**Week9 Report**

**Question (i) a**

input\_childSpeech\_trainingSet.txt

**Contents:** The dataset contains short, declarative sentences, often starting with "I" or simple commands like "Go Park" and "Look moon”. The phrases are mostly two- to five-word sentences without complex grammar. Common phrases like "More more," "I want cookie," and "All gone" appear multiple times, which suggests frequent repetition within the text, which is a characteristic often seen in datasets aimed at language acquisition or early reading.

**Vocabulary Size:** The vocabulary size should be around 50 to 100. Words are simple nouns, verbs, and descriptors, indicating a vocabulary focused on familiar objects, actions, and desires, which are typical in child-directed language.

**Dataset Length:** The dataset is 10,000 rows long and contains approximately 55,000 words

input\_childSpeech\_testSet.txt

**Contents:** Similar to the training set, belongs to child-directed language

**Vocabulary Size:** Similar to the training set, the vocabulary should be around 50 to 100

**Dataset Length:** The dataset is 1,000 rows long and contains approximately 5,000 words. Which means the test set contents is 1/10 of the training set

input\_shakespeare.txt

**Contents:** The dataset appears to contain dialogue structure from a play or dramatic script, with characters speaking in Old or Early Modern English. It features a conversational structure with lines attributed to various characters, including "First Citizen," "Second Citizen," "All," "MENENIUS," and so on. The content seems to focus on themes of conflict, social issues, and power dynamics, characteristic of classical literature by Shakespeare.

**Vocabulary Size:** Given the length and the complexity of Shakespearean language, the dataset likely contains several thousand unique words, possibly between 2,000 and 4,000.

**Dataset Length:** The dataset is 40,000 rows long and containsapproximately200,000 words in total

**Question (i) b**

1. Embedding Dimension (n\_embd): Currently set to 384, this parameter affects the dimensionality of each token's vector representation. Reducing it has a significant impact on parameter count while retaining most core functionality. Lowering it to 160 would be a good compromise, as it will still capture sufficient feature richness but with fewer parameters.

2. Number of Layers (n\_layer): Currently set to 6, dictating the depth of the transformer. Reducing it to 3 layers will cut down on parameters by about half, making the model shallower and faster without fully sacrificing depth.

3. Number of Attention Heads (n\_head): Currently set to 6, each head requires its own set of attention weights. Reducing to 4 heads minimizes the parameter count without drastically limiting the attention mechanism's expressiveness, which is sufficient for a model aimed at sub-1M parameters.

4. Block Size (block\_size): Currently set to 256, this determines the maximum context length for each sequence. Reducing it to 96 will reduce memory usage and speed up training, which is particularly useful if the input\_childSpeech\_trainingSet.txt dataset contains shorter sequences.

Batch Size (batch\_size): Although not directly influencing model parameters, reducing it from 64 to 32 will speed up training and lower memory consumption. This is practical for limited compute environments.

Dropout (dropout): While it doesn’t impact parameter count, reducing it slightly (from 0.2 to 0.1) can improve performance on smaller datasets and models by retaining more information flow through each layer.

Model parameters: 0.95492 M parameters