Optimizing fMRI Data Acquisition for Decoding Natural Speech with Limited Participants

Louis Jalouzot
CEA, ENS
Université Paris-Saclay
France
jalouzot.louis@gmail.com

Alexis Thual
karavela.ai
France

Yair Lakretz
ENS, EHESS, CNRS
Université PSL
France

Christophe Pallier
INSERM, CEA, CNRS
Université Paris-Saclay
France

Bertrand Thirion
INRIA, CEA
Université Paris-Saclay
France

Abstract

We present a systematic investigation into decoding perceived natural speech from fMRI data in a participant-limited setting. Using a publicly available dataset of eight participants [LeBel et al., 2023], we demonstrate that deep neural networks trained with a contrastive objective can effectively decode unseen natural speech by retrieving the embedding of perceived sentences from fMRI activity. We found that decoding performance directly correlates with the amount of training data available per participant. In this data regime, multi-subject training does not improve decoding accuracy compared to the single-subject approach. Additionally, training on similar or different stimuli across subjects has a negligible effect on decoding accuracy. Finally, we find that our decoders model both syntactic and semantic features, and that stories containing sentences with complex syntax or rich semantic content are more challenging to decode. While our results demonstrate the benefits of having extensive data per participant (deep phenotyping), they suggest that leveraging multi-subject data for natural speech decoding likely requires deeper phenotyping or a substantially larger cohort.

1 Introduction

Decoding percepts from brain activity is a central challenge in neuroscience. Foundational fMRI studies demonstrated the feasibility of decoding visual stimuli [Kamitani and Tong, 2005] and identifying simple linguistic units [Formisano et al., 2008]. Recent advances, fueled by deep-phenotyping datasets with extensive data per participant, have enabled more accurate decoding of complex stimuli like natural scenes [Ozcelik and VanRullen, 2023, Allen et al., 2022, Banville et al.] and natural language [Tang et al., 2023, Ye et al., 2025, LeBel et al., 2023]. However, a key obstacle remains: the high inter-subject variability of brain responses, which complicates developing generalizable decoders. While methods like functional alignment show promising results [Thual et al., 2023, Ho et al., 2023, Scotti et al., 2024], it is still unclear how to best leverage multi-subject data.

In this work, we focus on decoding perceived natural speech from fMRI, adopting a decoding-first approach to directly predict text representations from brain activity using a contrastive objective, a method proven effective in other brain decoding studies [Défossez et al., 2023, Scotti et al., 2023]. We make the following contributions:

1. We achieve up to 36% top-10 accuracy in sentence embedding retrieval, demonstrating that decoding performance increases with the amount of training data per participant and does not saturate even at 13.5 hours.
2. Training decoders on multiple subjects does not improve accuracy compared to single-subject models in this data regime.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

3. Varying the degree of stimulus overlap between subjects has minimal impact on multi-subject decoding performance.
4. Our analysis shows that decoders capture both syntactic and semantic features, but sentences with complex syntax or rich semantic content remain more difficult to decode.

