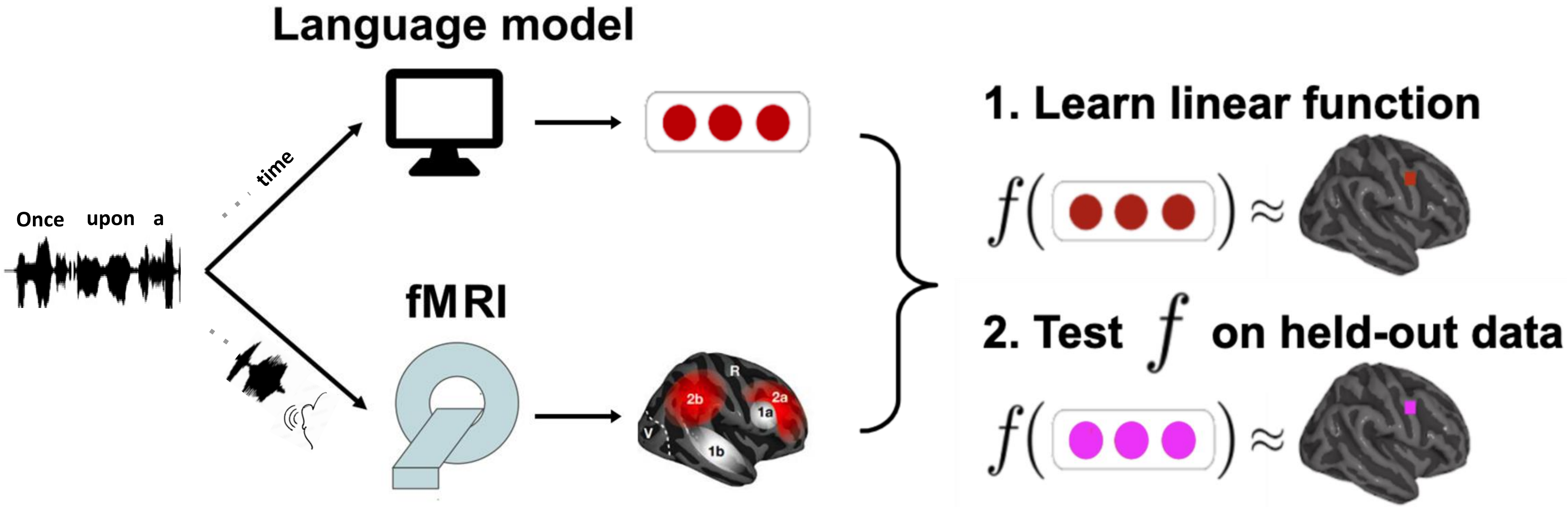
Joint processing of linguistic properties in brains and language models

