Professional (Series C) Ablation Studies Base Results

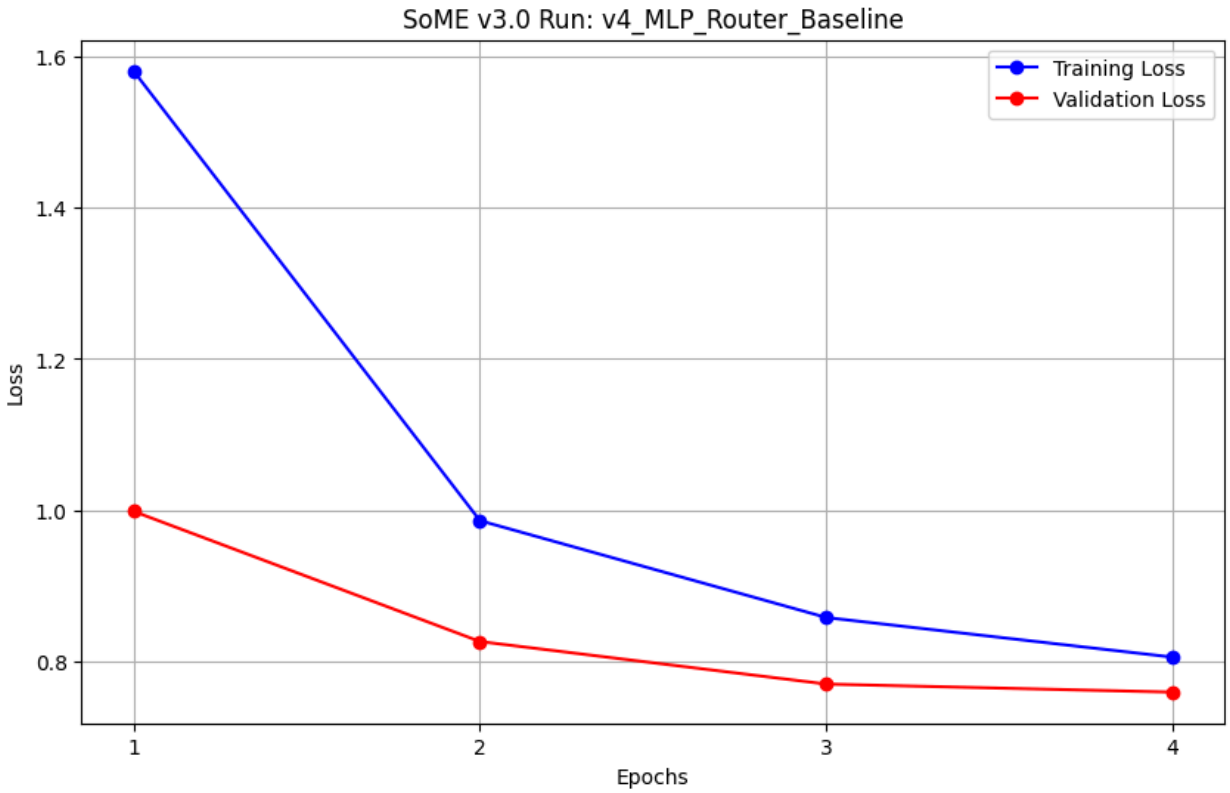
B1: (Prior run to act as a baseline)

- D_MODEL: 512
- NUM_LAYERS: 10
- NUM_HEADS: 8
- SEQ_LEN: 768
- BATCH_SIZE = 32
- VOCAB_SIZE = 8192
- SoME Parameters:
    - NUM_EXPERTS: 128
    - D_FFN: 1024
    - top_k: 4
    - alpha (Attraction): 0.01
    - beta (Peer-Pull): 0.005
    - delta: 0.001
    - theta_percentile: 0.05
    - ema_decay (Inertia): 0.99
- Training Parameters:
    - train_subset_size = 10000
    - val_subset_size = 2000
    - LEARNING_RATE = 6e-4
    - TRAINING_TEMP = 0.8
    - EPOCHS = 4
- Resulting Architecture:
    - Total parameters: 1373.57M
    - Trainable parameters: 29.42M (2.14%)
    - Total training steps: 1248
    - Using expert initialization method: default
    - Router: We'll use the MLP Router (d_model -> 2*d_model -> d_model)

