TOPOFORMER: BRAIN-LIKE TOPOGRAPHIC ORGANIZATION IN TRANSFORMER LANGUAGE MODELS THROUGH SPATIAL QUERYING AND REWEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial functional organization is a hallmark of biological brains: neurons are arranged topographically according to their response properties, at multiple scales. In contrast, representations within most machine learning models lack spatial biases, instead manifesting as disorganized vector spaces that are difficult to visualize and interpret. Here, we propose a novel form of self-attention that turns Transformers into "Topoformers" with topographic organization. We introduce *spatial querying* — where keys and queries are arranged on 2D grids, and local pools of queries are associated with a given key — and *spatial reweighting*, where we convert the standard fully connected layer of self-attention into a locally connected layer. We first demonstrate the feasibility of our approach by training a 1-layer Topoformer on a sentiment classification task. Training with spatial querying encourages topographic organization in the queries and keys, and spatial reweighting separately encourages topographic organization in the values and self-attention outputs. We then apply the Topoformer motifs at scale, training a BERT architecture with a masked language modeling objective. We find that the topographic variant performs on par with a non-topographic control model on NLP benchmarks, yet produces interpretable topographic organization as evaluated via eight different linguistic test suites. Finally, analyzing an fMRI dataset of human brain responses to a large set of naturalistic sentences, we demonstrate alignment between low-dimensional topographic variability in the Topoformer and human brain language network. Scaling up Topoformers further holds promise for greater interpretability in NLP research, and for more accurate models of the organization of linguistic information in the human brain.

1 INTRODUCTION

Biological brains are spatially organized, containing category-selective areas (Kanwisher, 2010), broad feature maps that tile individual cortical areas (Konkle & Oliva, 2012; Bao et al., 2020) and the cortex more broadly (Huth et al., 2012; 2016; Margulies et al., 2016), and large-scale distributed networks (Yeo et al., 2011; Braga et al., 2020). Particularly within brain regions, this spatial organization is one way in which the human brain, a vastly complex "black box", is more naïvely interpretable than modern deep neural networks (DNNs), whose units have functional properties organized without simple spatial priors. Recent work in computational neuroscience has bridged this gap in DNNs trained for vision, demonstrating that local smoothness or wiring cost minimization objectives can be incorporated into DNNs to encourage the development of smooth functional organization of responses, which can then be easily visualized in 2D (Lee et al., 2020a; Blauch et al., 2022; Keller & Welling, 2021; Doshi & Konkle, 2021; Margalit et al., 2023; Lu et al., 2023), building upon classic approaches (Kohonen, 1982; Jacobs & Jordan, 1992). In addition to simulating topographic properties within regions, topographic vision models have also explained the hierarchical organization of topographic information from earlier to later visual areas (Margalit et al., 2023; Lu et al., 2023). One topographic vision model has even demonstrated the emergence of spatial clusters corresponding to ventral, lateral, and dorsal streams of the visual system (Finzi et al., 2021). Collectively, topographic vision models are helping to unify a computational understanding of the functional organization of the visual system.

However, topographical priors have not yet been built into models of linguistic processing, despite tremendous progress in the development of natural language processing (NLP) models and their application in cognitive science and neuroscience Wilcox et al. (2020); Gauthier et al. (2020); Schrimpf et al. (2021); Caucheteux & King (2022); Goldstein et al. (2022a); Tuckute et al. (2023). In NLP, Transformer language models (LMs) have undoubtedly established themselves as the leading architecture for language tasks (Vaswani et al., 2017; Radford et al., 2018; Brown et al., 2020; OpenAI, 2023), displaying human-like language understanding and generation for the first time. In cognitive science and neuroscience, these LMs have emerged as the most quantitatively accurate models of human language processing. They generate probabilities of upcoming words that explain reading behavior of humans Wilcox et al. (2020); Merkx & Frank (2021); Shain et al. (2022), and their internal activations can explain the neural signals of humans reading or listening to naturalistic sentences or stories at the granularity of fMRI voxels and intracranial recordings Schrimpf et al. (2021); Goldstein et al. (2022b); Tang et al. (2023). Despite the success of these LMs, they remain difficult to interpret, and incomplete as models of brain function.

In the current work, our aim is to bridge these gaps by inducing a topographic organization of features within the Transformer architecture. We employ local-connectivity based approaches inspired by recent topographic vision models (Blauch et al., 2022; Keller & Welling, 2021) to the language domain, asking whether we can obtain topographic organization of linguistic representations within a Transformer architecture via spatial constraints. To do so, we introduce two computational motifs — *spatial querying* and *spatial reweighting* — to the self-attention layer, which encourage the development of topographic organization in separate components of the self-attention layer. We call Transformer models employing these constraints **Topoformers**. We show that we can scale these topographic motifs to a large BERT Topoformer model trained with a masked language modeling objective, and that topographic organization develops within each hierarchical layer of the network, without significantly compromising task performance. We interpret this topography using a novel suite of 8 semantic and syntactic tests. Last, we demonstrate that the topographic representations of the Topoformer can be aligned with the topographic representations of the human functionally-defined language network in multiple subjects. In summary, our work demonstrates for the first time that Transformer models can be trained to exhibit topographic organization similar to the human brain, and paves the way for further interpretability work leveraging spatial priors.

2 METHODS

In this study, we propose two approaches for enforcing topographic organization in a Transformer layer. Both methods rely on the use of local communication to introduce spatial constraints that encourage the formation of spatially organized linguistic representations.

2.1 SPATIAL QUERYING

We begin with the standard self-attention operation used by Vaswani et al. (2017). In this formulation, every token embedding is projected onto a set of queries, keys, and values, and the query of a given token is associated with a corresponding key of all other tokens. Spatial querying works by associating a local pool of queries with a given key. The locality is parameterized with a width parameter r_{SQ} determining the fraction of units in a given key's circular receptive field (RF). For simplicity, we examine the case of a simple non-weighted sum of queries. This is achieved by inserting a binary intermediate matrix $M \in \mathbb{R}^{d \times d}$, where d is the embedding dimension, and the columns of M determine the spatial pool of queries associating with a given key. Essentially, this makes it such that the dot product attention between a given pair of tokens is not between individual queries and keys, but local pools of queries and individual keys. This biases the representations of queries to be locally smooth, and the representations of keys to have a spatial correspondence with the queries. The local pooling of spatial querying can be visualized with a simple example, assuming a model dimension of $d = 3$, and 2 tokens.

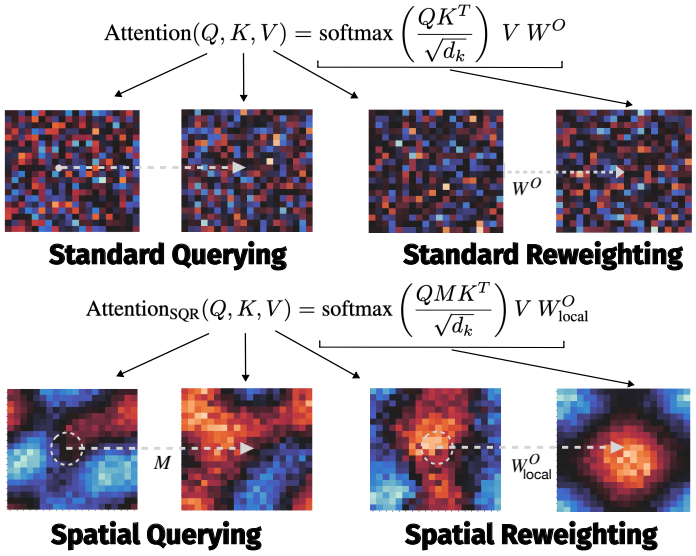


Figure 1: Spatial querying and reweighting operations in the "Topoformer".
The first row shows standard querying operations in the attention module of a single-head Transformer, and the second row shows the spatial counterparts used in the Topoformer. Standard querying associates a single query dimension of token *i* with a single key dimension of token *j*. In contrast, spatial querying associates a local pool of query dimensions with a given key dimension, through the intermediate local pooling matrix M. Standard (dense) reweighting applies a fully connected linear layer W^O to the outputs of a single attention head (typically to combine the outputs of multiple attention heads). In our formulation, we use a locally connected layer W^O_{local} in its place (spatial reweighting). While the figure illustrates querying for a pair of tokens and reweighting for a single token, when processing a full sequence, there is a 2D grid of the form shown here for each token. Each heatmap shows the second PC of responses (top: control model, bottom: SQR model).

$$
\begin{aligned}
QMK^T &= \begin{pmatrix} Q_{1,1} & Q_{1,2} & Q_{1,3} \\ Q_{2,1} & Q_{2,2} & Q_{2,3} \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} K_{1,1} & K_{2,1} \\ K_{1,2} & K_{2,2} \\ K_{1,3} & K_{2,3} \end{pmatrix} \\
&= \begin{pmatrix} Q_{1,1}+Q_{1,2} & Q_{1,2}+Q_{1,3} & Q_{1,1}+Q_{1,3} \\ Q_{2,1}+Q_{2,2} & Q_{2,2}+Q_{2,3} & Q_{2,1}+Q_{2,3} \end{pmatrix} \begin{pmatrix} K_{1,1} & K_{2,1} \\ K_{1,2} & K_{2,2} \\ K_{1,3} & K_{2,3} \end{pmatrix}
\end{aligned} \tag{1}
$$

We can see that, instead of the rows of the matrix multiplication containing individual queries, they now contain summed local pools of queries. This is the essence of spatial querying. The full self-attention equation with spatial querying (SQ) is given as follows:

$$
\text{Attention}_{\text{SQ}}(Q, K, V) = \text{softmax}\left(\frac{QMK^T}{\sqrt{d_k}}\right) V\, W^O \tag{2}
$$

For simplicity of visualization, we use a single attention head in the Topoformer implementation, but we retain the outer reweighting matrix W^O used in multi-head attention Vaswani et al. (2017). Our motivation for using single-head attention is to ensure that the dominant functional organization occurs *within* a head rather than *across* heads; without further constraints than Eq 2, organization across heads would be non-topographic and thus complicate interpretability and visualizations. Although the dimensionality of the model could be of any size, it's convenient in our implementation for it to be a perfect square, such that it can be reshaped into a $\sqrt{d} \times \sqrt{d}$ grid for visualization purposes. While we work with square grids, theoretically, any 1D, 2D, or 3D arrangements of units could be used to define the spatial position of units. For a visual explanation, Figure 1 compares the differences between standard operations within a self-attention block, and their spatial counterparts.