Subba Reddy Oota   Manish Gupta   Mariya Toneva

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

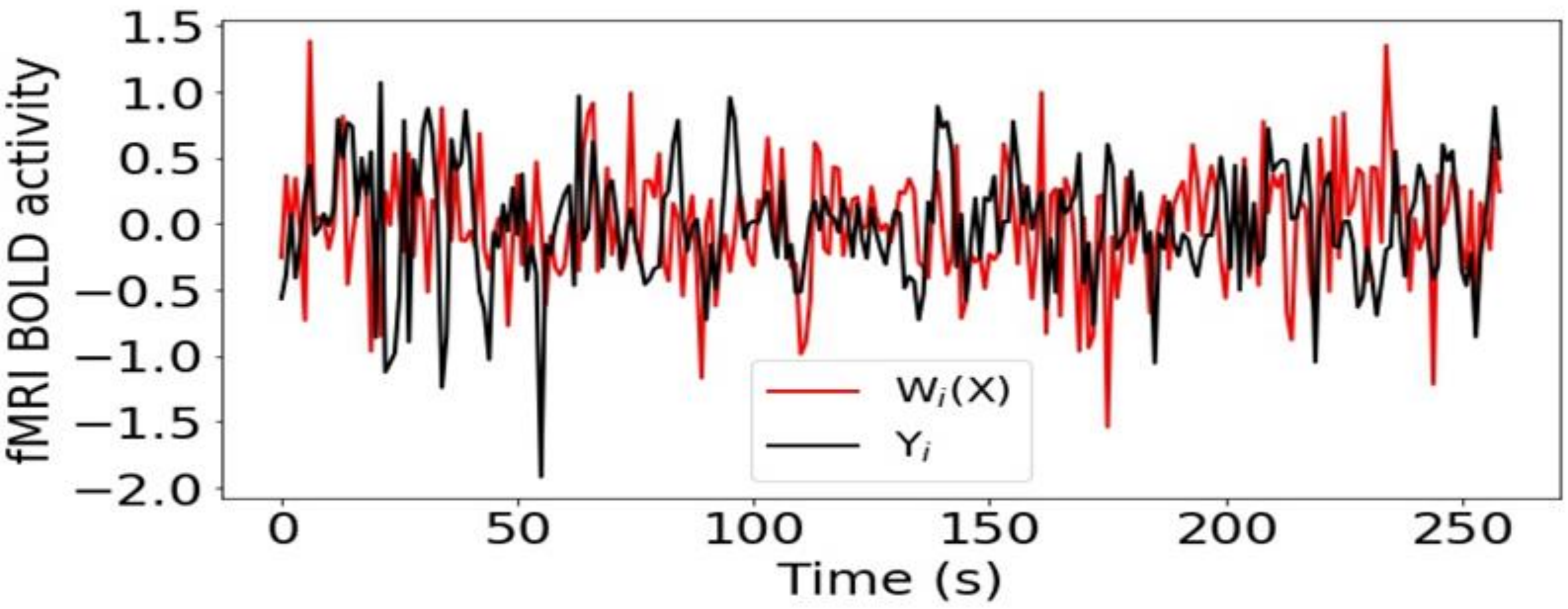
Language models (LMs) predict brain activity evoked by complex language (e.g. listening a story) to an impressive degree



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

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)

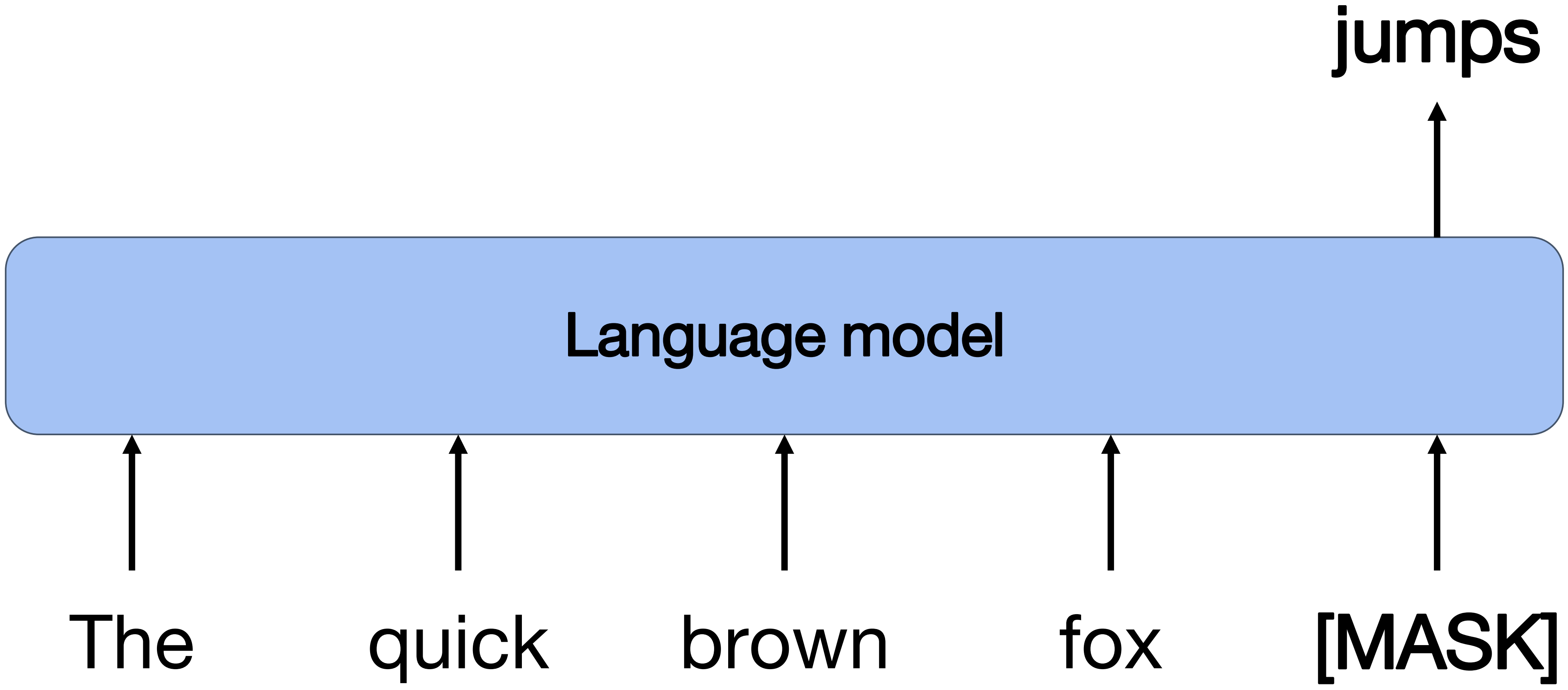
Language models (LMs) predict brain activity evoked by complex language (e.g. listening a story) to an impressive degree



