Brain-like functional organization in topographic transformer models of language processing

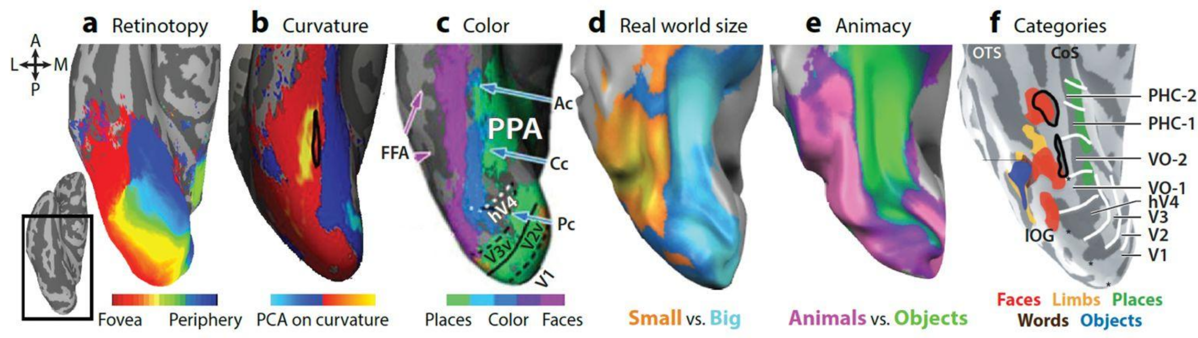
GitHub ICLR Re-Align paper

Taha Binhuraib[1], Greta Tuckute[2], Nicholas M. Blauch[3]

[1]Novus Technologies, [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, [3]Department of Psychology, Harvard University

The human brain is topographically organized

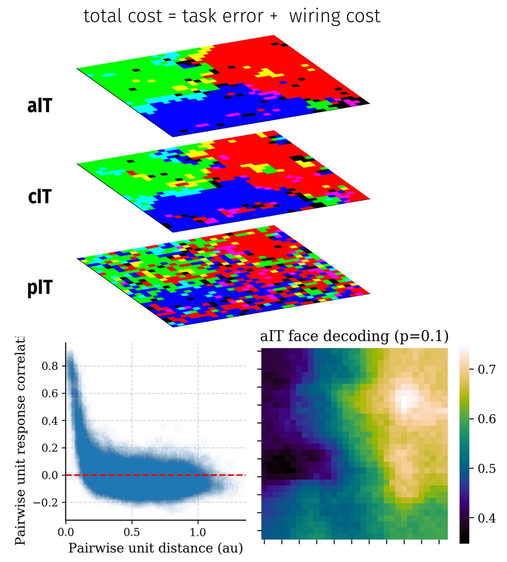
Topographic vision models have begun to explain the functional organization of the visual cortex



Arcaro & Livingstone, 2024

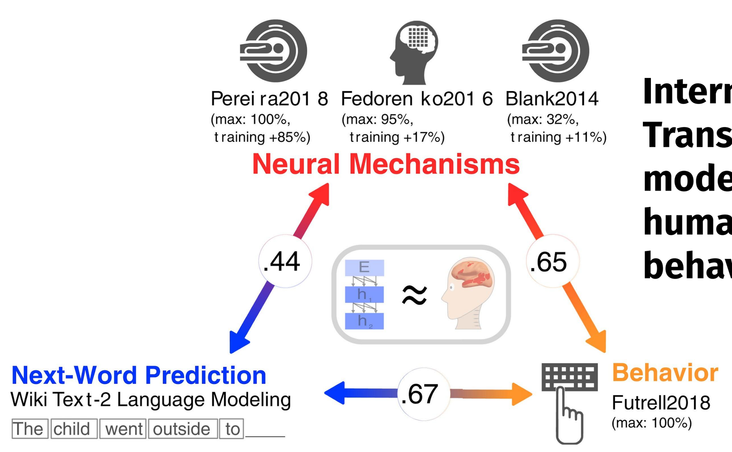
$$\mathcal{L} = \mathcal{L}_t + \lambda_w \mathcal{L}_w$$

total cost = task error + wiring cost



The Interactive Topographic Network (Blauch et. al 2022) simulates the emergence of functional specialization from **task optimization** under explicit biological **connectivity constraints**

Other models have provided converging support (TDANN, VTC-SOM, All-TNN, TVAE, etc.)



