicit maximal response in the language network	Sentences identified to elicit minimal response in the language network	Sentences sampled from diverse naturalistic text corpora
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.	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?	The judge spoke, breaking the silence. How to create a personalized chart. SEO- short for search engine optimization. Were all the vampire stories true? A king must have an heir. Her wisdom and foresight is evident. Wet hair clinging to her cheeks. He also admitted killing Mrs. Hengesbach..

Sentence Examples from each condition

Tuckute et al. Nat. Hum. Behav. 2024

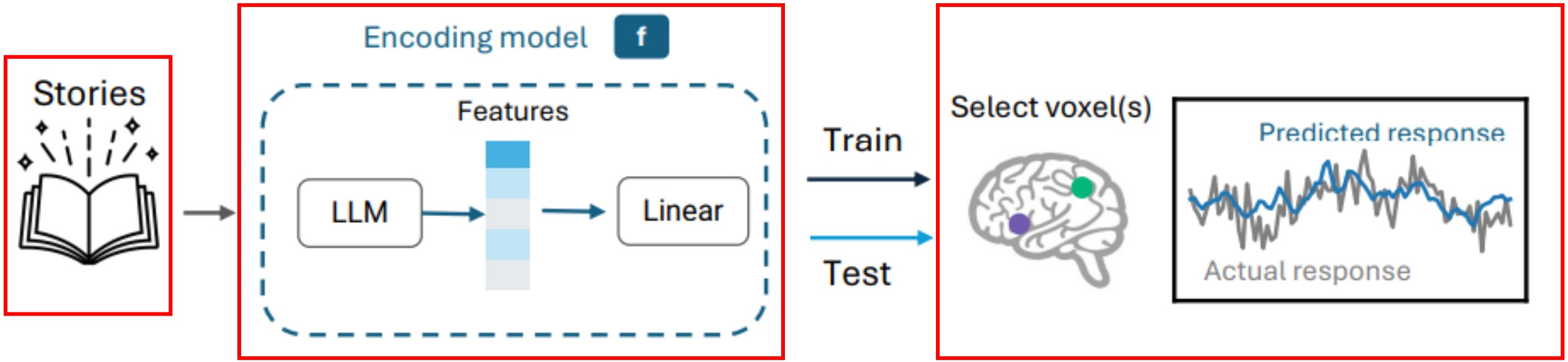
Model-selected sentences control language network responses



LLMs mimic human language (predict brain responses).
Also, enable noninvasive control of neural activity in the language network!

Tuckute et al. Nat. Hum. Behav. 2024

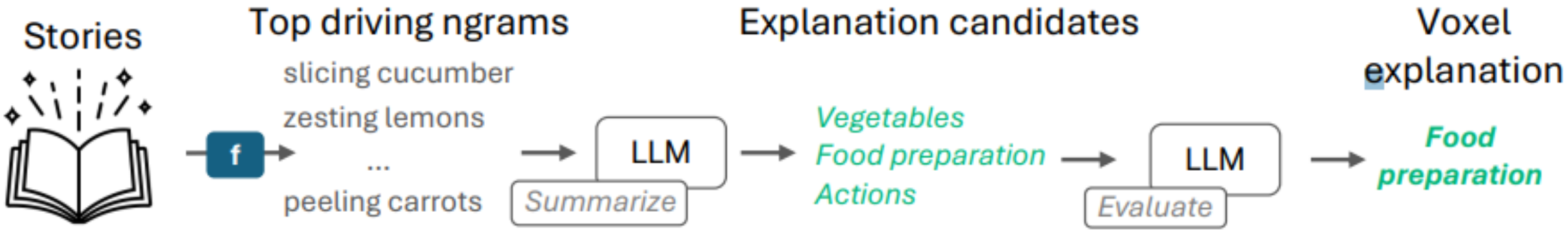
LMs are highly effective at predicting fMRI responses to language stimuli



What features of the language stimulus drive the response in each brain area?

Antonello et al. arXiv 2025
http://arxiv.org/pdf/2410.00812

Generating and evaluating natural-language explanations for single voxels (GEM-V)

