Abstract

Understanding how the brain represents language under varying cognitive loads is a central goal in cognitive neuroscience. Using fMRI data from Tuckute et al. (2024), we investigated this by training linear models to map between brain activity across six language-network ROIs and contextual sentence embeddings from BERT. We compared a low-effort "Suppressing" condition (simple, coherent sentences) with a high-effort "Driving" condition (often nonsensical sentences). Contrary to our initial hypothesis, we found that the high-effort "Driving" condition was not decodable using a general-purpose language decoder, performing at chance level. In contrast, the low-effort "Suppressing" condition was decodable, with performance consistently better than chance. A layer-wise decoding analysis further revealed that this accuracy for the "Suppressing" condition peaked at the middle layers of BERT (Layer 8), suggesting an alignment with compositional semantic representations. These results indicate that cognitive effort from processing incoherent stimuli does not produce a more robust semantic representation, but rather a distinct neural state inconsistent with standard models of language comprehension.[1]

1 Introduction

Mapping the neural basis of language is a primary objective in cognitive neuroscience. A powerful approach in this domain involves building computational models that aim to find a mapping between the brain's response to linguistic stimuli and the representations generated by artificial neural networks. These efforts largely fall into two complementary paradigms: encoding and decoding. Encoding models attempt to predict neural activity from the features of a stimulus, answering the question of *where* in the brain different types of information are represented (Huth et al., 2016). Conversely, decoding models attempt to predict the features of a stimulus from neural activity, answering the question of *what* information is robustly represented in the brain's overall state (Pereira et al., 2018).

The advent of deep language models, particularly large-scale Transformers like BERT (Devlin et al., 2019), has revolutionized this field. These models, pre-trained on vast text corpora, learn rich, contextual representations of language that have proven to be highly effective predictors of neural activity. This synergy allows researchers to move beyond simple linguistic features and test more nuanced hypotheses about how the brain processes complex, context-dependent meaning.

An open question in this field is how the brain's representations are affected by cognitive load. Theories of effortful processing suggest that when the brain works harder, it may engage neural resources more intensely -meaning with greater magnitude and a more specific activation pattern- potentially leading to a clearer neural representation. The study by Tuckute et al. (2024) provides a unique paradigm to investigate this, using language models to generate stimuli that either "drive" (elicit strong fMRI activation) or "suppress" (elicit weak activation) the language network. As confirmed by our own data exploration, these conditions correspond to high-effort processing of often nonsensical sentences and low-effort processing of simple, coherent sentences, respectively.

This project first establishes a methodological baseline through a series of structured tasks adapted from Pereira et al. (2018), comparing static and contextual embeddings. We then apply these methods to investigate the core cognitive question: does the increased neural effort associated with the "Driving" condition produce a more decodable neural signal? While we initially hypothesized that it would, our main finding is the opposite: the high-effort condition was undecodable, while the low-effort condition was successfully decoded, with accuracy peaking at the middle layers of BERT. This suggests that the nature of cognitive effort, rather than its magnitude alone, is the critical factor for forming a coherent neural representation of language.

2 Data and Methods

2.1 Datasets

Pereira et al. (2018) For our structured tasks, we used the publicly available data from Pereira et al. (2018). This includes fMRI data from subjects reading 180 distinct sentences, each related to a single word-concept. The high-dimensional data consists of 170,712 voxels per sentence. We also used the corresponding 300-dimensional GloVe vectors.

Tuckute et al. (2024) Our primary dataset for the open-ended task was sourced from Tuckute et al. (2024). It contains fMRI responses from 10 subjects to 1500 unique sentences. We pre-processed

[1] https://github.com/your-username/your-repo-link-here

this data by averaging responses across subjects and pivoting the data to create a matrix of 6 predefined language-network Regions of Interest (ROIs). The stimuli are categorized into three conditions: a "Baseline" set of 1000 natural sentences, a "Driving" set of 250 nonsensical sentences, and a "Suppressing" set of 250 simple, coherent sentences.

Condition	Example Sentence
Baseline	Sunday School classes begin at 9:30am.
Suppressing	We could play a card game.
Driving	Lack tose and tolerant on Yahoo!

Table 1: Example sentences from the three experimental conditions in the Tuckute et al. (2024) dataset.

2.2 Semantic Embeddings

We utilized three types of semantic embeddings:

- **GloVe**: 300-dimensional static vectors (Pennington et al., 2014).
- **Word2Vec**: 300-dimensional static vectors (Mikolov et al., 2013).
- **BERT**: 768-dimensional contextual vectors from the pre-trained `bert-base-uncased` model (Devlin et al., 2019). The final hidden state of the [CLS] token was used as the sentence-level vector. For our layer-wise analysis, we additionally extracted representations from all 12 layers.

2.3 Experimental Design

Our initial design intended to train a general-purpose decoder on the high-dimensional voxel data from Pereira et al. and test it on the Tuckute et al. data. This was unfeasible due to an input feature dimensionality mismatch (170,712 vs. 6). We therefore adopted a robust intra-dataset design, training our linear models exclusively on the large **"Baseline"** condition (1000 sentences) and testing on the held-out **"Driving"** and **"Suppressing"** conditions. This approach leverages the dataset's own control condition for training, allowing for a valid comparison of the two experimental conditions.

2.4 Evaluation Metrics

- **Rank Accuracy (Decoding):** Performance was measured by the rank of the true sentence vector when compared against all candidate vectors in its condition via cosine similarity. A lower average rank indicates better performance.
- **Pearson Correlation (Encoding):** Performance was measured by the Pearson correlation between the predicted 6-ROI activity pattern and the true pattern for each sentence. A higher average correlation indicates better performance.

3 Experiments and Results

3.1 Structured Tasks

We first performed a series of structured tasks to replicate and extend the findings of Pereira et al. (2018), establishing a solid methodological baseline for our main experiment.

3.1.1 Task 1: Word-Level Decoding

We began by comparing the performance of static GloVe and Word2Vec embeddings for decoding single-word concepts from the Pereira fMRI data. While the GloVe-based decoder achieved a slightly better overall average rank (57.36) than Word2Vec (58.42), Figure 1 reveals a critical difference in reliability. The GloVe model's performance is highly volatile across the 18 cross-validation folds, whereas the Word2Vec model is remarkably more stable, consistently performing well below the chance level.

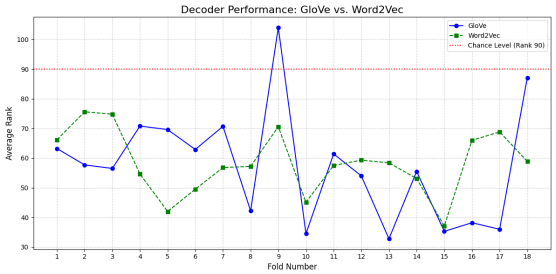

Figure 1: Decoder performance for static embeddings across 18 folds. **Conclusion:** While GloVe achieves a slightly better overall average rank, it is highly volatile. The Word2Vec-based decoder is more stable.

An analysis of the best and worst-decoded concepts (Table 2) revealed that both models performed best on frequent, concrete words and failed on abstract or morphologically complex words, suggesting that the distinctiveness of a concept's neural signature is key to its decodability.

We then tested the decoder's ability to generalize to full sentences. As shown in Figure 2, the decoder performed significantly better than chance on both

Model	Top 3 Best	Top 3 Worst
GloVe	do (1.00) food (1.00) time (2.00)	argumentatively (180.0) cockroach (176.0) applause (171.0)
Word2Vec	laugh (1.00) soul (1.00) stupid (1.00)	argumentatively (168.0) movie (148.0) tried (144.0)

Table 2: Best and worst decoded word concepts for static models (average rank in parentheses). **Conclusion:** Both models excel at concrete words and struggle with abstract ones.

Experiment 2 (avg. rank = 150.02 vs. chance 192) and Experiment 3 (avg. rank = 96.75 vs. chance 121.5) from Pereira. Performance was notably better for Experiment 3, suggesting that sentences from cohesive narratives elicit more decodable neural patterns. A topic-wise analysis (Table 3) confirmed that concrete topics (e.g., `body_part`) were more decodable than broad, abstract ones (e.g., `profession`).

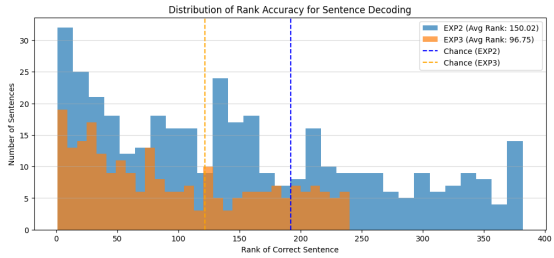

Figure 2: Distribution of rank accuracy for sentence decoding. **Conclusion:** The GloVe-based decoder performs better than chance (dashed lines) for both experiments, with superior accuracy on the cohesive narratives of Experiment 3.

Dataset	Top 3 Best Topics	Top 3 Worst Topics
Exp. 2	body_part (83.3) drink (87.9) human (94.9)	profession (233.3) vegetable (203.6) vehicles (191.8)
Exp. 3	opera (45.6) dreams (51.9) bone_fracture (57.4)	beekeeping (165.4) owl (156.6) pyramid (148.6)

Table 3: Most and least decodable topics from Experiments 2 and 3 (average rank in parentheses). **Conclusion:** The decoder excels at concrete or evocative topics and struggles with broad, abstract ones.

3.1.2 Task 2: Sentence Representation Decoding

To directly compare static versus contextual sentence representations, we trained decoders on Pereira's Experiment 3 data using both GloVe and BERT embeddings. As shown in Figure 3, the contextual BERT-based decoder achieved a significantly better average rank (60.10) than the static GloVe-based decoder (64.13), confirming the superiority of contextual embeddings for mapping to neural activity.

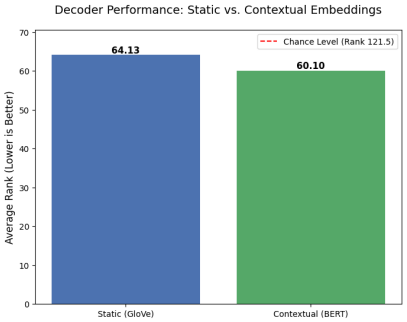

Figure 3: Decoder performance comparing static (GloVe) and contextual (BERT) sentence embeddings on Pereira's Experiment 3 data. **Conclusion:** The BERT-based decoder performs significantly better, demonstrating the advantage of contextual representations.

3.1.3 Task 3: Neural Encoding

Finally, we performed a neural encoding analysis to predict fMRI activity from sentence embeddings. Figure 4 shows the distribution of R^2 scores for both GloVe and BERT-based models. The GloVe model failed to predict any voxels above chance. In contrast, the BERT model successfully predicted activity in a subset of 30 voxels with an R^2 score greater than 0.03, further establishing the richer representational capacity of contextual models.

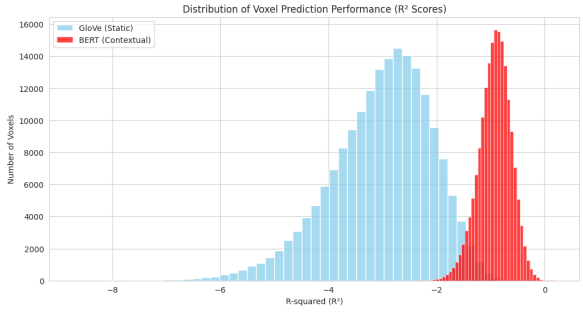

Figure 4: Distribution of voxel prediction performance (R^2) for encoding models. The plots show the full distribution and a zoomed-in view of positive R^2 scores. **Conclusion:** Contextual BERT embeddings can successfully predict voxel-level fMRI activity, while static GloVe embeddings cannot.

3.2 Open-Ended Task: Effect of Cognitive Load

Our open-ended task leveraged the validated methods to investigate the effect of cognitive load using the Tuckute et al. (2024) dataset. Our initial exploration confirmed that "Driving" stimuli elicit a strong positive fMRI response on average, while "Suppressing" stimuli elicit a negative response (see Appendix A), confirming they represent high and low effort states, respectively.

3.2.1 Decoding Fails for High-Effort Stimuli

Our main experiment tested whether the high-effort "Driving" condition was more decodable. We trained an unbiased decoder on the 1000 "Baseline" sentences and evaluated its performance on the final BERT layer embeddings for the "Driving" and "Suppressing" conditions. The results, shown in Figure 5, refuted our hypothesis. The decoder performed better for the low-effort "Suppressing" condition (avg. rank = 119.58) than for the "Driving" condition (avg. rank = 123.71). While "Suppressing" accuracy was better than chance (125), the "Driving" condition's performance was not.

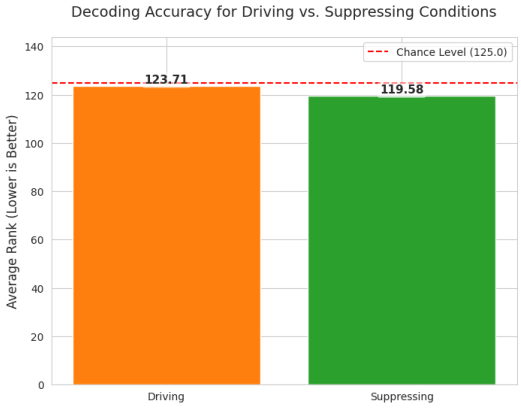

Figure 5: Decoding accuracy for Driving vs. Suppressing conditions using the final BERT layer. **Conclusion:** A general-purpose decoder performs better than chance (125) only for the low-effort Suppressing condition.

3.2.2 Initial Layer-wise Analysis Reveals a Representational Peak

To investigate this result, we first performed a targeted layer-wise decoding analysis using representative early (Layer 4), middle (Layer 8), and late (Layer 12) BERT layers. The results (Figure 6) showed a clear interaction. For the "Suppressing" condition, decoding accuracy peaked at the middle layers (avg. rank = 115.70 at Layer 8), suggesting an alignment with compositional semantic representations. The "Driving" condition remained undecodable at all levels.

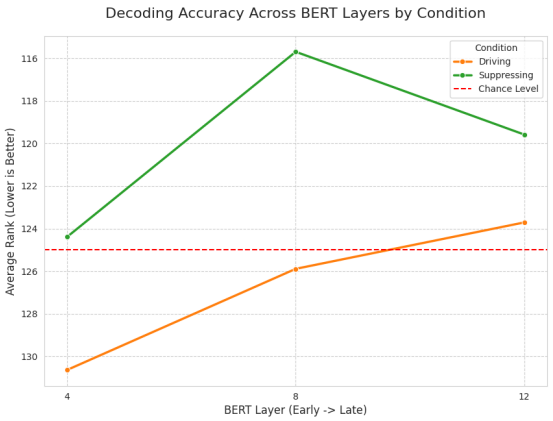

Figure 6: Initial decoding accuracy across three representative BERT layers. **Conclusion:** Accuracy for "Suppressing" sentences peaks at the middle layers.

3.2.3 Expansion: Full 12-Layer Sweep Confirms and Refines the Hierarchy

To gain a higher-resolution view of this processing hierarchy, we expanded our analysis to include all 12 layers of BERT. Table 4 provides a summary of the results, while Figure 7 offers a visual look at the full trend. This detailed analysis confirms that the "Driving" condition is undecodable across the entire model, while the "Suppressing" condition shows a clear U-shaped pattern, with peak performance at **Layer 10** (avg. rank = 115.48).

BERT Layer	Driving Rank	Suppressing Rank
1	133.08	126.40
2	136.87	119.01
3	137.57	122.78
4	130.64	124.39
5	131.68	124.55
6	129.37	124.21
7	124.33	116.87
8	125.89	115.70
9	123.46	115.55
10	121.74	**115.48**
11	122.91	117.00
12	123.71	119.58

Table 4: Average rank accuracy across all 12 BERT layers. The best performance (lowest rank) is bolded.

3.2.4 Expansion: Encoding and Within-Condition Analyses

To further characterize the neural signals, we conducted two final analyses. First, a complementary encoding analysis showed the "Suppressing" condition was more predictable (avg. correlation =

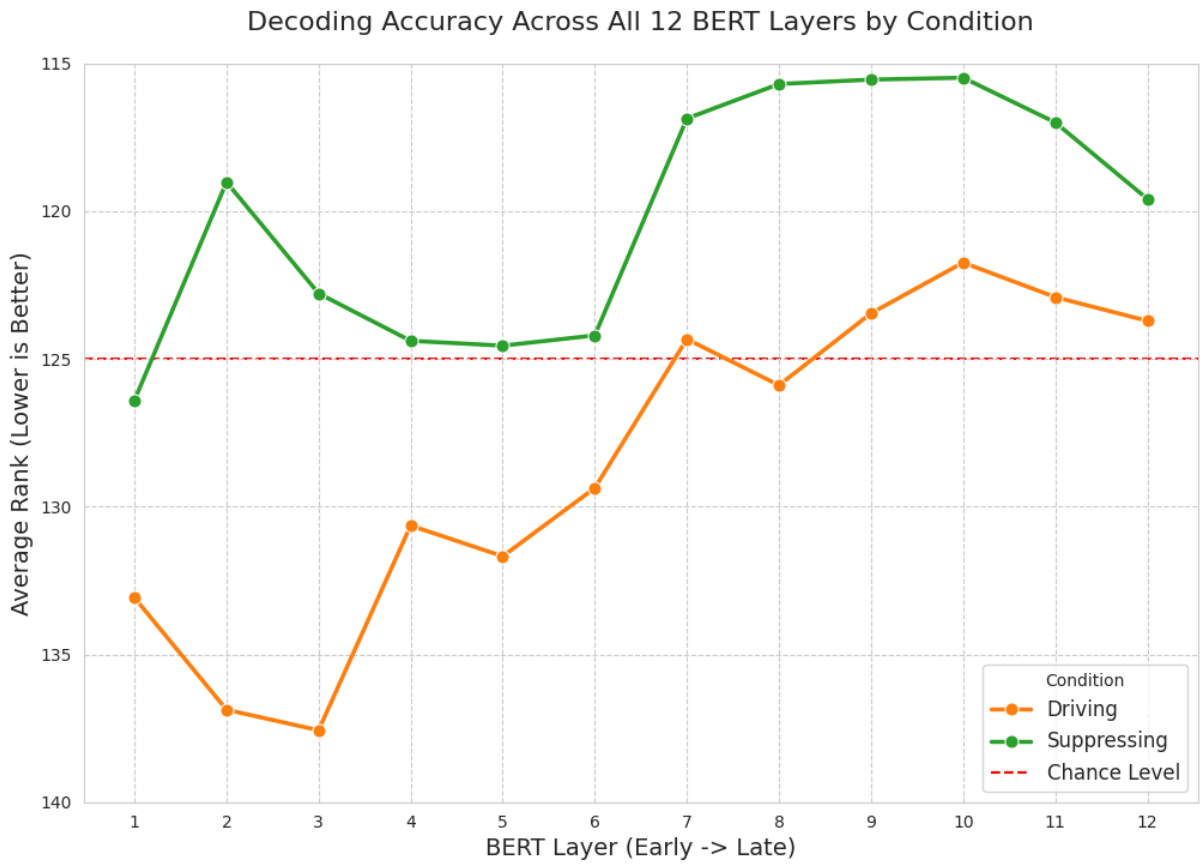

Figure 7: Full decoding accuracy sweep across all 12 BERT layers. **Conclusion:** The U-shaped curve for "Suppressing" sentences confirms a peak at the late-middle layers (specifically Layer 10), while "Driving" sentences remain at or below chance.

0.035) than the near-chance "Driving" condition (avg. correlation = 0.004), although this difference was not statistically significant ($p = 0.52$).

Second, to investigate if the "Driving" signal was simply noise, we ran a within-condition decoding analysis. The results are summarized in Table 5. Surprisingly, a decoder trained and tested only on "Driving" data performed better than chance (avg. rank = 117.05), similar to the "Suppressing" condition (avg. rank = 120.76). This indicates the neural signal for incoherent stimuli is consistent, but distinct from the neural patterns of normal comprehension.

Condition	Avg. Rank	Chance Level
Suppressing	120.76	125.0
Driving	117.05	125.0

Table 5: Within-condition decoding accuracy. **Conclusion:** Both conditions are decodable better than chance when trained on their own data, suggesting the "Driving" signal is internally consistent.

4 Discussion and Conclusion

Our project successfully implemented and evaluated a series of neural decoding and encoding models to investigate the brain's representation of language. The structured tasks provided a solid methodological foundation, replicating prior work and confirming the superiority of contextual embeddings (BERT) over static embeddings for mapping to neural activity.

Our primary finding from the open-ended task is that high cognitive effort, when induced by nonsensical stimuli, does not lead to a more decodable semantic representation. Our initial hypothesis was refuted; both decoding and encoding models performed significantly worse on the high-effort "Driving" condition when trained on a general-purpose baseline. This suggests a critical dissociation between the *magnitude* of neural activity—which, as shown in Appendix A, was higher for "Driving" stimuli—and the *coherence* of the underlying linguistic representation.

The surprising success of within-condition de-

coding for "Driving" stimuli adds a crucial layer of nuance to this interpretation. This analysis tested whether the signal was internally consistent by training and testing a decoder only on "Driving" data. The result that it performed better than chance (avg. rank = 117.05) suggests the brain enters a consistent and learnable "error state" when encountering incoherent language. This state, however, appears to exist in a different neural subspace from normal comprehension, as a decoder trained on "Baseline" sentences was unable to interpret it. The brain's effort in this context likely reflects processes like error-correction and parsing failure rather than the formation of a robust meaning.

Furthermore, the peak decoding accuracy for coherent ("Suppressing") sentences at the middle-to-late layers of BERT (peaking at Layer 10) is a significant finding. It suggests that the aggregate neural state across the language network aligns best with compositional semantic representations, where syntactic structure and meaning are actively being integrated, rather than with earlier syntactic or later, more abstract semantic layers.

Limitations and Future Work Several limitations should be considered when interpreting these results. Our initial, preferred experimental design was to train our decoder on the high-dimensional voxel data from Pereira et al. (2018) to create a truly unbiased, cross-dataset model. However, this was unfeasible due to a fundamental dimensionality mismatch between the training data (170,712 voxels) and the available test data from Tuckute et al. (2024), which was pre-processed into 6 ROI-level features.

This forced us to adopt an intra-dataset training paradigm (training on the "Baseline" condition). While this is a valid approach, it is possible that the "Baseline" condition, consisting of natural text, provided a training signal that was more similar to the "Suppressing" condition than the "Driving" condition, potentially contributing to the performance difference.

Moreover, the dimensionality of our input brain data (6 ROIs) is very low compared to the dimensionality of the target BERT embeddings (768). This makes the decoding task inherently challenging and likely explains why the overall rank accuracies, while better than chance, were not stronger. A more powerful analysis could be performed on the raw, voxel-level data, but processing and aligning such high-dimensional data was beyond the scope and resources of this project. Future work could leverage more advanced neuroimaging analysis pipelines to create a richer feature space from the raw fMRI data.

Finally, our analyses relied on linear models. Future studies could explore non-linear models (e.g., neural networks) to capture more complex relationships between brain activity and semantic representations, and a participant-level analysis could verify the consistency of our findings across individuals.

Conclusion In conclusion, despite the methodological limitations, our results robustly show that successful neural decoding depends not merely on cognitive effort, but on the nature of that effort and the successful formation of a coherent linguistic representation in the brain. The brain appears to enter a distinct, but internally consistent, state when processing incoherent language that is not well-described by standard language comprehension models.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. 2024. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561.

A Appendix: Data Exploration

Initial exploration of the Tuckute et al. (2024) dataset revealed the core properties of the experimental conditions. As shown in Figure 8, the fMRI signal distribution varied significantly across the three conditions. The "Driving" condition elicited a strong positive fMRI response on average, while the "Suppressing" condition elicited a negative response relative to the "Baseline."

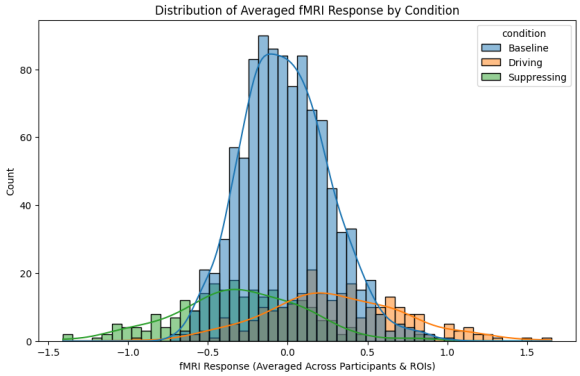

Figure 8: Distribution of averaged fMRI response across 6 ROIs for each of the three conditions. **Conclusion:** The "Driving" condition elicits a strong positive response, while "Suppressing" elicits a negative response relative to "Baseline."

Furthermore, we examined the stimuli themselves. Figure 9 shows that there were minor differences in sentence length across conditions, with "Driving" sentences being slightly longer on average and "Suppressing" sentences being slightly shorter.

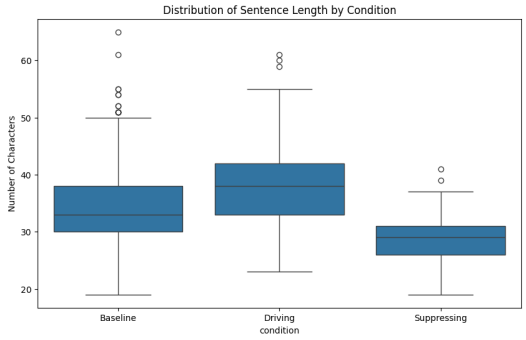

Figure 9: Distribution of sentence length (in characters) for each of the three conditions. **Conclusion:** "Driving" sentences are longest on average, while "Suppressing" sentences are shortest, though distributions overlap significantly.