Schrimpf et. al, 2021

Internal representations of Transformer large language models (LLMs) can predict human brain responses and behavior

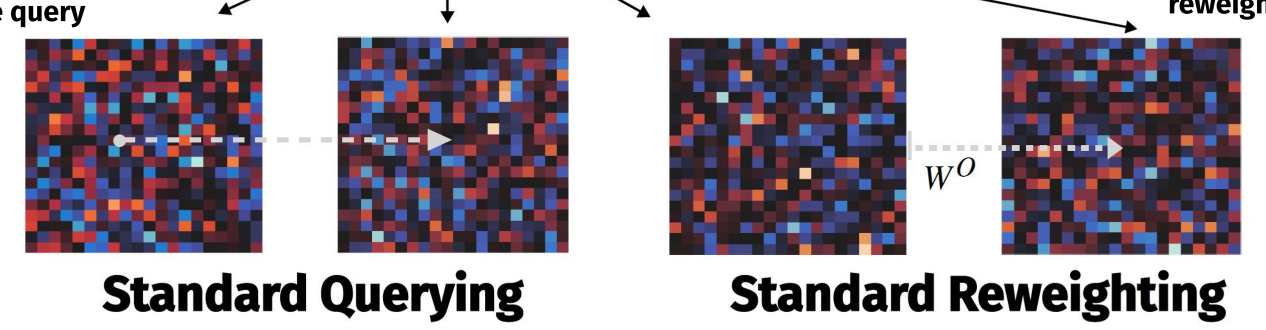
Question: Can we obtain topographic organization of linguistic representations within a transformer architecture?

Adding topographic priors to self-attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V\, W^O$$

Each key is associated with a single query

Fully-connected linear reweighting



$$\text{Attention}_{\text{SQR}}(Q, K, V) = \text{softmax}\left(\frac{QMK^T}{\sqrt{d_k}}\right) V\, W^O_{\text{local}}$$

Each key is associated with a *spatial* pool of queries

Locally-connected linear reweighting

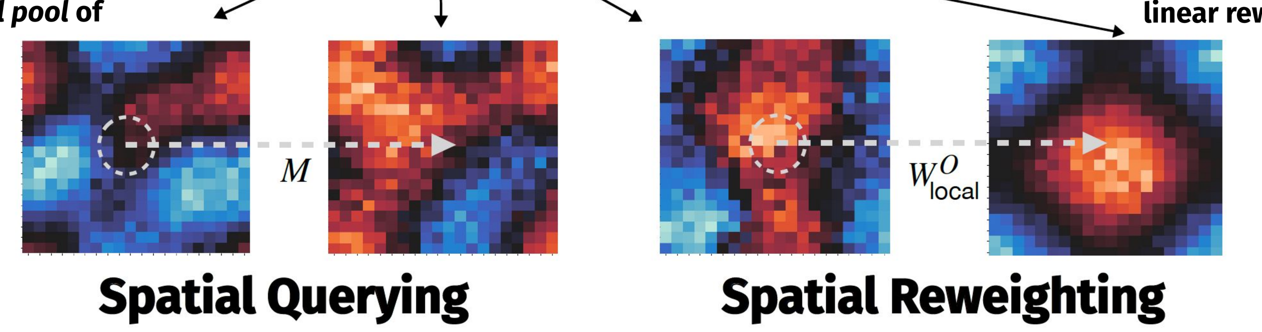


Figure 1: Spatial querying and reweighting operations in the "Topoformer"

Model training and evaluation

- Train a single-head 16-layer Topoformer BERT model with Masked Language Modeling objective following Geiping and Goldstein's (2022) *cramming* paradigm on the Bookcorpus-Wikipedia dataset.
- Evaluate task performance on the GLUE benchmark.

