5 Human Brain Data

5.1 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 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.

5.2 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; [16]). 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 [3] 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. 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 [18]. 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, we extracted voxels from language-selective regions in the brain. Language selectivity was defined based on an extensively validated language localizer task contrasting reading of *sentences* with *non-words strings* [7, 15]). We identified the top 10% language-selective voxels in 5 broad anatomical parcels 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]). These parcels delineate the expected gross locations of language-selective brain regions but are sufficiently large to encompass individual variability. The number of voxels in region of interest (ROI) was 75 for IFG, 37 for IFGorb, 47 for MFG, 163 for AntTemp, and 295 for PostTemp. In addition, we included a language network [netw] region (617 voxels), which consisted of all voxels in the aforementioned five regions, yielding a total of six regions of interest (ROIs) in our study.

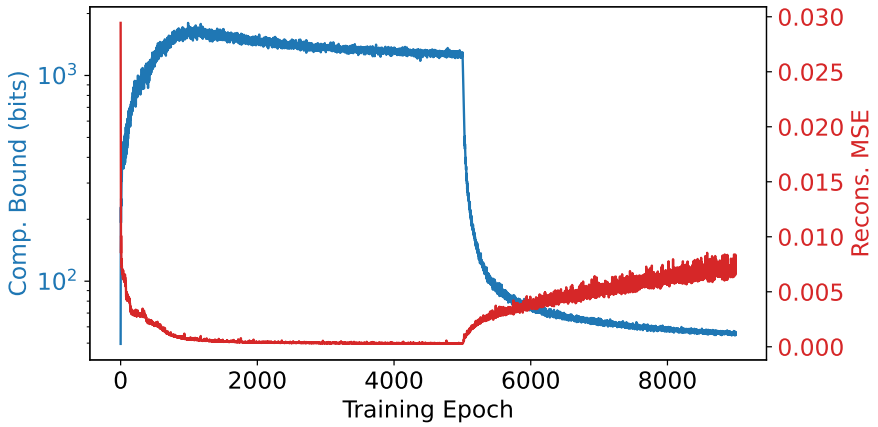


Figure 4: Training curves for a particular fv-VAE model, compressing IFG data for participant B. For the first 5,000 epochs, the model converged to high complexity (left axis) and low MSE (right axis). After epoch 5,000, we increased β, which decreased complexity and increased MSE.

6 Implementation Details

Here, we include further details about the fv-VAE model architecture, training process, and data sources used in our experiments.

Neural Architectures We used the same feedforward neural architecture for the fv-VAE models in all experiments. Anonymized code for replicating our experiments is included here, although given the sensitive nature of fMRI scans, we have not included the brain data in the repository.

A deterministic, feedforward encoder model mapped from an input, x, to a continuous hidden representation, h, via three fully-connected layers ReLU layers of size 1024, 512, and 64. We passed h through a single fully-connected 128-unit layer to generate μ, according to which we sampled a continuous latent representation $z \sim \mathcal{N}(\mu, I)$. Recall that this is similar to a standard VAE, but with a fixed unit variance.

The decoder mirrored the encoder model architecture: three fully-connected layers of size 512, 1024, and a final layer of the input size's dimension (which varied according to brain region). The first and second decoder layers used ReLU activations; the last layer used a sigmoid activation, as all fMRI data were normalized to be between 0 and 1.

Training fv-VAE Figure 4 depicts a typical training run, plotted here for the IFG region of participant B. Overall, the model was trained for 9,000 epochs, using batch size 250, using a default Adam optimizer with learning rate 0.001. For the first 5,000 epochs, we fixed $\beta = 1e-07$; this small but positive value allowed models to converge to low MSE values and mitigated numerical stability issues that arose if we set $\beta = 0$. As shown in Figure 4, for the first 5,000 epochs, the models converged to low MSE and high complexity. (Directly measuring the exact complexity is challenging, so we plotted the variational bound on complexity, computed via the KL divergence of two Gaussians.) After epoch 5,000, we increased β by $1e-08 \log(\text{epoch} - 5000)$ at each epoch. One could use a different annealing rate for β but, as evidenced by Figure 4, our chosen values tended to increase MSE and decrease complexity.

