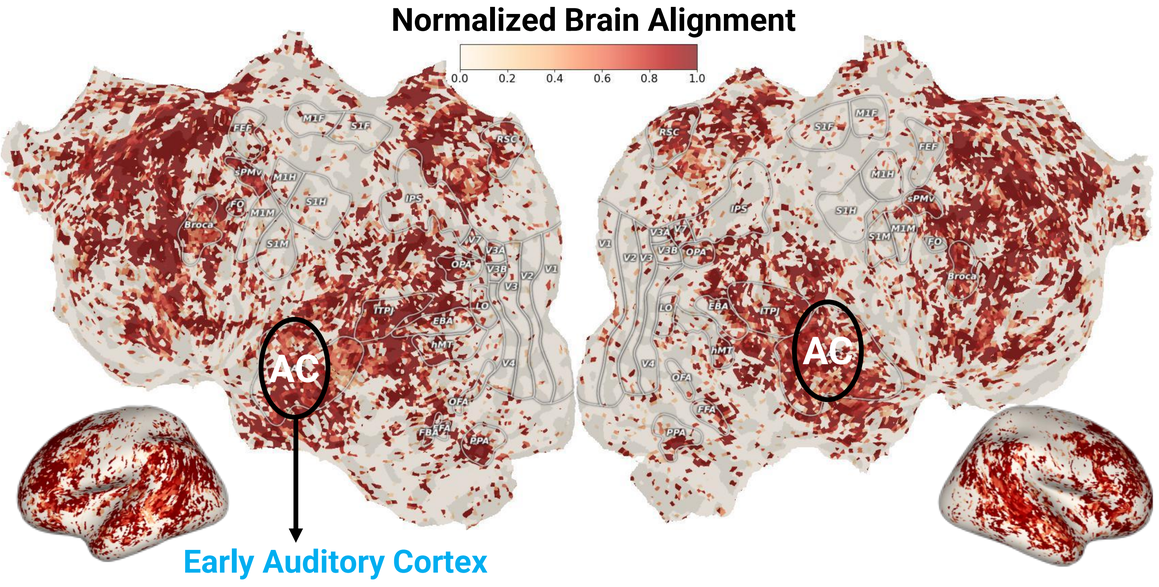
Speech language models lack important brain-relevant semantics

Subba Reddy Oota[1,2,3] Emin Çelik[3] Fatma Deniz[2] Mariya Toneva[3]

[1]Inria, France [2]TU Berlin, Germany [3]Max Planck Institute for Software Systems, Germany

ACL 2024 Bangkok, Thailand

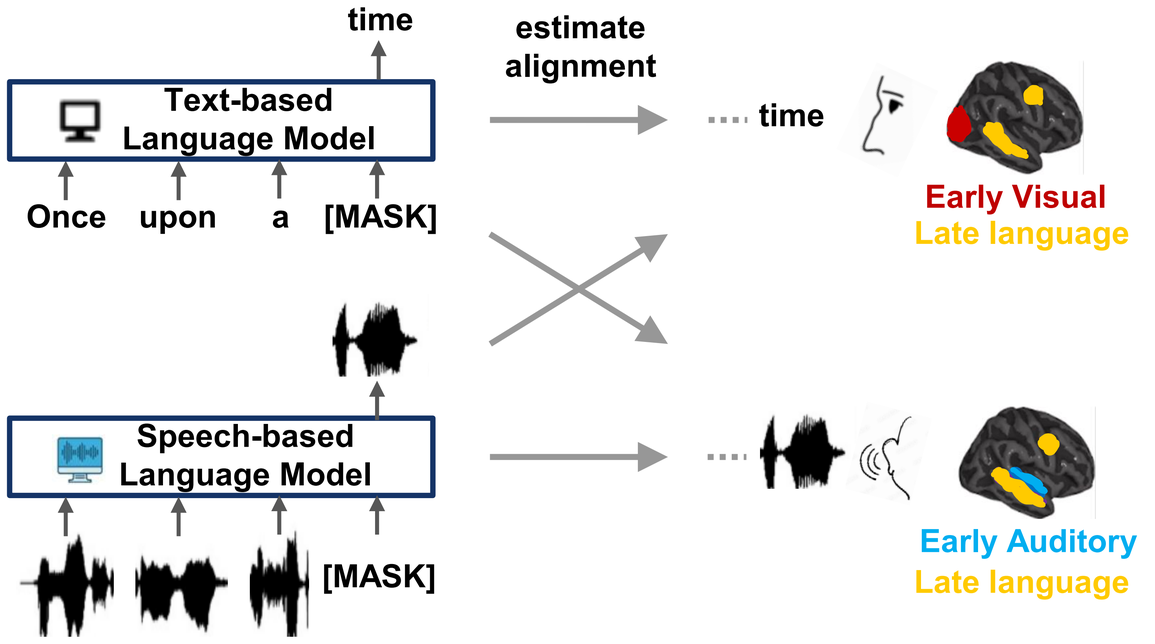
Text-based language models predict activity in many parts of the brain while listening, but also surprisingly in sound-specific regions



