# Language Models and Brain Alignment: Brain Encoding and Decoding


Subba Reddy Oota[1,2,3], Raju S. Bapi[4]

[1]Inria Bordeaux, France; [2]MPI for Software Systems, Germany; [3]TU Berlin, Germany,
[4]IIIT Hyderabad, India;

Subba.reddy.oota@tu-berlin.de, raju.bapi@iiit.ac.in


# Agenda

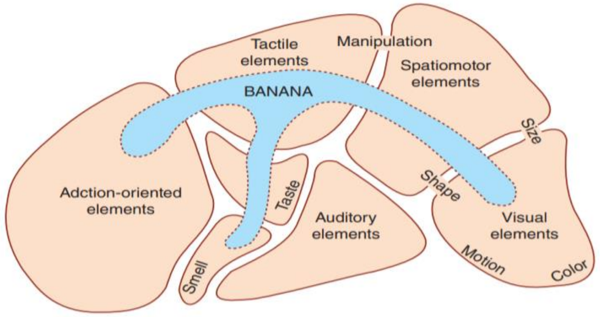
- Neuro-AI alignment: Introduction [1 hour 30 min]
  - Introduction to Brain encoding and decoding [30 min]
  - Types of Brain Recordings [15 min]
  - Types of Stimulus Representations [15 min]
  - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
  - Linguistic Brain Encoding [60 min]
    - Encoding schema
    - Pretrained language models and brain alignment
    - Challenges in using DL for cognitive science
  - Linguistic Brain Decoding [15 min]
  - Multimodal Brain Encoding [15 min]

# Agenda

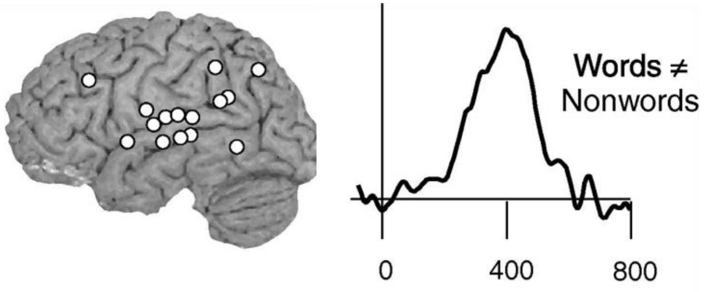
- Neuro-AI alignment: Introduction [1 hour 30 min]
  - Introduction to Brain encoding and decoding [30 min]
  - Types of Brain Recordings [15 min]
  - Types of Stimulus Representations [15 min]
  - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
  - Linguistic Brain Encoding [60 min]
    - Encoding schema
    - Pretrained language models and brain alignment
    - Task-based language models and brain alignment
    - Disentangling Syntax and Semantics
  - Linguistic Brain Decoding [15 min]
  - Multimodal Brain Encoding [15 min]

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

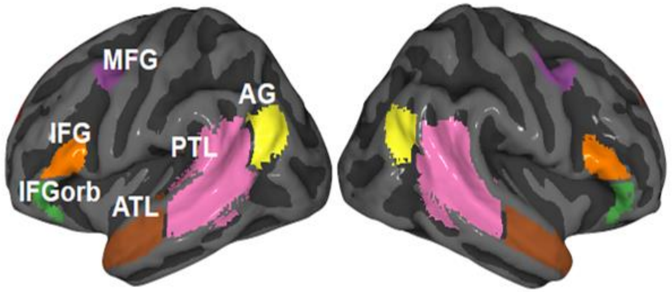
What

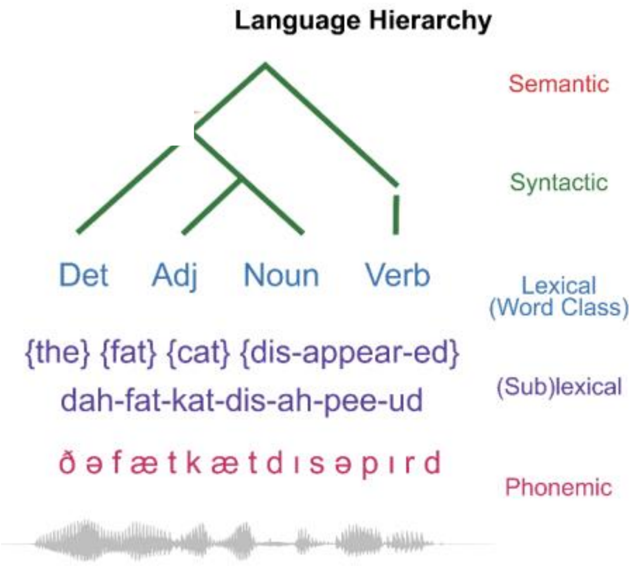



When




Where



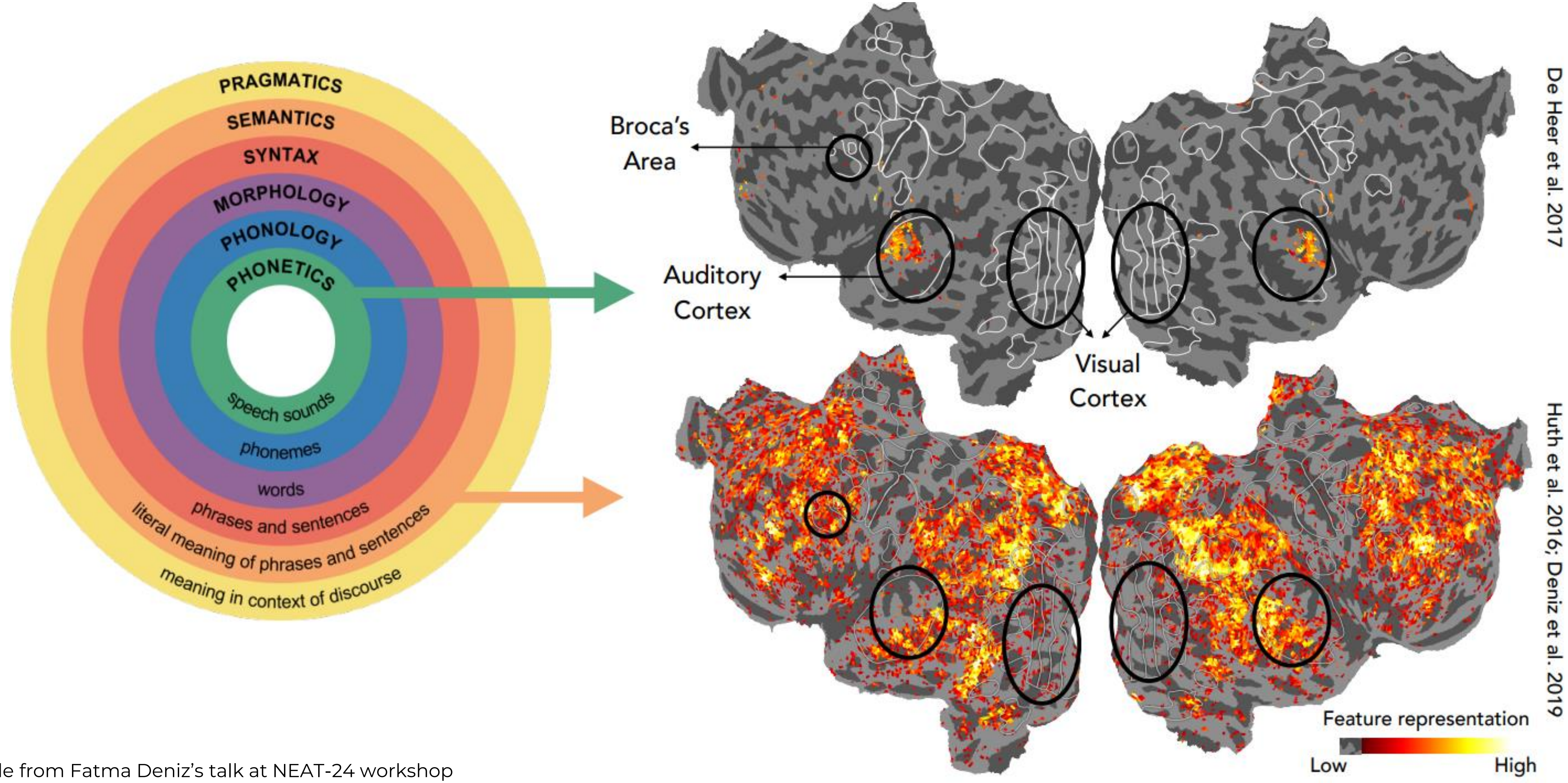




How

[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

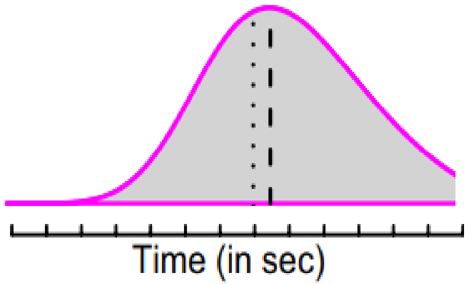
- How the human brain computes and encodes syntactic structures?
  - **Syntax:** how do words structurally combine to form sentences and meaning?

Controlled experiments

Structured complex sentence

I believe that you should accept the proposal of your new associate

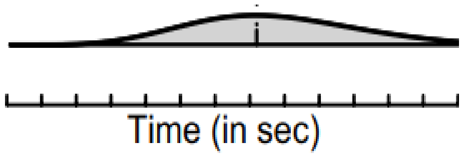
**BOLD response**




Word lists

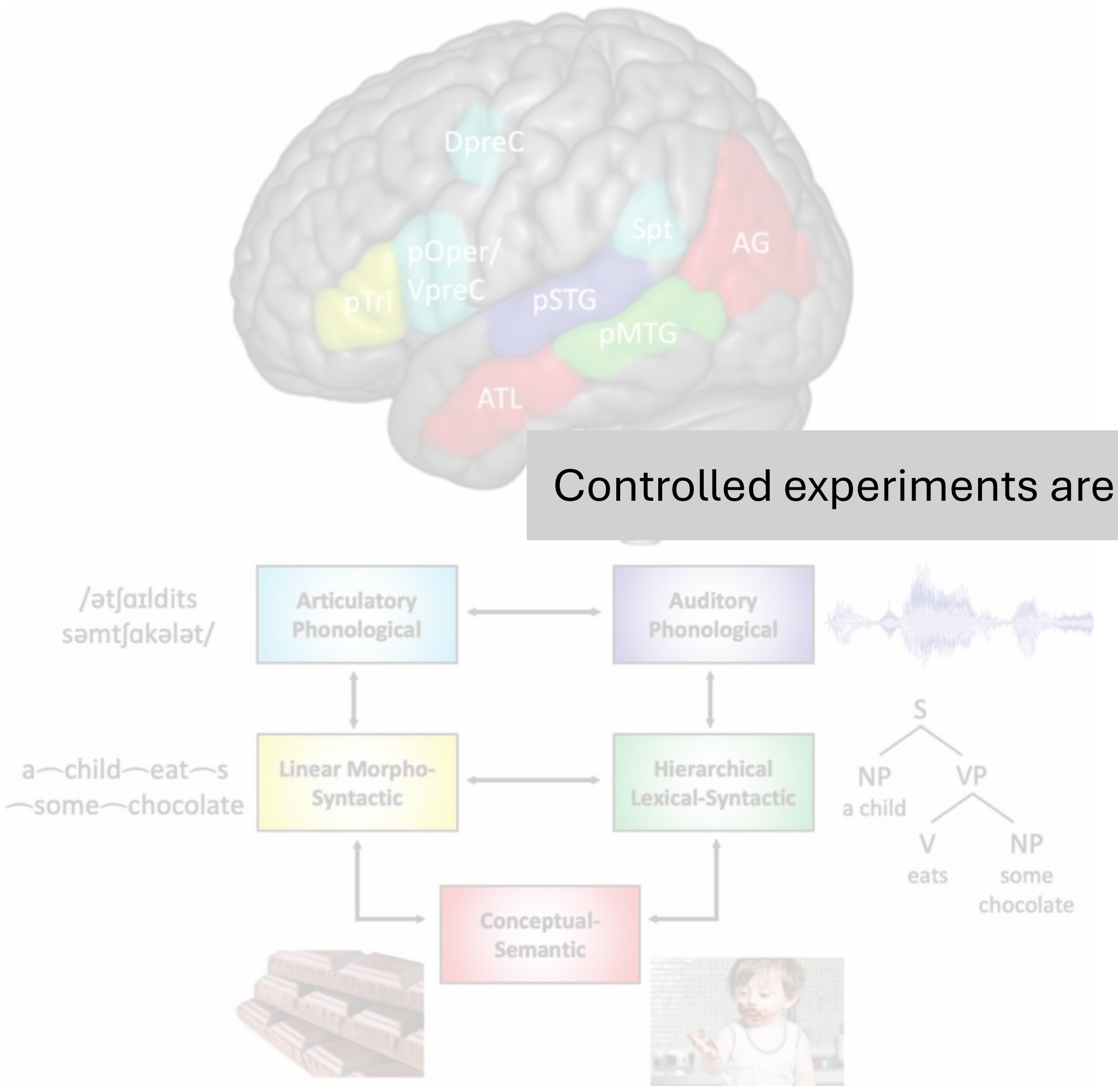
thing very tree where of watching copy tensed they states heart plus

**BOLD response**




Pallier et al. 2011

# Language organization in the brain



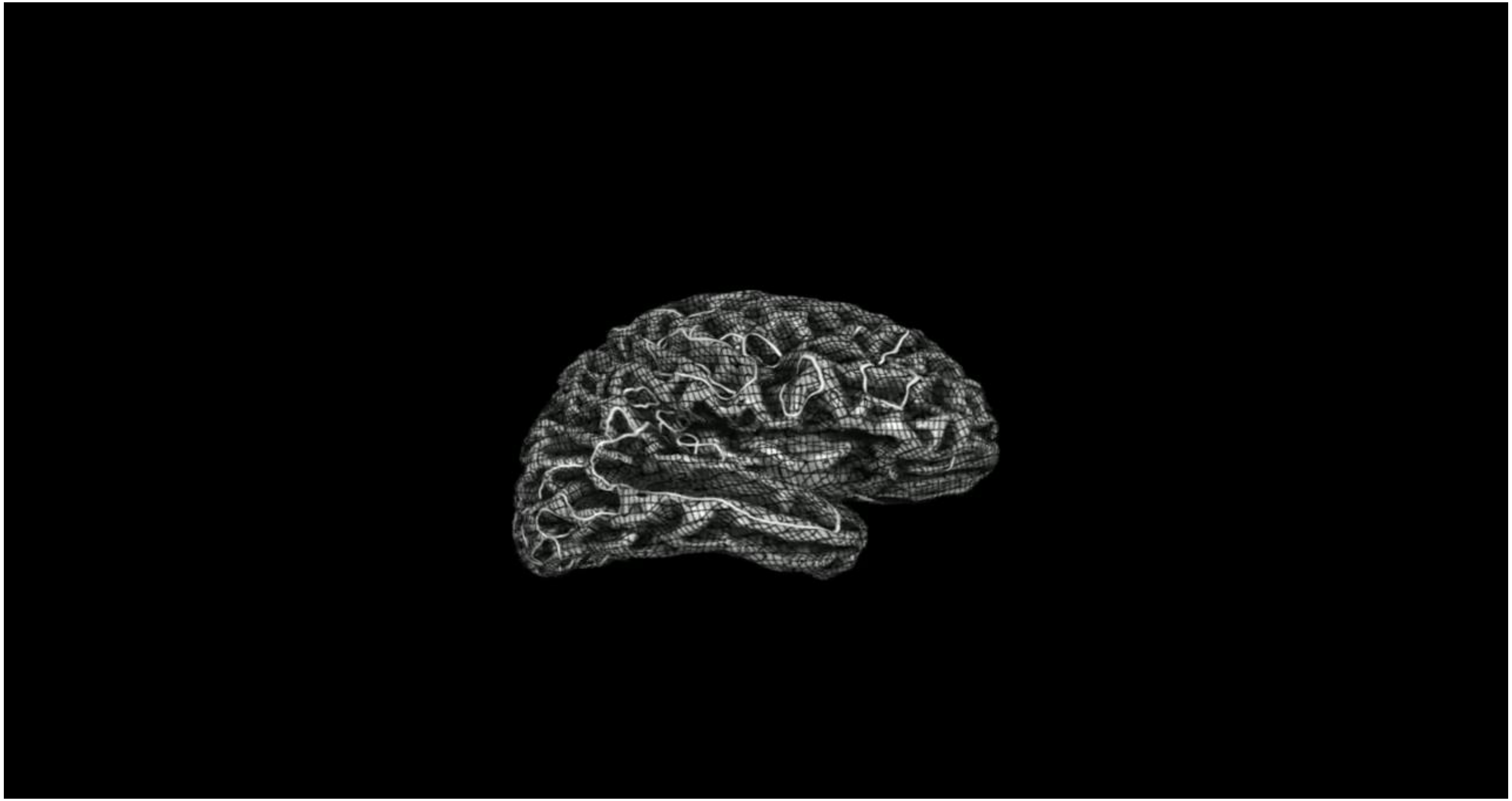

Controlled experiments are task-based and not ecological

…uage 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

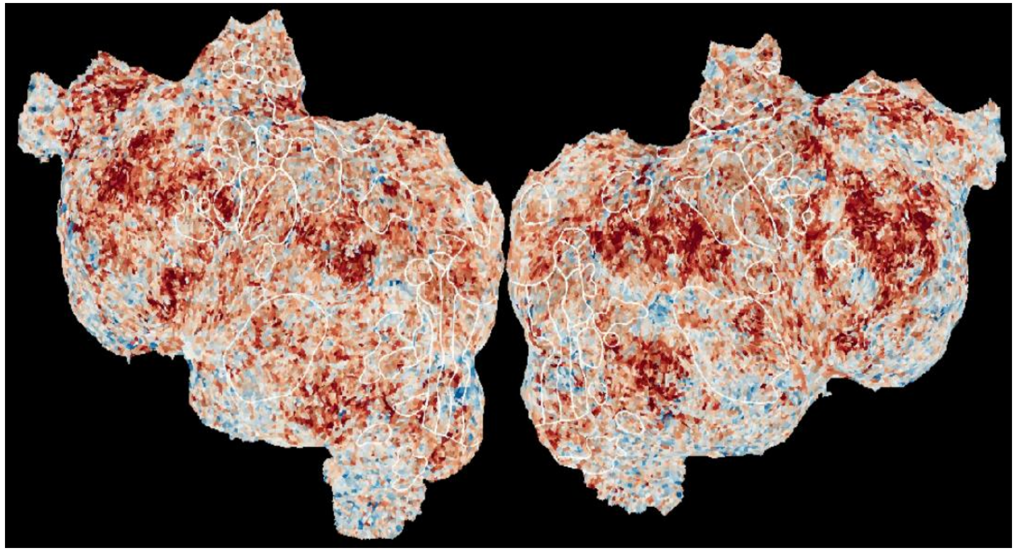
# What are we talking about when we talk about "mapping stimulus to the human brain"

How do we **perceive** the words?

Do **representations differ** when you read a book in **different languages**?

Do **concept** representations differ across **modalities**?

Where in the brain is **word meaning** represented?



Do **representations differ** when you **read or listen to a book**?

How does the brain combine multiple words across **different timescales** ?

(words —> sentences —> paragraphs)

Do **representations differ** when we learn **new languages**?

What is the **shared and unique information** explained by each modality?

# Deep learning models enable data-driven encoding models for naturalistic stimuli

DeepMind’s New AI Taught Itself to Be the World’s Greatest Go Player
Singularity Hub

Meet GPT-3. It Has Learned to Code (and Blog and Argue)
The New York Times

# 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

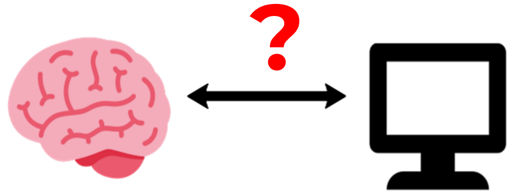
Lebel et al. 2022

Nastase et al. 2021

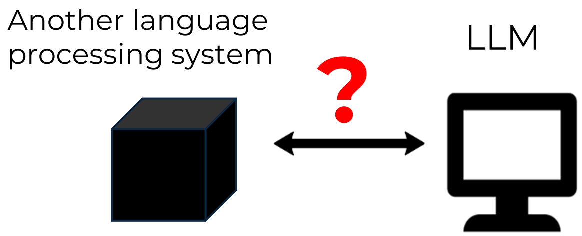
Li et al. 2022

Zhang et al. 2021

# How closely do LLM capabilities relate to those of the human brain?



1: methods to estimate alignment




2: neuroscience background

3: works on alignment between LLMs and brains, and reasons for alignment

4: works on reasons for alignment, and on improving alignment

**Questions very much encouraged!!**

# Agenda

- Neuro-AI alignment: Introduction [1 hour 30 min]
  - Introduction to Brain encoding and decoding [30 min]
  - Types of Brain Recordings [15 min]
  - Types of Stimulus Representations [15 min]
  - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
  - Linguistic Brain Encoding [60 min]
    - Encoding schema
    - Pretrained language models and brain alignment
    - Challenges in using DL for cognitive science
  - Linguistic Brain Decoding [15 min]
  - Multimodal Brain Encoding [15 min]

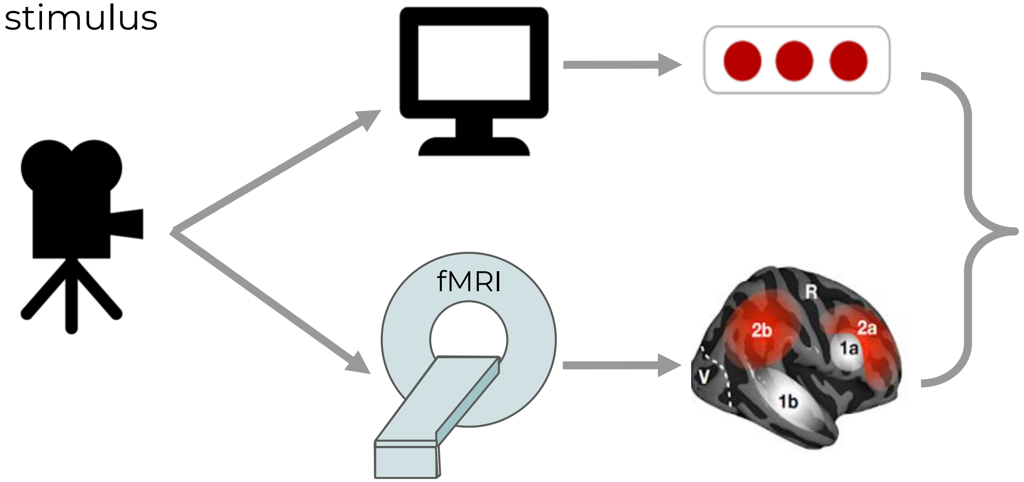
# Deep neural networks and brain alignment: brain encoding and decoding

Multimodal ecological stimulus

Deep learning system

A priori locations in DL system and brain

**Data-driven encoding model**



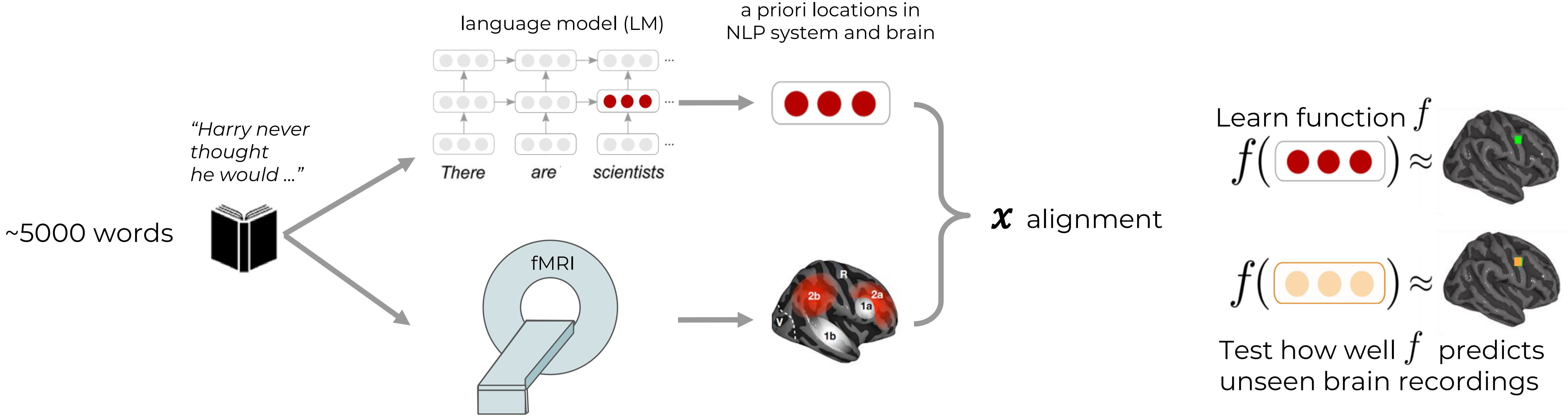

how are they related?