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)

Language models (LMs) are trained to predict missing words



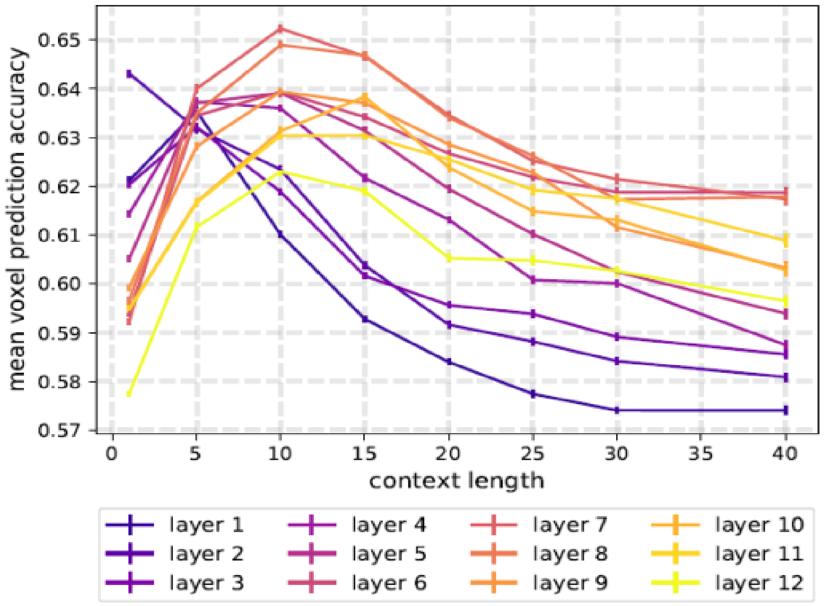
Language models (LMs) are trained to predict missing words

	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.

