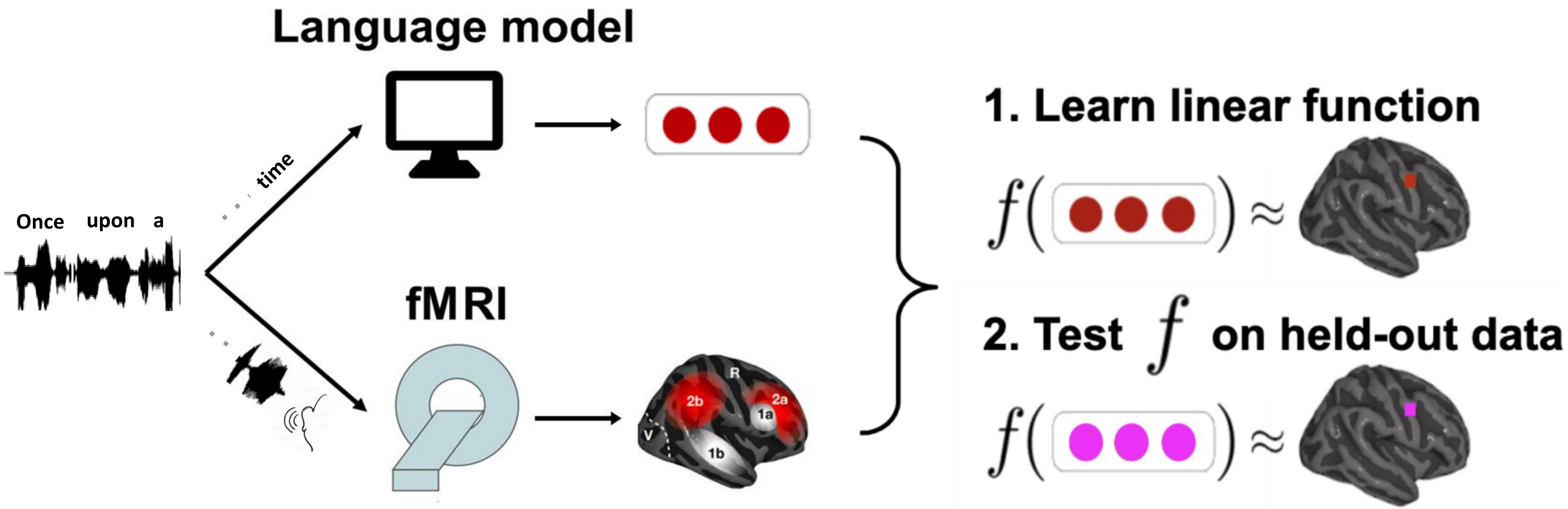
Characterizing similarities and differences between language processing in brains and language models.

Subba Reddy Oota
Ph.D. Student
Inria Bordeaux, France

Language models (LMs) predict brain activity evoked by complex language (e.g. listening a story) to an impressive degree



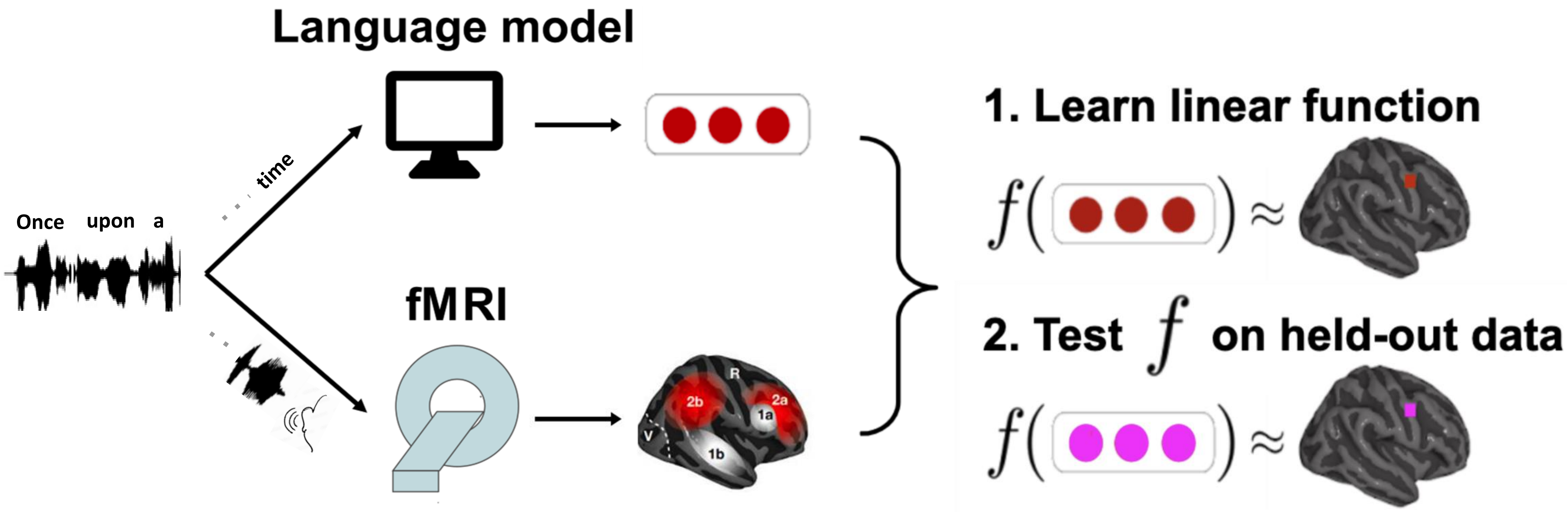
Brain alignment of a LM ⇒ how similar its representations are to a human brain's

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
…

Language models (LMs) predict brain activity evoked by complex language (e.g. listening a story) to an impressive degree

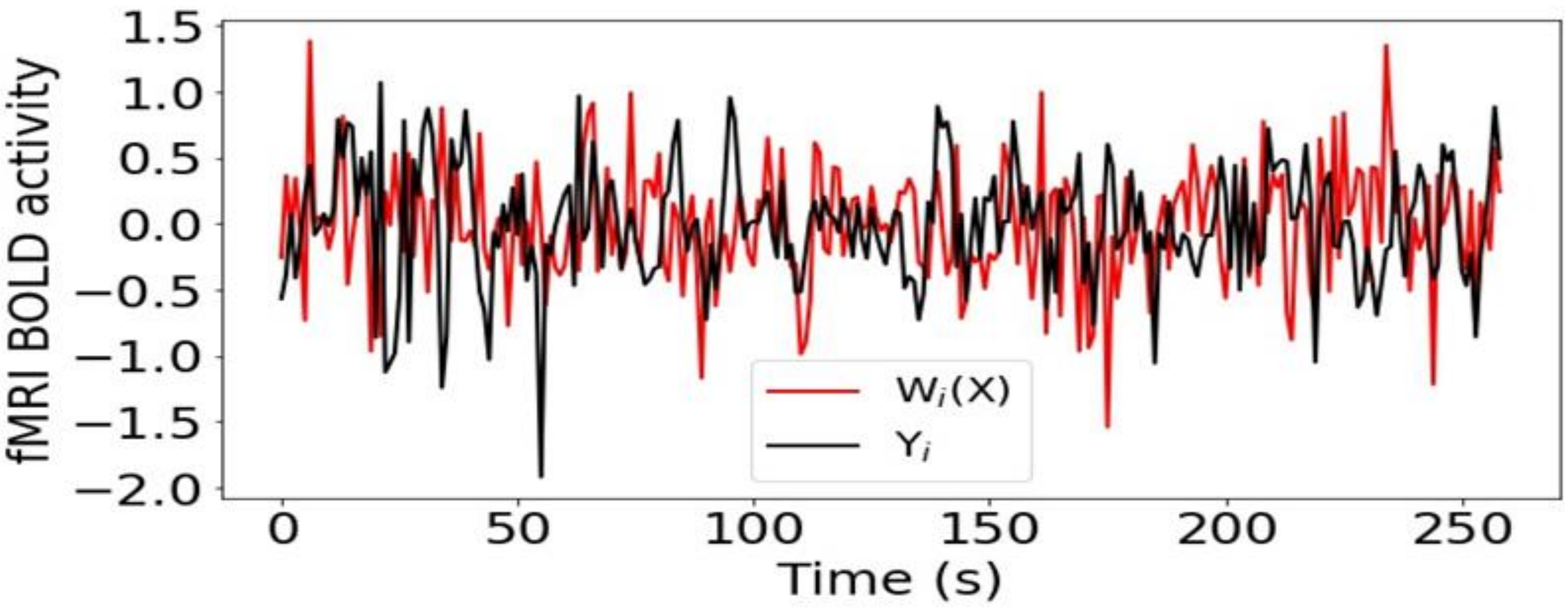


Wehbe et al. 2014,
Jain and Huth 2018,
Gauthier and Levy 2019

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

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

Language models (LMs) predict brain activity evoked by complex language (e.g. listening a story) to an impressive degree



Brain alignment of a LM ⇒ Why do language models have better brain alignment? What are the reasons?

Jain and Huth. Incorporating context into language encoding models for fMRI. (NeurIPS 2018)
Toneva and Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). (NeurIPS 2019)

Focusing on two research questions

What are the reasons for this better brain alignment with language model representations?

What types of information underlie the brain alignment of language models observed across brain regions?

What are the reasons behind better similarity between language models and brains?

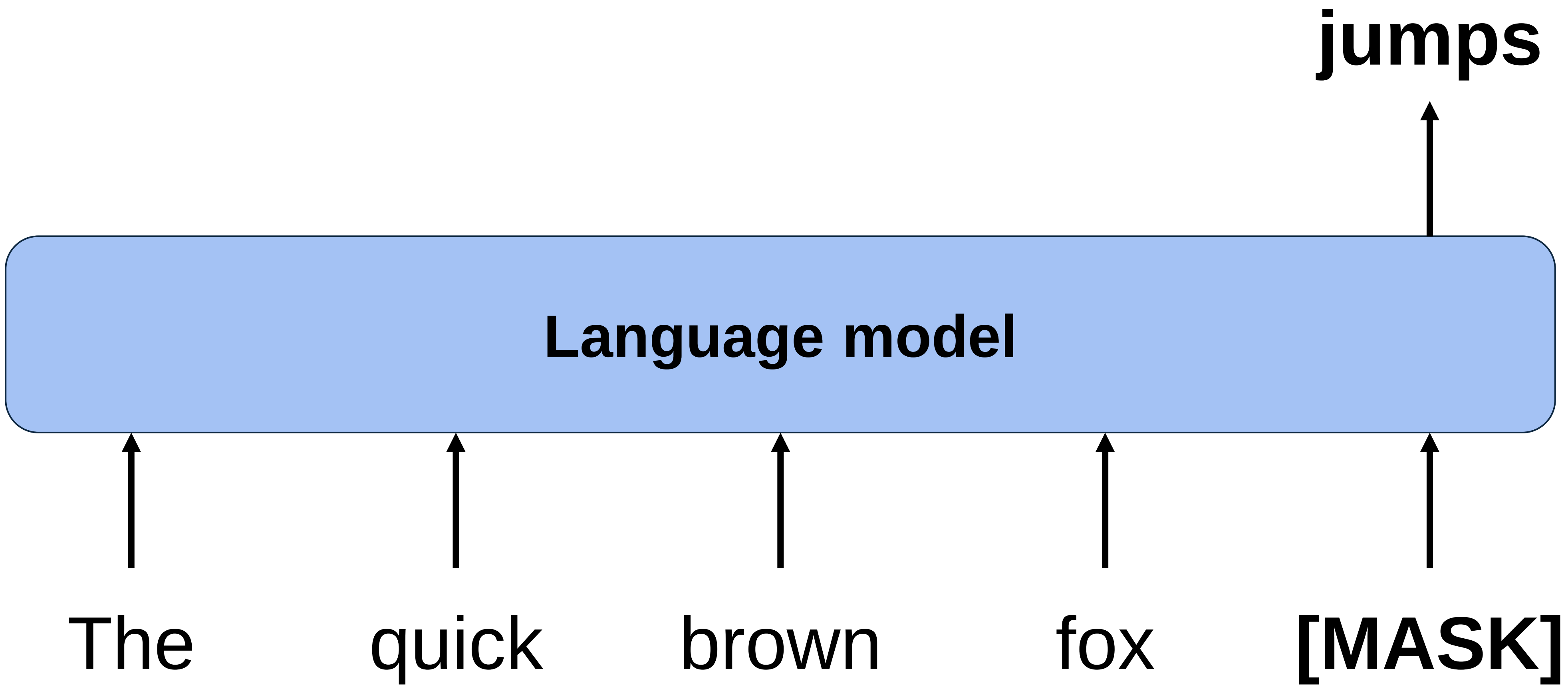
Joint processing of linguistic properties in brains and language models

Subba Reddy Oota Manish Gupta Mariya Toneva

Inria Microsoft MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

NEURAL INFORMATION PROCESSING SYSTEMS

Language models (LMs) are trained to predict missing words



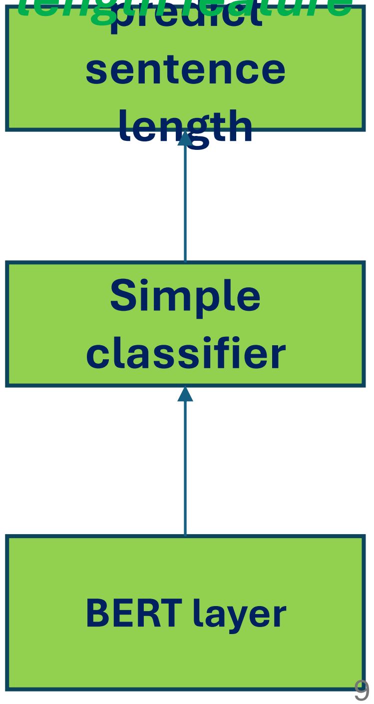
Interpreting BERT and beyond

- **Can we unveil the representations learned by BERT to linguistics structure?**
- Understand the reason behind the success of BERT but also its limitations.
- Guide the design of improved architectures.

Hierarchy of Linguistic Info - Setting

- Conneau et al., ACL'18 - Build diagnostic classifier to predict if a linguistic property is encoded in the given sentence representation.
- Features:
  - **Surface** – Sentence Length, Word Content
  - **Syntactic** – Bigram shift, Tree depth, Top constituent
  - **Semantic** – Tense, Subject Number, Object Number, Coordination Inversion and Semantic Odd Man Out.

