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.

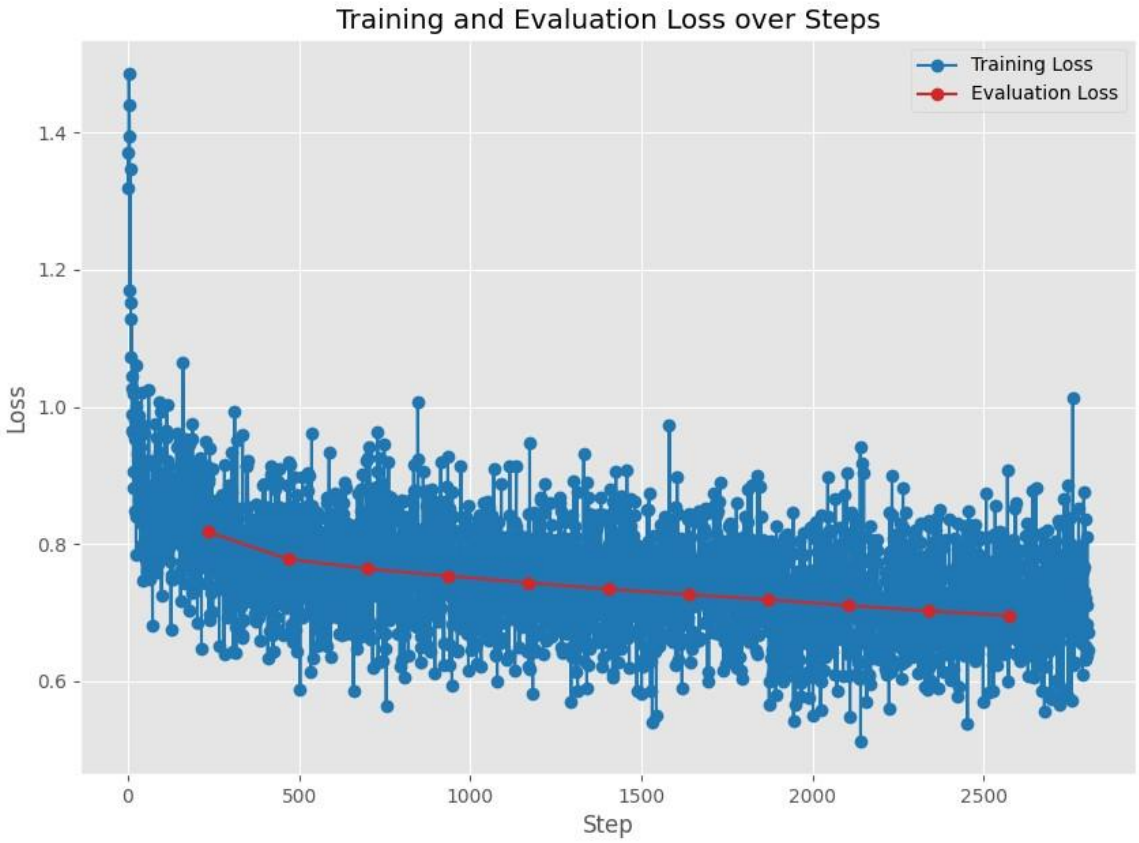
Loss Improvement Metrics

Metric	Value
Train Loss Improvement	52.88%
Validation Loss Improvement	14.90%

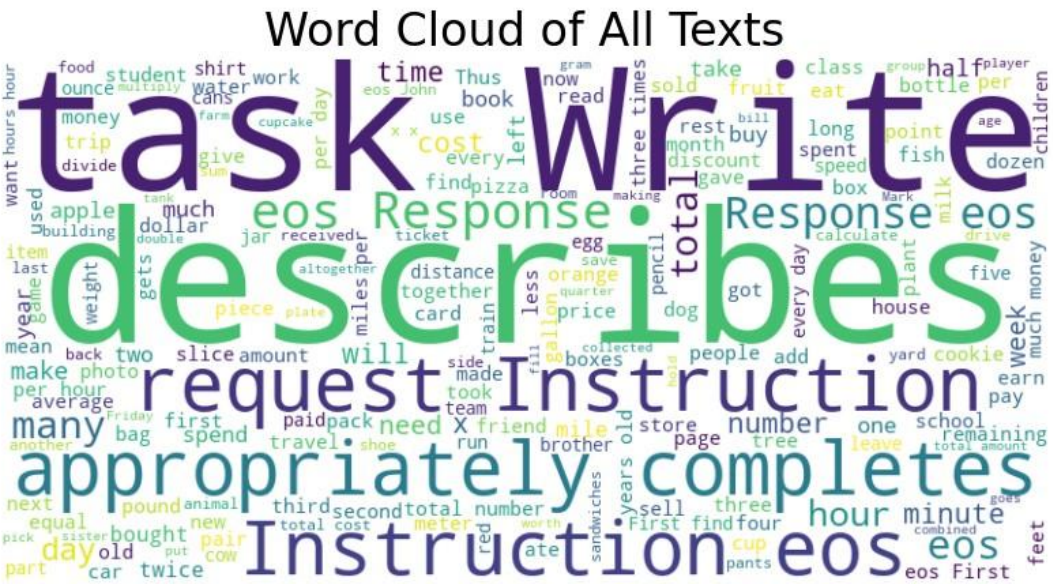
2B. Training Loss Over Steps