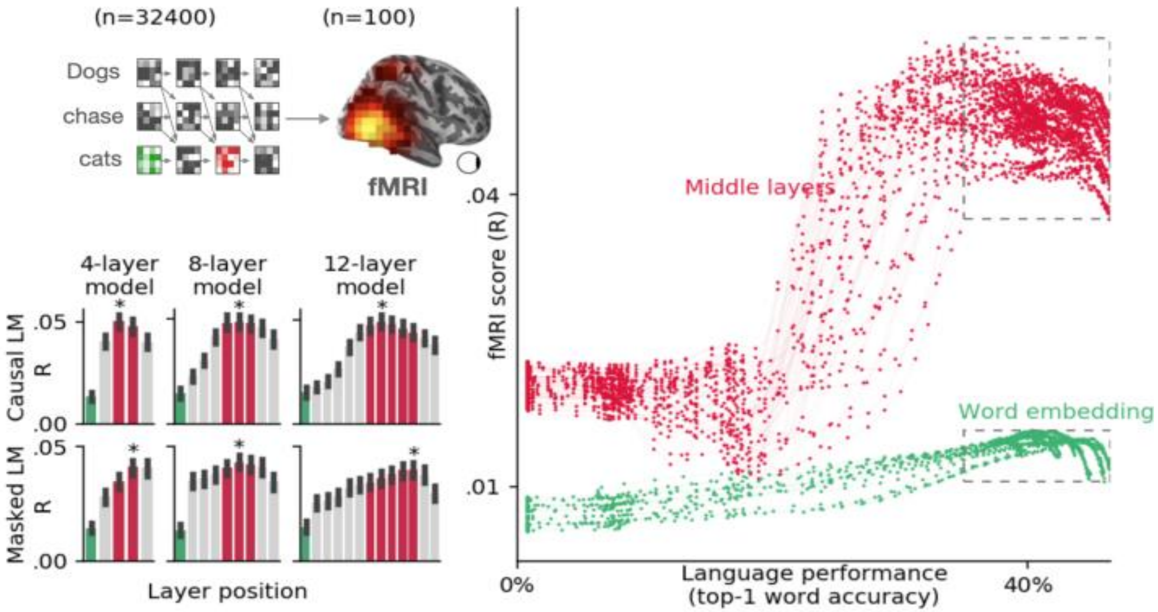
The strongest alignment with high-level language brain regions has consistently been observed in middle layers

