Exploring Thousands of Inferences on a Single Prompt

Charles Curt
Charles@Sliced-ai.com

July 2024

Abstract

This paper investigates the relationship between text generation hyperparameterstemperature, top_p, sequence length, and token lengthand their influence on the output diversity across thousands of inferences of a single prompt: "You are a time traveler in the year 10,000, describe what you see." The prompt was chosen for its open-ended nature, which allows a wide range of possible responses. Initial analysis reveals weak correlations between these hyperparameters and the generated outputs. Higher temperatures tend to produce incoherent or nonsensical outputs, while top_p and sequence length have more subtle influences. The study also introduces an initial attempt to expand the embedding space to look for more predictive abilities.

A novel Progressive Data Increment Method was employed due to its property of preventing overfitting and improving training stability, as demonstrated in Figure 6. This method allowed models to progressively train on increasing datasets, stabilizing validation loss while predicting temperature and top_p. Despite these efforts, the outputs remain too similar, offering only minimal predictive capability for the hyperparameters. However, there is potential in further expanding the embedding space to uncover hidden patterns.

1 Introduction

Large Language Models (LLMs) such as GPT-3.5-turbo, GPT-4o, and Llama3-7B utilize sampling hyperparameters like temperature, top_p, sequence length, and token length to influence the diversity and randomness of their outputs. Temperature controls the smoothness of the probability distribution from which tokens are sampled, while top_p restricts the token sampling to a cumulative probability mass. Sequence length and token length, which were also varied in this study, control the maximum number of tokens generated in a response and can influence the coherence and completeness of the output.

This study aims to examine whether these hyperparameters lead to observable differences in generated outputs, specifically through the lens of embedding and visualization techniques. By exploring weak correlations and testing advanced embedding and training methods, the objective is to identify patterns that could guide strategies for optimizing LLM performance in complex tasks.

The open-ended prompt used for this study was:

> *"You are a time traveler in the year 10,000, describe what you see."*

This prompt was chosen because of its potential to generate a wide variety of responses, allowing for a rich analysis of how temperature, top_p, sequence length, and token length influence the generation process.

2 Methodology

2.1 Data Generation

Thousands of responses were generated to the aforementioned prompt, systematically varying temperature, top_p, sequence length, and token length. Responses were generated using three different models: GPT-3.5-turbo, GPT-4o, and Llama3-7B. This multi-model approach ensured a more comprehensive analysis by examining the impact of model architecture in addition to hyperparameters.

2.2 Embedding Methods

The responses were embedded using OpenAIs embedding model. Initially, BERT-based embeddings were explored for the task, but they did not provide the resolution required for identifying meaningful separations in the output space. OpenAI embeddings were thus chosen for their superior clustering and differentiation capability, making them more suitable for this analysis.

2.3 Visualization and Correlation Analysis

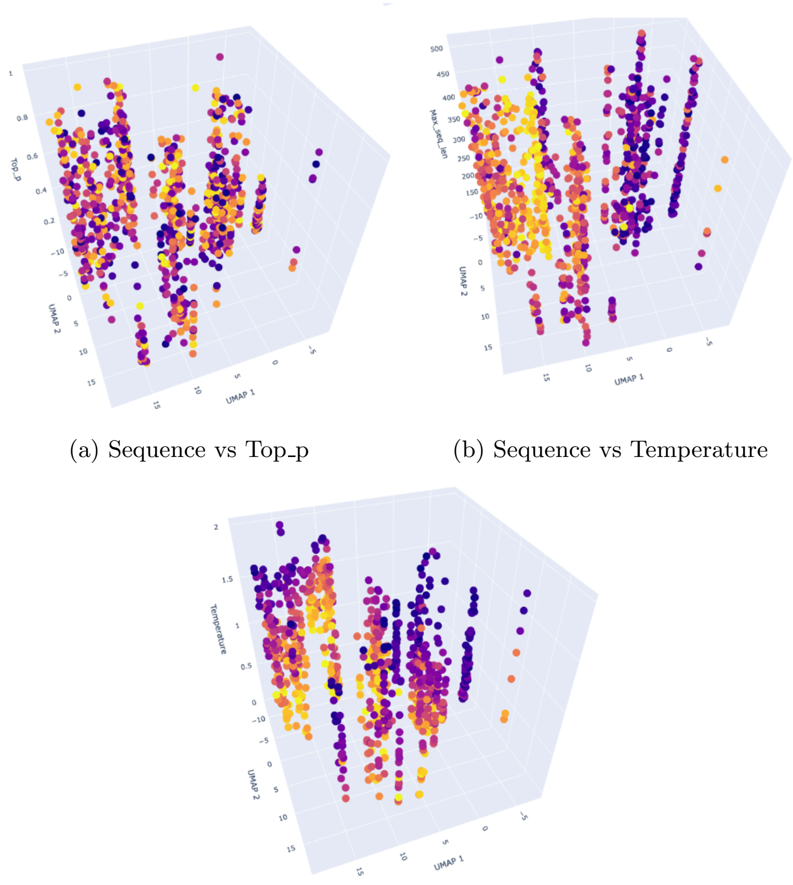
Dimensionality reduction was performed using UMAP (Uniform Manifold Approximation and Projection) to visualize how the responses cluster according to the hyperparameters, including temperature, top_p, sequence length, and token length. Correlation matrices were also generated to quantify relationships between UMAP dimensions and the various hyperparameters.

