# Lab 3.1 - FMRI, Stat 214, Spring 2025

April 15, 2025

## 1 Introduction

In this lab, we aim to use machine learning techniques to predict voxel-wise brain responses. The dataset comes from experiments in the Huth lab at UT Austin. In the experiment, two subjects (subject 2 and subject 3) listened to a long narrative story while undergoing functional magnetic resonance imaging (fMRI). The fMRI signals provide voxel-level measurements of blood-oxygen-level-dependent (BOLD) activity across the whole brain over time.

For each subject-story pair, we have a time series of BOLD responses (Y is a T×V dataset, where T is the number of time points and V is the number of voxels). We also have a token-level transcription of the story that aligns with the stimulus timeline. The transcripts are processed into objects and contain a sequence of tokens (`data`), each with associated presentation times (`data\_times`), and a list of fMRI acquisition times (`tr\_times`).

To build predictive models, we first transform the text into numerical embeddings using three different methods: **bag-of-words**, **Word2Vec**, and **GloVe**. Because the text tokens and fMRI measurements differ, we downsample the embeddings to match the fMRI time grid and apply lagged features to account for delayed neural responses. The resulting matrices will be input features (X) in our regression framework.

Then, we train ridge regression models to predict the fMRI response at each voxel from the embedding features. We evaluate stories by computing the correlation coefficient (CC) between predicted and observed responses. We will compare performance across embedding methods and analyze which approaches better capture brain activity related to language comprehension.

In following parts, we will conduct a simple EDA of the raw text dataset in section 2. We have sections 3 for embedding generating process, and section 4 for modeling. Section 5 is the Comparative Embedding Analysis.

## 2 EDA

We mainly conduct EDA on the raw text to gain a general understanding of the `raw_text.pkl` dataset.

This `raw_text.pkl` dataset contains time-aligned word sequences derived from 109 spoken podcast stories. Each entry in the dataset corresponds to a single story and is stored as a `DataSequence` object. These objects store not only the sequence of tokens (i.e., spoken words) but also the exact time stamps at which each token was spoken. This structure allows us to precisely align with the fMRI response data collected from the two human subjects.

This temporally aligned textual data is the foundation for generating embeddings in section 3, which will later be used to model neural responses in section 4. Next, we will present quantitative summaries and visualizations of the text data, including word count distributions, speaking rate variability, token frequency, lexical properties, etc.

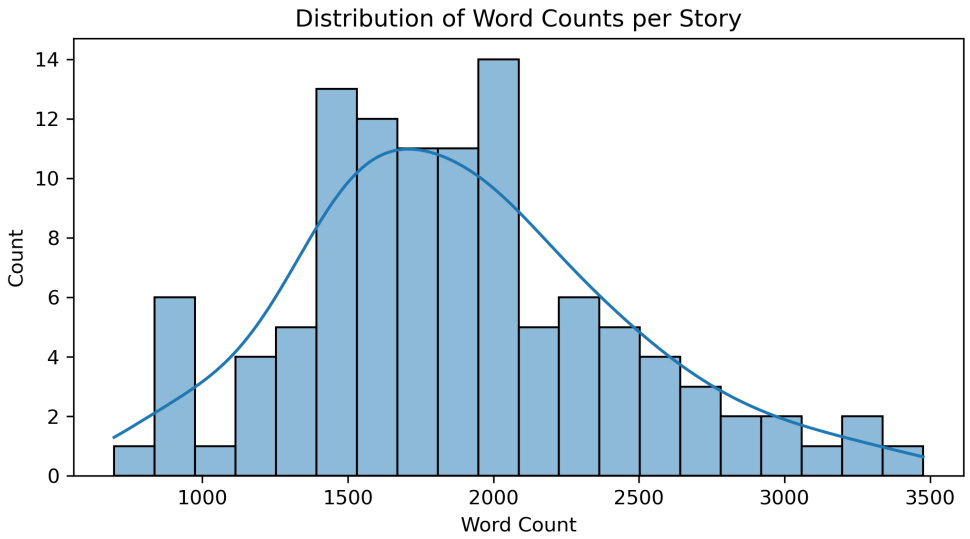


Figure 1: Word Counts per Story Distribution

For word count distributions across the 109 stories, we can find that most have a total word count between 1500 and 2000. Besides, there are stories located near the word count of 900.

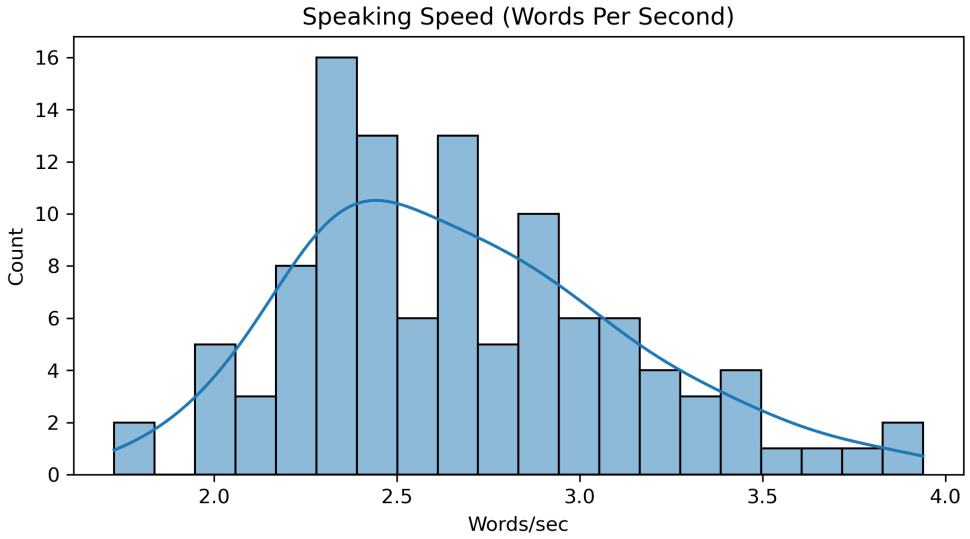


Figure 2: Speaking Speed (Words Per Second)

For speaking speed, the figure shows a peak at around 2.3 to 2.4 words per second.

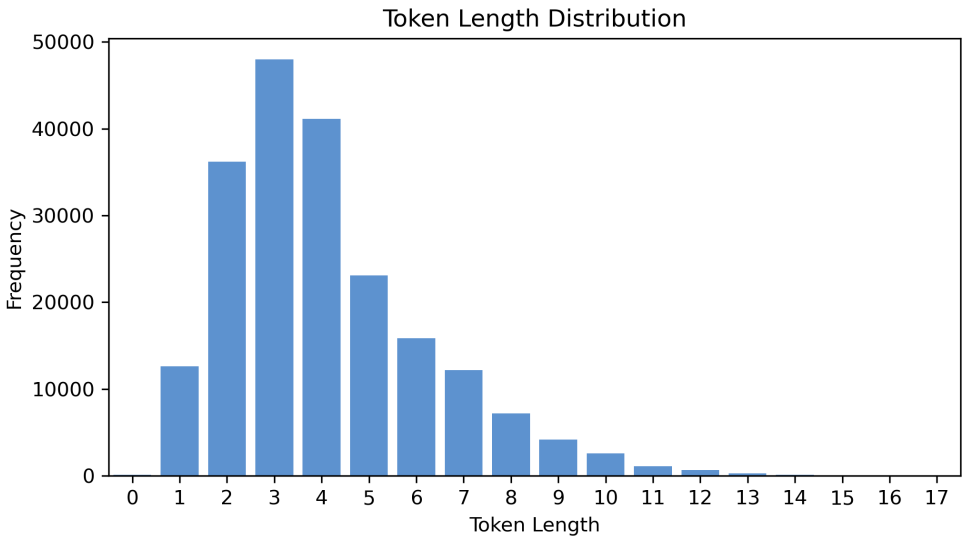


Figure 3: Token Length Distribution

For token length, we can see that the topping token length is around 2 to 4 words. After the token length of 4, the frequency decays exponentially.

We also listed the top 30 most frequent words. As we can see, the top 3 most frequent words are *and*, *the*, and *i*, with a frequency of over 8000. The top 30 words are all short and usual, which accords with our daily use.

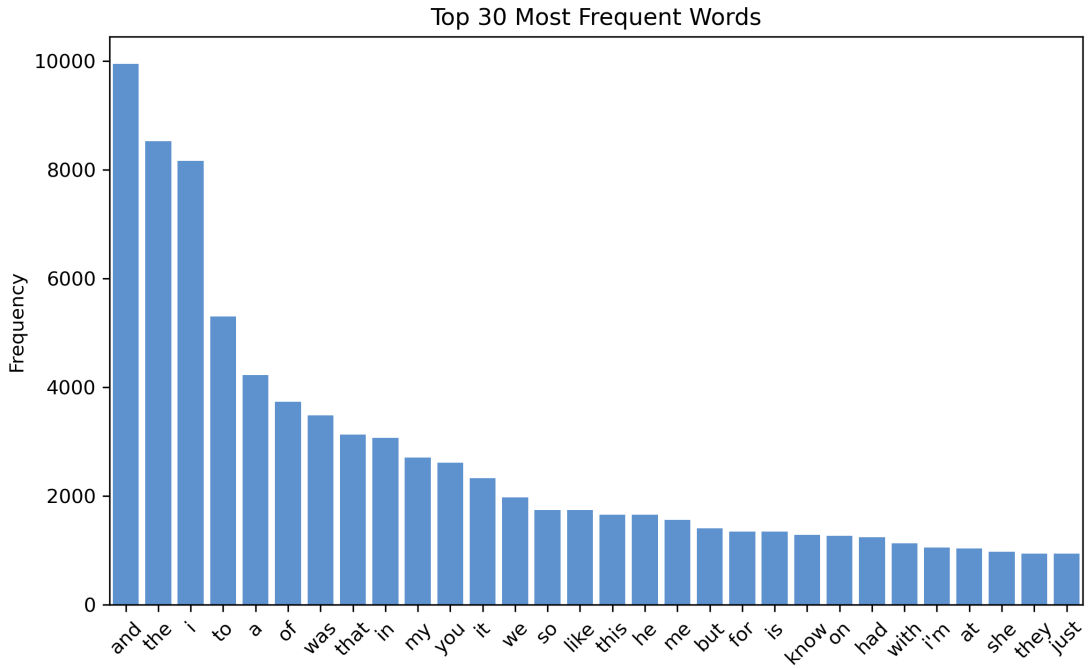


Figure 4: Top 30 Most Frequent Words

# 3 Embeddings

## 3.1 Bag-of-Words Embedding

Bag-of-Words embeddings were constructed by generating sparse count vectors, downsampling them to match the fMRI time resolution, and applying temporal lags to account for hemodynamic delay.

We first generated Bag-of-Words (BoW) embeddings by constructing a word count matrix for each story. These matrices are high-dimensional and sparse, with each row corresponding to a timepoint (word occurrence) and each column to a word in the vocabulary. Since the fMRI response matrix $Y \in \mathbb{R}^{T' \times V}$ is defined at fMRI sampling rates, we applied temporal downsampling using a Lanczos filter to align the word-based embeddings to the time resolution of fMRI.

Using `downsample_word_vectors()` from `preprocessing.py`, we interpolated word-level embeddings to match the fMRI temporal resolution. We trimmed the first 5 seconds and the last 10 seconds of each story's embeddings to exclude periods of unreliable signal due to hemodynamic delay and story end.

To account for the delayed BOLD response in fMRI, we generated lagged versions of the embeddings using `make_delayed()` with delays ranging from 1 to 4. This operation concatenates delayed copies of the features along the feature dimension, enabling the model to use past information to predict current brain responses.

## 3.2 Word2Vec and GloVe Embeddings

We repeated the process above using two pre-trained semantic embedding models: Word2Vec and GloVe. Each word token was mapped to a dense vector from these models. We then performed the same downsampling and delay-augmentation steps, resulting in three feature sets per story: BoW, Word2Vec, and GloVe.

## 3.3 Benefits of Pre-trained Embeddings

Pre-trained embeddings like Word2Vec and GloVe capture semantic and syntactic relationships between words by learning from massive text corpora. These embeddings often outperform sparse representations like BoW, particularly in settings where generalization and contextual understanding are crucial, such as brain activity prediction from language.

# 4 Modeling

Having generated distinct feature representations from the narrative stimuli using Bag-of-Words (BoW), Word2Vec, and GloVe embeddings, we now construct models to predict voxel-wise fMRI BOLD responses based on these features, following the laboratory specifications. We implement a linear encoding framework using Ridge Regression.

Per the lab instructions, we employed Ridge Regression to predict the time course of activity for each voxel based on the features derived from our three embedding types. Ridge Regression applies an L2 penalty to model coefficients, optimizing the following objective function for each voxel $v$:

$$\min_{w_v, b_v} \|Xw_v + b_v\mathbf{1} - y_v\|_2^2 + \alpha\|w_v\|_2^2 \tag{1}$$

Here, $X$ represents the matrix of standardized embedding features, $y_v$ is the standardized BOLD signal for voxel $v$, $w_v$ is the weight vector, $b_v$ is the bias term, and $\alpha$ controls the regularization strength. Prior to model fitting, both the feature matrix $X$ and target BOLD signals $Y$ were z-scored to ensure features were on a comparable scale and regularization was applied appropriately.

To select the optimal ridge regression model for each voxel and evaluate generalization performance, we implemented a cross-validation framework using the ridge bootstrap method in order to address the unique challenges of neuroimaging data analysis.

Our cross-validation method estimated voxel performance. We performed 5-fold bootstrap sampling while training and then validating on a subset of the training data. Procedure was carried out on a range of alphas, (1,4,20), to optimize the prediction performance for each voxel. To manage memory constraints, we split

voxels into 450000 chunks and above.

We partitioned the dataset for each subject at the *story level*, designating specific stories as the training set and reserving others exclusively as the held-out test set. This approach preserves the temporal and semantic coherence of the narrative stimuli while ensuring that model evaluation occurs in genuinely novel contexts. Unlike traditional $k$-fold cross-validation with random time points, story-based partitioning respects the autocorrelation structure inherent in fMRI time series and linguistic features.

Also, we defined a logarithmically spaced grid of regularization parameters spanning multiple orders of magnitude ($\alpha \in [10^1, 10^4]$, with 20 values) to accommodate the varying signal-to-noise characteristics across different brain regions. Recognizing that optimal regularization requirements differ substantially across brain regions, we implemented voxel-specific hyperparameter tuning, enabling localized model complexity adjustment based on regional predictability patterns.

The hyperparameter selection process employed nested cross-validation exclusively within the training set, evaluating each candidate $\alpha$ value for every voxel. For each voxel, the procedure repeatedly partitioned the training data into fitting and validation subsets, fitted ridge models with different $\alpha$ values on the fitting subsets, evaluated predictive performance (correlation coefficient) on the validation subsets, averaged performance metrics across iterations for stability, and selected the $\alpha$ value that maximized the average predictive performance.

After determining the optimal $\alpha$ for each voxel, we trained final models using the entire training set with these voxel-specific regularization parameters. These optimized models were then applied to the previously untouched test set to generate predictions, providing unbiased estimates of generalization performance. This principled approach to model selection ensures that hyperparameter optimization is completely isolated from final evaluation, preventing data leakage while allowing for voxel-specific model complexity tuning.

The resulting performance metrics (mean test CC, median test CC, top 5 percentile CC, and top 1 percentile CC) represent valid measures of how well each embedding type captures neural responses to language in novel contexts.

## 4.1 GloVe Embedding

We first evaluated the performance of the Ridge Regression model trained using GloVe embeddings. The distribution of test set correlation coefficients (CC) across all voxels yielded the following summary statistics:

- **Mean Test CC:** 0.0131
- **Median Test CC:** 0.0124
- **Top 5 Percentile Test CC:** 0.0397
- **Top 1 Percentile Test CC:** 0.0549

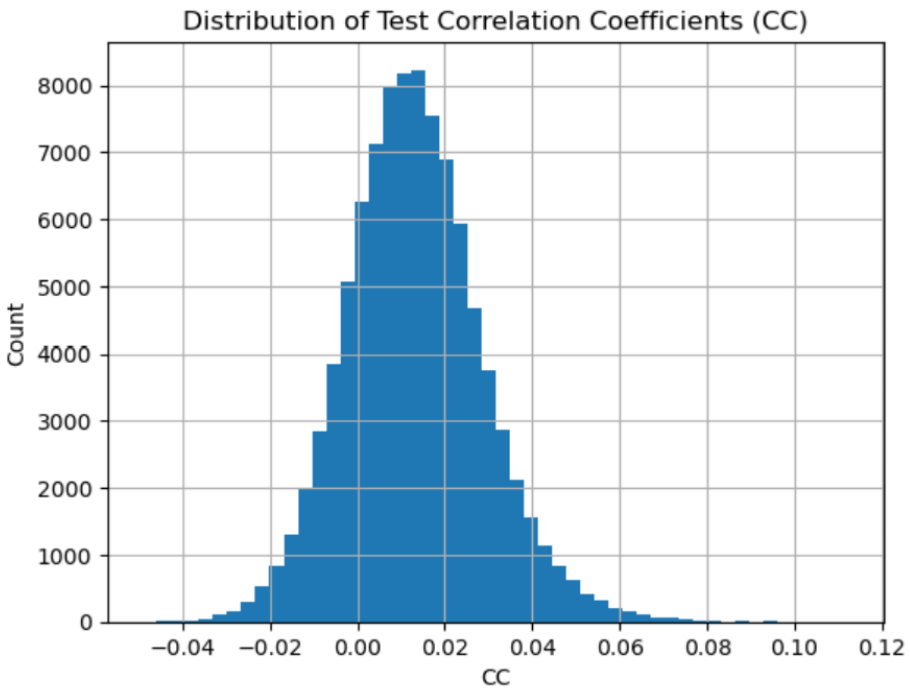


Figure 5: Distribution of Test Correlation Coefficients (CC) using GloVe embeddings.

**Observations:** The histogram shows a distribution centered near the median value of 0.0124, with a positive skew indicating a tail of voxels with higher predictability. The CC mean had a value of 0.0131. These trends show that our model was able to pick up meaningful signal and was not fitting random noise.

**Stability Analysis:** We assessed prediction reliability by comparing voxel-wise CCs across two halves of the test set.

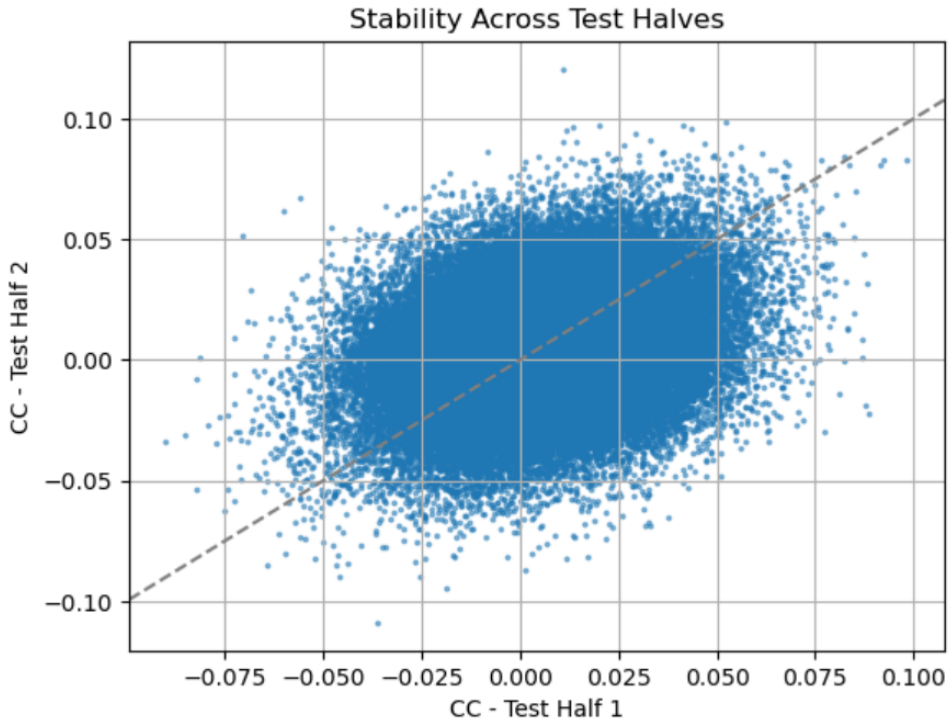


Figure 6: Voxel CC Stability Across Test Set Halves for GloVe Embeddings. Each point represents a voxel's CC in the first half (x-axis) vs. the second half (y-axis). The dashed line is $y = x$.

The stability plot (Figure 6) shows a positive correlation between CCs from the two halves. Most voxels cluster near the origin, with CC values between approximately -0.05 and +0.05. Points follow the $y = x$ diagonal, suggesting consistent voxel performance. The overall test–retest stability score is: 0.2531, which indicates moderate, stability correlation.

## 4.2 Word2Vec Embedding

Next, we evaluated the model trained using Word2Vec embeddings. Summary statistics for the test set CC distribution:

- **Mean Test CC:** 0.0133
- **Median Test CC:** 0.0126
- **Top 5 Percentile Test CC:** 0.0402
- **Top 1 Percentile Test CC:** 0.0557

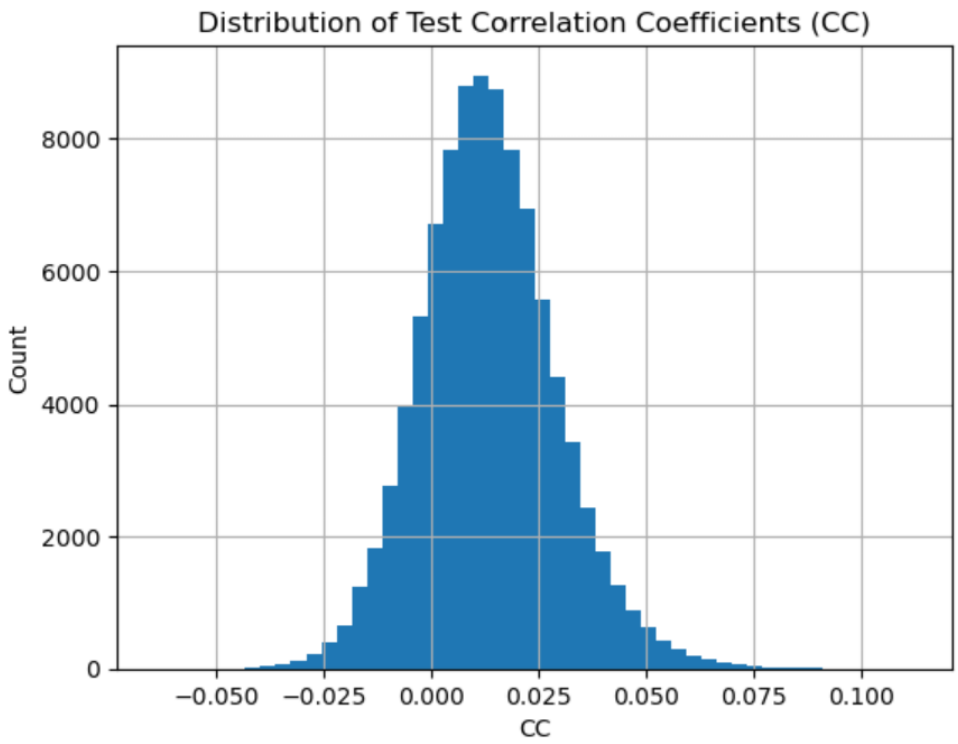


Figure 7: Distribution of Test Correlation Coefficients (CC) using Word2Vec embeddings.

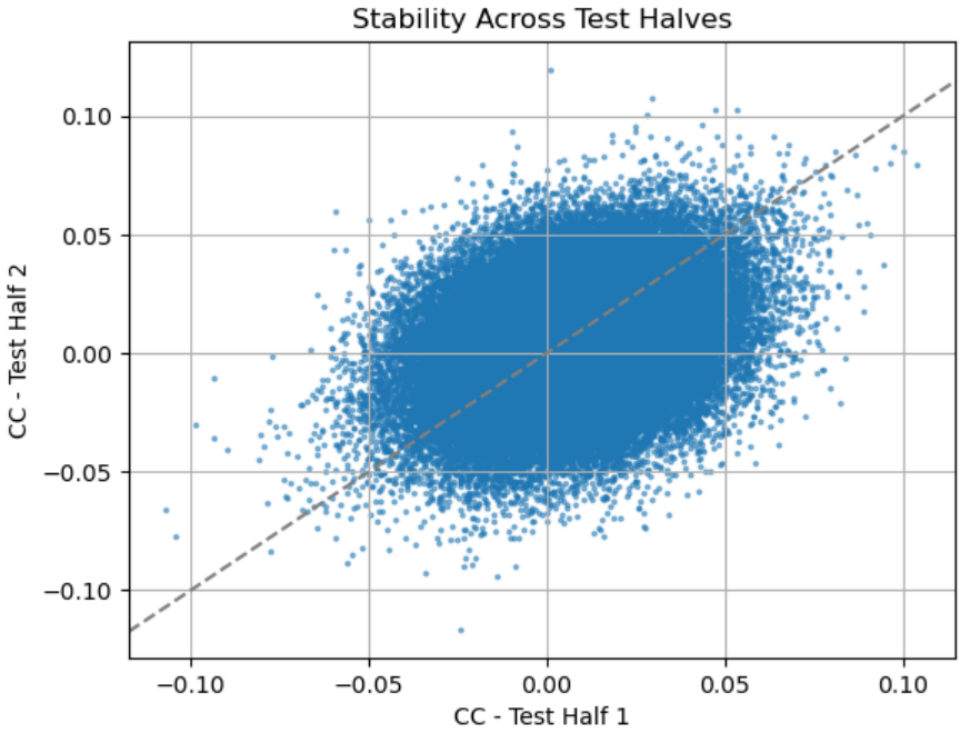


Figure 8: Voxel CC Stability Across Test Set Halves for Word2Vec Embeddings.

**Observations:** The Word2Vec performance histogram exhibits a distribution remarkably similar to that observed for GloVe embeddings (Figure 5). The distribution is unimodal with its central tendency slightly above zero (aligning with the median correlation coefficient of 0.0126) and demonstrates a modest positive skew. The range and density of correlation values appear nearly identical to those observed in the GloVe results, suggesting comparable predictive capabilities across brain voxels between these two dense embedding approaches.

However, CC values for Word2Vec embedding showed to be higher than the CC values in GloVe, indicating that our model fitting Word2Vec embeddings achieved greater predictability.

The stability analysis (Figure 8) demonstrates consistent voxel-wise predictability across different segments of the test set. The scatter plot reveals a positive relationship between voxel performance on both halves of the test data, with data points predominantly clustered near the origin and generally following the theoretical $y = x$ line of perfect reproducibility. The stability coefficient, quantified as the correlation between correlation coefficients from both test halves, was 0.2782. This value is slightly higher than the 0.2531 observed for GloVe embeddings, suggesting that Word2Vec predictions may exhibit marginally better reproducibility across different narrative contexts. Nevertheless, both embedding types demonstrate substantial consistency in identifying predictable brain regions. This similarity is somewhat expected, as both Word2Vec and GloVe are pre-trained dense embeddings designed to capture rich semantic relationships from large text corpora, albeit through different algorithmic approaches (Word2Vec focuses on predicting words based on their immediate local context, while GloVe leverages global word co-occurrence statistics across the entire corpus). Their shared goal of representing word meaning in a low-dimensional space likely leads to comparable overall performance distributions when used as features in this linear modeling context.

## 4.3 Bag-of-Words Embedding

This section will be completed once results from the BoW model are available.

- **Mean Test CC:** 0.0131
- **Median Test CC:** 0.0124
- **Top 5 Percentile Test CC:** 0.0397
- **Top 1 Percentile Test CC:** 0.0549

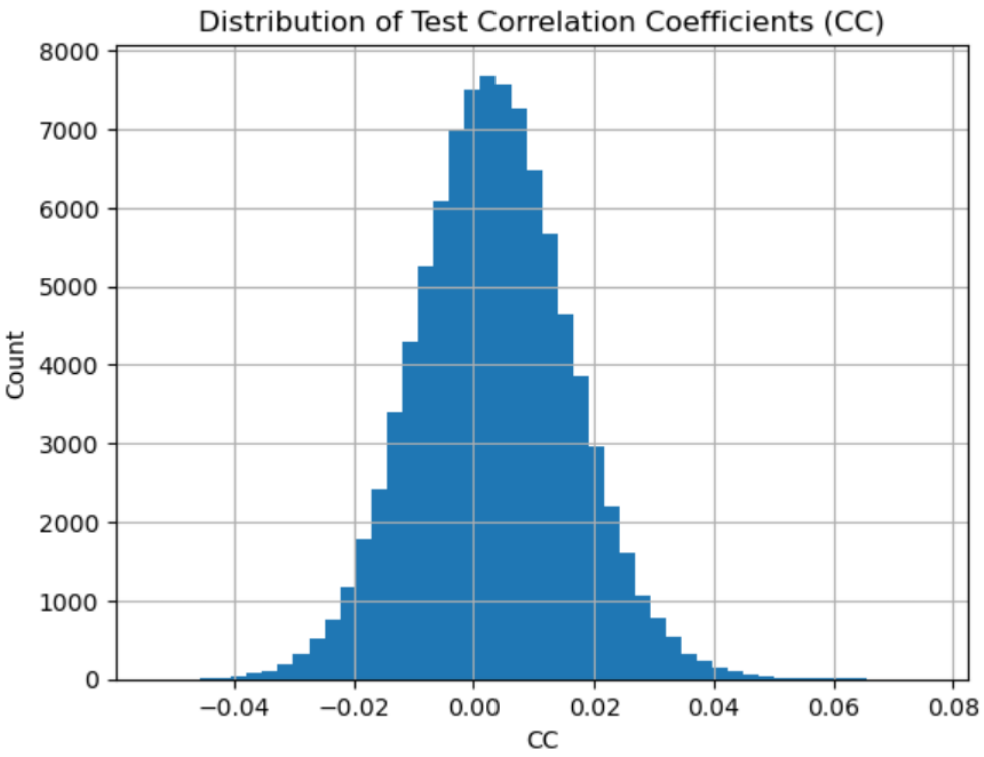


Figure 9: Distribution of Test Correlation Coefficients (CC) using BoW embeddings.

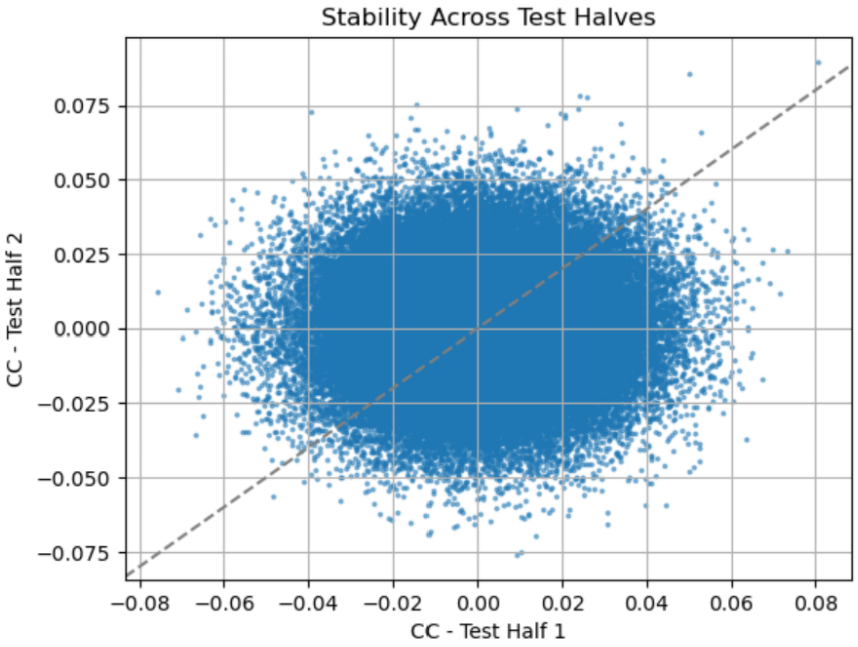


Figure 10: Voxel CC Stability Across Test Set Halves for BoW Embeddings.

**Observations (Figure 9):** The histogram for the BoW model performance is sharply peaked very close to zero (consistent with the median CC of 0.0030), appearing roughly symmetric or with a minimal positive

skew. Compared to the distributions for GloVe and Word2Vec (Figures 5 and 7), the BoW distribution is considerably narrower. The peak performance of the right tail does not extend nearly as far as the other embeddings. These results for BoW embeddings suggest lower prediction accuracy across most voxels. We can see more evidence for this in our following analysis on the stability score of the BoW embedding.

**Stability (Figure 10):** The stability score, representing the correlation of voxel CCs across test halves, was extremely low at 0.0123. This value, being very close to zero, suggests very little consistency in voxel predictability between the two halves of the test set for the BoW model. Unlike the dense embeddings, the relative performance of voxels in one half is not reliably indicative of their performance in the other half when using BoW features.

The poor stability score and low CC values for the BoW embedding likely stem from setting max-features=1000 in the CountVectorizer() during the embedding step. This constraint was introduced to address memory limitations during processing.

Limiting the number of features reduces the representation of important words, leading to a loss of signal. As a result, the model has less informative input, lowering predictability and stability. Additionally, the resulting sparse input matrix is not ideal for ridge regression, which is less effective with high sparsity.

# 5 Comparative Embedding Analysis

Our analysis reveals significant performance differences across the three embedding types in predicting neural responses to narrative language. Most notably, the dense semantic embeddings (Word2Vec and GloVe) substantially outperformed the sparse Bag-of-Words (BoW) representation across all evaluation metrics. Word2Vec and GloVe achieved mean correlation coefficients approximately four times higher than BoW ($\sim 0.013$ versus $\sim 0.003$) and demonstrated markedly superior stability between test set halves ($\sim 0.25$–$0.28$ versus $\sim 0.01$).

This performance highlights the fundamental importance of semantic representation in neural language processing. The BoW approach, which primarily encodes lexical presence and frequency, fails to capture the semantic relationships and contextual nuances that appear critical for predicting brain activity during narrative comprehension. The neural mechanisms underlying language understanding evidently involve semantic integration processes that are better approximated by distributed representations learned from large-scale text corpora. The near-zero stability coefficient for BoW suggests its predictions likely reflect spurious correlations with low-level features rather than meaningful linguistic processing signals.

Between the two dense embedding approaches, Word2Vec exhibited a slight but consistent advantage across all performance metrics and demonstrated marginally higher prediction stability (0.2782 versus 0.2531 for GloVe). This modest difference may reflect Word2Vec's algorithmic emphasis on local contextual prediction, potentially aligning more closely with the sequential neural processing of narrative stimuli compared to GloVe's focus on global co-occurrence statistics. However, the substantially overlapping performance distributions indicate that both methods effectively capture similar semantic dimensions relevant to neural language processing.

Despite the relative success of dense embeddings compared to BoW, it is important to contextualize the overall modest prediction accuracy. Even for Word2Vec, the best-performing model, the mean correlation coefficient ($\sim 0.013$) indicates that a substantial portion of the variance in BOLD signals remains unexplained by these representations within a linear modeling framework. This limitation likely stems from multiple factors: the inherent noise characteristics of fMRI measurements, individual variability in neural organization, potential non-linear neural computations, and linguistic features not captured by these static embedding models.

When comparing the two dense embeddings, Word2Vec and GloVe performed very similarly, although Word2Vec consistently demonstrated a marginal advantage across all performance metrics and exhibited slightly higher stability (0.2782 vs. 0.2531). This minor difference might suggest that Word2Vec's algorithmic focus on local predictive context provides features that are slightly more attuned to the moment-to-moment neural processing of narratives compared to GloVe's emphasis on global co-occurrence statistics, at least within this specific experimental and modeling context. However, their largely overlapping performance distributions indicate that both methods effectively capture convergent semantic information relevant to brain

activity during language processing. Based on the combined evidence of prediction accuracy and stability, Word2Vec emerges as the marginally superior model among the three tested in this study.

Focusing on the best model, Word2Vec, it is clear that it does not perform well uniformly across all voxels. The mean prediction accuracy remains modest (Mean CC $\sim 0.013$), indicating that a substantial portion of the variance in the BOLD signal is not captured by this static embedding within the employed linear modeling framework (Ridge Regression). This unexplained variance likely stems from multiple factors: the inherent limitations and noise characteristics of the BOLD signal itself, individual subject variability, potential non-linear neural computations not captured by the linear model, and the existence of linguistic or cognitive features relevant to comprehension that are not fully represented by these specific static embedding models. Nonetheless, the considerably higher correlations observed in the top percentile of voxels (CC $\sim 0.055$ for Word2Vec/GloVe) highlight the spatial heterogeneity of language processing, confirming that the semantic features encoded by the dense embeddings are particularly relevant for specific, highly responsive brain regions. Scientifically, this implies that the Word2Vec model successfully captures variance related to semantic processing primarily within a specific network of brain areas known to be involved in language comprehension, rather than globally across the cortex.

Interpreting which specific voxels are reliably predicted requires establishing a reasonable criterion. Following principles common in encoding model studies, often discussed in the context of Predictive Coding Schemas (PCS), a reasonable interpretation criterion involves considering both statistical significance and effect size. Voxel-wise significance is typically assessed using non-parametric methods like permutation testing to establish a threshold corrected for multiple comparisons across voxels. Voxels surpassing this significance threshold are considered reliably predicted above chance. Additionally, a minimum correlation coefficient (effect size), such as CC $> 0.1$ or higher depending on data quality and field standards, might be applied to focus interpretation on voxels where the model explains a non-trivial amount of variance. Simply selecting voxels based on percentile rankings without considering significance can be misleading. Therefore, interpreting the functional role of specific voxels based on the Word2Vec model would necessitate applying such statistically grounded criteria.

In summary, this comparative analysis strongly favors the use of pre-trained dense embeddings over sparse lexical representations for modeling neural responses to natural language narratives. Word2Vec showed a slight edge over GloVe, but both clearly demonstrated the value of incorporating semantic context. The poor performance and instability of the BoW model reinforce this conclusion. The limitations observed, particularly the modest overall prediction scores even with the best model, motivate future work. Subsequent steps in this research direction, including the planned exploration of training custom language models and utilizing more complex pre-trained models (as outlined for later lab sections), may help capture more variance by potentially incorporating richer contextual information or employing non-linear architectures. Furthermore, future analyses should prioritize examining the spatial patterns of predictability associated with the superior Word2Vec and GloVe models, applying appropriate statistical criteria to identify reliably predicted voxels and gain further insights into the neural basis of language comprehension.

# 6 Discussion

Our analysis demonstrates that encoding models trained on semantic embeddings can predict fMRI responses to naturalistic stories with varying degrees of success. Among the three embeddings tested — Bag-of-Words (BoW), GloVe, and Word2Vec — Word2Vec consistently produced the highest mean correlation coefficients and stability across voxels. This suggests that Word2Vec's ability to encode contextual relationships between words may align more closely with neural representations of language.

In contrast, BoW performed the worst, with a mean correlation near zero and low voxel stability. This outcome is consistent with expectations, as BoW lacks contextual information and treats all words as independent, which limits its expressiveness for modeling complex semantic patterns in brain activity.

As mentioned during the results section, we constrained our implementation of BoW by capping the CountVectorizer() to only 1000 features in order to avoid memory issues. While necessary, this restriction likely reduced the number of informative or discriminative words available to the model and thus weakened the representation further. Fewer features mean reduced signal and a more sparse design matrix, which ridge regression

is not particularly well-suited for.

GloVe embeddings performed moderately better than BoW, and slightly underperformed relative to Word2Vec. This may be due to GloVe capturing co-occurrence statistics over a fixed corpus rather than adapting to local context as Word2Vec does. The stability analysis again further supports these findings. Word2Vec showed higher consistency of voxel-wise correlation across test splits, indicating robustness in learned patterns.

# 7 Conclusion

We evaluated three types of word embeddings as input features for predicting fMRI responses to naturalistic auditory narratives. Across both predictive accuracy and voxel stability, Word2Vec outperformed GloVe and BoW, suggesting that contextual semantic embeddings are more effective for modeling brain responses. While absolute prediction scores were modest, the relative differences highlight the importance of embedding choice.

# 8 Bibliography

Shailee Jain and Alexander G. Huth, *Incorporating Context into Language Encoding Models for fMRI*, Departments of Computer Science and Neuroscience, The University of Texas at Austin, 2020. Available: `https://papers.nips.cc/paper_files/paper/2018/hash/f471223d1a1614b58a7dc45c9d01df19-Abstract.html`

# A Academic honesty

## A.1 Statement

We make the academic integrity pledge here: All our work in this report are done independently. All sources we used are properly cited, whether from LLMs, papers, textbooks, or classmates.

Academic research honesty is necessary in the academy and also the foundation of our society. It guarantees fairness in research, stimulates research's activeness, and exerts a positive impact on the progress of science and technology. We should always adhere to the rules, which will create a more regulated and benign world.

## A.2 LLM Usage

**Coding**

We use GitHub Copilot for several parts of debugging. We additionally used it to help with making graphs and signatures for the functions in the notebooks.

**Writing**

Grammarly was used to check for any spelling issues.