Toneva et al. 2019



BERT

Caucheteux et al. 2022

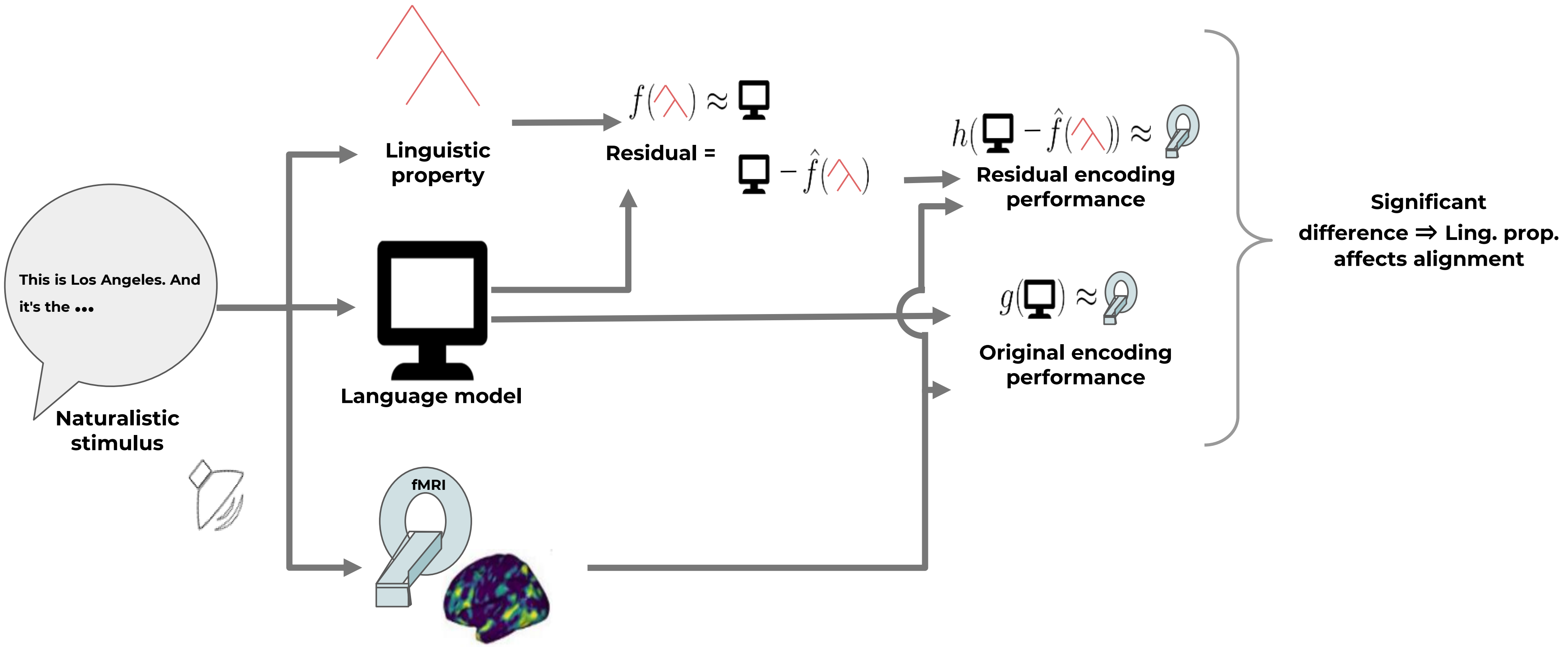


XLM

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

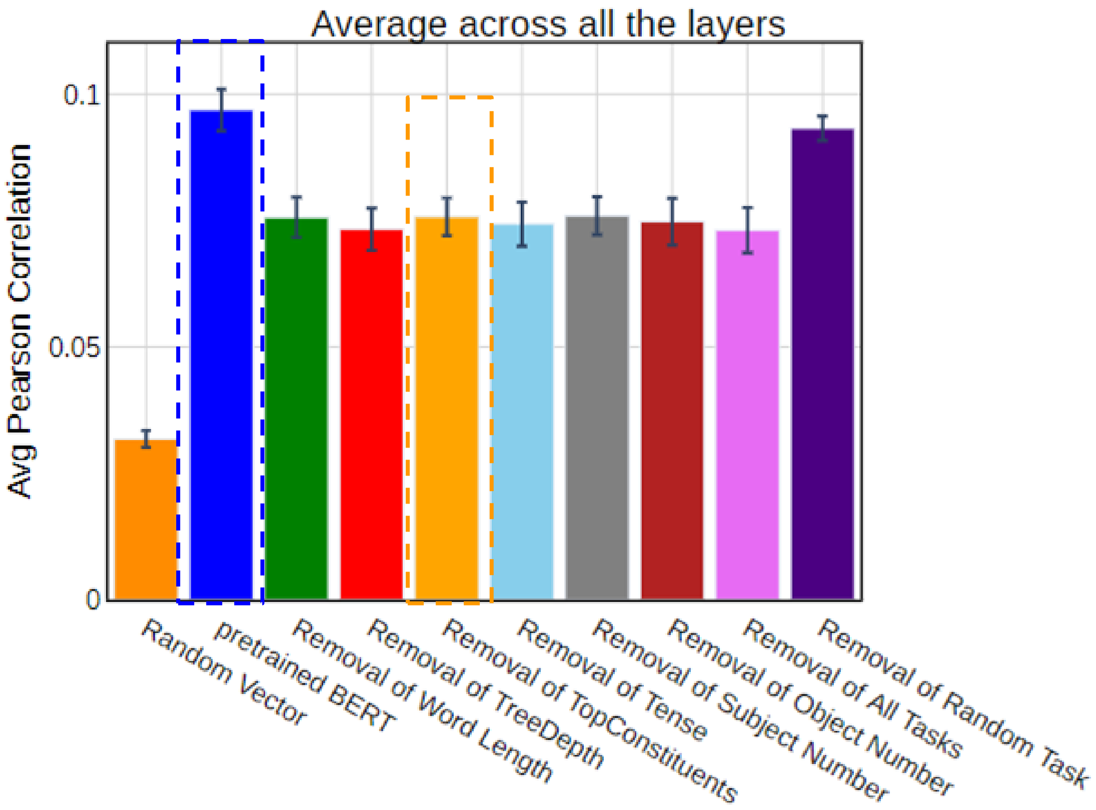
What are the reasons for this observed brain alignment?