If the prediction accuracy is good, then the model might be capturing the sentence length feature



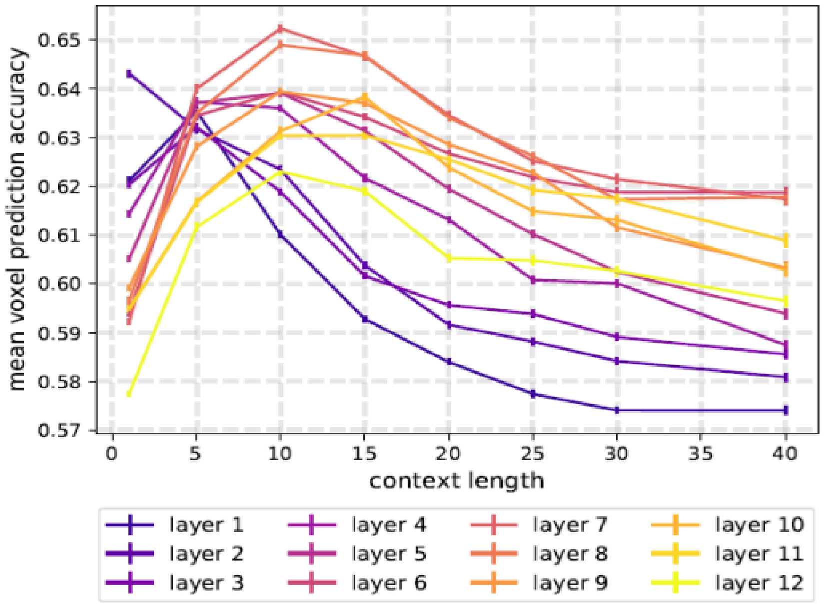
Language models (LMs) are trained to predict missing words

	Surface		Syntactic			Semantic				
Layer	**SentLen** (Surface)	**WC** (Surface)	**TreeDepth** (Syntactic)	**TopConst** (Syntactic)	**BShift** (Syntactic)	**Tense** (Semantic)	**SubjNum** (Semantic)	**ObjNum** (Semantic)	**SOMO** (Semantic)	**CoordInv** (Semantic)
1	93.9 (2.0)	24.9 (24.8)	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	49.9 (-0.1)	53.9 (3.9)
2	95.9 (3.4)	65.0 (64.8)	40.6 (11.3)	71.3 (16.1)	55.8 (5.8)	85.9 (23.5)	82.5 (15.3)	80.6 (17.1)	53.8 (4.4)	58.5 (8.5)
3	**96.2 (3.9)**	66.5 (66.0)	39.7 (10.4)	71.5 (18.5)	64.9 (14.9)	86.6 (23.8)	82.0 (14.6)	80.3 (16.6)	55.8 (5.9)	59.3 (9.3)
4	94.2 (2.3)	**69.8 (69.6)**	39.4 (10.8)	71.3 (18.3)	74.4 (24.5)	87.6 (25.2)	81.9 (15.0)	81.4 (19.1)	59.0 (8.5)	58.1 (8.1)
5	92.0 (0.5)	69.2 (69.0)	40.6 (11.8)	81.3 (30.8)	81.4 (31.4)	89.5 (26.7)	85.8 (19.4)	81.2 (18.6)	60.2 (10.3)	64.1 (14.1)
6	88.4 (-3.0)	63.5 (63.4)	**41.3 (13.0)**	83.3 (36.6)	82.9 (32.9)	89.8 (27.6)	**88.1 (21.9)**	82.0 (20.1)	60.7 (10.2)	71.1 (21.2)
7	83.7 (-7.7)	56.9 (56.7)	40.1 (12.0)	**84.1 (39.5)**	83.0 (32.9)	89.9 (27.5)	87.4 (22.2)	**82.2 (21.1)**	61.6 (11.7)	74.8 (24.9)
8	82.9 (-8.1)	51.1 (51.0)	39.2 (10.3)	84.0 (39.5)	83.9 (33.9)	89.9 (27.6)	87.5 (22.2)	81.2 (19.7)	62.1 (12.2)	76.4 (26.4)
9	80.1 (-11.1)	47.9 (47.8)	38.5 (10.8)	83.1 (39.8)	**87.0 (37.1)**	**90.0 (28.0)**	87.6 (22.9)	81.8 (20.5)	63.4 (13.4)	**78.7 (28.9)**
10	77.0 (-14.0)	43.4 (43.2)	38.1 (9.9)	81.7 (39.8)	86.7 (36.7)	89.7 (27.6)	87.1 (22.6)	80.5 (19.9)	63.3 (12.7)	78.4 (28.1)
11	73.9 (-17.0)	42.8 (42.7)	36.3 (7.9)	80.3 (39.1)	86.8 (36.8)	89.9 (27.8)	85.7 (21.9)	78.9 (18.6)	64.4 (14.5)	77.6 (27.9)
12	69.5 (-21.4)	49.1 (49.0)	34.7 (6.9)	76.5 (37.2)	86.4 (36.4)	89.5 (27.7)	84.0 (20.2)	78.7 (18.4)	**65.2 (15.3)**	74.9 (25.4)

BERT composes a **hierarchy of linguistic signals** ranging from surface to semantic features.

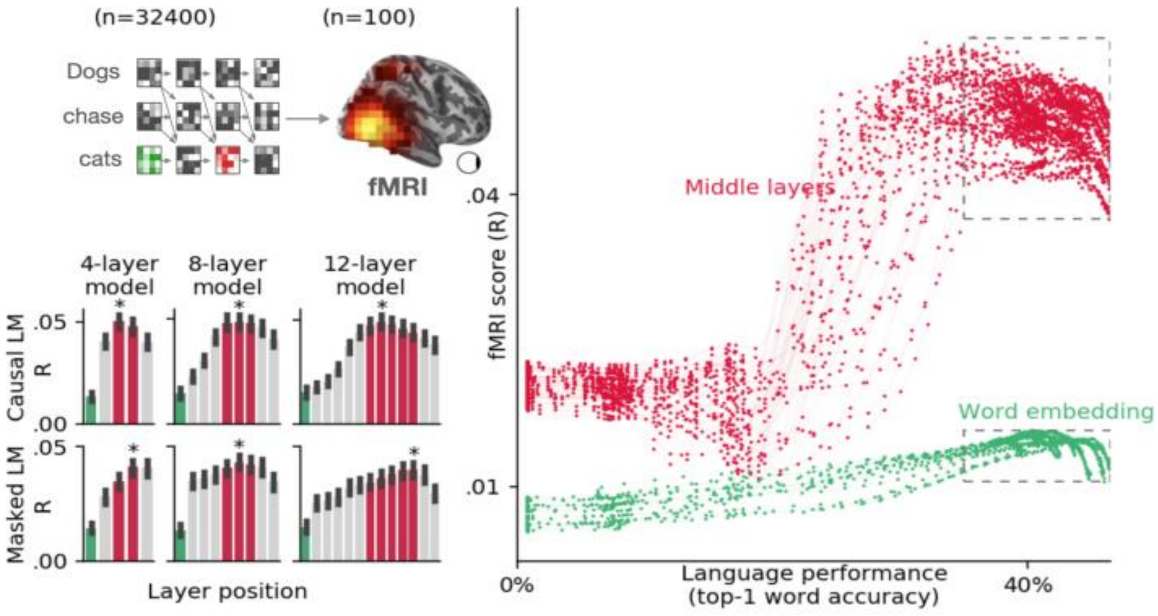
The strongest alignment with high-level language brain regions has consistently been observed in middle layers

Toneva et al. 2019



BERT

Caucheteux et al. 2022

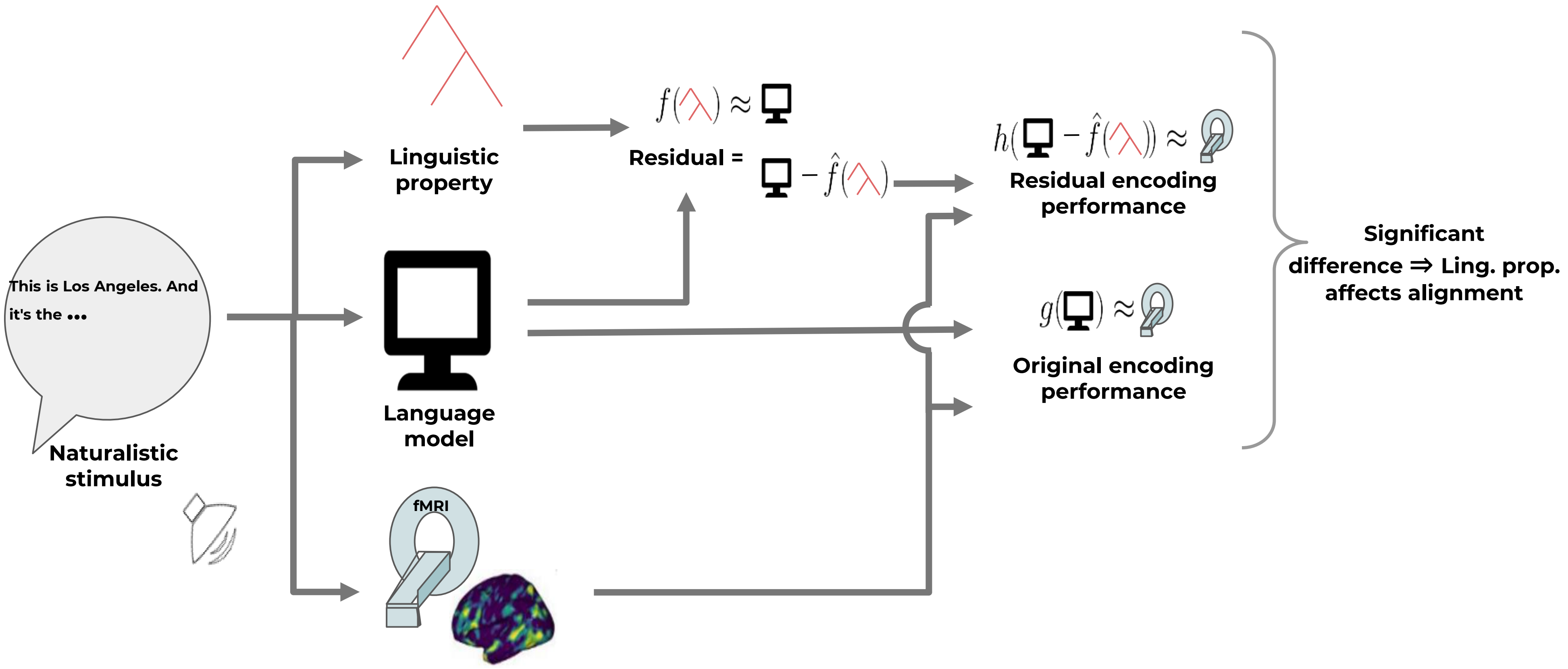


XLM

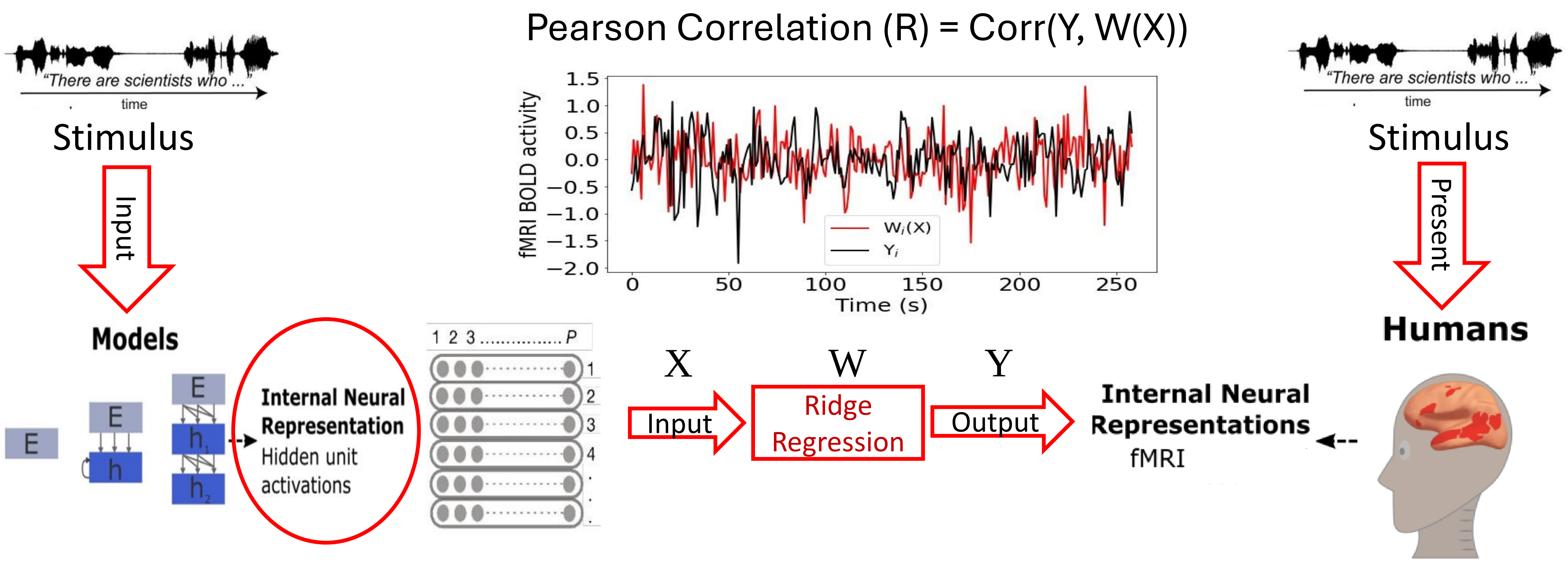
Across several types of large NLP systems, best alignment with fMRI in middle layers

What are the reasons for this observed brain alignment?

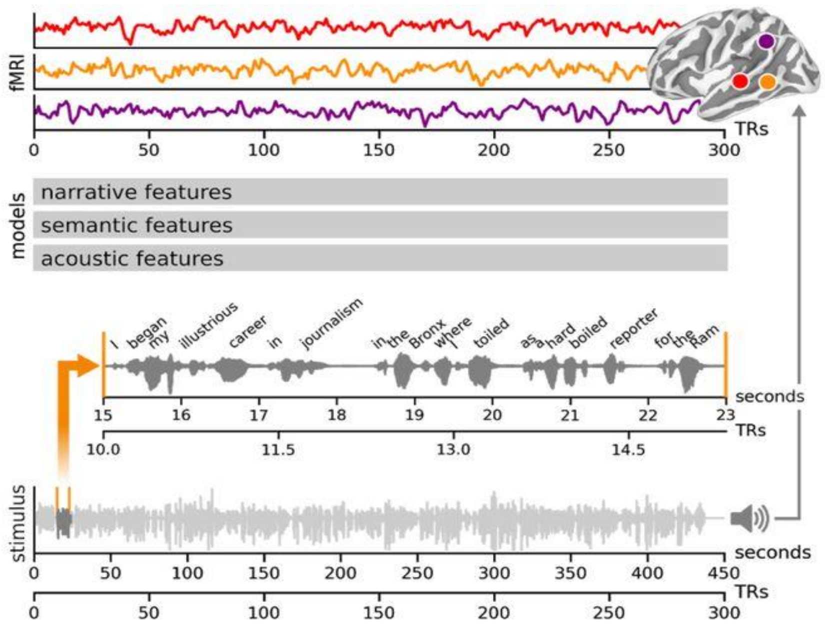
Investigate via a perturbation approach



Brain Encoding Schema?