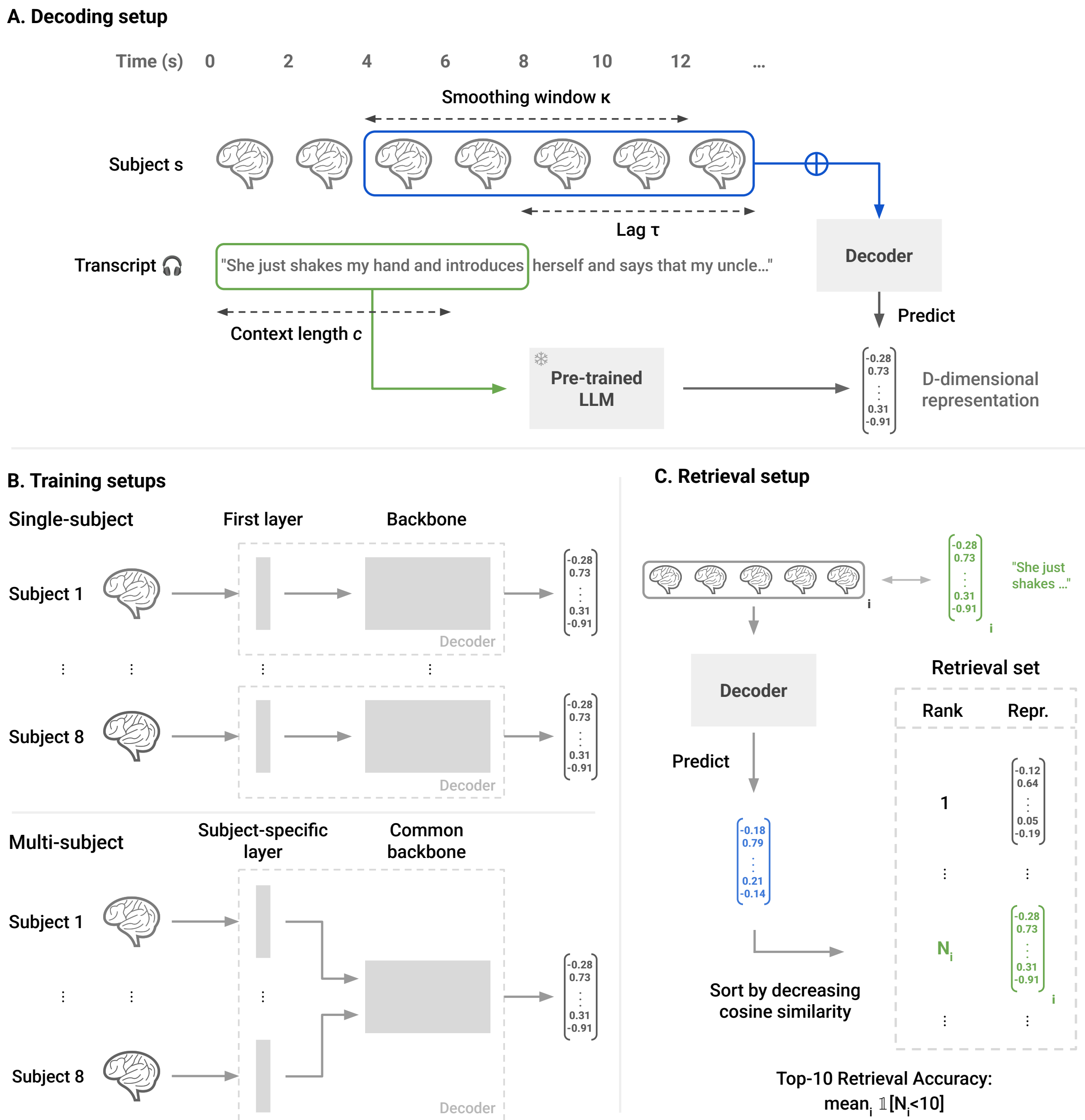

Figure 1: **Method for decoding natural speech from fMRI activity. A. Decoding setup** Deep Neural Networks are trained with a contrastive objective to predict text representations (derived from Large Language Models embeddings) from fMRI activity recorded as participants listened to natural speech. Key parameters include *context length* c, *lag* τ, and *smoothing* κ. **B. Subject approaches** We compare single-subject (one decoder per subject) and multi-subject (shared decoder backbone with subject-specific layers at the bottom) approaches. **C. Retrieval setup** Decoders are evaluated by ranking candidate chunks from a retrieval set based on cosine similarity between their representation and the predicted one. We compute top-10 accuracy.

2 Methods

We aim to train models that predict text representations (LLM embeddings) from fMRI activity using a contrastive objective, as summarized in Figure 1. The code will be made available upon publication.

Data and Task We use the dataset from LeBel et al. [2023], which contains 3T fMRI recordings of 8 participants listening to stories ($\sim$6 hours/subject). Three of these subjects have an additional $\sim$10.5 hours of data. To deal with the low temporal resolution of fMRI signal (one image is acquired every TR=2s), we chunk the perceived text into 2s segments according to acquisition windows. Our goal is to learn a mapping f from fMRI images $\boldsymbol{X}^s$ to LLM embeddings of the corresponding perceived chunks of text $\boldsymbol{Y}^s$. To account for hemodynamic delay, we predict the text embedding $\boldsymbol{Y}_t^s$ from a future fMRI volume $\boldsymbol{X}_{t+\tau}^s$ with τ=6s. Text representations are enriched by concatenating the transcripts of the current and $c = 3$ preceding text chunks before embedding.

Preprocessing fMRI data was preprocessed with `fmriprep` [Esteban et al., 2018], temporally smoothed (averaging over κ volumes), and standardized. This temporal averaging aims to reduce noise and improve the signal-to-noise ratio. For text, we used LLM2Vec [BehnamGhader et al., 2024] to generate $D = 4096$ dimensional embeddings, which were also normalized and standardized. To reduce dimensionality and focus on informative brain regions, we selected the top 4096 voxels for each subject based on their encoding performance. Specifically, we trained a Ridge regression model on the training set to predict fMRI activity from text embeddings, and selected voxels with the highest R^2 score on a separate validation set, thus avoiding circularity. We show results with 4096 voxels as it yields the best performance, but similar effects are observed with whole-brain data.

Evaluation Performance is evaluated using top-10 retrieval accuracy on a held-out test set from cross-validation, ensuring that training and test stimuli do not overlap. For each test sample, we rank all candidate embeddings from the test set (retrieval set) by their cosine similarity to the predicted embedding. Top-10 accuracy is the frequency with which the correct text embedding appears in the top 10 candidates. For comparability, we use 5-fold cross-validation for the subjects with less data and 15-fold for those with more, so that in both cases the retrieval set has approximately 2,000 samples. The variability across folds is reported with the average and 95% confidence intervals in our figures.
Moreover, we track two NLP similarity metrics between the ground truth and candidate text chunks from the retrieval set. We assess semantic similarity using GloVe bag-of-words cosine similarity on content words (nouns, verbs, adjectives). Syntactic similarity is assessed using Levenshtein distance on part-of-speech (POS) tagged chunks.

Training We model the mapping from fMRI to text representations using a 3-layer MLP with dropout, layer normalization, and skip connections inspired from Scotti et al. [2024]. This "Brain Decoder" is trained with a symmetric InfoNCE contrastive loss [Radford et al., 2021] (see Equation (1) in appendix for details). We use a learning rate scheduler and early stopping on the top-10 accuracy on a held-out validation set to mitigate overfitting. Hyperparameters (see Table 1 in appendix) were optimized via Bayesian search to maximize top-10 accuracy.
We compare a *Single-subject* approach (one decoder per participant) with a *Multi-subject* approach (a shared backbone with a subject-specific input layer). We also investigate the effect of stimulus overlap in the multi-subject setting by varying the proportion of shared stories during training.

3 Results

Baseline We compare our results to Ye et al. [2025]'s BrainLLM, which uses a similar dataset and setup. Although their model was not trained for retrieval, we adapted their predicted embeddings to our evaluation framework. BrainLLM achieves an average top-10 accuracy of 1.6% (chance=0.5%), providing a valuable baseline for comparison.

