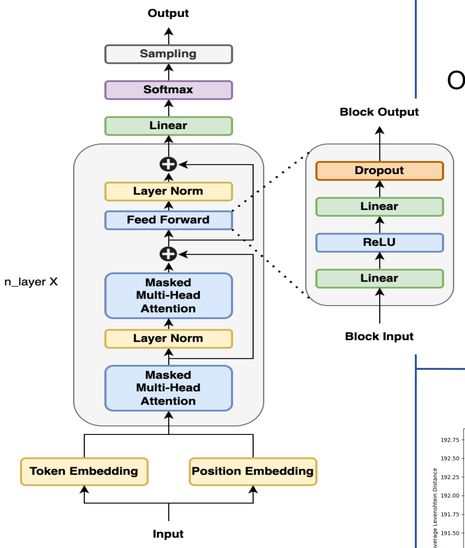
AI First Final Project

Jihwan Oh
joh413@gatech.edu

Github LinkedIn

Model Analysis (Part1)

Model Arch.



Tested Hyper Parameter

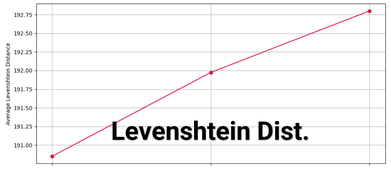
Tokenizer (Naive, BPE, GPT2)
Optimizer (SGD, Adam, AdamW)
(Momentum, Weight Decay)

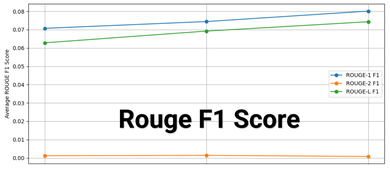
Learnig Rate & Dropout

# of Head, Layer, Embed

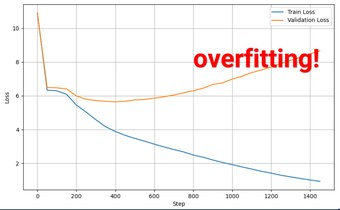
Data Block Size
Train data proportion

Metrics





Train & Validation Loss Graph

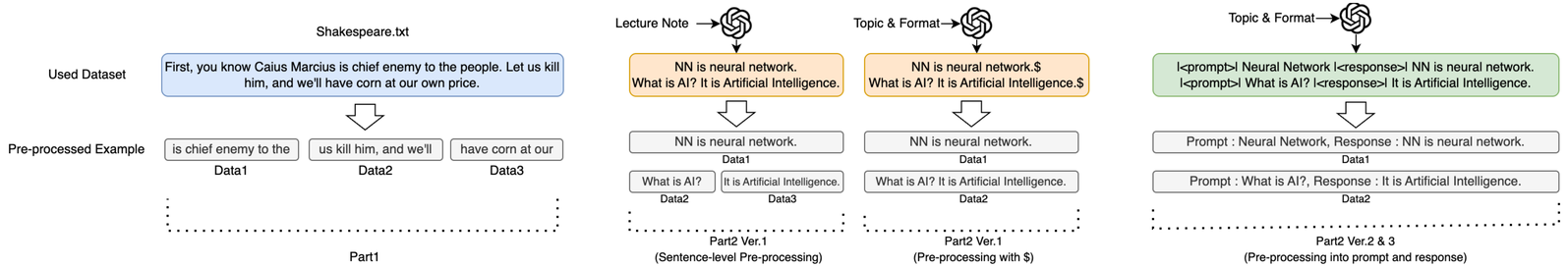


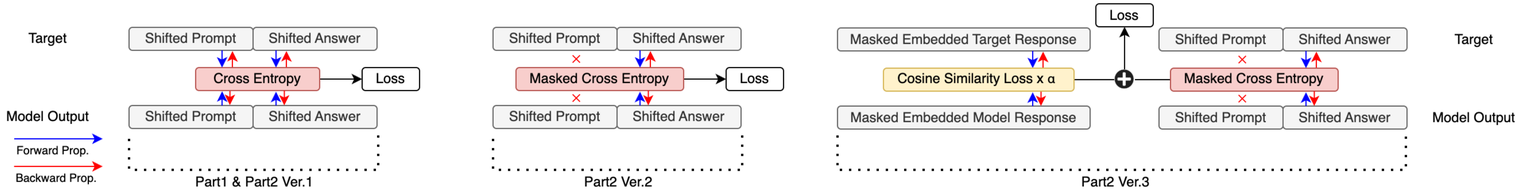
Enhancing Base Model (Pre-Part2)

1. Best Model Checkpointing & Early Stopping
2. Construct Dataset (AI, DL, ML)
3. Scale Model Size Appropriately
4. Randomly Shuffle the Dataset
5. Set Random Seed to Ensure Reproducibility

Question-Answering LLM Overview (Part2)

Goal: Develop a Question-Answering LLM with minimal reliance on open-source





Model Improvements (Ver.1)

1. Make QnA dataset (Total 5GB)
2. Sentence-Level Pre-Processing
   (Preserve coherent sentence structure)
3. Add Special Token, $
   (Combine question and answer into a single unit)

- Model response is not contextually relevant to the question

Model Improvements (Ver.2)

1. Make dataset with **|<prompt>|, |<response>|** tokens (Total 10GB, Motivated from Instruction Tuning)
2. Calculate Cross Entropy only on answer part
   (Focus model on generating good answers)

- Relevance slightly improved

- Model fails with minor prompt changes (e.g., synonym usage)

Model Improvements (Ver.3)

1. Add Cosine similarity loss
   (Learn Sentence level meaning)
   (More robust to minor variations)
2. Add BERTScore

- Relevance moderately improved

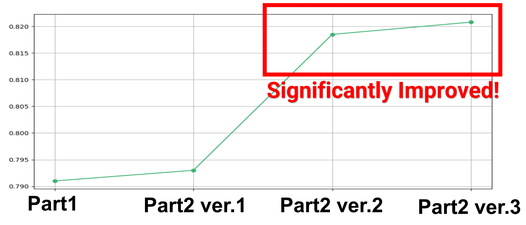
Add new Metric

- **Levenshtein distance** and **Rouge Score** focus on surface similarity. It can't evaluate the semantic alignment

- To evaluate **semantic fidelity**, I used **BERTScore**, which measures meaning alignment via contextual embeddings

Results

BERTScore



ROUGE Score