Finally, an autoencoder was trained to predict temperature, top_p, sequence length, and token length from the embeddings, and to explore the embeddings' potential for clustering and correlation.

3 Results and Analysis

3.1 UMAP Visualizations of Sequences and Hyperparameters

The relationship between the generated sequences and the hyperparameters was first visualized using UMAP. As shown in Figure 1, the expectation was to see distinct clusters corresponding to varying values of temperature, top_p, and sequence length. However, the results indicate weak correlations between the embeddings and these hyperparameters, with no strong separations apparent in the sequence visualizations.



(a) Sequence vs Top_p

(b) Sequence vs Temperature

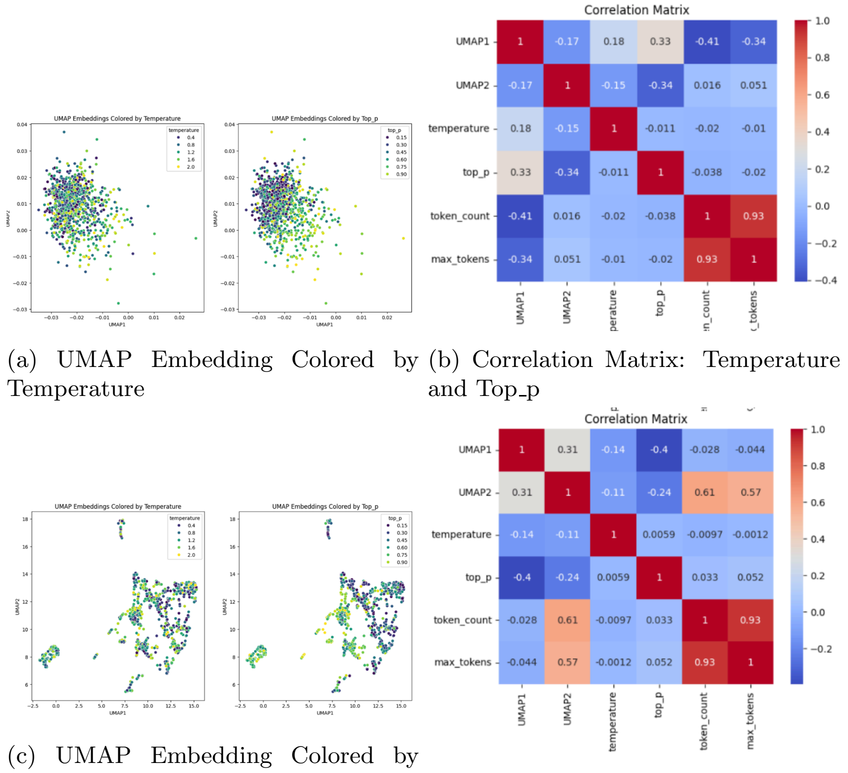
(c) Temperature vs Top_p

Figure 1: UMAP Visualizations of Sequence vs Top_p and Temperature. The visualizations show weak correlations between the generated sequences and hyperparameters.