Single-subject decoding performance scales with data quantity Single-subject decoding accuracy scales directly with the amount of training data (Figure 2). For the three participants with 13.5 hours of training data, we achieve an average top-10 accuracy of 27% (subject 3 reaching 36%), significantly outperforming the BrainLLM baseline (1.6%). For participants with only 4 hours of data, accuracy drops to 6%. The performance curve does not plateau, suggesting that acquiring even more data per subject could yield further gains. These results represent, to our knowledge, the first successful decoding of natural speech from fMRI using a contrastive objective. Our methodological choices were critical to this success (Figure 3). We incrementally improved a baseline MLP by adding a hemodynamic lag, using a text context window, smoothing fMRI data, switching to LLM2Vec embeddings, adopting a contrastive loss, and using a more refined "Brain Decoder" architecture.

Multi-subject training does not improve decoding Contrary to expectations, training on data from multiple subjects does not improve individual-subject decoding performance (Figure 4). With large models, it can even be detrimental,

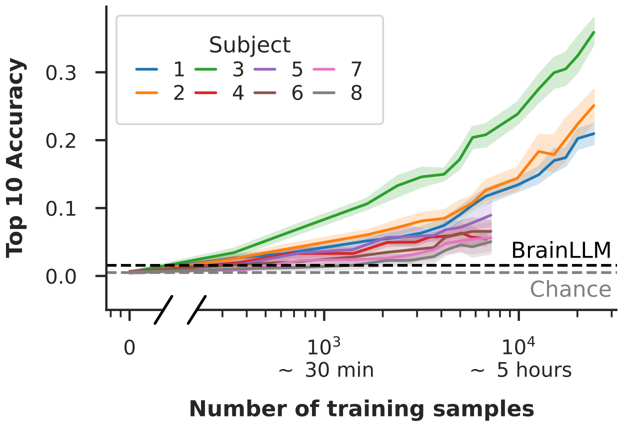

Figure 2: **Impact of training data amount on single-subject performance.** Cross-validated top-10 accuracy of single-subject decoders. The retrieval set contains ∼2k samples. Chance level (0.5%) and the BrainLLM baseline (1.6%) are shown for comparison.

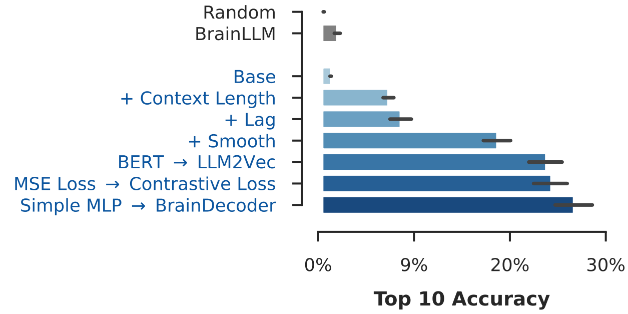

Figure 3: **Setup comparison.** Impact of various elements of the decoding setup on decoding performance. We start from a very crude version of our setup, namely "Base", which is essentially a simple MLP trained with MSE loss on BERT latents. Then each row corresponds to the previous setup with a modification described by its blue label.

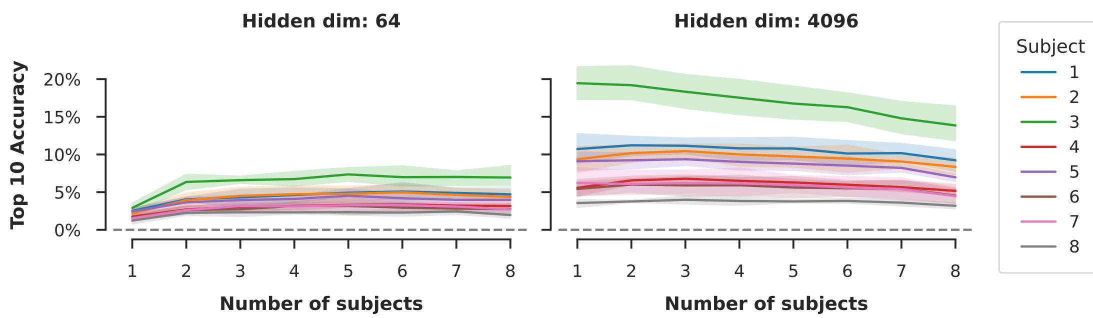

Figure 4: **Impact of the number of training subjects.** Multi-subject decoders were trained for all 255 combinations of the 8 subjects. For each subject (color), we plot the results for the combinations which achieved best accuracy for each number of subjects. We show results for small (left) and large (right) decoders.

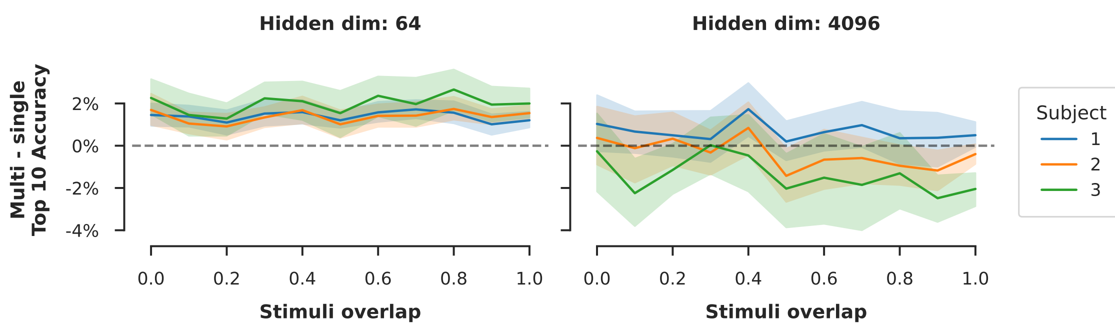

Figure 5: **Impact of training stimuli overlap.** We train multi-subject decoders while varying the ratio of overlapping stimuli between subjects. The graphics display the increment in accuracy over single-subject decoders for small (left) and large (right) decoders.

yielding lower accuracy than single-subject decoders. The effect is less pronounced for smaller models, but gains are marginal. This suggests that in our data regime (few subjects, deep phenotyping), the model struggles to overcome inter-subject variability, even if it generalizes to new stimuli for a given subject. Furthermore, we found that the degree of stimulus overlap between subjects during training has a negligible effect on multi-subject decoding performance (Figure 5), indicating that more target space diversity did not help the decoder generalize.

Decoders model both syntax and semantics We computed syntactic and semantic similarity between ground-truth text and retrieved candidates as a function of their embedding similarity rank to the decoder prediction (Figure 6).

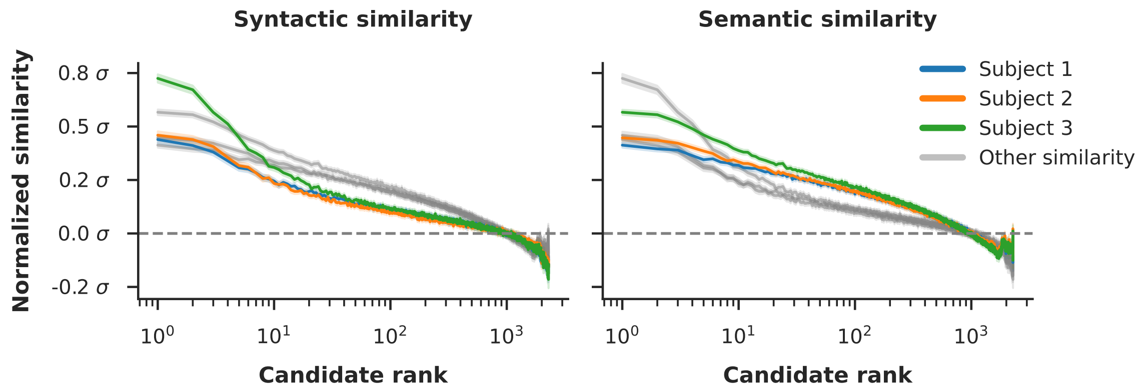

Figure 6: **Profiles of average syntactic/semantic similarities.** Syntactic (left) and semantic (right) similarities between ground-truth and candidate chunks, sorted by rank. Similarities are normalized by a random baseline.

Reassuringly, both similarities decrease as the rank of the candidate increases. This suggests that the decoders effectively model syntax and semantics and are not only relying on superficial features to achieve such performance. Also, the profile for syntactic similarity drops off more sharply than for semantic similarity, indicating the decoder is slightly better at differentiating syntactically dissimilar chunks than semantically dissimilar ones.

Simple text is easier to decode We analyzed the linguistic properties of well-decoded versus poorly-decoded texts. A qualitative analysis of the 10 best and worst decoded stories (see Table 2 in appendix) conducted with Gemini [Reid et al., 2024] revealed that stories with simple, conversational language, short sentences, and a direct narrative voice were decoded more accurately than those with complex syntax, formal language, and abstract ideas.

4 Discussion and Limitations

Our results achieve the highest top-10 retrieval accuracy reported for speech decoding and also provide clear insights for optimizing fMRI data acquisition with a small number of participants for this task. We demonstrate that a decoding-first approach with contrastive objectives can effectively predict LLM-derived text representations from fMRI data. Performance scales directly with the quantity of per-participant data without saturating, even with 13.5 hours of data.

Interestingly, multi-subject training failed to improve performance within our data set of eight individuals (with extensive data from only three), suggesting that inter-subject variability is too high for current decoders. This finding aligns with recent visual decoding findings, which indicate that increases in per-subject data outperform increases in subject count [Banville et al.]. This suggests that performance gains may require substantially larger cohorts or deeper phenotyping. Our finding that stimulus overlap between subjects has a minimal impact further supports the idea that inter-subject variability, rather than stimulus diversity, limits multi-subject performance.

Our analysis revealed that decoders leverage syntactic structure and semantic content, and they struggle more with linguistically complex stories. Stories written in simple, conversational language with short sentences were easier to decode than those with complex syntax and abstract ideas. This suggests the need for an investigation into training data biases and mitigation strategies.

Overall, our findings strongly suggest that collecting extensive data from a smaller number of participants is more effective than collecting less data from a larger number of subjects. This approach is especially beneficial when working with small participant groups, as is often the case in neuroimaging studies.

One limitation is that embeddings are predicted rather than text being reconstructed, though this avoids evaluation metrics dependent on generative models. fMRI's low temporal resolution remains a challenge. The main limitation is the quantity of the training data. To advance the boundaries of brain decoding, we advocate for larger datasets.

5 Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011016055) and benefited from state aid managed by the Agence Nationale de la Recherche ANR-24-CE17-1427 and innovation under the Specific Grant Agreement HORIZON-INFRA-2022-SERV-B-01.

References

E. J. Allen, G. St-Yves, Y. Wu, J. L. Breedlove, J. S. Prince, L. T. Dowdle, M. Nau, B. Caron, F. Pestilli, I. Charest, J. B. Hutchinson, T. Naselaris, and K. Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25(1):116–126, Jan. 2022. ISSN 1546-1726. doi: 10.1038/s41593-021-00962-x. URL `https://www.nature.com/articles/s41593-021-00962-x`.

H. Banville, Y. Benchetrit, S. d'Ascoli, J. Rapin, and J.-R. King. Scaling laws for decoding images from brain activity. doi: 10.48550/ARXIV.2501.15322. URL `https://arxiv.org/abs/2501.15322`.

P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados, and S. Reddy. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. 2024. URL `https://openreview.net/forum?id=IW1PR7vEBf#discussion`.

A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King. Decoding speech from non-invasive brain recordings, Oct. 2023. URL `http://arxiv.org/abs/2208.12266`.

O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. S. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. Dupre, M. Snyder, H. Oya, S. S. Ghosh, J. Wright, J. Durnez, R. A. Poldrack, and K. J. Gorgolewski. Fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16:111 – 116, 2018. URL `https://api.semanticscholar.org/CorpusID:54463721`.

E. Formisano, F. De Martino, M. Bonte, and R. Goebel. "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. *Science*, 322(5903):970–973, Nov. 2008. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1164318.

J. K. Ho, T. Horikawa, K. Majima, F. Cheng, and Y. Kamitani. Inter-individual deep image reconstruction via hierarchical neural code conversion. 271:120007, 2023. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2023.120007. URL `https://www.sciencedirect.com/science/article/pii/S1053811923001532`.

Y. Kamitani and F. Tong. Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5): 679–685, 2005.

A. LeBel, L. Wagner, S. Jain, A. Adhikari-Desai, B. Gupta, A. Morgenthal, J. Tang, L. Xu, and A. G. Huth. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10(1):555, Aug. 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02437-z. URL `https://www.nature.com/articles/s41597-023-02437-z`.

F. Ozcelik and R. VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion, June 2023. URL `http://arxiv.org/abs/2303.05334`.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision, Feb. 2021. URL `http://arxiv.org/abs/2103.00020`.

M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-B. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, I. Antonoglou, R. Anil, S. Borgeaud, A. M. Dai, K. Millican, E. Dyer, M. Glaese, T. Sottiaux, B.-j. Lee, F. Viola, M. Reynolds, Y. Xu, J. Molloy, J. Chen, M. Isard, P. Barham, T. Hennigan, R. McIlroy, M. Johnson, J. Schalkwyk, E. Collins, E. Rutherford, E. Moreira, K. W. Ayoub, M. Goel, C. Meyer, G. Thornton, Z. Yang, H. Michalewski, Z. Abbas, N. Schucher, A. Anand, R. Ives, J. Keeling, K. Lenc, S. Haykal, S. Shakeri, P. Shyam, A. Chowdhery, R. Ring, S. Spencer, E. Sezener, L. Vilnis, O. Chang, N. Morioka, G. Tucker, C. Zheng, O. Woodman, N. Attaluri, T. Kociský, E. Eltyshev, X. Chen, T. Chung, V. Selo, S. Brahma, P. Georgiev, A. Slone, Z. Zhu, J. Lottes, S. Qiao, B. Caine, S. Riedel, A. Tomala, M. Chadwick, J. C. Love, P. Choy, S. Mittal, N. Houlsby, Y. Tang, M. Lamm, L. Bai, Q. Zhang, L. He, Y. Cheng, P. Humphreys, Y. Li, S. Brin, A. Cassirer, Y.-Q. Miao, L. Zilka, T. Tobin, K. Xu, L. Proleev, D. Sohn, A. Magni, L. A. Hendricks, I. Gao, S. Ontan'on, O. Bunyan, N. Byrd, A. Sharma, B. Zhang, M. Pinto, R. Sinha, H. Mehta, D. Jia, S. Caelles, A. Webson, A. Morris, B. Roelofs, Y. Ding, R. Strudel, X. Xiong, M. Ritter, M. Dehghani, R. Chaabouni, A. Karmarkar, G. Lai, F. Mentzer, B. Xu, Y. Li, Y. Zhang, T. Paine, A. Goldin, B. Neyshabur, K. Baumli, A. Levskaya, M. Laskin, W. Jia, J. W. Rae, K. Xiao, A. He, S. Giordano, L. Yagati, J.-B. Lespiau, P. Natsev, S. Ganapathy, F. Liu, D. Martins, N. Chen, Y. Xu, M. Barnes, R. May, A. Vezer, J. Oh, K. Franko, S. Bridgers, R. Zhao, B. Wu, B. Mustafa, S. Sechrist, E. Parisotto, T. S. Pillai, C. Larkin, C. Gu, C. Sorokin, M. Krikun, A. Guseynov, J. Landon, R. Datta, A. Pritzel, P. Thacker, F. Yang, K. Hui, A. E. Hauth, C.-K. Yeh, D. Barker, J. Mao-Jones, S. Austin, H. Sheahan, P. Schuh, J. Svensson, R. Jain, V. Ramasesh, A. Briukhov, D. Chung, T. von Glehn, C. Butterfield, P. Jhakra, M. Wiethoff, J. Frye, J. Grimstad, B. Changpinyo, C. L. Lan, A. Bortsova, Y. Wu, P. Voigtlaender, T. N. Sainath, C. Smith, W. Hawkins, K. Cao, J. Besley, S. Srinivasan, M. Omernick,

C. Gaffney, G. Surita, R. Burnell, B. Damoc, J. Ahn, A. Brock, M. Pajarskas, A. Petrushkina, S. Noury, L. Blanco, K. Swersky, A. Ahuja, T. Avrahami, V. Misra, R. de Liedekerke, M. Iinuma, A. Polozov, S. York, G. van den Driessche, P. Michel, J. Chiu, R. Blevins, Z. Gleicher, A. Recasens, A. Rrustemi, E. Gribovskaya, A. Roy, W. Gworek, S. M. R. Arnold, L. Lee, J. Lee-Thorp, M. Maggioni, E. Piqueras, K. Badola, S. Vikram, L. Gonzalez, A. Baddepudi, E. Senter, J. Devlin, J. Qin, M. Azzam, M. Trebacz, M. Polacek, K. Krishnakumar, S.-y. Chang, M. Tung, I. Penchev, R. Joshi, K. Olszewska, C. Muir, M. Wirth, A. J. Hartman, J. Newlan, S. Kashem, V. Bolina, E. Dabir, J. R. van Amersfoort, Z. Ahmed, J. Cobon-Kerr, A. B. Kamath, A. M. Hrafnkelsson, L. Hou, I. Mackinnon, A. Frechette, E. Noland, X. Si, E. Taropa, D. Li, P. Crone, A. Gulati, S. Cevey, J. Adler, A. Ma, D. Silver, S. Tokumine, R. Powell, S. Lee, M. B. Chang, S. Hassan, D. Mincu, A. Yang, N. Levine, J. Brennan, M. Wang, S. Hodkinson, J. Zhao, J. Lipschultz, A. Pope, M. B. Chang, C. Li, L. E. Shafey, M. Paganini, S. Douglas, B. Bohnet, F. Pardo, S. Odoom, M. Rosca, C. N. dos Santos, K. Soparkar, A. Guez, T. Hudson, S. Hansen, C. Asawaroengchai, R. Addanki, T. Yu, W. Stokowiec, M. Khan, J. Gilmer, J. Lee, C. G. Bostock, K. Rong, J. Caton, P. Pejman, F. Pavetic, G. Brown, V. Sharma, M. Luvci'c, R. Samuel, J. Djolonga, A. Mandhane, L. L. Sjosund, E. Buchatskaya, E. White, N. Clay, J. Jiang, H. Lim, R. Hemsley, J. Labanowski, N. D. Cao, D. Steiner, S. H. Hashemi, J. Austin, A. Gergely, T. Blyth, J. Stanton, K. Shivakumar, A. Siddhant, A. Andreassen, C. L. Araya, N. Sethi, R. Shivanna, S. Hand, A. Bapna, A. Khodaei, A. Miech, G. Tanzer, A. Swing, S. Thakoor, Z. Pan, Z. Nado, S. Winkler, D. Yu, M. Saleh, L. Maggiore, I. Barr, M. Giang, T. Kagohara, I. Danihelka, A. Marathe, V. Feinberg, M. Elhawaty, N. Ghelani, D. Horgan, H. Miller, L. Walker, R. Tanburn, M. Tariq, D. Shrivastava, F. Xia, C.-C. Chiu, Z. Ashwood, K. Baatarsukh, S. Samangooei, F. Alcober, A. Stjerngren, P. Komarek, K. Tsihlas, A. Boral, R. Comanescu, J. Chen, R. Liu, D. Bloxwich, C. Chen, Y. Sun, F.-a. Feng, M. Mauger, X. Dotiwalla, V. Hellendoorn, M. Sharman, I. Zheng, K. Haridasan, G. Barth-Maron, C. Swanson, D. Rogozi'nska, A. Andreev, P. Rubenstein, R. Sang, D. Hurt, G. Elsayed, R.-s. Wang, D. Lacey, A. Ili'c, Y. Zhao, W. Han, L. Aroyo, C. Iwuanyanwu, V. Nikolaev, B. Lakshminarayanan, S. Jazayeri, R. L. Kaufman, M. Varadarajan, C. Tekur, D. Fritz, M. Khalman, D. Reitter, K. Dasgupta, S. Sarcar, T. Ornduff, J. Snaider, F. Huot, J. Jia, R. Kemp, N. Trdin, A. Vijayakumar, L. Kim, C. Angermueller, L. Lao, T. Liu, H. Zhang, D. Engel, S. Greene, A. White, J. Austin, L. Taylor, S. Ashraf, D. Liu, M. Georgaki, I. Cai, Y. Kulizhskaya, S. Goenka, B. Saeta, K. Vodrahalli, C. Frank, D. Cesare, B. Robenek, H. Richardson, M. Alnahlawi, C. Yew, P. Ponnapalli, M. Tagliasacchi, A. Korchemniy, Y. Kim, D. Li, B. Rosgen, K. Levin, J. Wiesner, P. Banzal, P. Srinivasan, H. Yu, c. Unlu, D. Reid, Z. Tung, D. Finchelstein, R. Kumar, A. Elisseeff, J. Huang, M. Zhang, R. Zhu, R. Aguilar, M. Gim'enez, J. Xia, O. Dousse, W. Gierke, S. Yeganeh, D. Yates, K. Jalan, L. Li, E. Latorre-Chimoto, D. D. Nguyen, K. Durden, P. Kallakuri, Y. Liu, M. Johnson, T. Tsai, A. Talbert, J. Liu, A. Neitz, C. Elkind, M. Selvi, M. Jasarevic, L. B. Soares, A. Cui, P. Wang, A. W. Wang, X. Ye, K. Kallarackal, L. Loher, H. Lam, J. Broder, D. Holtmann-Rice, N. Martin, B. Ramadhana, D. Toyama, M. Shukla, S. Basu, A. Mohan, N. Fernando, N. Fiedel, K. Paterson, H. Li, A. Garg, J. Park, D. Choi, D. Wu, S. Singh, Z. Zhang, A. Globerson, L. Yu, J. Carpenter, F. D. C. Quitry, C. Radebaugh, C.-C. Lin, A. Tudor, P. Shroff, D. Garmon, D. Du, N. Vats, H. Lu, S. Iqbal, A. Yakubovich, N. Tripuraneni, J. Manyika, H.-r. Qureshi, N. Hua, C. Ngani, M. A. Raad, H. Forbes, A. Bulanova, J. Stanway, M. Sundararajan, V. Ungureanu, C. Bishop, Y. Li, B. Venkatraman, B. Li, C. Thornton, S. Scellato, N. Gupta, Y. Wang, I. Tenney, X. Wu, A. Shenoy, G. Carvajal, D. G. Wright, B. Bariach, Z. Xiao, P. Hawkins, S. Dalmia, C. Farabet, P. Valenzuela, Q. Yuan, C. A. Welty, A. Agarwal, M. Chen, W. Kim, B. Hulse, N. Dukkipati, A. Paszke, A. Bolt, E. Davoodi, K. Choo, J. Beattie, J. Prendki, H. Vashisht, R. Santamaria-Fernandez, L. C. Cobo, J. Wilkiewicz, D. Madras, A. Elqursh, G. Uy, K. Ramirez, M. Harvey, T. Liechty, H. Zen, J. Seibert, C. H. Hu, A. Y. Khorlin, M. Le, A. Aharoni, M. Li, L. Wang, S. Kumar, A. Lince, N. Casagrande, J. Hoover, D. E. Badawy, D. Soergel, D. Vnukov, M. Miecnikowski, J. Šimša, A. Koop, P. Kumar, T. Sellam, D. Vlasic, S. Daruki, N. Shabat, J. Zhang, G. Su, K. Krishna, J. Zhang, J. Liu, Y. Sun, E. Palmer, A. Ghaffarkhah, X. Xiong, V. Cotruta, M. Fink, L. Dixon, A. Sreevatsa, A. Goedeckemeyer, A. Dimitriev, M. Jafari, R. Crocker, N. Fitzgerald, A. Kumar, S. Ghemawat, I. Philips, F. Liu, Y. Liang, R. Sterneck, A. Repina, M. Wu, L. Knight, M. Georgiev, H. Lee, H. Askham, A. Chakladar, A. Louis, C. Crous, H. Cate, D. Petrova, M. Quinn, D. Owusu-Afriyie, A. Singhal, N. Wei, S. Kim, D. Vincent, M. Nasr, I. Shumailov, C. A. Choquette-Choo, R. Tojo, S. Lu, D. d. L. Casas, Y. Cheng, T. Bolukbasi, K.-i. Lee, S. Fatehi, R. Ananthanarayanan, M. Patel, C. Kaed, J. Li, J. Sygnowski, S. Belle, Z. Chen, J. Konzelmann, S. Põder, R. Garg, V. Koverkathu, A. Brown, C. Dyer, R. Liu, A. Nova, J. Xu, J. Bai, S. Petrov, D. Hassabis, K. Kavukcuoglu, J. Dean, O. Vinyals, and A. Chronopoulou. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, Mar. 2024.

P. S. Scotti, A. Banerjee, J. Goode, S. Shabalin, A. Nguyen, E. Cohen, A. J. Dempster, N. Verlinde, E. Yundler, D. Weisberg, K. A. Norman, and T. M. Abraham. Reconstructing the Mind's Eye: fMRI-to-Image with Contrastive Learning and Diffusion Priors, May 2023. URL `http://arxiv.org/abs/2305.18274`.

P. S. Scotti, M. Tripathy, C. K. T. Villanueva, R. Kneeland, T. Chen, A. Narang, C. Santhirasegaran, J. Xu, T. Naselaris, K. A. Norman, and T. M. Abraham. MindEye2: Shared-Subject Models Enable fMRI-To-Image With 1 Hour of Data, Mar. 2024. URL `http://arxiv.org/abs/2403.11207`.

J. Tang, A. LeBel, S. Jain, and A. 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.

A. Thual, Y. Benchetrit, F. Geilert, J. Rapin, I. Makarov, H. Banville, and J.-R. King. Aligning brain functions boosts the decoding of visual semantics in novel subjects, 2023. URL `http://arxiv.org/abs/2312.06467`.

Z. Ye, Q. Ai, Y. Liu, M. de Rijke, M. Zhang, C. Lioma, and T. Ruotsalo. Generative language reconstruction from brain recordings. *Communications Biology*, 8(1):346, 2025.

Contrastive Loss

We use the symmetric InfoNCE loss, a contrastive loss function popularized by CLIP [Radford et al., 2021]. Given a batch of N pairs of predicted text embeddings from fMRI $\hat{\boldsymbol{y}}_i$ and ground-truth text embeddings $\boldsymbol{y}_i$, the loss is defined as:

$$\mathcal{L} = -\frac{1}{2N} \sum_{i=1}^{N} \left(\log \frac{\exp(s(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i)/\tau)}{\sum_{j=1}^{N} \exp(s(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_j)/\tau)} + \log \frac{\exp(s(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i)/\tau)}{\sum_{j=1}^{N} \exp(s(\hat{\boldsymbol{y}}_j, \boldsymbol{y}_i)/\tau)} \right) \tag{1}$$

where $s(\cdot, \cdot)$ is the cosine similarity and τ is a learnable temperature parameter. The first term inside the sum is the loss for predicting the correct text embedding given an fMRI-derived embedding, among all text embeddings in the batch. The second term is the loss for predicting the correct fMRI-derived embedding for a given text embedding.

Supplementary Tables

Hyperparameter	Value	Description
		Data Configuration
Context Length	3	Number of preceding text chunks concatenated (= 6s)
Temporal Smoothing	4	Number of preceding brain volumes averaged (= 8s)
Lag (s)	6	Delay between a brain volume and the target chunk
Top Encoding Voxels	4096	Number of voxels selected based on encoding performance
		Brain Decoder Configuration
Hidden Dimension	64 or 4096	Size of the hidden layers (both are considered in the experiments)
Number of Linear Layers	1	
Number of Residual Layers	1	
Total number of layers	3	Subject specific layer + Linear + Residual
Normalization Type	Layer	
Activation Function	GELU	
		Training Configuration
Temperature	0.7	Temperature of the contrastive loss
Learning Rate	1e-4	
Weight Decay	5e-4	
Patience	20	Number of epochs with no improvement before early stopping
Scheduler Patience	5	Number of epochs with no improvement before reducing the learning rate
Scheduler Factor	0.5	Factor by which the learning rate is reduced
Batch Size	1	Number of stories per batch
Max Epochs	200	

Table 1: **Hyperparameters of the DNNs** Comprehensive list of the hyperparameters used in the study. They were determined through a Bayesian grid search maximizing cross-validated top-10 accuracy on subjects 1, 2, and 3 in the multi-subject setup. In this setup small decoders (hidden dimension 64) have 250k parameters in their backbone and 250k (per subject) in the subject-specific layers. Large decoders (hidden dimension 4096) have 7e7 in addition to 1.7e7 (per subject) parameters.

Group	10 best decoded stories	10 worst decoded stories
Feature	Personal experience	Ideas & Reflection
Performance	Top-10 accuracy: mean 51%, min 44%, max 59%	Top-10 accuracy: mean 15%, min 5%, max 21%
Narrative Voice	Conversational, informal, direct address ("you know," "I mean")	More formal, polished, less direct address
Sentence Length	Primarily short, choppy sentences for immediacy and emotional impact	Longer, more complex sentences reflecting intellectual exploration
Vocabulary	Colloquial language, contractions, slang	Wider range of vocabulary, less colloquialism, more sophisticated phrasing
Use of Dialogue	Frequent, natural-sounding dialogue to advance the narrative and express emotion	Dialogue used, but less prevalent and often serves an illustrative function
Sentence Structure	Loose sentence structure with emphasis on emotion and immediacy	More complex sentence structures, often with clauses and sub-clauses
Tone	Immediate, intimate, emotionally charged, often reflecting raw feelings and vulnerability	Reflective, analytical, measured, with a more intellectual distance and a focus on conveying insights
Descriptive Language	Emphasis on sensory details and vivid imagery.	More use of metaphor and simile to add a higher level of detail and comparison to the text
Rhetorical Questions	Frequent use of rhetorical questions	Less use of rhetorical questions
Emphasis	Primarily to drive the narrative with a strong focus on the characters' internal experiences	Primarily to reflect, analyze and offer insights
Stories	kiksuya, itsabox, thefreedomridersandme, comingofageondeathrow, hangtime, liferеimagined, cautioneating, thatthingonmyarm, fromboyhoodtofatherhood, threemonths	notontheusualtour, breakingupintheageofgoogle, jugglingandjesus, theadvancedbeginner, alternateithicatom, forgettingfear, avatar, theshower, treasureisland, bluehope

Table 2: **Qualitative analysis of the impact of semantics and syntax** We use Gemini [Reid et al., 2024] to analyze the commonalities and differences between the 10 best and 10 worst performing stories in terms of semantics and syntax. Stories are sorted by the accuracy obtained from subject 3 and we display the mean/min/max accuracy for each group of stories.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately list the five main contributions of the paper. These claims are directly addressed and supported by the figures and analysis presented in the Results section.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Discussion and Limitations" section. It addresses limitations such as the small number of subjects, the model's bias towards syntax, the prediction of embeddings instead of text, and the low temporal resolution of fMRI.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper is empirical and does not contain theoretical results, theorems, or mathematical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper details the dataset, preprocessing steps, model architecture, training procedures, and evaluation metrics. Table 1 in the appendix provides a comprehensive list of hyperparameters, which should be sufficient for other researchers to reproduce the experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The assets used for this paper are publicly available: the fMRI dataset from LeBel et al. [2023] and the LLM2Vec model from BehnamGhader et al. [2024]. The code will be made available upon publication.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings are thoroughly described. The Methods section covers data splits and evaluation, while Table 1 in the appendix details hyperparameters which were selected with Bayesian search.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports 95% confidence intervals calculated across cross-validation folds. This is mentioned in the "Evaluation" paragraph of the Methods section and visualized in the result figures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The paper does not provide information on the computational resources (e.g., GPU type, memory, execution time) because they were not tracked during the experiments. However the project only required a single GPU with 40GB of memory.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research was conducted on a pre-existing, de-identified dataset, and we have adhered to the NeurIPS Code of Ethics throughout the research and writing process.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not contain a broader impacts section because of page constraints. While the work has positive implications for clinical applications, it does not discuss potential negative societal impacts, such as privacy concerns related to brain decoding technology.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release any new high-risk models or data. The research is based on a pre-existing dataset and the models are not being released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper properly credits the creators of the existing assets: the fMRI dataset from LeBel et al. [2023] and the language model from BehnamGhader et al. [2024]. However, it does not explicitly state the licenses or terms of use for these assets because of page constraints. The fMRI dataset is released under a CC0 license, and the LLM2Vec model is released under a MIT license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not introduce or release any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: This study uses a pre-existing dataset. The paper does not include details on participant instructions or compensation, which can be found in the original publication [LeBel et al., 2023].

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: This study uses a pre-existing dataset. The paper does not restate the IRB approval details from the original study, but refers to the original publication [LeBel et al., 2023] where this information is available.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper's core methodology relies on predicting LLM-derived text representations, and it explicitly states the use of LLM2Vec for this purpose in the Methods section. It also declares the use of Gemini for the qualitative analysis in the appendix (Table 2).

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.