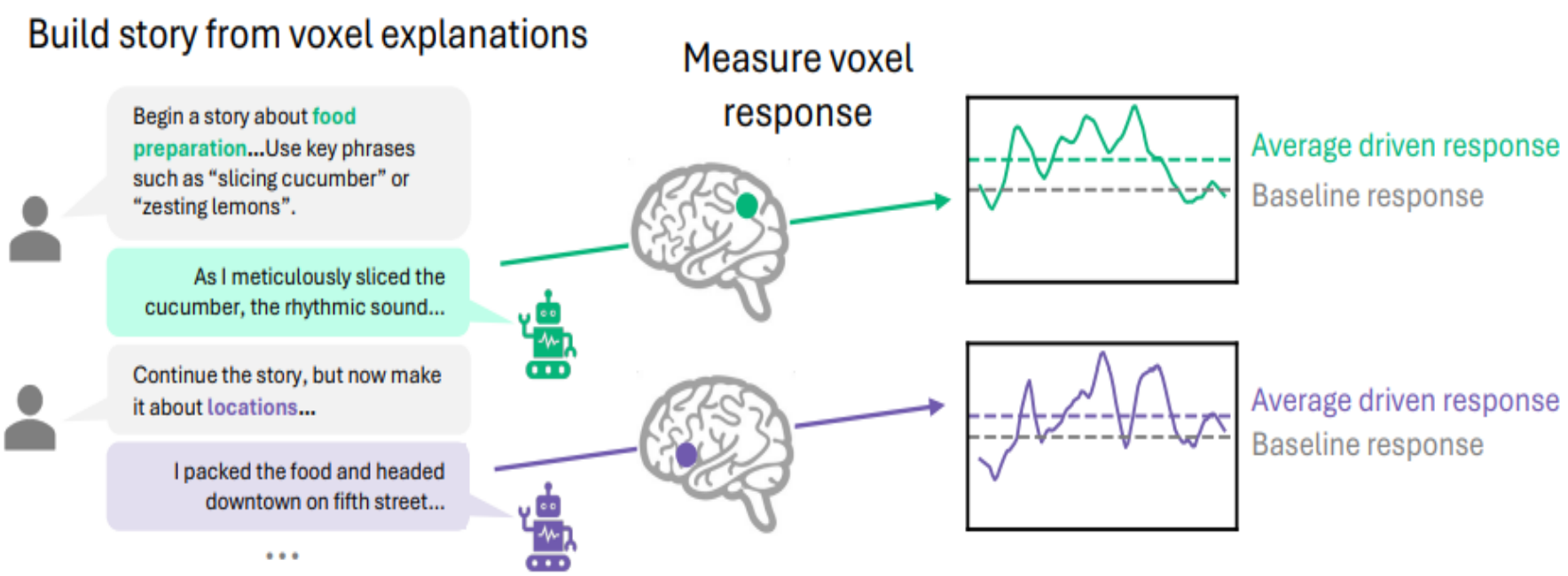


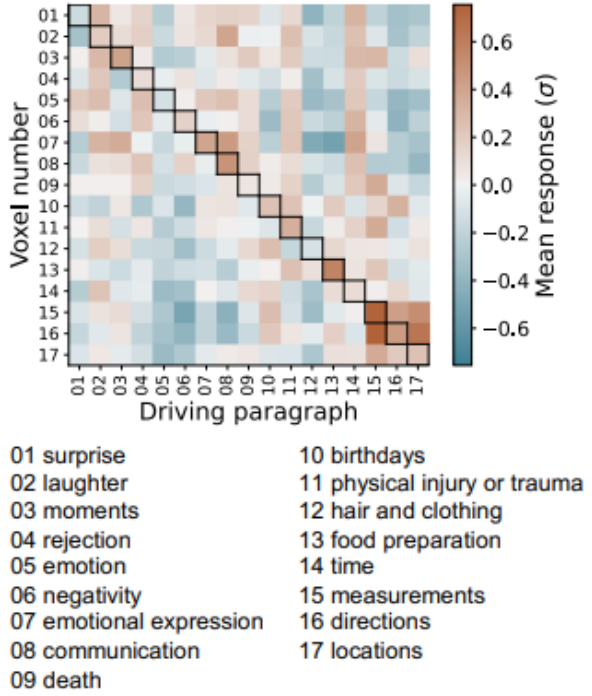
Are the generated explanations causally related to brain activation?

Antonello et al. arXiv 2025
http://arxiv.org/pdf/2410.00812

Build story from voxel explanations

For each subject, constructed stories using LLMs to drive 17 well-modeled voxels with diverse selectivity. These stories were then presented to the subjects in a second fMRI experiment.