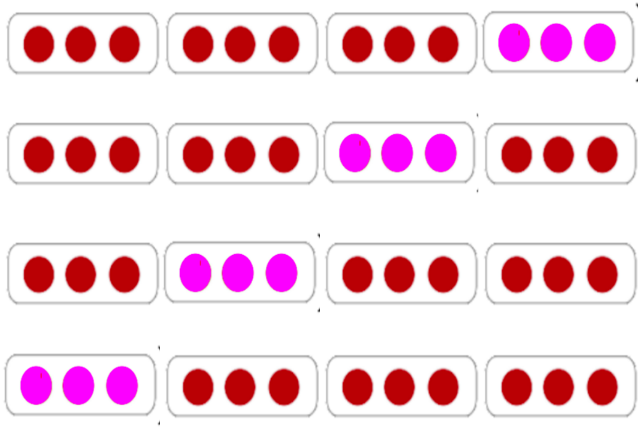
Brain alignment – 4-fold Cross-Validation + Ridge regression



(1) Narratives 21st year story

- 8267 words x 768 ⇒ LM representations
  - Downsample from word-level representations to TR-level (taken every 1.5s)
- 2226 fMRI time intervals x 768
  - Concatenate LM representations for previous 8 TRs ⇒ fMRI response from brain activity peaks about 6 seconds after stimulus onset
- 2226 fMRI intervals x 81924 voxels ⇒ fMRI predictions (same dimensions as actual brain activity)
- 18 participants

(2) Brain Alignment: Dataset Curation



(3) 4-fold Cross-Validation

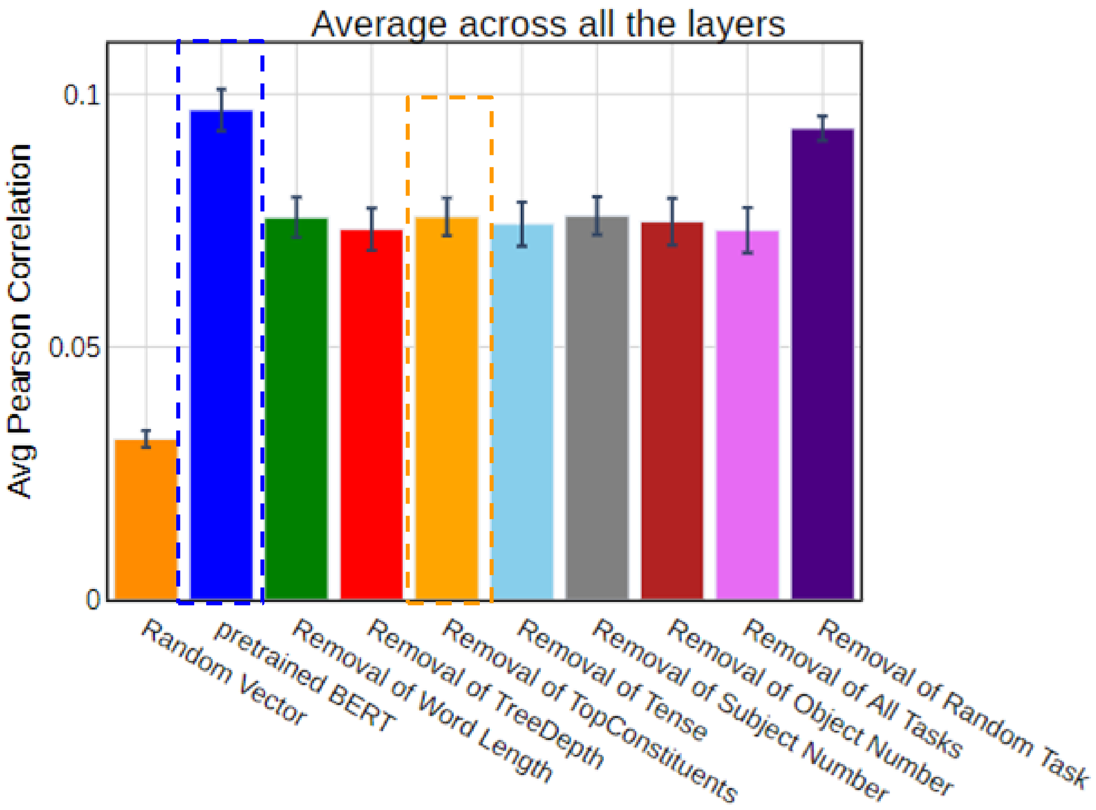
What are the reasons behind the success of LMs?

	Surface		**Syntactic**				**Semantic**					
Layers	Word Length 3-classes (Surface)		TreeDepth 3-classes (Syntactic)		TopConstituents 2-classes (Syntactic)		Tense 2-classes (Semantic)		Subject Number 2-classes (Semantic)		Object Number 2-classes (Semantic)	
	before	after	before	after	before	after	before	after	before	after	before	after
1	**74.67**	43.28	76.30	42.93	77.15	47.28	87.00	59.25	92.10	49.95	93.28	47.31
2	69.83	42.44	76.72	38.88	78.60	42.75	87.18	48.25	92.32	55.50	93.47	54.59
3	72.31	46.19	75.76	40.33	77.81	48.85	87.42	44.26	93.04	48.55	93.80	49.76
4	71.34	46.43	75.94	38.63	78.36	48.00	88.09	42.56	93.50	50.12	94.90	50.06
5	72.67	46.97	76.00	40.88	78.60	45.28	88.39	44.26	94.05	49.88	93.59	51.45
6	70.38	44.37	**79.02**	41.89	**80.23**	43.47	87.17	44.44	94.98	55.08	94.50	54.17
7	72.98	46.55	77.93	41.23	**80.23**	46.43	88.69	42.62	95.88	50.24	94.62	47.58
8	72.67	44.67	76.07	40.08	78.90	46.86	87.42	44.56	96.10	50.24	**95.10**	50.18
9	70.50	45.28	77.15	42.62	79.87	44.55	88.27	47.22	96.38	52.78	94.56	49.27
10	72.91	47.93	76.90	41.78	78.17	47.76	**88.94**	45.47	96.06	53.68	94.50	50.30
11	70.07	46.67	77.27	45.47	77.69	45.77	87.24	48.43	**96.94**	53.44	94.92	49.52
12	71.77	42.93	76.39	46.61	78.29	48.67	86.88	45.10	94.03	51.45	93.95	48.73

Successful removal of linguistic properties from pretrained BERT

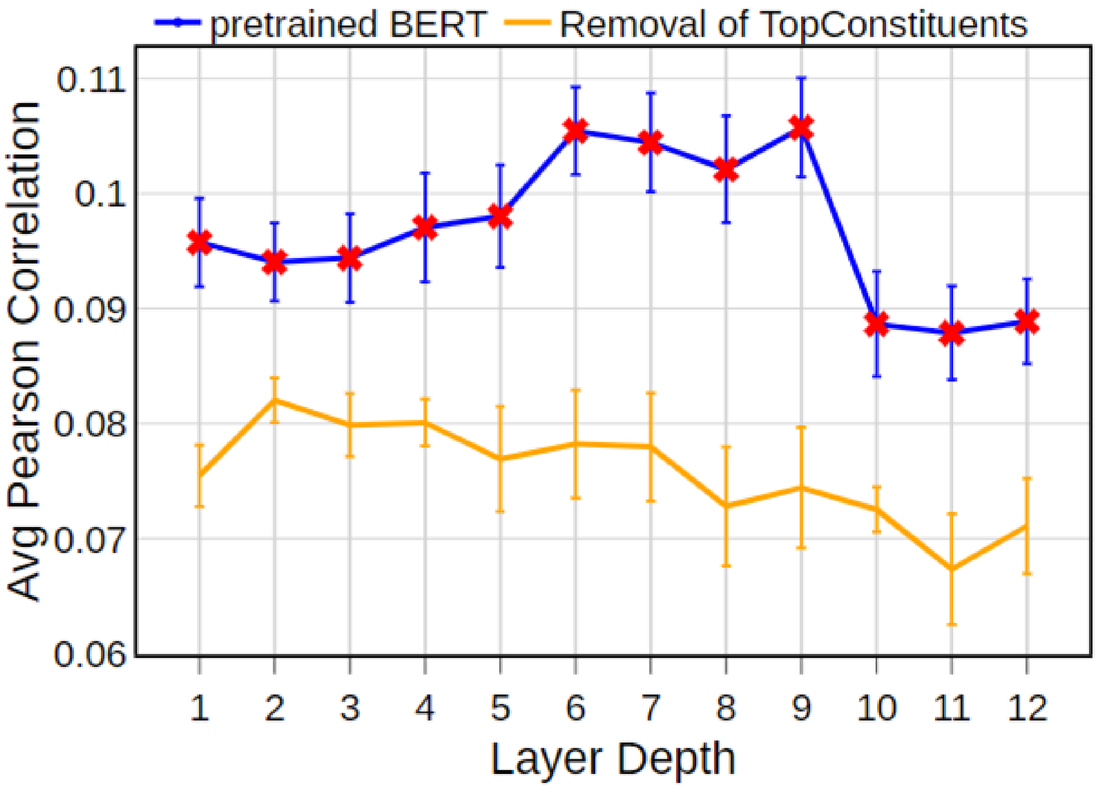
Does the removal of a linguistic property affects the alignment between a language model and the brain across all layers?

Result-1



Removal of each linguistic property leads to a significant decrease in brain alignment on average across layers.

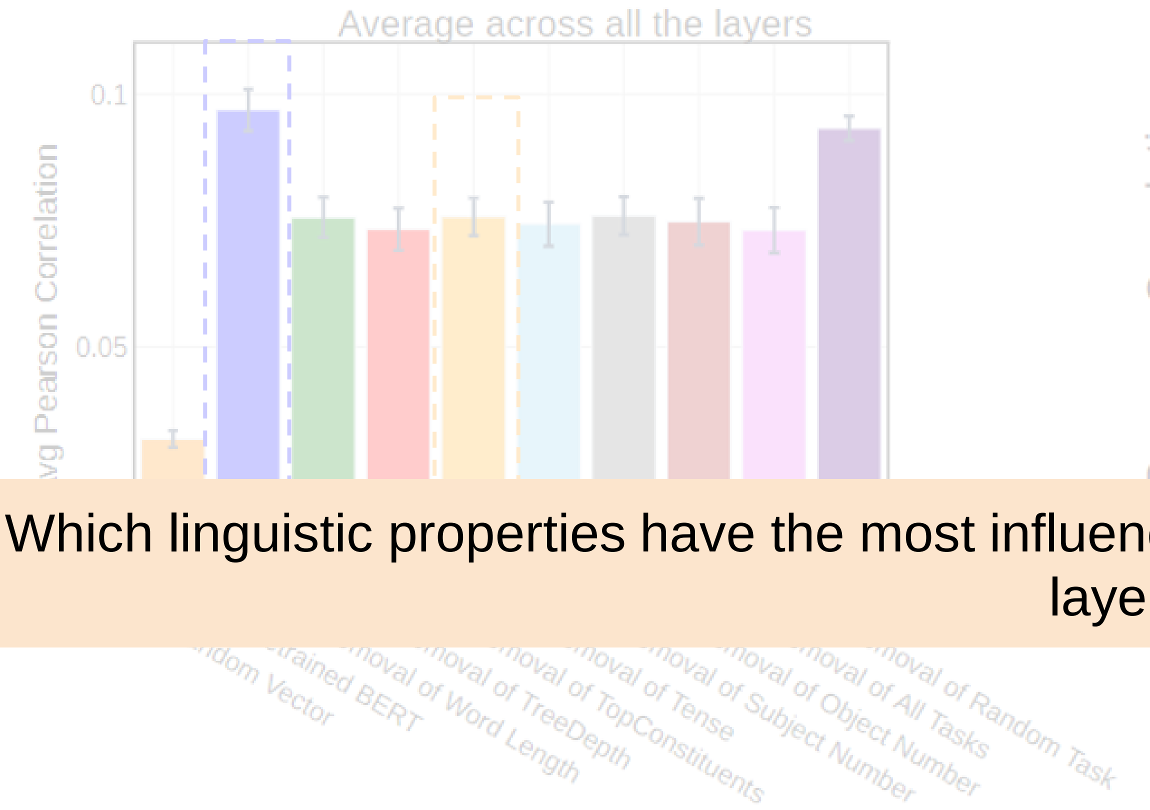
Result-1

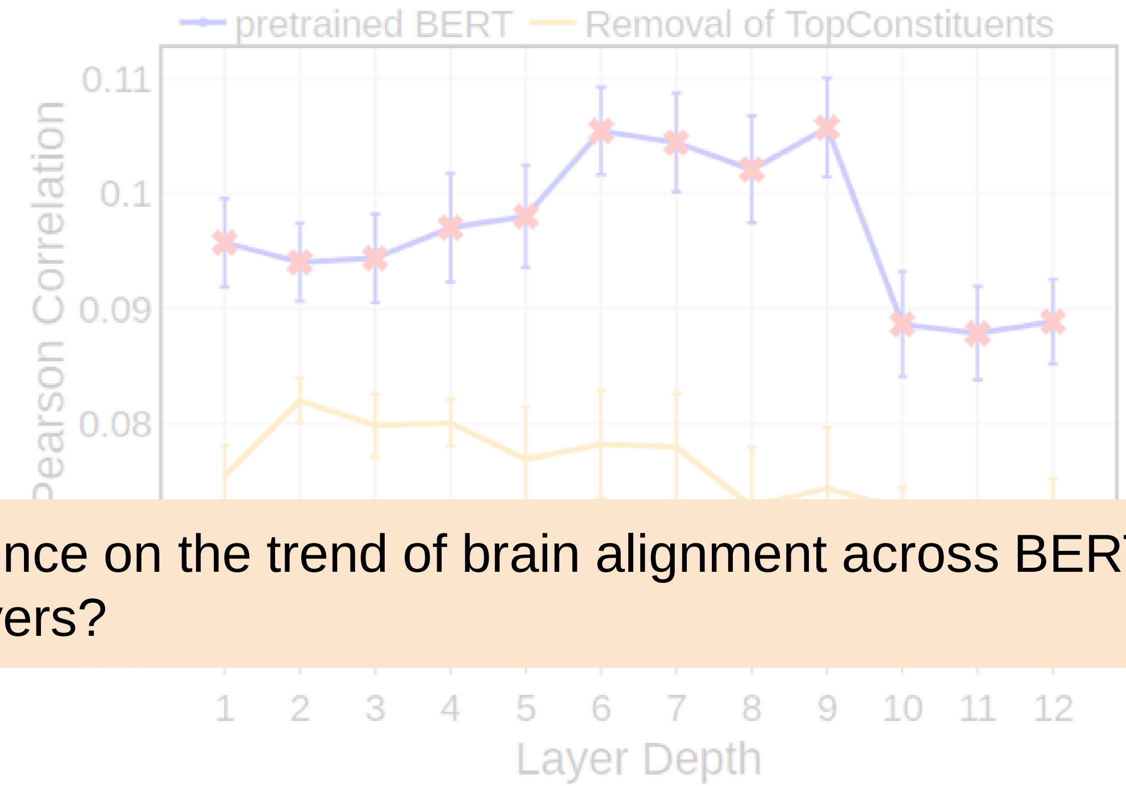


Removal of each linguistic property leads to a significant decrease in brain alignment on average across layers.

Greatest impact on brain alignment in the middle layers

Result-1





Which linguistic properties have the most influence on the trend of brain alignment across BERT layers?

Removal of each linguistic property leads to a significant decrease in brain alignment on average across layers.