Results:

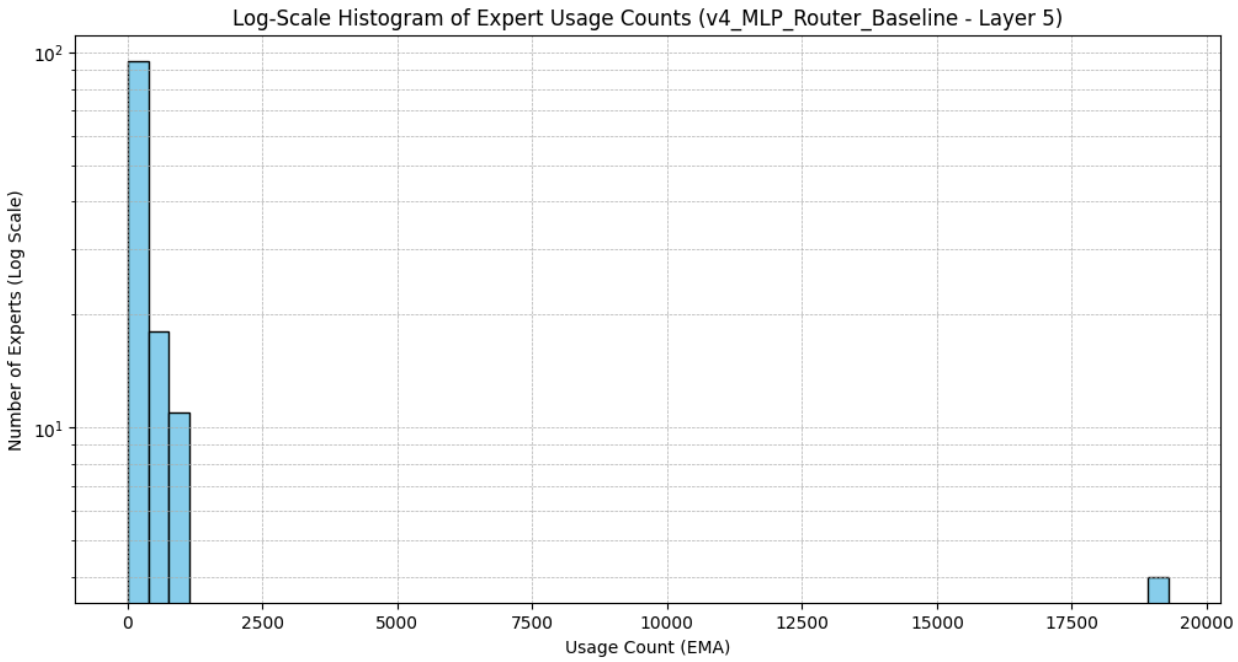
- Epoch 1:
    - Train Loss = 1.5803,
    - Val Loss = 0.9990,
    - Val Perplexity = 2.72
    - Middle Layer Expert Metrics:
        - Gini = 0.941,
        - Entropy = 3.272
- Epoch 2:
    - Train Loss = 0.9866,
    - Val Loss = 0.8270,
    - Val Perplexity = 2.29
    - Middle Layer Expert Metrics:

        - Gini = 0.923,
        - Entropy = 3.413
- Epoch 3:
    - Train Loss = 0.8587,
    - Val Loss = 0.7709,
    - Val Perplexity = 2.16
    - Middle Layer Expert Metrics:
        - Gini = 0.912,
        - Entropy = 3.635
- Epoch 4:
    - Train Loss = 0.8064,
    - Val Loss = 0.7602,
    - Val Perplexity = 2.14
    - Middle Layer Expert Metrics:
        - Gini = 0.914,
        - Entropy = 3.427