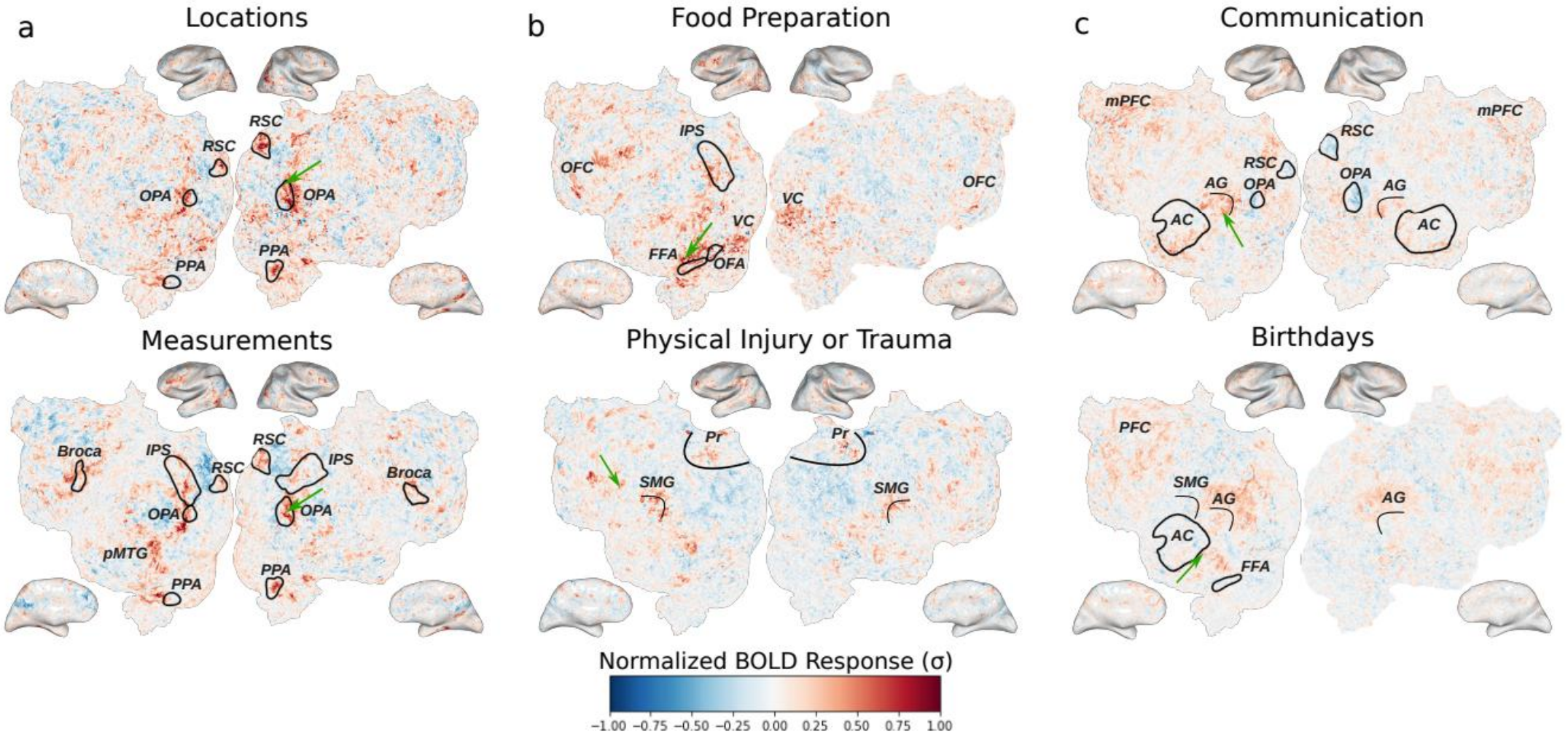




Are the generated explanations causally related to brain activation?

Antonello et al. arXiv 2025
http://arxiv.org/pdf/2410.00812

Responses to driving paragraphs reproduce known semantic contrasts

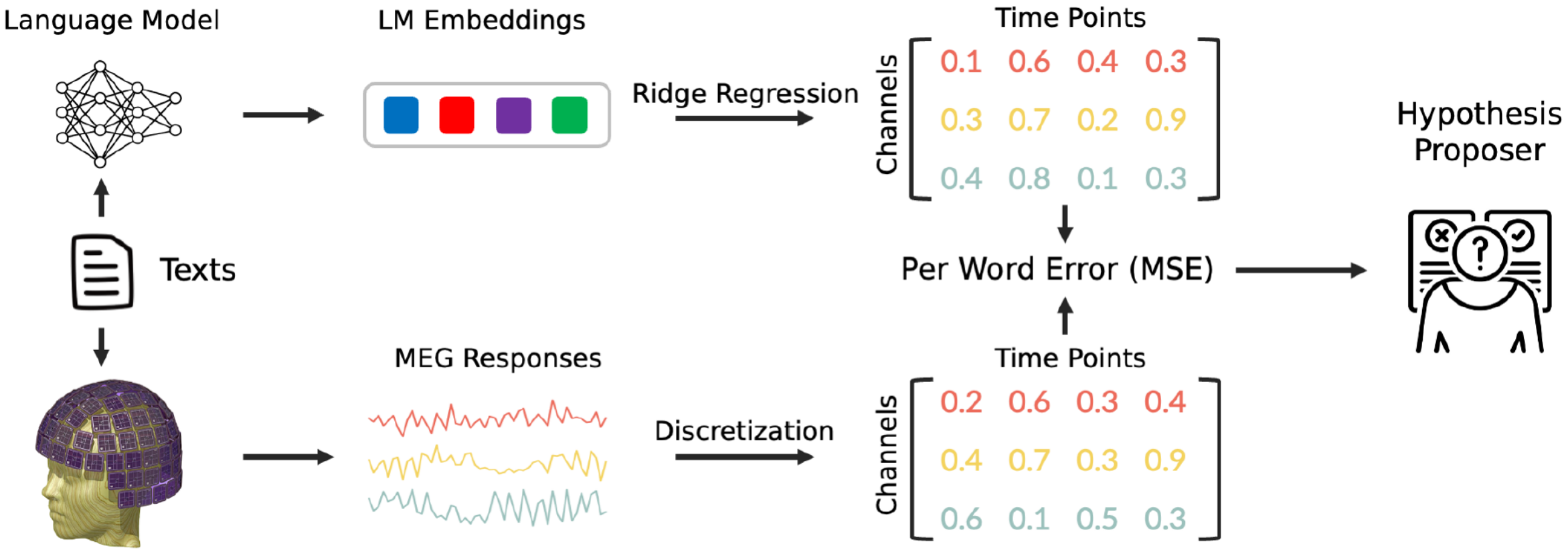


Are the generated explanations causally related to brain activation?

Antonello et al. arXiv 2025
http://arxiv.org/pdf/2410.00812

Divergences Between LMs and Human Brains

Participants either listened to or read narratives. LM embeddings from GPT-2 XL (1.5B) and Llama-2 7B. Ridge regression with cross-validation is used to compute prediction error per word.



NL hypotheses explaining the differences between two text corpora (D0, D1) are generated using a proposer (GPT-3) - verifier (FLAN-T5-XXL) system.

Zhou et al. NeurIPS 2024
https://arxiv.org/pdf/2311.09308

Divergences Between LMs and Human Brains

Validity measures the difference in certainty that the hypothesis is true between the two corpora.
Two topics chosen: **Social/Emotional Intelligence** and **Physical Commonsense**

Table 1: Top 10 hypotheses generated from the best layer of GPT-2 XL on the Harry Potter dataset

Hypothesis	**Validity**	***p*-value**
have a high level of emotional intensity	0.250	0.010
involve complex sentence structures or grammar	0.250	0.015
include emotional language or descriptions	0.238	0.008
have a high level of tension or conflict	0.237	0.023
have characters using body language or non-verbal cues	0.225	0.032
are emotionally charged, making it challenging for language models to accurately interpret the intended tone or sentiment	0.213	0.020
include conflicts between characters	0.200	0.035
have characters interacting with their environment	0.188	0.059
have complex sentence structures	0.175	0.081
have dialogue between characters with varying emotions	0.175	0.022

Divergence between human brains and LMs:
Brain responses not well explained by LMs because of limited representations of these topics

Zhou et al. NeurIPS 2024
https://arxiv.org/pdf/2311.09308

Divergences Between LMs and Human Brains

Each multiple-choice option is concatenated with the question to format it as a language modeling task.