Greatest impact on brain alignment in the middle layers

Result-2

Corr$_{task}$ (Δ probing accuracy$_{task}$, Δ brain alignment$_{task}$)

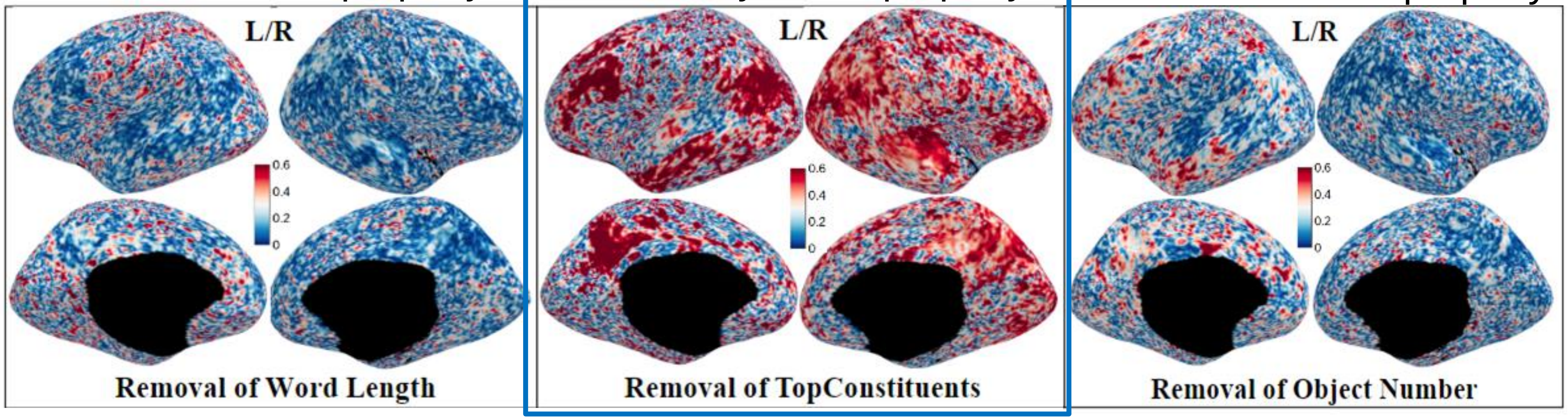
	Tasks	AG	ATL	PTL	IFG	IFGOrb	MFG	PCC	dmPFC	Whole Brain
	Word Length	0.261	0.264	0.220	0.355	0.129	0.319	0.143	0.100	0.216
Syntactic	TreeDepth	0.365	**0.421**	**0.458**	**0.442**	0.257	**0.436**	0.109	0.027	**0.443**
	TopConstituents	**0.489**	**0.421**	**0.464**	**0.516**	**0.453**	**0.463**	**0.459**	**0.463**	**0.451**
	Tense	0.226	0.283	0.307	0.325	0.345	0.339	**0.435**	0.122	0.248
	Subject Number	0.124	0.201	0.231	0.239	0.285	0.228	0.348	0.237	0.254
Semantic	Object Number	0.306	**0.392**	0.342	0.313	**0.503**	0.335	0.328	0.001	0.263

ROI-Level Analysis

Syntactic properties have the largest effect on the trend of brain alignment across model layers

Qualitative Analysis: Effect of each linguistic property

effect of surface property | effect of syntactic property | effect of semantic property



TopConstituent property is more localized to the canonical language regions in the left hemisphere and is more distributed in the right hemisphere.

Conclusions for neuro-AI research field

What are the reasons behind better similarity between language models and brains?

1. **AI-engineering**:
    - guide linguistic feature selection,
    - facilitate improved transfer learning,
    - help in the development of cognitively plausible AI architectures

2. **Computational modeling in Neuroscience**
    - enables cognitive neuroscientists to have more control over using language models as model organisms of language processing

3. **Model interpretability**
    - the addition of linguistic features by our approach can further increase the model interpretability using brain signals (Toneva & Wehbe 2019)

Artificial **Biological**

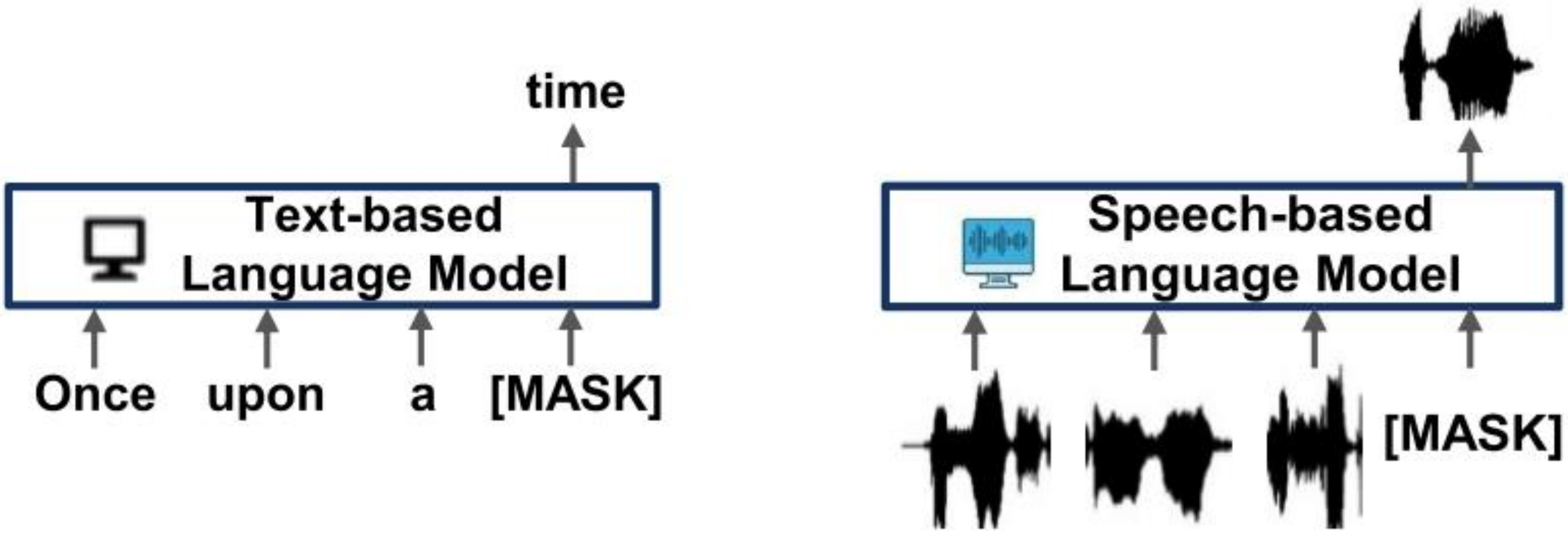
What types of information underlie the brain alignment of language models observed across brain regions?

Speech-based language models lack brain-relevant semantics

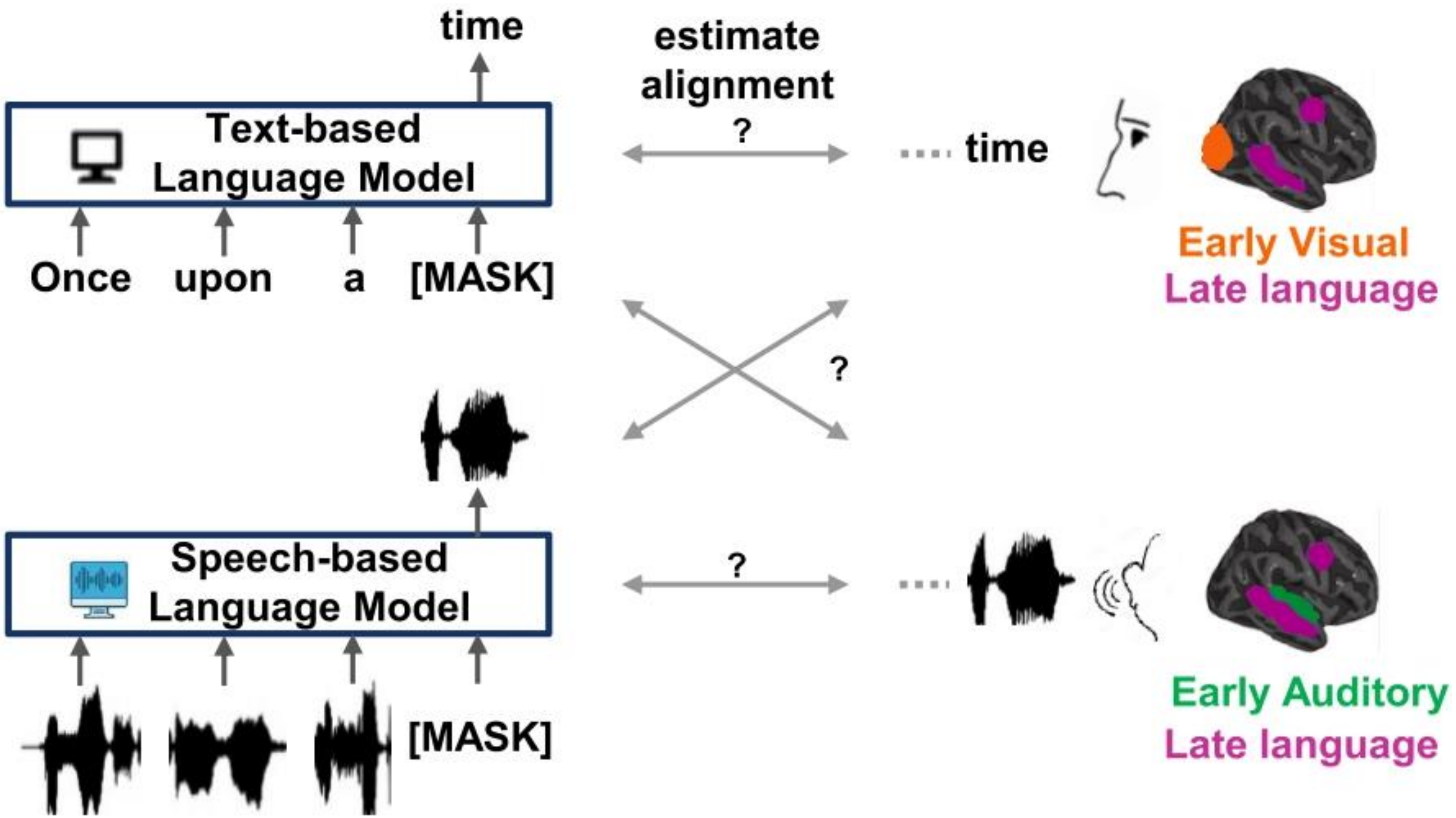
Subba Reddy Oota Emin Celik Fatma Deniz Mariya Toneva

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

Text- vs. Speech-based language models



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



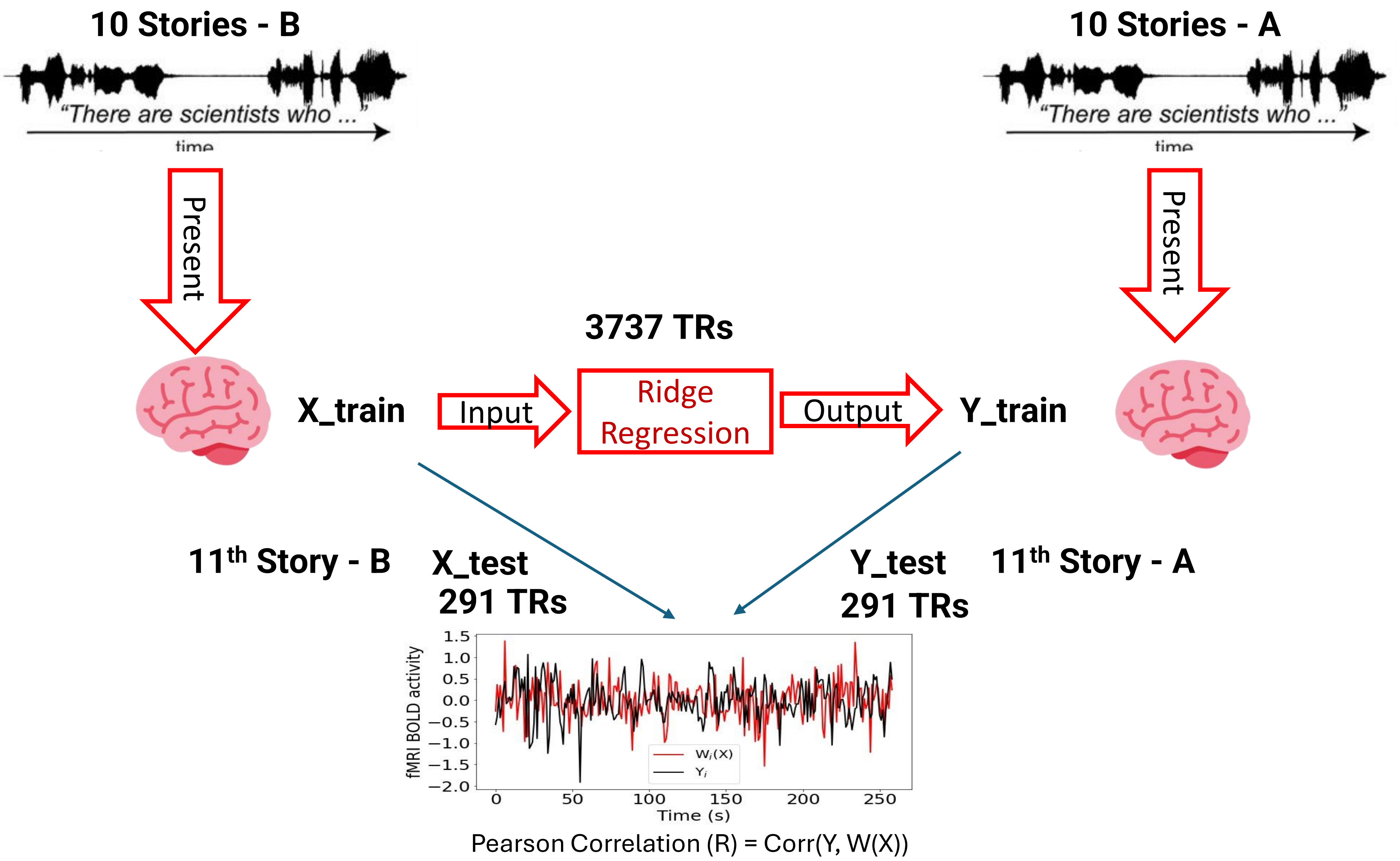
Datasets & Model

- Brain: fMRI recordings from Moth-Radio-Hour [Deniz et al. 2019]
    - Reading & Listening to the same short stories
    - N=6
- 3 text-based language models
    - BERT-base
    - GPT-2
    - FLAN-T5
- 2 speech-based language models
    - Wav2Vec2.0
    - Whisper

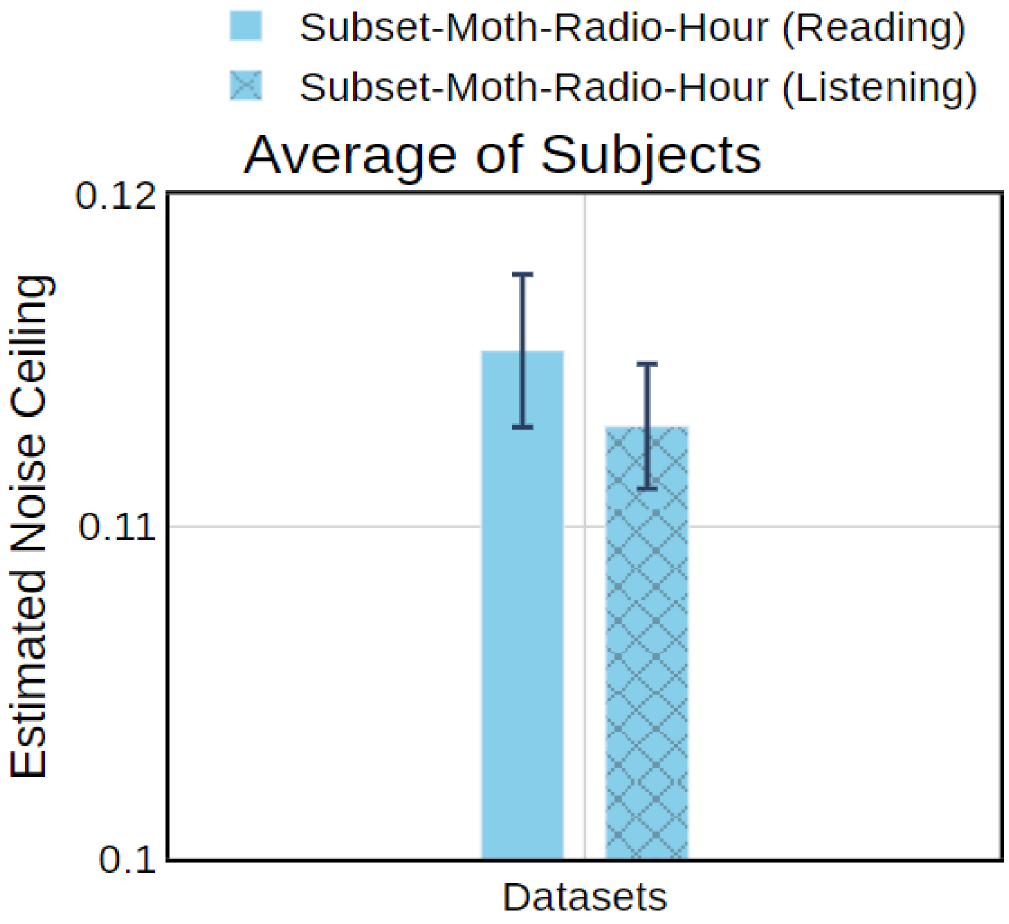
How can we quantify model predictions within a voxelwise encoding model?

