Lab 3.3 - FMRI, Stat 214, Spring 2025

Anonymous

May 16, 2025

1 Introduction

Understanding how the human brain processes natural language remains a central challenge in neuroscience and cognitive science. In previous phases of this project, we developed encoding models to predict voxel-level brain activity from language stimuli using fMRI data from [2]. In Lab 3.1, we compared Bag-of-Words, Word2Vec [4], and GloVe [5] embeddings; in Lab 3.2, we pre-trained a custom encoder using a masked language modeling objective to generate context-aware embeddings. Across both parts, these embeddings were used as inputs to ridge regression models to predict BOLD responses across voxels, providing insight into the relationship between linguistic representations and neural activity.

In this final phase, Lab 3.3, we shift focus to *fine-tuning and interpreting* a large pre-trained language model for our brain prediction task. Specifically, we begin by fine-tuning BERT [1] on the fMRI dataset to better align its embeddings with the neural responses observed during naturalistic story listening. We also explore parameter-efficient fine-tuning strategies such as Low-Rank Adaptation (LoRA) to assess whether we can achieve similar performance gains with reduced computational overhead. These fine-tuned models are then evaluated against the previously tested approaches from Labs 3.1 and 3.2.

In addition to predictive performance, this lab introduces an interpretability dimension. Using SHAP and LIME, we analyze which input words most strongly influence the model's predictions at well-predicted voxels. By focusing on voxels where prediction accuracy is highest, we aim to uncover linguistically meaningful features that contribute to model output. Comparing the outputs of SHAP and LIME across different stories and subsets of voxels allows us to assess the consistency and plausibility of the explanations provided, offering a new perspective on how language models—and potentially the brain—represent meaning during natural language processing.

2 EDA

In this section, we will conduct a primitive EDA on the dataset to gain a basic understanding of the data and its structure. The dataset consists of fMRI signals from two subjects, each listening to 101 stories. We divide the data into training/validation and test sets, with 75% of the stories (75 stories) used for training and validation, and 25% (26 stories) reserved for testing. The test data is reserved until the Test Performance section, and the EDA, embedding and modeling parts are conducted on the training/validation data only.

In this section, we conducted the EDA on the stories and fMRI signals separately.

2.1 Stories

The stories are stored in a "list of words" format, where each story is represented as a list of words, without punctuation or spaces. Figure 1 shows the distribution of story lengths, measured by the number of words, across the 75 stories in the training/validation set. The histogram indicates a roughly unimodal distribution with a slight right skew. The majority of stories cluster around the central tendency, with the mean length being 1893 words and the same median length. The similarity between the mean and median confirms the relatively mild skew. The shortest story contains 697 words, while the longest contains 3476 words, showing a considerable range in the duration of the stimuli presented.

The following shows a selected piece of one of the stories. The punctuations are added by us as it is not in

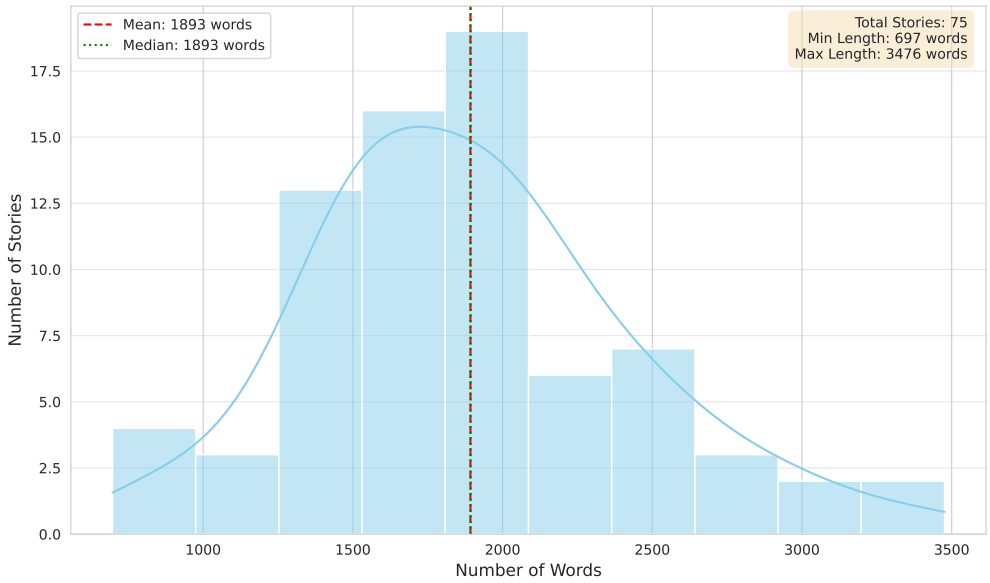


Figure 1: Distribution of story lengths in training/validation dataset.

the original text.

> "My story begins, I am driving my uh silver station Volvo uh from Brooklyn to my mother's house in Rosedale, Queens, on a hot mid-afternoon August day in two thousand and three. My mother uh is a widow. Uh, my father has passed away from lung cancer fifteen years before, nineteen eighty-eight, and she has not resumed dating. She has sworn off men in no uncertain terms. She has told me that, "I am never, ever going to wash another pair of men's underwear again. I am finished with the species. I'm done."

The story is a narrative piece, and the text is rich in detail and context. The language is conversational, with a mix of personal anecdotes and reflections. The use of "uh" indicates a speech pattern that is common in spoken language. That implies the corpus is informal and could differ from more formal ones that commonly used in NLP-related tasks.

2.2 fMRI Signals

The fMRI signal data captures the brain's response, measured via the Blood-Oxygen-Level-Dependent (BOLD) signal, as subjects listened to the stories. For each of the 75 stories comprising the training/validation set for a given subject, the data is organized into a two-dimensional numerical array. Each row in this array represents the brain activity across all measured voxels at a specific Time of Repetition (TR), which is the sampling interval of the fMRI scanner. Each column corresponds to a single voxel, with 94251 voxels for Subject 2 or 95556 for Subject 3, recorded per subject. Consequently, the dimensions of these arrays are $T \times V$, where V is constant, and T (the number of TRs) varies from story to story, reflecting the differing lengths of the narratives; indeed, as shown in Figure 2, there is a strong positive correlation between the number of words in a story and the duration of its corresponding fMRI recording in TRs.

To understand the basic characteristics of the BOLD signal itself, we examined the distribution of raw signal values across all voxels and time points within the training/validation set. Since plotting every single reading is impractical, we randomly sampled 10,000 individual signal values from the training data for each subject. Figure 3 displays the resulting histograms for these sampled raw BOLD values for Subject 2 and Subject 3. Both distributions appear highly similar, exhibiting an unimodal, approximately Gaussian form centered very close to zero. The spread or variance of the raw signals also seems comparable between the two subjects.

In addition to examining the average fMRI signal across the entire dataset, it is informative to visualize

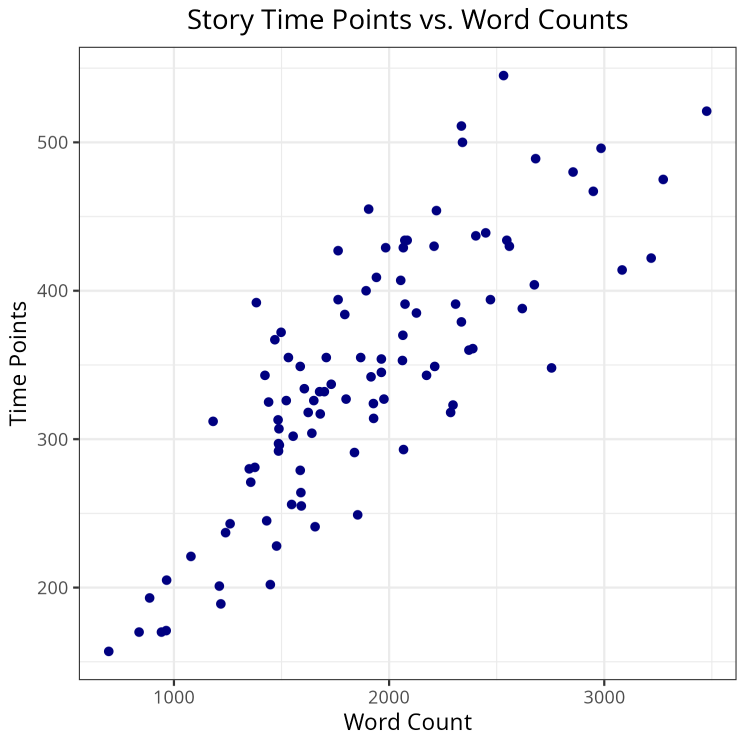


Figure 2: The number of words in a story is strongly correlated with the number of fMRI repetition time points.

the distribution of signals for each story individually. Figure 4 presents the distribution of several summary statistics (mean, median, interquartile range (IQR), minimum, and maximum) computed from the fMRI signal for each of the 101 stories, where each story is treated as a single data point. The central tendency and spread (IQR) appear highly consistent across stories, while the minimum and maximum values show much greater variability. This variation in extremes may reflect story-specific events or transient noise artifacts.

Figure 5 offers a more granular view by plotting the mean and IQR of the fMRI signal across all voxels at each time point within selected example stories. This enables a direct comparison of the temporal signal profiles between Subject 2 and Subject 3 over the course of each story. Despite some fluctuations, the mean signal values for both subjects consistently oscillate within a relatively narrow range (approximately –1 to 1), not only in the stories shown here but across the full set.

In stark contrast, the maximum signal value at each time point, shown in Figure 6, reveals substantial variability in peak activity levels over the course of a single story. Although some of this variation may be driven by noise, there appears to be a moderate degree of correlation or shared structure in the timing of peak signal fluctuations between the two subjects listening to the same story.

Another thing to note is that the fMRI signals of Subject 2 contain some NaN values, which are not present in Subject 3. However, the proportion of NaN values is very small, only 5.9×10^{-6} of the total number of values. Due to the small proportion, we simply imputed them with the global mean to prepare the prediction and expect no significant impact on the model performance no matter which imputation method we use.

3 Embeddings

The primary objective of this stage is to transform the raw textual narratives from the podcast stories into numerical feature vectors aligned with the temporal resolution of the fMRI data (i.e., one feature vector per Repetition Time, TR). This process largely mirrors the methodology employed in Lab 3.2, involving the

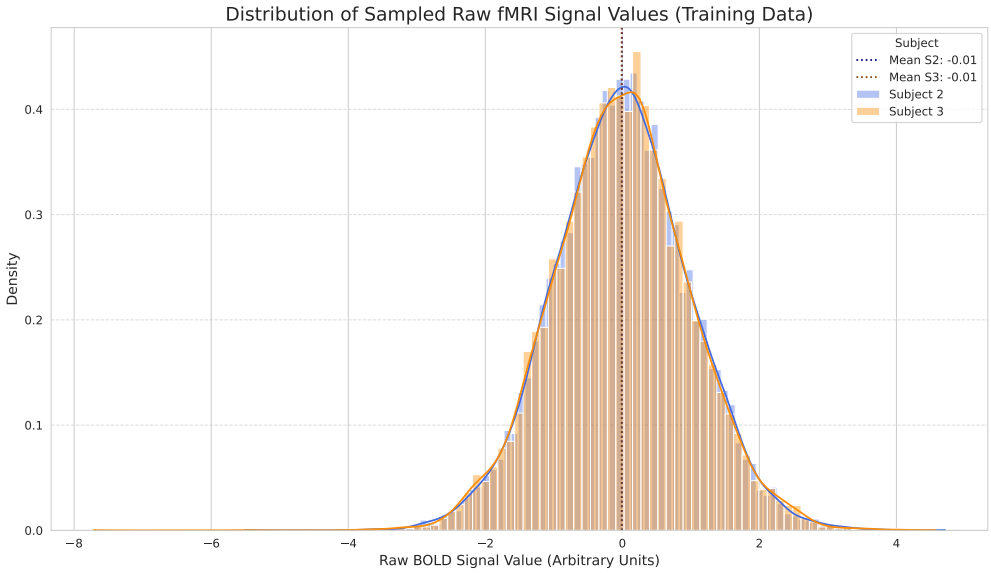


Figure 3: Distribution of fMRI signals in training/validation dataset.

extraction of word-level embeddings and subsequent temporal aggregation. However, a key distinction in this lab is the use of the pre-trained BERT model [1] as the source of token representations. The model can be the original BERT model or a fine-tuned version, which will be described in the following sections.

The overall pipeline for generating TR-level features is as follows:

1. **Tokenization:** Each story's text is tokenized using the pre-trained tokenizer associated with BERT. This tokenizer segments words into subword units.
2. **Token Embedding Extraction with Sliding Window:** The pre-trained BERT has a maximum input sequence length of 512 tokens. Since the stories are often longer than this limit, a sliding window approach is implemented to process the entire token sequence of each story and obtain token-level embeddings (final hidden states) from the BERT model. This process is detailed in Section 3.1.
3. **Word-Level Aggregation:** For words represented by multiple subword tokens, their respective token embeddings (obtained from the sliding window output) are averaged to produce a single vector representation for each word. For words represented by a single token, their token embedding is used directly.
4. **TR-Level Aggregation with Lanczos Resampling:** The sequence of word embeddings, timed according to their occurrence in the story, is then resampled to match the fMRI TR timings. This is achieved using Lanczos interpolation, identical to the method in Lab 3.2, which computes a weighted average of word embeddings temporally proximal to each TR.

The resulting TR-level feature vectors serve as the input for the ridge regression models predicting fMRI voxel activity. The use of a pre-trained BERT model provides rich, context-aware initial embeddings, while the sliding window mechanism allows us to leverage this model for sequences exceeding its native context length.

3.1 Sliding Window for Token Embedding Extraction

The BERT model is restricted to processing input sequences of at most 512 tokens. To derive token embeddings for entire stories, which typically exceed this length, we implement a sliding window technique. This technique processes the full token sequence of a story in manageable, overlapping segments.

The procedure is as follows:

1. **Chunking:** The complete sequence of input token IDs for a story is divided into chunks of a fixed window size (set to 512 tokens, matching the BERT model's limit).

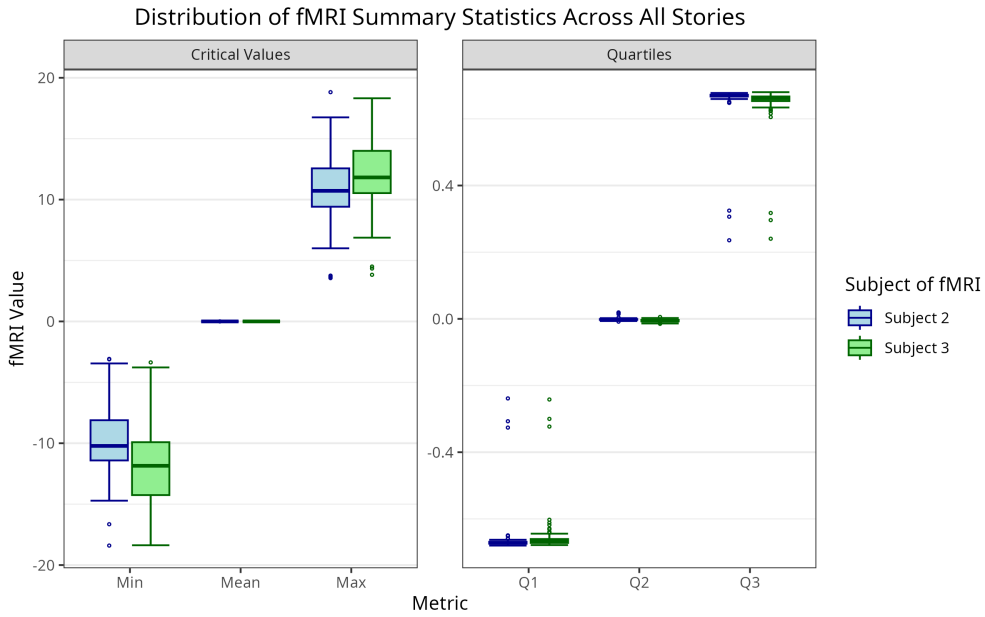


Figure 4: Across all 101 stories, the mean, median, and interquartile range (Q1–Q3) of fMRI signals are tightly distributed, while the minima and maxima show wide variation.

2. **Striding and Overlap:** These chunks are extracted with a specified stride (set to 256 tokens). A stride smaller than the window size ensures that consecutive windows overlap. This overlap is beneficial as tokens appearing near the boundaries of one chunk will also appear in more central, and thus potentially better contextualized, positions in adjacent chunks.
3. **BERT Processing:** Each chunk of tokens, along with its corresponding attention mask segment, is passed independently through the pre-trained BERT. The model outputs the final hidden states (embeddings) for every token within that chunk.
4. **Aggregation of Token Embeddings:** Since a single token from the original long sequence might appear in multiple overlapping windows, its representation is computed by averaging the hidden states obtained for that token from all the chunks in which it was processed. This is achieved by summing the embedding vectors for each token position across all chunks covering it and then dividing by the number of times that position was included in a chunk.

The result of this sliding window process is a sequence of token embeddings with the same length as the original tokenized story, where each token embedding $\mathbf{e}_i \in \mathbb{R}^{768}$ has been informed by a local context of up to 512 tokens. These aggregated token embeddings are then used for the word-level aggregation, and finally for the TR-level aggregation.

3.2 Differentiability of the Embedding Process

The embedding process described above is fully differentiable, meaning that the gradients can be backpropagated through the entire pipeline. This allows for the possibility of fine-tuning the BERT, enabling it to adapt to the fMRI data and the language processing tasks at hand. By allowing gradients to flow through the token embedding, sliding window, and aggregation steps, we can optimize the entire embedding process jointly with pretrained BERT.

4 Modeling - Pre-Trained Embeddings

4.1 Modeling Approach

We create a predictive model to predict fMRI levels for each voxel using the pre-trained embeddings we have generated. Specifically, we fit a ridge regression model. This modeling approach contains the parameter

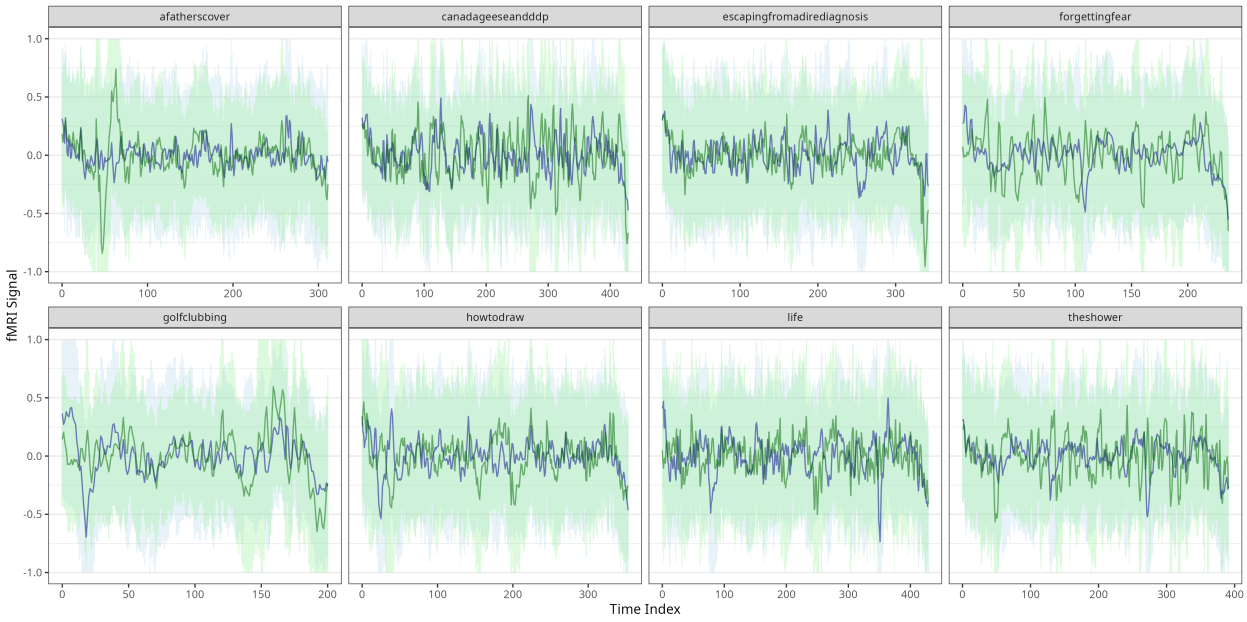


Figure 5: Mean fMRI signal (line) and interquartile range (shaded area) for Subject 2 and Subject 3 at each time point, plotted separately for each story. Mean values for both subjects generally oscillate between –1 and 1.

alpha, which controls the penalty term on the model's weights as L2 (squared) loss.

We start by fitting a regression model for Subject 2, for each different embedding - Bag of Words, GloVe, and Word2Vec. Using the cross-validation strategy described in the next section, we find the best alpha hyperparameter value for regularization.

4.2 Model Evaluation Strategy

We utilize a standard k-fold cross-validation strategy to develop our predictive models. The split is done at a story level instead of a TR level to mimic the real-world scenario where the model is trained on a set of stories and then evaluated on unseen stories. 60% of the stories are used for training and validation, and the remaining are reserved for testing and remain untouched until the final evaluation.

For each fold, the bag-of-words is retrained on the training data to avoid data leakage, while the pre-trained embeddings are applied before the data split as they are fixed and independent of the training data.

The metric we use to evaluate the model performance is the correlation coefficient (CC) between the predicted and actual fMRI signals, which is a standard metric in the context of fMRI signals. We do this per voxel, giving us a voxel-wise CC. This is the metric that our model is trained to optimize for.

This strategy is designed to mimic the real-world scenario with the best efforts to avoid data leakage and measure the model's generalization performance.

4.3 Results

Our cross-validation results for hyperparameter tuning are shown in Tables 1, 2, and 3 for the models trained with the Word2Vec, GloVe, and Bag of Words embeddings, respectively. These are performance metrics on the validation set.

From this cross-validation process, the best model ended up being the Bag of Words model which used an alpha of 1000. It had a mean test CC of 0.0009, median CC of 0.0009, Top 1 percentile CC of 0.0311, and Top 5 Percentile CC of 0.0215. Overall, the CC is low (our model has a limited ability to predict fMRI levels

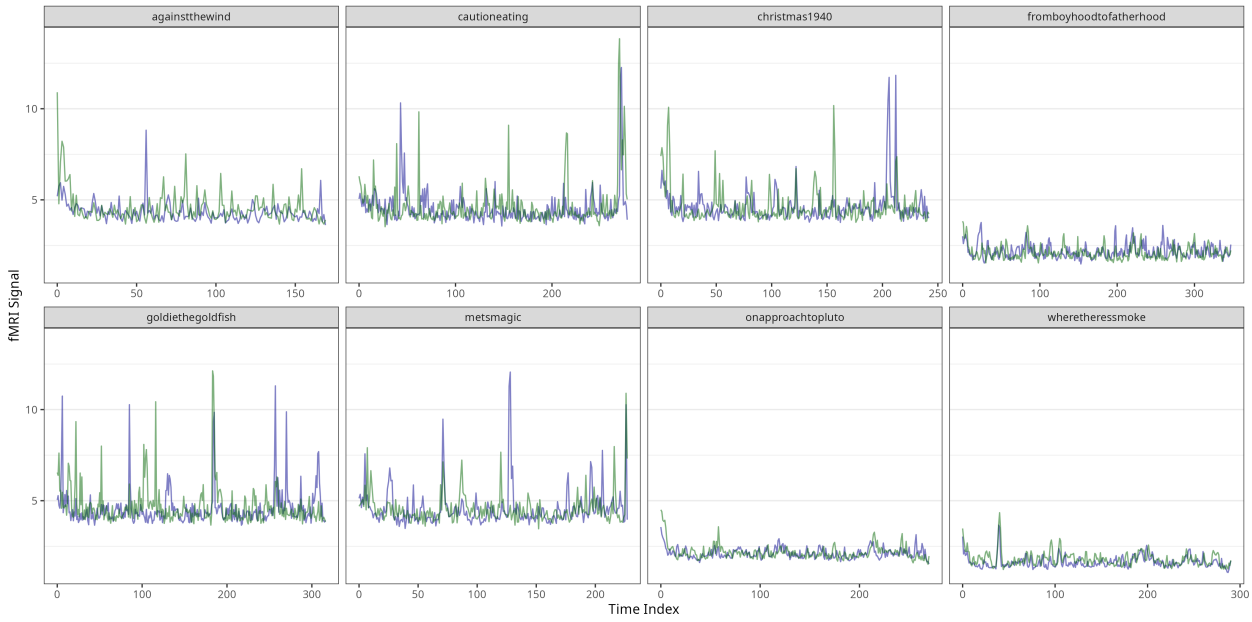


Figure 6: Maximum fMRI signals for Subject 2 and Subject 3 are highly moderate across several stories, despite substantial noise.

Table 1: Performance metrics for Word2Vec at different values of `alpha`. Best alpha: 1000 (Mean CV CC = 0.0057).

Alpha	Mean CC	Median CC	Top1 CC	Top5 CC
0.1	0.0035	0.0032	0.0462	0.0325
1	0.0036	0.0032	0.0462	0.0325
10	0.0036	0.0032	0.0463	0.0325
100	0.0039	0.0035	0.0478	0.0330
1000	0.0057	0.0051	0.0518	0.0364

well), and slightly higher than a random guessing.

4.4 Detailed Evaluation & Analysis

For the model with the best embedding, which was Bag of Words, we performed a more detailed evaluation. We examine the distribution of test CC across voxels. We generate a list of CCs (one for each voxel), then visualize see how the CC is distributed across voxels. We are looking to see how differently the model performs on some voxels in comparison with others if they all have similar performance, or if there is a skew/outlier voxels, etc. Figure 7 describes the distribution of CC across voxels. We can see that the distribution is relatively symmetric, with no major skew. Numerically, the distribution of CC is centered around 0.0009, with a 25th percentile above -0.01 and a 75th percentile at approximately 0.01. An important note is that there are numerous outlier voxels on either side (based on the 1.5*IQR outlier threshold), which suggests that the spread of CC across voxels is relatively large. The positive outliers are stronger/slightly further from the center than the negative outliers.

So, this model does not perform the same across all voxels. Scientifically, this implies that prediction in different voxels of the brain has varying levels of difficulty. This could stem from that some brain areas (voxels) are irreverent from language processing, or at least language processing in listening to these stories, leading to noise that cannot be predicted with the story text, while some others actively respond to the

Table 2: Performance metrics for GloVe at different values of `alpha`. Best alpha: 1000 (Mean CV CC = 0.0067).

Alpha	Mean CC	Median CC	Top1 CC	Top5 CC
0.1	0.0048	0.0044	0.0474	0.0337
1	0.0048	0.0044	0.0474	0.0337
10	0.0058	0.0046	0.0479	0.0341
100	0.0055	0.0051	0.0496	0.0352
1000	0.0067	0.0061	0.0535	0.0377

Table 3: Performance metrics for BoW at different values of `alpha`. Best alpha: 1000 (Mean CV CC = 0.0230).

Alpha	Mean CC	Median CC	Top1 CC	Top5 CC
0.1	0.0063	0.0062	0.0451	0.0338
1	0.0084	0.0082	0.0486	0.0368
10	0.0098	0.0095	0.0514	0.0392
100	0.0141	0.0135	0.0614	0.0476
1000	0.0230	0.0213	0.0851	0.0676

story. However, more domain knowledge is required to justify the hypothesis.

We want to have a reasonable interpretation criterion for interpreting voxels according to PCS. We want to make sure that the predictions are meaningfully better than by chance. One option is to only select voxels in the top x percentile of the observed distribution of CCs (i.e. for $x = 5$ for the 5th percentile). The reason to select the top voxels is that we know they respond to the stories actively, while others could be just random noise, as stated before. In terms of stability, we would ideally want to be able to predict voxels well across different stories, subjects, etc. We could check this by examining each voxel's CC across model performance for different stories, or different models for different subjects. Lastly, we want the full prediction process to be reproducible and computed reliably. These conditions align with the three main parts of PCS.

4.5 Stability Analysis

We conduct a stability analysis by examining performance across different subjects. In this case, we compare Subject 2, which has been detailed so far, with another model trained on Subject 3. We train and test a model on Subject 3 using the same stories as Subject 2. We can then compare the distributions of CCs to see how stable the process is across different subjects.

Our final Subject 3 model uses the Bag of Words embedding and an alpha hyperparameter of 1000 for Ridge. After the training, the final test set performance for Subject 3 was a mean CC of 0.0017, median CC of 0.0014, top 1 percentile CC of 0.0320, and a top 5 percentile CC of 0.0213. This is shown in Table 4. In comparison with Subject 2, we note that the top 1 percentile and top 5 percentile CCs are very similar, which suggests stable results. The mean and median CC are better for Subject 3, though not by a large margin that would suggest high instability.

Table 4: Test Performance Metrics for Subject 2 and Subject 3

Subject	Mean Test CC	Median Test CC	Top 1 Percentile CC	Top 5 Percentile CC
Subject 2	0.0009	0.0009	0.0311	0.0215
Subject 3	0.0017	0.0014	0.0320	0.0213

We also visualize the distribution of CC across voxels to have a deeper understanding and comparison with Subject 2. This is shown in Figure 8. Visually, the distributions of CC across voxels look very similar between Subject 2 and Subject 3. The medians and quartiles, and outliers also show no major differences.

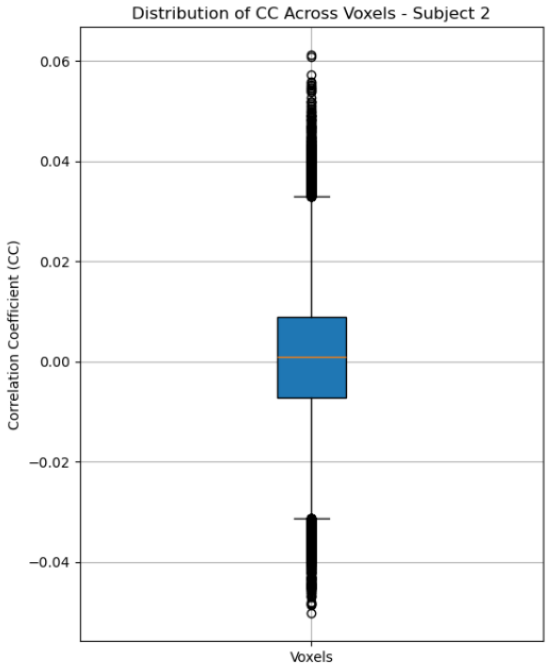


Figure 7: Distribution of CC across voxels for Subject 2 using Bag of Words embeddings.

This suggests a generally stable result.

5 Modeling - Encoder Embedding

In this section, we now fit a linear model using the embeddings generated from the BERT-style encoder, as opposed to the pretrained embeddings. Doing so will allow us to examine if using an encoder shows potential for better predictions than pretrained embeddings, and how they differ. We follow a similar cross validation process, testing values across the 0.1 - 1000 range for the Ridge hyperparameter alpha.

5.1 Encoder Embedding Model - Hyperparameter Selection

We utilized the same cross-validation strategy as in the previous section. The hyperparameter training results for selecting the alpha value are shown in Table 5. Based on the CC scores, we selected alpha=1000 as the best value.

Table 5: Performance metrics for encoder embeddings at different values of α. Best alpha: 1000 (Mean CV CC = -0.0052).

Alpha	Mean CC	Median CC	Top1 CC	Top5 CC	Top10 CC
0.1	-0.0055	-0.0060	0.0382	0.0233	–
1	-0.0055	-0.0060	0.0382	0.0233	–
10	-0.0056	-0.0060	0.0382	0.0233	–
100	-0.0052	-0.0058	0.0393	0.0240	–
1000	-0.0052	-0.0058	0.0393	0.0240	–

5.2 Encoder Embedding Model - Results

On the test set, the best model for the encoder embedding (with alpha=1000) reached a mean CC of 0.0060 for Subject 2 and a mean CC of 0.0114. Interestingly, for the encoder embeddings, we see not only significantly greater performance overall but also a larger difference in magnitude between the CC of Subject 2 and Subject 3, which will be examined in further depth. These results are shown in Table 6.

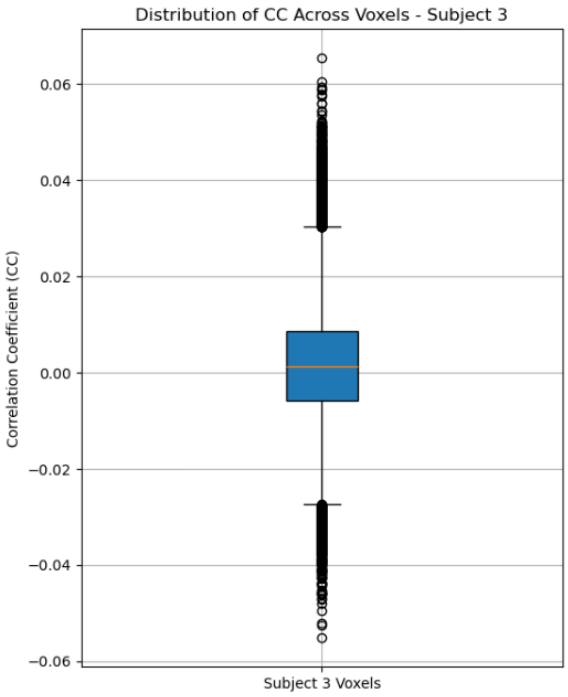


Figure 8: Distribution of CC across voxels for Subject 3 using Bag of Words embeddings.

5.3 Encoder Embedding Model - Detailed Evaluation & Analysis

We examine the distribution of test CC for the encoder embedding Ridge model across voxels. Figure 9 describes the distribution of CC across voxels for Subject 2. We can see that the distribution is somewhat symmetric, with a slight upward skew. Numerically, the distribution of CC is centered around 0.007, with a 25th percentile around -0.01 and a 75th percentile at approximately 0.04. An important note is that there are numerous outlier voxels on either side (based on the 1.5*IQR outlier threshold), which suggests that the spread of CC across voxels is relatively large. The positive outliers are more numerous and stronger/further from the center than the negative outliers, reflecting a slight upward skew. One interesting note is that this distribution of CC is generally shifted more upward than the pre-trained embedding (i.e., Bag of Words) model, and has more positive outliers.

So, this model does not perform the same across all voxels, just as we saw with the model using pre-trained embeddings. Scientifically, this implies that prediction in different voxels of the brain has varying levels of difficulty. This could stem from that some brain areas (voxels) are irrelevant to language processing, or at least language processing in listening to these stories, leading to noise that cannot be predicted with the story text, while some others actively respond to the story. To investigate this further, more domain knowledge would be useful.

As stated earlier, we want to have a reasonable interpretation criterion for interpreting voxels according to PCS. We want to make sure that the predictions are meaningfully better than by chance. One option is to only select voxels in the top x percentile of the observed distribution of CCs (i.e., for $x = 5$ for the 5th percentile). The reason to select the top voxels is that we know they respond to the stories actively, while others could be just random noise, as stated before. In terms of stability, we would ideally want to be able to predict voxels well across different stories, subjects, etc. We could check this by examining each voxel's CC across model performance for different stories, or different models for different subjects. Lastly, we want the full prediction process to be reproducible and computed reliably. These conditions align with the three main parts of PCS.

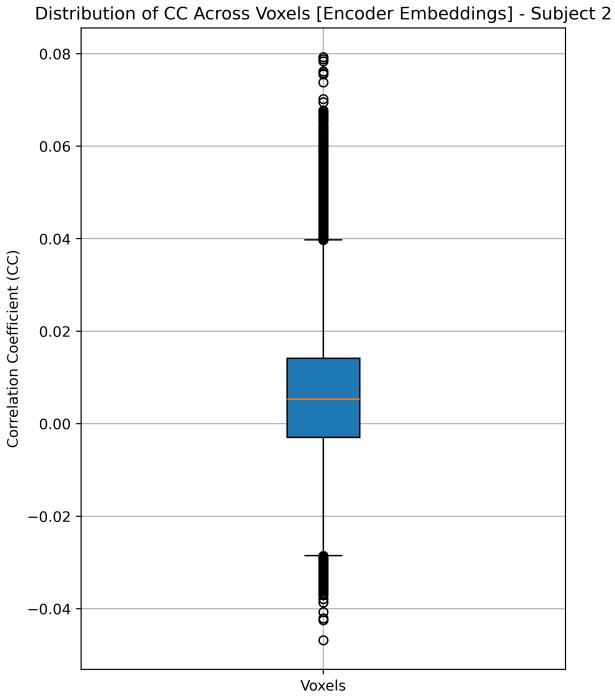


Figure 9: Distribution of CC across voxels for Subject 2 using Encoder embeddings.

5.4 Pre-Trained vs. Encoder Embeddings Comparison

We compare the results of the encoder embedding against the pre-trained embeddings by examining the CC of the encoder model and the Bag of Words model (which was the best pre-trained model). We see that the encoder embeddings are able to reach a higher CC. This is true across all thresholds of CC, including mean, median, Top 1%, and Top 5% CC. In particular, the top performance of voxels (for Top 1% and Top 5% are closer) is closer, while the mean and median CCs are much higher for the model using encoder embeddings. This can mean that the encoder embeddings generally provide better performance throughout more voxels, whereas the pre-trained embeddings only provide comparable performance at their best voxels. As mentioned, the distribution of CC is generally shifted more upward for the encoder embedding than the pre-trained embedding (i.e., Bag of Words) model, and has more positive outliers. This is shown when comparing the boxplots of the CC distribution across voxels in Figures 7 and 9.

In summary, different embedding methods do not perform equally well across the voxels. The encoder embeddings tend to have better performance across more voxels, whereas the pre-trained embeddings tend to have far worse performance for most voxels. However, the pre-trained embeddings still provide comparable performance to the encoder embeddings for the top-performing voxels for both models.

5.5 Encoder Embedding Model - Stability Analysis

We conduct a stability analysis for the encoder embedding model by examining performance across different subjects. In this case, we compare Subject 2, which has been discussed so far, with another model trained on Subject 3. We train and test a model on Subject 3 using the same stories as Subject 2. We can then compare the distributions of CCs to see how stable the process is across different subjects.

Our final Subject 3 model uses an alpha hyperparameter of 1000 for Ridge. After the training, the final test set performance for Subject 3 was a mean CC of 0.0114, median CC of 0.0093, top 1 percentile CC of

Table 6: Comparison - Test Performance Metrics for Bag-of-Words vs. Pre-Trained Embeddings

Embedding	Subject	Mean CC	Median CC	Top 1 % CC	Top 5 % CC
Pre-Trained (Bag-of-Words)	2	0.0009	0.0009	0.0311	0.0215
	3	0.0017	0.0014	0.0320	0.0213
Encoder	2	0.0060	0.0053	0.0416	0.0290
	3	0.0114	0.0093	0.0681	0.0426

0.0681, and a top 5 percentile CC of 0.0426. This is shown in Table 6. In comparison with Subject 2, the median and median CCs are significantly better for Subject 3, as they are almost double that of Subject 2. The top 1 percentile and top 5 percentiles are closer, though still show a noticeable gap.

We also visualize the distribution of CC across voxels to have a deeper understanding and comparison with Subject 2. This is shown in Figure 10. Visually, the distributions of CC across voxels look somewhat similar between Subject 2 and Subject 3, as the medians and quartiles are comparable. However, Subject 3 shows a much stronger right skew and has stronger and more numerous outliers in the higher CC value region. This suggests that Subject 3 has many better, higher-performing voxels. Overall, the differences suggest that there may be some moderate instability in the result of the encoder embedding model across subjects, so further stability testing could be useful.

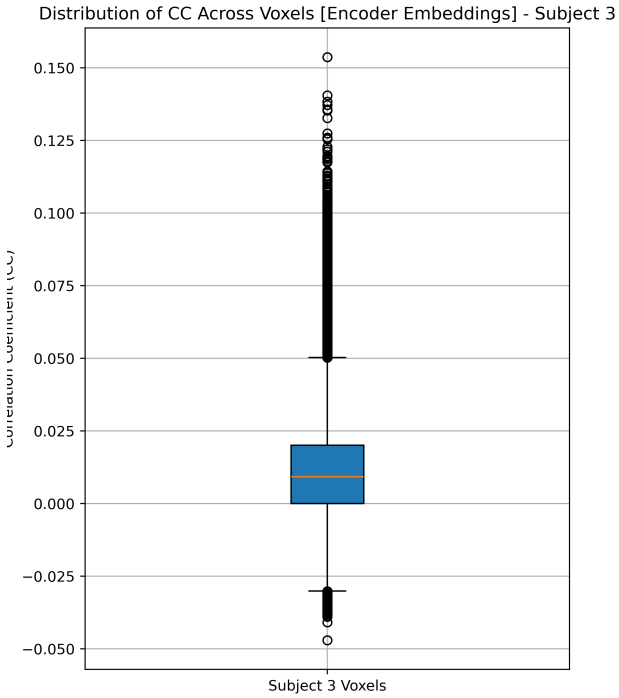


Figure 10: Distribution of CC across voxels for Subject 3 using Encoder embeddings.

6 Modeling - Pre-Trained BERT and LoRA

In this section, we fit predictive models using the pre-trained BERT, both with a regression head only and finetuning the entire BERT with LoRA. We will compare the performance of these models to the previous

models using pre-trained embeddings and encoder embeddings.

6.1 Model Selection Criteria

The core objective of our modeling phase is to predict fMRI voxel activity using features derived from textual stimuli. To evaluate and select the best-performing models, we employ a strategy largely consistent with previous labs, using story-level splits and focusing on the model's ability to generalize to unseen data. The primary metric for this evaluation remains the performance (measured by CC) on a held-out validation set.

A key departure from Lab 3.1 and 3.2 in our model selection process for this lab (Lab 3.3) is the shift from k-fold CV to a fixed train-validation-test split. Specifically, the available stories are divided into a training set (60% of stories), a validation set (20%), and a test set (20%).

The rationale for using a validation set instead of k-fold CV is twofold:

1. **Computational Cost:** Fine-tuning BERT, even with LoRA, is expensive. Performing k-fold CV would multiply this already significant computational burden by a factor of k, making extensive hyperparameter tuning or model exploration prohibitively expensive.
2. **Early Stopping in Deep Learning:** A common and crucial practice in training deep learning models is early stopping. This technique involves monitoring the model's performance on a validation set and picking the model state that achieves the best performance on this set. Implementing a clear and consistent early stopping mechanism is more straightforward with a single, fixed validation set than within a k-fold CV framework because each fold in a CV may pick a different model state based on its own validation set performance.

Therefore, for both modeling approaches in this section, model selection and hyperparameter tuning are guided by MSE loss on the fixed validation set. The model configuration that yields the lowest validation MSE for each subject is selected as the best-performing model for that subject. The final evaluation of these selected models is then performed on the reserved test set. This approach allows for practical model development while still providing a robust mechanism for generalization assessment and overfitting mitigation.

6.2 Regression Head with Pretrained BERT Embeddings

In this initial modeling phase, we utilize the embeddings generated from the pre-trained BERT as fixed features to predict fMRI voxel activity. The BERT model itself is not fine-tuned at this stage; only a linear regression head is trained on top of these static BERT embeddings. The process for obtaining these TR-level features from the stories was detailed in the previous sections. Each TR is thus represented by a 768-dimensional vector (the hidden size of BERT).

For each subject, a separate linear regression model is trained to map these 768-dimensional TR-level feature vectors to the BOLD signal activity across all their respective voxels.

Given the large number of features and target voxels, and consistent with common practices for training neural network components, we employ gradient descent to optimize the weights of these linear regression heads. This contrasts with the closed-form solution often used for traditional ridge regression when the dataset size allows. Instead of an explicit L2 penalty term in the loss function (as in standard ridge regression), we achieve regularization through weight decay in the optimizer.

The training procedure for these regression heads is as follows:

- **Optimizer:** AdamW [3] is used, incorporating weight decay directly into the optimization step, acting as L2 regularization.
- **Learning Rate:** A learning rate of 2×10^{-3} was employed.
- **Weight Decay:** To investigate the impact of regularization, we experimented with three different weight decay values: 10^{-1}, 10^{-2}, and 10^{-3}.
- **Loss Function:** MSE between the predicted and actual fMRI signals for each subject.
- **Epochs and Early Stopping:** The models were trained for a maximum of 100 epochs. The state of the linear regression head that achieved the lowest MSE on the validation set was saved as the best model for each subject. This early stopping mechanism helps prevent overfitting.

The performance of the best models will be evaluated on the test set in the following sections.

6.3 LoRA Fine-tuning of BERT

To further adapt the pre-trained BERT model to our specific fMRI prediction task, we employ Low-Rank Adaptation (LoRA) [**hu2021lora**], a parameter-efficient fine-tuning technique. Instead of updating all the weights of the large BERT model, LoRA introduces small, trainable low-rank matrices into specific layers, significantly reducing the number of trainable parameters while often achieving performance comparable to full fine-tuning. The original weights of the BERT model remain frozen.

The embedding extraction process remains unchanged from the previous section, where we used the pre-trained BERT model to generate TR-level features. The only difference is that we now propagate gradients through the embedding extraction pipeline and capture the gradients of the LoRA parameters. This is possible because the extraction pipeline is fully differentiable.

The LoRA configuration and training procedure are as follows:

- **LoRA Target Modules:** We applied LoRA to the query and value projection matrices within each attention layer of the BERT model. The key and feed-forward dense layers were not modified with LoRA adapters in this setup, aligning with the standard practice in LoRA finetuning.
- **LoRA Rank (r):** We experimented with two LoRA ranks: $r = 4$ and $r = 8$. A lower rank results in fewer trainable parameters.
- **LoRA Alpha (α):** The LoRA scaling factor α was set to $2 \times r$, a common practice in LoRA.
- **LoRA Dropout:** A dropout rate of 0.1 was applied to the LoRA layers.
- **Bias Term for LoRA:** No bias term was added to the LoRA layers, nor was the bias term in the original BERT model modified.

For the regression heads (the linear layers mapping the 768-dimensional BERT output to voxel activities for each subject), we initialized them with the weights obtained from the best-performing "Regression Head" trained in the previous section, with a weight decay of 10^{-1}. This initialization strategy is intended to leverage the knowledge learned during the regression head training phase, providing a warm start for the classifiers, preventing the BERT's weights from being breaked by the regression head training.

The joint optimization of LoRA parameters and the (initialized) regression heads followed this procedure:

- **Optimizer:** AdamW [3].
- **Learning Rate:** A base learning rate of 2×10^{-3} was used, scaled linearly by the ratio of the current batch size (15) to a reference batch size of 75.
- **Weight Decay:** A weight decay of 10^{-1} was applied to all trainable parameters.
- **Training Data and Batching:** The model was trained on the designated training set (60% of stories), processed in mini-batches of 15 stories. The order of stories was shuffled at the beginning of each epoch.
- **Loss Function:** MSE between predicted and actual fMRI signals.
- **Epochs and Early Stopping:** Training proceeded for a maximum of 100 epochs. For each subject, the LoRA adapter weights and the corresponding regression head weights that yielded the lowest MSE on the validation set were saved independently.

The performance of the best models will be evaluated on the test set in the following sections.

6.4 Model Comparison - Pre-trained vs Fine-tuned (LORA)

In this section, we compare the performance of the pre-trained BERT model vs the model that we finetune using LORA. The results for CC across voxels, split by various metrics like Mean, Median, Top 1%, and Top 5% CC, are highlighted in Table 7. We consider these different metrics (as in prior labs) to have a more robust sense of how the models are performing across voxels, and whether certain models have more variance in their voxel performance.

We can see that overall, fine-tuning does improve the performance of the model as expected. However, the improvement is very relatively small in magnitude. For subject 2, the Mean CC improves by around 0.002, which represents an approximately 10% change. The upper echelons of voxels in terms of performance, meaning the Top 10% and Top 5% CC's, improve by a smaller relative change. Overall, we see a notable but small improvement performance across voxels, both throughout the mean/median and the best-performing voxels.

These experiments are repeated for Subject 3 in Table 8. We generally see the same takeaways, just with consistenly worse performances across both models and all metrics. This may indicate that it is harder to predict fMRI levels for Subject 3. Another note is that the improvement between pretrained and finetuned at the top-performing voxels for Subject 3 is relatively smaller than the improvement we see in Subject 2.

Table 7: Voxel-wise test-set correlation (CC) for Subject 2: fine-tuned vs. pretrained model

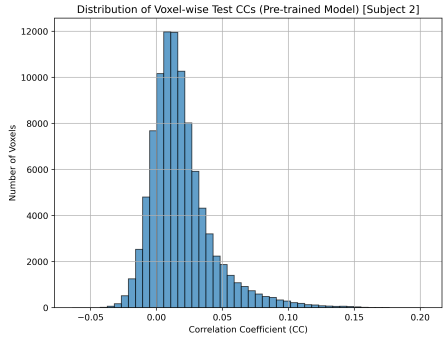
Model	**Mean CC**	**Median CC**	**Top 1% CC**	**Top 5% CC**
Pretrained	0.018732	0.014398	0.103731	0.064962
Fine–tuned	0.020627	0.015985	0.108326	0.069247

Table 8: Voxel-wise test-set correlation (CC) for Subject 3: fine-tuned vs. pretrained model

Model	**Mean CC**	**Median CC**	**Top 1% CC**	**Top 5% CC**
Pretrained	0.005317	0.005367	0.044171	0.030436
Fine–tuned	0.005947	0.005888	0.045137	0.031242

Additionally, we wanted to explore in greater depth how performance looks across the distribution of voxels for each model. These can be compared through a histogram of the CC distribution across voxels, shown in Figure 11. Through this, we can see that both the pretrained and finetuned model CC's follow roughly the same distribution in shape. The pretrained model, however, has a higher concentration of CC around the 0.01 region.

(a) Pre-trained Model



(b) Fine-tuned Model

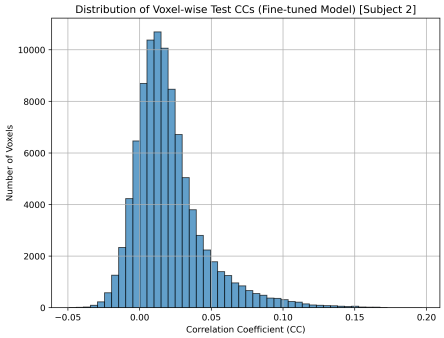


Figure 11: Test CC Distribution for the Pre-trained Model (left) and Fine-tuned Model (right) for Subject 2.

6.5 Model Comparison - All Models

We now compare the performance of the fine-tuned model and pre-trained model with all models from parts 3.1 and 3.2. This includes the model with encoder embeddings, Word2Vec, GloVe, and Bag of Words. The results are shown in Table 9

For the purpose of stability and completeness of our analysis, this comparison was repeated for Subject 3. These results are shown in Table 10.

We also compare against the visual distributions of the previous models (in Labs 3.1 and 3.2) to understand the specific performance across voxels. We can see that the pretrained and finetuned models have a more skewed distribution than the previous models, which have a more centered distribution. In general, the previous models also exhibit lower performance than the pretrained and finetuned models, which have more

Table 9: Voxel-wise test-set correlation coefficients (CC) for **Subject 2**.

Model	**Mean CC**	**Median CC**	**Top 1% CC**	**Top 5% CC**
Finetuned	0.020627	0.015985	0.108326	0.069247
Pretrained	0.018732	0.014398	0.103731	0.064962
Encoder	0.005989	0.005274	0.041602	0.029038
Word2Vec	0.005648	0.004466	0.050116	0.032150
GloVe	0.004951	0.004253	0.043254	0.029343
BoW	0.000904	0.000870	0.031131	0.021515

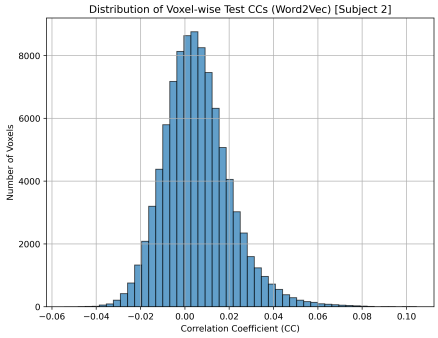
Table 10: Voxel-wise test-set correlation coefficients (CC) for **Subject 3**.

Model	**Mean CC**	**Median CC**	**Top 1% CC**	**Top 5% CC**
Finetuned	0.005947	0.005888	0.045137	0.031242
Pretrained	0.005317	0.005367	0.044171	0.030436
Encoder	0.011400	0.009270	0.068056	0.042578
Word2Vec	0.008742	0.006996	0.062659	0.038738
GloVe	0.008662	0.006905	0.061016	0.038540
BoW	0.001653	0.001390	0.032009	0.021303

parameters and are pretrained on a larger corpus. This is expected, as the pretrained and finetuned models are more complex and capable of capturing more complex patterns in natural language.

However, for Subject 3, the performance is quite different. The Encoder-based model, Word2Vec, and GloVe end up having better performance than the finetuned and pretrained models. This could indicate some instability in the result between the two subjects. Additionally, it could indicate that simpler baselines work better for Subject 3 prediction.

(a) Word2Vec Model



(b) Encoder Model

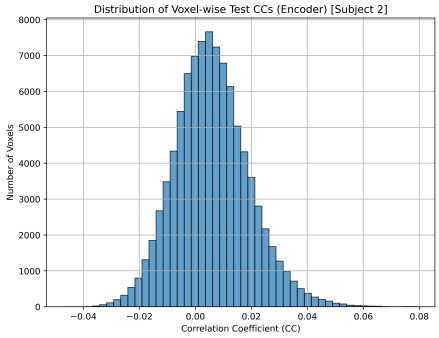


Figure 12: Test CC Distribution for the Word2Vec Model (left) and Encoder Model (right) for Subject 2.

7 Interpretation

In this section, we apply SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to interpret the fine-tuned BERT model's predictions by identifying the most influential words for two different test stories: `buck` and `lawsthatchokecreativity`. Annotated versions of each story, highlighting word-level importances, are included in Appendix B and Appendix C.

7.1 Voxel Selection

Although the fine-tuned model generates predictions for all 94,251 voxels in Subject 2 and 95,556 voxels in Subject 3, we limit our interpretability analysis to the voxels where the model performs well. This selective approach is crucial: methods like SHAP and LIME attempt to assign attribution based on a model's behavior, so they are only meaningful when the model's predictions are reliable. Interpreting poorly performing voxels would be misleading because feature attributions would reflect noise rather than a meaningful signal.

To identify the most informative voxels, we rank them by the CC between predicted and actual fMRI responses across time points. We then select voxels whose performance exceeds the 99.5th percentile (i.e., the top 0.5% of correlation scores). Table 11 shows the corresponding CC thresholds for each subject and story. From the eligible voxels, we randomly sample a subset for SHAP and LIME interpretation to keep the analysis computationally manageable.

Table 11: Top 0.5% CC Voxel Thresholds

Story	Subject 2	Subject 3
`buck`	0.216	0.294
`lawsthatchokecreativity`	0.159	0.206

7.2 SHAP & LIME Implementation

Both SHAP and LIME explain model predictions by perturbing the input features and measuring how those changes affect the output, attributing influence to each feature accordingly. In our case, the model takes BERT embeddings as input and outputs fMRI predictions, so we define a wrapper function around the model that returns predictions only for a subset of high-performing voxels. This allows SHAP and LIME to focus their explanations on the most informative regions of the brain.

For SHAP, we use `KernelExplainer`, a model-agnostic method that estimates Shapley values by fitting a locally weighted linear model around each prediction. For LIME, we use `LimeTabularExplainer`, which similarly fits local linear models to approximate feature importances. Both methods require a background dataset to serve as a reference distribution for generating perturbed samples. To ensure consistency and interpretability, we use the BERT embeddings from several training set stories as the background.

After applying SHAP and LIME, we obtain a three-dimensional array of explanation values with shape `(num_chunks, embedding_dim, num_voxels)`. To summarize these into word-level importance scores for each voxel, we take the absolute value of the explanation values and then average across the embedding dimension. The result is a matrix of shape `(num_chunks, num_voxels)` that reflects the relative importance of each word chunk to each voxel's predicted activation.

7.3 Test Story 1: `buck`

As shown in Figure 13, the word rankings produced by SHAP and LIME for `buck` are fairly consistent between Subject 2 and Subject 3, although the correlation is weaker than that observed for the raw attribution scores. This suggests that both SHAP- and LIME-based word importances are relatively stable across listeners to this story.

Interestingly, Figure 14 indicates that the words most highly ranked by SHAP differ substantially from those most highly ranked by LIME. In fact, their word rankings are weakly negatively correlated, implying that SHAP and LIME emphasize different features of the input when assigning importance. This points to complementary but largely divergent interpretations of model behavior by the two methods and suggests

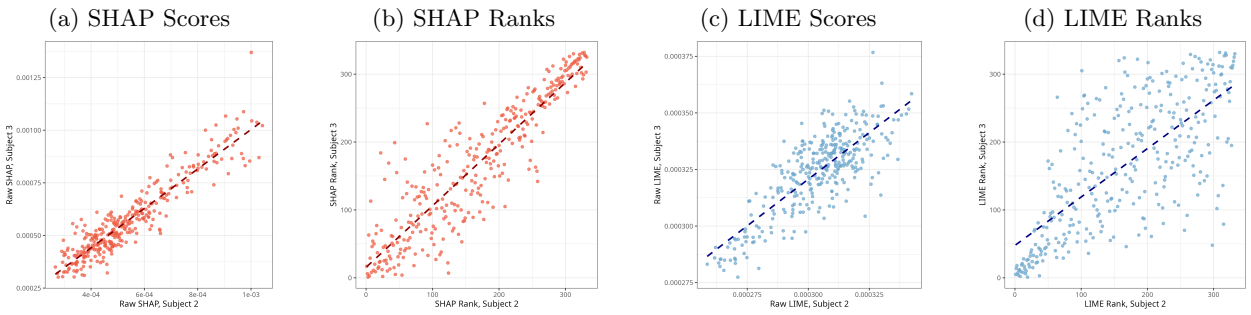


Figure 13: Scatterplots comparing word-level SHAP (red) and LIME (blue) values between Subject 2 and Subject 3. Scores (left) show strong cross-subject agreement, while ranks (right) exhibit greater dispersion—indicating consistency in which words matter, but differences in their relative importance.

instability.

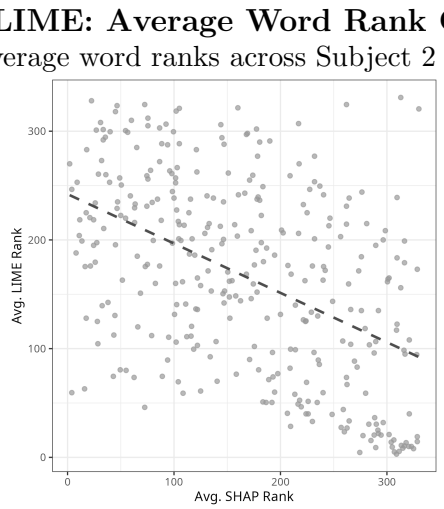


Figure 14: The weak negative correlation $r = -0.35$ between SHAP ranks (x-axis) and LIME ranks (y-axis) suggests that the two methods often prioritize different sets of words, even when averaged across subjects.

In `buck`, SHAP tends to highlight:

- **Emotional experiences** (e.g., "to eat from this spread", "no one had ever asked me what i want")
- **Prison dynamics** (e.g., "prison", "walk the one gate after another", "hundred guys waiting", "[out] of prison they know you on parole")
- **Identity markers** (e.g., "know this ink work this is prison ink", "now i speak a lot of different gangster languages", "understand blood i even speak a little")
- **Transitional phrases** (e.g., "so", "because for the past ten years", "and leave out of here")

Meanwhile, LIME tends to highlight:

- **Temporal markers** (e.g., "after twenty six years", "corridor for the last time", "through that last gate")
- **Quantitative details** (e.g., "thirty seven eighty seven five report", "with two hundred dollars in it", "no more than eight feet away")
- **Action sequences** (e.g., "they load us on a van", "i get to r and r", "washing dishes my wife says")

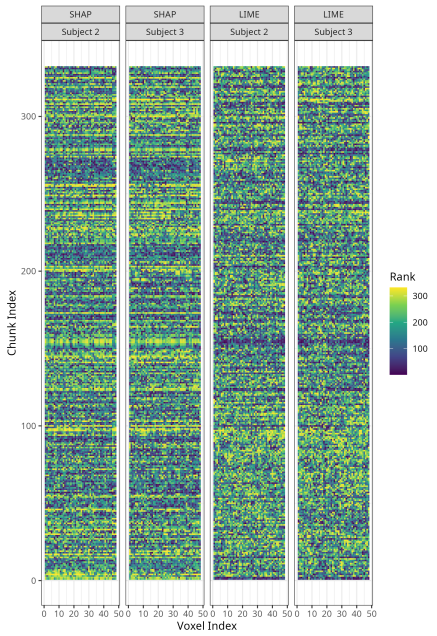
Curiously, words identified by both methods almost exclusively involve **interactions with the speaker's wife** (e.g., "gave her a quick kiss whispered in", "call out to my wife i said hey babe", "i could see her she was snapping pictures", "and he tells my wife wait").

This pattern suggests that these two methods can provide complementary insights about which textual features drive fMRI responses: SHAP may better capture emotional and thematic elements, while LIME may be more effective at identifying sequentially important narrative details. For a more comprehensive list of words identified by each method, see Appendix B.

Voxelwise Consistency of SHAP and LIME Word Ranks

Distribution of word rankings across individual voxels and subject-score combinations

(a) Voxelwise Rank Heatmap



(b) Interquartile Range of Word Ranks

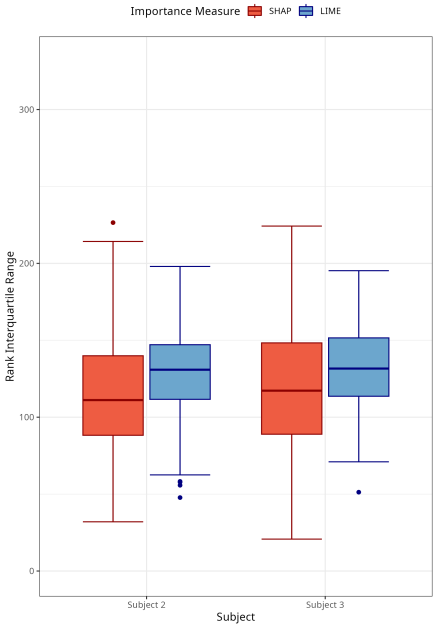


Figure 15: Word ranks assigned by individual voxels show a high degree of consistency within and across subjects. The heatmap (left) displays voxelwise rankings for 332 word chunks, revealing horizontal banding patterns that reflect agreement in word importance. The boxplot (right) summarizes the spread of these ranks within chunks. For the average word, the IQR spans only ∼125 ranks, suggesting relatively stable importance scores across voxels.

Regarding differences across voxels, Figure 15 reveals visible horizontal banding in the heatmaps for both SHAP and LIME, indicating that word rankings are relatively consistent across voxels within and between subjects. The accompanying boxplot shows that SHAP produces slightly lower median IQRs for word ranks across voxels, suggesting less variability and lower voxel sensitivity for the average word. Conversely, LIME exhibits a tighter distribution of IQR values, indicating that while its average variability may be higher, the spread among the least consistently ranked words is smaller.

7.4 Test Story 2: `lawsthatchokecreativity`

As shown in Figure 16, the word rankings produced by SHAP and LIME for `lawsthatchokecreativity` are fairly consistent between Subject 2 and Subject 3, although the correlation is weaker than that observed for the raw attribution scores. This suggests that both SHAP- and LIME-based word importances are relatively stable across listeners to this story.

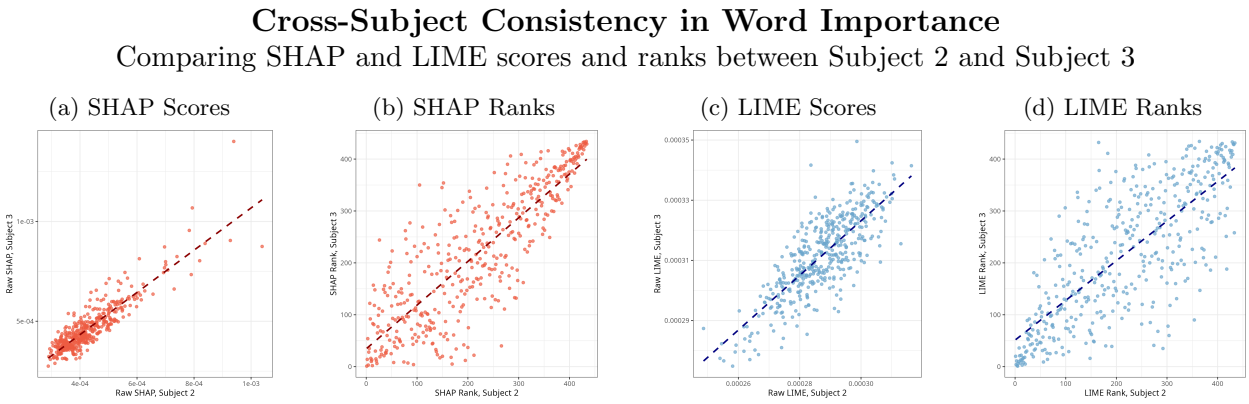


Figure 16: Scatterplots comparing word-level SHAP (red) and LIME (blue) values between Subject 2 and Subject 3. Scores (left) show strong cross-subject agreement, while ranks (right) exhibit greater dispersion—indicating consistency in which words matter, but differences in their relative importance.

Interestingly, Figure 17 indicates that the words most highly ranked by SHAP differ substantially from those most highly ranked by LIME. In fact, their word rankings are weakly negatively correlated, implying that SHAP and LIME emphasize different features of the input when assigning importance. This points to complementary but largely divergent interpretations of model behavior by the two methods.

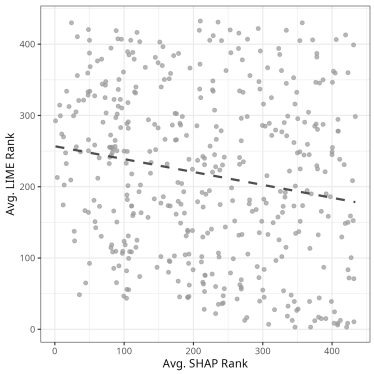


Figure 17: The weak negative correlation $r = -0.12$ between SHAP ranks (x-axis) and LIME ranks (y-axis) suggests that the two methods often prioritize different sets of words, even when averaged across subjects.

In `lawsthatchokecreativity`, SHAP tends to highlight:

- **Concrete entities and specific references** (e.g., "john philip sousa", "united states capitol", "four hundred and forty eight percent", "supreme court considered")
- **Technical legal terminology** (e.g., "will be eliminated by a process of evolution", "copyright law at its core regulates", "judgement of fair use")
- **Action-oriented phrases** (e.g., "singing the songs", "take sounds and images", "traveled to this place")

Meanwhile, LIME tends to highlight:

- **Abstract concepts** (e.g., "a culture which is top down owned", "it is a literacy", "weird time it's kind of age of prohibitions")
- **Temporal and spatial references** (e.g., "across the country well in nineteen forty five", "at that time this legal", "and in nineteen forty one")
- **Cultural commentary** (e.g., "this is a picture of culture", "celebrating amateur culture", "it is how our kids think")
- **Technology-related expressions** (e.g., "with access to a fifteen hundred dollar computer", "instinct the technology produces", "can't stop our kids from using")

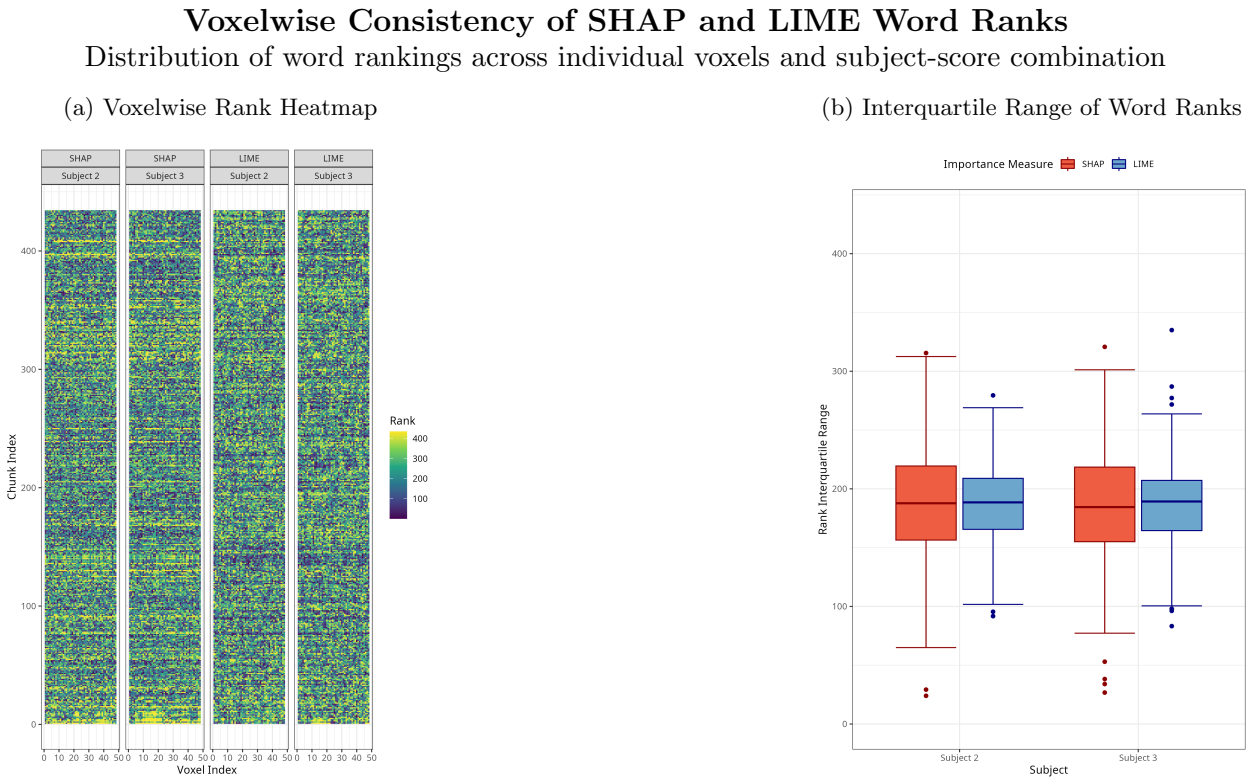


Figure 18: Word ranks assigned by individual voxels show a high degree of consistency within and across subjects. The heatmap (left) displays voxelwise rankings for 434 word chunks, revealing horizontal banding patterns that reflect agreement in word importance. The boxplot (right) summarizes the spread of these ranks within chunks. For the average word, the IQR spans only ∼180 ranks, suggesting relatively stable importance scores across voxels.

However, words identified by both methods include:

- **Key thesis statements** (e.g., "i'm gonna tell you three stories", "now so instead what we need is")
- **Narrative transitions** (e.g., "and that's where the story of", "extremism on the other a fact we should have many many")
- **Opinion indicators** (e.g., "i think much more important much", "to fight for i as any good")

This pattern suggests that fMRI responses may be driven by both specific semantic content (captured by SHAP) and broader narrative structure (captured by LIME), reflecting the multi-level processing that occurs in the brain during language comprehension. For a more comprehensive list of words identified by each method, see Appendix C.

Regarding differences across voxels, Figure 18 reveals visible horizontal banding in the heatmaps for both SHAP and LIME, indicating that word rankings are relatively consistent across voxels within and between subjects. The accompanying boxplot shows that SHAP produces slightly lower median IQRs for word ranks across voxels, suggesting less variability and lower voxel sensitivity for the average word. Conversely, LIME exhibits a tighter distribution of IQR values, indicating that while its average variability may be higher, the spread among the least consistently ranked words is smaller.

8 Conclusion

This lab demonstrated the efficacy of leveraging pre-trained language models for predicting fMRI responses to narrative stimuli. Our experiments focused on utilizing BERT, first with static embeddings and then through parameter-efficient fine-tuning using LoRA.

The results indicate that embeddings from the pre-trained BERT model significantly outperform those from earlier labs (custom encoders or simpler static embeddings like Bag-of-Words) for Subject 2, and show competitive results for Subject 3. This highlights the benefit of the rich contextual representations learned by larger models. Furthermore, fine-tuning these BERT embeddings with LoRA provided an additional, albeit modest, improvement in predictive accuracy across both subjects, underscoring the value of task-specific adaptation even with limited trainable parameters. The choice of LoRA rank ($r = 8$) and initializing regression heads from prior experiments proved beneficial.

Interpretability analysis using SHAP and LIME on the fine-tuned LoRA-BERT model identified words relevant to the narrative content that influenced predictions for well-modeled voxels. While SHAP and LIME offered somewhat divergent perspectives—SHAP often emphasizing thematic content and LIME highlighting structural or specific details—their combined insights suggest that the model attends to various meaningful linguistic features. Word importance showed reasonable consistency across subjects and voxels, indicating some shared neural processing patterns captured by the model.

In summary, this work confirms that pre-trained transformers, further refined with techniques like LoRA, provide a robust framework for modeling neural language processing. They offer improved predictive power and, through interpretability methods, can shed light on the linguistic features driving brain activity.

References

[1] Jacob Devlin et al. "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)*. 2019, pp. 4171–4186.

[2] Shailee Jain and Alexander Huth. "Incorporating context into language encoding models for fMRI". In: *Advances in neural information processing systems* 31 (2018).

[3] Ilya Loshchilov and Frank Hutter. "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017).

[4] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[5] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

A Academic honesty

A.1 Statement

We affirm that the work in this report is entirely my own. We have not copied from any unauthorized sources, and all contributions from classmates, external sources, or tools are acknowledged. Academic research honesty is necessary because it ensures fairness, builds trust in scholarly work, and reflects personal integrity. Misrepresenting work undermines academic standards and disrespects the time and effort of peers and educators. Maintaining honesty in research fosters a learning environment where collaboration and progress can thrive authentically.

A.2 LLM Usage

We used ChatGPT to assist in clarifying concepts, creating visualizations, checking grammar, and improving the structure of our explanations. No content of the report or code was generated by the LLM without our review, editing, and refinement. We ensured that all content was written and understood by us, and the LLM was used as a tool to enhance our work rather than replace our understanding, we take full responsibility for all content in the report.

B buck Full Text

Words discovered by SHAP in red, LIME in blue, or both in **bold**. Threshold based on 75^{th} percentile word ranks.

so as i sit down to eat from this spread that has been given in my honor i can't help but to f notice the two feelings that i'm having one i'm a little nervous two i'm very excited because you see in the morning i'll be paroling from state prison after twenty six years this spread is given to me not as like to say hey homie we appreciate you nah it was a spread to say we don't never wanna see your ass in here again and they broke out the finest of the finest top ramen chili in a can uh roast beef in a can you know this this top line of stuff right so after the spread has been had it's time for me to give away some property such as my tv my radio and some self help books that helped me a lot so they'll be better utilized left behind nine o'clock rolls around it's lights out you don't have to go to sleep but you do gotta get off the day room floor it's cool by me i'm leaving in the morning so i may have gotten like three hours worth of sleep that night i slept ready roll and that means i slept with my clothes on so once they call my name i'm out of there bye see you later so five forty five rolls around the night man says cyprien d thirty seven eighty seven five report to r and r i walk down that corridor for the last time walk the one gate after another till i get to r and r and that's receiving and release and i'm on the release end at this point i get to r and r and there's about seven to six other guys waiting for the same thing they start fingerprinting me mugshots and everything i'm cool like time to go so they load us on a van and we drive through one gate after another gate till we get to the final gate the cop says alright everybody off the van you'll step to the man at the base of the tower there and you'll give him some personal information so we did and he gave each man an envelope with two hundred dollars in it better known as gate money we all got back onto the van and drove through that last gate we made it around to the visitor's parking lot and he asked is there anyone with a ride i do i was the only one to get off the van these other guys that two hundred dollar gate money they had to spend it on buying their bus tickets maybe back to southern california or even further up north but i was the lucky one my wife was waiting there with her camera **i could see her she was snapping pictures** of me like a paparazzi i walk briskly towards her **gave her a quick kiss whispered in** her ear let's get out of here before they say we made a mistake we got in the car drove away i saw a seven eleven i was like hey babe pull over lemme i wanna go in and get something so i went in and i grabbed a pack of big red gum i hadn't had chewing gum in twenty six years big red was my favorite we got back in the car drove away we were on the road for about thirty to forty five minutes and i s i spotted a target i said hey babe that's target i saw that on the uh tv pull over lemme i wanna go in there so we did the first thing that caught my eye was how big the shopping basket was it was bigger than the space that i had to store my personal property while in prison i grabbed the basket and went down the aisle grabbing various things such as underwear socks t-shirt toothbrush toothpaste you name it i grabbed it well there goes that two hundred dollar gate money so back in the car my wife says so what do you want for dinner whoa she's stumped me with that one because for twenty six years no one had ever asked me what do i want for dinner my dinner came through a thirteen by thirteen hole in the wall you stick your hand in there and pull out a tray and what's ever on there that's what you eat so she says i know what to fix so she drove we made it to the grocery store went into the grocery store this gigantic store berkeley bowl so i made it to the produce section i'm stuck there are so many oranges to choose from and the reason and i'm stuck just looking at the oranges is because for the past ten years in prison they had taken oranges away from us because they were used to make wine so there's like a bumpy orange a smooth orange a navel orange i even met my first blood orange and at the same time this wave hit me of alone and feeling like i don't know a nervous feeling hit me like i'm the only person in this store who's paroled from a prison today after twenty six years i'm a little stuck a little so i'm looking around and i spot this ink work it's not the typical ink work that i see out here now like with colors no i know this ink work this is prison ink he has a prison guard tower on him a meltic clock with time i know that work i've seen that many times so i walk over to him and i whisper so what yard did you parole from he said oh me i got i paroled from pelican bay like four months ago i'm like and they know you got out of prison they know you on parole aw yeah they not tripping here this is berkeley man nah nobody's tripping well i walked away from him with the understanding that if he can get a job so can i so i met my wife at the checkout once again there's another stunned moment for me there are twelve checkout lines and there's somebody on the other end bagging the groceries in prison there was one window maybe thirteen by thirteen again and there was like three hundred guys waiting to get into this one window with their lists i didn't have to worry about an alarm going off or anything no getting down we're gonna get our groceries and leave out of here that was my thought cool so we made it through the made it through the checkout went back to the car got in the car and we're on the road and we come to a stoplight and this homeless guy had a sign that says anything will help so my wife reaches into the cup holder and gets ready to hand him some change and all of a sudden he gets a phone call **and he tells my wife wait** i gotta take this call and my thoughts was like ih if he got a cellphone i gotta get one so we drive away and my wife uh she just drives real fast right and we make it to the neighborhood just as fast as we got to the store and she pulls into the neighborhood and she points out the various neighbors' house um this is where fred and mike live the brewers live here and i'm like okay and we get to our house and she said this is us and we live at the edge of the forest i'm like cool cool and so we went inside and she gives me the grand tour and so she says well i'm gonna prepare this dinner i said cool so she made steak and lobster with small potatoes nothing out of a can folks nothing out of a packet either she made this meal we sat down we ate and i had my first glass of wine with my meal not them i'm not saying i've never had wine before i've never had wine with a meal this is some grown folks' shit i thought right so after the meal was finished we're cleaning up washing dishes my wife says you wanna get further acclimated i'm like yeah she said take out the garbage no problem i reach down tie that garbage bag up throw it over my shoulder and out to the front door i go when i open up the door no more than eight feet away from me was a ten point buck just standing there looking at me look at it standing staring at me now had we been on a prison **yard me and this buck woulda had issues** now i speak a lot of different gangster languages i'm fluent in crip i understand blood i even speak a little sh olo vato loco but i don't speak deer so i shut the door i **call out to my wife i said hey babe** there's somebody here to see you she came to the door opened it and saw there was the buck she grabbed the trashbag and walked right past it i said damn she speaks deer she came back and we both look out the door together at this buck who's still standing there just looking just majestic as could be and my wife says you don't need to trip him he's just one neighbor coming to welcome you to the neighborhood cool i thought like well the fellas aren't gonna believe this shit when i tell them this

C `lawsthatchokecreativity` Full Text

Words discovered by SHAP in red, LIME in blue, or both in **bold**. Threshold based on 75^{th} percentile word ranks.

i'm gonna tell you three stories on the way to one argument that's **going to tell you a little bit about how we open user generated content up for business so here's the** first story nineteen oh six this man john philip sousa traveled to this place the united states capitol to talk about this **technology what he called the quote** talking machines sousa was not **a fan of the talking** machines this is what he had to say these talking machines are going to ruin artistic development of music in this country when **i was a boy in front of every house in the summer evenings you** would find young people together singing the songs of the day **or the old songs** today you hear these infernal machines going night and day we will not have a vocal cord left sousa said the vocal cords will be eliminated by a process of evolution **as was the tail of** man when he came from the ape this is a picture of culture we could describe it using modern computer terminology as a kind of read write culture it's a culture where people participate in the creation and the re creation of their culture in that sense it's read write sousa's fear was that we would lose that capacity because of these quote infernal machines they would take it away and in its place we'd have the opposite of read write culture what we could call read only culture culture where creativity was consumed but the consumer is not a creator a culture which is top down owned where the vocal cords of the millions have been lost now as you look back at the twentieth century at least in what we think of as the quote developed world hard not to conclude that sousa was right never before in the history of human culture had it been as professionalized never before as concentrated never before has creativity of the millions been as effectively displaced and displaced because of these quote infernal machines the twentieth century was that century where at least for those places we know the best culture moved from this read write to read only existence so second cg land is a kind of property is property it's protected by law as lord blackstone described it land is protected by trespass law **for most of the history of trespass** law by presuming it protects the land all the way down below and to an indefinite extent upwards now that was a pretty good system for most of the history of the regulation of land until this technology came along and people began to wonder were these instruments trespassers as they flew over land without clearing the rights of the farms below as they traveled across the country well in nineteen forty five supreme court got a chance to address that question two farmers thomas lee and tinie causby who raised chickens had a significant complaint because of these technologies the complaint was that their chickens followed the pattern of the airplanes and flew themselves into the walls of the barn when the airplanes flew over the land and so they appealed to lord blackstone to say these airplanes were trespassing since time immemorial the law had said you can't fly over the land without permission of the landowner so this flight must stop well the supreme court considered this hundred years tradition and said **in an opinion written by justice douglas** that the causbys must lose the supreme court said the doctrine protecting land all the way to the sky has no place in the modern world otherwise every transcontinental flight would be subject would subject the operator to countless trespass suits common sense a rare idea in the law but here it was common sense revolts at the idea common sense cg finally cg before the internet the last great terror to rain down on the content industry was a terror created by this technology broadcasting a new way to spread content and therefore a new battle over the control of the businesses that would spread content now at that time the entity the legal cartel that controlled the performance rights for most of the music that would be broadcast using these technologies was ascap they had an excusi exclusive license on the most popular content and they exercised it in a way that tried to demonstrate to the broadcasters who really was in charge so between nineteen **thirty one and nineteen thirty nine they** raised rates by some four hundred and forty eight percent until the broadcasters finally got together and said okay enough of this and in nineteen thirty nine a lawyer sydney kaye started something called broadcast music incorporated know it as bmi and bmi was much more democratic **in the art that it would include within** its repertoire including african american music for the first time in the uh repertoire but most important was that bmi took public domain works and made arrangements of them which they gave away for free to their subscribers so that in nineteen forty when ascap threatened to double their rates the majority of broadcasters switched to bmi now ascap said they didn't care the people will revolt they predicted because the very best music was no longer available because they had shifted to the second best public domain provided by bmi well they didn't revolt and in nineteen forty one ascap cracked and the important point to recognize is that even though these broadcasters were broadcasting something you would call second best that competition was enough to break at that time this legal cartel over access to music okay three stories here's the argument in my view the most significant thing to recognize about what this internet is doing is its opportunity to revive the read write culture that sousa romanticized digital technology is the opportunity for the revival of these vocal cords that he spoke so passionately to congress about user generated content spreading in businesses in extraordinarily valuable ways celebrating amateur culture by which i don't mean amateurish culture i mean culture where people produce for the love of what they're doing and not for the money i mean the culture that your kids are producing all the time for when you think of what sousa romanticized in the young people together singing the songs of the day of the old songs you should recognize what your kids are doing right now taking the songs of the day and the old songs and remixing them to make them something different it's how they understand access to this culture i'm not talking about nor justifying people taking other people's content in wholesale and distributing it without the permission of the copyright owner i'm talking about people taking and recreating using other people's content using digital technologies to say things differently it is now anybody with access to a fiften hundred dollar computer who can take sounds and images from the culture around us and use it to say things differently these tools of creativity have become tools of speech it is a literacy for this generation this is how our kids speak it is how our kids think it is what your kids are as they increasingly understand digital technologies and their relationship to themselves now in response to this new use of culture using digital technologies the law has not greeted this sousa revival with very much common sense instead the architecture of copyright law and the architecture of digital technologies as they interact have produced the presumption that these activities are illegal because if copyright law at its core regulates something called copies then in the digital world the one fact we can't escape is that every single use of culture produces a copy every single use therefore requires permission without permission you are a trespasser common sense here though has not yet revolted in response to this response that the law has offered to these forms of creativity instead what we've seen is something much worse than a revolt there's a growing extremism that comes from both sides in this debate in response to this conflict between the law and the use of these technologies one side builds new technologies **such as one recently announced that will enable them to automatically take** down from sites like youtube any content that has any copyrighted content in it whether or not there's a judgment of fair use that might be applied to the use of that content and on the other side among our kids there's a growing copyright abolitionism a generation that rejects the very notion of what copyright is supposed to do rejects copyright and believes that the law is nothing more than an ass to be ignored and to be fought at every opportunity possible the extremism on one side begets **extremism on the other a fact we should have learned many many** times over and both extremes in this debate are just wrong now the balance that i try **to fight for i as any good** liberal try to fight for first by looking to the government total mistake right looked first to the courts and the legislatures to try to get them to do something to make the system make more sense it failed partly because the courts are too passive partly because the legislatures are corrupted by which i don't mean that there's bribery operating to stop real change but more the economy of influence that governs how congress functions means that policymakers here will not understand this until it's too late to fix it so we need something different we need a different kind of solution and the solution here in my view is a private solution a solution that looks to legalize what it is to be young again and to realize the economic potential of that **and that's where the story of bmi** becomes relevant because as bmi demonstrated competition here can achieve some form of balance the same thing can happen now we don't have a public domain to draw upon **now so instead what we need is** two types of changes first that artists and creators embrace the idea choose that their work be made available more freely so for example they can say their work is available freely for non commercial this amateur type of use but not freely for any commercial use and second we need the businesses that are building out this read write culture to embrace this opportunity expressly to enable it so that this ecology of free content or freer content can grow on a neutral platform where they both exist simultaneously so that more free can compete with less free and the opportunity to develop the creativity in that competition can teach one the lessons of the other now i would talk about one particular such plan that i know something **about but i don't want violate ted's first** commandment of selling so i'm not gonna talk about this at all i'm instead just gonna remind you of the point that bmi teaches us that artist choice is the key for new technology having an opportunity to be open for business and we need to build artist choice here if these new technologies are to have that opportunity but let me end with something **i think much more important much** more important than business it's the point about how this connects to our kids we have to recognize they're different from us it is technology that has made them different and as we see what this technology can do we need to recognize you can't kill the instinct the technology produces we can only criminalize it we can't stop our kids from using it we can only drive it underground we can't make our kids passive again we can only make them quote pirates and is that good we live in this weird time it's kind of age of prohibitions where in many areas of our life we live life constantly against the law ordinary people live life against the law and that's what i we are doing to our kids they live life knowing