Alignment between LMs and brains depends on semantics and syntactic structure

Abraham Jacob Fresen[1], Micha Heilbron[1,2]

MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

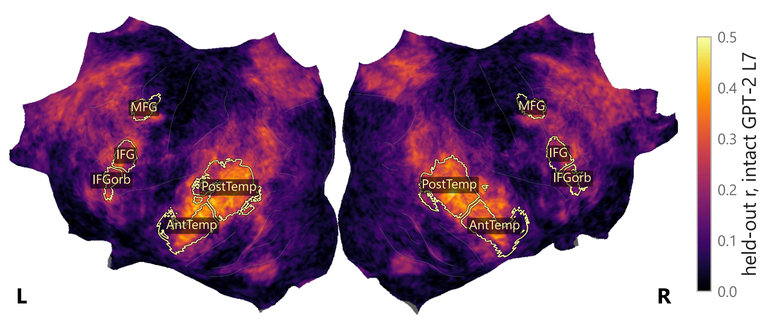
CONCLUSIONS

- Complex syntactic features contribute to brain alignment beyond word classes.
- Linguistic features provide alignment beyond surface-level features such as word rate and surprisal.
- Low-dimensional semantic features outperform all other feature sets.

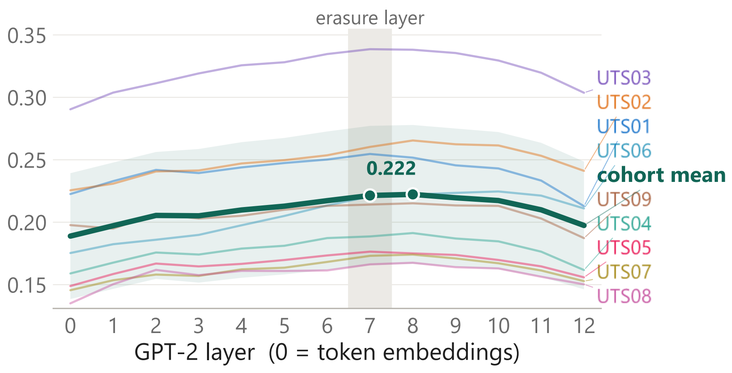
RESULTS

Concept	K	Decode before	Decode after	Δr	Δr, labels permuted
POS	16	+0.85	-0.02	-0.039	-0.019
Labels	44	+0.65	0.00	-0.052	-0.024
Labels x Depth	86	+0.42	-0.02	-0.063	-0.025
Semantic block	6	+0.77	0.00	-0.079	-0.017
Word rate	1	+1.00	-0.20	-0.015	—

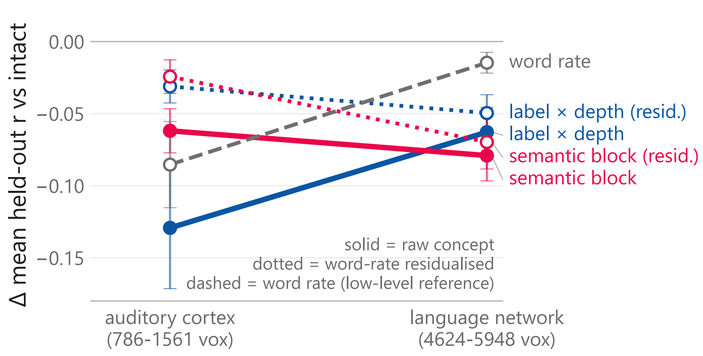
GPT-2 aligns across the language network