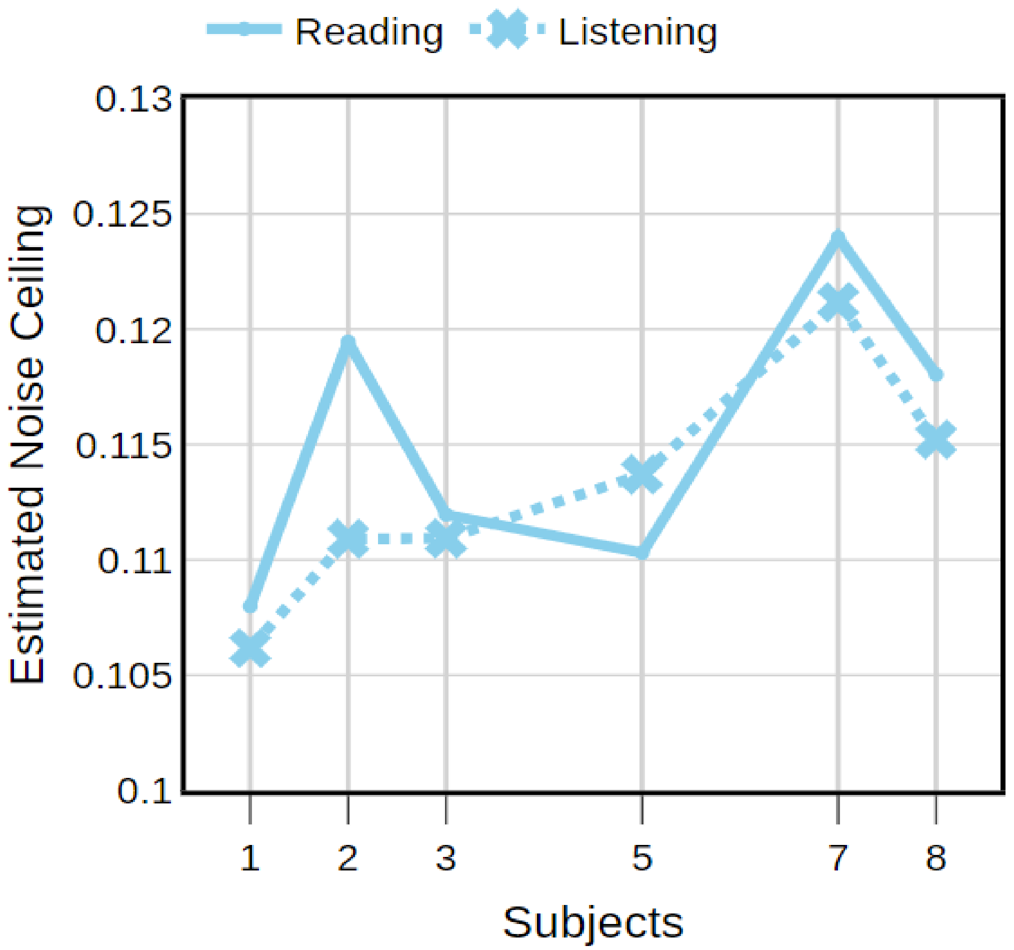
- To quantify model predictions within a voxelwise encoding model, we can compute variance:
- **Estimate noise ceiling:**
  - If we have data from multiple participants, we can predict the brain activity of one pariticipant using the data from remaining participants.
  - This can offer an upper bound for each voxel for a target participant, and it is related to a quantity called the noise ceiling estimate.
  - Normalized predictivity: percent of explained variance (model predictions/noise ceiling estimate)

Estimate Noise Ceiling: shared info between participants (A&B)



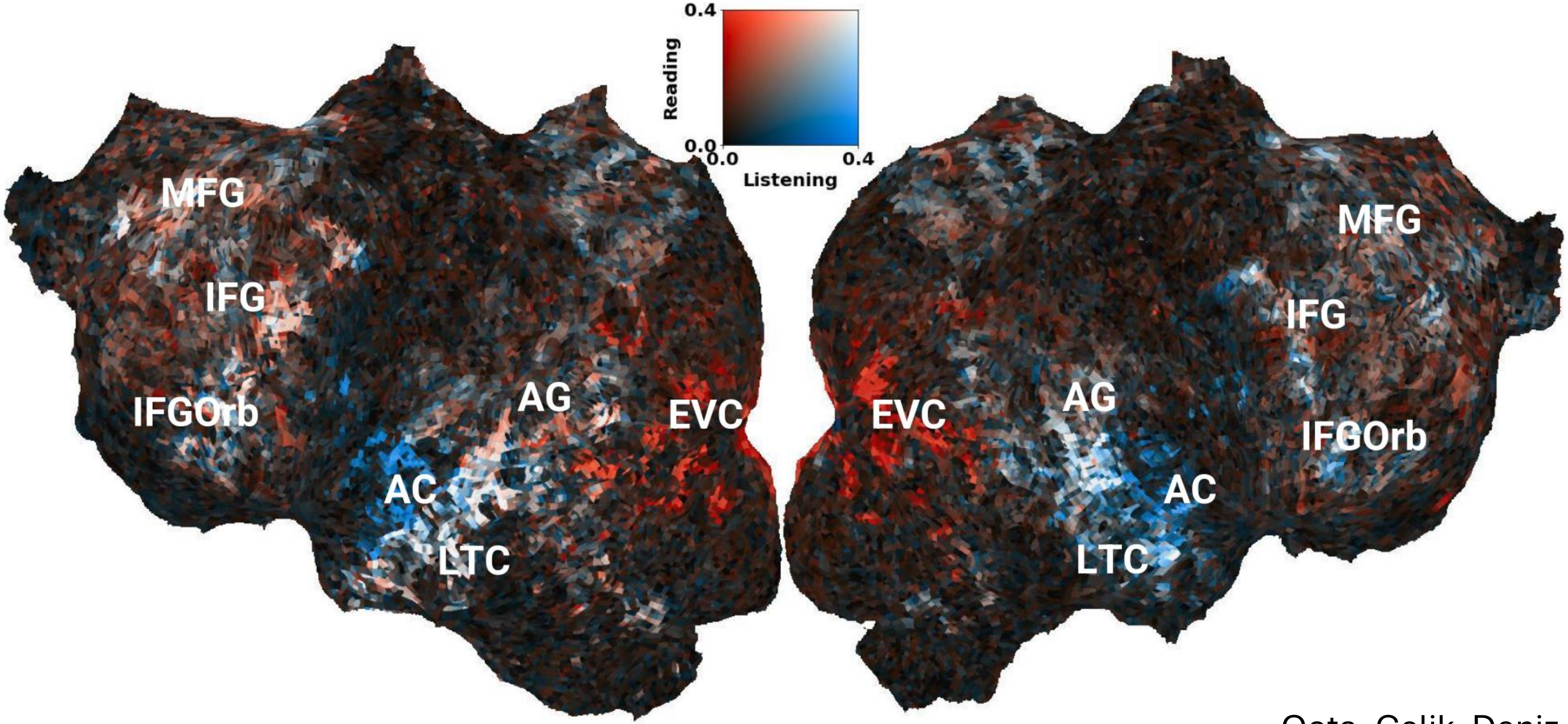
Estimated Noise Ceiling





S08: Estimated Noise Ceiling (Reading vs. Listening)

BLUE-AC and Orange-VC voxels are well predicted in estimated noise-ceiling.

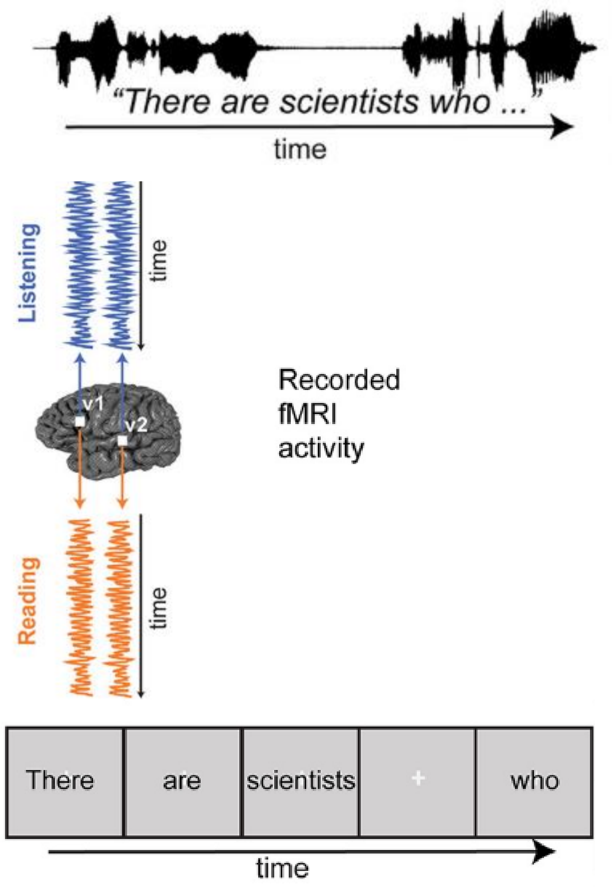


Oota, Celik, Deniz & Toneva,
[Under Review]

Research questions

1. Text-based language models are trained on written text
   - **Why they have impressive performance in early auditory cortex?**

Oota, Celik, Deniz & Toneva,
[Under Review]


"There are scientists who ..."
time
Listening
time
v1
v2
Recorded
fMRI
activity
Reading
time
There
are
scientists
+
who
time

Low-level Stimulus Features

Textual Features

Letters
Num of Letters
Num of Words
Word Length STD

Speech Features

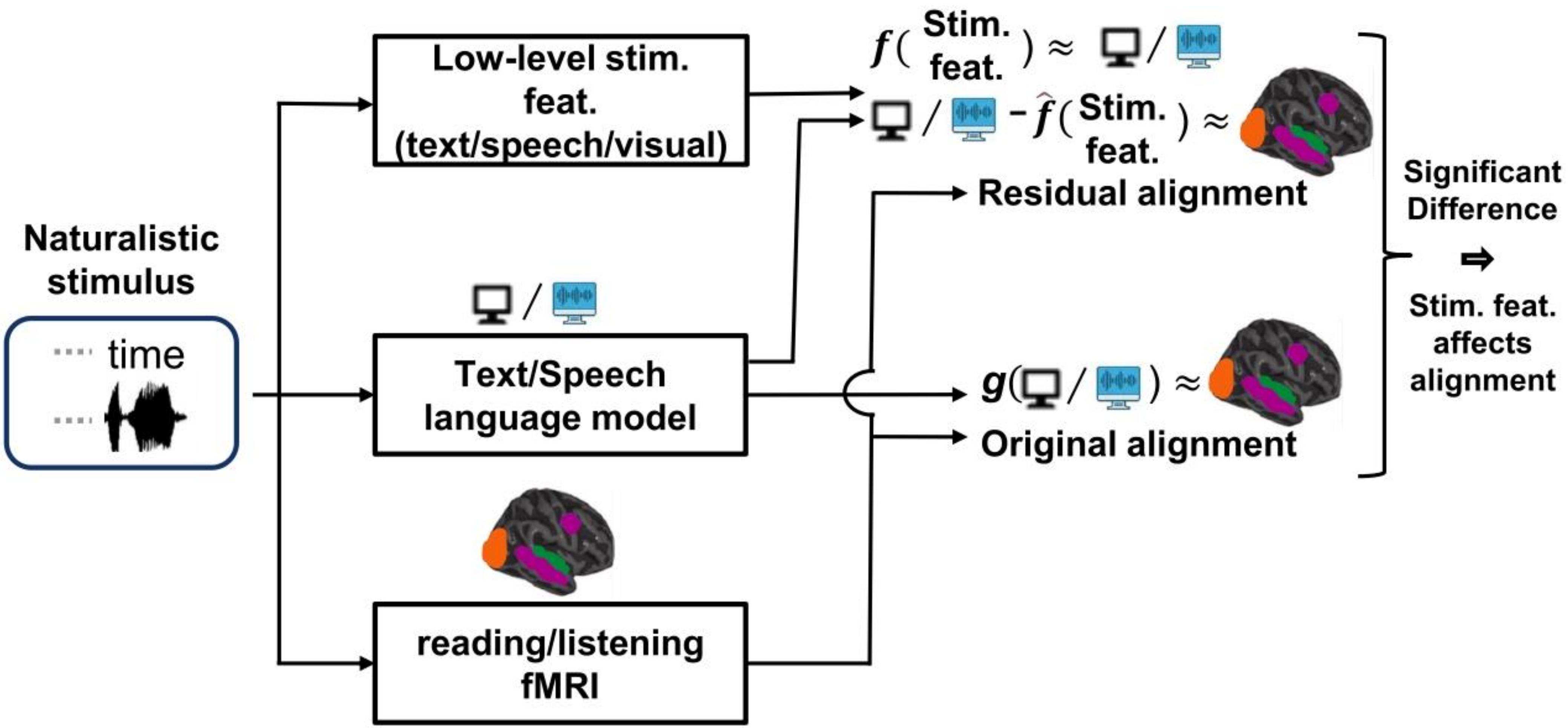
Phonemes
Num of Phonemes
Diphones
FBANK
MFCC
Mel-Spectrogram
PowSpec
Phonological
Articulation

