**Team:**

Will Schulte

Ahmad Sesay

Shi Feng Wu

Terrence Fitzhugh

**Masked Language Model (MLM) BERT Text Filler**

Our project was to create a MLM model using the BERT python library, and fine tune it on the complete works of Jane Austen to test its accuracy and viability on both her own novels and on other types of texts. We used three models throughout our project, the pretrained bert-base-uncased, a model fine tuned on Jane Austen for 10 epochs, and a model fine tuned for 3 epochs. The purpose of the model is to fill in missing tokens from text based on the context around it, as shown below:

**Jane Austen Model 10 Epochs Test (“Persuasion, 1817”):**

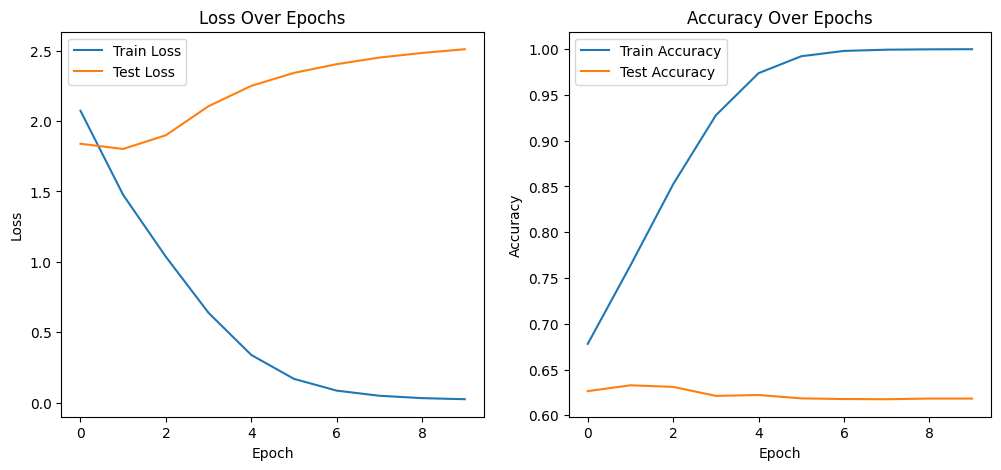
| Snippet from “Persuasion” (1818): | Filled in: |
| --- | --- |
| Lady Russell was fond of Bath, in short, and disposed to think it must  suit them all; and as to her young friend's health, by passing all the  warm months with her at Kellynch Lodge, every danger would be avoided;  and it was in fact, a change which must do both health and spirits  good. Anne had been too little from home, too little seen. Her spirits  were not high. A larger society would improve them. She wanted her to  be more known. | Lady russell was fond of Bath, in short, and seemed to think it must  suit them all; and as to her young friend's health, by spending all the  warm months with her at Kellynch Lodge, every danger would be avoided;  and it was in effect, a change which must make both health and spirits  good. Anne had been too little from home, too little seen. Her spirits  were not high. A larger society would improve them. She wanted her to  be more known. |

After giving the Jane Austen model a snippet from “Persuasion”, the model did not fully reproduce the same words as the original text. However, the words that were put in place of the original words still managed to keep the sentence structure grammatically correct and fitting with the context of the text. The models don’t often get the text-specific token correct, aside from some exceptions in cases where only one or few tokens fit. This can be seen when given phrases or grammatical structures that are the same across all texts such as the above example with “Her spirits were” and “She wanted her to”. The model also successfully filled in the name of “Lady Russell” as she is the most prominent character in the novel whose name starts with that prefix.

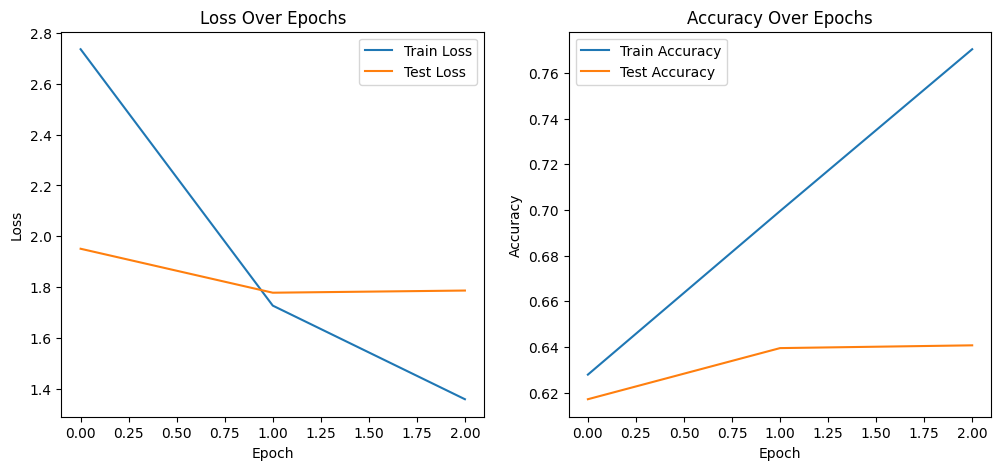
The model also shows that its predictions are influenced by the context around them. For example, the sentence “I went to the [MASK] the other day and bought an apple.” gets filled in to “I went to the market the other day and bought an apple.”. This makes sense in the context that one would go to the market or store to buy food. This means that the model is not only looking to find grammatically correct tokens to fill, it is also looking for the ones that make the most contextual sense.

When training the model, we realized the complexity of the bert-base-uncased model we were using was causing the model to overfit as shown in the two graphs below. By the time the training had reached the fifth epoch, the model had displayed almost 100% accuracy with very low training loss. After observing this, we attempted to find more simplistic models to use like distiledBERT however the issue of overfitting given the same hyperparameters persisted. To combat this, we decided to train the models on three epochs (shown in the second graph image) to reduce overfitting and record their accuracy.

Loss and accuracy for the Jane Austen model trained over 10 epochs:

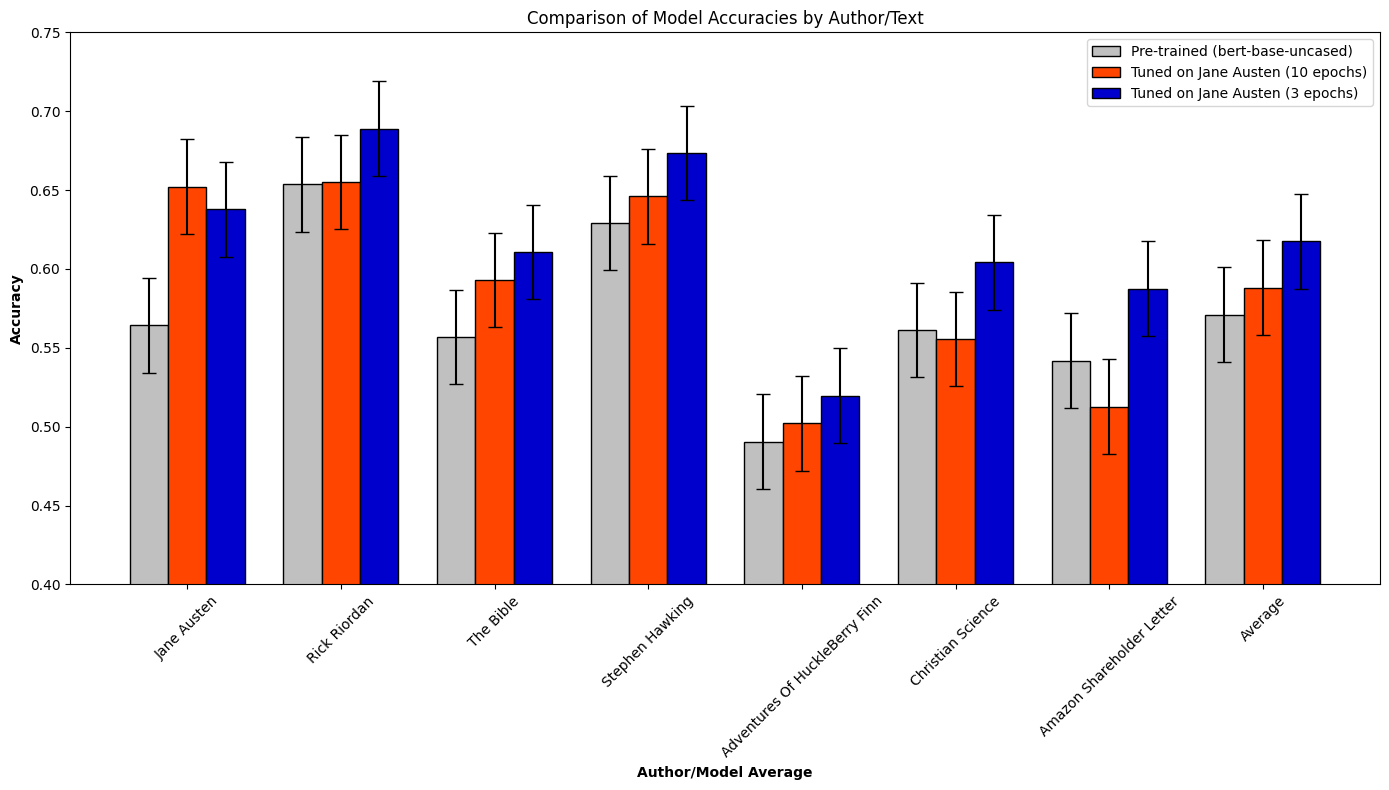


Loss and accuracy for the Jane Austen model trained over 3 epochs:



During our testing phase, we used three different models. The first model was the pre-trained bert-base-uncased model, the second was the overfitted Jane Austen model, and the third was the Jane Austen model trained over three epochs. From the data that we retrieved, the overtrained model did the best on works from Jane Austen which was to be expected since it trained the longest on her works. A side effect of this training was that while the overfitted model did better than the base model, the Austen model trained on three epochs had higher accuracy in most works outside of Austen's works. This data shows us why it is important not to train a model for too long as it will lose its ability to function at higher accuracies in subjects outside the training data.

Accuracies for three models with various texts:



One thing that we observed while testing the models was the total amount of masks and the frequency of masks in a single sentence. When testing both models, we found that when multiple masks are used in a sentence, the model does not have the context to generate grammatically correct tokens. We showcase two examples of this below.

**Multiple Mask Examples:**

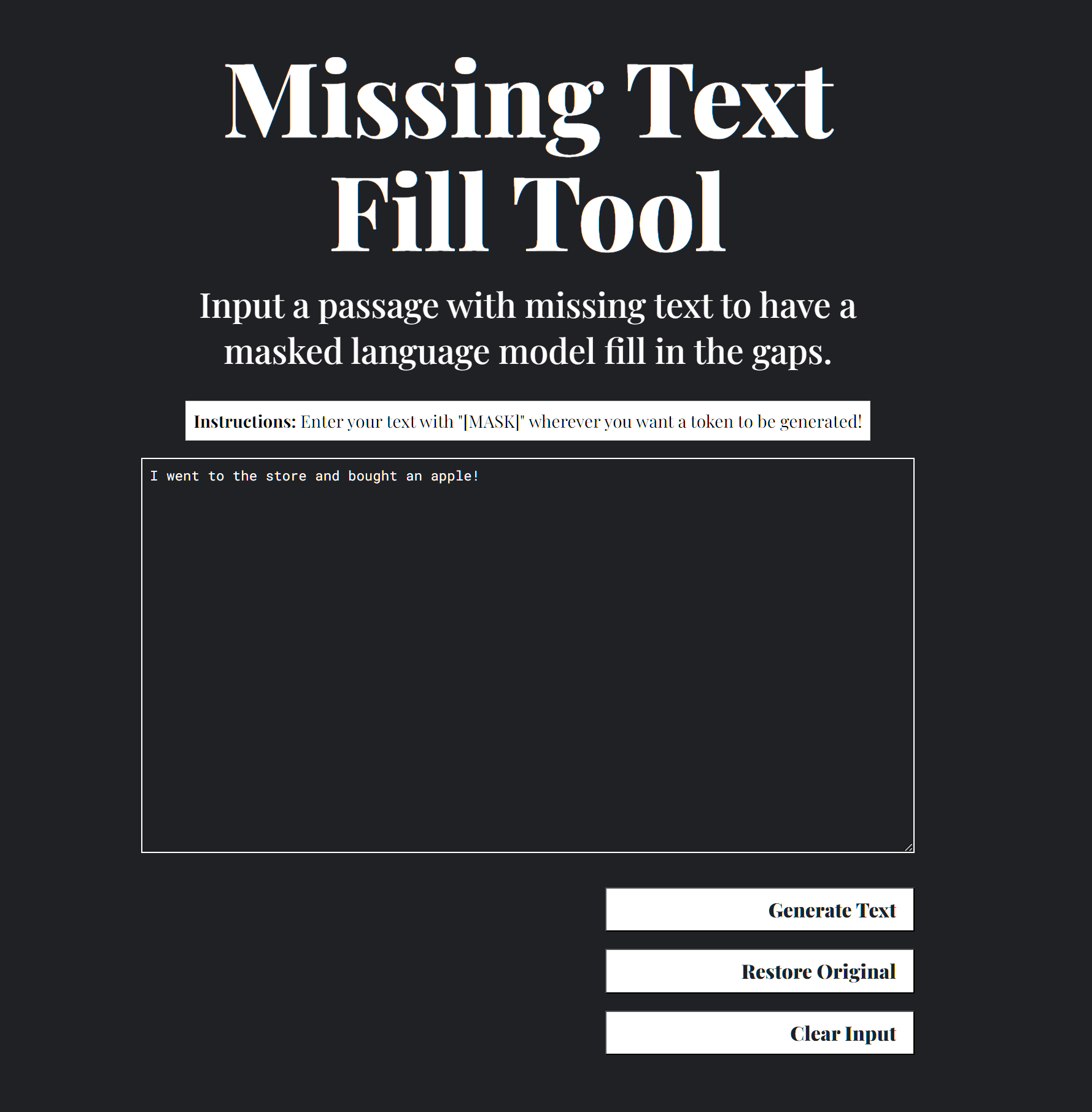
**Yellow highlighted tokens show where the text was masked, red highlighted tokens show the filled in text from the model.**

| Sample Text: | Filled in: |
| --- | --- |
| "Hello, my name is [MASK]. I [MASK] [MASK] [MASK] [MASK] [MASK]. I also [MASK] [MASK] [MASK]. [MASK] [MASK] [MASK] [MASK]" | Hello, my name is jennifer. I am have my , johnson. I also have a you. hello hello hello . |

| Sample Text: | Filled in: |
| --- | --- |
| The Gardiners stayed a week at Longbourn; and what with the Philipses,  the Lucases, and the officers, there was not a day without its  engagement. Mrs. Bennet had so carefully provided for the entertainment  of her brother and sister, that they did not once sit down to a family  dinner. | The Gardiners , , , lord Longbourn; and for the the others,  the others, and the officers, there was not a day without its  engagement. Mrs. Bennet was so so such the the entertainment  of her brother and sister, that they did not once sit down to a family  dinner. |

**GUI:**

A webpage was made using flask to run a test version of the model that allows users to enter their own sentences while indicating where tokens need to be filled, and the model would then generate the complete version of the text.



**Purpose and reflection:**

The purpose of the model was to be able to restore texts with missing parts quickly and without need for human intervention. While the model may not be completely accurate, it is accurate enough to fix missing data with passable words that are grammatically correct for the context. This model could also be used in many other circumstances such as a grammar checker to replace words and even in conjunction with other language models to improve contextual understanding. The model does accomplish its purpose by being able to restore blanks in text, but it is still limited in its applications. The model cannot consistently recreate a token that accurately matches what was masked. However, it can fix broken texts so that they remain grammatically correct. This fulfills the goals of the model to an extent as the user will be able to better understand the restored text. However, the issue of large chunks of texts missing remains an issue as it prevents the model from having the necessary context to select grammatically correct tokens.

Our biggest problem with our project is the quick overfitting of data that comes from the relatively small training set. In the future, we might have chosen a dataset that has more to work with or attempted to stretch the data further through random noise, as even the complete works of Jane Austen is too little data for such a complicated model to effectively train on. The utility of this project would also benefit greatly from the ability to effectively fill in multiple missing tokens at once, at least being able to maintain grammatical correctness. This would require a much larger context window than what we are currently using if we would want to maintain contextual stability as well.

**Contributions:**

GUI: Shi Feng

Data Collection and Processing: Terence

ML Training: Will

Analysis and Writeup: Ahmad