2.2 SPATIAL REWEIGHTING

Spatial querying only imposes a topographic relationship between queries and keys to encourage the development of topographic organization of the values and self-attention outputs (hereafter fc_out), we convert the outer reweighting matrix W^O to a locally connected layer W^O_{local}. By using locally connectivity in W^O_{local}, we encourage the model to learn more localized feature representations in the values and attention outputs (analogous to what spatial querying does for the queries and keys). We parameterize local connectivity using a width parameter r_{SR} that determines the fraction of units within a given unit's circular receptive field (RF). Our locally connected linear layer $W^O_{local} \in \mathbb{R}^{d \times d}$ is situated in our network as follows:

$$\text{Attention}_{\text{SQR}}(Q, K, V) = \text{softmax}\left(\frac{QMK^T}{\sqrt{d_k}}\right) V \, W^O_{\text{local}} \tag{3}$$

Preliminary experiments demonstrated the need to use large positive weights to fully encourage the development of topographic organization. Thus, we initialize $W^O = |W^O_i * 10|$, where W^0_i is a standardly initialized PyTorch linear layer. This operation, denoted as spatial reweighting, has the effect of enhancing local correlations, commonly viewed as a hallmark of topographic organization (Lee et al., 2020b; Blauch et al., 2022; Margalit et al., 2023). These excitatory feedforward connections mimic the dominant role of excitatory pyramidal neurons in between-area cortical communication in biological brains (Laszlo & Plaut, 2012; Blauch et al., 2022).

3 RESULTS

3.1 TRAINING A 1-LAYER TOPOFORMER ON A SUPERVISED TASK

We begin by training 1-layer single-head encoder-only Topoformer (bidirectional attention) on the IMDB sentiment analysis dataset Maas et al. (2011), which classifies movie reviews as having a positive or negative overall sentiment. The local connectivity in the spatial querying and spatial reweighting operations are controlled through a hyperparameter r_{SR} that sets the radius of spatial receptive fields (RFs). We investigated the effect of different RF sizes in the **Topoformer-SQR** model, finding that smaller r_{SQ} values yield better accuracy and topography, while the network is more robust to the r_{SR} hyperparameter (see Appendix A.9 for details). In the following, we report results for an RF size of 0.3 for the SQ model and 0.1 for the SQR model. We set $d = 400$.

3.1.1 TOPOFORMER-SQ

We first describe the results of a model using only spatial querying (Topoformer-SQ). Following training, Topoformer-SQ achieved an accuracy of 0.81 on the IMDB sentiment test set. In comparison, an identical 1-layer Transformer model without spatial querying achieved an accuracy of 0.83. To probe its topographic organization, we conducted selectivity analyses and principal component analysis (PCA) to investigate the unit activation patterns in the different layers, as shown in Figure 2 (see Appendix A.1, A.3 for details). Our selectivity analysis was designed to contrast the response magnitudes for positive and negative sentiment sentences. As expected, this analysis revealed a topographic organization in the keys and query sublayers, but not in the values of fully-connected layers. We next performed PCA to assess generic forms of topographic variability in the model representations. We found that the weights of the first two principal components (PCs) exhibited a smooth topography in the keys and queries, with the second PC spatially aligned to the selectivity for both representations. This demonstrates that the network has learned to organize its dominant modes of variability spatially, a hallmark of topographic functional organization.

3.1.2 TOPOFORMER-SQR

We next trained a model incorporating both spatial querying and reweighting (Topoformer-SQR). This model achieved a test set accuracy of 0.75 on the IMBD sentiment test set, slightly lower than the Topoformer-SQ model. We performed identical probing analyses to those performed in the previous section, highlighting the results for the values and fully-connected (fc_out) representations in Figure 2B. We found that the Topoformer-SQR exhibited more pronounced topographic organization in the values and fc_out layers compared to Topoformer-SQ. This suggests that the local

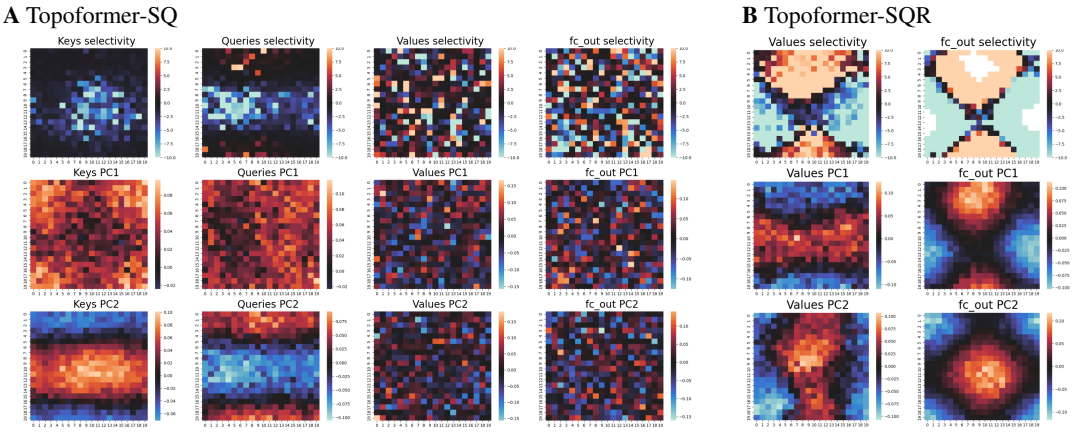


Figure 2: Topographic organization across sublayers with spatial querying and reweighting. **A.** Topoformer-SQ produces topography in the keys and queries, but not the values or self-attention outputs. Each column shows a different sublayer representation within a self-attention block (keys, queries, values, and fc_out). The representations were obtained by averaging across the tokens in each sentence from the IMBD sentiment classification test set (Maas et al., 2011). The first row shows selectivity for positive vs. negative sentiment sentences. The second and third rows show the PC weights for the first and second components, respectively. **B.** Topoformer-SQR produces topography in the values and self-attention outputs. The format is the same as for **A.**, but for brevity we show only the values and self-attention outputs as the keys and queries show a similar topographic organization from spatial querying.

connectivity matrix W^{O}_{local} (see Figure 1 A., and equation 3) successfully enforced a topographic correspondence between the values and attention outputs, as predicted.

3.2 Scaling up: Topoformer-BERT

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 multi-head and single-head non-topographic BERT control models and Topoformer-BERT, each trained with the Cramming procedure (Geiping & Goldstein, 2022).

We next scaled up the Topoformer motifs to train a BERT model using a Masked Language Modeling objective (**Topoformer-BERT**). We followed the training paradigm introduced by (Geiping & Goldstein, 2022). We trained a 16-layer BERT model on the Bookcorpus-Wikipedia dataset (Zhu et al., 2015) for 12 hours (see Appendix A.2 for more details). To provide a control for our Topoformer-BERT model, we trained a standard, non-topographic single-head BERT model with identical parameters and training procedure as our Topoformer-BERT (besides the lack of topographical motifs, Appendix A.7). To evaluate the models' performance on natural language tasks, we followed the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2019) procedure as described in (Geiping & Goldstein, 2022), testing each model on all tasks besides WNLI as in (Devlin et al., 2019). Critically, we observed that the task performance of Topoformer-BERT on the GLUE Benchmark was similar to that of the non-topographic model counterpart, suggesting that our added spatial constraints were not significantly hindering task performance Table 1. Having established that Topoformer-BERT is capable of performing linguistic tasks, we move on to characterizing the topographic organization in Topoformer-BERT.

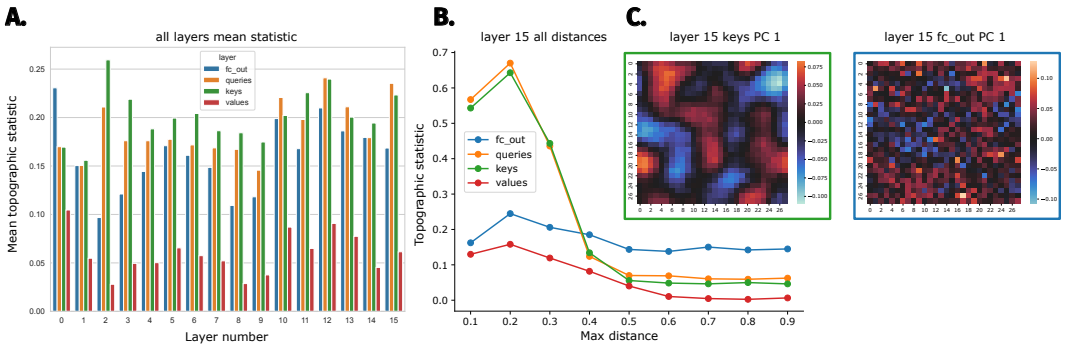


Figure 3: Topographic organization across all layers of Topoformer-BERT.
The generic topography statistic is given by Equation 4. **A.** Mean statistic $\bar{t}_g$ computed over a range of maximum distances **B.** Statistic $t_{g,d}$ computed at each of several maximum distances for layer 15 **C.** Visualization of the first principal component (PC) weights for keys and fc_out sublayers.

First, we systematically quantified the topography of each the 16 Topoformer-BERT layers using a statistic that relates the degree of correlation with the distance between pairs of units (Appendix A.3.2, Equation 4). A high value of the statistic t_g indicates that nearby units tend to be more correlated in their response pattern across sequences than distant units. We plot the mean statistic $\bar{t}_g$ over distance thresholds for all layers in Figure 3A, and the distance-threshold-specific statistic $t_{g,d}$ for layer 15 (Figure 3B. In general, the keys and queries have the greatest degree of topographic organization, and the values show the weakest organization. Nevertheless, each is consistently above 0, driven by very local decay in correlation, as seen in the analysis of $t_{g,d}$ across different maximum distances (Figure 3.