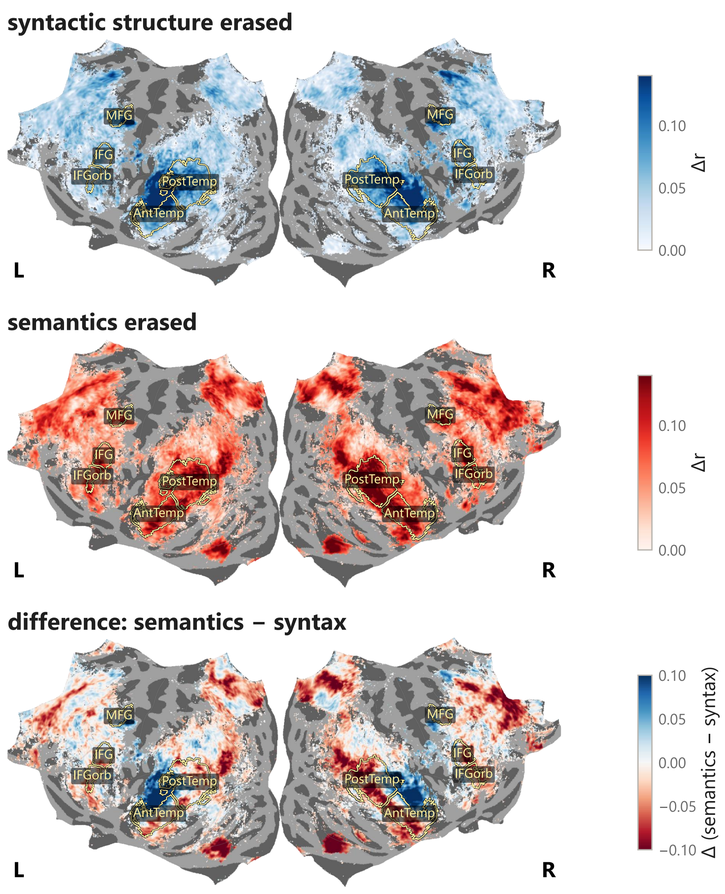
Alignment peaks in the middle layers



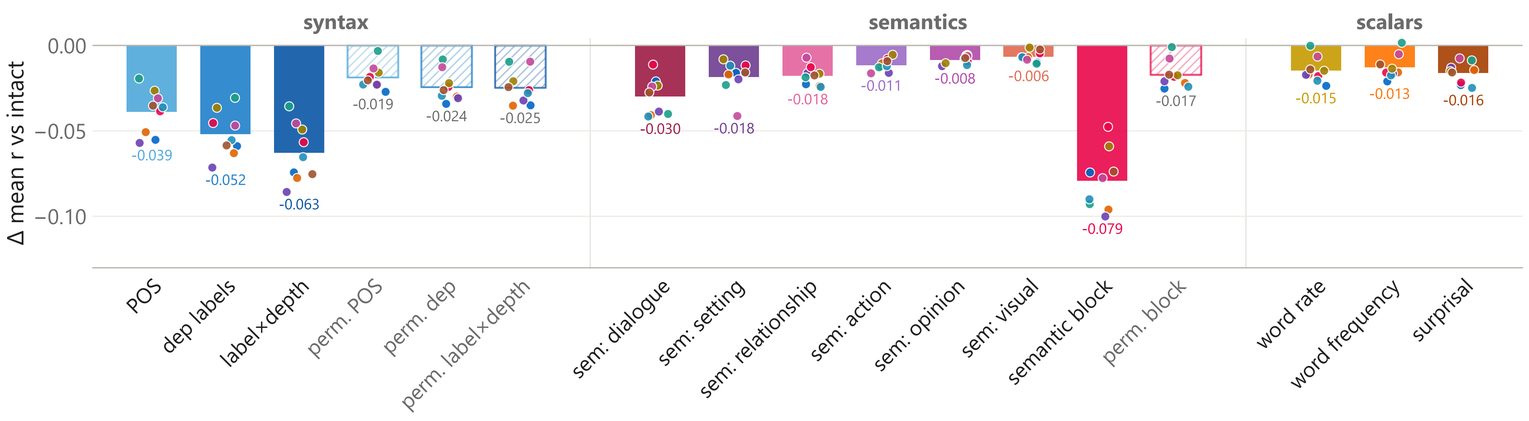
Language-network effects are not word rate



Alignment drop after LEACE erasure



Every erasure costs alignment; semantics most, syntactic structure more than word class