The lack of clear clusters in these visualizations led to further investigation using correlation matrices and autoencoders to quantify the relationships between the embeddings and the hyperparameters.

3.2 Correlation Matrices and UMAP Embeddings

To further analyze the weak correlations, correlation matrices were computed between UMAP dimensions and the hyperparameters (temperature, top_p, sequence length, token length). As shown in Figure 2, while some relationships between temperature and top_p exist, they are not strong enough to yield distinctive clustering.



	UMAP1	UMAP2	perature	top_p	n_count	_tokens
UMAP1	1	-0.17	0.18	0.33	-0.41	-0.34
UMAP2	-0.17	1	-0.15	-0.34	0.016	0.051
temperature	0.18	-0.15	1	-0.011	-0.02	-0.01
top_p	0.33	-0.34	-0.011	1	-0.038	-0.02
token_count	-0.41	0.016	-0.02	-0.038	1	0.93
max_tokens	-0.34	0.051	-0.01	-0.02	0.93	1

(a) UMAP Embedding Colored by Temperature

(b) Correlation Matrix: Temperature and Top_p

	UMAP1	UMAP2	temperature	top_p	token_count	max_tokens
UMAP1	1	0.31	-0.14	-0.4	-0.028	-0.044
UMAP2	0.31	1	-0.11	-0.24	0.61	0.57
temperature	-0.14	-0.11	1	0.0059	-0.0097	-0.0012
top_p	-0.4	-0.24	0.0059	1	0.033	0.052
token_count	-0.028	0.61	-0.0097	0.033	1	0.93
max_tokens	-0.044	0.57	-0.0012	0.052	0.93	1

(c) UMAP Embedding Colored by Top_p

(d) Correlation Matrix (repeat data)

Figure 2: UMAP Embedding and Correlation Matrix for Temperature and Top_p. The correlations between hyperparameters and embeddings are weak.

The matrices confirm that temperature, top_p, and sequence length exert only weak influences on the structure of the embeddings. The impact of model architecture on clustering was subsequently explored.

3.3 Cluster Analysis by Model

Clustering analysis based on model architecture was performed to determine if the differences in architecture (GPT-3.5-turbo, GPT-4o, and Llama3-7B) created more significant separations in the embeddings. As seen in Figure 3, model architecture played a greater role in clustering than hyperparameters, suggesting that the choice of model has a more pronounced effect on the generated outputs.

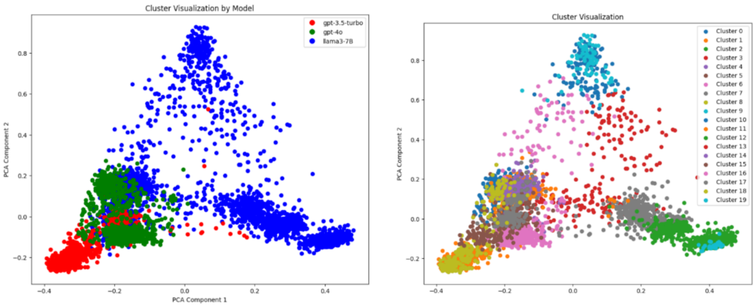


Figure 3: Cluster Visualization by Model: GPT-3.5-turbo, GPT-4o, and Llama3-7B. Model architecture has a greater impact on clustering than temperature, top_p, and sequence length.

The fact that model architecture creates more noticeable separations than temperature, top_p, or sequence length supports the idea that these hyperparameters alone are not sufficient to guide meaningful clustering.

3.4 Autoencoder Predictions of Temperature, Top_p, and Sequence Length

An autoencoder was trained to predict temperature, top_p, and sequence length from the embeddings. As seen in Figure 4, while the autoencoder was able to weakly predict the values, the predictions were often inaccurate, particularly for outliers.

The weak predictions suggest that certain outputs cannot be easily mapped to their corresponding temperature, top_p, or sequence length values, indicating that these hyperparameters may not have a consistent impact on the output structure.