Table 3: Datasets for Fine-Tuning with Sample Questions and Answers (Correct Answer in Bold)

Dataset	Type	Num train	Options	Sample question	Sample answers
Social IQa	Social/Emotion	33.4k	3	Sydney had so much pent up emotion, they burst into tears at work. How would Sydney feel afterwards?	1. affected 2. **like they released their tension** 3. worse
PiQA	Physical	16.1k	2	When boiling butter, when it's ready, you can	1. Pour it onto a plate 2. **Pour it into a jar**

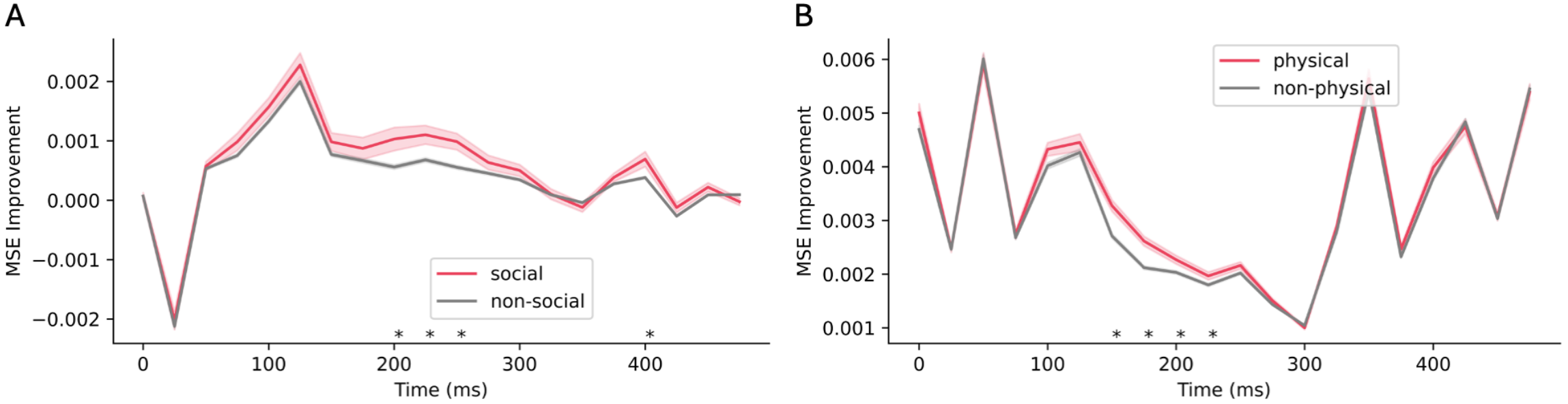
Fine-tuning!

Zhou et al. NeurIPS 2024
https://arxiv.org/pdf/2311.09308

Divergences Between LMs and Human Brains

Can targeted fine-tuning improve LM-brain alignment?

Comparison of improved MSE between **(A) social** and **(B) physical words** and those outside each category evaluated on models fine-tuned on corresponding datasets. Positive values denote lower MSEs in the fine-tuned model.



LMs differ from human language processing in social/emotional intelligence and physical commonsense

Fine-tuning indeed improves alignment for words annotated with that category!

Zhou et al. NeurIPS 2024
https://arxiv.org/pdf/2311.09308

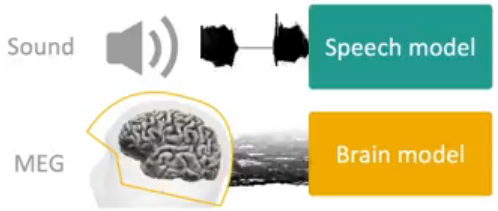
Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- **Brain Decoding** [30 min]
- Summary and Future Trends [10 min]

What is Brain Decoding?

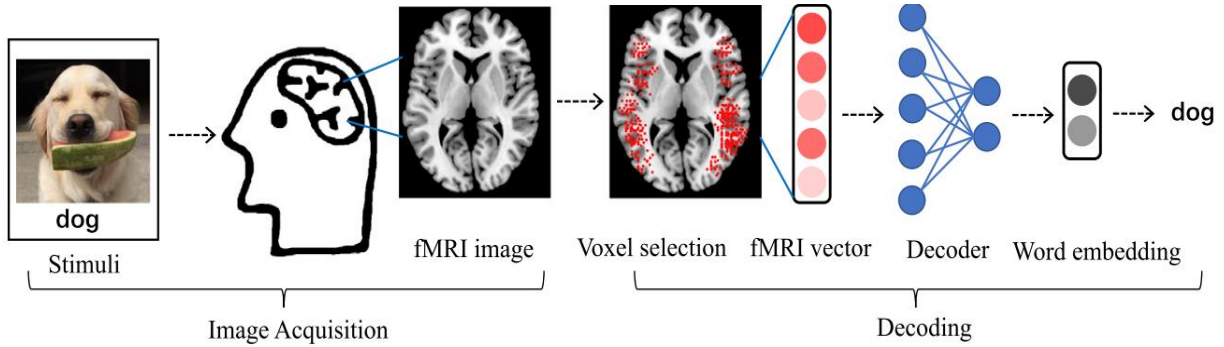
- Can we reconstruct the stimulus, given the brain response?
- Can you read the mind from fMRI responses!
- Or at least tell what the person saw?