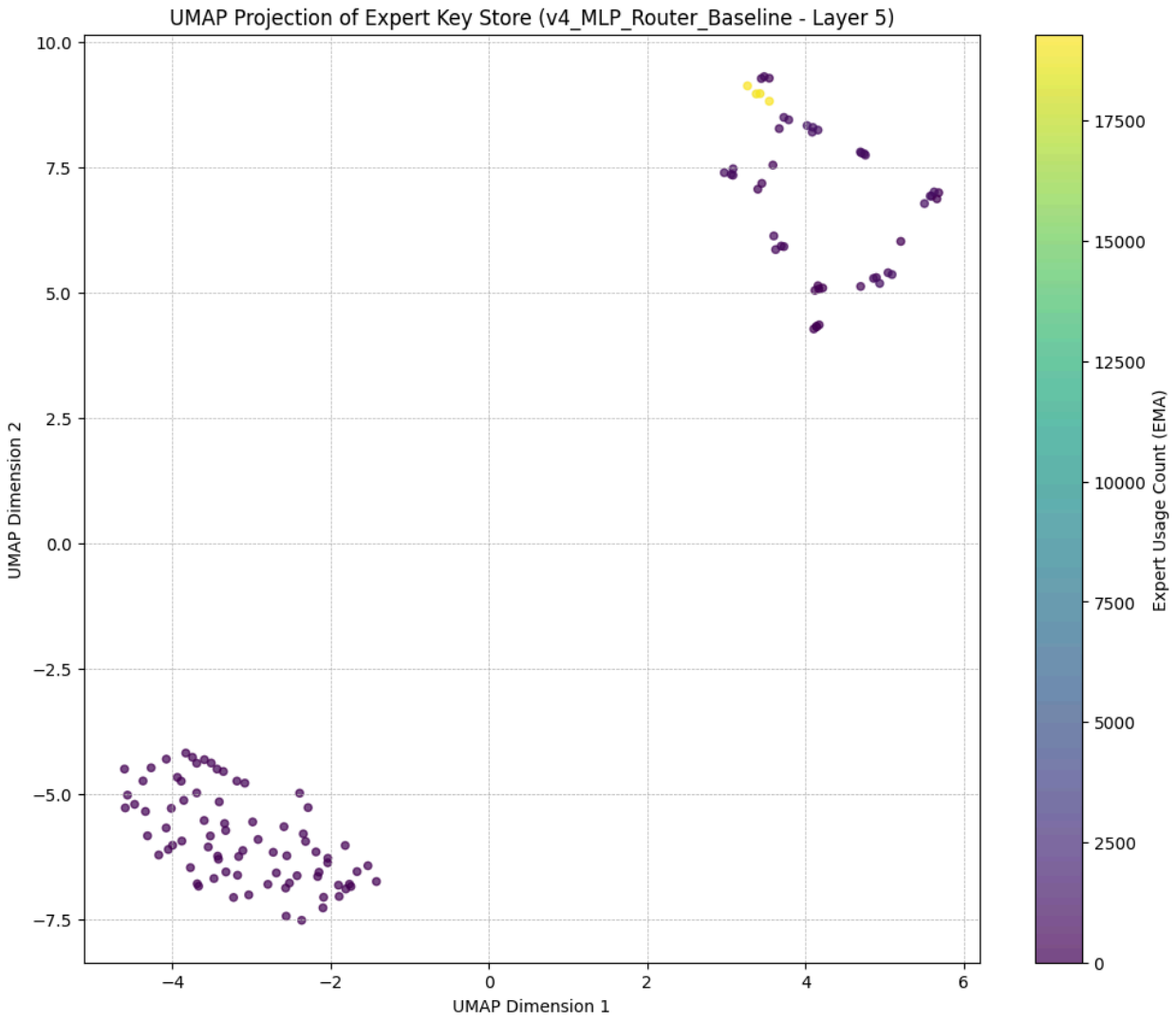


Aggregate Utilization Analysis (from Middle Layer)

- Expert Usage (Layer 5): 128/128 (100.00%)
- Final Gini Coefficient (Layer 5): 0.9143
- Final Shannon Entropy (Layer 5): 3.4272 (Max: 7.0000)



Key Store Structure Visualization (from Middle Layer)



