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.

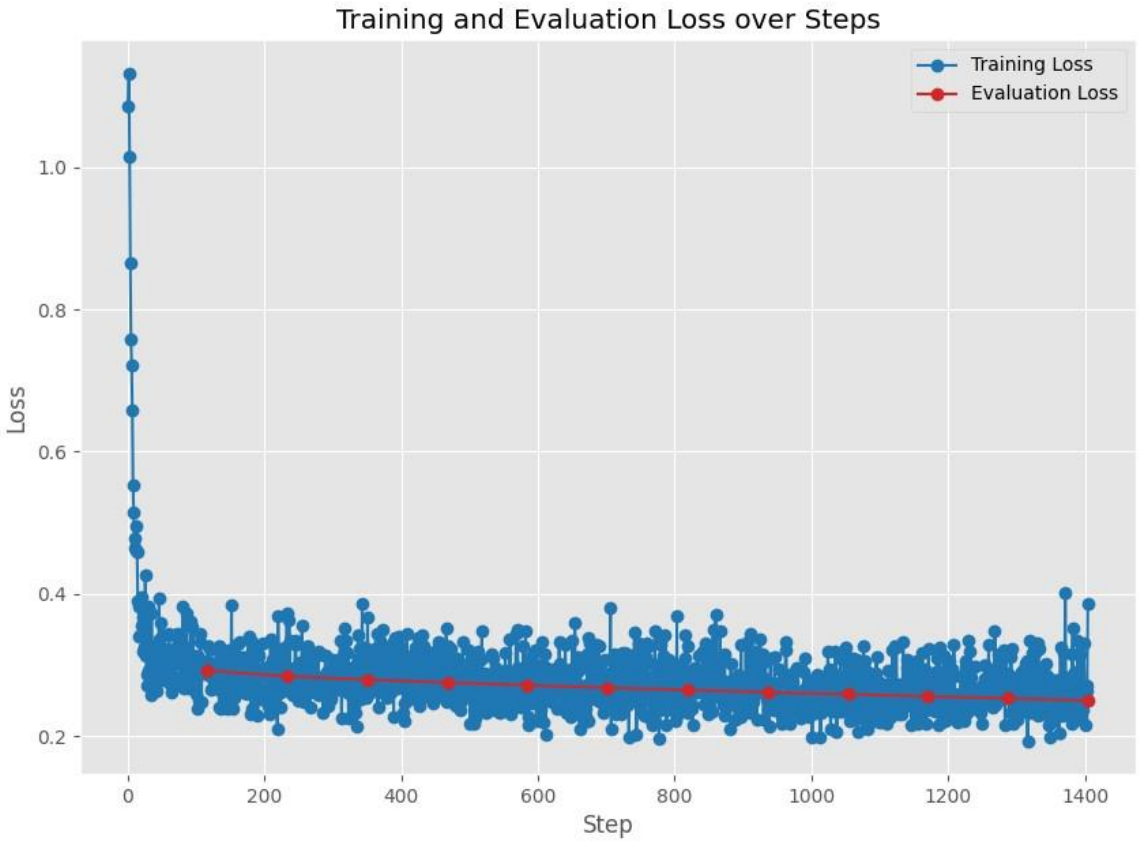
Loss Improvement Metrics

Metric	Value
Train Loss Improvement	64.51%
Validation Loss Improvement	14.43%

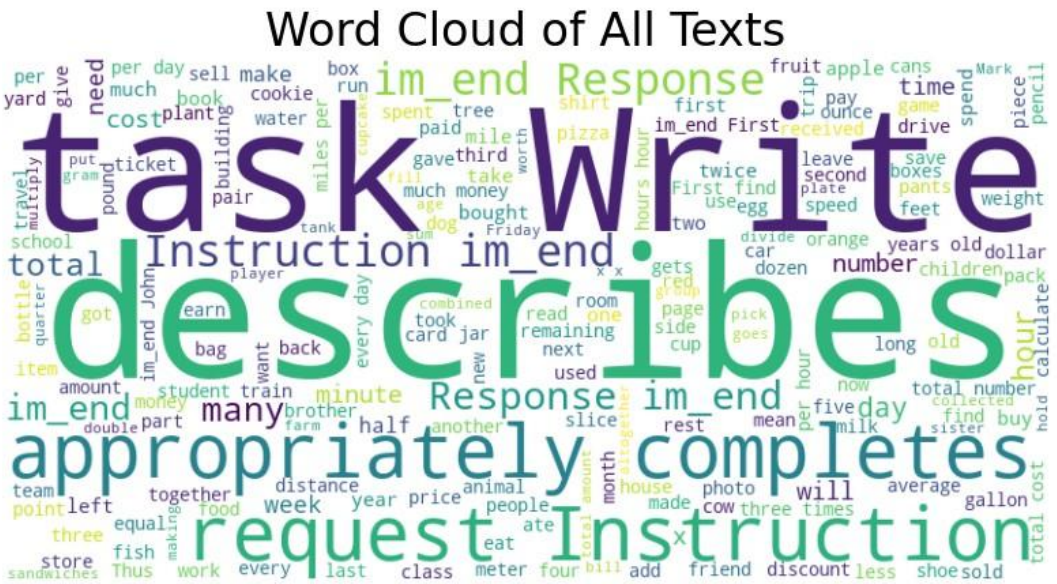
2B. Training Loss Over Steps