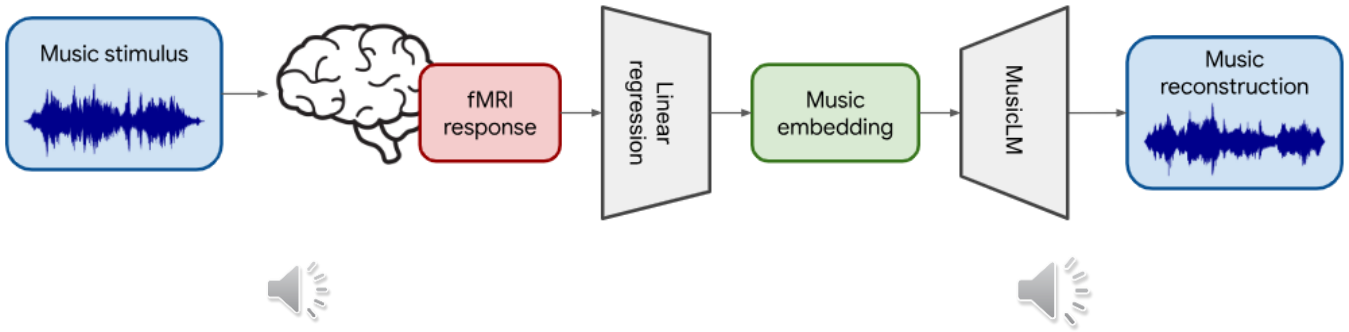
Linguistic Decoding



Decoding speech from non-invasive brain recordings

Défossez, Caucheteux, Rapin, Kabeli & King (2022)
arxiv.org/pdf/2208.12266





Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli & Jean-Rémi King. "Decoding speech perception from non-invasive brain recordings" Nature Machine Intelligence 2023.

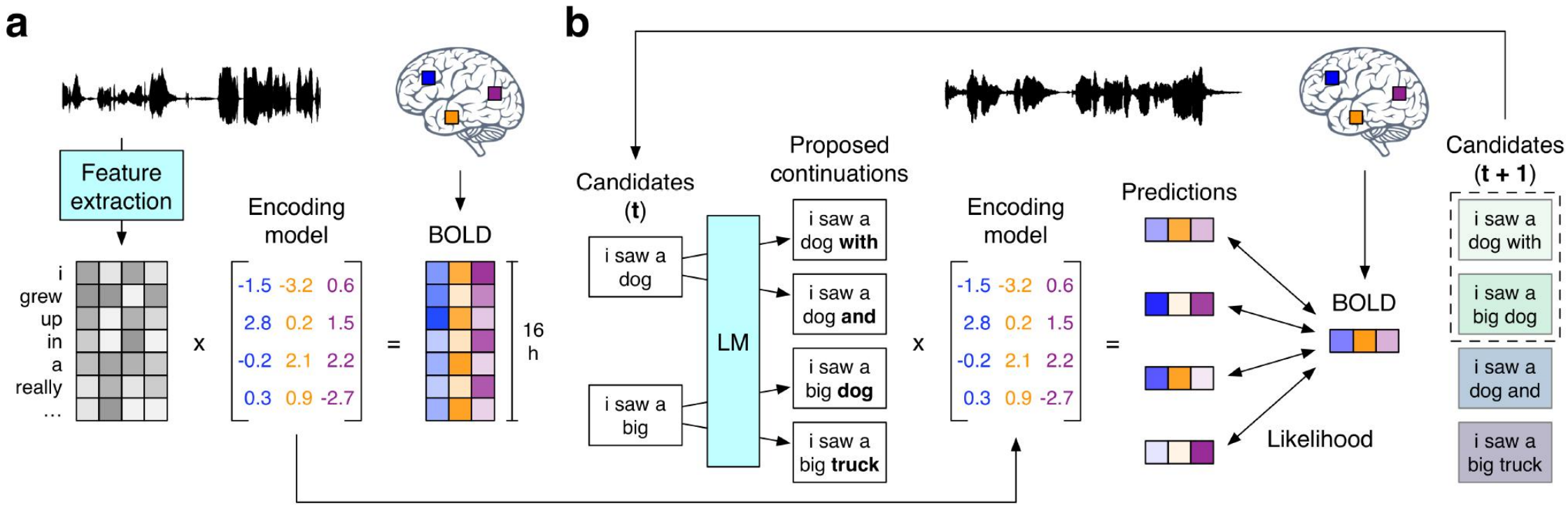
Timo I. Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, Shinji Nishimoto. "Brain2Music: Reconstructing Music from Human Brain Activity" Arxiv 2024.

Outline

- Semantic reconstruction of continuous language from non-invasive brain recordings
- Generative Language reconstruction from non-invasive brain recordings

Continuous Language Decoder

- Stimuli: Moth-Radio-Hour, Short-movie-clips
- Stimulus representation: GPT2 language model
- Brain recording & modality: fMRI, listening