3.5 Autoencoder Results: Reproducing Embeddings

The autoencoder was also trained to reproduce the embeddings themselves in an attempt to see if the model could learn to cluster the data based on the underlying structure. As shown in Figure 5, the model exhibited a weak ability

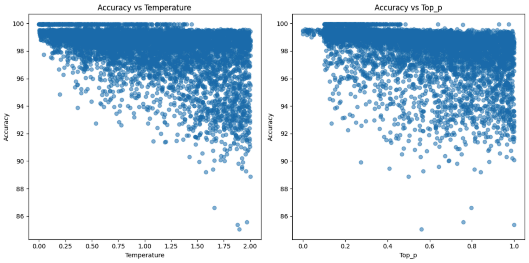


Figure 4: Autoencoder Predictions of Temperature and Top_p, showing weak correlation between the actual values and predictions. Many outliers were unpredictable.

to separate the embeddings into clusters, but the clusters were not distinct enough to be practically useful.

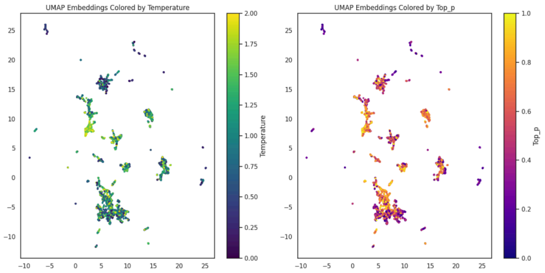


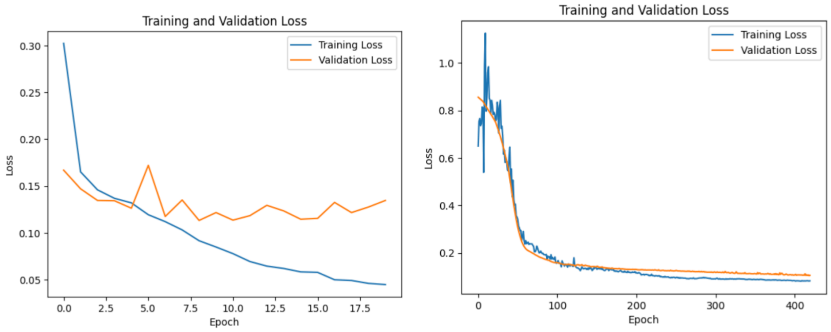
Figure 5: Autoencoder Training to Reproduce Embeddings, showing weak clustering of the data based on some underlying structure.

Although there is potential in using autoencoders to identify latent structures, the results show that the current embedding space does not allow for strong separations based on the hyperparameters.

3.6 Training Loss Comparison: Standard vs Progressive Data Increment

Finally, a comparison was made between standard training methods and the Progressive Data Increment Method. As demonstrated in Figure 6, the Progressive Data Increment Method stabilized the loss curve and reduced overfitting, improving generalization in the model's predictions.

The Progressive Data Increment Method's ability to maintain low training and validation loss suggests it is a promising technique for training large models on complex data without overfitting.



(a) Standard Training Loss

(b) Progressive Training Loss

Figure 6: Training and Validation Loss: Standard vs Progressive Data Increment. The progressive method prevents overfitting and improves generalization.

4 Discussion

The results from this study present several key insights. First, temperature, top_p, sequence length, and token length produce only weak correlations in the embedding space, as confirmed by UMAP visualizations, correlation matrices, and the inability of the autoencoder to strongly predict these values. This contradicts the theoretical assumption that varying these hyperparameters would lead to clear structural differences in the outputs.

Second, model architecture has a much more pronounced effect on clustering than temperature, top_p, or sequence length. The outputs generated by different models, such as GPT-4o and Llama3-7B, showed clearer separations, implying that the models underlying structure is more influential in determining the variability of output responses.

Finally, the Progressive Data Increment Method offered a significant improvement in generalization, preventing overfitting even with larger data sets. This method shows potential for training models on progressively increasing data sizes.

5 Conclusion and Future Work

This study sought to identify correlations between temperature, top_p, sequence length, token length, and text output in large language models. While weak correlations were found, particularly with higher temperatures producing incoherent outputs, the overall effect of these hyperparameters on the embedding space was minimal. The models' architecture had a far greater impact on output variability.

Autoencoders showed some potential in weakly predicting temperature, top_p, and sequence length values and in clustering data based on embeddings, but further improvements are needed to uncover stronger correlations. Future work will focus on expanding the embedding space and exploring more advanced models to identify latent structures in generated text.