Second, we took a step towards interpreting the emergent topographical structure in Topoformer-BERT. Specifically, we evaluated the selectivity of the unit activations to a set of eight test suites targeting different linguistic properties. All eight test suites consisted of 76 sentences each, and were either based on carefully designed minimal pair sentences based on prior work (Gauthier et al. (2020); Hu et al. (2020); Misra et al. (2023)) or were designed by us to control for the number of words and sentence surprisal (see Appendix A.8 for information and sentence examples for each test suite).

The first suite, **Intactness** tests intact sentences versus their scrambled counterparts, thereby degrading both linguistic form (syntax) and meaning (semantics). The next suites test more targeted linguistic properties: Suites 2 through 4 test three different dimensions of *meaning* that have been extensively investigated in prior work, as specified below. Suite 2 tests **Animacy** (sentences with animate vs. inanimate meanings; Naselaris et al. 2009; Connolly et al. 2012; Konkle & Caramazza 2013, suite 3 tests **Concreteness** (sentences with concrete vs. abstract meanings; Binder et al. 2005; Fiebach & Friederici 2004), and suite 4 tests **Visuomotor** properties (sentences with visual vs. motor meanings; Desai et al. 2010; Lynott et al. 2020). The next suite (5) tests **Semantic acceptability** using minimal pair sentences (Conceptual Minimal Pair Sentences; Misra et al. 2023). The final three suites test three different dimensions of *form* using suites from SyntaxGym Gauthier et al. (2020); Hu et al. (2020): Suite 6 tests **Agreement** (Subject-Verb Number Agreement), suite 7 tests **Licensing** (Reflexive Number Agreement), and suite 8 tests **Garden-Path** ambiguity (Verb Transitivity).

We performed selectivity analyses for these eight test suites (Figure 4). These analyses intuitively ask whether a given unit shows a preference for a particular contrast (e.g., animate versus inanimate sentences). We evinced a strong topographic organization according to broad semantic categories (top row, Figure 4), both in terms of significant topographic selectivity, as well as significant decodability of condition from the distributed pattern of activities. Intriguingly, the selectivity patterns were different across contrasts, implying that semantic distinctions are represented in topographic activity pattern differences across categories. It is important to note that despite the strongly significant selectivity, the mean activity patterns were highly similar across categories within each contrast (Appendix A.9.3, Figure 12): rather than indicating contrasting hot spots of activation for animate and inanimate content, for example, the rank order of unit activities tends to be similar

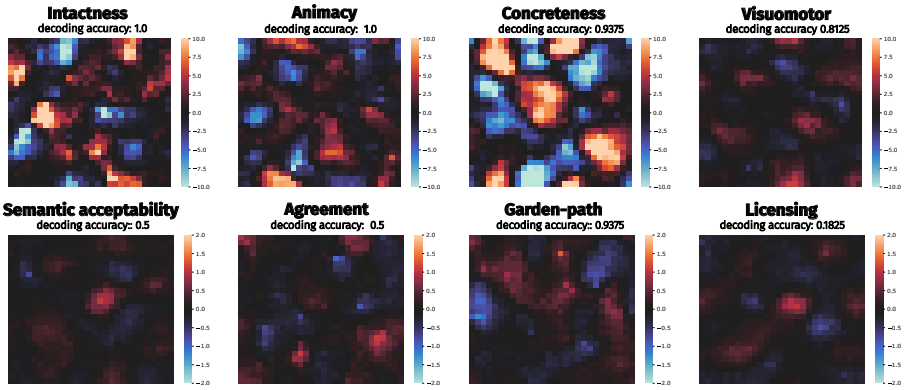


Figure 4: Selectivity-based interpretation of topographic organization in Topoformer-BERT. Each panel shows the selectivity of Topoformer-BERT layer 15 (keys), for a given contrast. Each test suite contains two contrasting conditions each with a set of sentences; unit activities are computed as the mean over tokens for each sentence, and the conditions are contrasted with a t-test. We plot the selectivity significance value (see Appendix A.3), where $s = 2$ corresponds to positive selectivity with $p = 0.01$, and $s = -2$ corresponds to negative selectivity with the same significance level. The first row contains sentences with natural variability, whereas the bottom row contains results from constructed minimal pairs differing in only one word across conditions. To ensure visibility of effects regardless of size, we used different statistic ranges for plotting of each row: $s = 10$ for the top row, and $s = 2$ for the bottom row.

across sentences, and selectivity across conditions is indicative of small yet distinct deviations from the dominant pattern. This is not unique to the Topoformer (Appendix A.9.3, Figure 12E.), however, the Topoformer allows a uniquely intuitive visualization of selectivity patterns in 2D, aiding interpretability.

Second, we turned to more controlled test suites constructed using pairs of minimally different sentences in line with prior work in psycholinguistics and natural language processing (e.g., Linzen et al. 2016; Warstadt et al. 2020). As expected, the effects were lower (bottom row, Figure 4) relative to sentences only matched on length and surprisal (top row). We evidenced weak topographic selectivities to sentences with correct syntactic agreement versus those that do not, for example. The weak selectivities were also reflected in the fact that the condition of interest could not be decoded accurately from the patterns of key activation.

In summary, we quantified the extent of generic topographical organization across all sublayers across the full Topoformer-BERT model, and honed in on selectivities of the topographic organization of the final layer. We analyzed eight different linguistic properties, finding strong effects for naturalistic *semantic* dimensions: animacy, concreteness, and visuomotor properties. Future work should investigate the organization of finer-grained semantic dimensions, as well as more extensive tests for syntactic knowledge.

3.3 Modeling the topographic organization of the human language network

To assess the topographic organization of language in the human brain, we recorded brain responses using event-related fMRI from N=5 participants (4 female, native English speakers) during a sentence reading task. Participants read 1,000 6-word, corpus-extracted sentences that were selected to maximize semantic and stylistic diversity (see A.4). Following standard preprocessing, we used a set of five language masks (”parcels”) that denote brain regions within which most or all individuals in prior studies Fedorenko et al. (2010); Lipkin et al. (2022) showed activity for an extensively validated language localizer contrast between reading of sentences and non-word strings (Fedorenko et al., 2010). For each participant, within these anatomical parcels, we then computed individual functionally-defined regions by comparing responses to sentences and non-words, and taking all voxels with at least weak preferences for sentences ($t > 1$). We then restricted our analyses to these voxels, henceforth the ”language network”. To determine that the language network exhibits spatial smoothness, as in the model, we computed the generic topographic statistic t_g (Equation 4) on

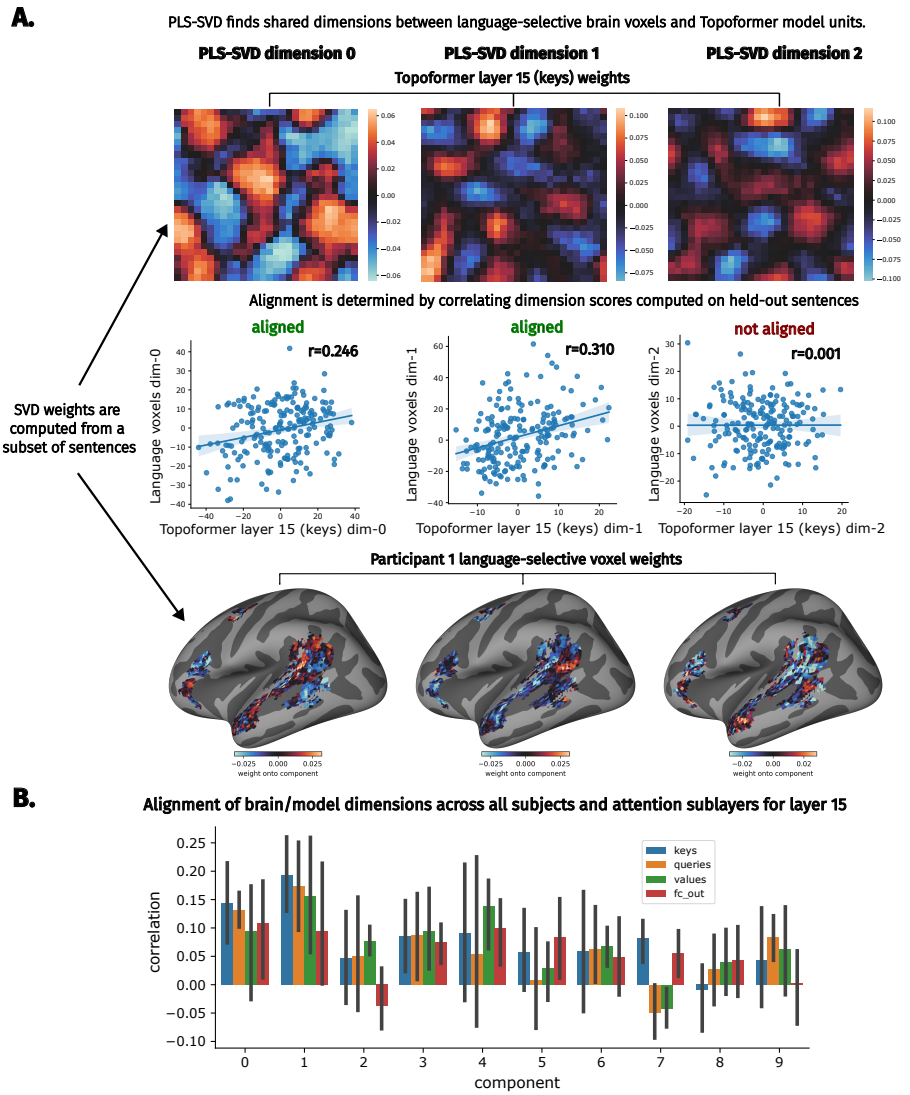


Figure 5: Alignment of topographic representations in the human language network and Topoformer-BERT model. A. Illustration of the PLS-SVD alignment approach for a single participant and model sublayer representation. **B.** Alignment quantified across all 10 components, and each sublayer of Topoformer-BERT layer 15. The alignment of components is computed as the correlation of respective cross-validated PLS-SVD component scores for brain and model representations. Error bars are 95% confidence intervals over 5 participants.

unsmoothed brain responses within the functionally-defined language network of each participant, splitting the network into 5 spatial subregions (see Appendix A.4). We compared this statistic to a null distribution, using shuffled brain responses, finding that the t_g value for each cluster fell outside this null distribution, indicating significant decay in unit response correlations with distance.

