Increasing Brain-LLM Alignment via Information-Theoretic Compression

Mycal Tucker*
mycal@mit.edu

Greta Tuckute*
gretatu@mit.edu

Abstract

Recent work has discovered similarities between learned representations in large language models (LLMs) and human brain activity during language processing. However, it remains unclear what information LLM and brain representations share. In this work, inspired by a notion that brain data may include information not captured by LLMs, we apply an information bottleneck method to generate compressed representations of fMRI data. For certain brain regions in the frontal cortex, we find that compressing brain representations by a small amount increases their similarity to both BERT and GPT2 embeddings. Thus, our method not only improves LLM-brain alignment scores but also suggests important characteristics about the amount of information captured by each representation scheme.

1 Introduction

Artificial large language models (LLMs) have emerged as the most accurate models of human language processing. LLMs generate probabilities of upcoming words that predict human reading patterns [27, 22, 17] and internal LLM representations can predict brain signals of humans reading or listening at the granularity of fMRI voxels and intracranial recordings [21, 4, 9]. Such LLMs have fundamentally shifted the neuroscience of language: for the first time, we have models that are able to predict brain activity to the extent where LLMs are used to simulate experiments *in silico* [11], generate sentences for driving or suppressing brain responses [26], or infer the story content that an individual was listening to [23]. All these studies leverage the predictive power of LLMs. However, relatively few studies ([1, 8]) investigate the *representational alignment* between LLM internal states and human brain units. In our work, we first investigate the representational similarity between two widely used LLM architectures [6, 20] and human fMRI voxels for a very large (1,000) set of diverse, naturalistic sentences. Second, we ask whether we can leverage an information bottleneck (IB) approach [24, 2] to generate compressed representations of brain activity and, in doing so, increase representational alignment between LLMs and humans. Surprisingly, we find that, in some frontal regions of the brain, compressing brain data increases alignment with LLM representations. This finding suggests that brain data encode information that LLM representations do not contain. More broadly, our work establishes the use of information theoretic tools as a "dial" to modify representational complexity and to better unify representational spaces, including those from artificial and biological language processing.

2 Approach

We investigate representational alignment between brain responses during language processing, and LLM representations. In compressing brain representations, we seek to explore how the two systems represent linguistic information in a unified manner (Figure 1).

*Equal contribution by both authors.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

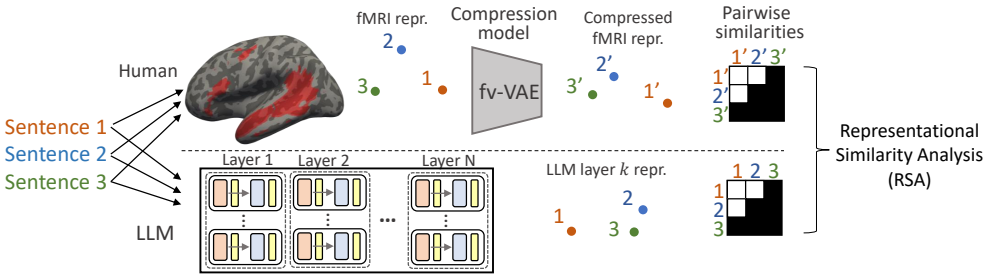


Figure 1: Overall approach: we measured fMRI brain activations (top) and LLM activations (bottom) for 1,000 sentences. By compressing brain representations using a variational information bottleneck method, we sought to improve representational alignment with LLMs.

Human Brain Data We recorded brain activity in event-related 3T fMRI from N=5 participants (4 female, aged 21-30, 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. Participants read each sentence presented on the screen one at a time for 2s with an inter-stimulus interval of 4s. Blood oxygenation level dependent (BOLD) responses to each sentence were estimated using GLMsingle [18] (5 noise regressors and a ridge regression fraction of 0.05). To mitigate the effect of collecting data across two separate scan sessions, the estimates from each session were z-scored for each voxel in each participant. We were interested in brain responses from language-selective areas of the brain, so we identified the language network functionally in each participant using 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]). In addition, we included a language network [netw] region, which consisted of all voxels in these 5 regions, yielding a total of 6 regions of interest (ROIs) in our study. All participants gave informed written consent in accordance with the requirements of an institutional review board. Further information can be found in Appendix 5 and [26].