BERT Model	MNLI	SST-2	STSB	RTE	QNLI	QQP	MRPC	CoLA	GLUE
multihead	83.0/83.2	91.6	84.8	54.7	88.5	86.9	86.4	43.7	78.1
1 head	81.1/81.5	90.0	82.1	51.2	87.6	86.7	84.8	47.5	76.9
Topoformer	80.1/80.1	90.9	75.1	51.2	86.6	86.0	81.5	46.3	75.31

Table 1: Comparison of GLUE performance between non-topographic BERT control models and Topoformer-BERT

Interpreting the emergent topography

	Test Suite	Category	Example
semantic contrasts	Intactness	Intact	She scored 2 goals in the soccer game.
		Scrambled	Soccer scored game. the She in 2 goals.
	Animacy	Animate	The gnu galloped across the savanna, majestic and swift.
		Inanimate	The oven's warm glow promised delicious, freshly baked bread.
	Concreteness	Concrete	She peeled the banana slowly, savoring its sweet, ripe aroma.
		Abstract	Her motive for volunteering was purely altruistic and kind.
	Visuomotor	Visual	To solve problems, I often visualize them in my mind.
		Motor	His grip on the rope tightened as he climbed higher.
	Semantic Acceptability	Acceptable	A sunflower has yellow petals.
		Unacceptable	A peanut has yellow petals.
minimal pair syntax contrasts	Agreement	Matched	The authors that hurt the senator are good.
		Mismatched	The authors that hurt the senator is good.
	Licensing	Matched	The authors that liked the senator hurt themselves.
		Mismatched	The authors that liked the senator hurt himself.
	Garden-Path	Ambiguous	As the criminal shot the woman with her young daughters yelled at the top of her lungs.
		Unambiguous	As the criminal fled the woman with her young daughters yelled at the top of her lungs.

Table 2: Overview of test suites with sentence examples. Each test suite had 38 sentences in each category, for a total of 76 sentences in each suite.