To extract brain data at varying levels of compression, we saved checkpoints of fv-VAE models during training, after epoch 5,000. Specifically, we used checkpoints every 100 epochs from epoch 5,000 to 6,000, and every 500 epochs from epoch 6,500 to 9,000 (all ranges inclusive). We used more frequent sampling in the earlier epochs, as MSE tended to increase more quickly in that region. Lastly, for each checkpoint, we computed the actual compressed representation for each sentence by passing it through the fv-VAE model and recording the output, μ. By recording μ, rather than sampling from a Gaussian centered at μ, we reduced noise in subsequent RSA analysis.

GPT2-XL data In the main paper, we described how we generated BERT embeddings using the [CLS] token. In additional experiments, we compared brain data to representations from the unidirectional-attention Transformer GPT2-XL model [19] (48 layers, embedding dimension of $1,600$), available via the HuggingFace library (Wolf et al. [28], Transformers version 4.11.3) To generate a single representation for an entire sentence, we used the representation of the last token in the GPT model.

7 Variational Autoencoders

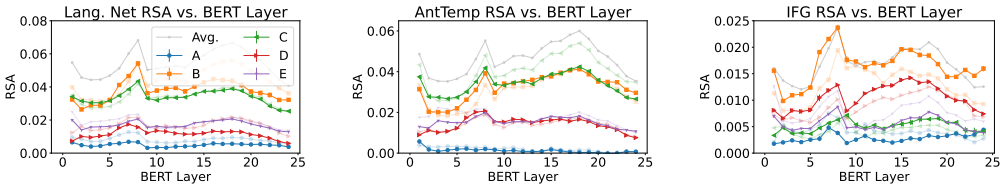
Here, we include an extended discussion of variational autoencoders (VAEs) [12] and our extension to fixed-variance VAEs. In a traditional VAE, an encoder is characterized a deterministic feedforward network that maps from an input, x, to parameters of a Gaussian distribution: $\mu(x), \Sigma(x)$. Using the "reparametrization trick," one samples a latent representation, z, from the Gaussian distribution, and z is used to generate a reconstruction of x via a decoder network.

Overall, the VAE training loss comprises a reconstruction loss (e.g., MSE) and a bound on the complexity of representations: $I(X; Z)$. Equation 3 establishes this complexity loss.

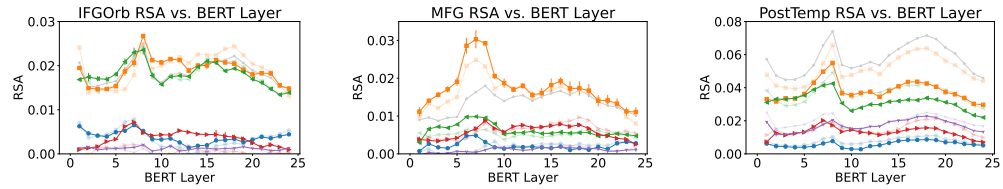
$$\begin{aligned} I(X;Z)_{\text{VAE}} &= D_{\text{KL}}[\mathbb{P}(X,Z) \| \mathbb{P}(X)\mathbb{P}(Z)] \\ &= D_{\text{KL}}[\mathbb{P}(Z|X)\mathbb{P}(X) \| \mathbb{P}(X)\mathbb{P}(Z)] \\ &= D_{\text{KL}}[\mathcal{N}(\mu(x), \Sigma(x)) \| \mathbb{P}(Z)] \\ &\leq \Sigma(x)^2 + \mu(x)^2 - \log(\Sigma(x)) - \frac{1}{2} \end{aligned} \tag{3}$$

The first two lines include definitions of complexity, using the KL divergence of the joint distribution from the product of its marginals. The third line follows from the nature of the VAE architecture, wherein we sample z from a Gaussian distribution. Lastly, the fourth line sets an upper bound on the complexity of representations by assuming that $\mathbb{P}(Z)$ is a unit Normal distribution, centered at the origin.