Large Language Model Representations We obtained sentence representations, for each of the 1,000 sentences, from the pre-trained BERT-large-cased model [5] (24 layers, embedding dimension of 768). To obtain a summary representation of each sentence, we used the classification token, `[CLS]`. In the main paper, we focused on results using BERT representations, but we also conducted experiments using GPT2-XL embeddings (further details in Appendix 6).

Information Bottleneck In our approach, we use Information Bottleneck (IB) methods to generate compressed representations of brain data. In classic IB literature, one considers a stochastic encoder $q(z|x)$ and decoder $p(\hat{x}|z)$ that seek to reconstruct an input, x, via a lossy representation, z [24]. IB systems weigh a trade-off between competing terms: maximizing informativeness – $I(X;\hat{X})$, how well an input can be reconstructed from z – and minimizing complexity – $I(X;Z)$, how many bits about an input are contained in z. This tradeoff is expressed more formally via the optimization

$$\max \quad I(X;\hat{X}) - \beta I(X;Z) \tag{1}$$

where X is an input, Z is a representation of X, and $\hat{X}$ is a reconstruction of the input, given Z. By varying the scalar weight, β, one can control representational complexity.

Substantial prior work establishes exact and approximate methods for solving Equation 1 for varying β, generating encoders across a spectrum of complexity values [24, 2, 10, 25]. Similar to such work, we propose a neural network architecture that supports variational bounds on complexity, which we use to generate compressed representations of brain data for improved representational alignment.

Neural Compression Model: Fixed Variance VAE We adopted a variational approach to the IB optimization, trading off informativeness and complexity, using a neural architecture based upon Variational Autoencoders (VAE) [12]. In a VAE, given an input, x, a deterministic encoder outputs

parameters to a Gaussian distribution (a mean, μ, and a diagonal standard deviation matrix, Σ), from which a latent variable, z, is sampled: $z \sim q_\theta(z|x) = \mathcal{N}(\mu(x), \Sigma(x))$ [12, 10]. This Gaussian parametrization of the encoder supports a variational bound on complexity (see details in Appendix 7.

In our work, we propose a novel neural network method, the fixed-variance VAE (fv-VAE), that makes a simple modification to existing VAE architectures to set $\Sigma(x) = I$, where I is the identity matrix. Using this constant variance, the bound on complexity in Equation 3 is easily updated to remove terms dependent upon $\Sigma(x)$ and yield the overall fv-VAE training loss:

$$\mathcal{L}(x, \hat{x}) = ||x - \hat{x}^2|| + \beta\mu(x)^2 \tag{2}$$

where the first term represents the mean squared error (MSE) of a reconstruction, and the second term is the simplified bound on complexity, weighted by a scalar β. By varying β, one may limit the amount of information about x encoded in z.

The small difference between VAE and fv-VAE affords important benefits for Representational Similarity Analysis (RSA; [14]). In a standard VAE, $\Sigma(x)$ can vary by dimension in the latent space, leading to some dimensions appearing more "stretched" than others. The relative scales of such dimensions are ignored in RSA using similarity metrics such as Euclidean distance or Pearson correlation, as these metrics weigh each feature dimension equally. Thus, in our work, we used fv-VAE to fix a constant scale across dimensions.

3 Experiments

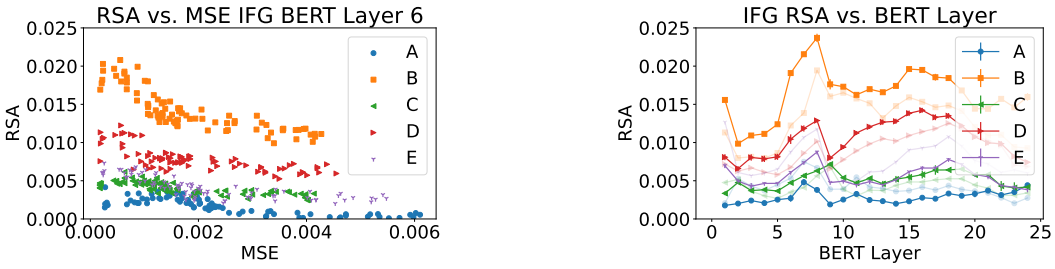
Training We trained fv-VAE models with latent dimension 128 to reconstruct the fMRI data described in Section 2: for five participants and each of the six brain ROIs, we minimized the reconstruction MSE of the 1,000 sentences (with the additional complexity loss). Models were trained for 5,000 epochs to first converge to low MSE; after epoch 5,000, β increased every epoch to penalize representation complexity, which in turn increased MSE. By saving checkpoints as β increased, and by replicating training across five random seeds, we generated a suite of compressed brain representations at different complexity levels. Implementation details are included in Appendix 6.