To determine whether the topography of the human language network is linguistically meaningful and corresponding to that of Topoformer-BERT, we performed representational alignment using partial least squares singular value decomposition (PLS-SVD). Given z-scored brain responses X and model embeddings Y, PLS-SVD finds joint low-dimensional embeddings X_c and Y_c by computing the SVD on the covariance matrix X^TY as $X^TY = U\Sigma V$ such that the left singular vectors $U = W_x$ are the component weights from brain responses and the right singular vectors $V^T = W_y$ are the component weights from model embeddings. The component scores are then given as $X_c = XW_x$ and $Y_c = YW_y$, where $X_c^{(i)}$ and $Y_c^{(i)}$ are the i-th aligned component scores.

Given the spatial organization of both brain and Topoformer responses, we can visualize the SVD weights of individual brain and model components $W_x^{(i)}$ and $W_y^{(i)}$, respectively, reshaped into their native spatial format. To compute the alignment of components, we perform a cross-validated analysis that ensures generalization, where SVD is computed using 80% of the sentences, and the scores are computed for the remaining 20% of the data. These scores can then be correlated across brain and model, for each dimension, to determine their alignment.

Figure 5A plots example alignments between the first three brain and model components, using the first participant and the Topoformer-BERT layer 15 (final layer, zero-indexed) keys representation. We see that the first two components are strongly aligned, as well as strongly topographically organized in both model and brain spaces. The third component is not aligned, despite being spatially organized in each representational space. Figure 5B repeats this analysis for all participants and sublayers, using layer 15 again. In general, the first two components were significantly aligned for each sublayer, whereas later components were less likely to be aligned. This result demonstrates that the low-dimensional variability can be aligned in the topographic representations of the human language network and Topoformer language model. The fact that we used functionally-defined language regions suggests that there is spatial functional organization even *within* this relatively functionally homogeneous brain network (e.g., Blank & Fedorenko (2020); Fedorenko et al. (2020)), rather than simply across different functional networks with heterogeneous response profiles, similar to the organization that emerges in the Topoformer model.

To determine the specificity of this alignment, we performed an identical analysis using a control network and untrained Topoformer-BERT variant (Appendix A.6). Alignment, as well as voxel encoding model prediction, was significantly greater between the trained Topoformer-BERT and language network compared to a non-language control network and an untrained model, highlighting the linguistic nature of the alignment.

4 Discussion

Here, we introduced the first topographically organized Transformer language models, "Topoformers". Across small and large models, we found that these spatial querying and reweighting operations produced topographic organization in Topoformer models trained on natural language processing tasks. This organization was revealed with specific hypotheses by contrasting different linguistic properties, as well as generically through PCA. Finally, analyzing brain responses to a large number of sentences in the human language network, we uncovered topographic variability with low-dimensional alignment to that found in the Topoformer-BERT model.

Introducing topography into language models may improve interpretability in NLP. We took some initial steps with our suite of tests, but the interpretability problem is far from solved. One issue is that of "polysemanticity," whereby units are involved in the representation of several distinct concepts (Bricken et al., 2023). Despite strong semantic selectivity, we found that Topoformer-BERT's activations were highly overlapping across categories, similar to non-topographic models. While our 2D visualizations aided interpretability of selectivity, efforts to improve "monosemanticity" or to encourage disentangled representations (Higgins et al., 2021) may prove fruitful in yielding even more interpretable topographic organizaton when combined with the Topoformer motifs. Additionally, topographically constrained sparse autoencoders might allow for greater interpretability of entangled representations in Topoformers (Cunningham et al., 2023; Bricken et al., 2023).

Introducing topography is also necessary to imrpove the biological realism of models of language processing in the human cortex, and understand how biological constraints (e.g. wiring cost) shape the emergence and organization of the language network. Future work should aim to scale the approach towards foundation-model level. One critical insight is that scale will not only improve the performance of these models, but also improve their brain predictivity (Schrimpf et al., 2021). In parallel, a greater focus on biological plausibility may prove fruitful for basic neuroscientific investigation of the language network (Jain et al., 2023).

This work marks the beginning of topographic modeling of language processing. We hope that other researchers will be persuaded to embrace topography in language models, and push the development and use of Topoformers along several new directions.

REFERENCES

John Ashburner and Karl J. Friston. Unified segmentation. *NeuroImage*, 26:839–851, 2005. doi: 10.1016/j.neuroimage.2005.02.018.

Pinglei Bao, Liang She, Mason Mcgill, and Doris Y. Tsao. A map of object space in primate inferotemporal cortex. *Nature*, (January 2019), 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2350-5.

J. R. Binder, C. F. Westbury, K. A. McKiernan, E. T. Possing, and D. A. Medler. Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, 17(6): 905–917, June 2005. ISSN 0898-929X. doi: 10.1162/0898929054021102.

Idan A. Blank and Evelina Fedorenko. No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, 219:116925, October 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.116925. URL https://www.sciencedirect.com/science/article/pii/S1053811920304110.

Nicholas M Blauch, Marlene Behrmann, and David C Plaut. 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), jan 2022. ISSN 0027-8424. doi: 10.1073/pnas.2112566119. URL http://www.pnas.org/lookup/doi/10.1073/pnas.2112566119.

Rodrigo M. Braga, Lauren M. DiNicola, Hannah C. Becker, and Randy L. Buckner. Situating the left-lateralized language network in the broader organization of multiple specialized large-scale distributed networks. *Journal of Neurophysiology*, 124(5):1415–1448, November 2020. ISSN 1522-1598. doi: 10.1152/jn.00753.2019.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*, July 2020. URL http://arxiv.org/abs/2005.14165. arXiv: 2005.14165.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):134, December 2022. ISSN 2399-3642. doi: 10.1038/s42003-022-03036-1. URL https://www.nature.com/articles/s42003-022-03036-1.

Andrew C Connolly, J Swaroop Guntupalli, Jason Gors, Michael Hanke, Yaroslav O Halchenko, Yu-Chien Wu, Hervé Abdi, and James V Haxby. The representation of biological classes in the human brain. *Journal of Neuroscience*, 32(8):2608–2618, 2012.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

Rutvik H Desai, Jeffrey R Binder, Lisa L Conant, and Mark S Seidenberg. Activation of sensory–motor areas in sentence comprehension. *Cerebral Cortex*, 20(2):468–478, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, June 2019.

Fenil Doshi and Talia Konkle. Organizational motifs of cortical responses to objects emerge in topographic projections of deep neural networks. *Journal of Vision*, 21(9):2226, September 2021. ISSN 1534-7362. doi: 10.1167/jov.21.9.2226.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194, aug 2010. doi: 10.1152/jn.00032.2010. URL http://dx.doi.org/10.1152/jn.00032.2010.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203: 104348, October 2020. ISSN 1873-7838. doi: 10.1016/j.cognition.2020.104348.

Christian J. Fiebach and Angela D. Friederici. Processing concrete words: fMRI evidence against a specific right-hemisphere involvement. *Neuropsychologia*, 42(1):62–70, January 2004. ISSN 0028-3932. doi: 10.1016/S0028-3932(03)00145-3. URL https://www.sciencedirect.com/science/article/pii/S0028393203001453.

Dawn Finzi, Jesse Gomez, Marisa Nordt, Alex A Rezai, Sonia Poltoratski, and Kalanit Grill-Spector. Differential spatial computations in ventral and lateral face-selective regions are scaffolded by structural connections. *Nature Communications*, 12(1):2278, apr 2021. doi: 10.1038/s41467-021-22524-2. URL http://dx.doi.org/10.1038/s41467-021-22524-2.

Jon Gauthier, Jennifer Hu, Ethan Gotlieb Wilcox, Peng Qian, and Roger Philip Levy. Syntaxgym: An online platform for targeted evaluation of language models. In *Annual Meeting of the Association for Computational Linguistics*, 2020. URL https://api.semanticscholar.org/CorpusID:220060899.

Jonas Geiping and Tom Goldstein. Cramming: Training a language model on a single gpu in one day, 2022.

Matthew F. Glasser, Timothy S. Coalson, Emma C. Robinson, Carl D. Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F. Beckmann, Mark Jenkinson, Stephen M. Smith, and David C. Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, August 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature18933. URL http://www.nature.com/articles/nature18933.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, March 2022a. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-022-01026-4. URL https://www.nature.com/articles/s41593-022-01026-4.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Rose Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner K. Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Y. Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25:369 – 380, 2022b.

Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, 12(1):6456, nov 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26751-5. URL https://www.nature.com/articles/s41467-021-26751-5.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy. A Systematic Assessment of Syntactic Generalization in Neural Language Models, May 2020. URL http://arxiv.org/abs/2005.03692. arXiv:2005.03692 [cs].

Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, dec 2012. doi: 10.1016/j.neuron.2012.10.014. URL http://dx.doi.org/10.1016/j.neuron.2012.10.014.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, apr 2016. doi: 10.1038/nature17637. URL http://dx.doi.org/10.1038/nature17637.

R A Jacobs and M I Jordan. Computational consequences of a bias toward short connections. *Journal of Cognitive Neuroscience*, 4(4):323–336, 1992. doi: 10.1162/jocn.1992.4.4.323. URL http://dx.doi.org/10.1162/jocn.1992.4.4.323.