In our fixed-variance VAE (fv-VAE), we set the variance of a traditional VAE encoder as the identity matrix, but otherwise follow the normal sampling mechanism and training loss. The training loss, in particular, simplifies when replacing $\Sigma(x)$ and removing constant terms, to only include $\mu(x)^2$. We note, however, that the fv-VAE method is not simply an L2-regularized model; it samples latent representations, which is a necessary component for establishing complexity bounds.

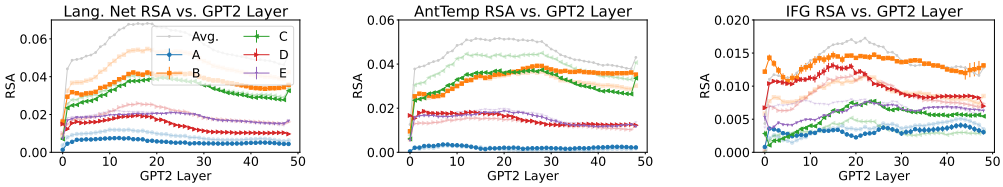


(a) BERT netw (b) BERT AntTemp (c) BERT IFG

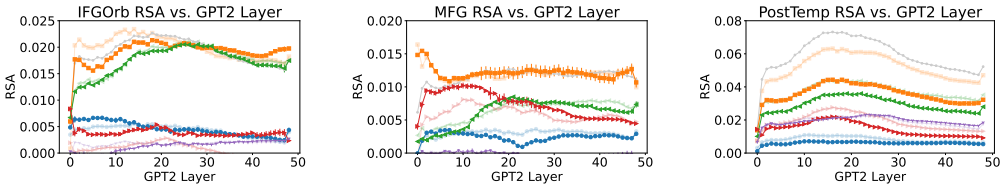


(d) BERT IFGorb (e) BERT MFG (f) BERT PostTemp

Figure 5: RSA scores comparing compressed (bold) and uncompressed (faded) brain representations, across BERT layers. As a further baseline, we include RSA scores using the averaged similarity matrix across participants. Compression increased RSA scores for some frontal regions, but not temporal regions.



(a) GPT2 netw (b) GPT2 AntTemp (c) GPT2 IFG



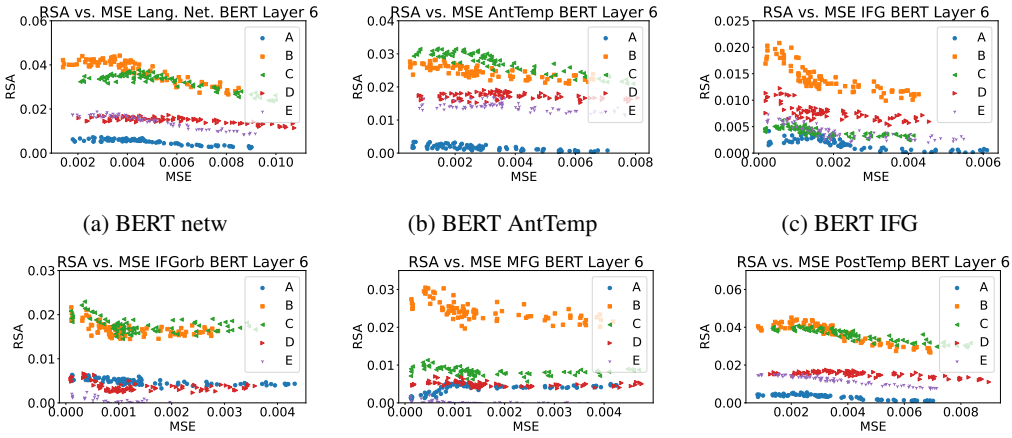
(d) GPT2 IFGorb (e) GPT2 MFG (f) GPT2 PostTemp

Figure 6: RSA scores between participant fMRI data and GPT2-XL embeddings. As in Figure 5, bold colors represent RSA scores for compressed data; faded colors represent uncompressed data. Trends largely mirror results from BERT: we observed some increases in RSA for participants B, C, and D in frontal regions.