Evaluation We used Representational Similarity Analysis (RSA; [13]) to evaluate representational similarity between brain and LLM representations. RSA characterizes the similarities for all pairs of sentences across all features in either the brain or the LLM representation space. Intuitively, RSA assesses the correspondence between human and LLM representations under the assumption that all features contribute equally to capture the population-level representation-space geometry.

To evaluate RSA, we first generated compressed fMRI representations for each of the 1,000 sentences in our dataset (generated as $\mu(x)$ using a trained fv-VAE). Next, we computed the Pearson similarity of all sentence pairs of the compressed representations, yielding a square similarity matrix. Lastly, we calculated the RSA metric via the Spearman correlation coefficient between the upper triangulars of the brain similarity matrix and the LLM similarity matrix at a given layer. Overall, this RSA metric corresponded to the notion of "are sentences that are similar in the compressed brain representation space similar in the LLM embedding space." Figure 1 depicts this overall process.

Inferior Frontal Gyrus (IFG) Results Figure 2 depicts the similarity between LLM representations and compressed brain responses (the frontal IFG language region), demonstrating how compressing representations increased RSA scores for some participants. Figure 2 a shows the RSA scores vs. MSE of the compressed brain representations for BERT Layer 6. The MSE value captures the reconstruction error of the trained fv-VAE model used to generate the representations; greater MSE corresponds to greater compression. At a high level, RSA scores for all participants (different colors) decrease at the highest MSE values. However, critically, for low MSE values, RSA scores for participants B, C, and D *increase* as MSE increases. That is, compressing the fMRI data to a certain extent increases similarity to BERT representations.

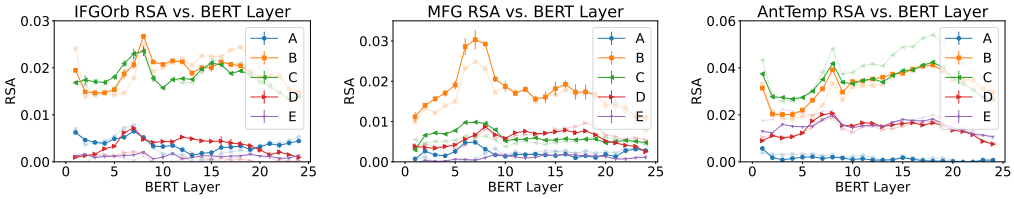
Figure 2 b shows systematic improvements across all BERT layers, for participants B, C, and D. Each faded line represents RSA scores for non-compressed fMRI data from each participant; bold lines represent RSA scores using compressed fMRI representations (RSA scores were highly significant (Appendix 9)). We chose the optimal level of compression via 5-way cross-validation, selecting



(a) RSA vs. MSE BERT Layer 6

(b) RSA vs. BERT Layer

Figure 2: RSA scores between the inferior frontal gyrus (IFG) and BERT embeddings. a) For BERT representations from one layer, we calculated the RSA for compressed brain data (increased MSE reflects increased compression). For some subjects, like B and D, small increases in compression improved RSA. b) Using the optimal MSE for each layer, we compared compressed RSA scores (bold) to uncompressed (faded) across BERT layers. Subjects B, C, and D had improved alignment.



(a) IFGorb

(b) MFG

(c) PostTemp

Figure 3: RSA for frontal (a, b) and temporal (c) regions, as a function of BERT embedding layer. In frontal regions, compressing fMRI data improved alignment, particularly for participants B and D.

the optimal MSE for a single fv-VAE model and averaging the corresponding RSA values for other models at that MSE. For participants for whom the bold lines are above the faded lines (B, C, and D), we find a consistent benefit in compressing brain representations.