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 -9.16%, suggesting that further training could potentially lead to improvements.

Overall a percentage difference of 52.88% 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.



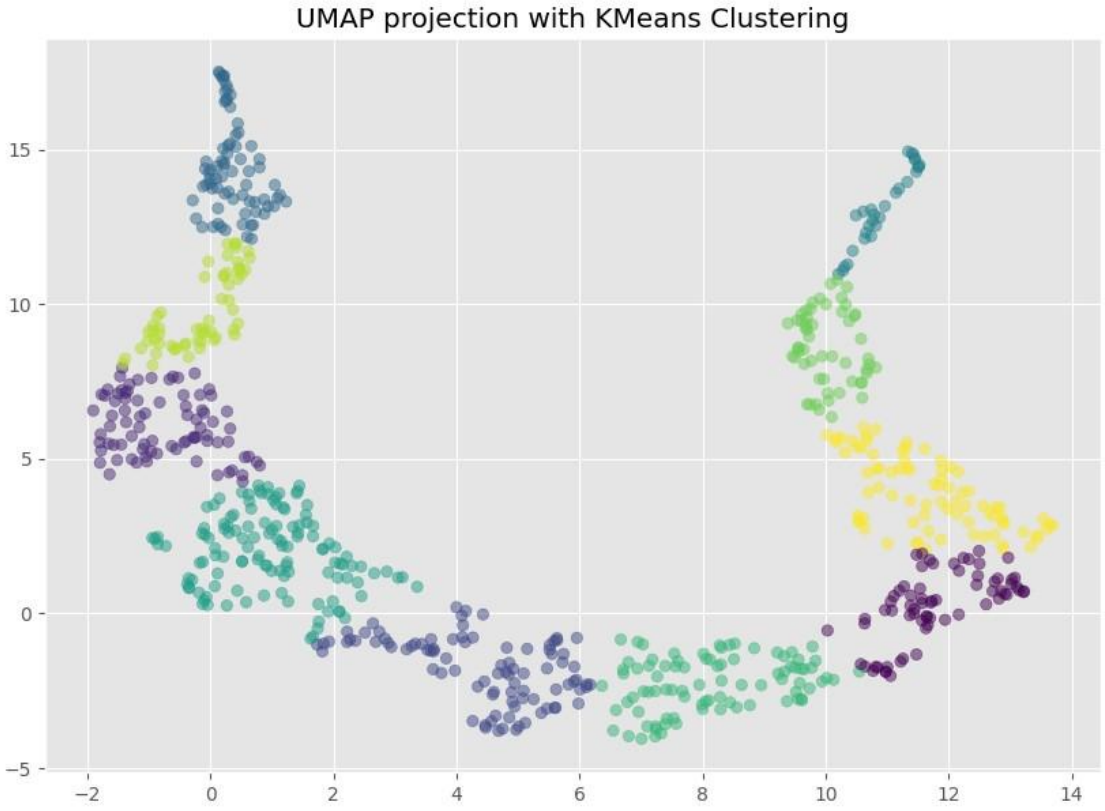
3. Word Cloud of Text Data



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.