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



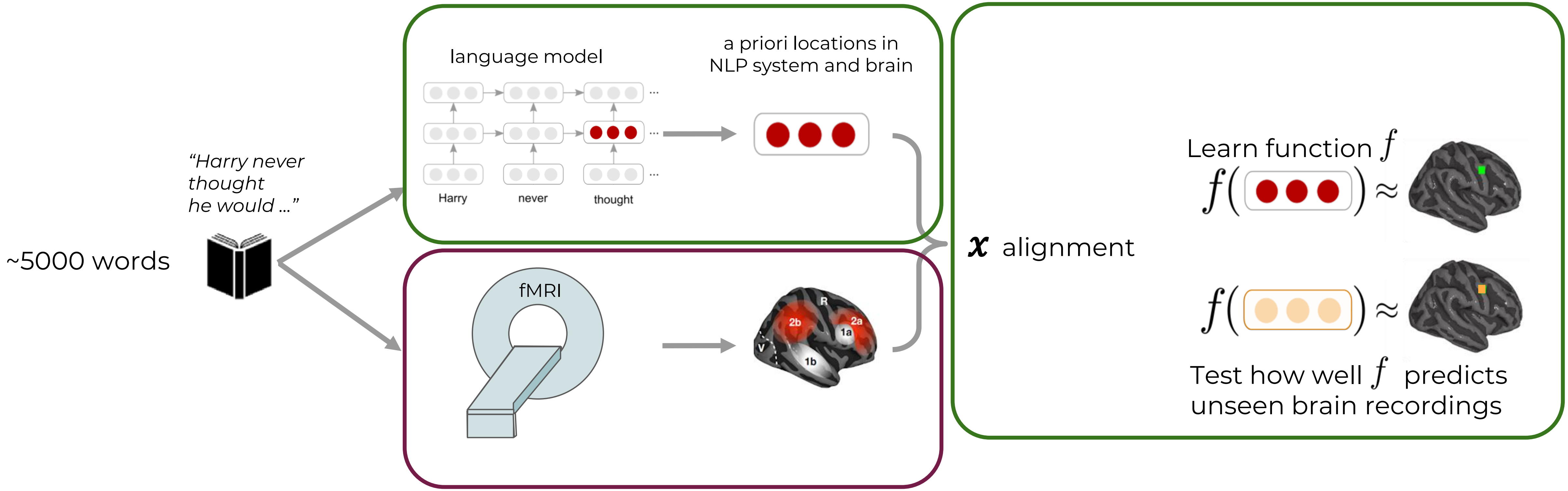

Brain alignment of a LM ⇒ how similar its representations are to a human brain's

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

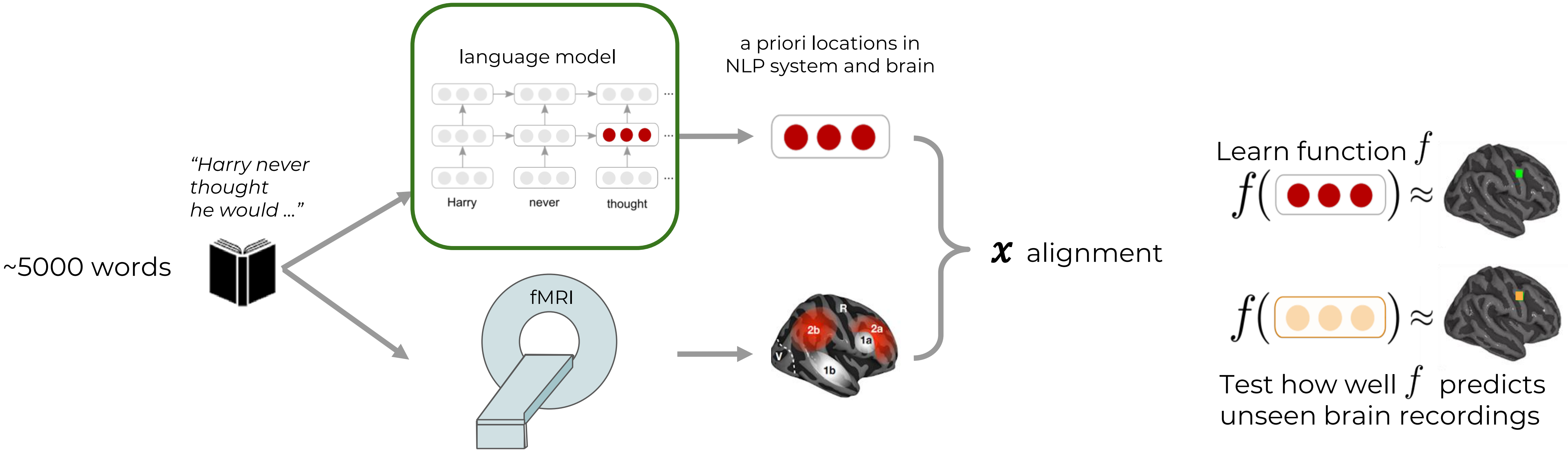
Toneva and Wehbe 2019,
Caucheteux et al. 2020,
Toneva et al. 2020

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

# LLMs, estimating alignment, evaluation

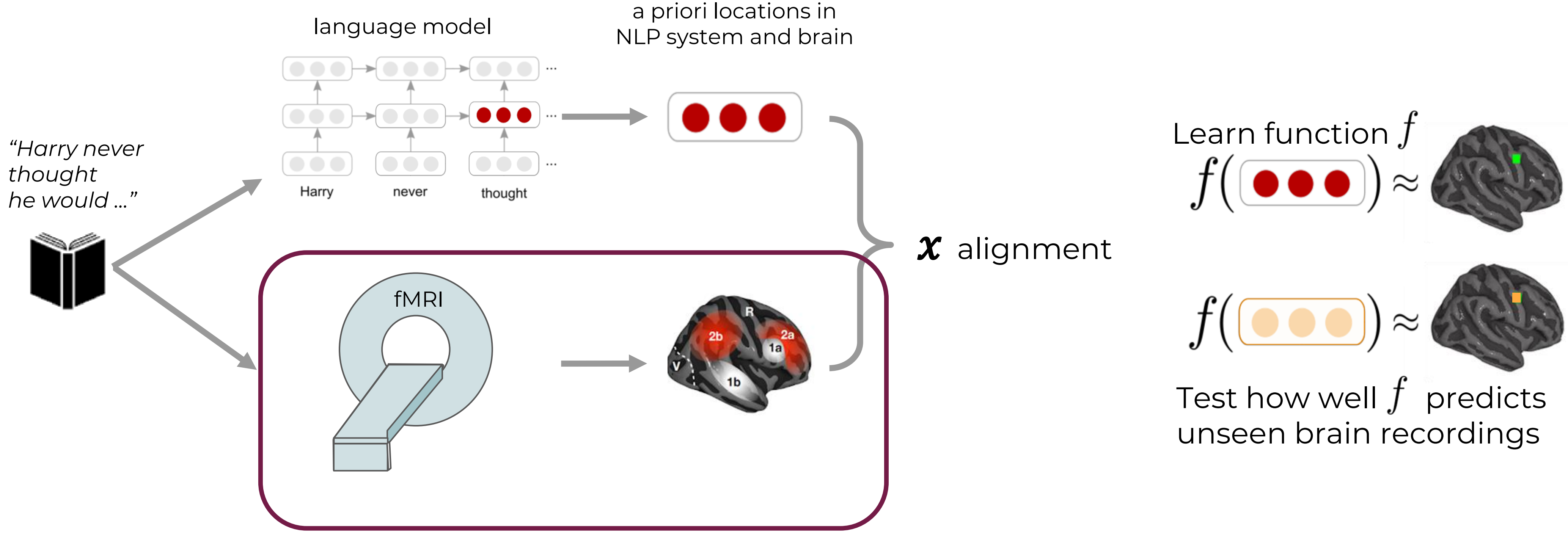


# Part 1: LLMs + extracting representations

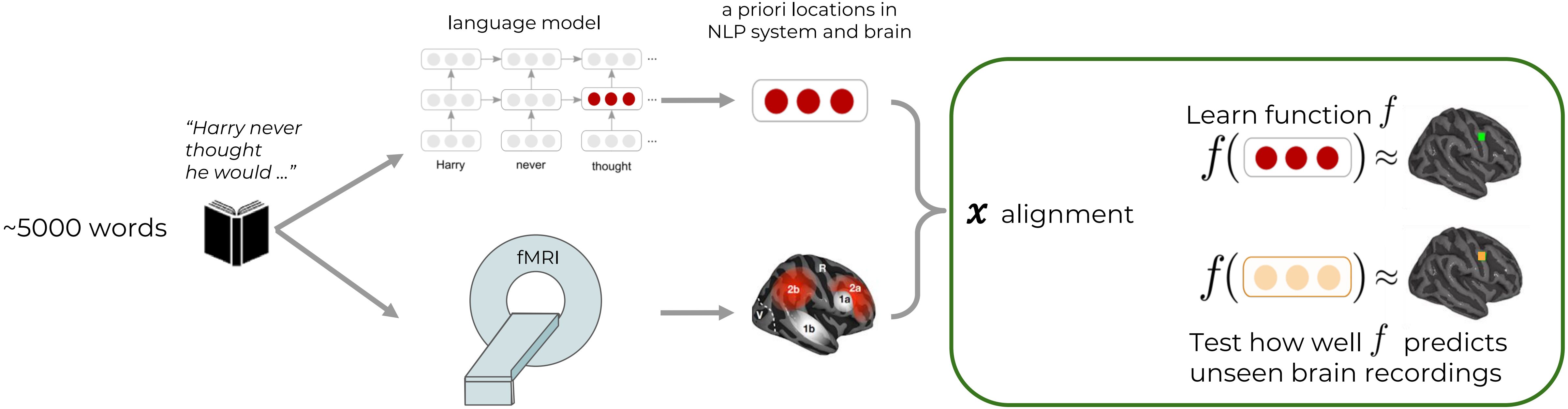


# LLMs, estimating alignment, evaluation

~5000 words



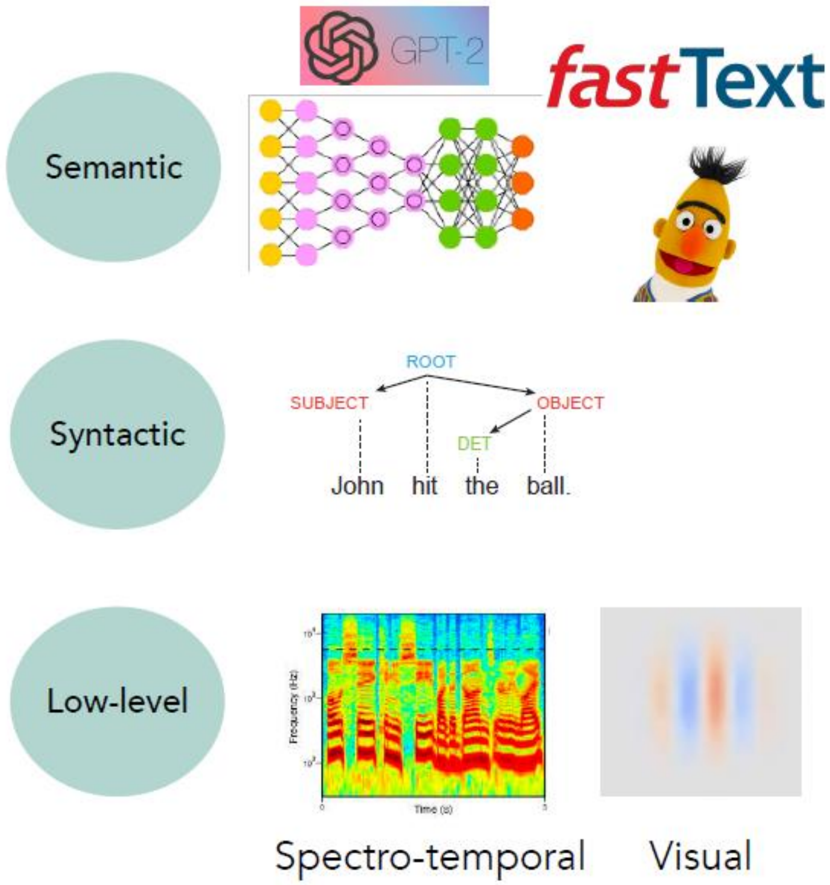
# Estimating brain-LM alignment + evaluation

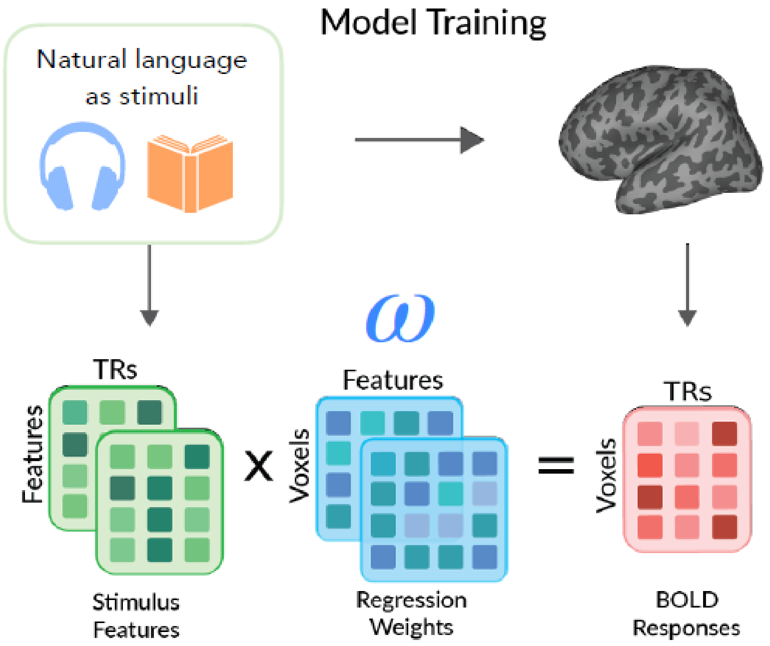


# Agenda

- Neuro-AI alignment: Introduction [1 hour 30 min]
  - Introduction to Brain encoding and decoding [30 min]
  - Types of Brain Recordings [15 min]
  - Types of Stimulus Representations [15 min]
  - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
  - Linguistic Brain Encoding [60 min]
    - Encoding schema
    - Pretrained language models and brain alignment
    - Challenges in using DL for cognitive science
  - Linguistic Brain Decoding [15 min]
  - Multimodal Brain Encoding [15 min]

# Pretrained language models and brain alignment



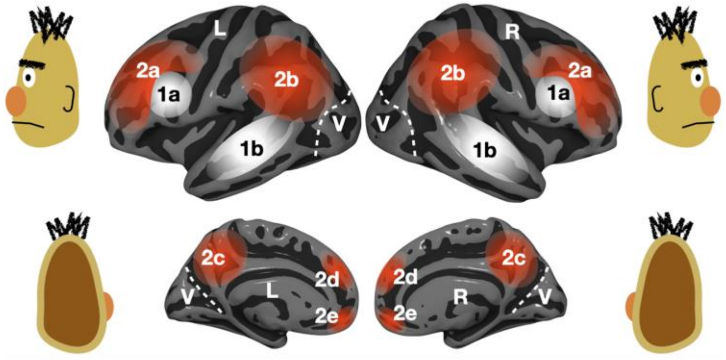



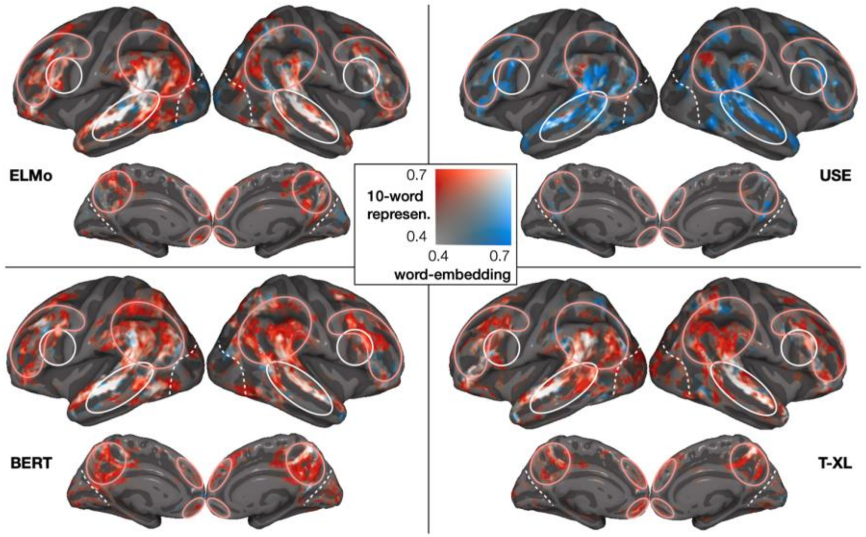

**Regression weights** map from feature space to brain responses.

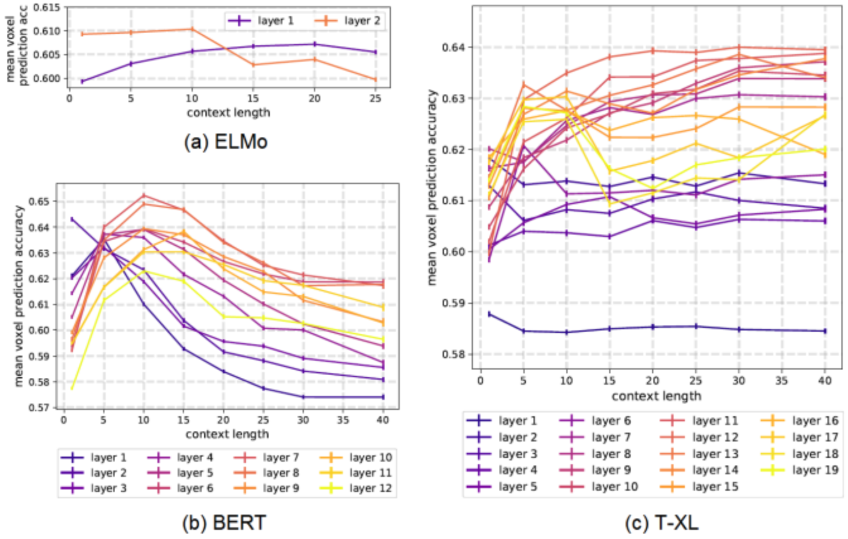
Comparison of semantic feature spaces from PLMs with traditional word embeddings

# Language: work utilizing DL progress

- Stimuli: one chapter of Harry Potter
- Stimulus representation: derived from **pretrained** NLP systems
- Brain recording & modality: fMRI, reading




across several types of large NLP systems, best alignment with fMRI in middle layers

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). Advances in Neural Information Processing Systems, 32.