Shailee Jain, Vy A. Vo, Leila Wehbe, and Alexander G. Huth. Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, pp. 1–65, jan 2023. ISSN 2641-4368. doi: 10.1162/nol_a_00101. URL https://direct.mit.edu/nol/article/doi/10.1162/nol_a_00101/114613/Computational-language-modeling-and-the-promise-of.

Joshua B. Julian, Evelina Fedorenko, Jason Webster, and Nancy G. Kanwisher. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage*, 60:2357–2364, 2012. URL https://api.semanticscholar.org/CorpusID:18168257.

Nancy Kanwisher. Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences of the United States of America*, 107(25):11163–11170, 2010. ISSN 1091-6490. doi: 10.1073/pnas.1005062107.

T. Anderson Keller and Max Welling. Topographic VAEs learn equivariant capsules. *arXiv*, sep 2021. URL https://arxiv.org/abs/2109.01394.

Roozbeh Kiani, Hossein Esteky, Koorosh Mirpour, and Keiji Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, 97(6):4296–4309, jun 2007. doi: 10.1152/jn.00024.2007. URL http://dx.doi.org/10.1152/jn.00024.2007.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982. ISSN 0340-1200. doi: 10.1007/{BF00337288}. URL http://link.springer.com/10.1007/{BF00337288}.

Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *The Journal of Neuroscience*, 33(25):10235–10242, jun 2013. doi: 10.1523/{JNEUROSCI}.0983-13.2013. URL http://dx.doi.org/10.1523/{JNEUROSCI}.0983-13.2013.

Talia Konkle and Aude Oliva. A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, 74(6):1114–1124, June 2012. ISSN 08966273. doi: 10.1016/j.neuron.2012.04.036.

Sarah Laszlo and David C Plaut. A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, 120(3):271–281, mar 2012. ISSN 1090-2155. doi: 10.1016/j.bandl.2011.09.001. URL http://dx.doi.org/10.1016/j.bandl.2011.09.001.

Hyodong Lee, Eshed Margalit, Kamila M. Jozwik, Michael A. Cohen, Nancy Kanwisher, Daniel L. K. Yamins, and James J. DiCarlo. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. Preprint, Neuroscience, July 2020a.

Hyodong Lee, Eshed Margalit, Kamila M. Jozwik, Michael A. Cohen, Nancy Kanwisher, Daniel L. K. Yamins, and James J. DiCarlo. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *BioRxiv*, jul 2020b. doi: 10.1101/2020.07.09.185116. URL http://biorxiv.org/lookup/doi/10.1101/2020.07.09.185116.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope H. Kean, Olessia Jouravlev, Lara I. Rakocevic, Brianna Pritchett, Matthew Siegelman, Caitlyn Hoeflin, Alvince L. Pongos, Idan Asher Blank, Melissa Kline Struhl, Anna A. Ivanova, Steven Shannon, Aalok Sathe, Malte Hoffmann, Alfonso Nieto-Castanon, and Evelina Fedorenko. Probabilistic atlas for the language network based on precision fmri data from ¿800 individuals. *Scientific Data*, 9, 2022.

Zejin Lu, Adrien Doerig, Victoria Bosch, Bas Krahmer, Daniel Kaiser, Radoslaw M. Cichy, and Tim C. Kietzmann. End-to-end topographic networks as models of cortical map formation and human visual behavior: Moving beyond convolutions. *arXiv preprint arXiv:2308.09431*, August 2023.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, 52:1271–1291, 2020.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015.

Kyle Mahowald and Evelina Fedorenko. Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *NeuroImage*, 139:74–93, 2016. URL https://api.semanticscholar.org/CorpusID:5171423.

Eshed Margalit, Hyodong Lee, Dawn Finzi, James J DiCarlo, Kalanit Grill-Spector, and Daniel L. K. Yamins. A unifying principle for the functional organization of visual cortex. *bioRxiv*, pp. 2023.05.18.541361, 2023.

Daniel S Margulies, Satrajit S Ghosh, Alexandros Goulas, Marcel Falkiewicz, Julia M Huntenburg, Georg Langs, Gleb Bezgin, Simon B Eickhoff, F Xavier Castellanos, Michael Petrides, et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 113(44):12574–12579, 2016.

D. Merkx and S. L. Frank. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 12–22, 2021. doi: 10.18653/v1/2021.cmcl-1.2.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2928–2949, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.213.

Thomas Naselaris, Ryan J Prenger, Kendrick N Kay, Michael Oliver, and Jack L Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–915, 2009.

Alfonso Nieto-Castañón. *Handbook of Functional Connectivity Magnetic Resonance Imaging Methods in CONN*. Hilbert Press, February 2020. ISBN 978-0-578-64400-4. doi: 10.56441/hilbertpress.2207.6598.

OpenAI. GPT-4 Technical Report, March 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, and Christophe Pallier. Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps, July 2022. URL http://arxiv.org/abs/2207.03380. arXiv:2207.03380 [cs].

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jacob S. Prince, Ian Charest, Jan W. Kurzawski, John A. Pyles, Michael J. Tarr, and Kendrick Norris Kay. Improving the accuracy of single-trial fmri response estimates using glmsingle. *eLife*, 11, 2022.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018. URL https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, November 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2105646118. URL http://www.pnas.org/lookup/doi/10.1073/pnas.2105646118.

C. Shain, C. Meister, T. Pimentel, R. Cotterell, and R. P. Levy. Large-scale evidence for logarithmic effects of word predictability on reading time. *Preprint*, 2022. URL https://doi.org/10.31234/osf.io/4hyna.

Luke Tait, Ayşegül Özkan, Maciej J Szul, and Jiaxiang Zhang. A systematic evaluation of source reconstruction of resting meg of the human brain with a new high-resolution atlas: Performance, precision, and parcellation. *Human Brain Mapping*, 42(14):4685–4707, 2021.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26:858–866, 2023. doi: 10.1038/s41593-023-01007-x. URL https://www.nature.com/articles/s41593-023-01007-x.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. Driving and suppressing the human language network using large language models, May 2023. URL https://www.biorxiv.org/content/10.1101/2023.04.16.537080v3. Pages: 2023.04.16.537080 Section: New Results.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pp. 5999–6009, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, July 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00321. URL https://doi.org/10.1162/tacl_a_00321.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and R. Levy. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. *ArXiv*, June 2020. URL https://www.semanticscholar.org/paper/On-the-Predictive-Power-of-Neural-Language-Models-Wilcox-Gauthier/fccfc6839777fe6f06197548dbe4bacb48b1a14b.

BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books, June 2015. URL http://arxiv.org/abs/1506.06724. arXiv:1506.06724 [cs].

A APPENDIX

A.1 FEASIBILITY MODEL METHODOLOGY (TOPOFORMER-SQ AND TOPOFORMER-SQR)

A.1.1 TRAINING PARADIGM

We trained a 1-layer encoder-only Transformer on the IMDB sentiment analysis dataset Maas et al. (2011), which classifies movie reviews as having a positive or negative overall sentiment. We utilized the average of all tokens as the input to a binary classifier, as is standard practice, and optimized a cross-entropy loss. We trained for 20 epochs, which was sufficient to reach convergence. We used an Adam optimizer (Kingma & Ba, 2017) with a learning rate of 0.001. A batch size of 128 was used for the SQ model, while a batch size of 256 was used for the SQR model (SQR was less stable and required the larger batch size; similar results were seen across batch sizes for SQ).

A.1.2 SELECTIVITY ANALYSES

After completing the training process, we conducted a selectivity analysis which is ubiquitously used in neuroscience. We probed the Q, K, and V layers to gain insights into their topographic organization and how the model attends to different aspects of the input data. By analyzing the responses of each domain (i.e., sentences with positive and negative sentiment) versus the others using a two-tailed t-test, we obtained the test statistic t, the significance value p of the test, and the sign of the test statistic $s = \text{sign}(t)$. We then computed the selectivity as $-s\log_{10}(p)$, which provided a quantification of our model's selectivity for each domain.

A.2 LARGE-SCALE MODEL METHODOLOGY (TOPOFORMER-BERT)

A.2.1 TRAINING METHODOLOGY

We used a batch size of 4096, and Adam optimizer (Kingma & Ba, 2017) with weight decay of 0.01, $\epsilon = 10^{-12}$, $\beta_1 = 0.9$ $\beta_2 = 0.98$, gradient clipping at a clip value of 0.5.

A.2.2 SELECTIVITY ANALYSES

Since the large-scale models were not trained on a categorization task, we had to devise a particular hypothesis of linguistic relevance for the large-scale Topoformers. To accomplish this, we investigated eight test suites that each target different properties of linguistic input (see Section 3.2 and Appendix 2). To obtain the decoding accuracy scores showcased in Figure 4, we employed Principal Component Analysis (PCA) as a dimensionality reduction technique on the activations of the sub-representations' keys. We then utilized 50 principal components to train a logistic regression classifier for decoding category types. For instance, in the case of **Animacy**, our objective was to predict whether the activations originated from an animate or inanimate sentence. The dataset underwent an 80-20 split for training and testing, respectively.

A.3 GENERAL ANALYSIS METHODOLOGY

A.3.1 PRINCIPAL COMPONENT ANALYSIS

In addition to selectivity analyses, for all models, as well as brains, we aimed to uncover generic patterns of activation in the layers of our model without any specific hypothesis in mind. To this end, we performed principal component analysis (PCA) on the activations of each layer individually. By analyzing the PCs, we were able to identify the dominant modes of variation in the activation data, and gain deeper insights into the structure of the activations. To visualize these patterns of activation, we reshaped the weights of the first two PCs to the same size as our cortical map. This allowed us to compare the patterns of activation across different layers and gain a deeper understanding of the topographic organization of our model.