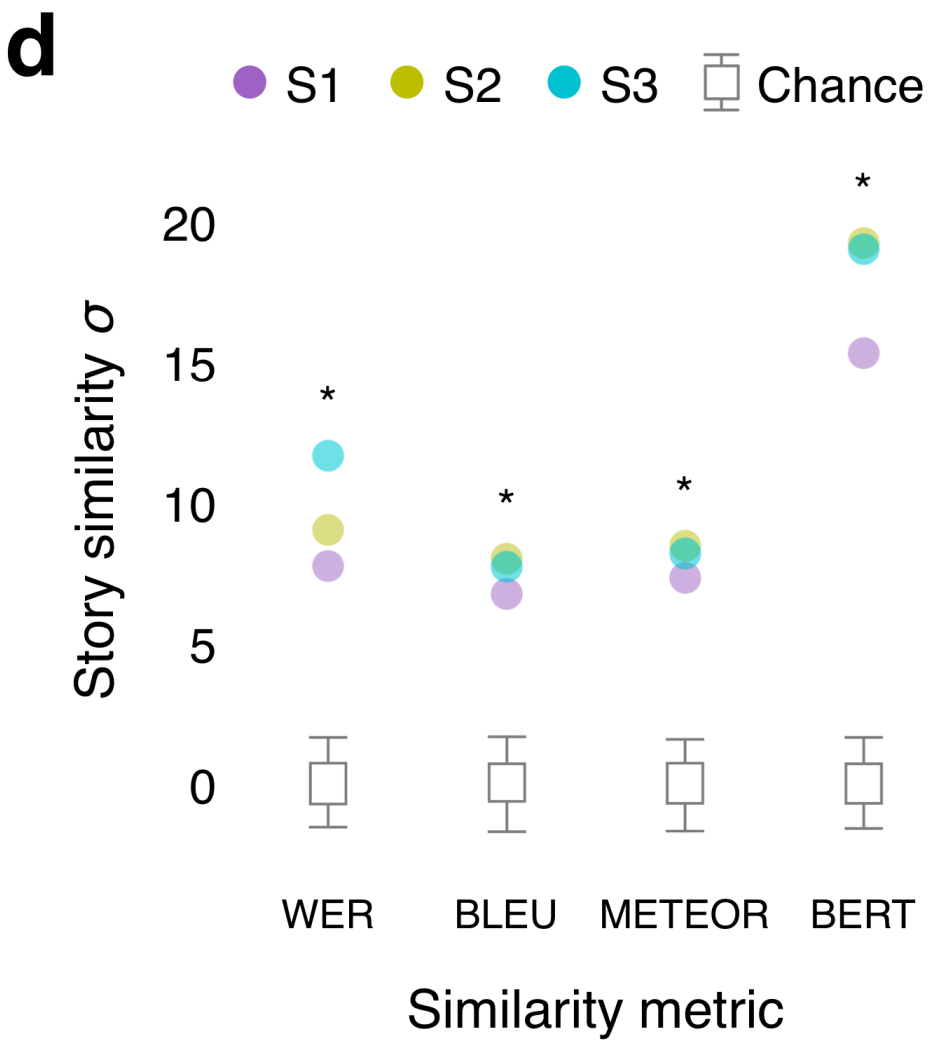


Jerry Tang, Amanda LeBel, Shailee Jain & Alexander G. Huth "Semantic reconstruction of continuous language from non-invasive brain recordings" Nature Neuroscience 2023.

Continuous Language Decoder

Jerry Tang, Amanda LeBel, Shailee Jain & Alexander G. Huth "Semantic reconstruction of continuous language from non-invasive brain recordings" Nature Neuroscience 2023.

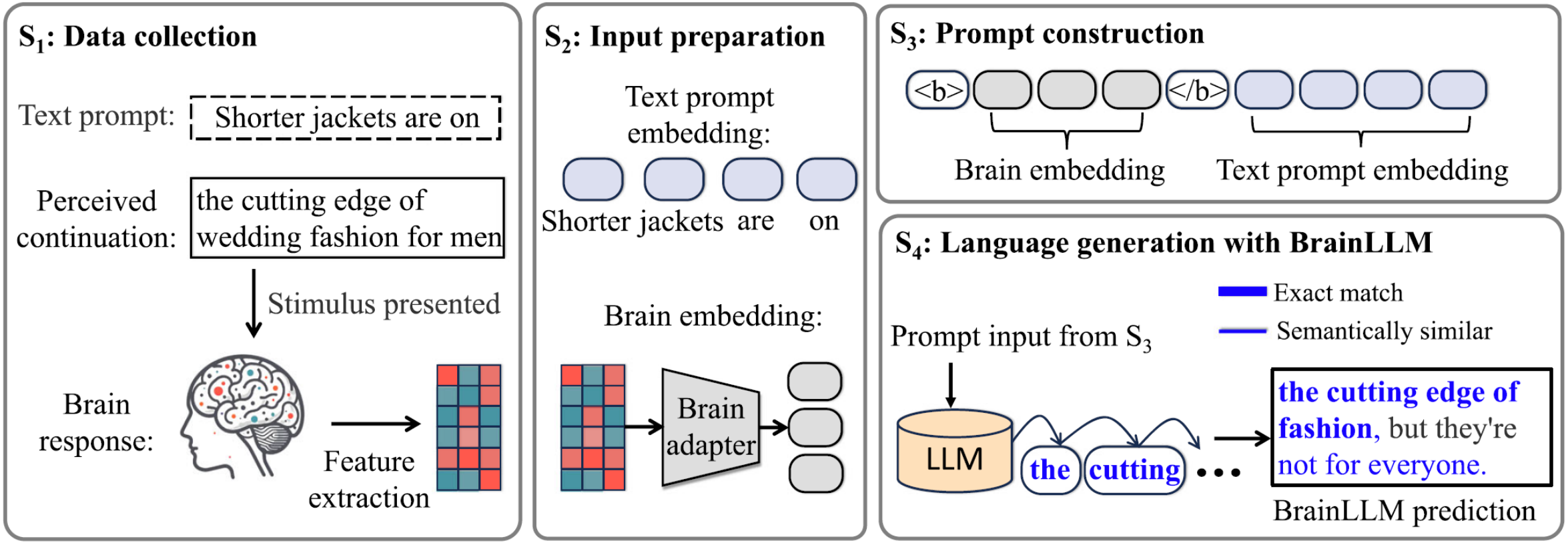
Continuous Language Decoder



Jerry Tang, Amanda LeBel, Shailee Jain & Alexander G. Huth "Semantic reconstruction of continuous language from non-invasive brain recordings" Nature Neuroscience 2023.