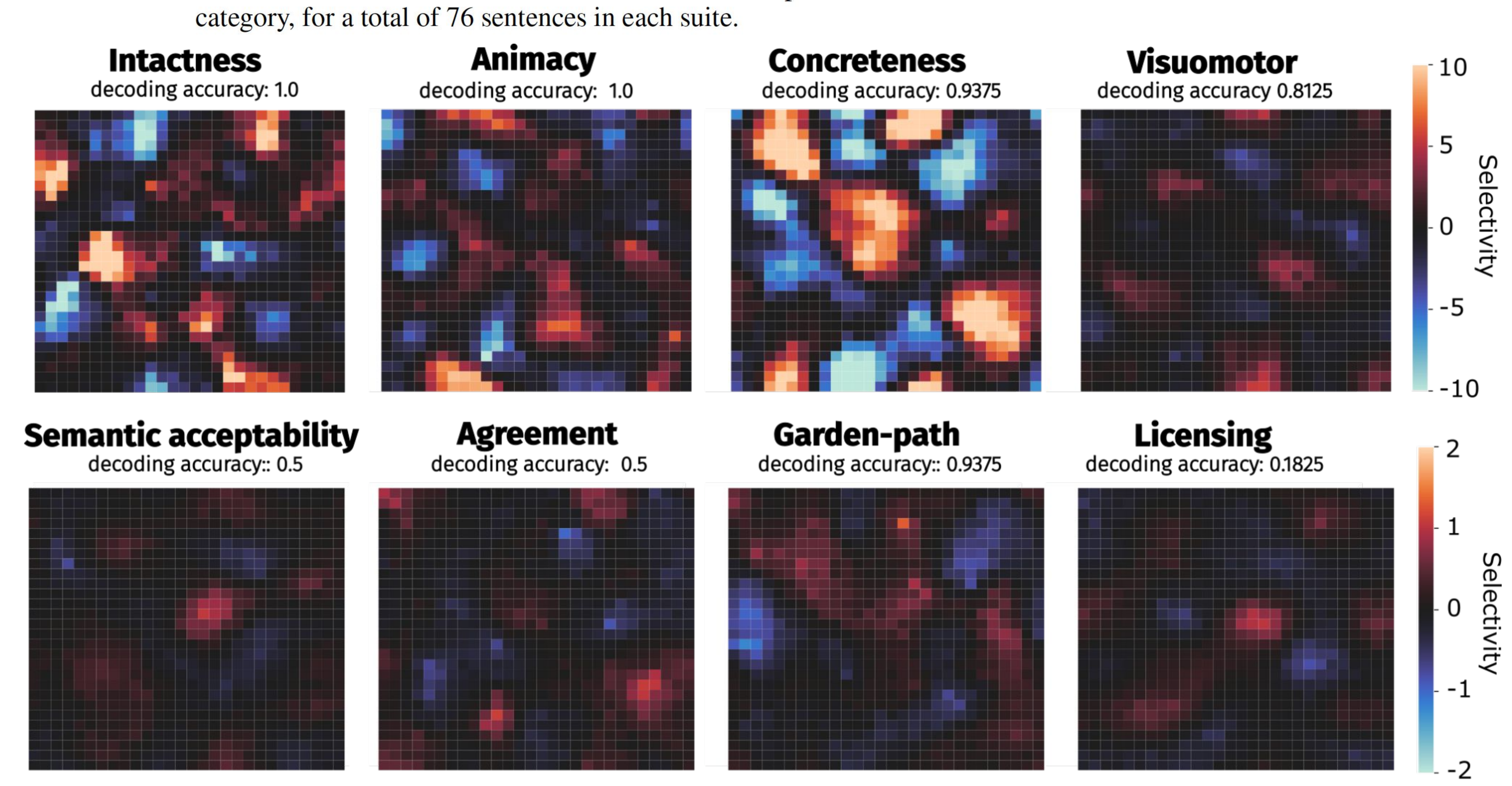


Figure 3: Selectivity-based interpretation of topographic organization in Topoformer-BERT.

Visualizing topography

characterize topography at multiple spatial scales

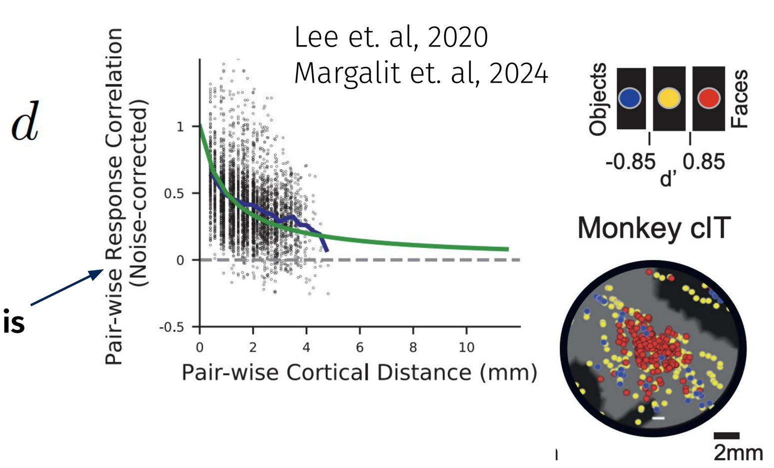
$$t_{g,d} = r_s(-R_{i,j}, \mathcal{D}_{i,j}) \;\; \forall \;\; i,j : D_{i,j} < d$$