A.3.2 QUANTIFICATION OF TOPOGRAPHY

We compute the generic topographic statistic t_g as a measure of the general distance dependence of pairwise response correlations, based on the notion that local correlation is a hallmark of topographic

organization (Kiani et al., 2007; Lee et al., 2020b). Given the Pearson correlation matrix $R_{i,j} = r_p(\boldsymbol{a}_i, \boldsymbol{a}_j)$ — where $\boldsymbol{a}_i$ gives the activity vector of unit i over sentences, $r_p(\boldsymbol{x}, \boldsymbol{y})$ gives the Pearson correlation of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $r_s(\boldsymbol{x}, \boldsymbol{y})$ gives the Spearman rank correlation of $\boldsymbol{x}$ and $\boldsymbol{y}$ — and the pairwise Euclidean distances $\mathcal{D}_{i,j}$ computed in volumetric space, along with a maximum distance over which to compute the statistic d, we compute the generic statistic as follows:

$$t_{g,d} = r_s(-R_{i,j}, \mathcal{D}_{i,j}) \ \ \forall \ \ i, j : D_{i,j} < d \tag{4}$$

Because topographic organization can exist at multiple scales, with long-range correlation violating the locally distance-decaying correlation, we may wish to compute $t_{g,d}$ at a range of maximum distances, as in Figure 3. Thus, for those analyses, we computed a vector $\boldsymbol{t}_g = \{t_{g,d_0}, ..., t_{g,d_n}\}$ over a linearly spaced range maximum distance values $\boldsymbol{d}$. We summarized this vector by taking the mean $\bar{t}_g = \frac{1}{n}\sum_i \boldsymbol{t}_{\boldsymbol{g}}^{\boldsymbol{i}}$ (Figure 3A), and plotting the vector against $\boldsymbol{d}$ (Figure 3B). When no maximum distance is used, we refer to the statistic as simply t_g, as used in the brain analyses.

A.4 Human Brain Data

A.4.1 Participants and Acquisition

We recorded brain responses using fMRI from N=5 participants during a sentence reading task. The participants were neurotypical native speakers of English (4 female), aged 21 to 30 (mean 25; std 3.5), all right-handed. Participants read 1,000 6-word, corpus-extracted sentences that were selected to maximize semantic and stylistic diversity. Participants completed two scanning sessions where each session consisted of 10 runs of the sentence reading experiment (sentences presented on the screen one at a time for 2s with an inter-stimulus interval of 4s, 50 sentences per run) along with additional tasks. Participants were exposed to the same set of 1,000 sentences (no repetitions), but in fully randomized order. Structural and functional data were collected on the whole-body, 3 Tesla, Siemens Prisma scanner with a 32-channel head coil. T1-weighted, Magnetization Prepared RApid Gradient Echo (MP-RAGE) structural images were collected in 176 sagittal slices with 1 mm isotropic voxels (TR = 2,530 ms, TE = 3.48 ms, TI = 1100 ms, flip = 8 degrees). Functional, blood oxygenation level dependent (BOLD) were acquired using an SMS EPI sequence (with a 90 degree flip angle and using a slice acceleration factor of 2), with the following acquisition parameters: fifty-two 2 mm thick near-axial slices acquired in the interleaved order (with 10% distance factor) 2 mm × 2 mm in-plane resolution, FoV in the phase encoding (A $\ll$ P) direction 208 mm and matrix size 104 × 104, TR = 2,000 ms and TE = 30 ms, and partial Fourier of 7/8. All participants gave informed written consent in accordance with the requirements of an institutional review board.

A.4.2 Data Preprocessing and First-Level Modeling

fMRI data were preprocessed using SPM12 (release 7487), and custom CONN/MATLAB scripts. Each participant's functional and structural data were converted from DICOM to NIfTI format. All functional scans were coregistered and resampled using B-spline interpolation to the first scan of the first session. Potential outlier scans were identified from the resulting participant-motion estimates as well as from BOLD signal indicators using default thresholds in CONN preprocessing pipeline (5 standard deviations above the mean in global BOLD signal change, or framewise displacement values above 0.9 mm; Nieto-Castañón (2020)). Functional and structural data were independently normalized into a common space (the Montreal Neurological Institute [MNI] template; IXI549Space) using SPM12 unified segmentation and normalization procedure Ashburner & Friston (2005) with a reference functional image computed as the mean functional data after realignment across all timepoints omitting outlier scans. The output data were resampled to a common bounding box between MNI-space coordinates (-90, -126, -72) and (90, 90, 108), using 2 mm isotropic voxels and 4th order spline interpolation for the functional data, and 1 mm isotropic voxels and trilinear interpolation for the structural data. Last, the functional data were smoothed spatially using spatial convolution with a 4 mm FWHM Gaussian kernel or no smoothing (we investigated both to ensure that are claims about topographic organization could not be explained by smoothing). A General Linear Model (GLM) was used to estimate the beta weights that represent the blood oxygenation level dependent (BOLD) response amplitude evoked by each individual sentence trial using GLMsingle Prince et al. (2022) (fixation was modeled implicitly, such that all timepoints that did not correspond to one of the conditions (sentences) were assumed to correspond to a fixation period). Within the GLMsingle framework, the HRF which provided the best fit to the data was identified for each voxel (based on

the amount of variance explained). Data were modeled using 5 noise regressors and a ridge regression fraction of 0.05. The 'sessionindicator' option in GLMsingle was used to specify how different input runs were grouped into sessions. By default, GLMsingle returns beta weights in units of percent signal change by dividing by the mean signal intensity observed at each voxel and multiplying by 100. Hence, the beta weight for each voxel can be interpreted as a change in BOLD signal for a given sentence trial relative to the fixation baseline. After first-level modeling, the voxels within a set of 5 masks ("parcels") were extracted. These parcels were derived from n=220 independent participants using a Group-Constrained Subject-Specific (GSS; Julian et al. (2012) based on an extensively validated language localizer contrast between reading of sentences and non-word strings Fedorenko et al. (2010); Mahowald & Fedorenko (2016); Lipkin et al. (2022). These parcels delineate the expected gross locations of language-selective brain regions but are sufficiently large to encompass individual variability. The parcels are in the left hemisphere, three frontal parcels (inferior frontal gyrus [IFG], its orbital portion [IFGorb], and middle frontal gyrus [MFG]) and two temporal ones (anterior temporal [AntTemp], posterior temporal [PostTemp]). The mean number of voxels in these five parcels were (they differ slightly across participants because of lack of spatial coverage in the functional acquisition sequence): IFG=743 (SD=0); IFGorb=364 (SD=13.4); MFG=462 (SD=0); AntTemp=1623 (SD=4.1); PostTemp=2948 (SD=0). The parcels are available for download at **ANONYMIZED**. As a control brain region, we extracted voxels within a set of motor- and supplementary motor areas in the left hemisphere. Specifically, we used the Glasser parcellation Glasser et al. (2016) to extract responses within 5 motor parcels: 1, 2, 3a, 3b, and 4. Moreover, we identified a set of supplementary regions using the grouping category "Paracentral lobular and mid-cingulate cortex" in Tait et al. (2021) which consisted of 8 additional parcels: 24dd, 24dv, 6mp, 6ma, SCEF, 5m, 5L, and 5mv, yielding a total of 13 Glasser parcels of interest. The mean number of voxels in these 13 parcels were 3753 (SD=363.2).

A.5 QUANTIFYING BRAIN TOPOGRAPHY

Here, we quantified the spatial smoothness of brain representations using the generic topographic statistic t_g, in each language ROI and a control ROI (see A.4 for detail on ROI definition). Due to differences in sizes across ROIs, we use the version computed without a maximum distance, over the full range of voxel pairs in each ROI. Statistics are shown in Figure 6A. To determine the significance of these statistic values, which were low in some cases, a null permutation analysis was performed by shuffling the voxel responses of each region 100 times with respect to their positions, and computing t_g. This allowed us to construct a null distribution against which to compare real t_g values. As can be seen in Figure 6, the variability of this null distribution is very small. All of the brain data fell outside the 95th percentile, indicating that each brain region — including the control region — exhibited significant spatial smoothness in responses to natural sentences.

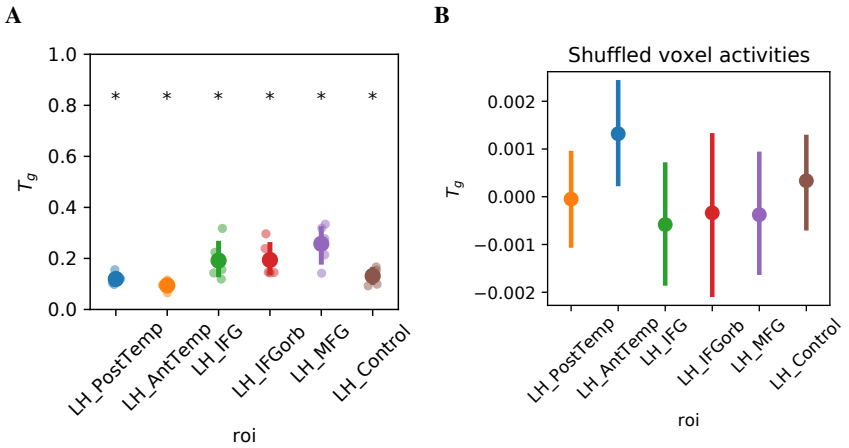
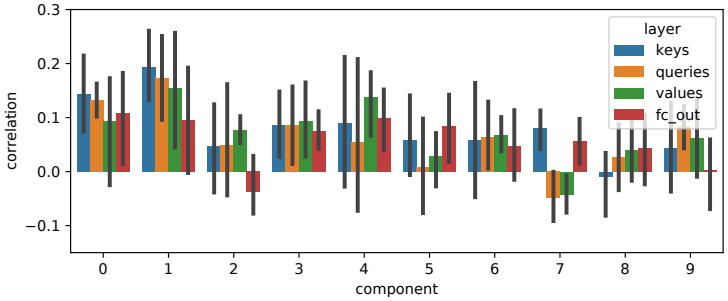


Figure 6: Generic topographic statistic of unsmoothed brain responses. **A.** Generic topographic statistic (t_g) for each ROI of the language network, and a control ROI. Error bars plot 95% confidence intervals. Stars indicate significance compared to the null distribution shown **B.**, using $alpha = 0.05$. **B.** Null distributions of t_g computed for each ROI using shuffled voxel locations with 100 permutations per participant and ROI. Error bars plot 95% confidence intervals.