Investigate via a perturbation approach



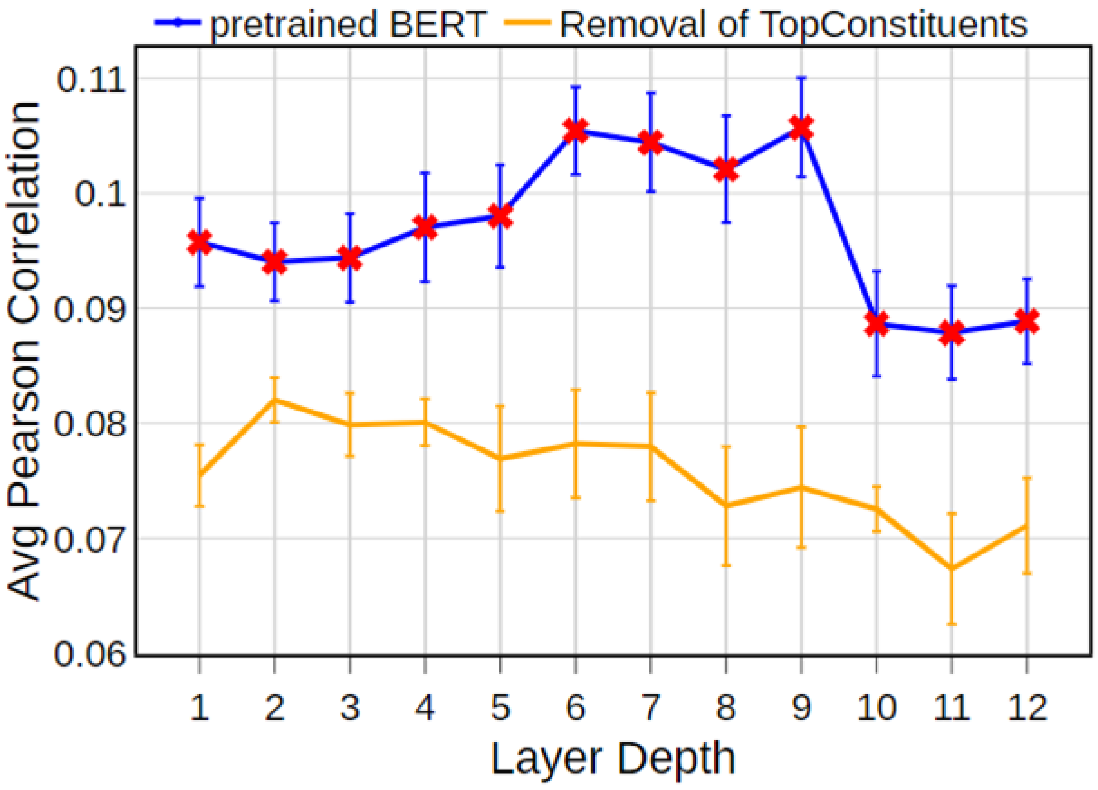
Does the removal of a linguistic property affects the alignment between language model and the brain across all layers?

