SoME Ablation runs v3

V1:

- Model Dimensions:
  - D_MODEL=512,
  - NUM_HEADS=8,
  - NUM_LAYERS=10
- SoME Parameters:
  - num_experts=128,
  - d_ffn=1536,
  - top_k=8
- Training Schedule:
  - 4 Epochs,
  - LR=8e-4
- Data Configuration:
  - SEQ_LEN=768,
  - BATCH_SIZE=32,
  - VOCAB_SIZE=8192,
  - train_subset_size=20000
  - Dataset = roneneldan/TinyStories
- Model & Training Scale:
  - Total Parameters: 2037.44M (2.04 Billion)
  - Trainable Parameters: 21.55M (1.06% of total)
  - Total Training Steps: 2500
- Expert Initialization:
  - init_method="sparse”
- Results:
  - The Gini coefficient started high at 0.763 and showed a clear, healthy decrease to 0.721.
  - The Shannon Entropy started low at 4.814 and showed a clear, healthy increase to 4.894.
  - Training Loss: Decreased steadily from 1.1118 at the end of Epoch 1 to 0.6255 at the end of Epoch 4.
  - Validation Loss: Decreased steadily from 0.7924 at the end of Epoch 1 to 0.6475 at the end of Epoch 4.
  - Validation Perplexity: Showed significant improvement, reducing from 2.21 to 1.91.

V2:

- Model Dimensions:
  - D_MODEL=512,
  - NUM_HEADS=8,
  - NUM_LAYERS=10
- SoME Parameters:

  - num_experts=512,
  - d_ffn=512,
  - top_k=8
- Training Schedule:
  - 4 Epochs,
  - LR=8e-4
- Data Configuration:
  - SEQ_LEN=768,
  - BATCH_SIZE=32,
  - VOCAB_SIZE=8192,
  - train_subset_size=20000
  - Dataset = roneneldan/TinyStories
- Model & Training Scale:
  - Total Parameters: 2711.15M
  - Trainable Parameters: 21.55M (1.06% of total)
  - Total Training Steps: 2500
- Expert Initialization:
  - init_method="sparse”
- Results:
  - The Gini coefficient started at a literally unbelievable 0.980 and stayed there. A Gini of 1.0 is perfect inequality. This is as close to a "winner-take-all" system as is physically possible.
  - The Shannon Entropy started at a record-low 3.604 and, after a brief, desperate attempt to rise, collapsed back down.
  - Training Loss: a slight increase from 1.7921 at the end of Epoch 1 to 1.8434 at the end of Epoch 4.
  - Validation Loss: a slight increase from 1.5772 at the end of Epoch 1 to 1.6268 at the end of Epoch 4.
  - Validation Perplexity: Showed significant reduction in performance, reducing from 4.84 to 5.09.

V3:

- Model Dimensions:
  - D_MODEL=512,
  - NUM_HEADS=8,
  - NUM_LAYERS=10
- SoME Parameters:
  - num_experts=512,
  - d_ffn=512,
  - top_k=16
- Training Schedule:
  - 4 Epochs,
  - LR=8e-4
- Data Configuration:

    - SEQ_LEN=768,
    - BATCH_SIZE=32,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 2711.15M
    - Trainable Parameters: 21.55M (1.06% of total)
    - Total Training Steps: 2500
- Expert Initialization:
    - init_method="sparse”
- Results:
    -