summarize topography over all scales

$$t_g = \{t_{g,d_0}, \ldots, t_{g,d_n}\}$$

$$\bar{t}_g = \frac{1}{n}\sum_i t_g^i$$

Local correlation is a good index of topography



Quantification of topography in all layers of Topoformer-BERT

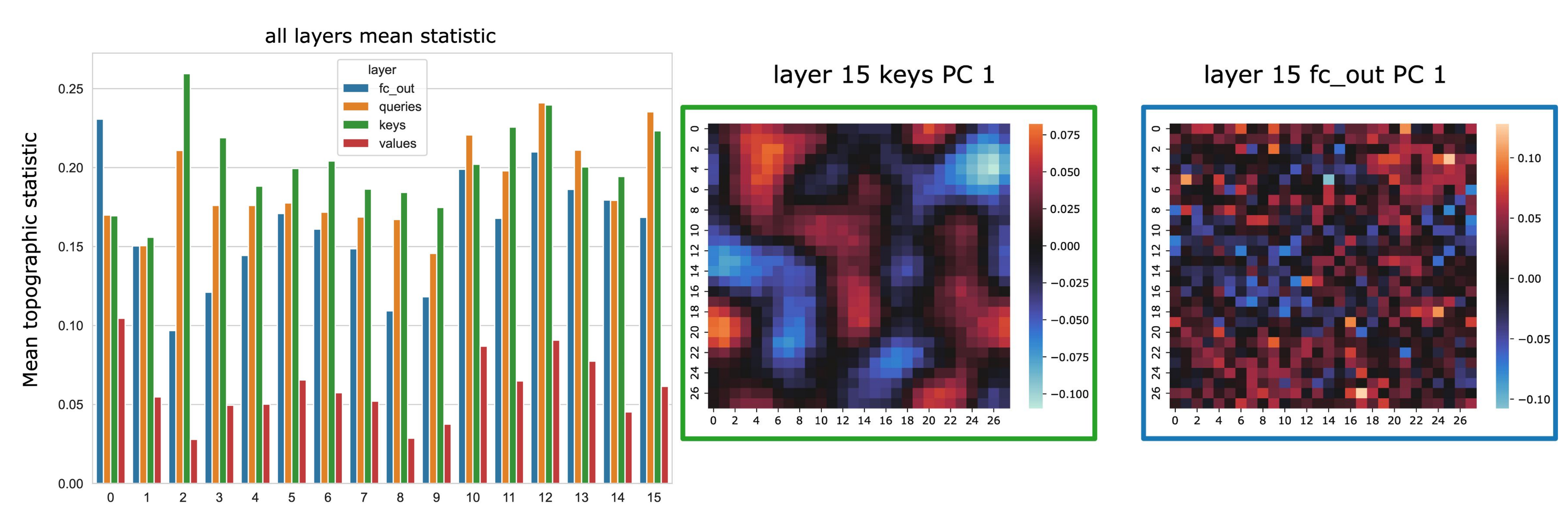


Figure 2: Topographic organization across all layers of Topoformer-BERT.

Brain-model alignment via cross-decomposition

PLS-SVD: SVD on the cross-covariance matrix of X and Y to find aligned components

$$X^T Y = U \Sigma V$$

brain responses (n x p); model embeddings (n x m); brain loadings (p x d); singular values (d x d diag); model loadings (transpose) (d x m)

With held-out data compute scores:

$$X_x = XU$$

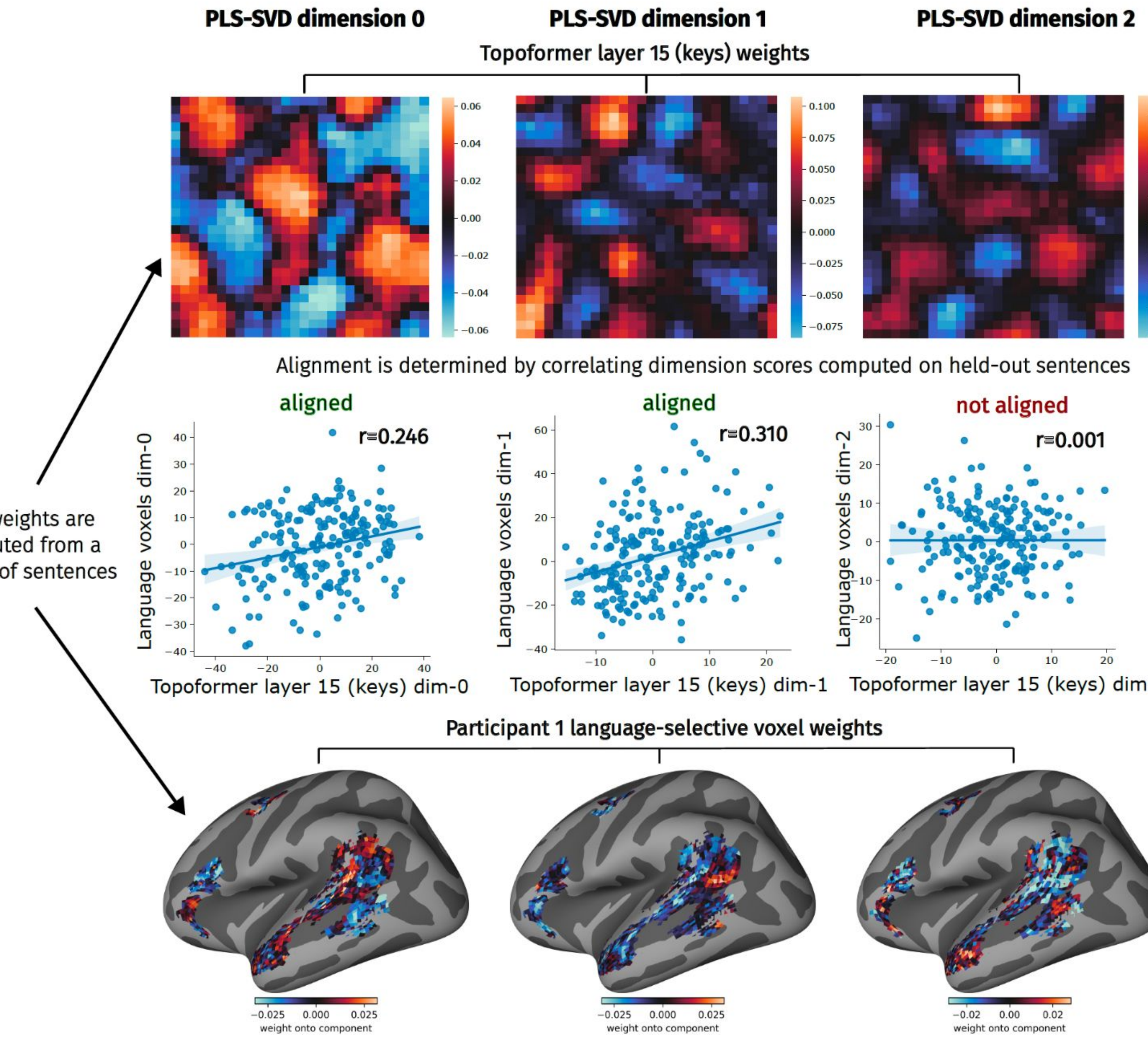
$$Y_c = YV^T$$

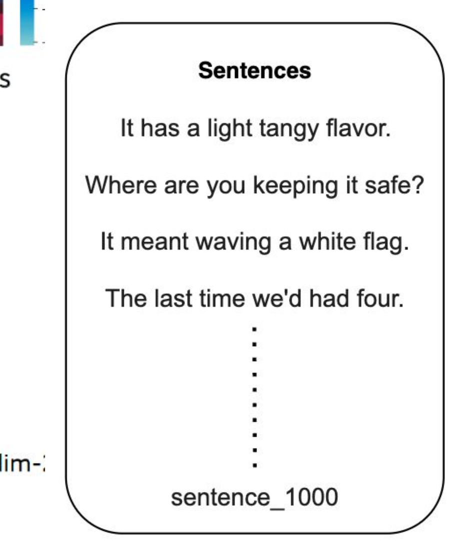
Compute alignment of i-th components

$$a^{(i)} = r(X_c^i, Y_c^i)$$

A.