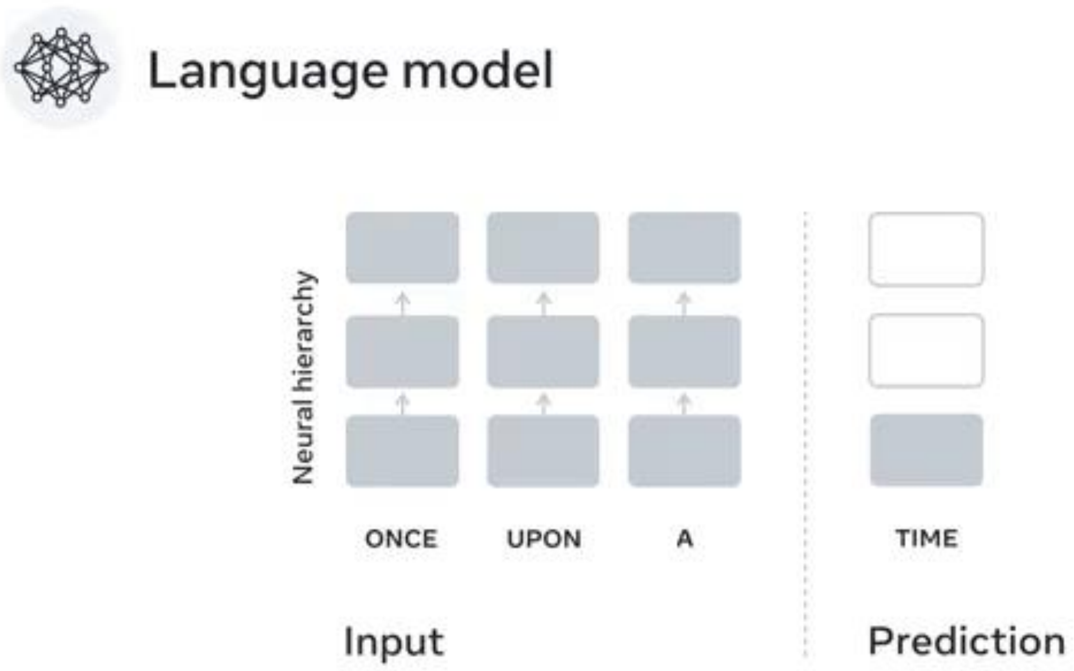
# Language: work utilizing DL progress

- Stimuli: sentences
- Stimulus representation: derived from pretrained NLP systems
- Brain recording

best alignment with fMRI & MEG in middle layers

better performance at ...ing next word -> better ...ion of fMRI & MEg





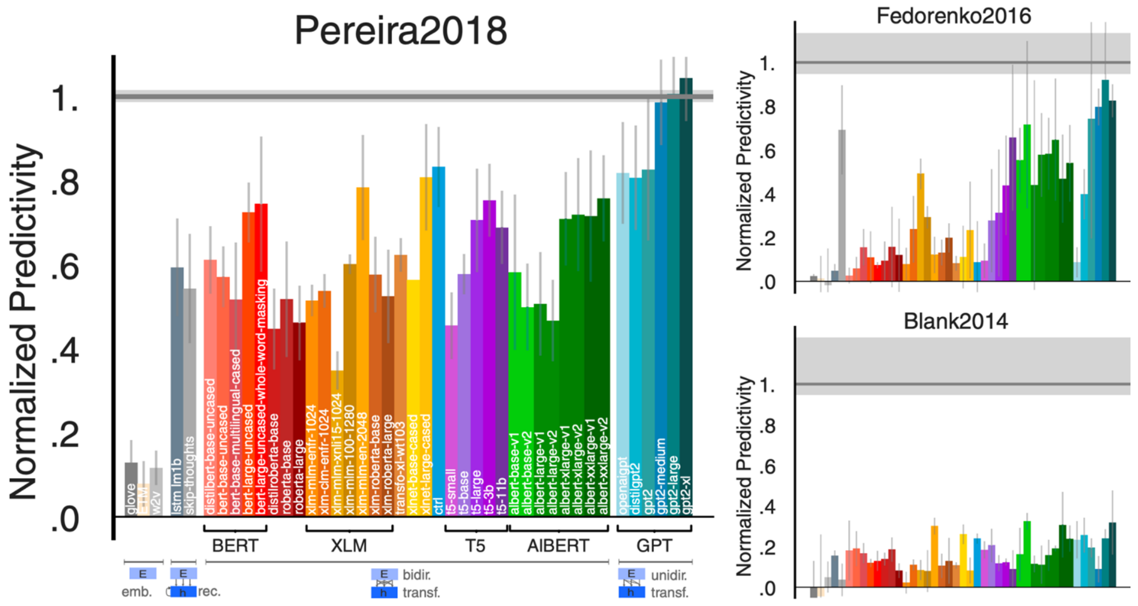


Caucheteux, Charlotte, and Jean-Rémi King. "Brains and algorithms partially converge in natural language processing." Communications biology 5, no. 1 (2022): 1-10.

# Language: work utilizing DL progress

- Stimuli: sentences, passages, short story
- Stimulus representation: derived from pretrained NLP systems (BERT, GPT-2, T5 , and XLM)
- Brain recording & modality: fMRI & ECoG, reading & listening

some NLP systems can predict fMRI and ECoG up to 100% of estimated noise ceiling

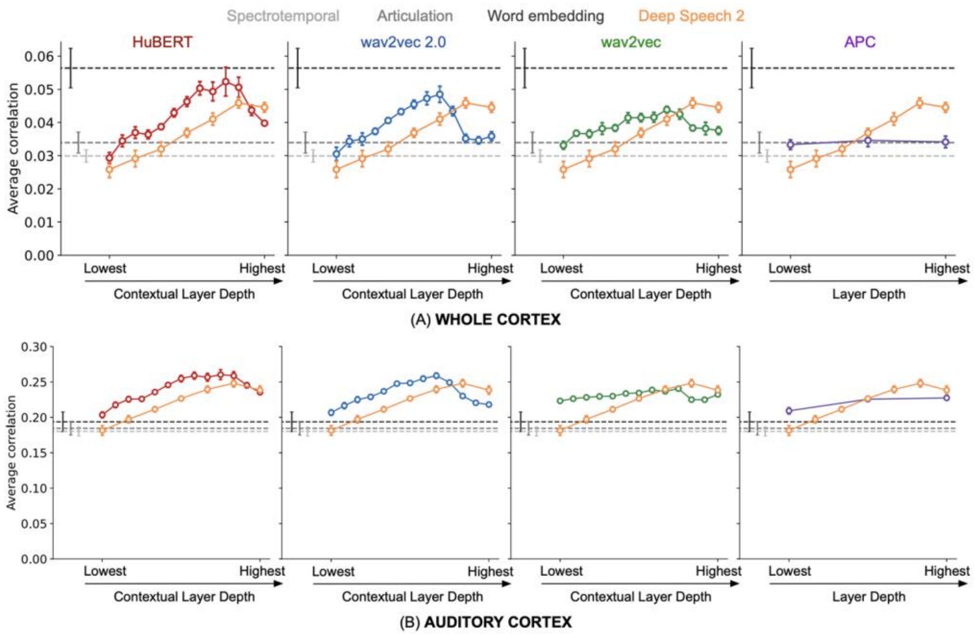


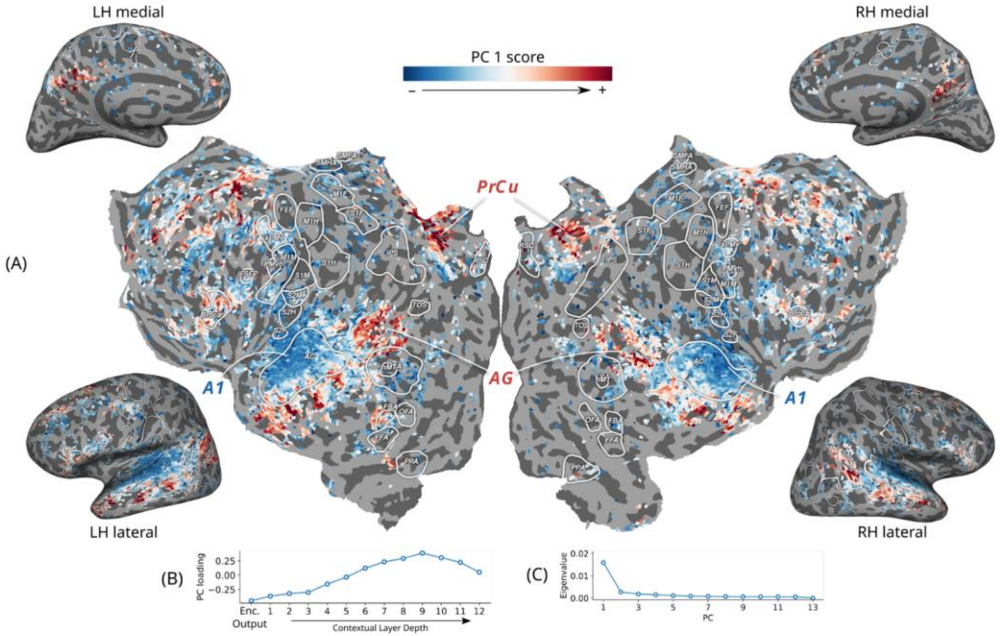

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. "The neural architecture of language: Integrative modeling converges on predictive processing." Proceedings of the National Academy of Sciences 118, no. 45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

# Audio: work utilizing DL progress

- Stimuli: Moth Radio Hour
- Stimulus representation: derived from pretrained **self-supervised speech models**
- Brain recording & modality: fMRI, listening

Middle layers of self-supervised speech models predict auditory cortex the best

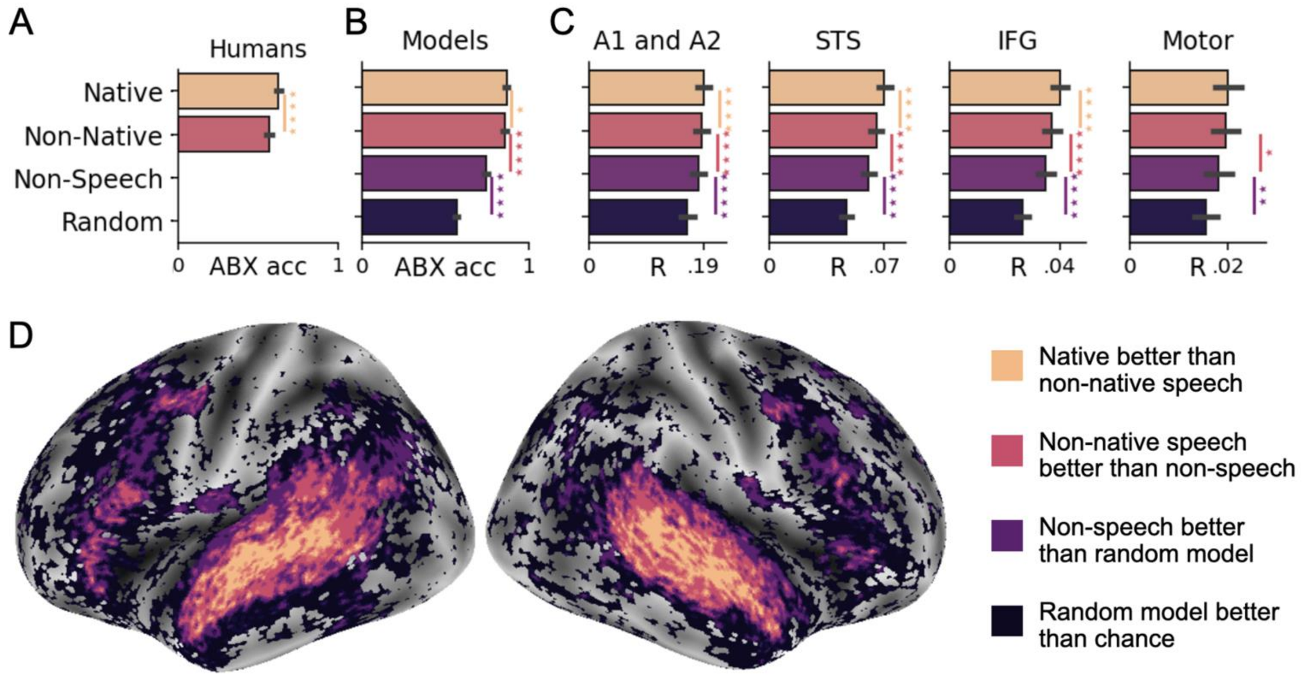



Vaidya, Aditya R., Shailee Jain, and Alexander G. Huth. "Self-supervised models of audio effectively explain human cortical responses to speech." ICML (2022).

# Audio: work utilizing DL progress

- Stimuli: audio books
- Stimulus representation: derived from pretrained self-supervised speech model
- Brain recording & modality: fMRI, listening in 3 languages (Eng, Fr, Mandarin)




Self-supervised speech models reveal specialization for native sounds in the STS and MTG;

IFG and AG show more general specialization for speech rather than native-language

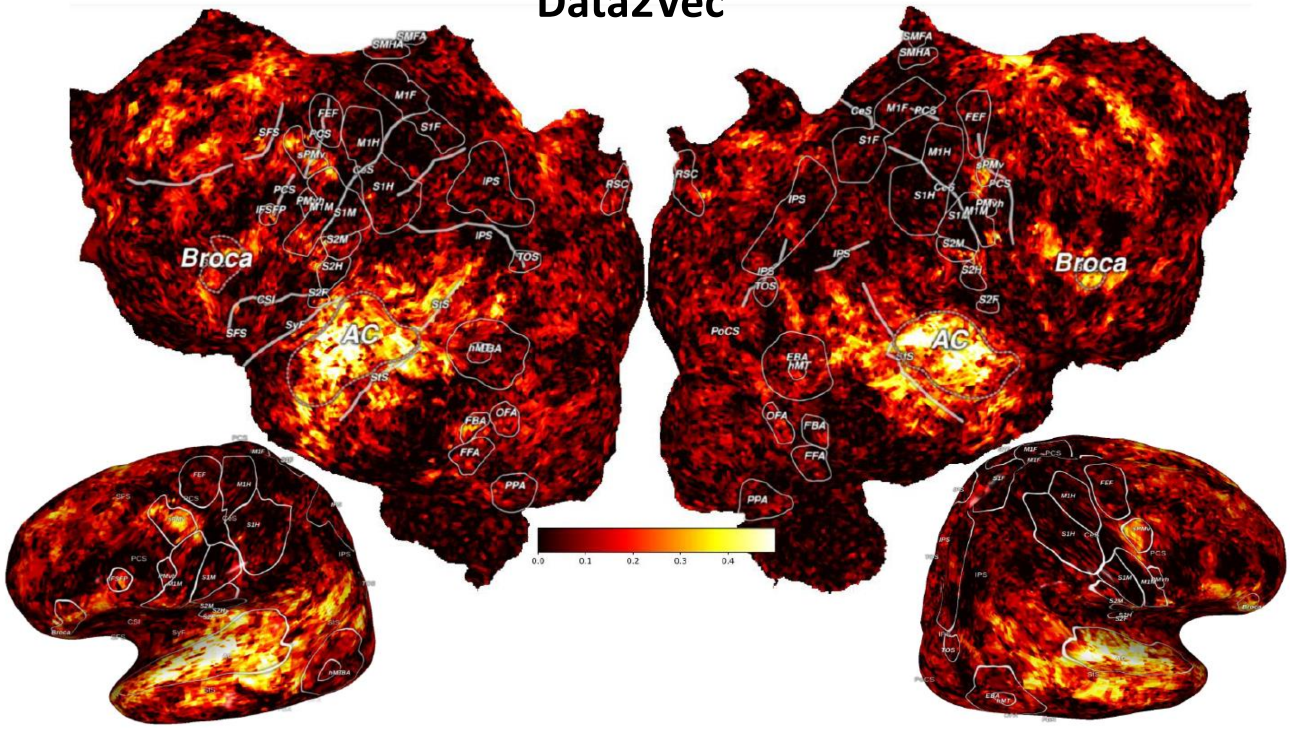
Millet, Juliette, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. "Toward a realistic model of speech processing in the brain with self-supervised learning." arXiv preprint arXiv:2206.01685 (2022).

# Audio: work utilizing DL progress

Contrastive and predictive models encode the information better than the generative and the traditional low-level acoustic baselines, and VGGish models.

- Stimuli: Moth-Radio-Hour
- Stimulus representation: derived from 5 basic + 25 pretrained self-supervised speech models
- Brain recording & modality: fMRI

**Data2Vec**

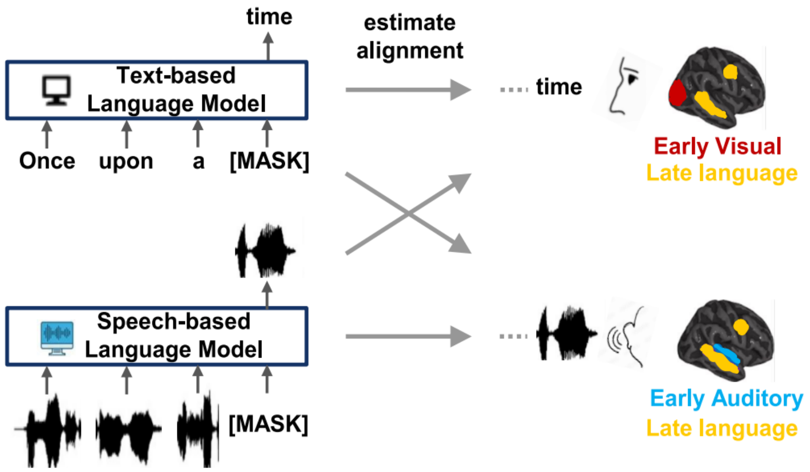


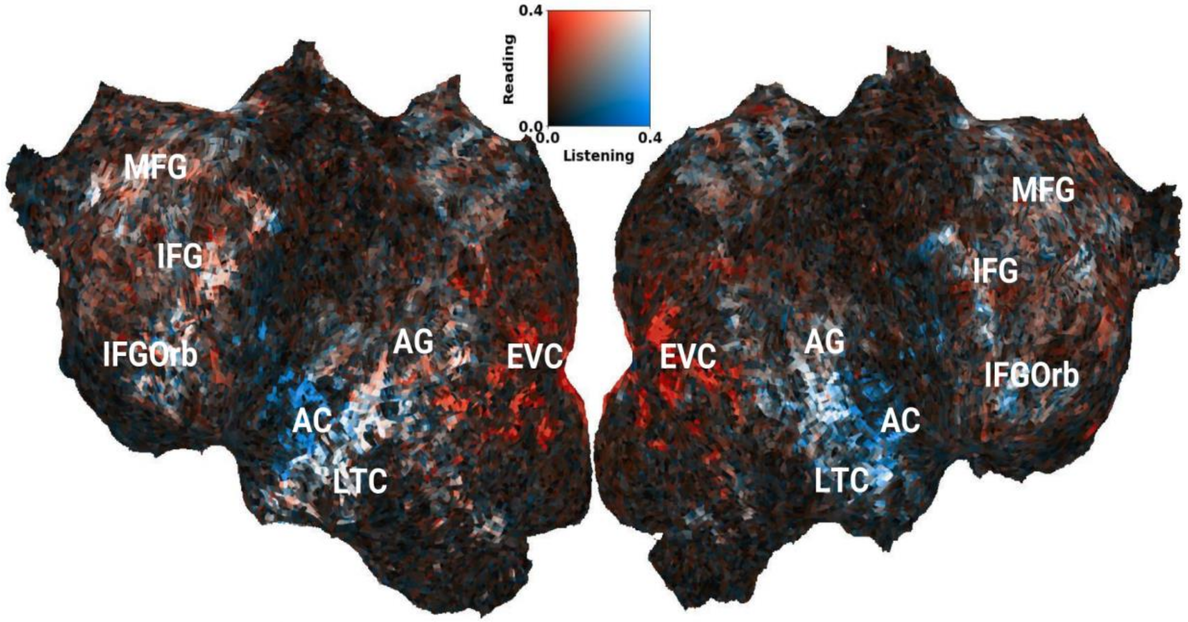

| Category | Model | AC | Broca | Whole Brain |
|---|---|---|---|---|
| Traditional non-DL & non-SS DL Methods | Spectrogram | 0.0545 | 0.0511 | 0.0495 |
| | Filter bank | 0.0477 | 0.0450 | 0.0498 |
| | Mel | 0.0489 | 0.0515 | 0.0511 |
| | MFCC | 0.0495 | 0.0520 | 0.0517 |
| | VGGish | **0.1612** | **0.0785** | **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 | **<u>0.2683</u>** | **<u>0.1756</u>** | **<u>0.0886</u>** |
| | DistilHuBERT | 0.2323 | 0.1101 | 0.0738 |
| | LightHuBERT | 0.2328 | 0.1102 | 0.0737 |

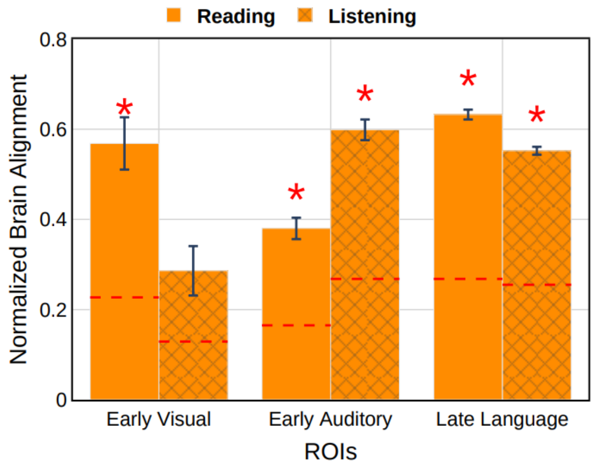
Subba Reddy Oota, Khushbu Pahwa, Mounika Marreddy, Manish Gupta, and Bapi S. Raju. "Neural architecture of speech" ICASSP-2023

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

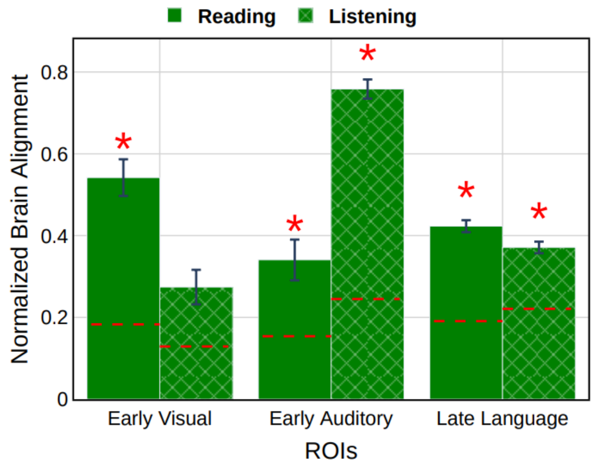
- Stimuli: Subset-Moth-Radio-Hour
- Stimulus representation: pretrained NLP models and speech models
- Brain recording & modality: fMRI, Reading, Listening