Generative Language Decoder

Four main Stages in Language Generation with brain recordings.



BrainLLM

Ye et al. (2025). Generative language reconstruction from brain recordings. Commun Biol. https://doi.org/10.1038/s42003-025-07731-7

Generative Language Decoder

Examples of language generation with BrainLLM and its controls (PerBrainLLM).
Text in blue and bold indicates that the generated content and the ground truth (perceived continuation) are manually annotated as semantically similar and an exact match, respectively.

Text prompt	Continuation	BrainLLM prediction	Control prediction
Shorter jackets are on	the cutting edge of wedding fashion for men	**the cutting edge of fashion**, but they're not for everyone	their way out **of** style, but they're still popular.
A wall is a	solid structure that defines and sometimes protects an area	**structure that defines and sometimes protects an area**	vertical **structure** made of stone, brick or concrete
I'm just stand- -ing there like	the proverbial deer in headlights	a **deer in the headlights**	an idiot
she was like petite I could have	folded her up and put her my pocket	picked **her up** with one hand	driven her to work every day

BrainLLM can be used in an auto-regressive manner to **reconstruct a 10 min-long language stimulus**!!
Brain language interfaces for mapping functional representations of language perception in the brain.

Ye et al. (2025). Generative language reconstruction from brain recordings. Commun Biol.
https://doi.org/10.1038/s42003-025-07731-7

Agenda

- Introduction to the tutorial [10 min]
- Introduction to Brain Encoding and Decoding [50 min]
- Brain Encoding: Scaling Laws, Multilinguality, Multimodal and Instruction-tuned Models [60 min]
- Coffee Break & Networking [30 min]
- Brain-informed Fine-tuning of Language Models [30 min]
- Brain-based Interpretability and Causal Testing of AI Models [20 min]
- Brain Decoding [30 min]
- **Summary and Future Trends** [10 min]

Summary & Future Directions

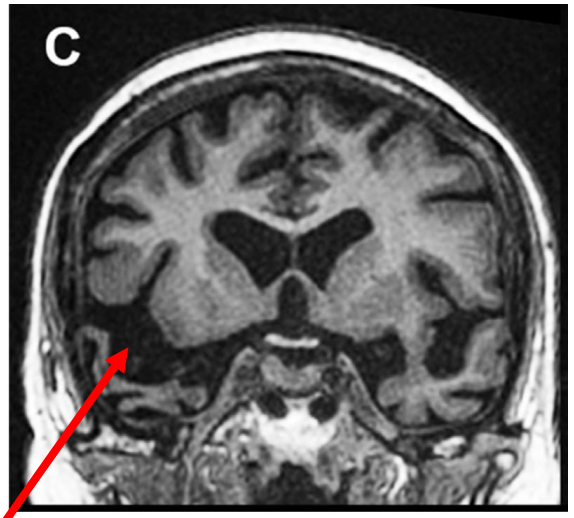
- Exciting times: publicly accessible neuroimaging data of various tasks starting to be avaliable now!
  - Opportunities:
    - Data ahead of theory, so it's an open field for theoretical and methodological innovation!
    - Encoding models can be interpreted as process models constraining brain-computational theories (Kriegeskorte and Douglas, 2019).
    - Decoding models serve as a test for the presence of information in neural responses (Karamolegkou et al., 2023)
    - Decoding is relevant for cognitive neuroscientists interested in how semantic information is represented in the brain.
    - Computational linguists are interested in the cognitive plausibility of distributional models. (Minnema & Herbelot, ACL 2019)
    - DL is helpful in uncovering patterns in brain responses and may lead to theories of information organization in the brain.
  - Challenges:
    - Neuroimaging data is more complex, noisy as compared to classical datasets used by DL researchers
    - The differences between the models in terms of **architectural variability** and **variability in pretraining methods** have an impact on the outcomes. More tightly controlled comparisons are required to better isolate the effects of these factors.
    - Disentangling contributions of multiple sources of info impacting brain predictions

DNNs & Brain Damage

- DL models of encoding and decoding have not yet been put through the brain-damage experiments. Ex. Semantic Dementia

Animal habitat task.
The patient is asked:
Where would you find this?

Stimulus	Response
	I don't know
	In the house. It's a dog
	Outside the house. There are lots of them. They fly about.
	In water. It's got a bushy tail so it's good at swimming.
	In the house. It's somebody's son.



Rt Ant Temporal Lobe Damage (Patient 8)

Stimulus	Response
FROG	In water
COW	On a farm
DUCK	On ponds. I see them on the river when I go walking.
SQUIRREL	In the woods, in the country. They are wild.
MONKEY	In trees, in Africa.

Do DL Models exhibit such degradation with damage to units?