PLS-SVD finds shared dimensions between language-selective brain voxels and Topoformer model units.





5 participants, reading 1,000 sentences in fMRI

Tuckute et al., 2024

B.

Alignment of brain/model dimensions across all subjects and attention sublayers for layer 15

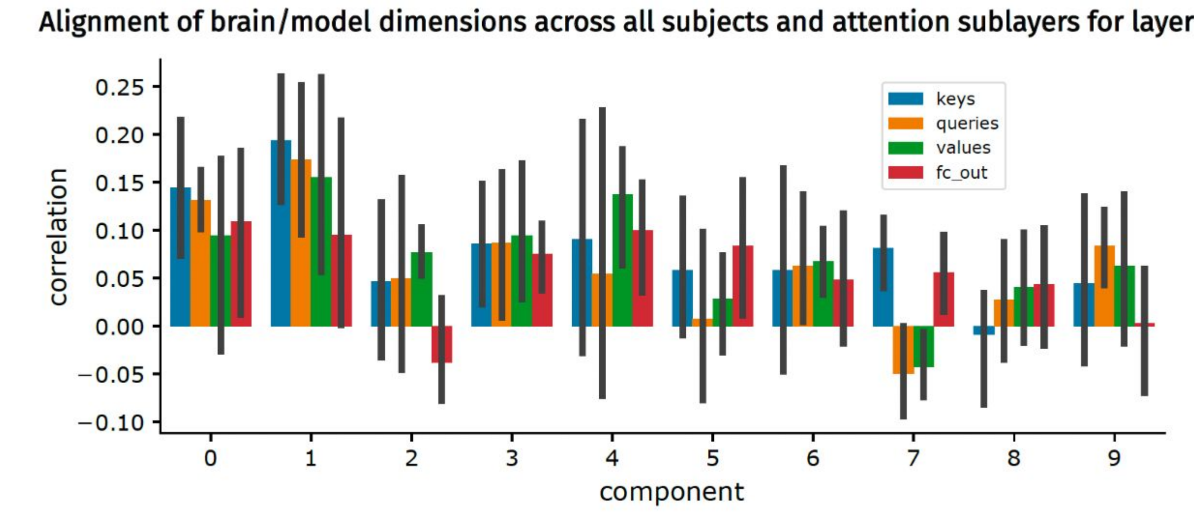


Figure 4: Alignment of topographic representations in the human language network and Topoformer-BERT

Conclusions

- Topoformers produce a topographic organization of linguistic representations
- There is significant alignment in the topographic components of the model and human language network, but both resist neat interpretation (so far)
- Our work provides a new perspective on graded functional topography within the language network, which should be investigated further

Ongoing and future work

- Apply to Topoformer to more domains (vision, audition, ...)
- Residual stream topography, wiring minimization, areally mappable models
- Advance interpretability of model (e.g. SAEs) and brain topography (e.g. more data)
- Unify principles of functional organization within and across brain areas

References:

- Blauch, N. M., Behrmann, M., & Plaut, D. C. (2022). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. Proceedings of the National Academy of Sciences of the United States of America, 119(3). https://doi.org/10.1073/pnas.2112566119
- Geiping, J., & Goldstein, T. (2023). Cramming: Training a Language Model on a single GPU in one day. In Proceedings of the International Conference on Machine Learning (pp. 11117-11143). PMLR.
- Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118. https://doi.org/10.1073/pnas.2105646118
- Arcaro, M., & Livingstone, M. (2024). A Whole-Brain Topographic Ontology. Annual Review of Neuroscience, 47. Annual Reviews.
- Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., & Fedorenko, E. (2024). Driving and suppressing the human language network using large language models. Nature Human Behaviour, 8(2), 1–18.
- Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., & DiCarlo, J. J. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. BioRxiv. https://doi.org/10.1101/2020.07.09.1851
- Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. K. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. Neuron, 112(14), 2435–2451.e7. https://doi.org/10.1016/j.neuron.2024.04.018