Research questions

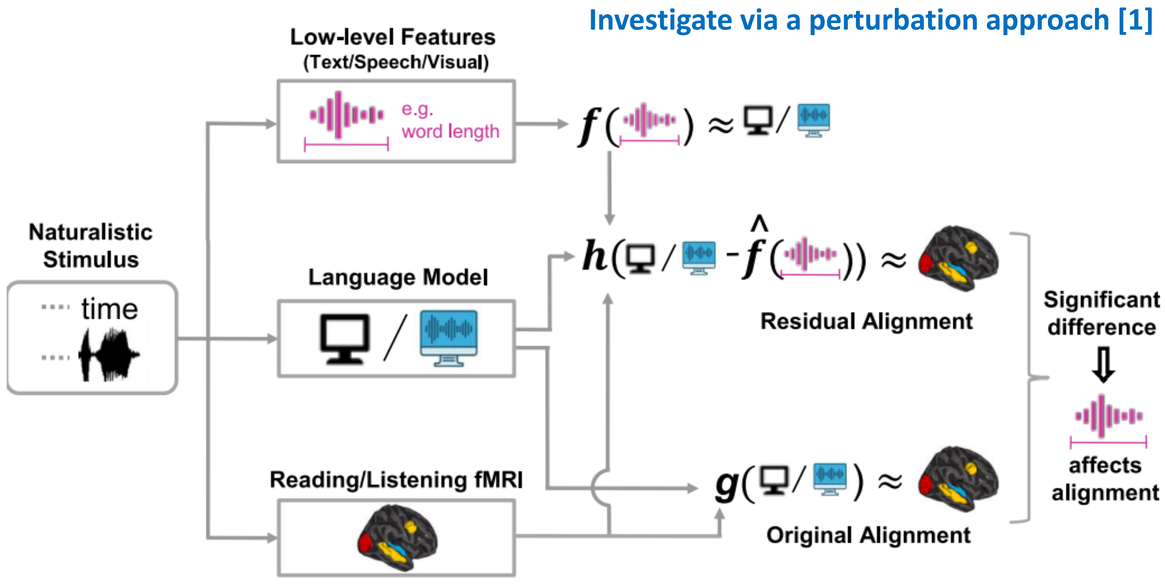
1. Why do **text-based language models** (📄) predict early auditory cortices to an impressive degree?
2. What types of information do language models truly predict in the 🧠?
3. How is the alignment between brain recordings and language model representations affected by the elimination of low-level stimulus features?

Text- vs. Speech-based language models



What types of information lead to high brain alignment?

Investigate via a perturbation approach [1]



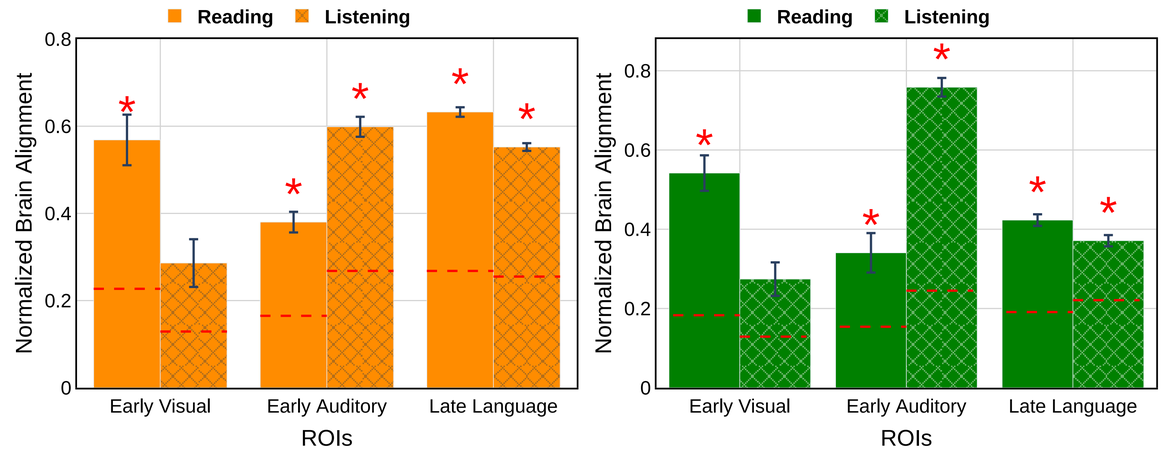
Naturalistic Story Datasets [2]

Dataset	Modality	# Subjs	1-TR	# TRs
Subset-Moth-Radio-Hour	Reading	6	2.0045s	4028
Subset-Moth-Radio-Hour	Listening	6	2.0045s	4028

Low-level stimulus features

Textual	Number of letters, Number of words, and Word length std
Speech	Number of phonemes, MFCC, Phonological, Articulation, PowSpec, Mel, DiPhones and fBank
Visual	Motion energy

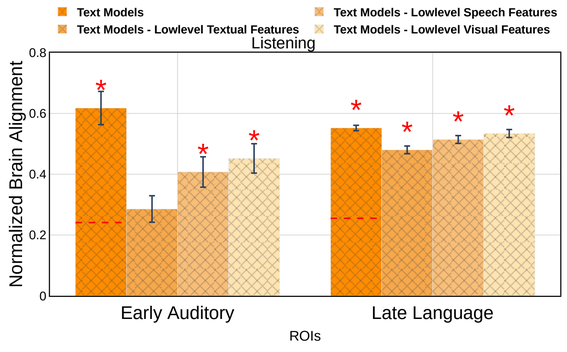
Both types of language models show high brain alignment in late language regions and significantly predict in early sensory regions



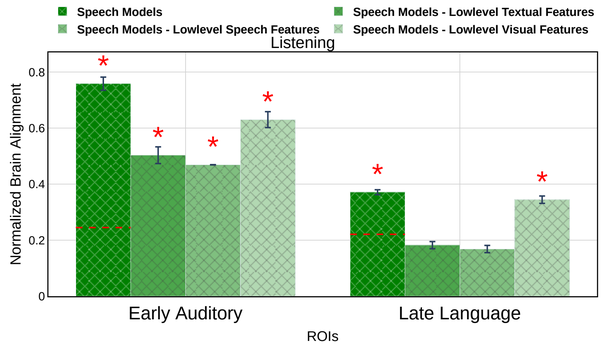
(a) Text Models (b) Speech Models

Removal of low-level features results in divergent brain activity predictions

Text models' alignment in early auditory due to low-level textual features

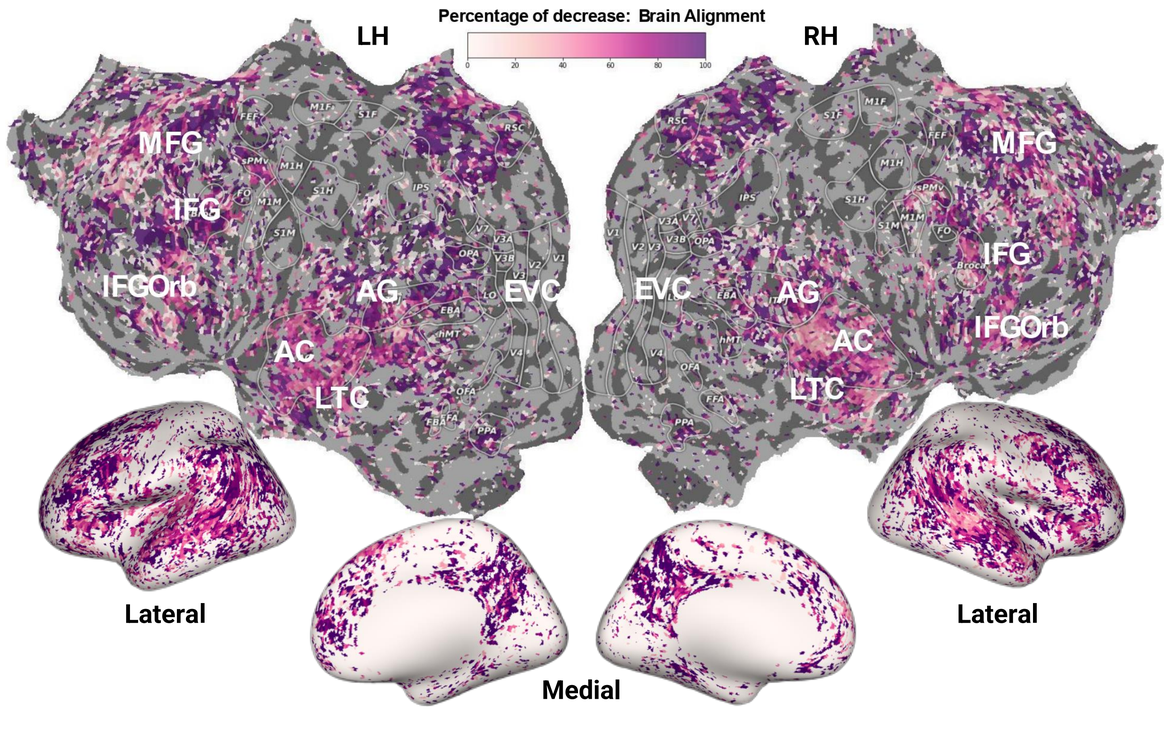


Speech models' alignment in late language entirely due to low-level stimulus feature



(a) Text Models (b) Speech Models

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



Conclusions for neuro-AI research field

- **Speech models** (🔊) useful for modeling early listening: investigate speech models to learn more about early auditory cortex.
- **Text models** (📄) useful for modeling late language in both listening and reading.
- But, more work to do for a complete end-to-end model of reading and listening in 🧠.

References

[1] Toneva et al. 2022. Combining computational controls with natural text reveals aspects of meaning composition. Nature computational science.

[2] Deniz et al. 2019. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. Journal of Neuroscience.

Acknowledgements

This work was partially funded by the German Research Foundation (DFG) - DFG Research Unit FOR 5368.