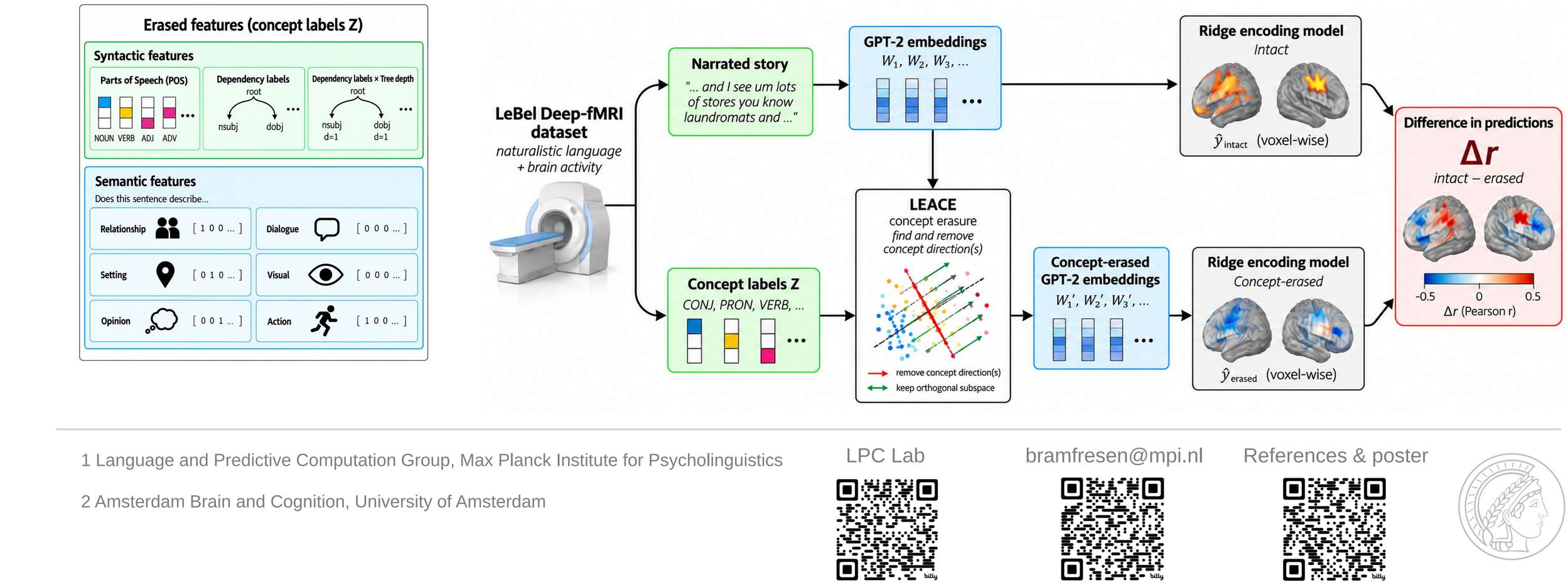


METHODS

- LeBel dataset: high-quality fMRI of nine participants listening to 27 stories[14].
- GPT-2 hidden states of the transcripts.
- Three syntactic features of increasing detail and six semantic features (GPT-4 labels from Benara et al.[16]).
- Each concept is erased from the hidden states with LEACE[17]; erasure is verified by decoding.
- Ridge encoding models[14] are fitted on intact and erased features and tested on a held-out story.
- Contribution = Δr per voxel, erased minus intact, compared against random-subspace and permuted-label controls.

BACKGROUND

- Transformers predict human brain responses to language remarkably well[1,2]. But what information drives this alignment?
- The candidates are debated: semantics[5,6] or syntax[7,8], surface features[9,10] or deeper meaning[11,12].
- By erasing these features from GPT-2 layers and measuring how this affects brain prediction, we ask how much each contributes and where.



1 Language and Predictive Computation Group, Max Planck Institute for Psycholinguistics

2 Amsterdam Brain and Cognition, University of Amsterdam

LPC Lab

bramfresen@mpi.nl

References & poster