Result-1



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

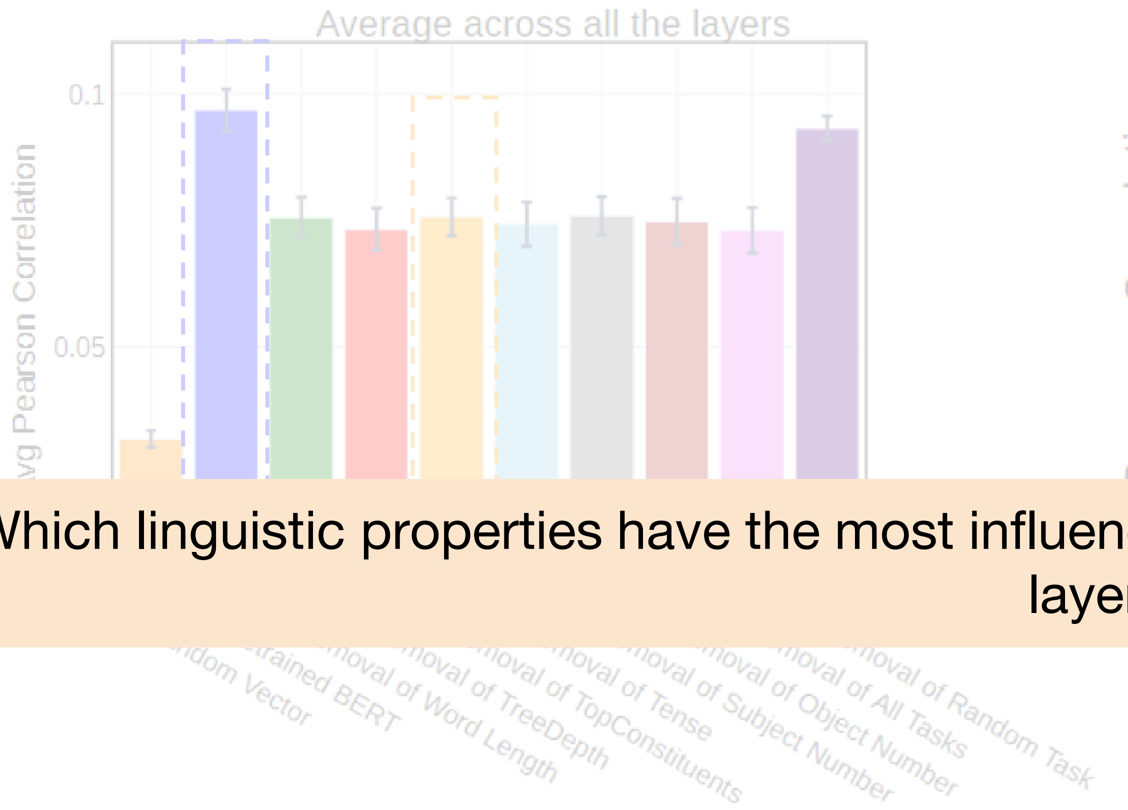
Result-1

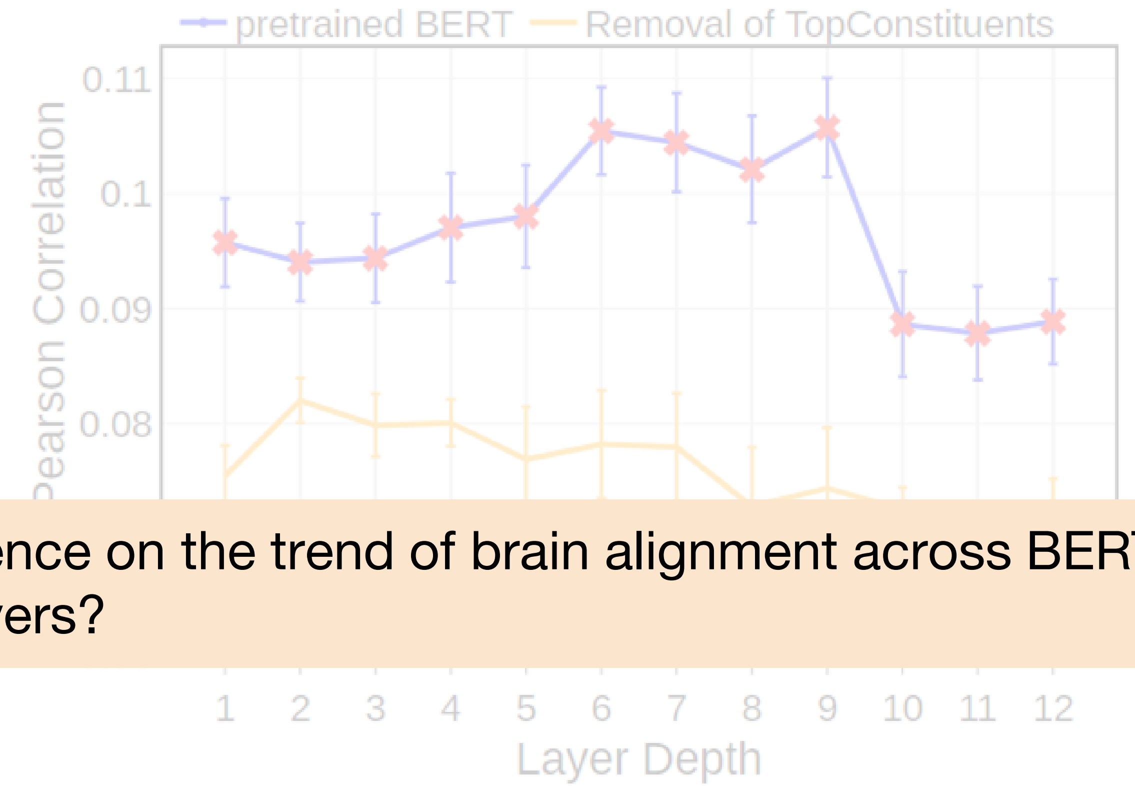


Removal of each linguistic property leads to a significant decrease in brain alignment on average across layers.

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

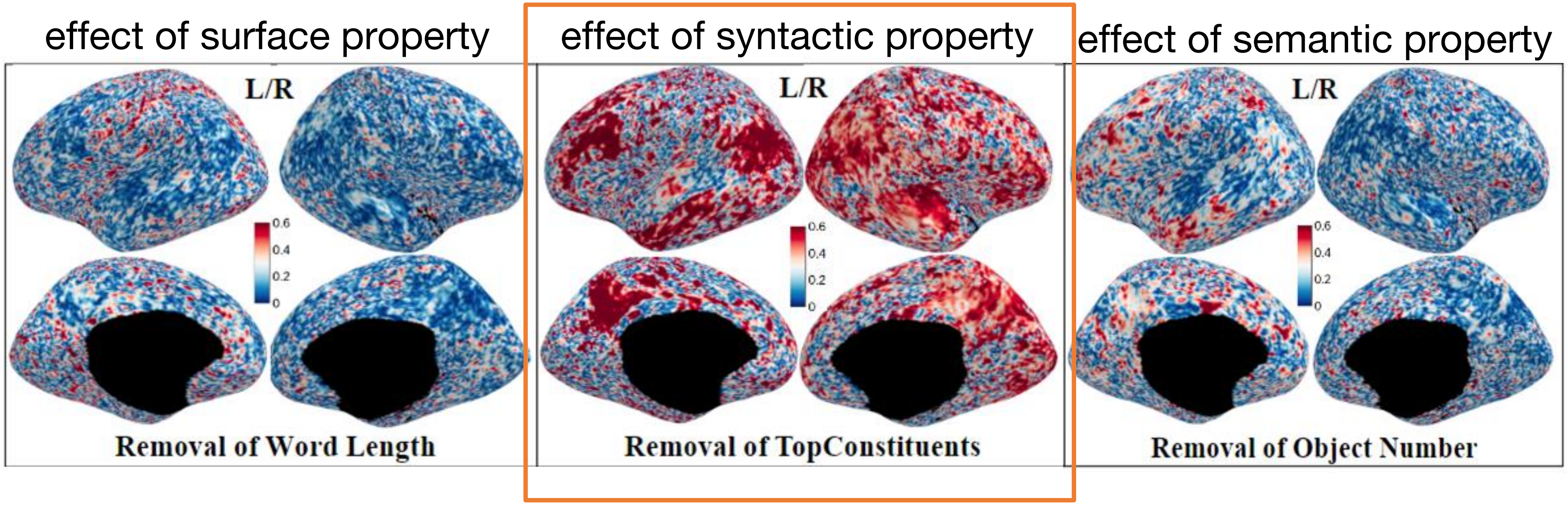
$$\text{Corr}_{\text{task}}\ (\Delta\ \text{probing accuracy}_{\text{task}}\ ,\ \Delta\ \text{brain alignment}_{\text{task}})$$

	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**
Syntactic	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



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

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)

Joint Processing of linguistic properties in brains and language models (NeurIPS 2023)

Subba Reddy Oota **Manish Gupta** **Mariya Toneva**

Bridging AI and Neuroscience (BrAIN) group

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS