Finetuning Analysis Report

1. Introduction

This report provides an in-depth analysis of the finetuning job and an dataset analysis. Using the finetuned model and the dataset, we generated text embeddings and applied various techniques to visualize and interpret the data. The analysis includes t-SNE visualization, UMAP projection with KMeans clustering, anomaly detection, and topic identification using KMeans clustering. Additionally, we generated a word cloud to visualize the most frequent words in the dataset. The report aims to provide insights into the underlying structure of the text data and identify potential outliers or anomalies. The following sections present the results of the analysis along with visualizations and interpretations.

2A. Training and Validation Loss Improvements

The table below shows the percentage improvement in training and validation losses over the course of training.

Loss Improvement Metrics

Metric	Value
Train Loss Improvement	21.75%
Validation Loss Improvement	26.51%

2B. Training Loss Over Steps

The learning curve has not plateaued, indicating that the model was still making progress in learning from the training data. The percentage change in training loss over the last 5 steps is approximately 29.60%, suggesting that further training could potentially lead to improvements.

Overall a percentage difference of 21.75% in the training loss was observed. The training loss plot is crucial for understanding the model's learning progress over time. A downward trend in the loss indicates that the model is learning and improving its predictions. If the last few evaluation losses are relatively constant, it suggests that the model's learning has plateaued, indicating that the model may have reached its learning capacity on the given dataset. However, if the evaluation loss continues to decrease, even subtly, it could mean that the model may benefit from additional training epochs, as there's potential for further convergence. In practice, a balance must be struck to avoid overfitting, where the model learns the training data too well, including its noise and outliers, which can harm generalization to new data.
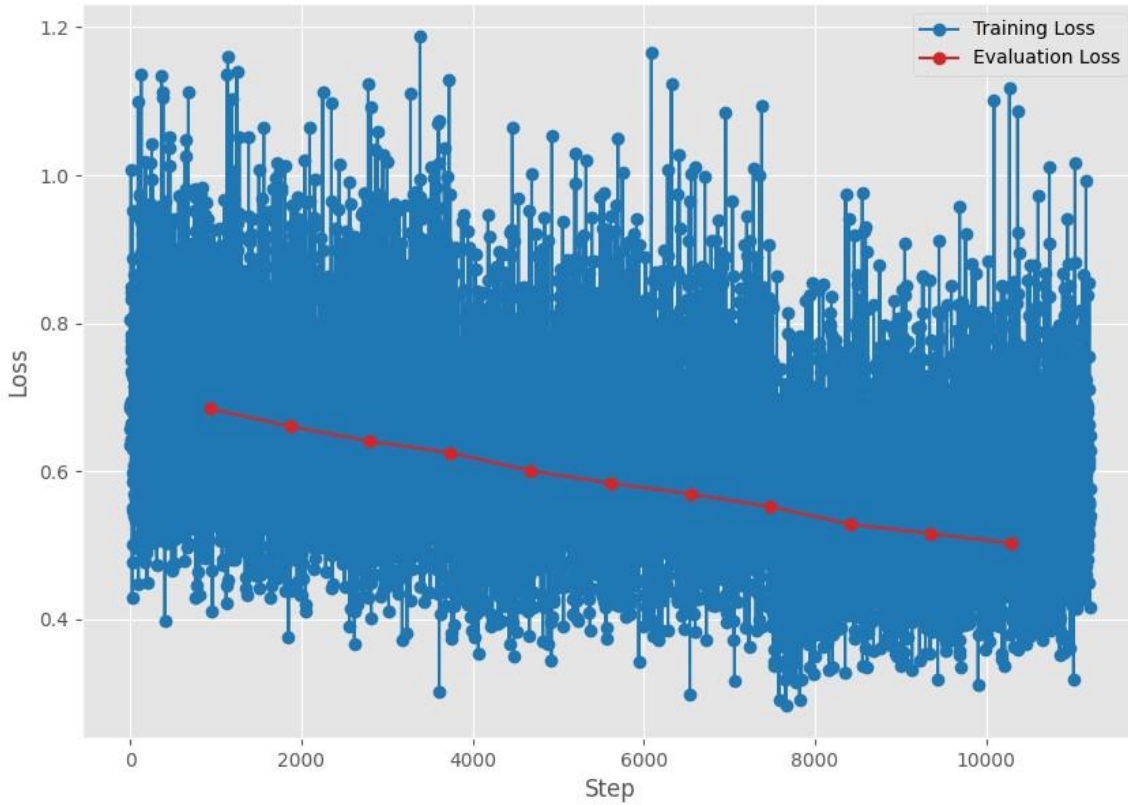
3. Word Cloud of Text Data
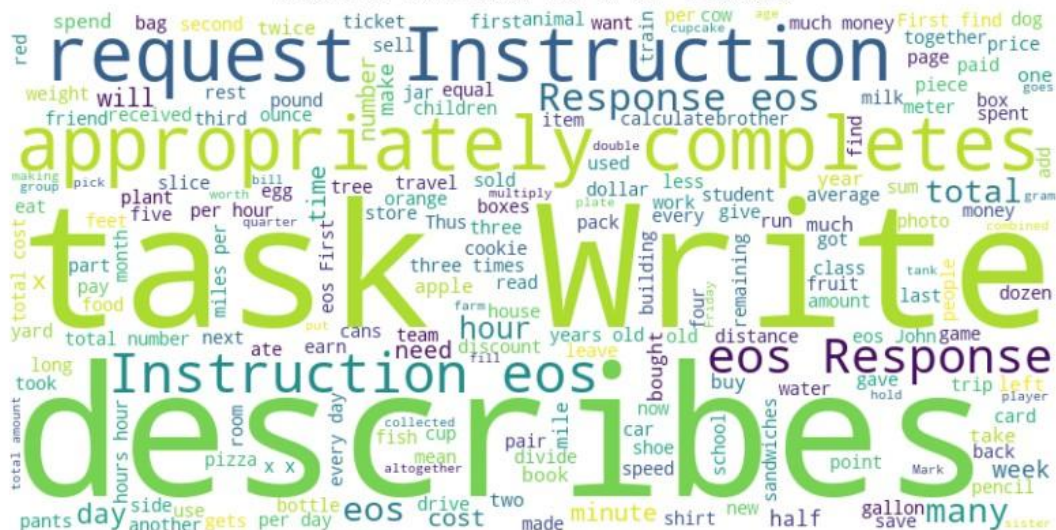
The word cloud visualization provides a quick and intuitive way to understand the most prominent terms and themes present in the dataset. The size of each word in the cloud corresponds to its frequency across the text corpus; larger words indicate higher frequency. Such a visualization allows us to gauge the focus and recurring topics in the dataset at a glance.

In reviewing this word cloud, notice which terms are most dominant. These terms often give insight into the overarching subject matter or discourse within the data. It's also useful for identifying any unexpected or unusual words that may warrant further investigation. In some cases, a word cloud can reveal biases, commonalities, or trends that might not be immediately evident from a cursory reading of the texts.

4. UMAP with KMeans Clustering
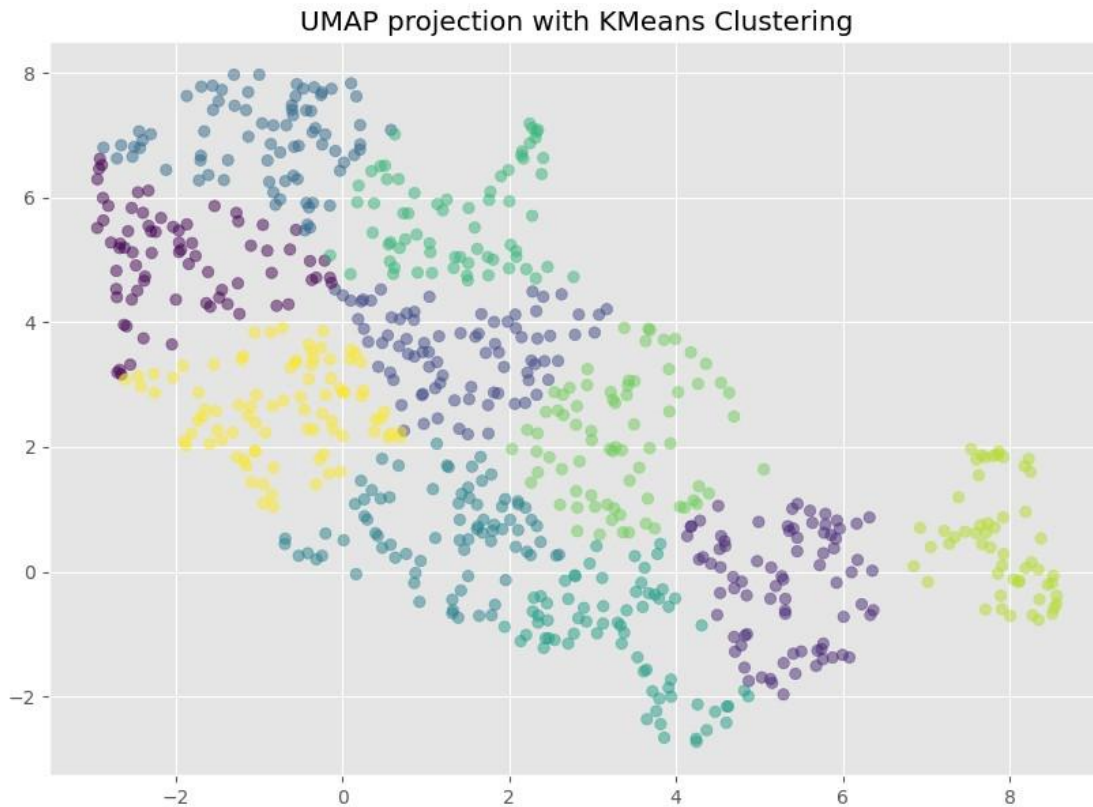
UMAP, combined with KMeans clustering, identifies distinct topics or themes within the data. Each color in the plot represents a different cluster, suggesting groupings of texts with similar content.

6. Anomaly Scores Distribution

This histogram shows the distribution of anomaly scores given by the Isolation Forest method, this method is used to identify the severity of anomalies in the dataset. Anomalies are data points that significantly differ from the norm, potentially indicating errors or unique patterns. if the distribution is skewed to the right, it indicates a higher number of anomalies. if the distribution is uniform, it indicates a lower number of anomalies.

Anomaly Scores Histogram
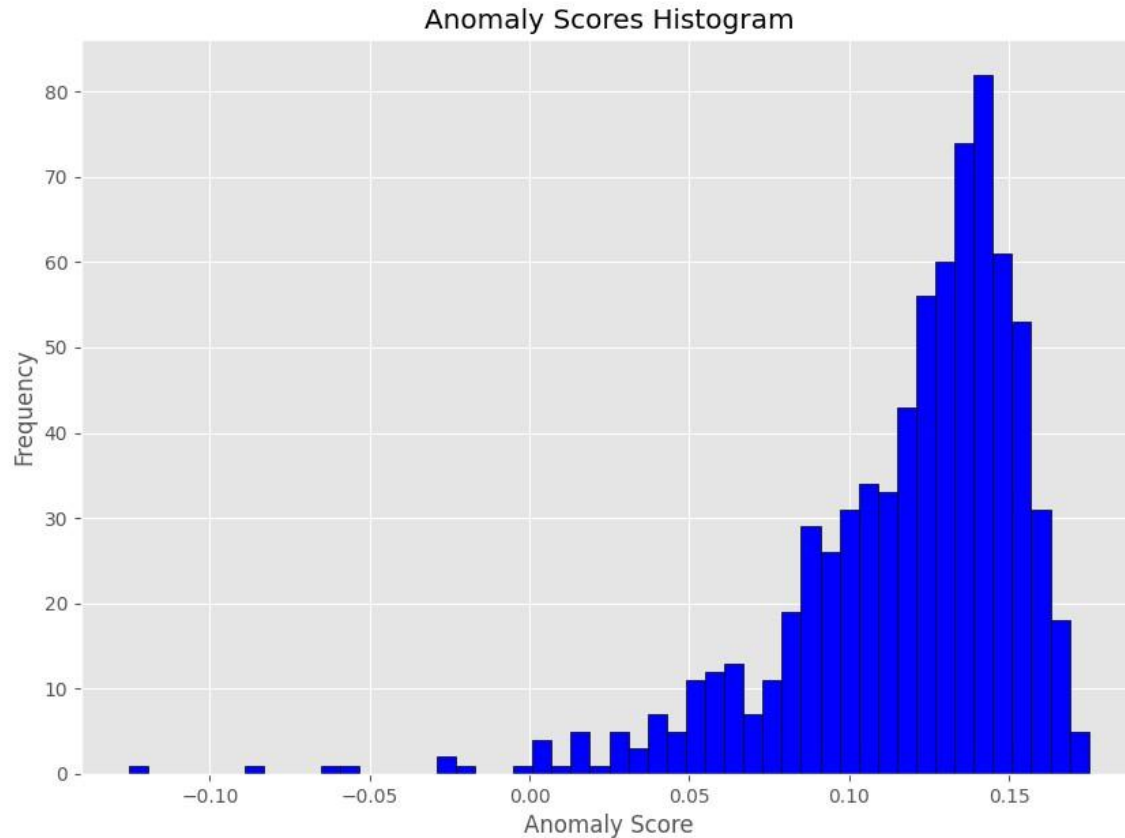
5. Outliers Detection

Outliers detection is a critical step in data analysis, especially when dealing with large text datasets. Outliers can significantly influence the outcomes of statistical analyses and predictive modeling, leading to skewed results. In this context, outliers are unusual or atypical texts that deviate markedly from the majority of the data. Identifying these outliers helps in understanding the data's underlying structure and possibly uncovering rare but insightful patterns or anomalies that could be of interest. For this analysis we used the Isolation Forest algorithm, a SOTA anomaly detection technique, to compute anomaly scores for each text. These were then applied to systematically identify outliers. This method excels in detecting data points that diverge from the norm, highlighting texts that exhibit unique characteristics compared to the typical patterns observed in the dataset. The following texts are identified as outliers:

1. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>Mark has 12 candy bars in total between Mars bars, Snickers, and Butterfingers. He has 3 Snickers and 2 Mars bars. How many Butterfingers does he have?<eos>

###Respon...

1.   Below is an instruction that describes a task. Write a response that appropriately completes the request. ###Instruction:<eos>Rocky boxed 190 fights in his career. 50 percent of his fights were knockouts, and 20 percent of the knockouts were in the first round. How many knockouts did Rocky hav...

2.   Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>You can buy 4 apples or 1 watermelon for the same price. You bought 36 fruits evenly split between oranges, apples and watermelons, and the price of 1 orange is $0.50. Ho...

3.   Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>Hasan is packing up his apartment because hes moving across the country for a new job. He needs to ship several boxes to his new home. The movers have asked that Hasan av...

4.   Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>Harry participates in the auction of a classic painting. The auction starts at $300, Harry is the first to bid, adding $200 to the starting value, a second bidder doubles...

5.   Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>The gauge on a water tank shows that the tank is 1/3 full of water. To fill the tank, 16 gallons of water are added. How many gallons of water does the tank hold when ful...

6.   Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>James has 5 dozen boxes of matches. Each box contains 20 matches. How many matches does James have?<eos>

###Response:<eos>He has 5*12=<<5*12=60>>60 boxes of matches
T...

7.   Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>Ralph has $54.00 worth of products in his cart. At the register, he asks if he could have a 20% discount on an item with a small issue. This item is $20.00 to start. T...