(a) Text Models



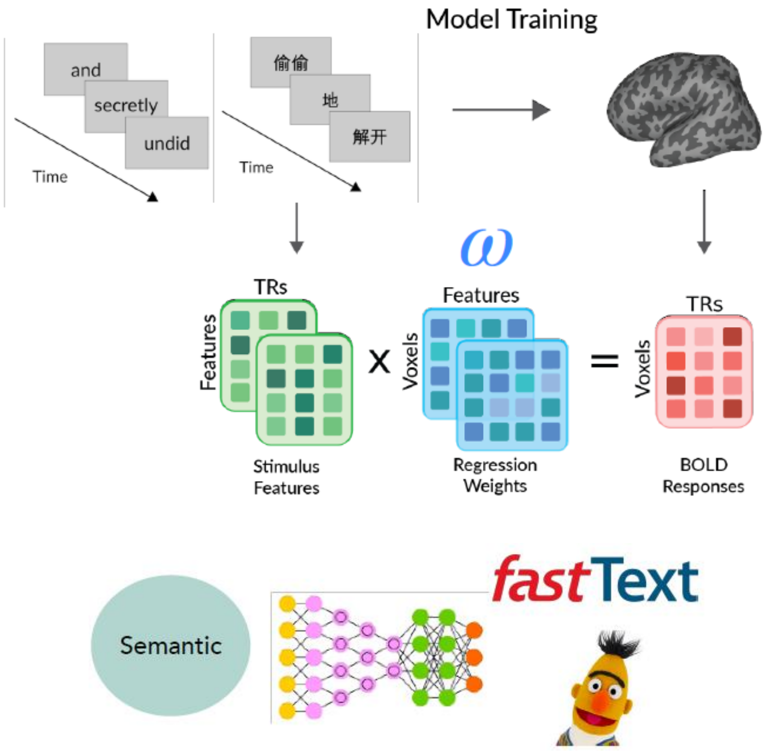

(b) Speech Models

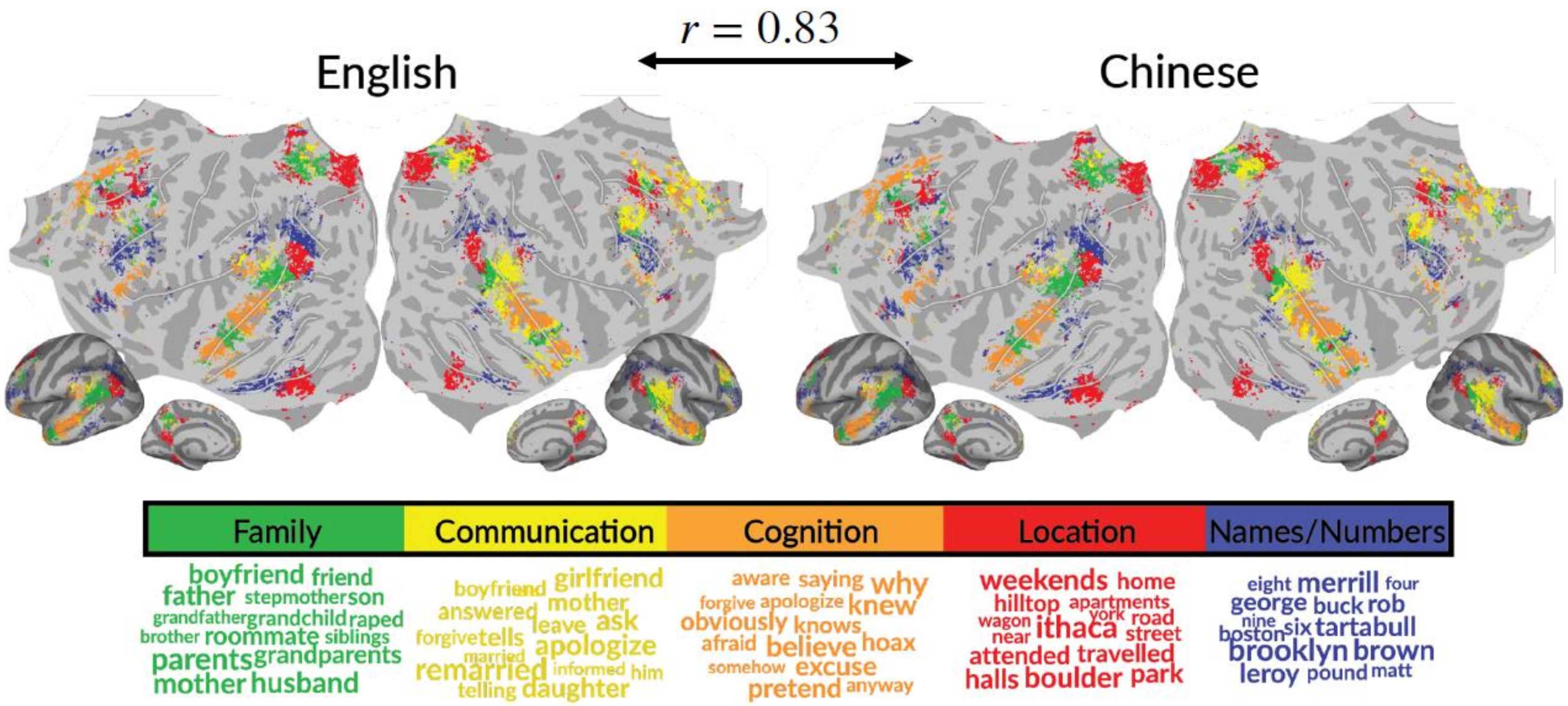
- **Late language regions**: Both types of models show high brain alignment with **late language regions**, but **speech models** trails behind **text models**
- Highly predict early visual and auditory areas.

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

# 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


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. Use 🧠 to evaluate how well representations from 🔤 (static vs. recurrent vs. pretrained) can predict representations of the 🧠 during language comprehension
2. **Speech models (🔊)** useful for modeling **early listening (🎧)**: investigate speech models to learn more about AC
3. **Text models (🔤)** useful for modeling **language processing** in both 🎧 and 📖
4. **Semantic representations** are independent of the modality (📖 or 🎧) and distributed across language regions
5. Across several types of pretrained language models, best alignment with fMRI/MEG in middle layers
6. **Text models (🔤)** predict fMRI recordings significantly better than speech models **(🔊)**
7. **Semantic representation** within individuals are mostly **shared across Chinese and English**

# Agenda

- Neuro-AI alignment: Introduction [1 hour 30 min]
  - Introduction to Brain encoding and decoding [30 min]
  - Types of Brain Recordings [15 min]
  - Types of Stimulus Representations [15 min]
  - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
  - Linguistic Brain Encoding [60 min]
    - Encoding schema
    - Pretrained language models and brain alignment
    - Challenges in using DL for cognitive science
  - Linguistic Brain Decoding [15 min]
  - Multimodal Brain Encoding [15 min]

# Challenges in using DL for cognitive science

- Not designed to specifically model brain processing

NLP systems: Designed to predict upcoming words

*Harry never thought* ***???***

*Harry never thought he* ***???***

*Harry never thought he would* ***???***

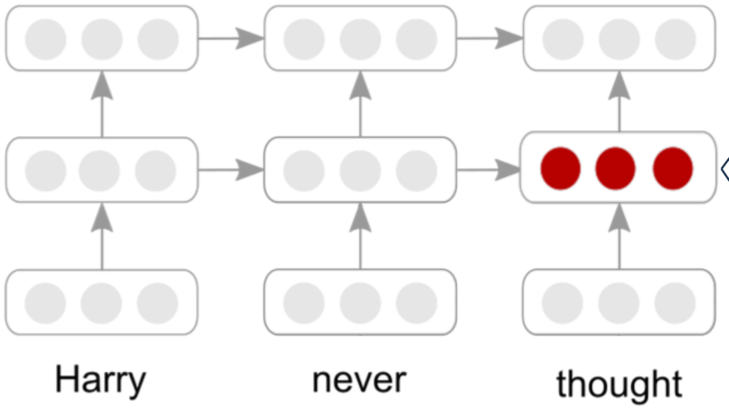
...

# Challenges in using DL for cognitive science

- Not designed to specifically model brain processing
  - Training DL models using brain recordings
  - Task-based modeling

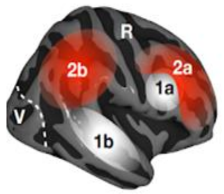
# Challenges in using DL for cognitive science

- Not designed to specifically model brain processing
  - Training DL models using brain recordings
  - Task-based modeling
- Can be difficult to interpret due to multiple sources of information




part-of-speech
+
semantic role
+
dependence on
other words
+
...

?



# Challenges in using DL for cognitive science

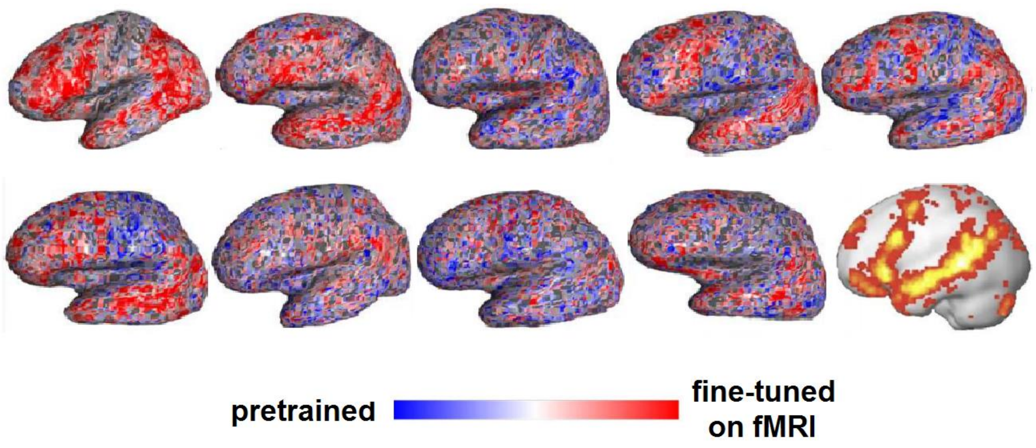
- Not designed to specifically model brain processing
  - Training DL models using brain recordings
  - Task-based modeling
- Can be difficult to interpret due to multiple sources of information
  - Disentangling contributions of different info sources to brain predictions

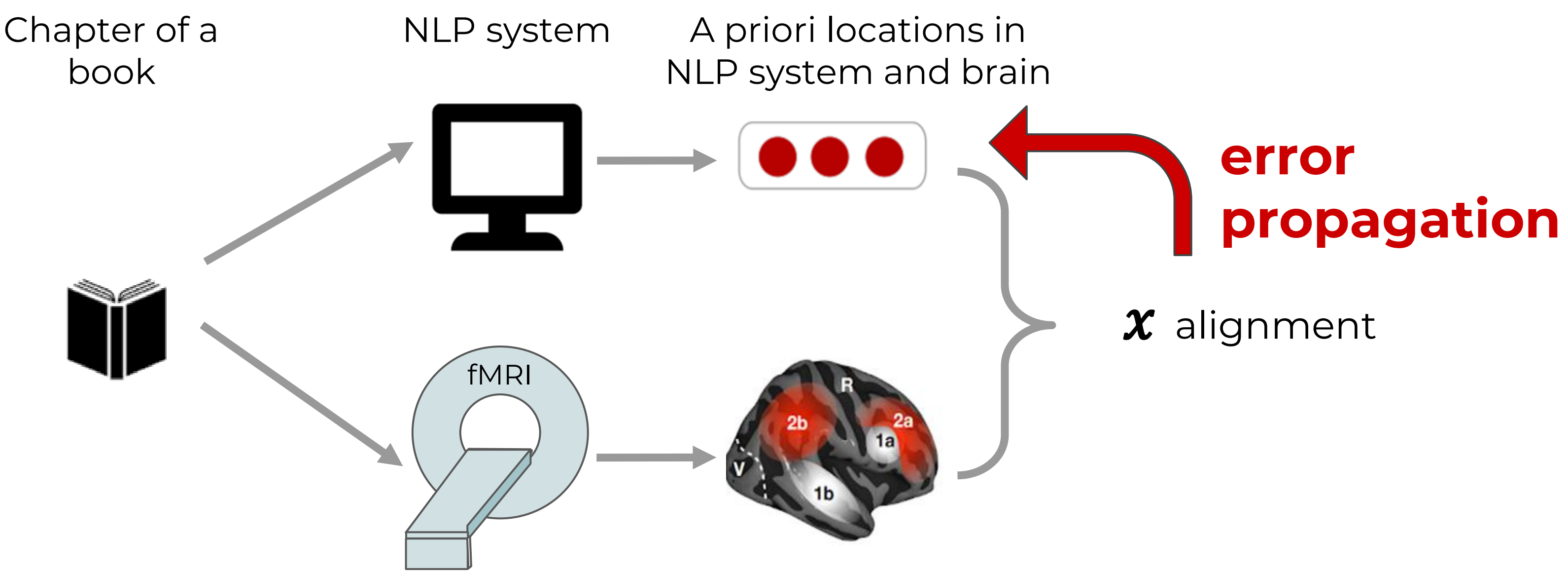
# Agenda

- Neuro-AI alignment: Introduction [1 hour 30 min]
  - Introduction to Brain encoding and decoding [30 min]
  - Types of Brain Recordings [15 min]
  - Types of Stimulus Representations [15 min]
  - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
  - Linguistic Brain Encoding [60 min]
    - Encoding schema
    - Pretrained language models and brain alignment
    - Challenges in using DL for cognitive science
      - Training DL models using brain recordings
      - Task-based language models and brain alignment
      - Disentangling Syntax and Semantics
  - Linguistic Brain Decoding [15 min]
  - Multimodal Brain Encoding [15 min]

# Training DL models using brain recordings

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

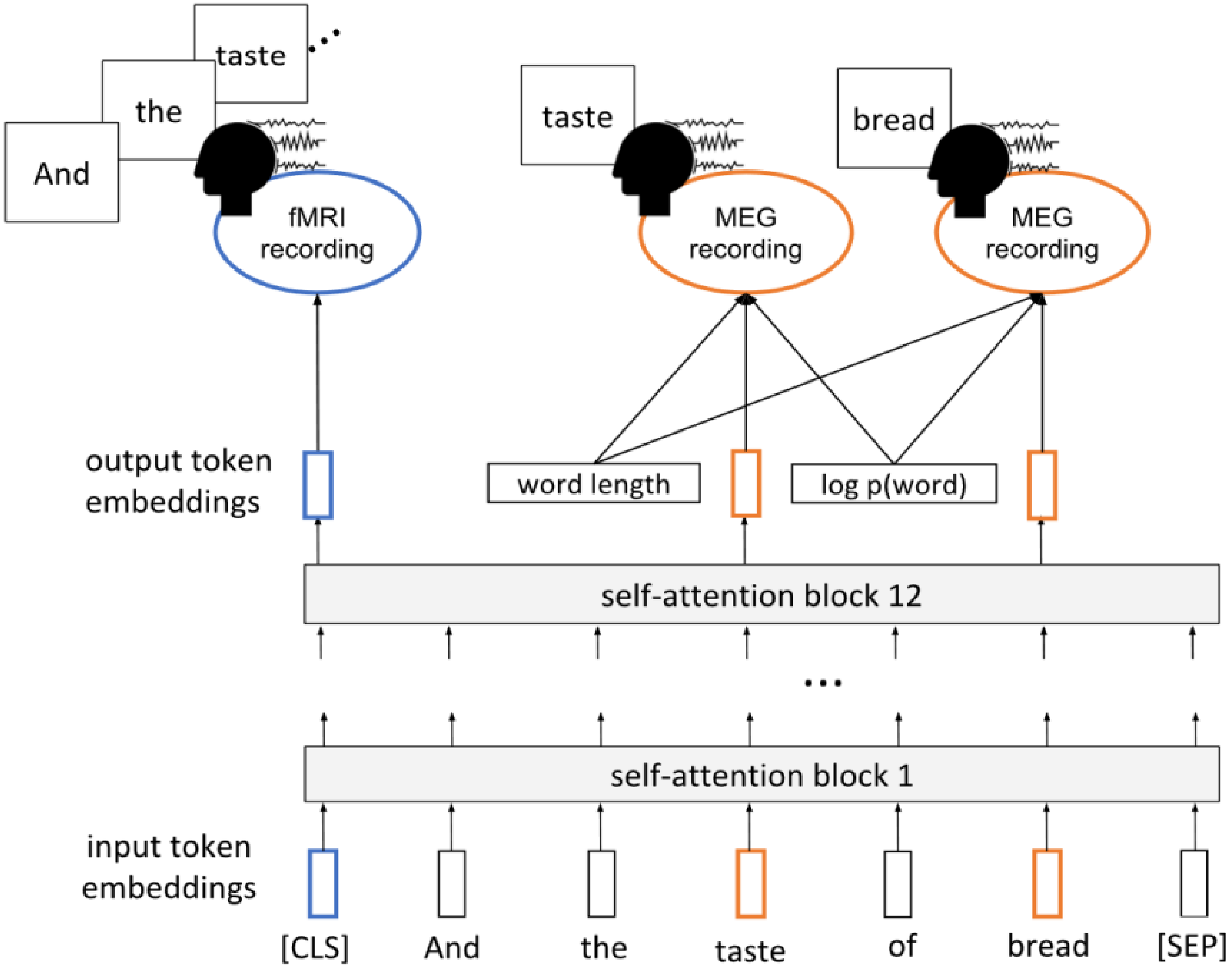



Brain-optimized NLP model predicts unseen fMRI recordings better, especially in canonical language regions

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

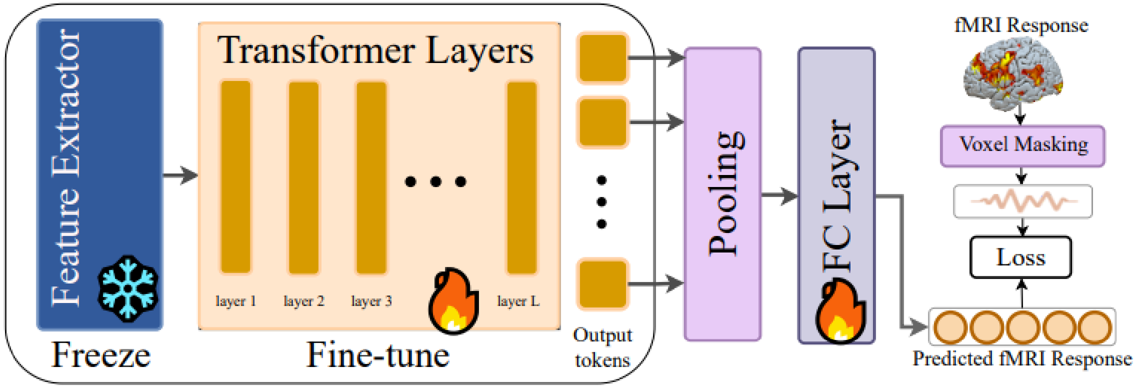



| Metric | Vanilla | MEG | Joint |
|---|---|---|---|
| CoLA | 57.29 | 57.63 | **57.97** |
| SST-2 | 93.00 | **93.23** | 91.62 |
| MRPC (Acc.) | 83.82 | 83.97 | **84.04** |
| MRPC (F1) | 88.85 | **88.93** | 88.91 |
| STS-B (Pears.) | **89.70** | 89.32 | 88.60 |
| STS-B (Spear.) | **89.37** | 88.87 | 88.23 |
| 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 |

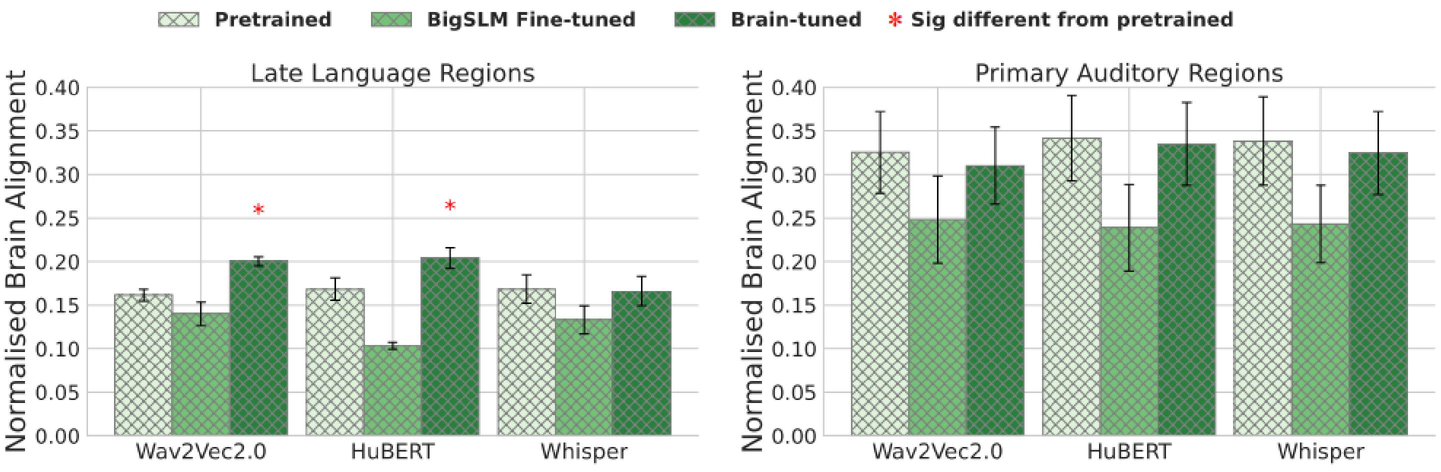
Schwartz, Dan, Mariya Toneva, and Leila Wehbe. "Inducing brain-relevant bias in natural language processing models." Advances in neural information processing systems 32 (2019).

# Training Speech models using brain recordings

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

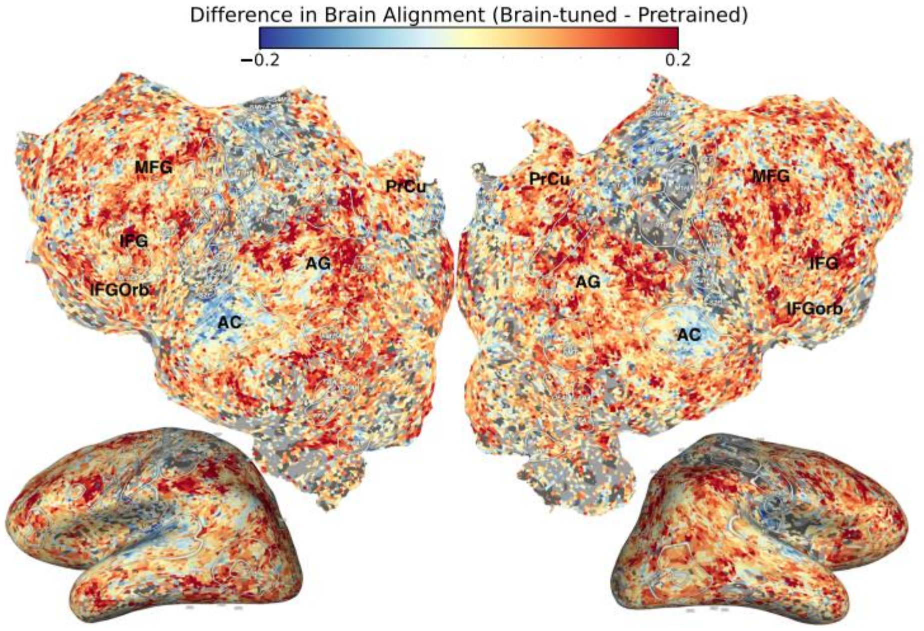



(a) Proposed brain-tuning approach




(a) Normalized alignment for late language regions

(b) Normalized alignment for primary auditory

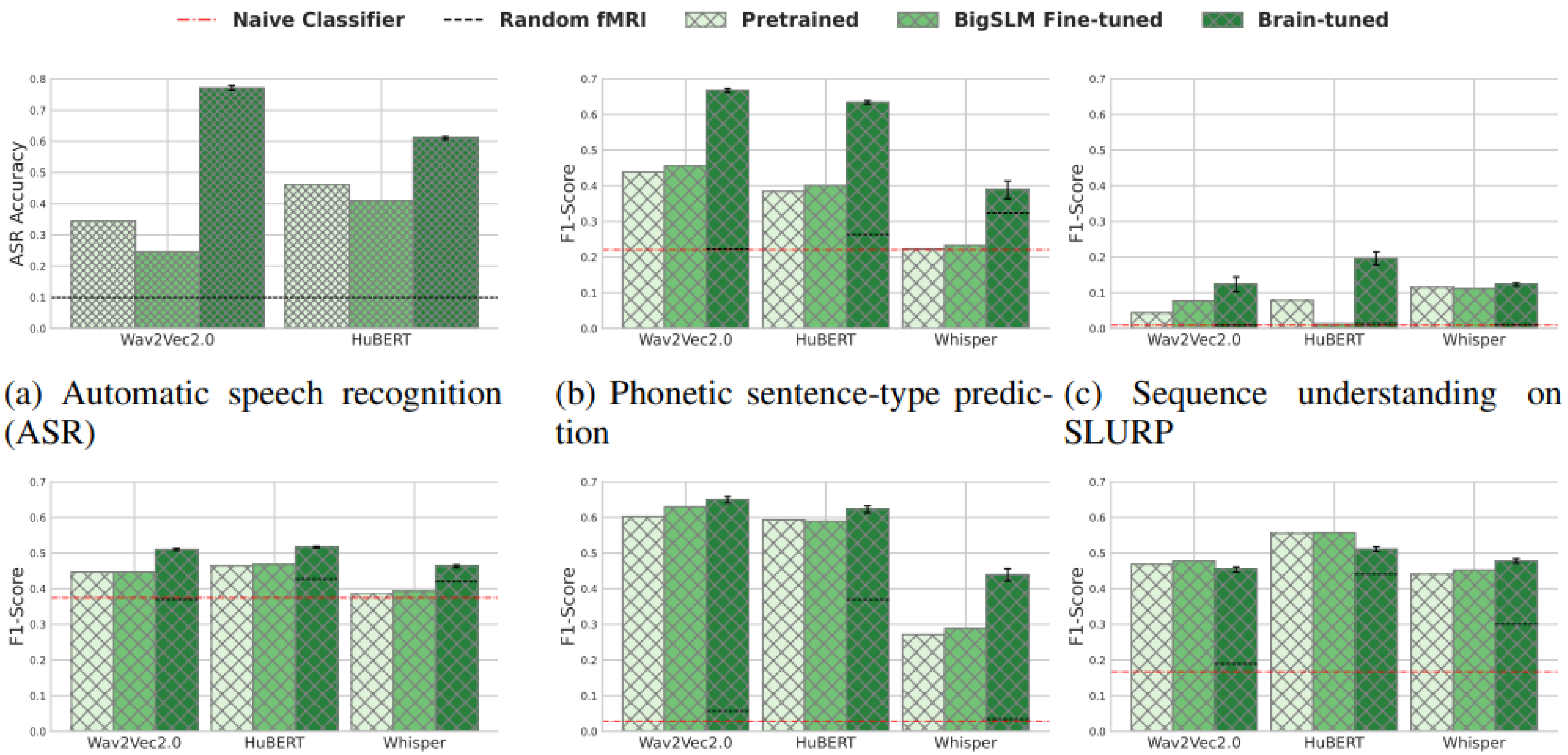



(c) Difference in brain alignment due to brain-tuning of Wav2vec2.0

- 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

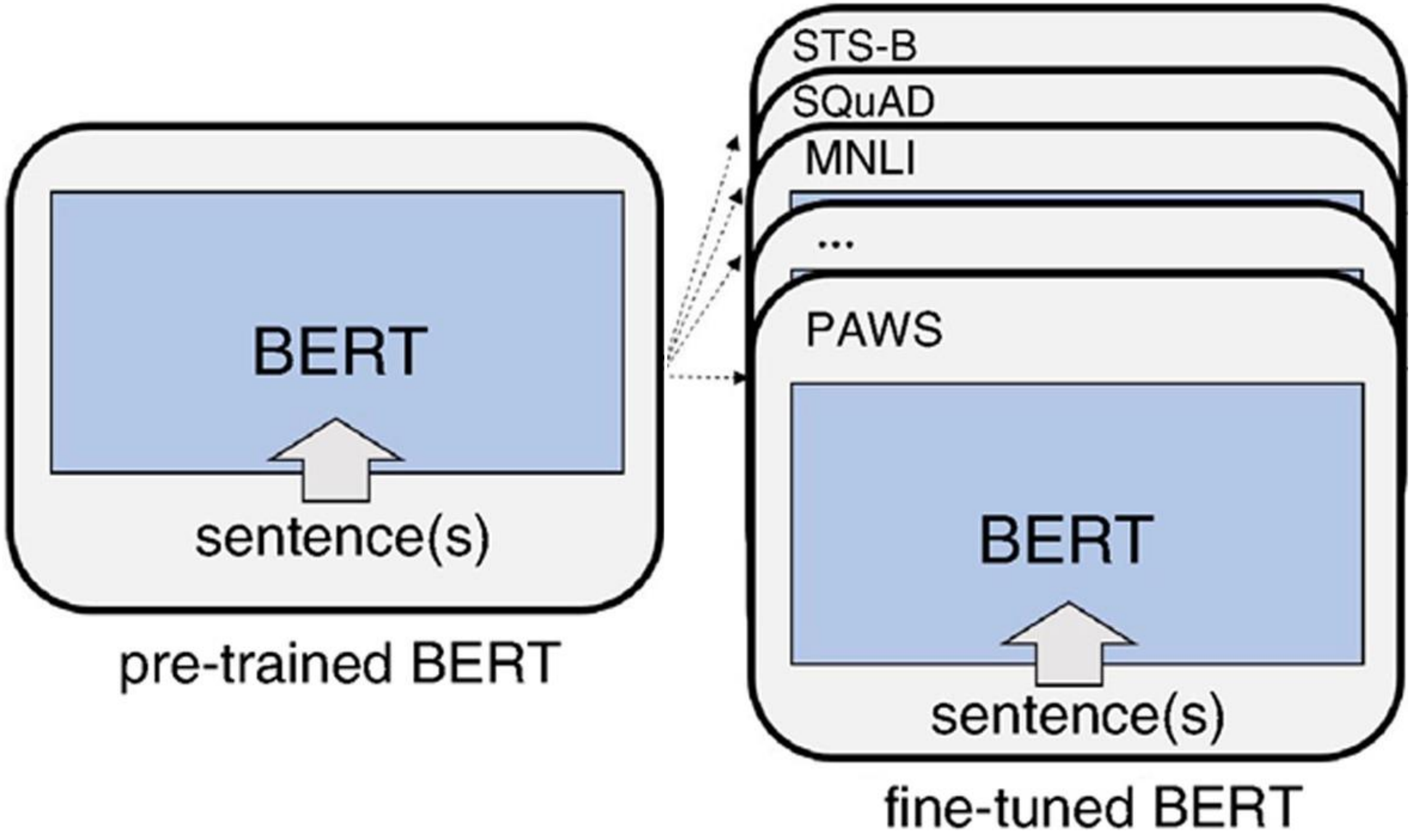
- Brain-tuned models show consistent improvement over the baselines, with biggest gains in more semantic tasks (ASR and phonetic sentence-type prediction)

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

# Agenda

- Neuro-AI alignment: Introduction [1 hour 30 min]
  - Introduction to Brain encoding and decoding [30 min]
  - Types of Brain Recordings [15 min]
  - Types of Stimulus Representations [15 min]
  - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
  - Linguistic Brain Encoding [60 min]
    - Encoding schema
    - Pretrained language models and brain alignment
    - Challenges in using DL for cognitive science
      - Training DL models using brain recordings
      - Task-based language models and brain alignment
      - Disentangling Syntax and Semantics
  - Linguistic Brain Decoding [15 min]
  - Multimodal Brain Encoding [15 min]

# Can task-specific language models better predict fMRI brain activity?




Tasks

- Paraphrase
- Summrisation
- Question Answering
- Sentiment Analysis
- NER
- Word Sense Disambiguation
- Natural Language Inference
- Semantic Role Labeling
- Coreference Resolution
- Shallow Syntax
- Pretrained BERT

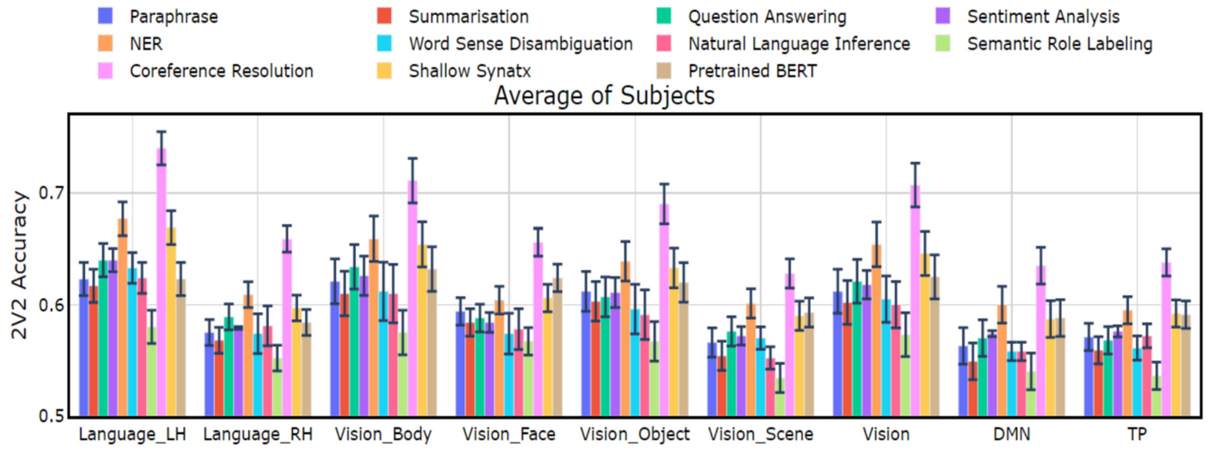
Syntactic

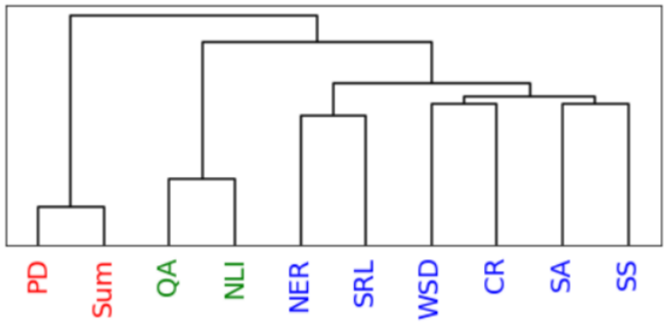
Devlin et al. 2019, Bowon et al. 2020

# Tasks affect processing: NLP

- Stimuli: passages and narratives
- Stimulus representation: task-optimized NLP models for a range of tasks
- Brain recording & modality: fMRI, reading & listening of different stimuli

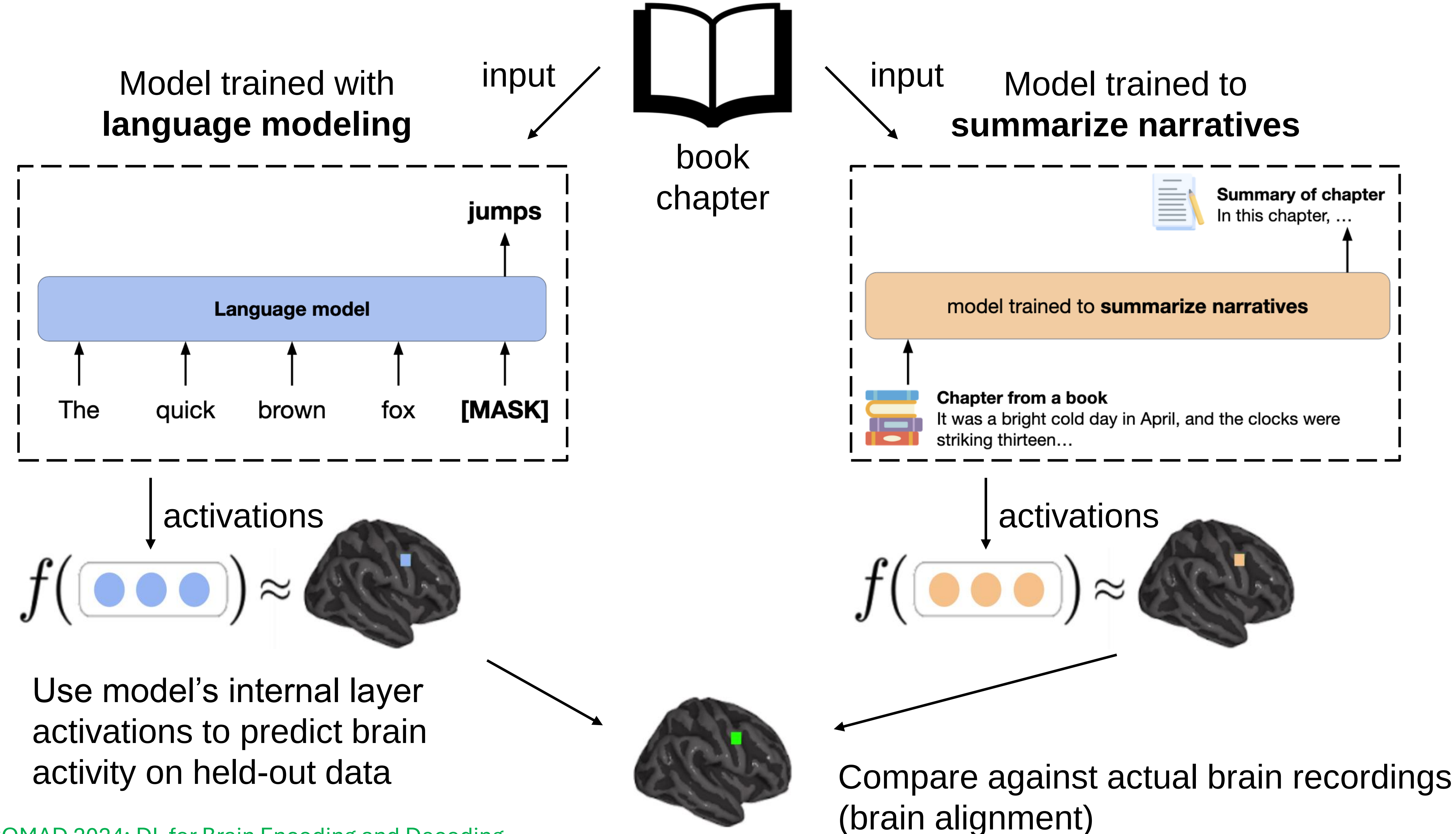
Reading fMRI best explained by coref. resolution, NER, shallow syntax parsing
Listening fMRI best explained by paraphrasing, summarization, NLI



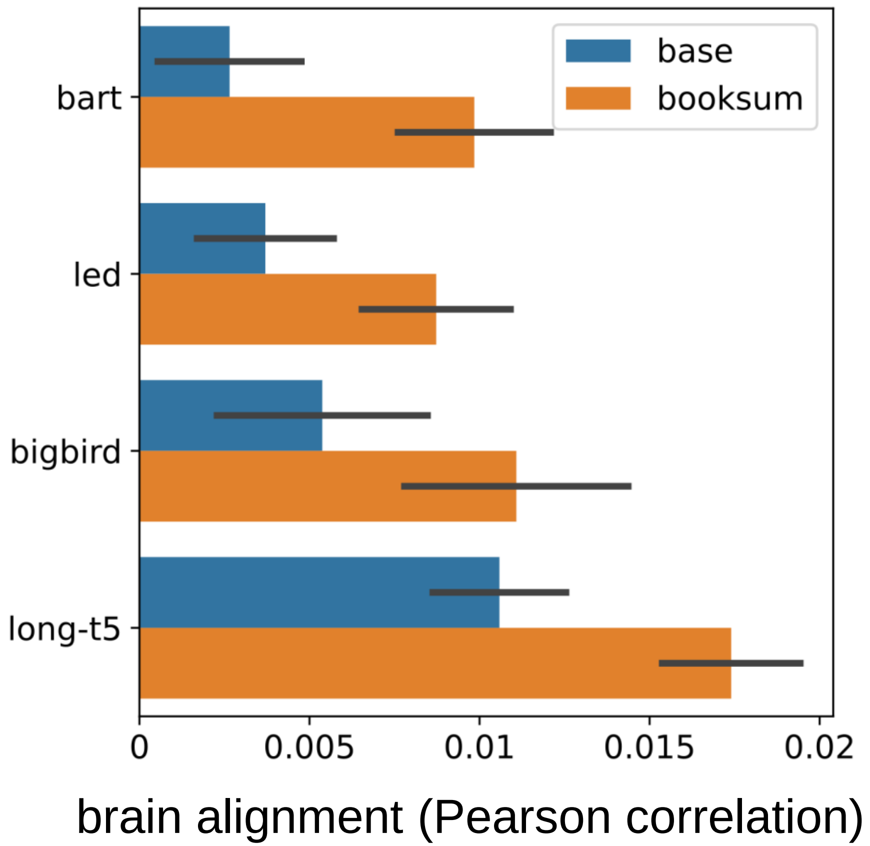




Oota, Subba Reddy, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampudi. "Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity?." NAACL (2022).

# How to build better Language models?



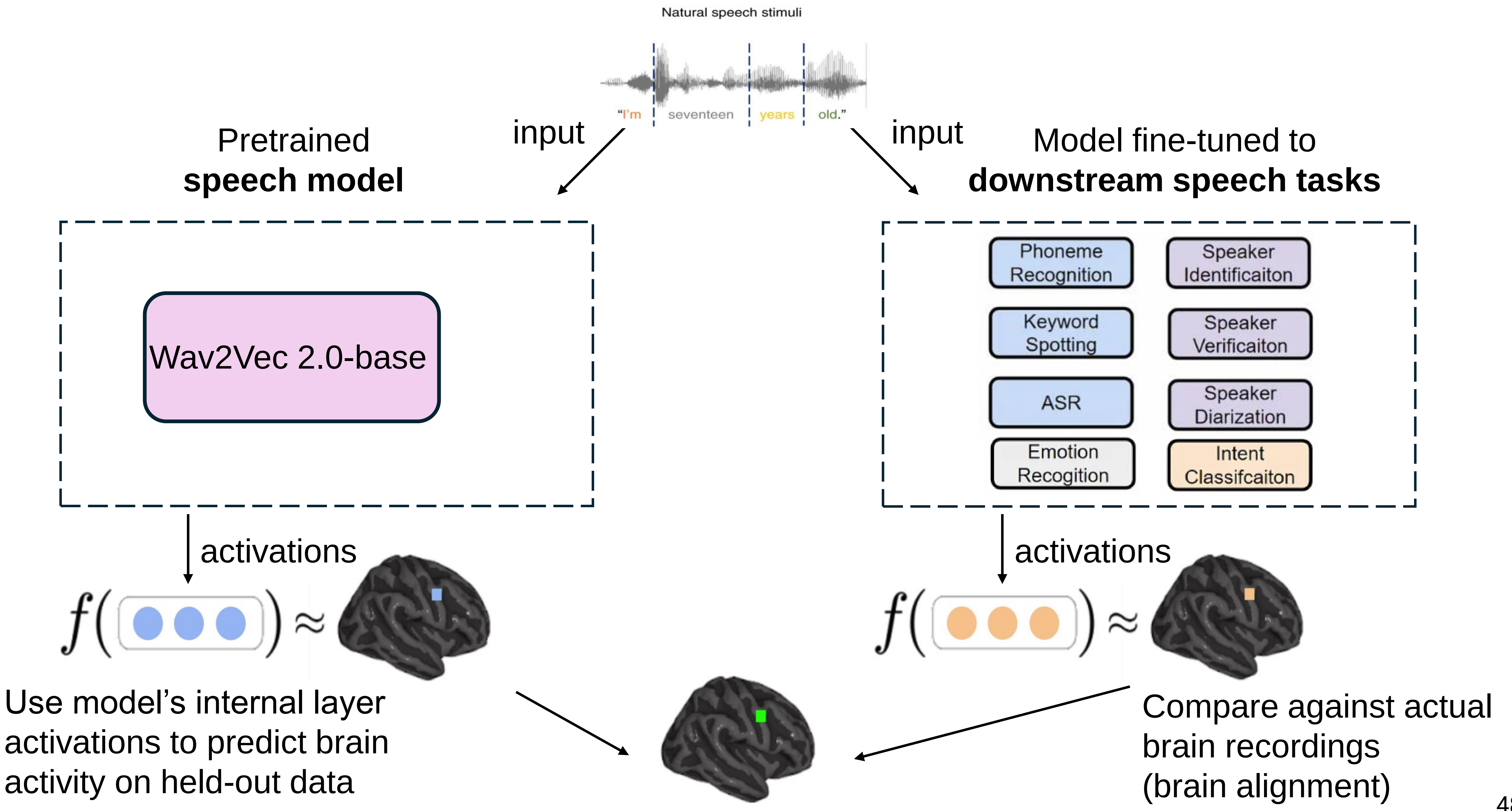
# Result: Summarize narratives → Greater brain alignment 🧠



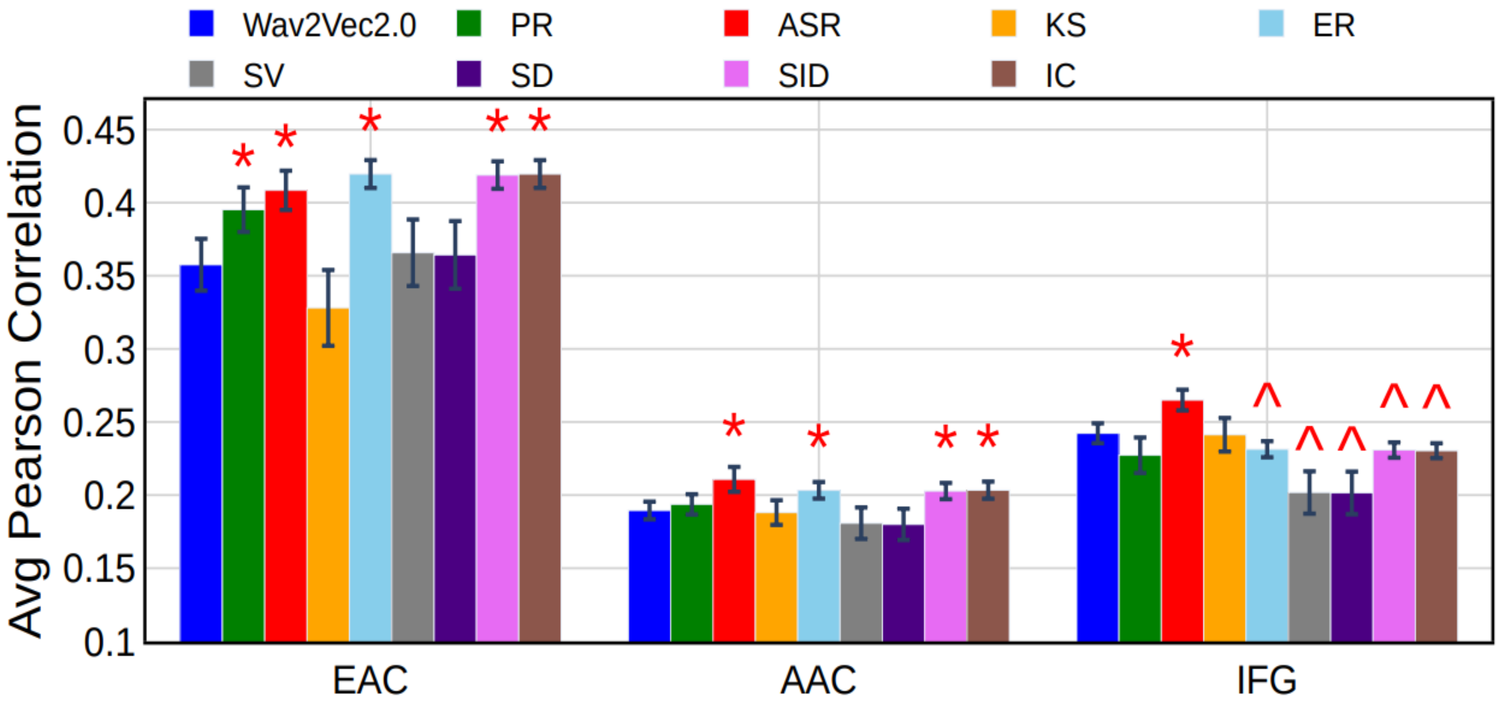

Training language models to summarize narratives improves brain alignment

this is the title of our paper!

# Tasks affect processing: Speech



# Region level alignments




- All speech tasks are better aligned with EAC compared to AAC and IFG regions.
- Finetuning on ER, SID and IC leads to the best alignment for the early auditory cortex
- Finetuning on ASR provides the best encoding for the auditory associative cortex and language regions.


Oota, Subba Reddy, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Raju Surampudi. "Speech Taskonomy: Which Speech Tasks are the most Predictive of fMRI Brain Activity?" INTERSPEECH (2023).

# Agenda

- Neuro-AI alignment: Introduction [1 hour 30 min]
  - Introduction to Brain encoding and decoding [30 min]
  - Types of Brain Recordings [15 min]
  - Types of Stimulus Representations [15 min]
  - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
  - Linguistic Brain Encoding [60 min]
    - Encoding schema
    - Pretrained language models and brain alignment
    - Challenges in using DL for cognitive science
      - Training DL models using brain recordings
      - Task-based language models and brain alignment
      - Disentangling Syntax and Semantics
  - Linguistic Brain Decoding [15 min]
  - Multimodal Brain Encoding [15 min]

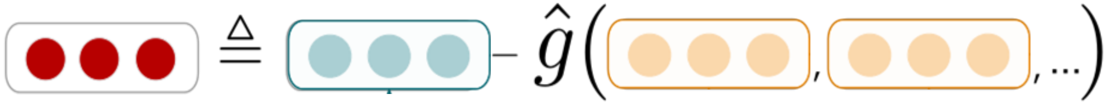
# Disentangling contributions of different info sources to brain predictions

*“Mary finished the apple”*

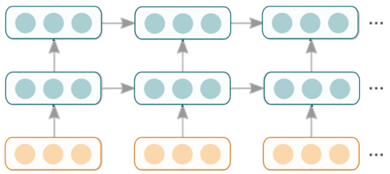
**supra-word meaning** may contain concept of:

- eating
- apple core
- …

Isolating supra-word meaning is a type of intervention



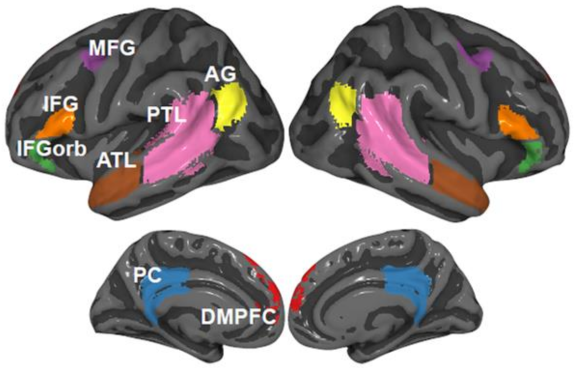

supra-word meaning



Toneva, Mariya, Tom M. Mitchell, and Leila Wehbe. "Combining computational controls with natural text reveals new aspects of meaning composition." BioRxiv (2020).

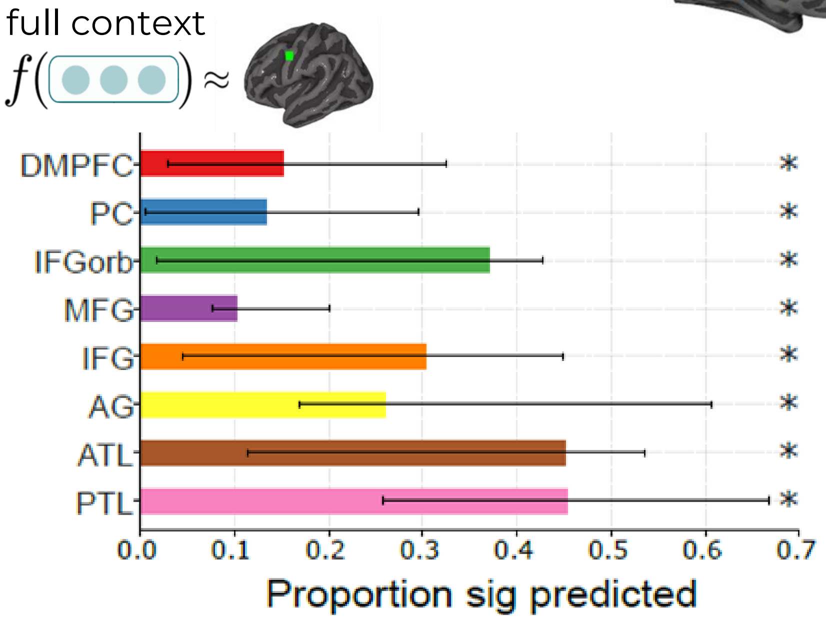
# Disentangling contributions of different info sources to brain predictions

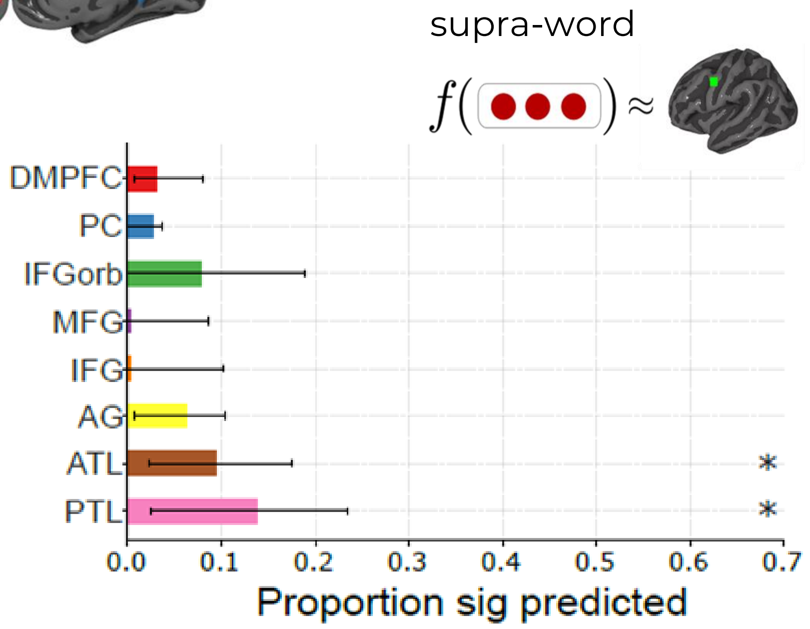
- Stimuli: one chapter of Harry Potter
- Stimulus representation: disentangled embeddings from pretrained NLP models
- Brain recording & modality: fMRI & MEG, reading




Bilateral PTL and ATL process supra-word meaning

Word-level information important for prediction of most language regions



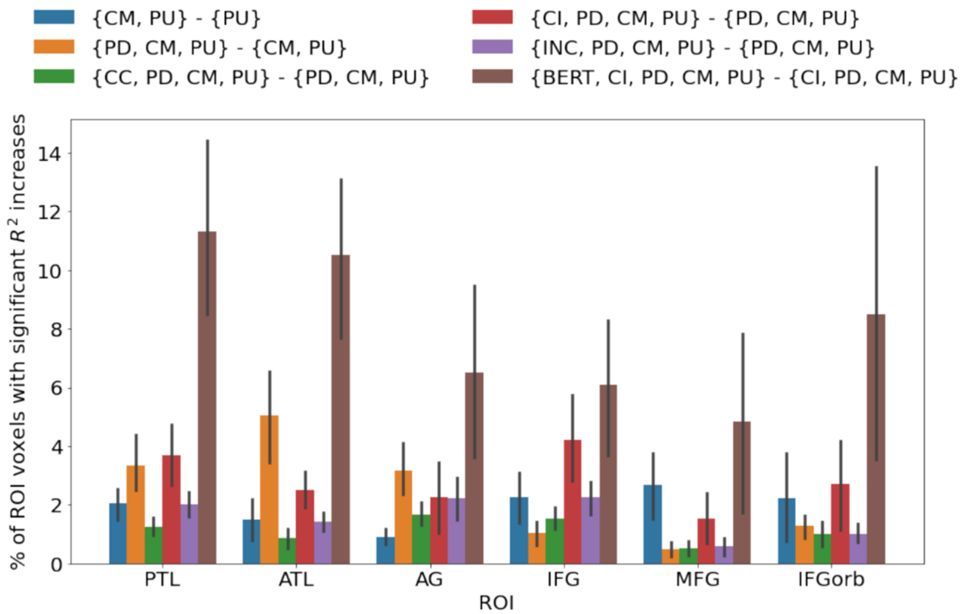



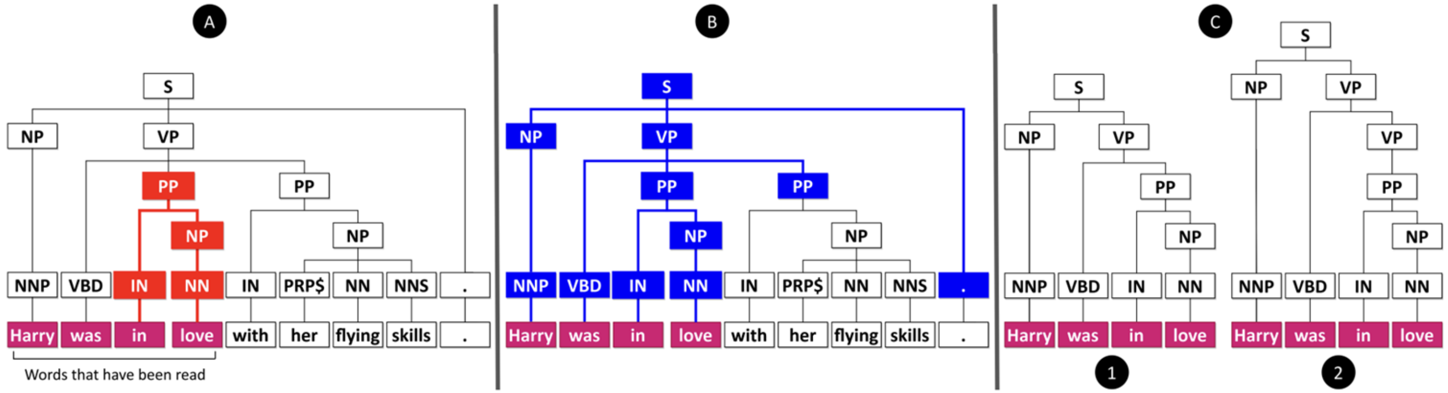


Toneva, Mariya, Tom M. Mitchell, and Leila Wehbe. "Combining computational controls with natural text reveals new aspects of meaning composition." BioRxiv (2020).

# Disentangling contributions of different info sources to brain predictions

- Stimuli: one chapter of Harry Potter
- Stimulus representation: syntactic tree representations & pretrained NLP model
- Brain recording & modality: fMRI, reading



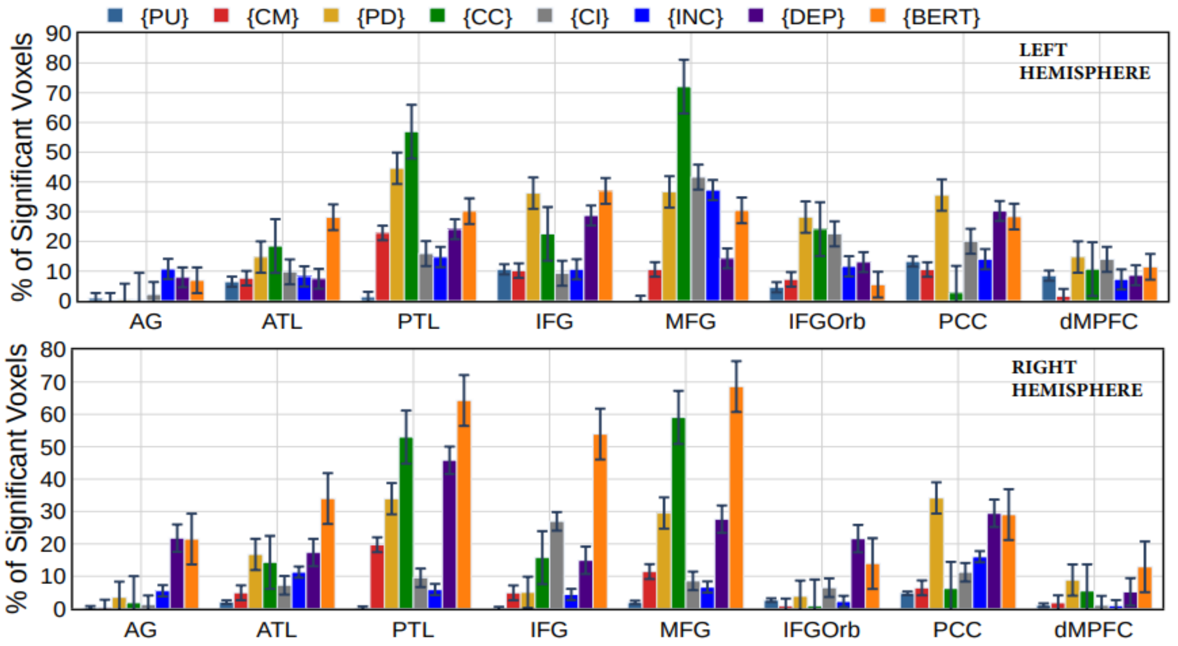




Syntactic structure-based features explain additional variance in language regions over complexity metrics

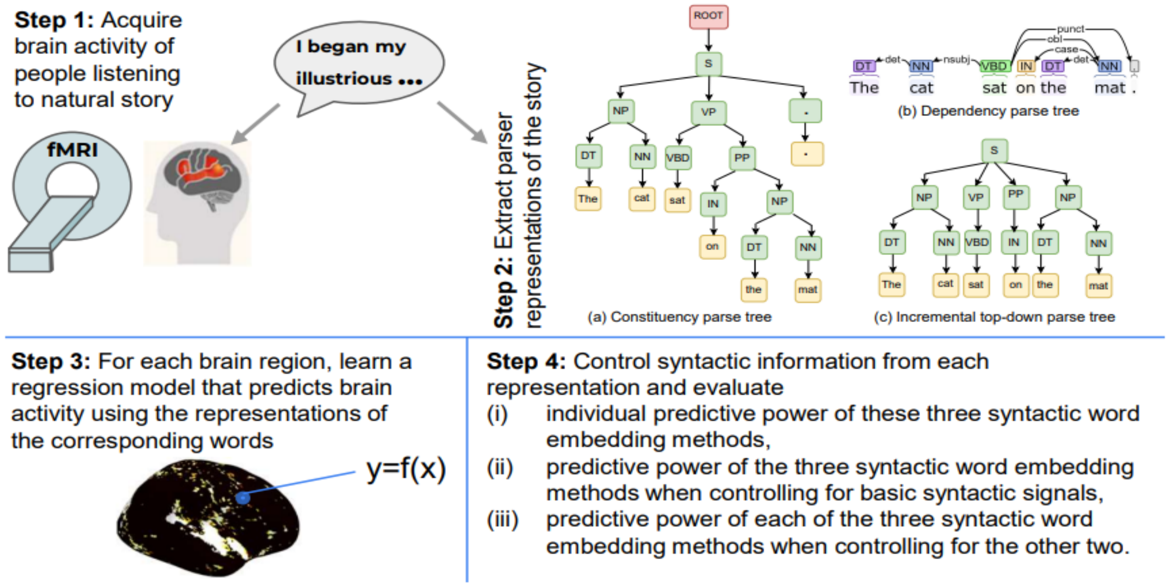
Regions predicted by syntactic and semantic are difficult to distinguish


Reddy, Aniketh Janardhan, and Leila Wehbe. "Can fMRI reveal the representation of syntactic structure in the brain?." Advances in Neural Information Processing Systems 34 (2021): 9843-9856.

# Disentangling contributions of different info sources to brain predictions




- Stimuli: Narratives
- Stimulus representation: syntactic tree representations & pretrained NLP model
- Brain recording & modality: fMRI, listening



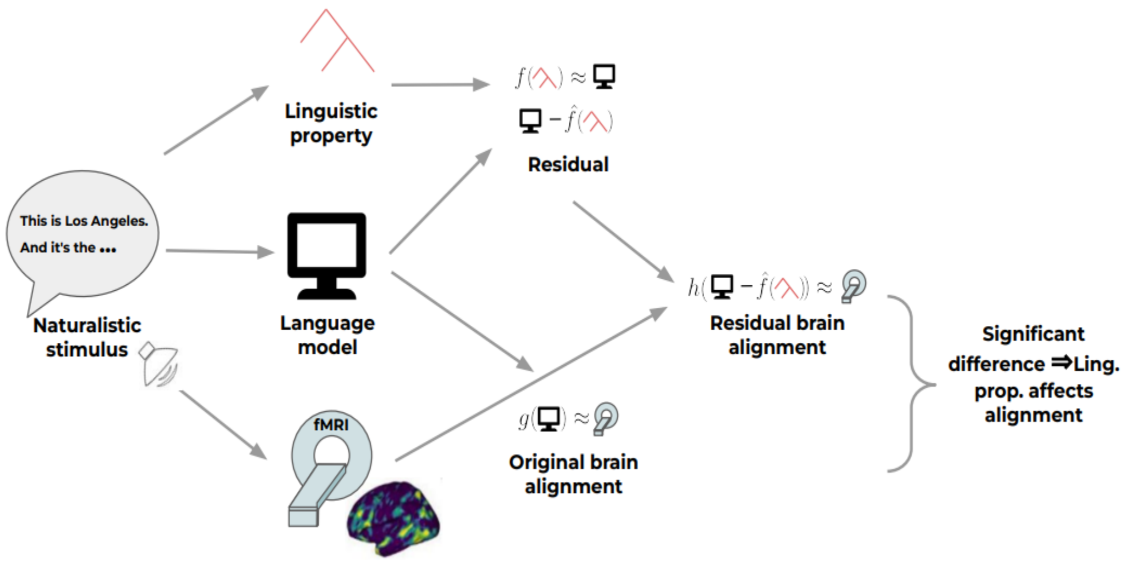

Constituency tree structure is better in temporal cortex and MFG, while Dependency structure is better in AG and PCC,

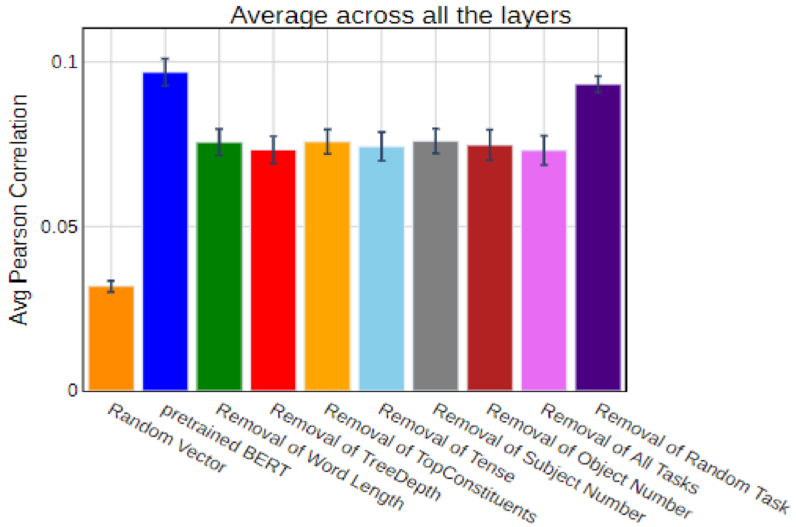
Regions predicted by syntactic and semantic are difficult to distinguish

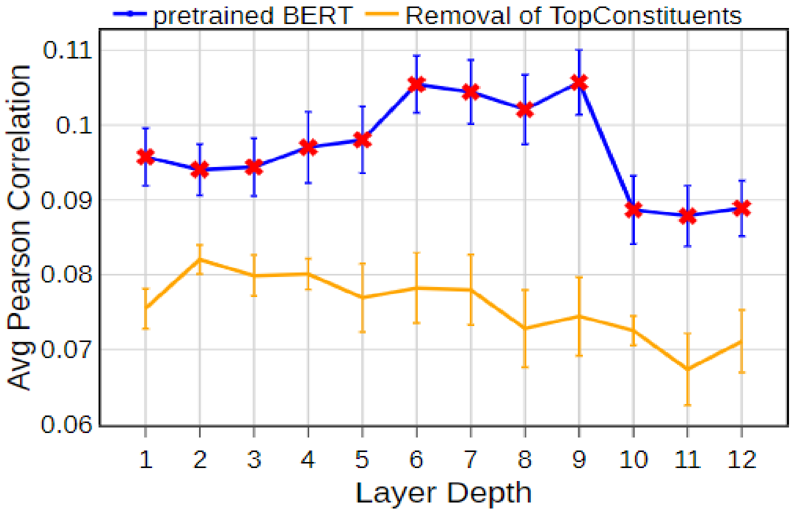
Oota, Subba Reddy et al. 2022 "How distinct are Syntactic and Semantic Representations in the Brain During Sentence Comprehension?" ACL 2023

# Joint processing of linguistic properties in brains and language models

- Stimuli: Narrative Stories
- Stimulus representation: pretrained NLP model and removal of linguistic properties
- Brain recording & modality: fMRI, Listening
- Questions: What linguistic properties underlie brain alignment, across all layers but also specifically in middle layers?



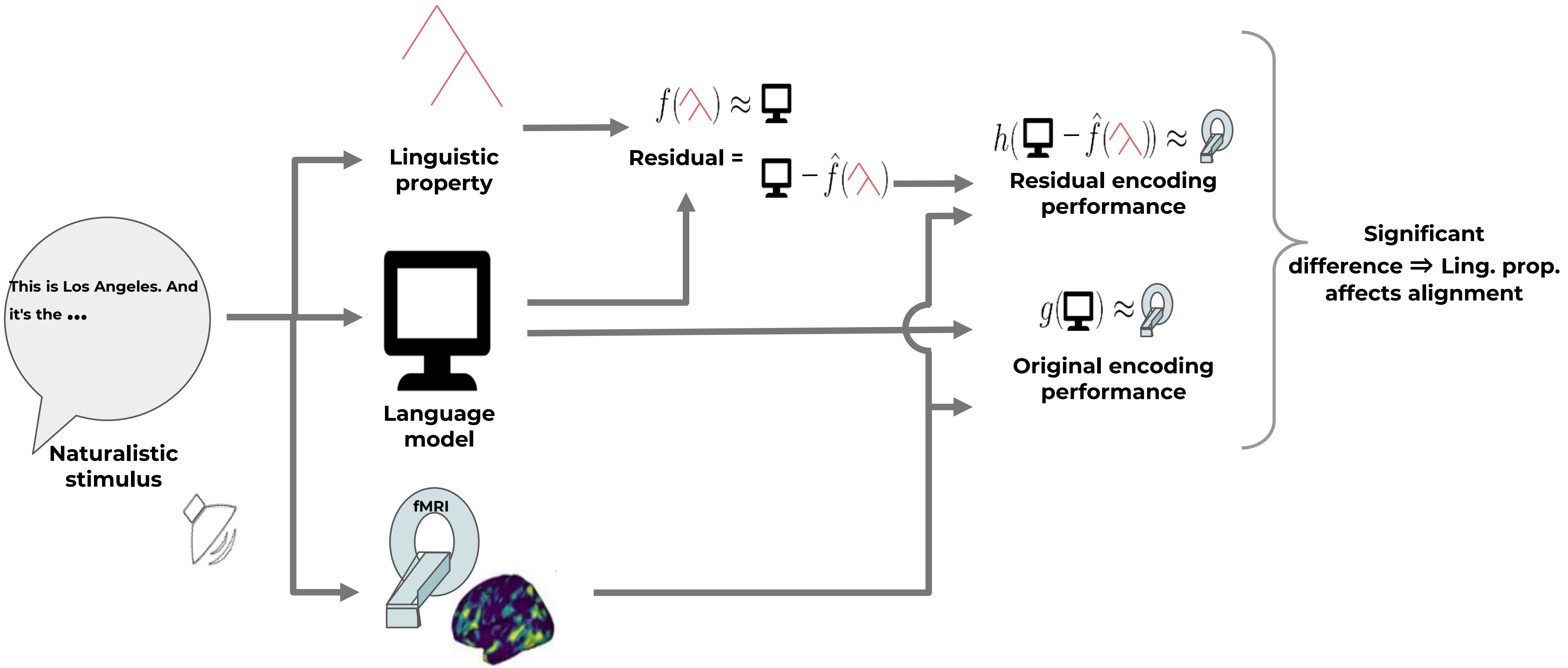




Top constituents and Tree Depth contribute the most to the alignment trend across layers

Oota, Subba Reddy, Gupta, Manish and Toneva, Mariya. "Joint processing of linguistic properties in brains and language models." NeurIPS 2023.

# What are the reasons for this observed brain alignment?

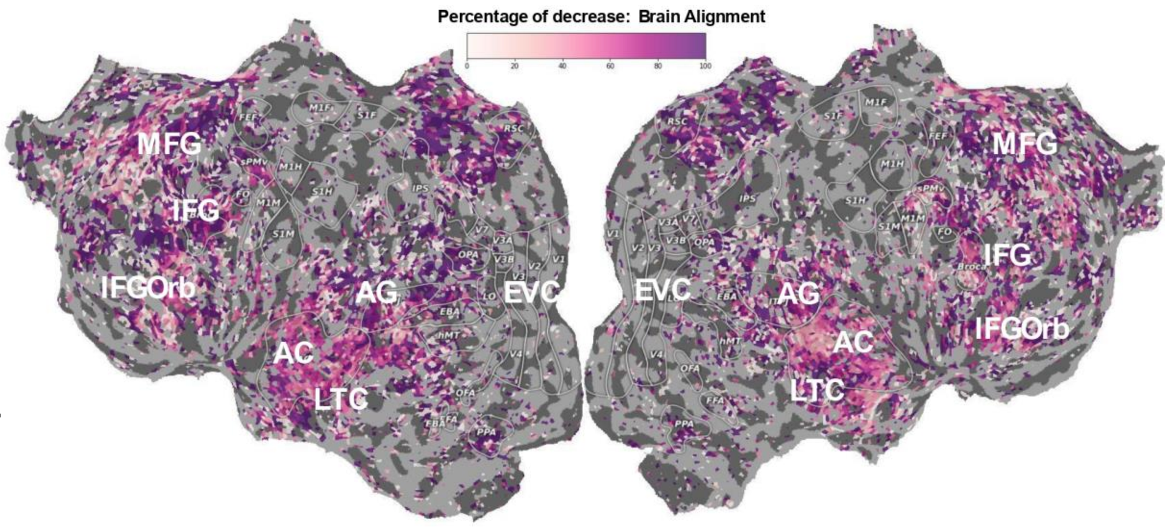
**Investigate via a perturbation approach**

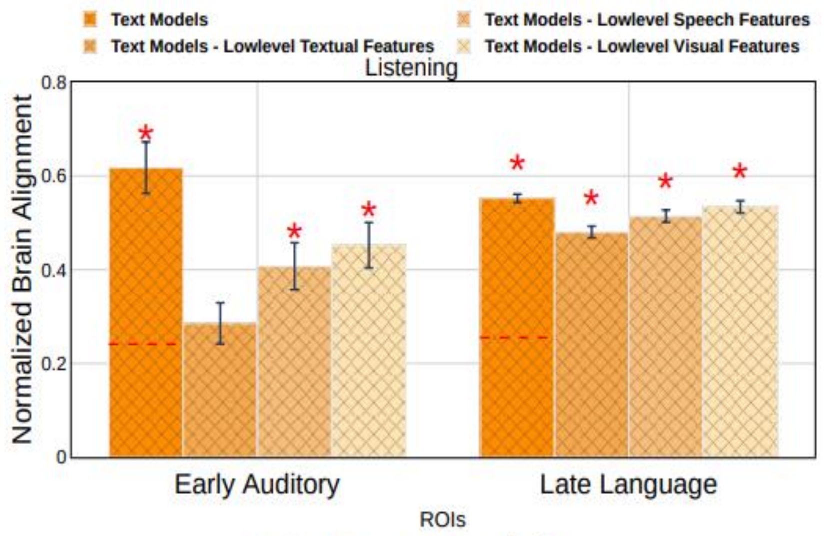



Oota, Subba Reddy, Gupta, Manish and Toneva, Mariya. "Joint processing of linguistic properties in brains and language models." NeurIPS 2023.

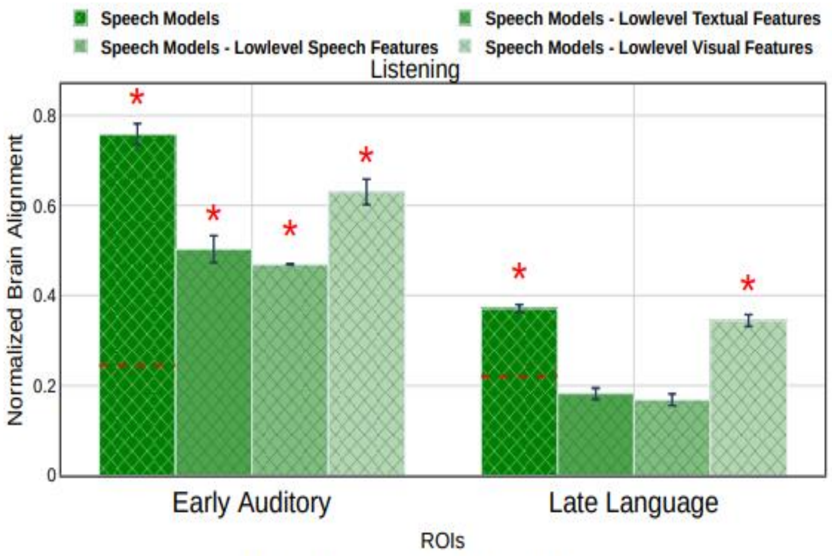
# Speech language models lack important brain relevant semantics

- Stimuli: Narrative Stories
- Stimulus representation: pretrained NLP model and speech models
- Brain recording & modality: fMRI, Reading, Listening
- **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?




(a) Text Models



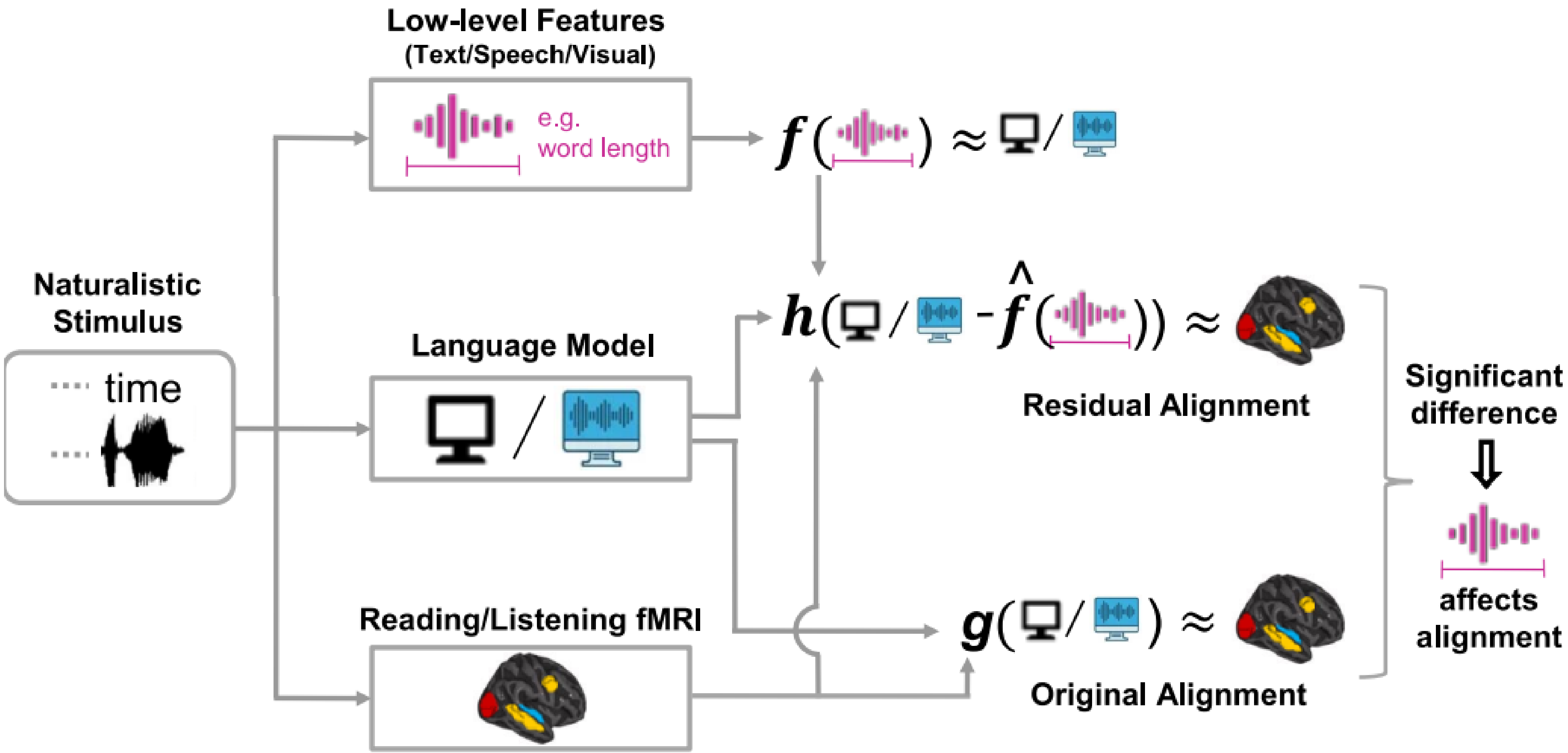

(b) Speech Models

- **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.

# What types of information lead to high brain alignment?

**Investigate via a perturbation approach**




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

# Conclusions for neuro-AI research field

1. **Text models ( ) :** alignment with **early auditory cortex (AC)** during listening and **early visual cortex (VC)** during reading is due to **low-level textual features**

2. **Speech models ( ) :** high alignment with **early auditory cortex (AC)** is only **partially explained** by **low-level speech features**.

3. Language regions predicted by **syntactic and semantic representations** are difficult to distinguish

4. **Syntactic properties** contribute the most to the alignment trend across middle layers of language model.

5. **Past word context** is crucial in obtaining significant brain predictivity results.

6. **Booksum models'** representations of Characters, Emotions and Motions are more aligned to the brain than the base models' representations.

7. **Brain-tuned models** show consistent improvement over the baselines, with biggest gains in more semantic tasks.

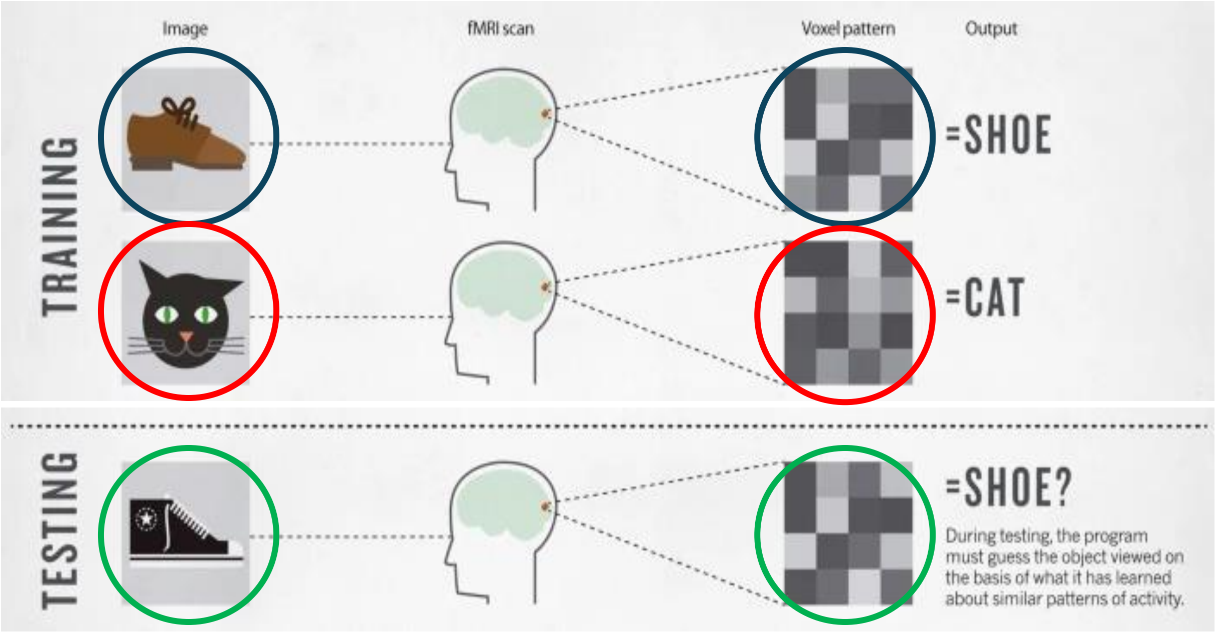
# Agenda

- Neuro-AI alignment: Introduction [1 hour 30 min]
    - Introduction to Brain encoding and decoding [30 min]
    - Types of Brain Recordings [15 min]
    - Types of Stimulus Representations [15 min]
    - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
    - Linguistic Brain Encoding [60 min]
        - Encoding schema
        - Pretrained language models and brain alignment
        - Challenges in using DL for cognitive science
        - Training DL models using brain recordings
        - Task-based language models and brain alignment
        - Disentangling Syntax and Semantics
    - Linguistic Brain Decoding [15 min]
    - Multimodal Brain Encoding [15 min]

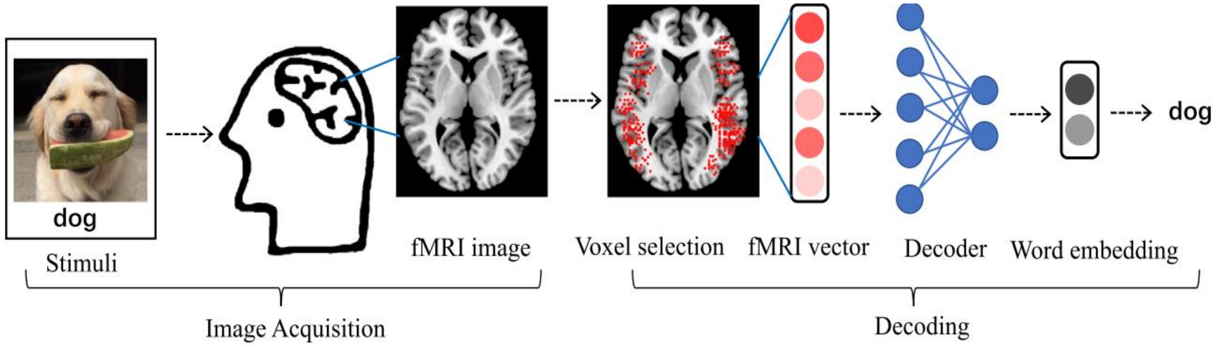
# What is Brain Decoding?

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

Visual Task

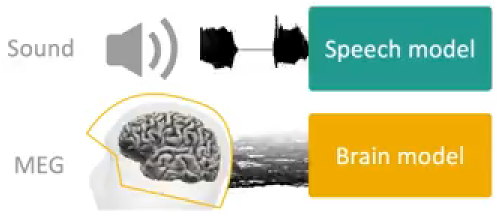



Language Task



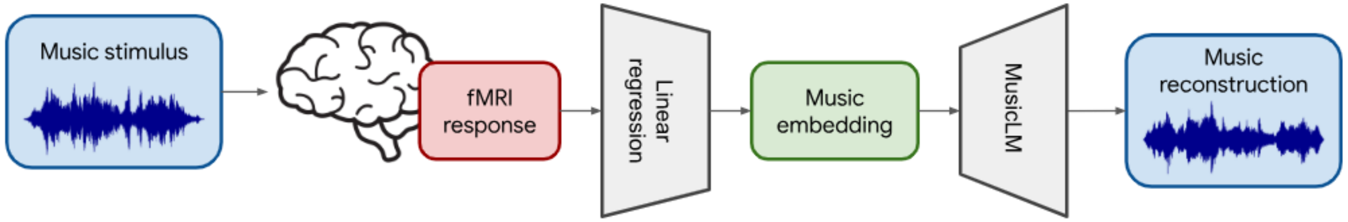

Smith et al., 2011, Wang et al. 2019

# Linguistic Decoding




**Decoding speech from non-invasive brain recordings**

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

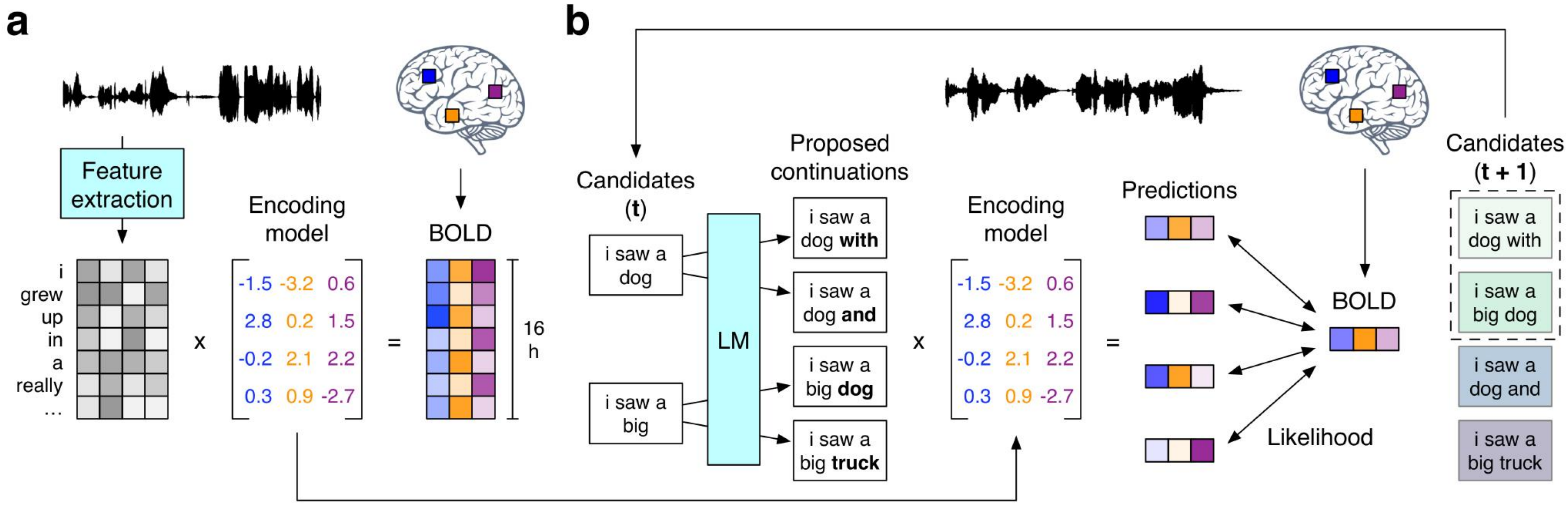



Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli & Jean-Rémi King. "Decoding speech perception from non-invasive brain recordings"  Nature Machine Intelligence 2023.

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

# 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

**c**

| Actual stimulus | Decoded stimulus |
|---|---|
| *i got up from the air mattress and pressed my face against the glass of the bedroom window expecting to see eyes staring back at me but instead finding only darkness* | i just continued to walk up to the window and open the glass i stood on my toes and peered out i didn't see anything and looked up again i saw nothing |
| *i didn't know whether to scream cry or run away instead i said leave me alone i don't need your help adam disappeared and i cleaned up alone crying* | started to scream and cry and then she just said i told you to leave me alone you can't hurt me i'm sorry and then he stormed off i thought he had left i started to cry |
| *that night i went upstairs to what had been our bedroom and not knowing what else to do i turned out the lights and lay down on the floor* | we got back to my dorm room i had no idea where my bed was i just assumed i would sleep on it but instead i lay down on the floor |
| *i don't have my driver's license yet and i just jumped out right when i needed to and she says well why don't you come back to my house and i'll give you a ride i say ok* | she is not ready she has not even started to learn to drive yet i had to push her out of the car i said we will take her home now and she agreed |

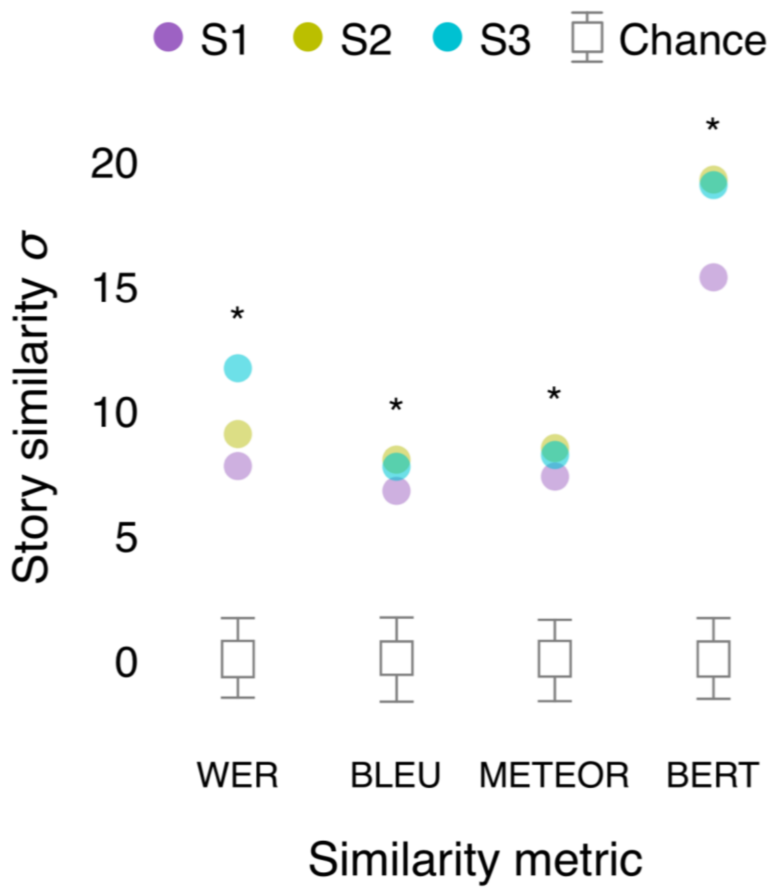
Exact

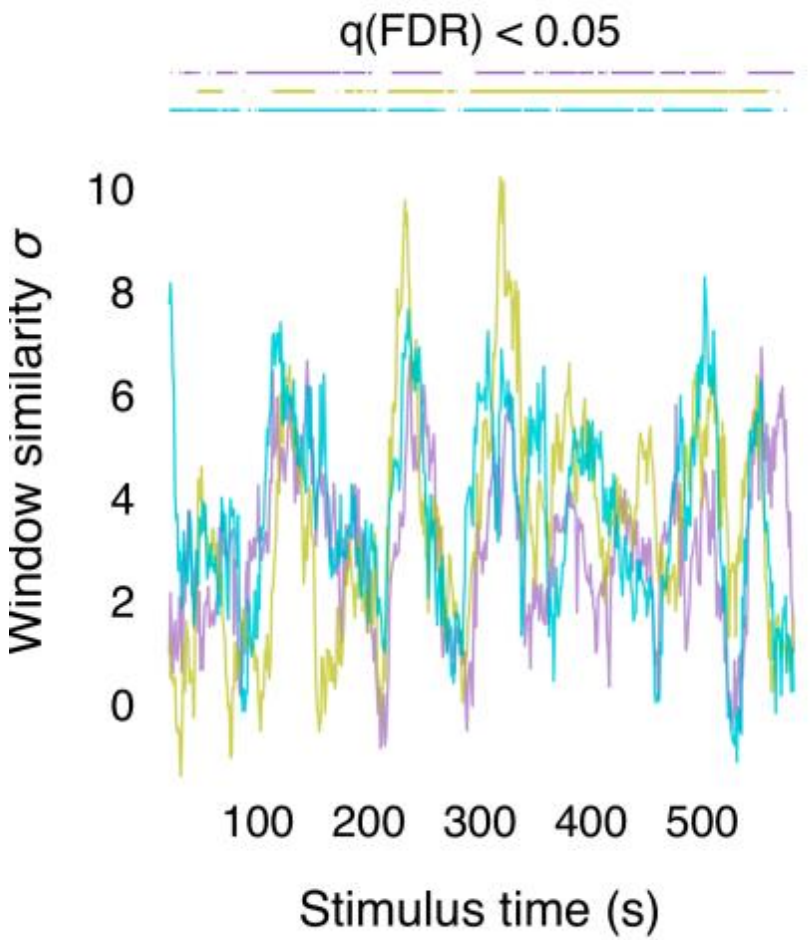
Gist

Error

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.

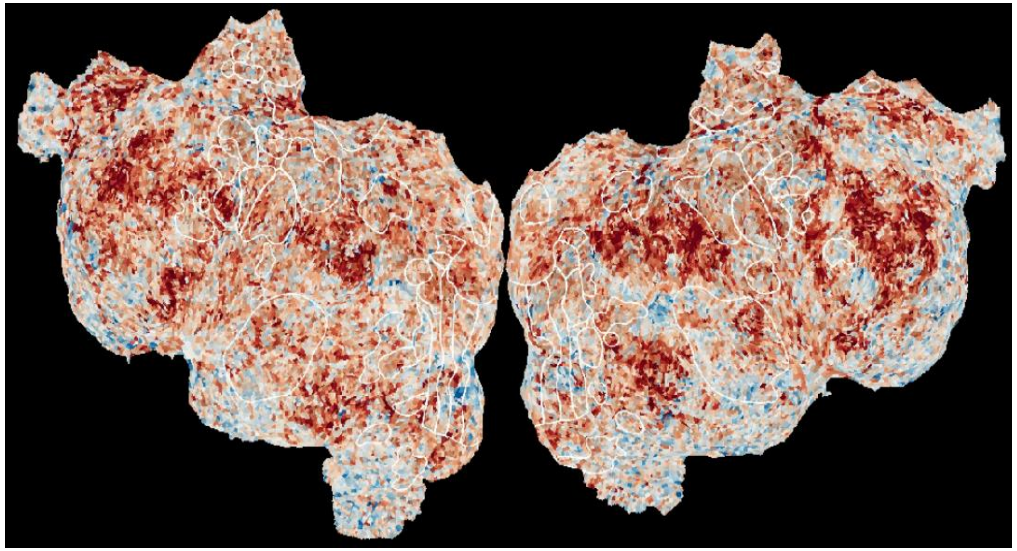
# Agenda

- Neuro-AI alignment: Introduction [1 hour 30 min]
    - Introduction to Brain encoding and decoding [30 min]
    - Types of Brain Recordings [15 min]
    - Types of Stimulus Representations [15 min]
    - Methodology [30 min]
- Coffee break [30 min]
- Language and Brain: Deep Learning for Brain Encoding and Decoding [1 hour 30 min]
    - Linguistic Brain Encoding [60 min]
        - Encoding schema
        - Pretrained language models and brain alignment
        - Challenges in using DL for cognitive science
            - Training DL models using brain recordings
            - Task-based language models and brain alignment
            - Disentangling Syntax and Semantics
    - Linguistic Brain Decoding [15 min]
    - Multimodal Brain Encoding [15 min]

# What are we talking about when we talk about "mapping stimulus to the human brain"

How our brain **separates** and **integrates** information across modalities through a hierarchy of early sensory regions to higher cognition (language regions)?

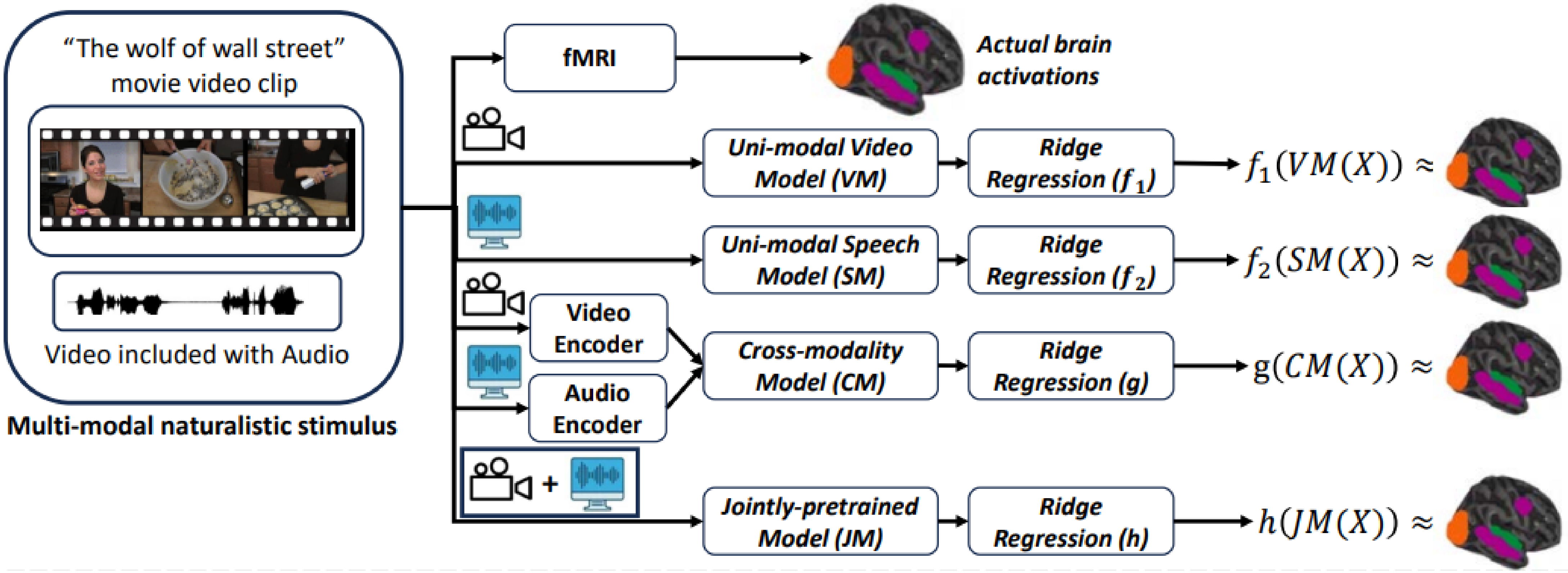
Do **concept** representations differ across **modalities**?



Where in the brain is **unimodal** and **multimodal** information represented?

What is the **shared and unique information** explained by each modality?

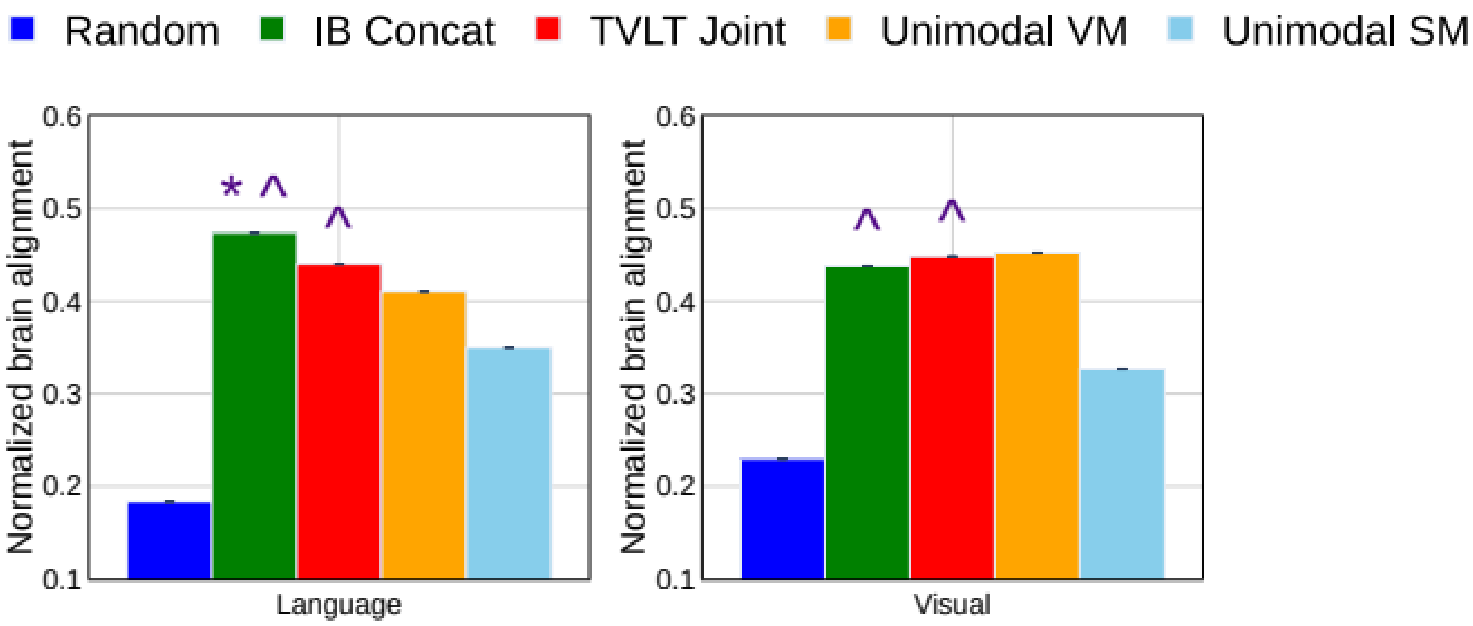
# How well multimodal models predict brain activity evoked by multimodal stimuli?




How our brain **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" Under Review.

# Surprising Trends in Brain Alignment: Unimodal vs. Multimodal Models




- Multi-modal effects: In general, multimodal models have better predictivity in the language regions
- Unimodal effects: Unimodal models have higher predictivity in the early sensory regions (visual and auditory).

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

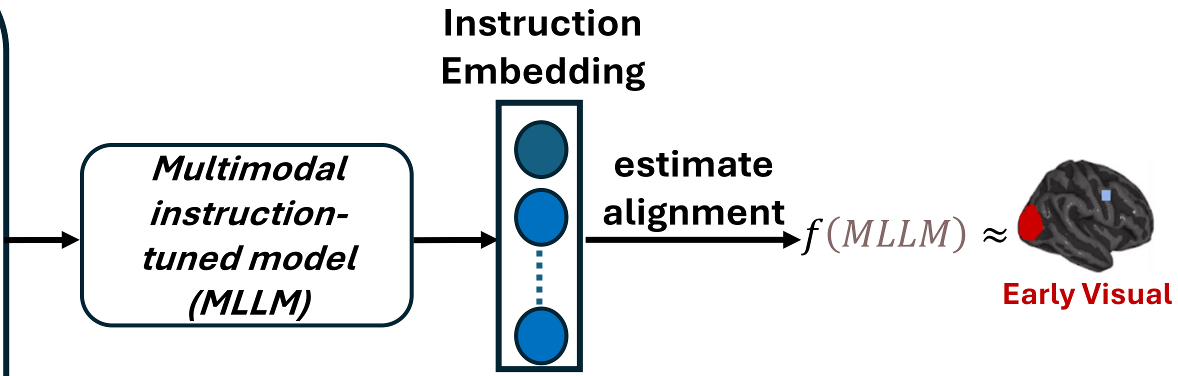
# Correlating instruction-tuning (in multimodal models) with vision-language processing (in the brain)

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

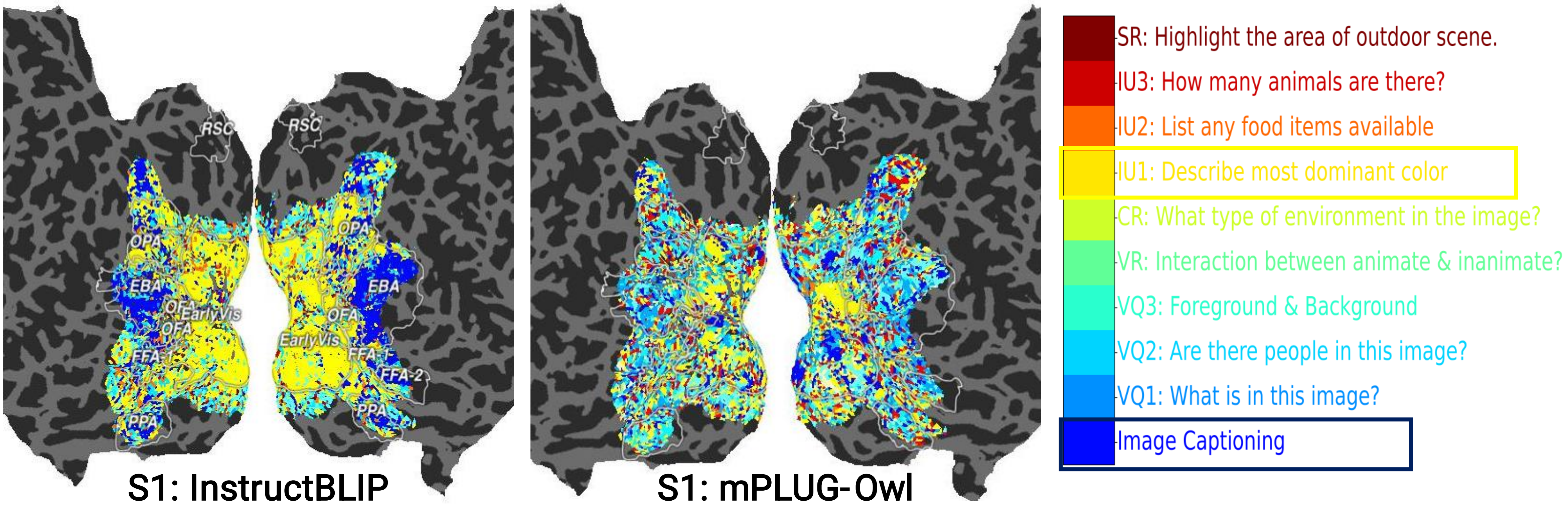



| Task | Description |
|---|---|
| Image Understanding | IU1: Describe the most dominant color in the image<br>IU2: List any food items visible.<br>IU3: How many animals are there in the image? |
| Visual Question Answering | VQ1: What is in this image?<br>VQ2: Are there any people in this image? If yes, describe them.<br>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? |

Do multimodal instruction-tuned models prompted using natural language instructions lead to better brain alignment and differentiate instruction-specific representations?

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)" Under Review.

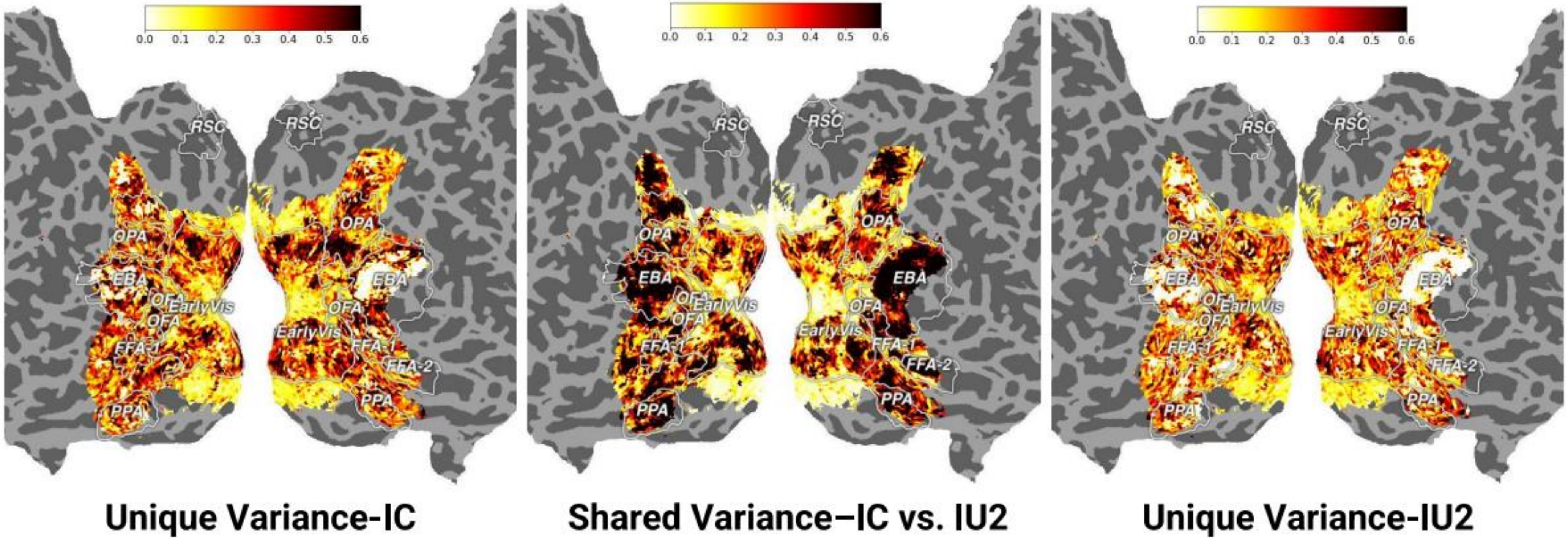
# Which task-specific instructions are highly correlated to visual function localizers?




- 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)" Under Review.

# 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, while IC retains some unique variance.
- High overlap between IC and IU2 in higher visual areas but lower overlap in early visual cortex.

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)" Under Review.

# Conclusions for neuro-AI research field

1. Both **cross-modal and jointly pretrained models** demonstrate significantly improved brain alignment with **language regions** compared to visual regions when analyzed against unimodal video data.

2. **Multi-modal models** to capture **additional information**—either through knowledge transfer or integration between modalities—which is crucial for multi-modal brain alignment

3. The **differences between the models** in terms of **architectural variability** and **variability in pretraining methods**, this suggests that future work could benefit from more tightly controlled comparisons to better isolate the effects of these factors.

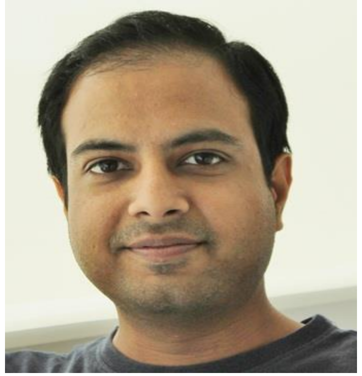
4. Several **task-specific** instructions leading to improved brain alignment between fMRI recordings and MLLMs, **not all instructions** were relevant for brain alignment.

# Collaborators


Subba Reddy Oota


Khushbu Pahwa


Maneesh Singh


Manish Gupta


Mariya Toneva


Bapi Raju Surampudi


Fatma Deniz


Xavier Hinaut


Frederic Alexandre