Frontal Brain Region Results We performed similar RSA analyses for compressed representations for frontal (IFG, IFGorb, and MFG) and temporal (AntTemp and PostTemp) brain regions of interest. Compressing frontal region representations increased alignment for participants B, C, and D, but we found no such benefit for temporal regions. For example, for BERT layers < 10, compressed scores for participant C improved over the uncompressed RSA scores in regions IFGorb and MFG (Figures 3 a and b, respectively), but RSA scores for AntTemp were largely unchanged for compressed data (Figure 3 c). Further experiments, computing RSA scores between GPT2 embeddings and compressed brain representations corroborate this trend: frontal regions showed some improvements from compression whereas temporal regions did not. Notably, the lack of increased alignment for *all* brain regions highlights that the positive results for frontal brain regions cannot be solely attributed to simple explanations such as all compressed brain representations being more similar to LLM representations. Results for all regions, for both BERT and GPT2 RSA scores, are included in Appendix 8. Most scores were highly significant (Appendix 9).

4 Contributions

In this work, we proposed an information bottleneck method to generate compressed representations of brain activity during sentence processing and showed that, for some brain regions, such compression increased alignment with LLM representations. These early findings suggest that traditional fMRI methods employed during language experiments capture data that LLMs do not represent; future work may improve LLMs or seek to better understand sources of variation in brain activity. Our work hints at ideas that warrant further investigation such as 1) characterizing differences between brain regions that have previously been treated as a single language network, and 2) understanding what information remains in optimally-compressed brain representations. Ultimately, we advocate for an information-theoretic approach to understand and align representation spaces.

References

[1] Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. Blackbox meets blackbox: Representational similarity & stability analysis of neural language models and brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, January 2019. doi: 10.18653/v1/W19-4820. URL https://doi.org/10.48550/arXiv.1906.01539.

[2] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. URL https://arxiv.org/abs/1612.00410.

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

[4] 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.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

[6] 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.

[7] Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castanon, Susan L. Whitfield-Gabrieli, and Nancy G. Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104 2:1177–94, 2010.

[8] Ariel Goldstein, Avigail Dabush, Bobbi Aubrey, Mariano Schain, Samuel A. Nastase, Zaid Zada, Eric Ham, Zhuoqiao Hong, Amir Feder, Harshvardhan Gazula, Eliav Buchnik, Werner Doyle, Sasha Devore, Patricia Dugan, Daniel Friedman, Michael Brenner, Avinatan Hassidim, Orrin Devinsky, Adeen Flinker, and Uri Hasson. Brain embeddings with shared geometry to artificial contextual embeddings, as a code for representing language in the human brain. 2022. doi: 10.1101/2022.03.01.482586. URL https://doi.org/10.1101/2022.03.01.482586.

[9] 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, 2022.

[10] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

[11] Shailee Jain, Vu Anh Vo, Leila Wehbe, and Alexander G. Huth. Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, pages 1–27, 2023. doi: 10.1162/nol_a_00101.

[12] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008.

[14] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008. ISSN 1662-5137. URL https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008.

[15] Benjamin Lipkin, Greta Tuckute, Josef Affourtit, Hannah Small, Zachary Mineroff, Hope Kean, Olessia Jouravlev, Lara Rakocevic, Brianna Pritchett, Matthew Siegelman, Caitlyn Hoeflin, Alvincé Pongos, Idan A. Blank, Melissa Kline Struhl, Anna Ivanova, Steven Shannon, Aalok Sathe, Malte Hoffmann, Alfonso Nieto-Castañón, and Evelina Fedorenko. Probabilistic atlas for the language network based on precision fMRI data from >800 individuals. *Scientific Data*, 9(1): 529, August 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01645-3. URL https://www.nature.com/articles/s41597-022-01645-3. Number: 1 Publisher: Nature Publishing Group.

[16] 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.

[17] Byung-Doh Oh and William Schuler. Transformer-based lm surprisal predicts human reading times best with about two billion training tokens. *arXiv preprint arXiv:2304.11389*, April 2023. doi: 10.48550/arXiv.2304.11389. URL https://doi.org/10.48550/arXiv.2304.11389. Cite as: arXiv:2304.11389 [cs.CL] (or arXiv:2304.11389v1 [cs.CL] for this version).

[18] 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.

[19] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[20] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24, 2019.

[21] 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.

[22] 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.