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 46.21%, suggesting that further training could potentially lead to improvements.

Overall a percentage difference of 64.51% 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.



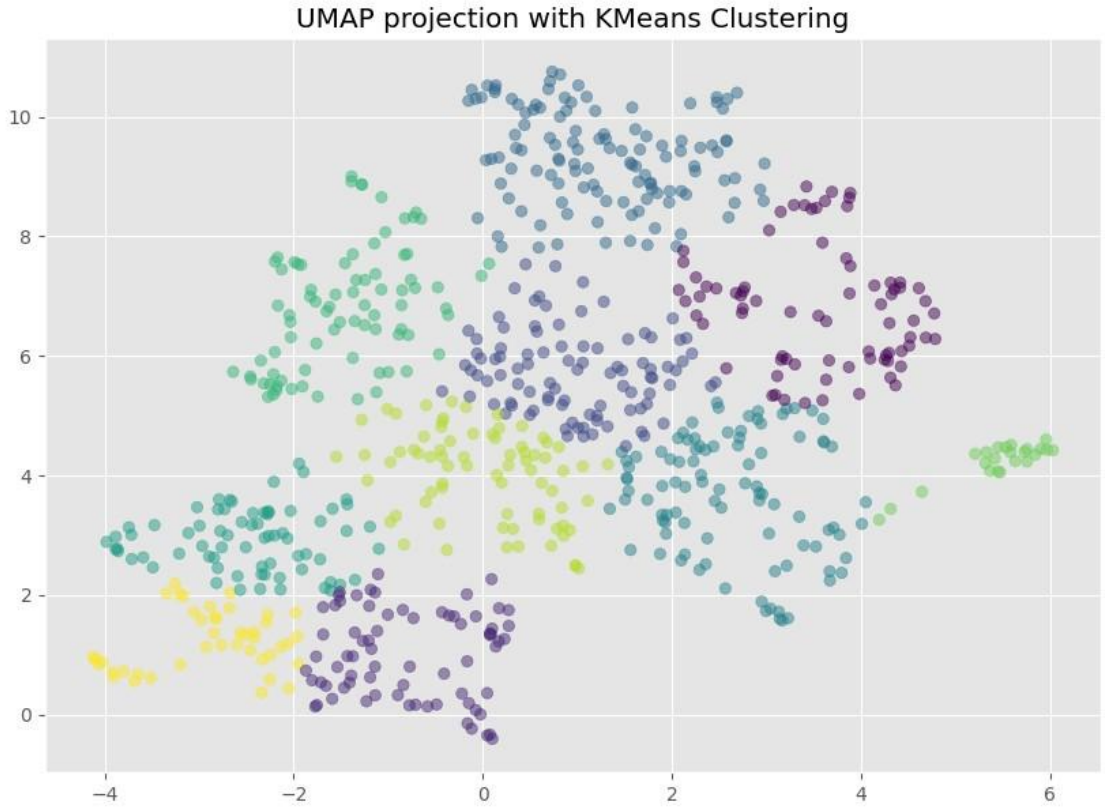
3. Word Cloud of Text Data



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.

