# Enhancing Spatial Transcriptomic Inference with LLM Loss Functions

*Sidh Jaddu*

*Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center*

Dartmouth Health

## ABSTRACT

- Skin cancer is a highly potent illness, and diagnosis is sometimes ineffective
- Deep learning-based Spatial Transcriptomic (ST) methods are more cost-effective; however, in many cases, they lack a contextual understanding, so LLM loss functions can be used to solve this problem
- This study improves upon existing ST inference models by enhancing their LLM loss functions

## INTRODUCTION

Skin cancer is the most common cancer in the United States, with one out of every five Americans developing skin cancer during their lifetimes.

**Current Research**

- Existing diagnostic methods often miss early-stage tumors or fail to map tumor behavior at the molecular level accurately.
- ST methods greatly help solve these issues, but they are costly and, thus, are difficult to upscale.
- Deep learning-based ST methods are more cost-effective; however, in many cases, they lack contextual understanding or have outputs that are not biologically plausible.
- **ST Inference Models w/ LLM Loss Function:**
  - To reduce the costs of producing ST data, inference models were proved to be used to infer ST data from Visium Slides.
  - In addition, LLM loss functions were also shown to enhance the contextual understanding of the models.

**Goal: Improve upon existing ST inference models by enhancing their LLM loss functions.**

## METHODS

**Data and Materials:**

- STs from 16 Visium Slides of skin tissue collected from various patients from the DHMC
- Additional STs of skin tissue from 5 patients from the Cell Atlas
- Lambda Labs: 8x Tesla V100 GPUs

**Data Preprocessing:**

- The data, which was in the AnnData format, was filtered based on a list of predetermined 1000 spatially variable genes
- Then, the data was normalized and the genes were mapped to token IDs

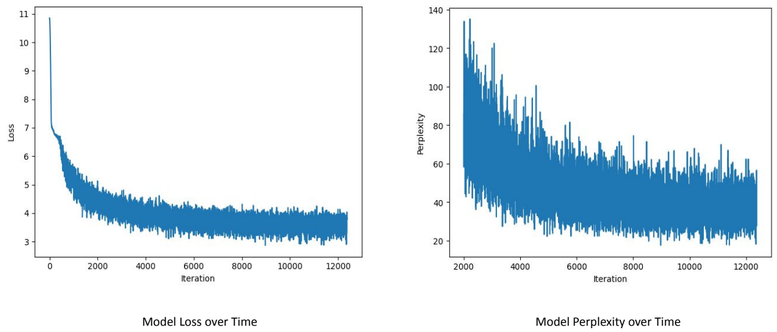
**Improving the LLM Loss Function:**

- This study sought to improve upon the current LLM loss function by using more data and exploring other LLM models
- Existing LLM loss functions are based on only GPT2 and trained on the only the STs from the 16 Visium Slides
- This study further fine-tunes the existing LLM model by using STs from additional patients and hyperparameter tuning
- The data augmentation increased the data from the initial 40,000 cells to over 60,000 cells in the combined data
- It also explores using other LLM models, including ALBERT and XLNet

## RESULTS

Initially, the model displayed the following performance metrics with GPT2:

- Mean loss - 4.333, Perplexity - 76.192

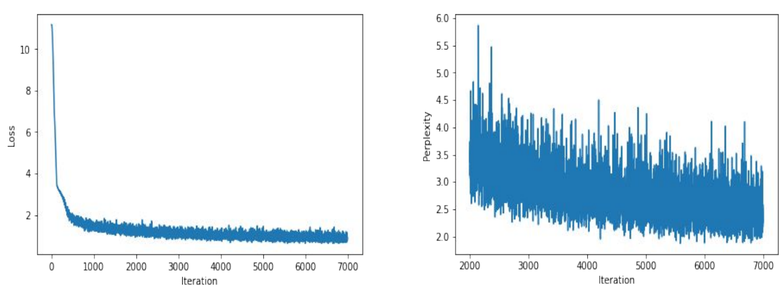



Model Loss over Time  Model Perplexity over Time

Figure 1. GPT2 Loss and Perplexity During Training with the Original Dataset

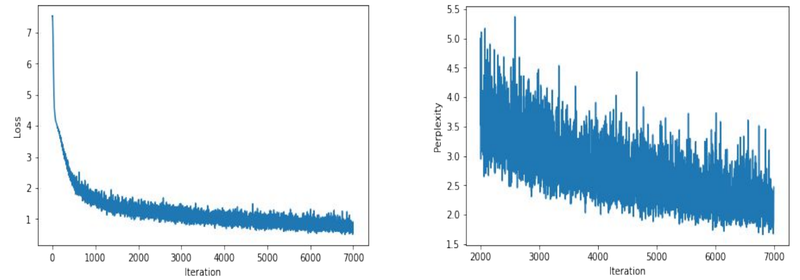
With the augmented data, the models had the following performance metrics:

- GPT2: Mean Loss - 1.298, Perplexity - 3.663
- ALBERT: Mean Loss - 1.286, Perplexity - 3.619
- XLNet: Mean Loss - 0.189, Perplexity - 1.209



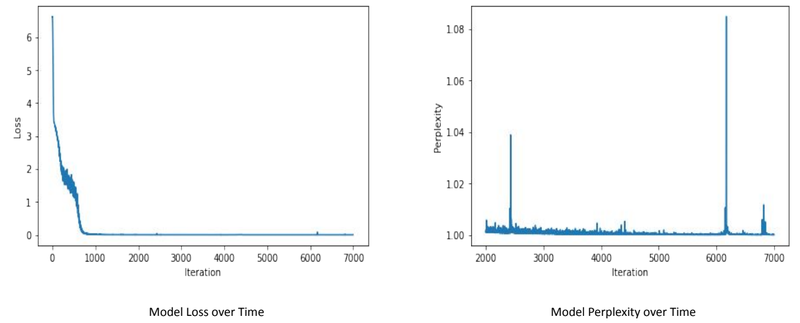

Model Loss over Time  Model Perplexity over Time

Figure 2. GPT2 Loss and Perplexity During Training with the Augmented Dataset




Model Loss over Time  Model Perplexity over Time

Figure 3. ALBERT Loss and Perplexity During Training with the Augmented Dataset




Model Loss over Time  Model Perplexity over Time

Figure 4. XLNet Loss and Perplexity During Training with the Augmented Dataset

## CONCLUSION

**Impact of Additional ST Data and Other Models:**

- The increase of training data did significantly improve LLM model performance regardless of model type used
- XLNET outperformed GPT2 and ALBERT
- These improvements will improve ST Inference

**Limitations:**

- Was not able to explore more advanced models like GPT4
- Improved upon loss function, but the improved loss function was not validated in the original ST inference model

**Future Work:**

- Investigate other types of cells, not just skin
- Explore fine-tuning more advanced LLM models

**Acknowledgements**: Dr. Joshua Levy and Gokul Srinivasan

edit.