[23] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866, May 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01304-9. URL https://www.nature.com/articles/s41593-023-01304-9. Number: 5 Publisher: Nature Publishing Group.

[24] Naftali Tishby, Fernando C. Pereira, and William Bialek. The Information Bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999.

[25] Mycal Tucker, Roger P. Levy, Julie Shah, and Noga Zaslavsky. Trading off utility, informativeness, and complexity in emergent communication. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=O5arhQvBdH.

[26] 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.

[27] E. G. Wilcox, J. Gauthier, J. Hu, P. Qian, and R. Levy. On the predictive power of neural language models for human real-time comprehension behavior. *ArXiv*, 2020.

[28] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Acknowledgments and Disclosure of Funding

This work was supported by the Amazon Fellowship from the Science Hub (administered by the MIT Schwarzman College of Computing) (G.T., M.T.), the International Doctoral Fellowship from American Association of University Women (AAUW) (G.T.), the K. Lisa Yang ICoN Center Graduate Fellowship (G.T.).

G.T. is also grateful for mentorship and support from Evelina Fedorenko, and M.T. thanks Julie Shah for her support throughout this project.

5 Human Brain Data

5.1 Participants and Acquisition

We recorded brain responses using fMRI from N=5 participants during a sentence reading task (see [26]). 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.

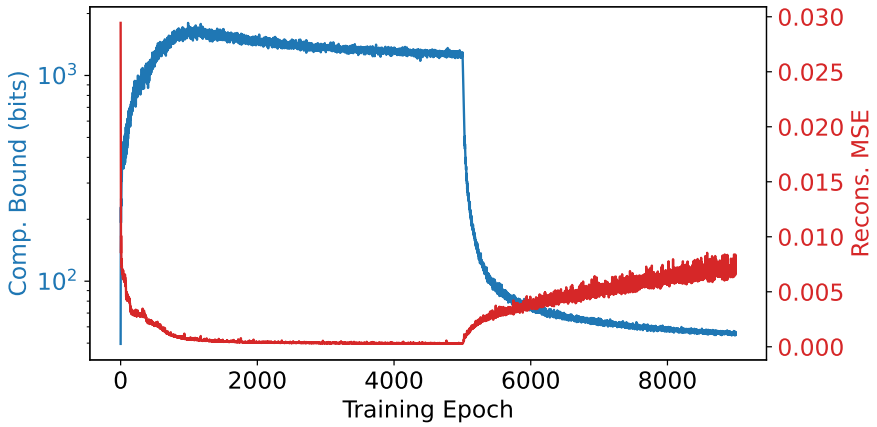


Figure 4: Training curves for a particular fv-VAE model, compressing IFG data for participant B. For the first 5,000 epochs, the model converged to high complexity (left axis) and low MSE (right axis). After epoch 5,000, we increased β, which decreased complexity and increased MSE.

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. Code for replicating our experiments is included at https://github.com/mycal-tucker/brain-compression, 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

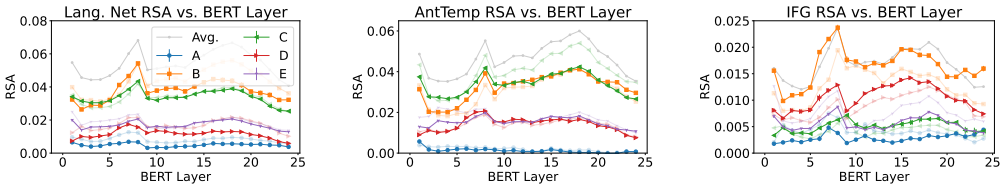
Here, we include an extended discussion of variational autoencoders (VAEs) [12] and our extension to fixed-variance VAEs. In a traditional VAE, an encoder is characterized a deterministic feedforward network that maps from an input, x, to parameters of a Gaussian distribution: $\mu(x), \Sigma(x)$. Using 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.