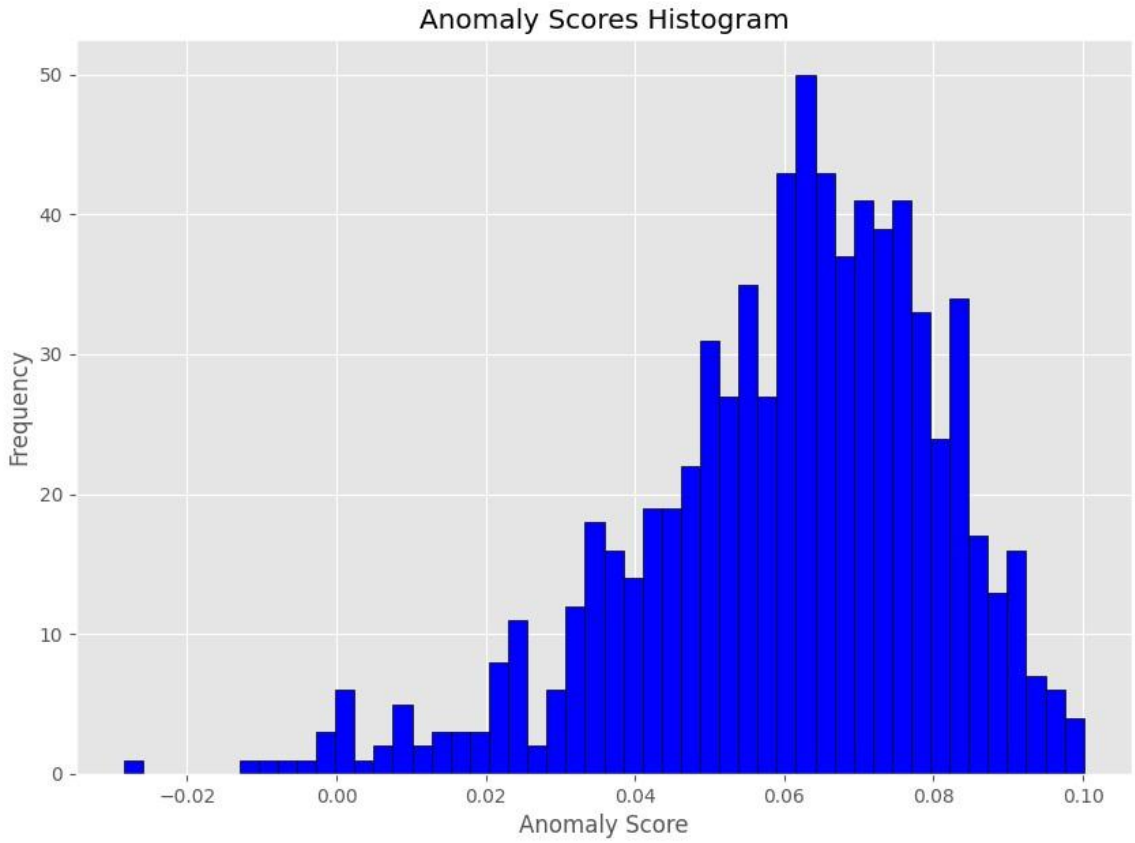
4. UMAP with KMeans Clustering



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:<|im_end|>Ben's potato gun can launch a potato 6 football fields. If a football field is 200 yards long and Ben's dog can run 400 feet/minute, how many minutes will it take hi...
2. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<|im_end|>Marsha works as a delivery driver for Amazon. She has to drive 10 miles to deliver her first package, 28 miles to deliver her second package, and half that long to d...
3. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<|im_end|>Hannah is making banana bread. She needs to use 3 cups of flour for every cup of banana mush. It takes 4 bananas to make one cup of mush. If Hannah uses 20 bananas, ...
4. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<|im_end|>One logger can cut down 6 trees per day. The forest is a rectangle measuring 4 miles by 6 miles, and each square mile has 600 trees. If there are 30 days in each mon...
5. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<|im_end|>A truck drivers heavy semi truck can go 3 miles per gallon of gas. The truck driver needs to put gas in his truck at one gas station, but wants to put the minimum am...
6. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<|im_end|>Three snails raced across a rain-soaked sidewalk. The first snail raced at a speed of 2 feet per minute. The second snail raced at twice the speed of the first sna...
7. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<|im_end|>Karen is paddling her canoe up a river against the current. On a still pond, Karen can paddle 10 miles per hour. The river flows in the opposite direction at 4 miles...
8. Below is an instruction that describes a task. Write a response that appropriately completes the request.

###Instruction:<|im_end|>Lighters cost $1.75 each at the gas station, or $5.00 per pack of twelve on Amazon. How much would Amanda save by buying 24 lighters online instead of at the gas sta...