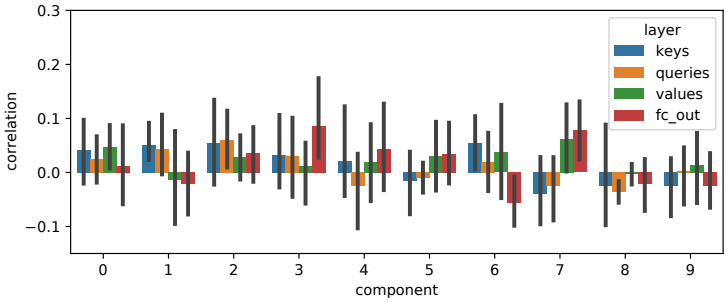
A.6 BRAIN-MODEL CONTROL ANALYSES

In the main text, we performed a PLS-SVD analysis to assess the alignment of human language network and model components. Here, we repeated this analysis considering two important controls: i) motor-related control brain regions, and ii) untrained Topoformer-BERT. First, we confirmed that the alignment with Topoformer-BERT (Figure 7A) is stronger in the language network than in a control brain network (motor-related regions, see Appendix A.4 for details; Figure 7B). Particularly, whereas significant alignment was seen for the first two components in the language network, this was not true for most sublayers when comparing to the control brain network (exception: weak significant alignment of values with component 0, and keys with component 1) with substantially reduced alignment overall. Second, we analyzed an untrained version of the Topoformer-BERT model. We found that this untrained model demonstrated some alignment with the first (0th) component, however this was highly variable across participants and not statistically significant (Figure 7C.). When comparing to the control brain network, the untrained model showed very little alignment (Figure 7D). Thus, the alignment was particularly strong between the trained Topoformer-BERT model and language network.

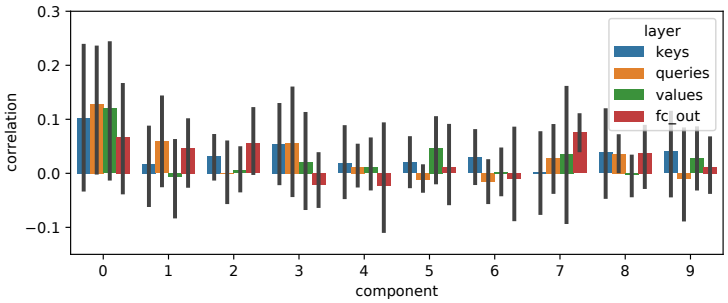
A Trained Topoformer-BERT, language network



B Trained Topoformer-BERT, control network



C Untrained Topoformer-BERT, language network



D Untrained Topoformer-BERT, control network

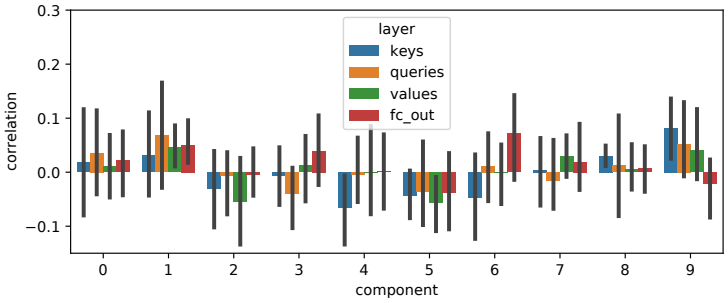


Figure 7: PLS-SVD alignment results across two control analyses. We follow the same analysis approach used in Figure 5, for each combination of trained/untrained Topoformer, and language/control networks. Error bars show 95% CI over all participants' voxels.

A.6.1 ENCODING MODEL ANALYSES

Since our PLS-SVD alignment analysis is less common for comparing representations in the neuroscience literature than other analyses, we opted to perform an additional encoding model analysis for greater comparability to prior work.

The encoding model analysis approach is used to predict a given voxel's response based on a representational space. We employed ridge regression, or L2-regularized least squares regression. Formally, let us consider the model embedding space $X \in \mathbb{R}^{n \times d}$, where each of n rows is a d-dimensional vector consisting of the model's representation of a given sentence. A given voxel's responses are given as $y \in \mathbb{R}^n$. For each voxel, we can formulate our regularized regression problem as finding a vector $\hat{w} \in \mathbb{R}^d$ such that

$$\hat{w} = \arg\min_{w} \|Xw - y\|_2^2 + \lambda \|w\|_2^2, \tag{5}$$

where $\| \cdot \|_2$ is the Euclidean norm. The λ multiplier is a hyperparameter that specifies the relative weight of the regularization term in the loss. This value is chosen using leave-one-sentence-out cross-validation using 80% of the total data, as summarized below

$$\hat{\lambda} = \arg\min_{\lambda} \frac{1}{n} \sum_{k=1}^{.8*n} \left[\|X^{(k)} \hat{w} - y^{(k)}\|_2^2 + \lambda \|\hat{w}\|_2^2 \right], \tag{6}$$

where $X^{(k)}$ and $y^{(k)}$ are held out cross-validation data at the k-th fold. This cross-validation is performed efficiently using the scikit-learn function `RidgeCV` (Pedregosa et al., 2011). The remaining 20% of the data is used to test the generalization of the encoding model, by predicting held-out voxel responses. We report the correlation of held-out voxel responses with predictions, and take the average over all voxels from all participants.

We perform this encoding model analysis for both trained and untrained Topoformer-BERT models, for voxels in both language and motor-related, non-linguistic control brain networks. The results are seen in Figure 8. We see that the trained Topoformer provides superior neural fits, and that language network voxels are predicted better than the control network. This confirms the previous results found using the PLS-SVD alignment approach. We note that the relatively low performance of the encoding model is partly attributed to the fact that we included a large subset of voxels. The results obtained via PLS-SVD in Figures 5 and 7 indicate that the shared variability *across* voxels aids the alignment between brain and model. Finally, we note that although the untrained Topoformer-BERT had lower predictivity performance than the trained counterpart, it was still above zero, in line with prior findings Schrimpf et al. (2021); Caucheteux & King (2022); Pasquiou et al. (2022).

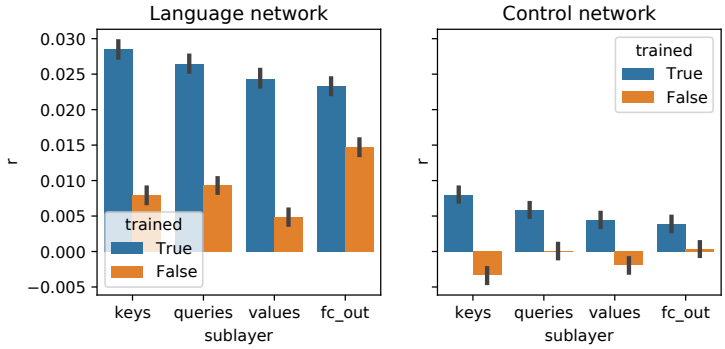


Figure 8: Encoding model prediction results across two control analyses: Comparison between the language network versus a motor-related control brain network (motor-related brain regions, and comparison of the trained versus untrained Topoformer-BERT model.

A.7 LACK OF TOPOGRAPHIC ORGANIZATION IN THE CONTROL MODEL

We trained a single-head BERT model using the same dataset and training procedure as employed for the **Topoformer-BERT**. Notably, the control model was trained with the intentional exclusion of spatial querying or Reweighing mechanisms. All other aspects of the model, including architecture and training parameters, remained consistent with those of the Topoformer-BERT. Subsequently, we followed the outlined procedure depicted in Figure 2 to extract activations corresponding to sublayer representations from the control model.

As expected, the outputs from this non-topographic control model exhibited a lack of discernible topographic organization. This deficiency arises due to the absence of spatial querying or reweighting mechanisms within the model architecture.

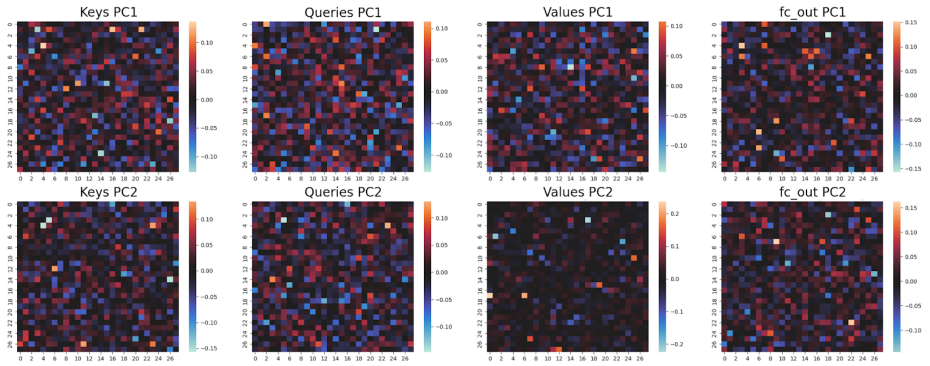


Figure 9: Lack of topographic organization in the control model. Here, we can see that none of the (keys, queries, values, and fc_out) were topographically organized.

A.8 TEST SUITES

We evaluated a set of eight test suites targeting different linguistic properties. An overview of these suites along with examples can be seen in Table 2.

Test Suite	Category	Example
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.
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.

All eight test suites consisted of 76 sentences each, with 38 sentences in each category.