Visual Features

Motion Energy

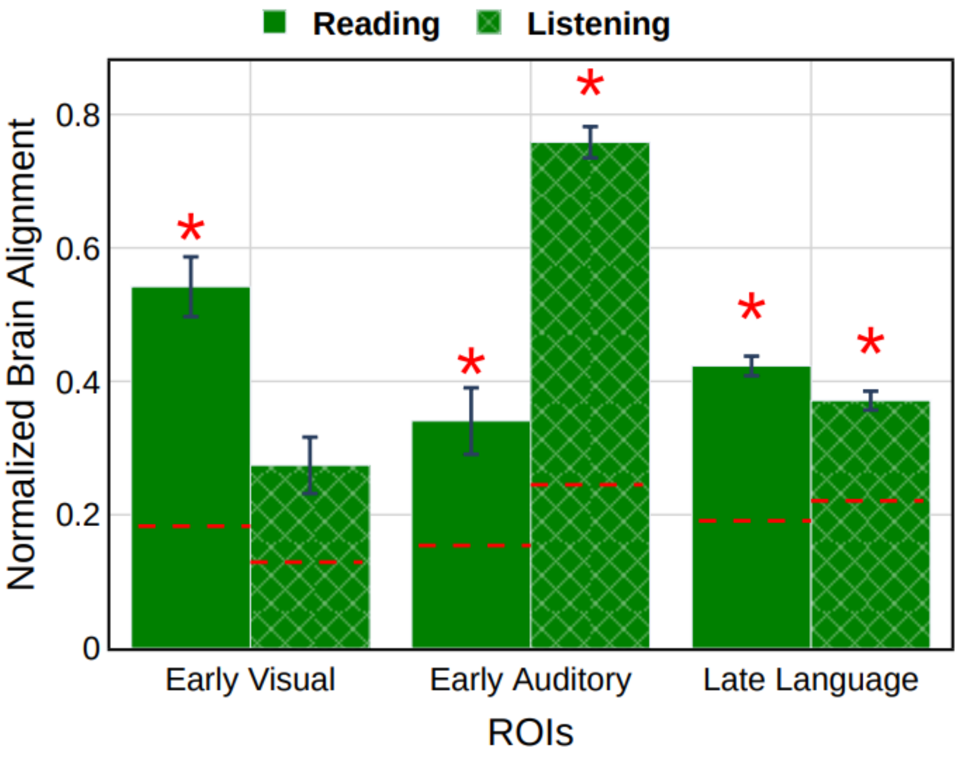
What are the reasons for this observed brain alignment?

Investigate via a perturbation approach



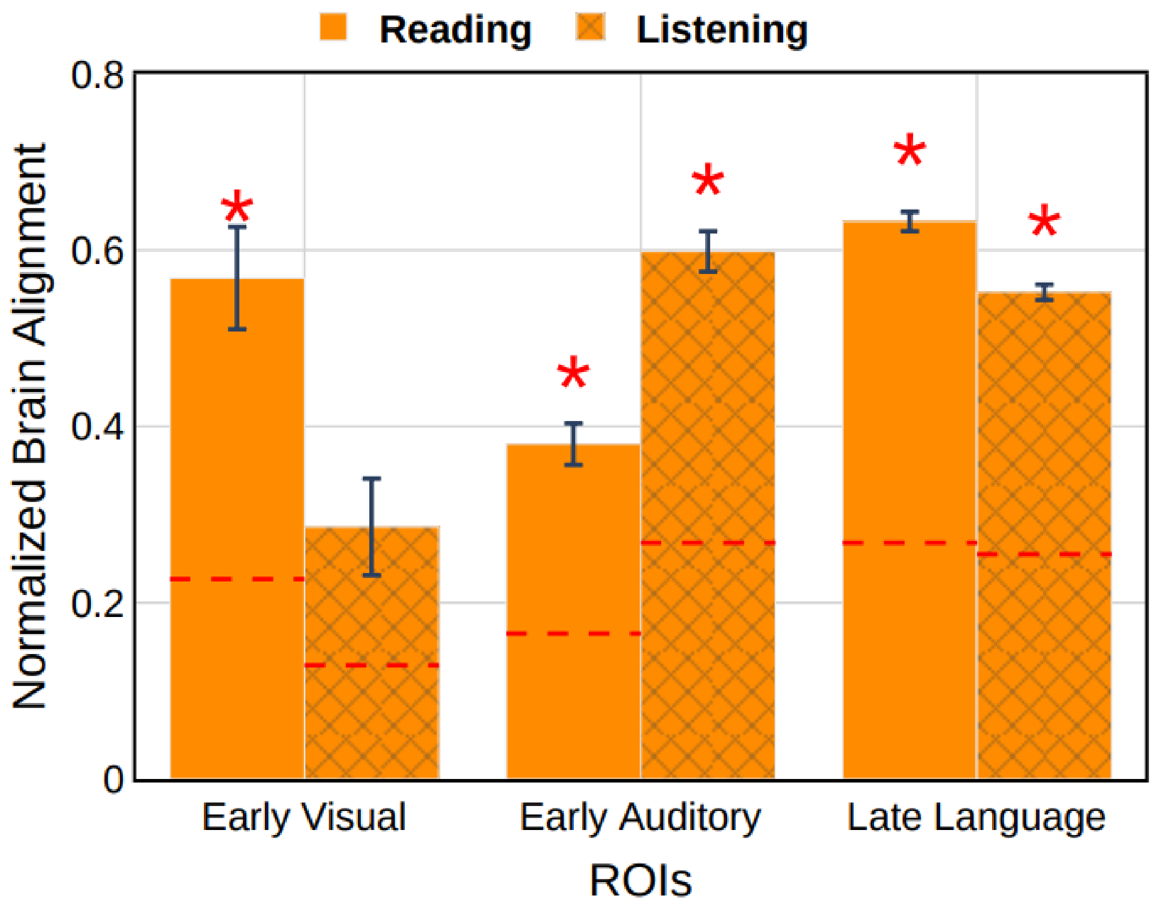
Text vs. Speech-based language models & brain alignment

(a) Text Models

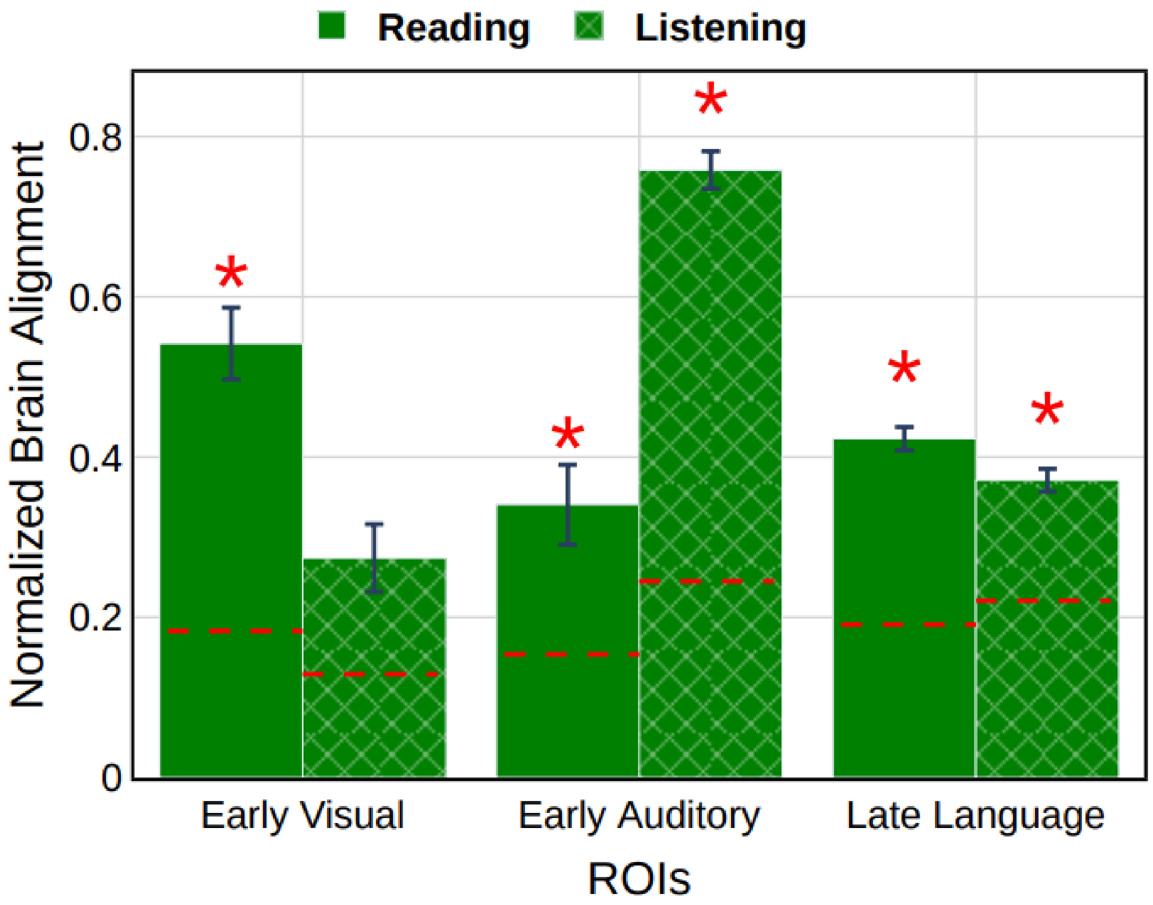


(b) Speech Models

- Late language regions: speech-based language models have a very high brain alignment both during reading and listening, but trails behind text-based models.
- Speech model representations highly predict early visual and auditory areas. What types of information present in these model resulting high brain alignment?



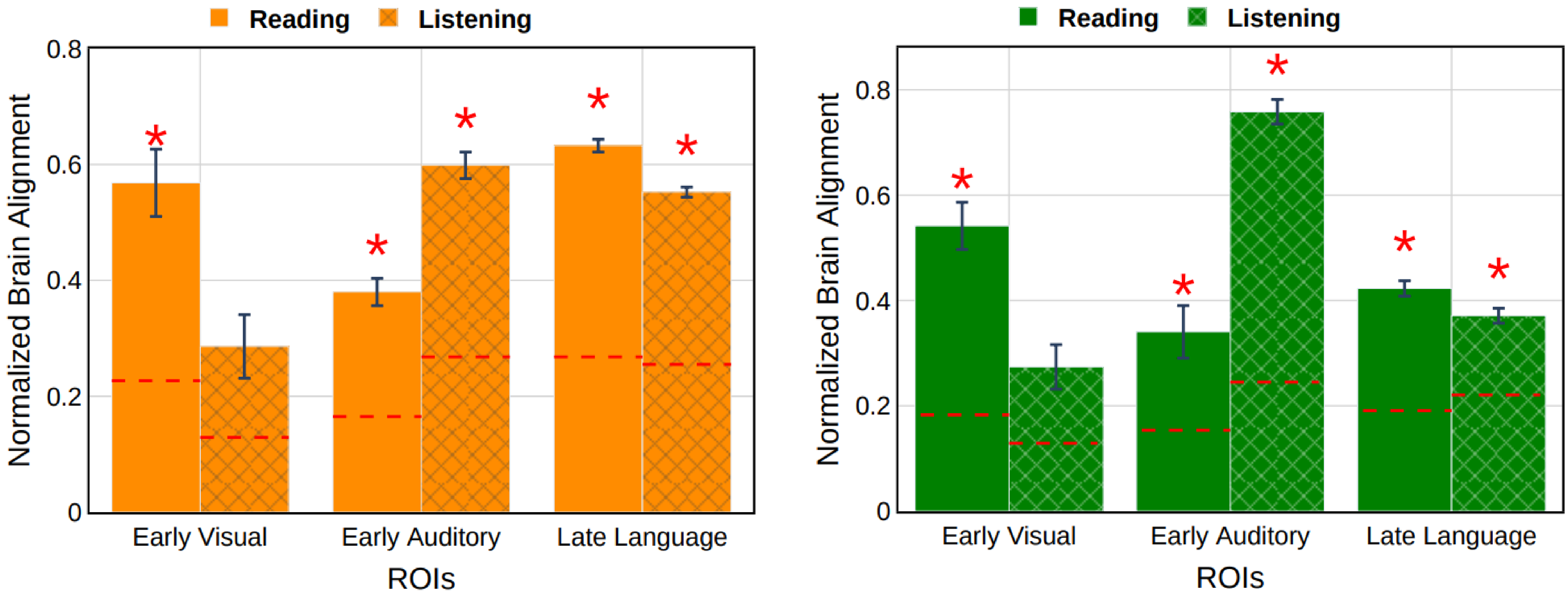
(a) Text Models



(b) Speech Models

- **Late language regions**:
  - high brain alignment both during reading and listening
- **Highly predict early visual and auditory areas.**

- **Late language regions**:
  - high brain alignment both during reading and listening,
  - but **trails behind text-based models**.
- **Highly predict early visual and auditory areas.**