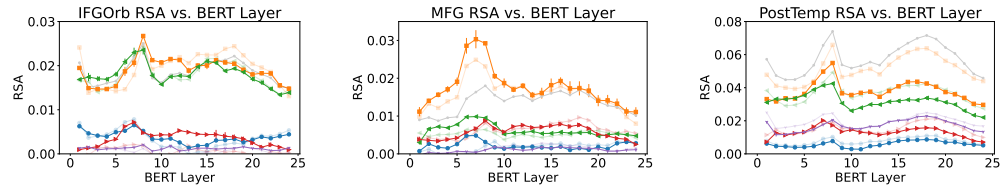
$$\begin{aligned} I(X;Z)_{\text{VAE}} &= D_{\text{KL}}[\mathbb{P}(X,Z) \| \mathbb{P}(X)\mathbb{P}(Z)] \\ &= D_{\text{KL}}[\mathbb{P}(Z|X)\mathbb{P}(X) \| \mathbb{P}(X)\mathbb{P}(Z)] \\ &= D_{\text{KL}}[\mathcal{N}(\mu(x), \Sigma(x)) \| \mathbb{P}(Z)] \\ &\leq \Sigma(x)^2 + \mu(x)^2 - \log(\Sigma(x))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

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.

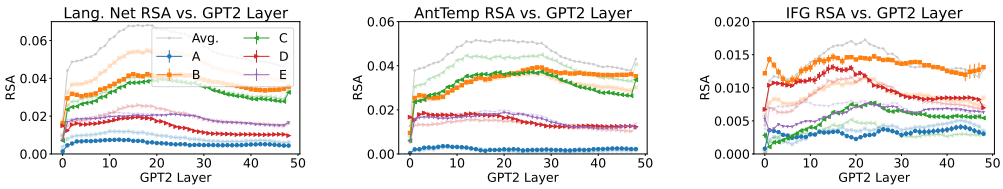


(a) BERT netw (b) BERT AntTemp (c) BERT IFG

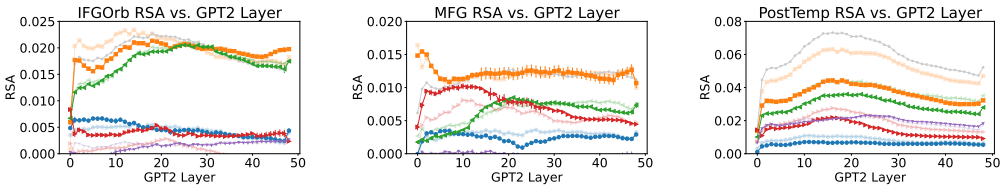


(d) BERT IFGorb (e) BERT MFG (f) BERT PostTemp

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



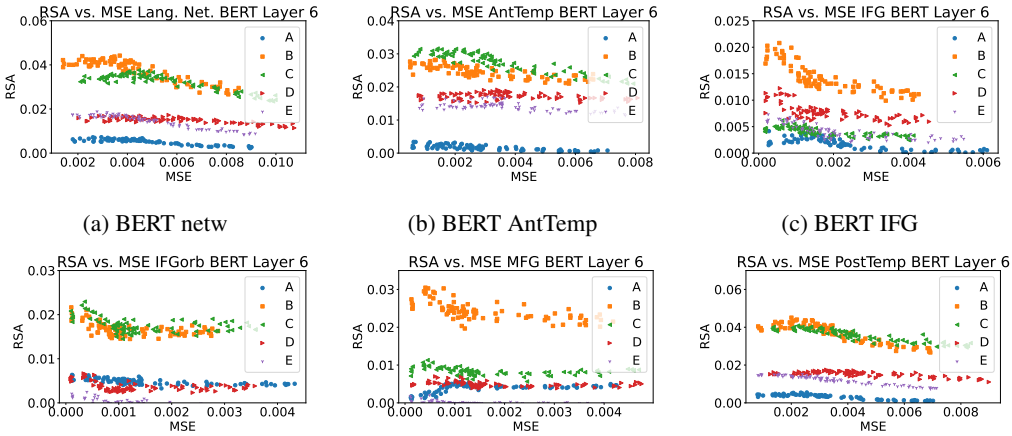
(d) GPT2 IFGorb (e) GPT2 MFG (f) GPT2 PostTemp