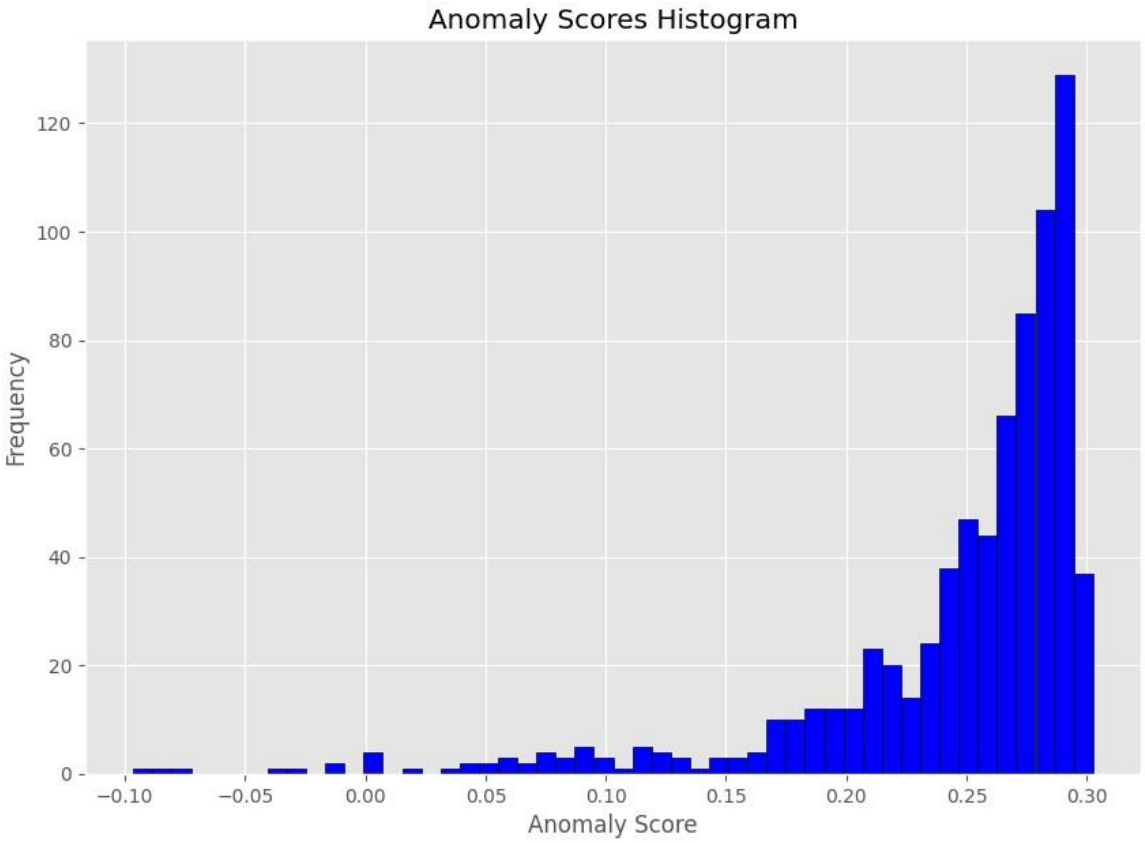
4. UMAP with KMeans Clustering



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.



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>An eagle can fly 15 miles per hour; a falcon can fly 46 miles per hour; a pelican can fly 33 miles per hour; a hummingbird can fly 30 miles per hour. If the eagle, the fa...
2. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>Mona brought 20 cookies to share in class. Jasmine brought 5 fewer cookies than Mona. Rachel brought 10 more cookies than Jasmine. How many cookies altogether did Mona, J...
3. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>A certain tree was 100 meters tall at the end of 2017. It will grow 10% more than its previous height each year. How long has the tree grown from 2017 until the end of 20...
4. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>A movie theater charges $5 for matinee tickets, $7 for evening tickets, and $10 for opening night tickets. A bucket of popcorn costs $10. On Friday, they had 32 matinee c...
5. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>In a school, the number of participants in the 2018 Science Quiz Bowl was 150. There were 20 more than twice the number of participants in 2019 as there were in 2018. In ...
6. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>Debbie works at a post office packing boxes to mail. Each large box takes 4 feet of packing tape to seal, each medium box takes 2 feet of packing tape to seal, and each s...
7. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>William's class set a goal each week of the number of cans of food that is to be collected. On the first day, 20 cans were collected. Then the number of cans increased by...
8. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<eos>Susie has 572 beans in the jar. One-fourth of them are red and one-third of the remaining beans are white. Then half of the remaining are green beans. How many green bean...