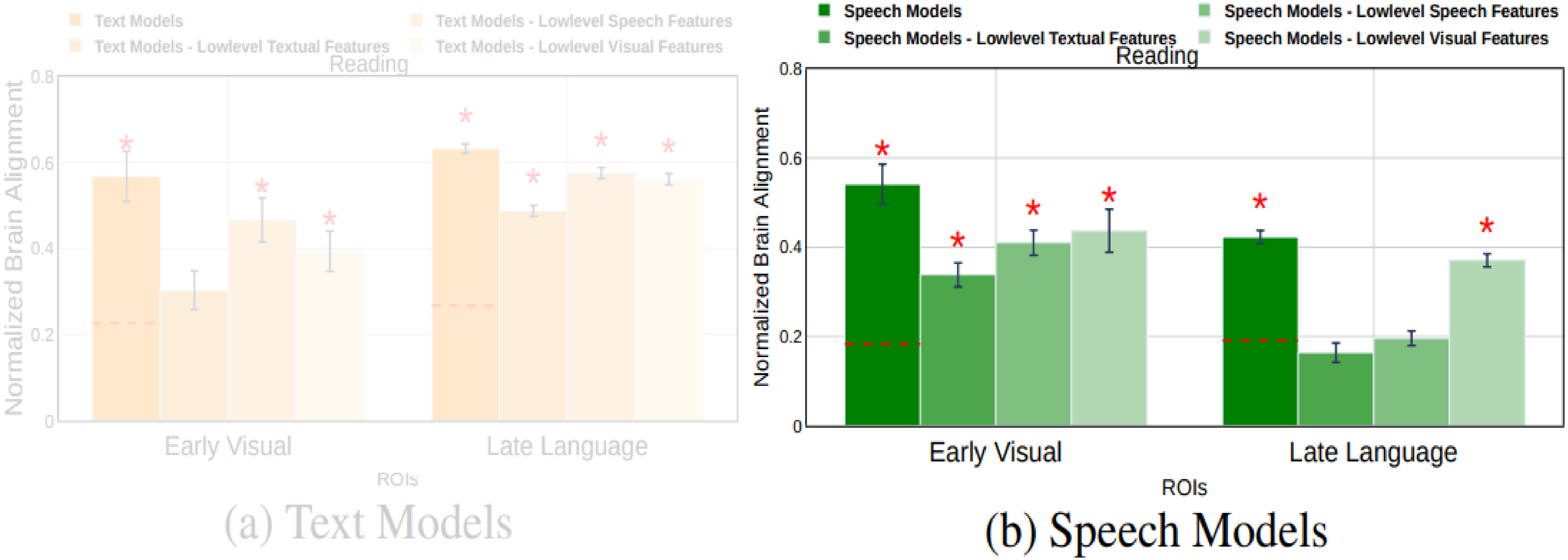


(a) Text Models

(b) Speech Models

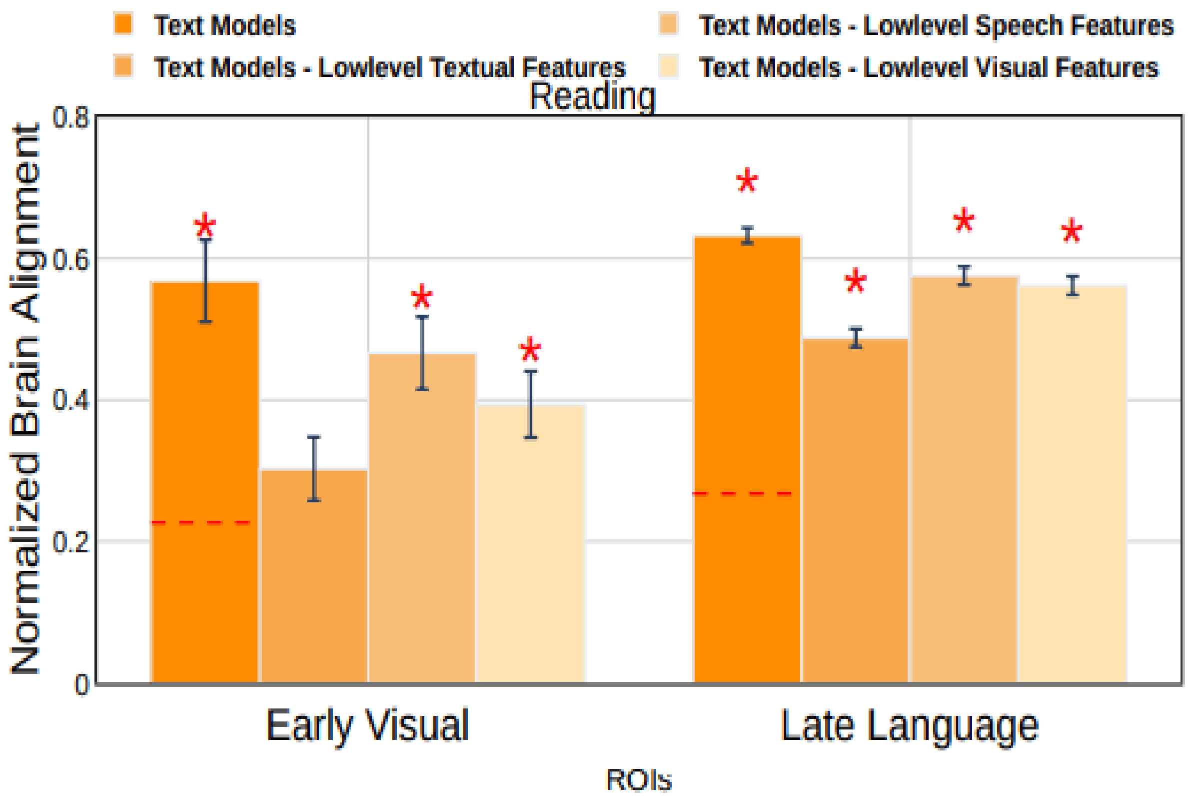
- Both models show high brain alignment with **late language regions**, but **speech models** trails behind **text models**
- Both models highly predict **early visual** and **auditory regions.**

Reading condition in early visual & late language regions

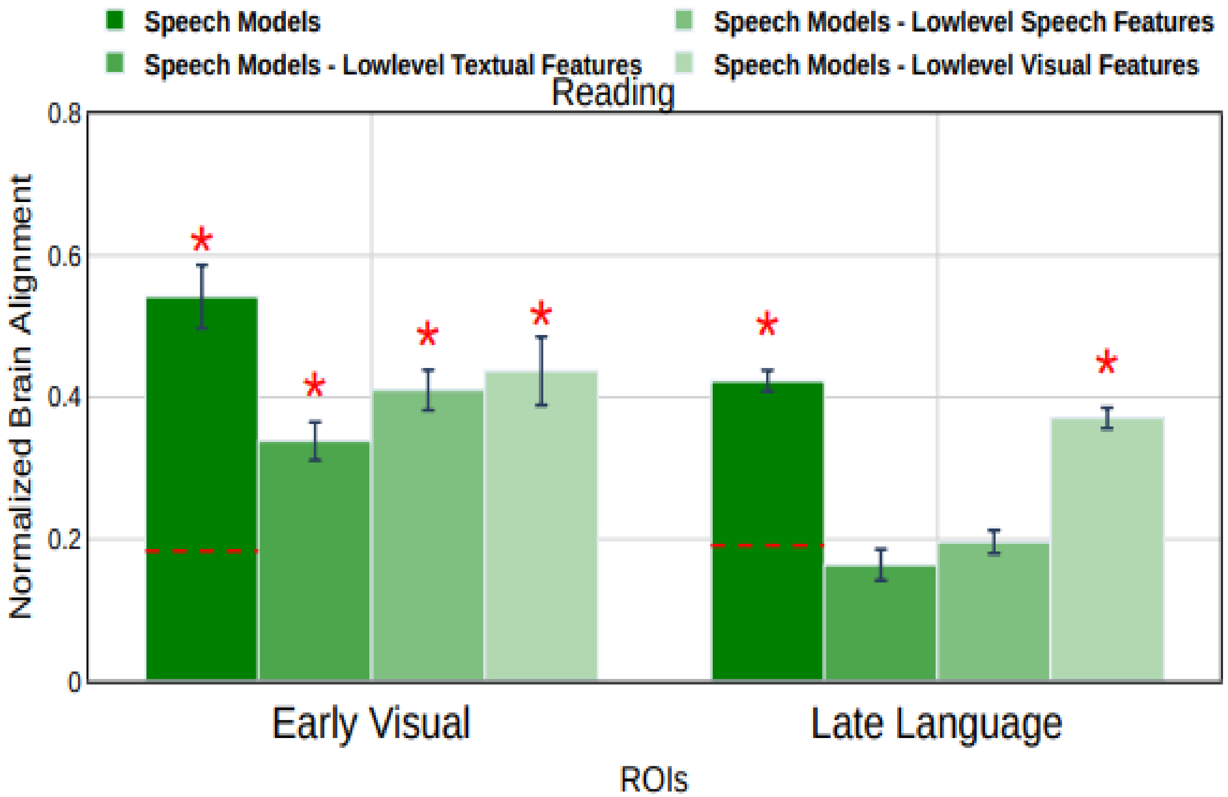


(a) Text Models (b) Speech Models

- Alignment of speech-based language models with late language regions is explained by low-level stimulus features, suggesting this brain alignment is not due to brain-relevant semantics.
- Alignment in the early visual regions is partially explained by low-level stimulus features



(a) Text Models



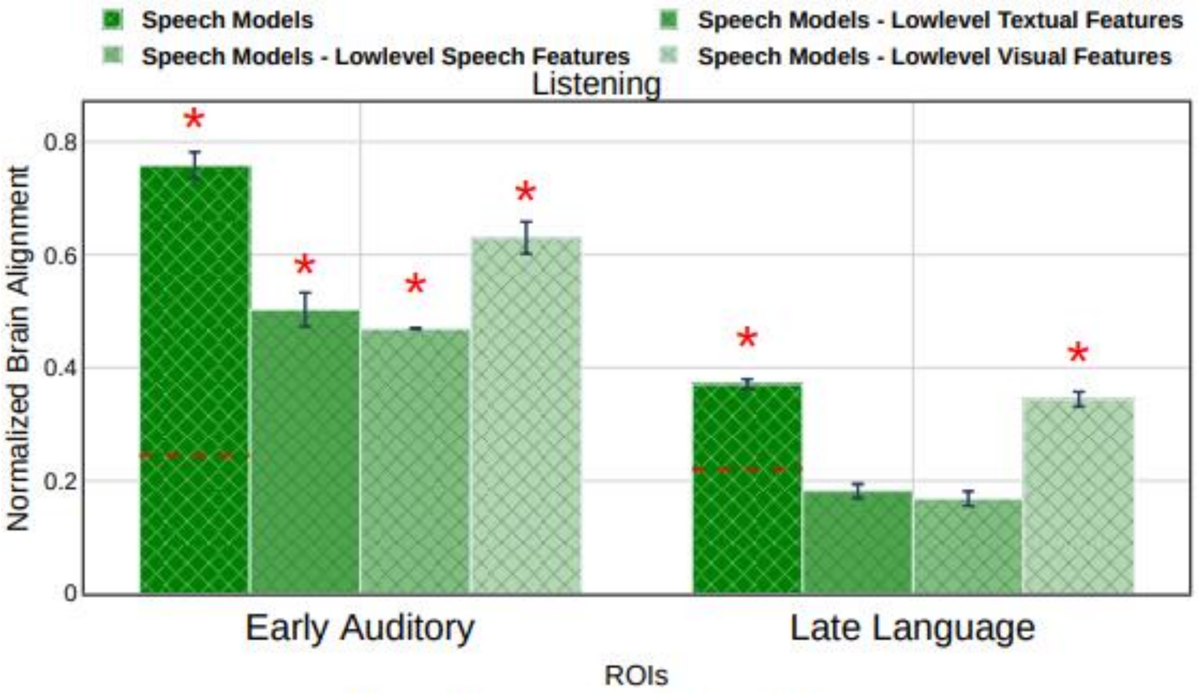
(b) Speech Models

- **Late language regions:** alignment is **not due to low-level stimulus features**, suggesting this brain alignment is due to **brain-relevant semantics**.
- **Early visual regions:** alignment is **largely** explained by **low-level textual features**

- **Late language regions:** alignment is **due to low-level stimulus features**, suggesting this brain alignment is **not due to brain-relevant semantics.**
- **Early visual regions:** alignment is **partially** explained by **low-level stimulus features**

Listening condition in early auditory & late language regions

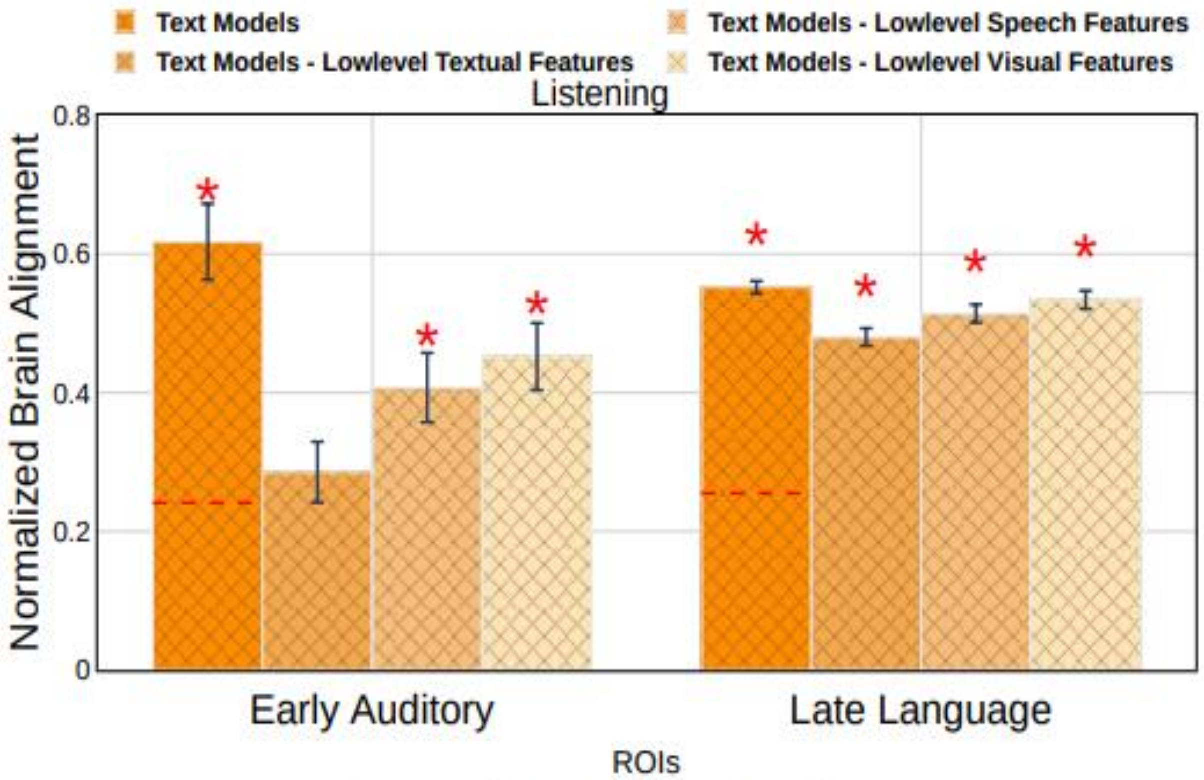
(a) Text Models



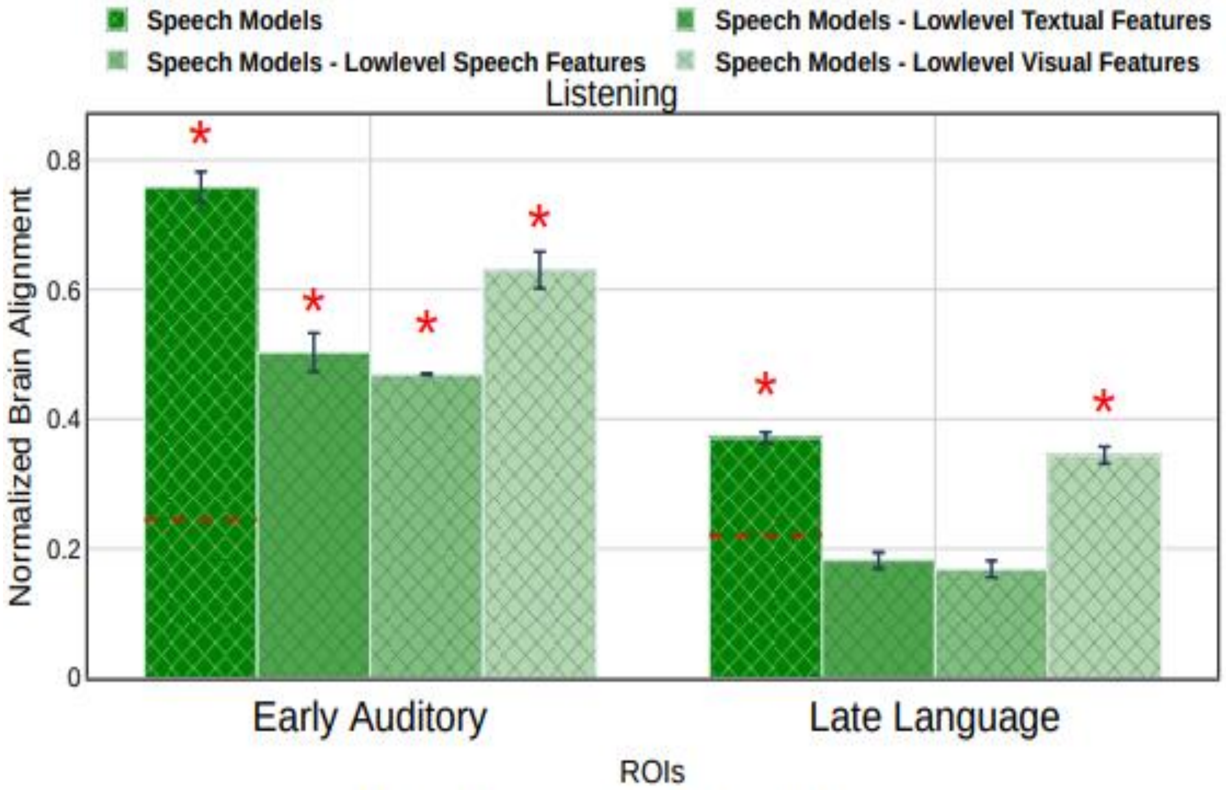
(b) Speech Models

- Alignment of text-based language models with late language regions is not due to low-level stimulus features,