8 Additional Results

In the main paper, we included some of the key results from our approach, highlighting RSA scores for particular regions of interest. Here, we present more complete results, including RSA scores using BERT and GPT2 embeddings, for all five regions of interest, as well as the overall language network (netw). Results for BERT and GPT2 are included in Figures 5 and 6, respectively.

As in the main paper, each colorful line represents the RSA scores for a particular participant using compressed (bold) or uncompressed (faded) fMRI data. In addition to such analysis, we included a "averaged" baseline, for which we computed the average similarity matrix across all participants before calculating the RSA score. For example, for the AntTemp region, we computed the (1000×1000) Pearson similarity matrix for each of the five participants, averaged the five matrices, and computed the RSA score between the BERT similarity matrix and the averaged participant similarity matrix.



(a) BERT netw (b) BERT AntTemp (c) BERT IFG

(d) BERT IFGorb (e) BERT MFG (f) BERT PostTemp

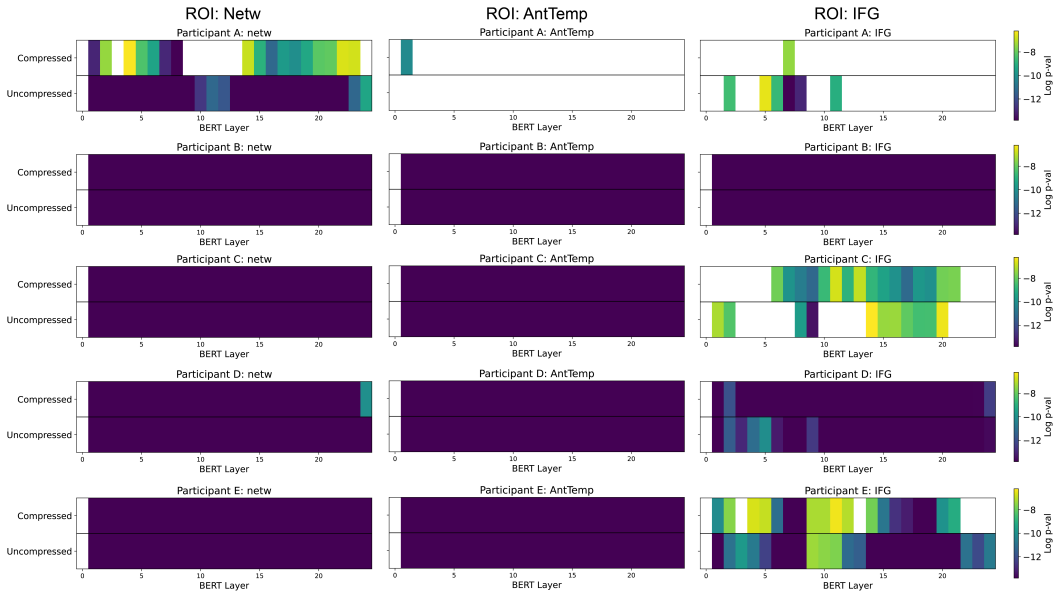
Figure 7: RSA vs. MSE using BERT Layer 6 embeddings. In several brain regions, small increases in MSE led to increases in RSA, suggesting benefits to compressing brain data.

Figures 5 and 6 jointly speak to the robustness of our results by displaying similar trends for different LLM embeddings. That is, for both BERT and GPT embeddings, we observed increased RSA scores for compressed brain representations in frontal regions, for participants B, C, and D, but not in temporal regions.