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

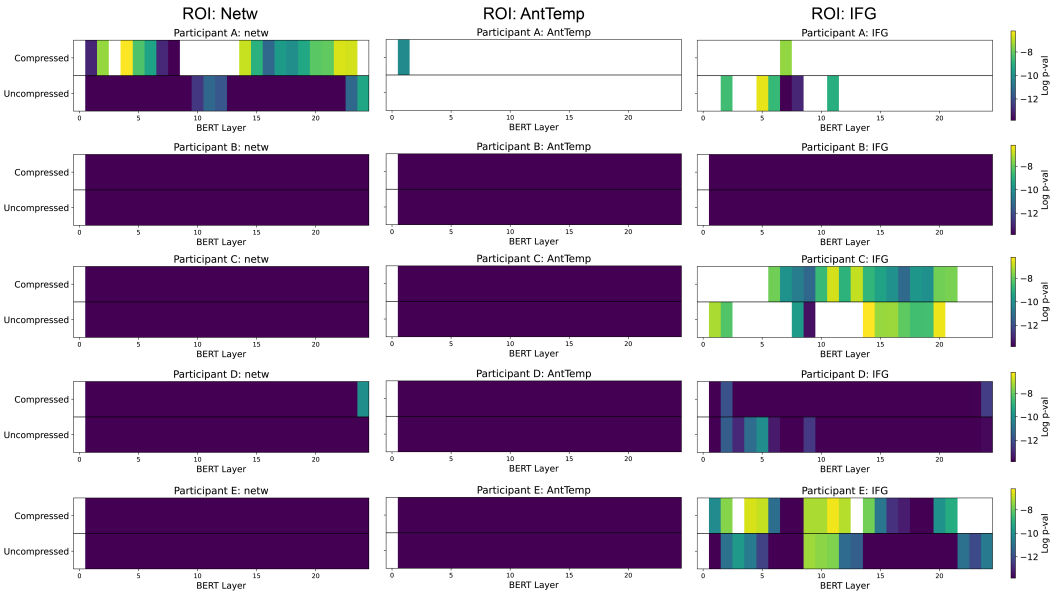
Figure 7: RSA vs. MSE using BERT Layer 6 embeddings. In several brain regions, small increases in MSE led to increases in RSA, suggesting benefits to compressing brain data.

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.

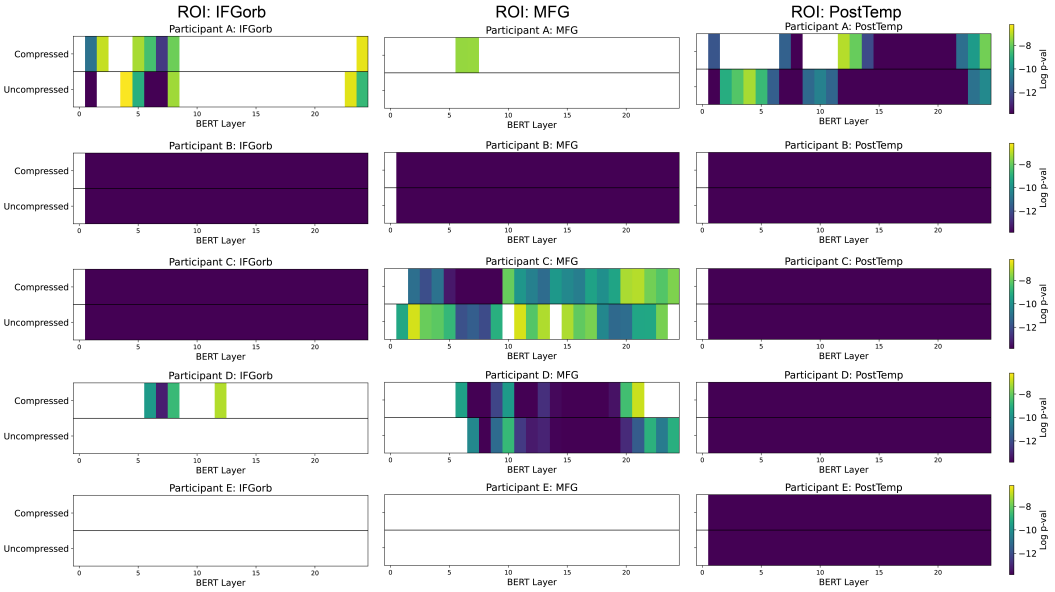
Figure 7 provides a snapshot of the benefits conferred by compressing brain data. Each figure mirrors Figure 2 a in plotting RSA vs. MSE, for embeddings from BERT layer 6. Increases in RSA as MSE increases indicates that compressing brain data increases alignment with LLM representations. Several brain 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.