The first suite, **Intactness** consisted of intact sentences versus their scrambled counterparts, thereby degrading both linguistic meaning (syntax) and meaning (semantics). Capitalization and final sentence punctuation was retained in the scrambled sentences. Suites 2 through 4 evaluated three different dimensions of *meaning* that have been extensively investigated in prior work, as specified next. Suite 2, **Animacy**, consisted of sentences with animate vs. inanimate meanings Naselaris et al. (2009); Connolly et al. (2012); Konkle & Caramazza (2013). We sampled 76 animate/inanimate word categories from Konkle & Caramazza (2013). Specifically, 38 animate words were randomly sampled from the "Big-Animate" and "Small-Animate" categories (from 120 words in total), and 38 inanimate words were randomly sampled from the "Big-Inanimate", and "Small-Inanimate" categories (from 120 words in total). ChatGPT (version 4) was prompted to generate sentences about each of these words, approximately 10 words long. Suite 3, **Concreteness**, consisted of sentences with concrete vs. abstract meanings Binder et al. (2005); Fiebach & Friederici (2004). We randomly sampled 38 concrete and 38 abstract word categories from Binder et al. (2005) (from 50 words in total in each category), and similarly prompted ChatGPT to generate 10-word sentences about each of these words. Suite 4, **Visuomotor**, consisted of sentences with visual vs. motor meanings Desai et al. (2010); Lynott et al. (2020). We took all available visual and motor verbs from Desai et al. (2010) (23 in each category). For the remaining 15 words for each category (to obtain the standard number of 38 sentences in each category), we sampled words from the Lancaster Sensorimotor Norms Lynott et al. (2020). Specifically, for the visual category, we selected the 15 top-rated "Visual" words. For the motor category, we averaged across the "Foot leg", "Hand arm", "Head", "Mouth", and "Torso" ratings and selected the 15-top rated words excluding inappropriate words and different forms of the same word stem. We similarly prompted ChatGPT to generate 10-word sentences about each of these words.

For all the three semantic test suites (2, 3, 4) we matched the number of words between categories, and tested that the surprisal of the sentences in each category (within a suite) were not significantly different from each other to avoid a confound of overall sentence surprisal (evaluated by two-sided, unpaired t-tests; $p > 0.11$).

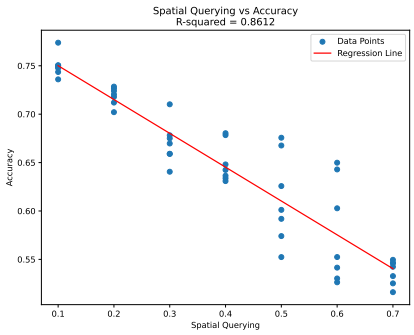
Suite 5, **Semantic acceptability** consisted of minimal pair sentences (Conceptual Minimal Pair Sentences Base; Misra et al. 2023) with 38 acceptable sentences and 38 unacceptable sentences. The remaining three suites (6, 7, 8) evaluated three different dimensions of *form* using suites from SyntaxGym Gauthier et al. (2020); Hu et al. (2020): Suite 6 consisted of matched/mismatched **Agreement** sentences (Subject-Verb Number Agreement; `https://syntaxgym.org/test_suite/items?test_suite=261`), suite 7 consisted of matched/mismatched **Licensing** sentences (Reflexive Number Agreement; `https://syntaxgym.org/test_suite/items?test_suite=260`), and suite 8 consisted of **Garden-Path** ambiguous sentences (Verb Transitivity; `https://syntaxgym.org/test_suite/items?test_suite=270`).

A.9 Receptive field (RF) size analysis

A.9.1 Effect on performance

In this section, we discuss our methodology for choosing a receptive field (RF) size. First, we train the 1-layer (**Topoformer-SQR**) across a sweep of different RF sizes for both spatial querying (r_{SQ}) and spatial reweighting (r_{SR}). As seen in 10, we can see a clear inverse relationship between the r_{SQ} ($R^2 = 0.8612$), in line with the idea that larger spatial querying pools reduce the amount of information due to the local averaging of queries within the pool. However, we found that the size of spatial reweighting r_{SR} has no measurable effect on model performance r_{SR} ($R^2 = 0.0137$). This deviates from the result for spatial querying because there is no averaging performed in spatial reweighting. Future work could explore a weighted version of spatial querying that would be expected to show less of a decrement in performance with increasing RF size. Importantly, substantial topographic organization can be seen with a very small SQ value, such that performance decrements are minimal.

A Spatial querying



B Spatial reweighting

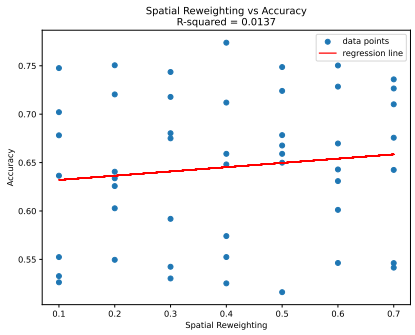
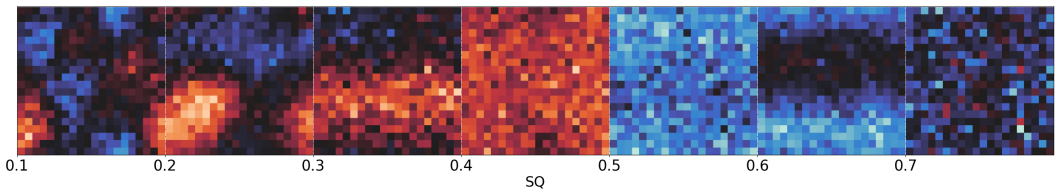


Figure 10: RF size effect on accuracy for the IMDB sentiment test set

A.9.2 Qualitative evaluation

Using the models from the previous section, we plotted a matrix grid containing the weights of the second PC across all RF values. For brevity, we visualized the the keys and fc_out sublayer representations using the same procedure mentioned in Figure 2, so as to determine the effects of RF size on both spatial querying and reweighting. As seen in Figure 11, having smaller r_SQ values results in much stronger topographic organization in the keys, in line with the better performance; larger values of r_{SQ} were associated with poor performance and minimal organization, suggesting that the model struggled to learn representations in the presence of large averaging pools. However, also similar to the performance results, the spatial reweighting RF r_{SR} had minimal effect on the topography, with topographic organization developing in the fc_out sublayer across a large range of RF sizes.

A Keys PC2



B fc_out PC2

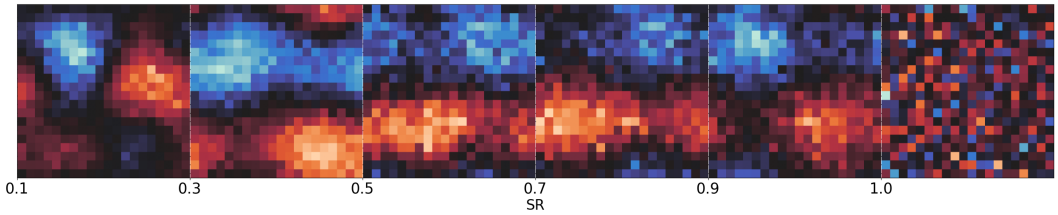


Figure 11: Effect of receptive field size on topographies. Panel **A** illustrates the impact of the spatial querying RF width $r_S Q$ on the topography of keys, while panel **B** visualizes the effect of spatial reweighting RF width r_{SR} on the topography of fc_out.

A.9.3 MEAN ACTIVATIONS

In this section, we extract mean activations over tokens for each sentence in four test suites, generating relevant mean activation profiles. As illustrated in Figure 12, these mean activations exhibit a high degree of correlation. To explore whether this phenomenon is unique to Topoformer-BERT, we conduct a comparative analysis of keys sublayer representations for animate and inanimate sentences across three distinct models. The results, depicted in Figure 12, reveal a high correlation among models, including an off-the-shelf all-mini-LM-L6-v2 Reimers & Gurevych (2019), demonstrating consistent responses across different categories.

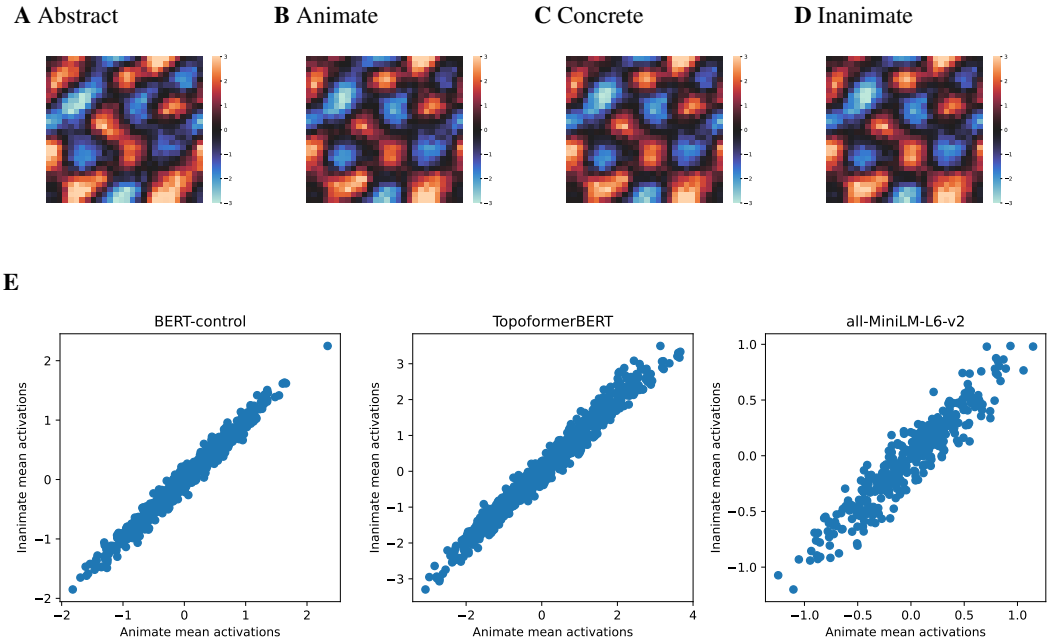


Figure 12: Mean activations in Topoformer-BERT layer 15 (keys). On 4 different test suites. Mean activations are highly correlated across all categories. Panel **E** Compares the keys sublayer activations for sentences about animate and inanimate information, across 3 different models in the final layer of Topoformer-BERT, BERT-control, and an off-the-shelf model all-miniLM-L6-v2 (Reimers & Gurevych, 2019) activations. Each model shows highly correlated responses.