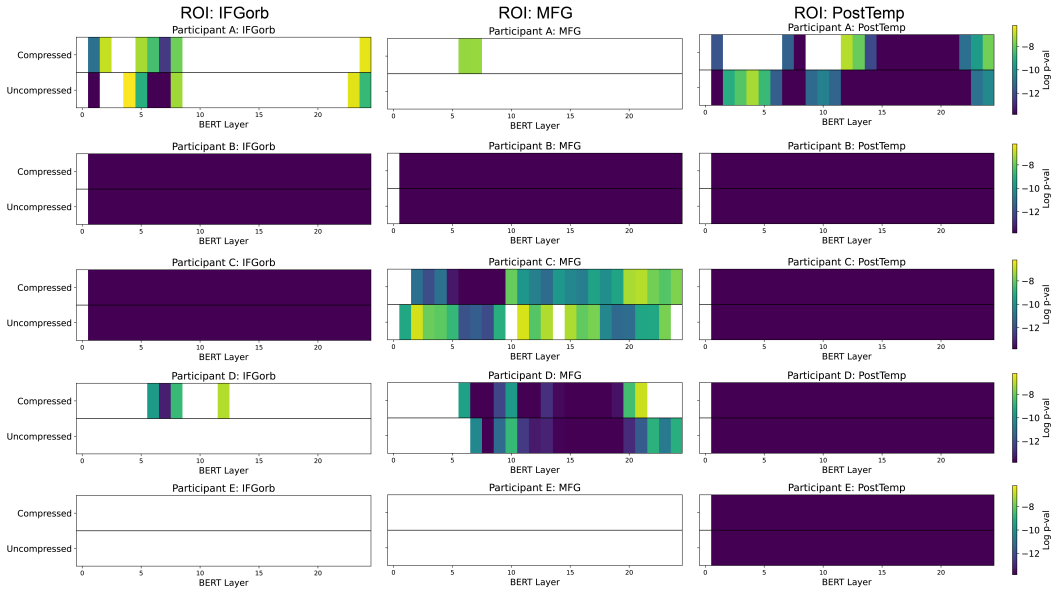
Figure 7 provides a snapshot of the benefits conferred by compressing brain data. Each figure mirrors Figure 2 a in plotting RSA vs. MSE, for embeddings from BERT layer 6. Increases in RSA as MSE increases indicates that compressing brain data increases alignment with LLM representations. Several brain regions, including, interestingly, temporal regions, produce such curves. For example, considering the full language network (Figure 7 a), RSA for participant B peaks for an MSE of approximately 0.004 – greater than the minimum MSE of 0.002. These results offer tantalizing but incomplete evidence that compressing brain data could improve alignment for all brain regions. We hope to continue to investigate such effects in future work.

9 Statistical Analysis

We provide statistical significance values associated with the brain-LLM RSA scores (obtained via the Spearman correlation coefficient) for the main BERT analyses (Figures 2, 3, and 5). Each heatmap in Figure 8 shows the log p-value for each ROI (columns) for each participant (rows). Each heatmap has two rows, corresponding to the p-values for the compressed and uncompressed RSA scores, across all BERT layers. Lighter values indicate less significant RSA scores. The upper bound (yellow) of the color scale is $\log(0.05/25)$ which corresponds to a Bonferroni corrected p-value (correction for number of layers); the lower bound (dark purple) is fixed at $\log(0.000001)$. Blank areas correspond to non-significant scores. Most scores were highly significant, as evident from the dark panels.



(a) Log of p-values associated with RSA scores for the Netw, AntTemp, and IFG ROIs (columns) for all participants (rows), across all BERT layers.



(b) Log of p-values associated with RSA scores for the IFGorb, MFG, PostTemp ROIs (columns) for all participants (rows), across all BERT layers.

Figure 8: P-values associated with RSA scores for all ROIs and all participants across BERT layers.