Model Performance: Comparing ReLU with Letter Tokenization and GeLU with Word Tokenization

Performance Comparison

Using ReLU and Letter Tokenization

Training and Validation Loss:

- Continued improvement until step 4500.
- step 4500: train loss 2.0460, val loss 2.1017

Observations:

The model showed a steady decrease in both training and validation loss, indicating effective learning and generalization over time.

Using GeLU and Word Tokenization

Training and Validation Loss:

- Continued improvement until step 4500.
- step 4500: train loss 3.4707, val loss 3.9442

Observations:

The change to word tokenization and GeLU activation likely impacted the model's ability to capture more meaningful patterns in the data, typically resulting in improved performance metrics.

The Final training loss and the validation loss however is greater in GeLU

Qualitative Analysis

Word tokenization has enhanced the model's ability to generate coherent and contextually appropriate text compared to letter-level tokenization.

GeLU activation is known for its smooth, non-linear properties, which often lead to better convergence and performance in deep learning models, particularly in GPT-like architectures.

Training Efficiency

Training Time: Model using GeLU activation functions require more computation but yield better performance.