- Alignment of speech-based language models with late language regions is explained by low-level stimulus features, suggesting this brain alignment is not due to brain-relevant semantics.
- Alignment in the early auditory regions is partially explained by low-level stimulus features



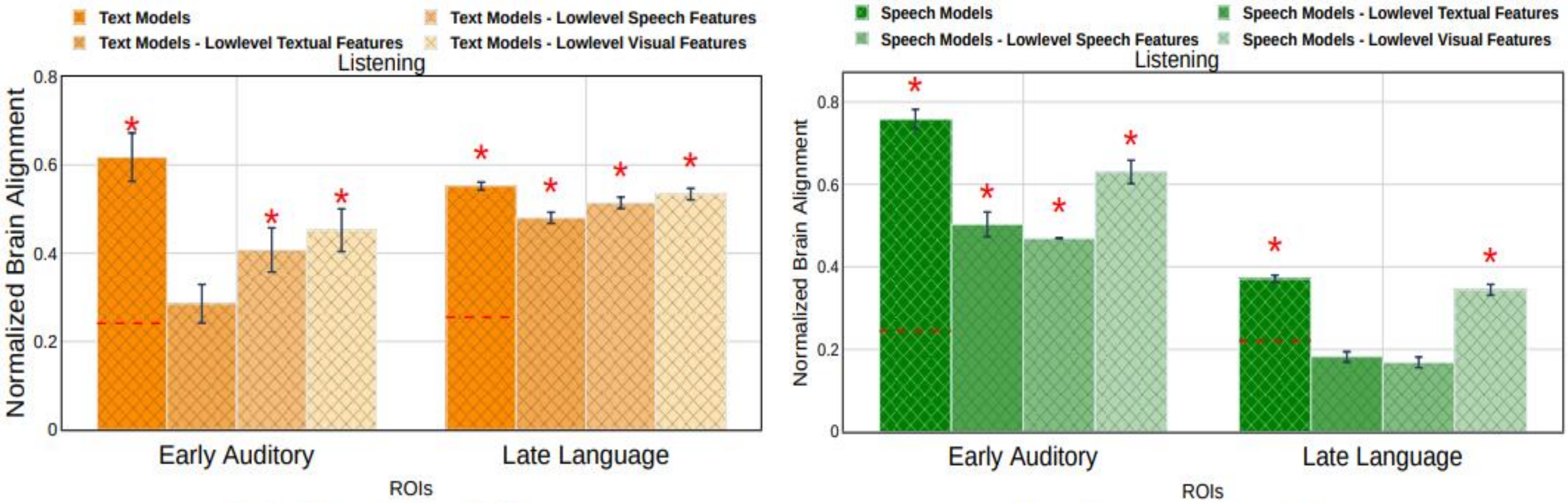
(a) Text Models



(b) Speech Models

- **Late language regions alignment:**
  - due to **brain-relevant semantics,** not simple **low-level stimulus features**
- **Early auditory regions alignment: largely** explained by **low-level textual features**

- **Late language regions alignment:**
  - due to **low-level stimulus features**, not **brain-relevant semantics.**
- **Early auditory regions alignment: partially** explained by **low-level stimulus features**



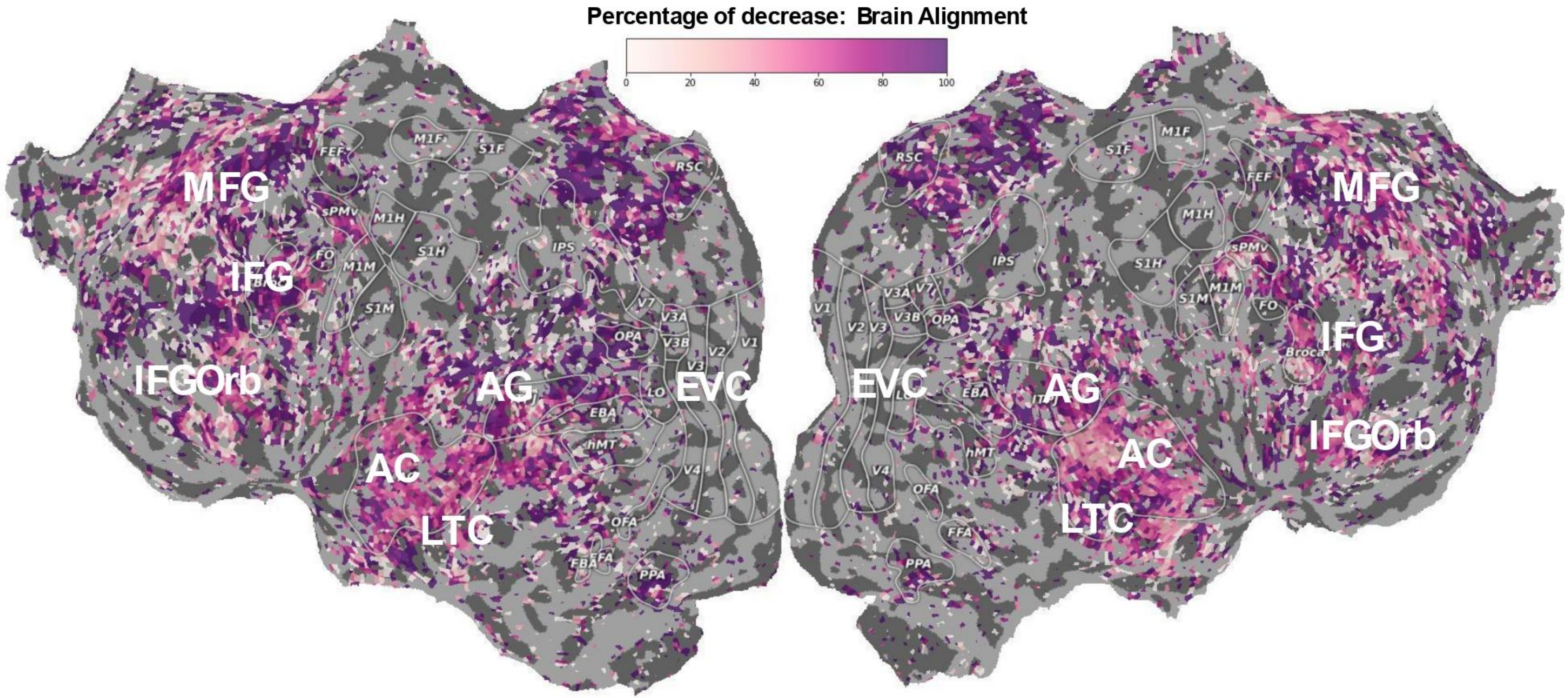
(a) Text Models

(b) Speech Models

- **Text models** alignment with **late language regions** due to **brain-relevant semantics**, while **speech models** alignment due to **low-level stimulus features**.

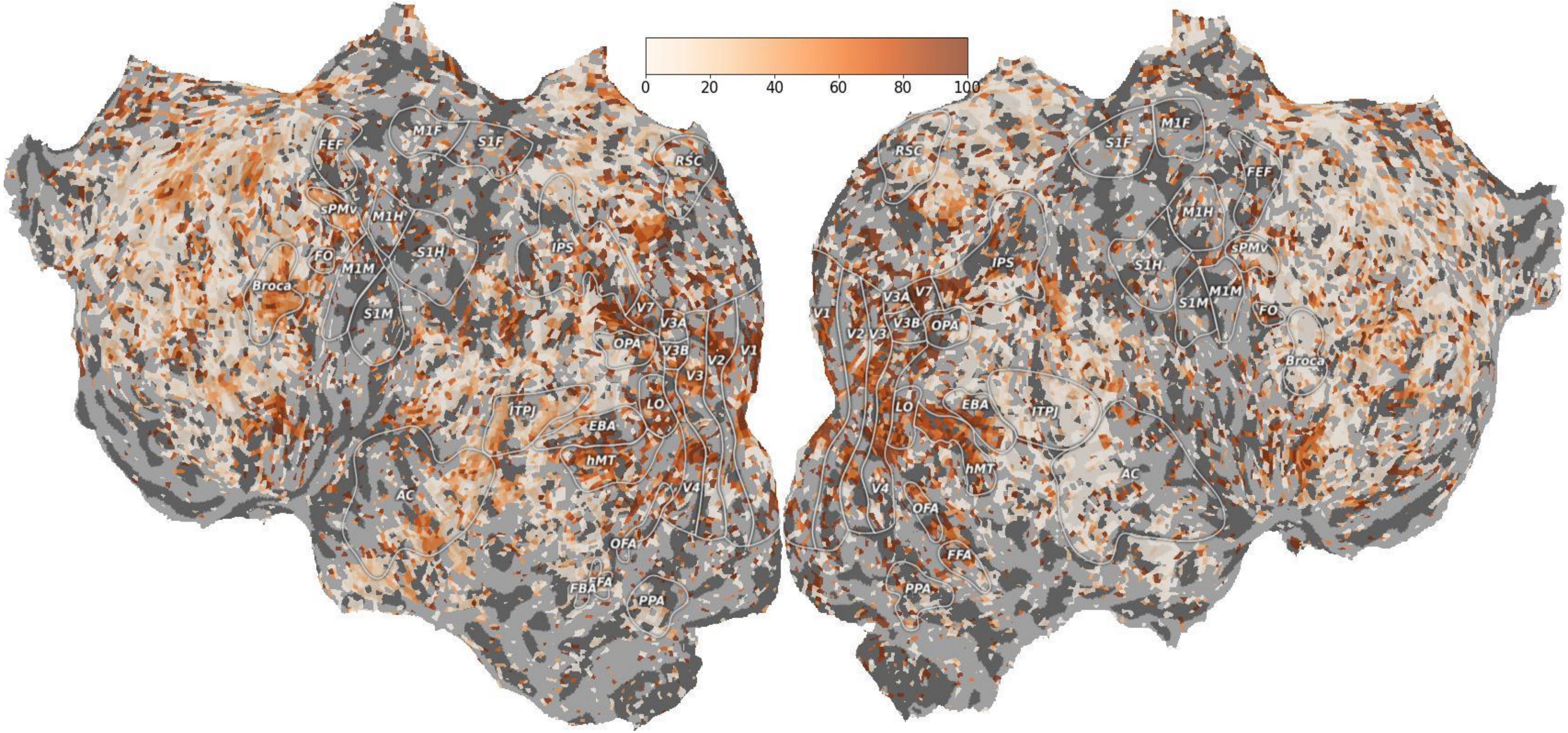
- **Text models** alignment with **early auditory regions** mostly due to **low-level textual features,** while **speech models** alignment is only **partially explained** by these features.

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



Oota, Celik, Deniz & Toneva, [Under Review]

Text-based language models have more information shared with late language regions beyond number of letters feature.



Oota, Celik, Deniz & Toneva,
[Under Review]

Conclusions for neuro-AI research field

The alignment of **text models** with **auditory** and **speech models** with **visual** regions is entirely due to low-level stimulus features

Unexpected alignment of models with sensory regions corresponding to the incongruent modality

Text models: alignment with visual and auditory, entirely due to low-level textual features
Speech models: greater alignment with auditory than visual, this difference cannot be entirely accounted by low-level stimulus features

Models align differently with their corresponding sensory regions, this variance is not explained by low-level stimulus features

The impact of low-level stimulus features:

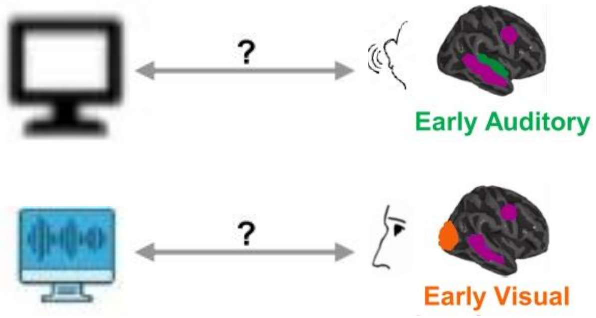
text model alignment is marginal,

speech models alignment is entirely driven by low-level features

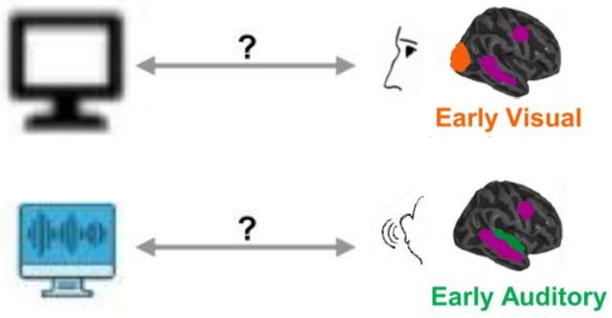
Text- and speech-based models show substantial alignment with late language regions

Conclusions for neuro-AI research field

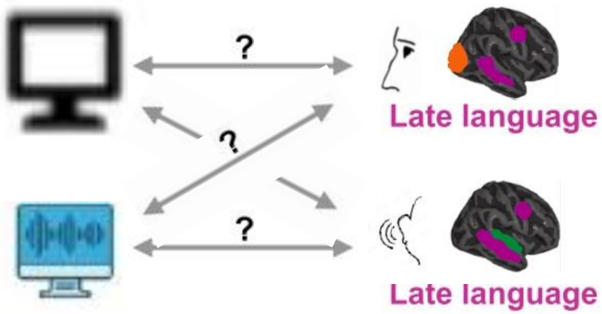
The **surprising alignment** of **models** with **incongruent modality** sensory regions is driven by **low-level features**



Models show **varying alignments** with their **corresponding sensory regions**



The impact of **low-level stimulus features**: **text model** alignment is **marginal**, **speech models** alignment is entirely driven by these features



Contributions and findings

- The unexpected alignment of models with sensory regions associated with the incongruent modality (i.e. text models with auditory regions and speech models with visual regions) is entirely due to low-level stimulus features

- Models exhibit varied alignment with the respective sensory regions, and this variance cannot be ascribed to low-level stimulus features alone

- While both text- and speech-based models show substantial alignment with late language regions, the impact of low-level stimulus features on text model alignment is marginal, whereas for speech-based models, alignment is entirely driven by these low-level features