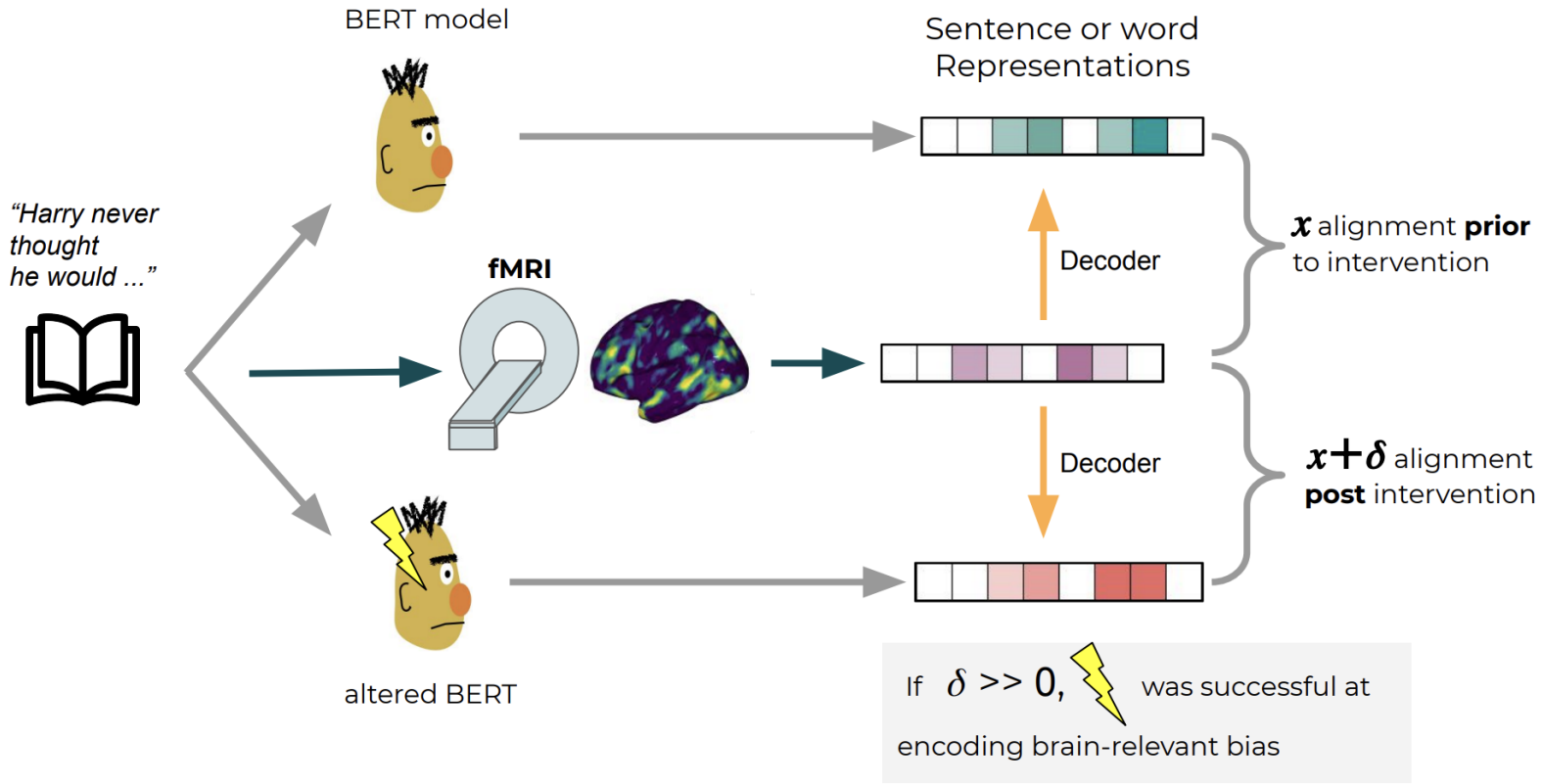
Snowden, Harris, Thompson, Kobylecki, Jones, Richardson, Neary (2018). Semantic dementia and the left and right temporal lobes, *Cortex*, 107(188-203). https://doi.org/10.1016/j.cortex.2017.08.024.

DNNs as a model organism

What information is necessary or sufficient to perform a task? To predict brain activity?

Model *organism*, and not simply a model, because DNNs have evolved separately from the human brain

Model organisms allow for direct interventions, which cannot be done in humans



Abdou, Mostafa, Ana Valeria González, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard. "Does injecting linguistic structure into language models lead to better alignment with brain recordings?." *arXiv preprint arXiv:2101.12608* (2021).

References

- Nishimoto, Shinji, et al. "Reconstructing visual experiences from brain activity evoked by natural movies." *Current biology* 21.19 (2011): 1641-1646.
- Anumanchipalli, Gopala K., Josh Chartier, and Edward F. Chang. "Speech synthesis from neural decoding of spoken sentences." Nature 568.7753 (2019): 493-498.
- Schrimpf, Martin, et al. "The neural architecture of language: Integrative modeling converges on predictive processing." Proceedings of the National Academy of Sciences 118.45 (2021): e2105646118.
- Wehbe, Leila, et al. "Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses." PloS one 9.11 (2014): e112575.

Thanks!

- Questions
  - Subba.Oota@adialab.ae
  - tanchak@iitd.ac.in
  - gmanish@microsoft.com
  - raju.bapi@iiit.ac.in
- Connect with us:
  - https://www.linkedin.com/in/subba-reddy-oota-11a91254/
  - https://tanmoychak.com/
  - http://aka.ms/manishgupta, https://sites.google.com/view/manishg/
  - https://sites.google.com/view/bccl-iiith/home

A big thank you!




Tutorial, Code and Material:

Material from AAAI 2026 Tutorial would be uploaded soon!

(Past): Deep Learning for Brain Encoding and Decoding, Cogsci-2022

https://tinyurl.com/DL4Brain

(Past): Language and the Brain: Deep Learning for Brain Encoding and Decoding, IJCNN 2023

https://tinyurl.com/DLBrainIJCNN2023

(Past): Deep Neural Networks and Brain Alignment: Brain Encoding and Decoding, IJCAI 2023

https://tinyurl.com/DLBrainIJCAI2023