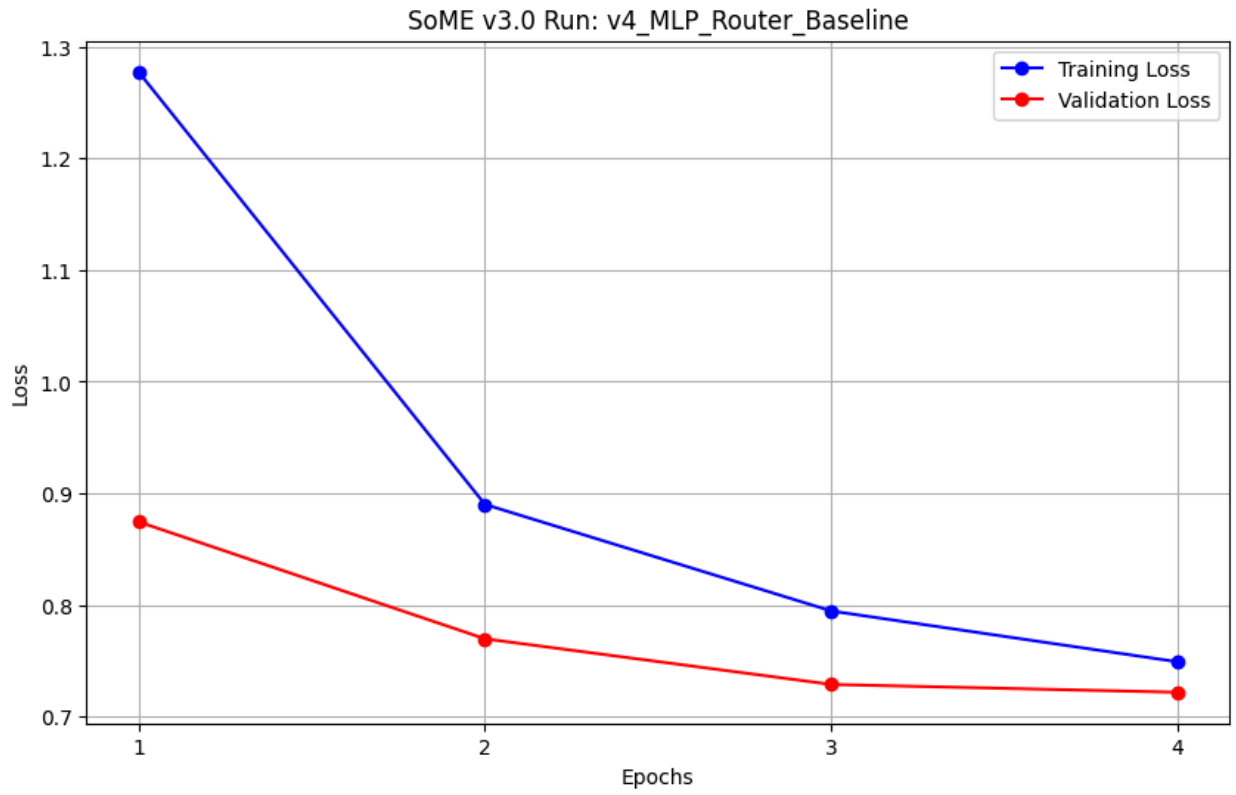
C1:

- D_MODEL: 512
- NUM_LAYERS: 10
- NUM_HEADS: 8
- SEQ_LEN: 768
- BATCH_SIZE = 32
- VOCAB_SIZE = 8192
- SoME Parameters:
  - NUM_EXPERTS: 64
  - D_FFN: 1024
  - top_k: 1
  - alpha (Attraction): 0.01
  - beta (Peer-Pull): 0.005
  - delta: 0.001

    - theta_percentile: 0.05
    - ema_decay (Inertia): 0.99
- Training Parameters:
    - train_subset_size = 10000
    - val_subset_size = 2000
    - LEARNING_RATE = 6e-4
    - TRAINING_TEMP = 0.8
    - EPOCHS = 4
- Resulting Architecture:
    - Total parameters: 701.50M
    - Trainable parameters: 29.42M (4.19%)
    - Total training steps: 1248
    - Using expert initialization method: default
    - Router: We'll use the MLP Router (d_model -> 2*d_model -> d_model)

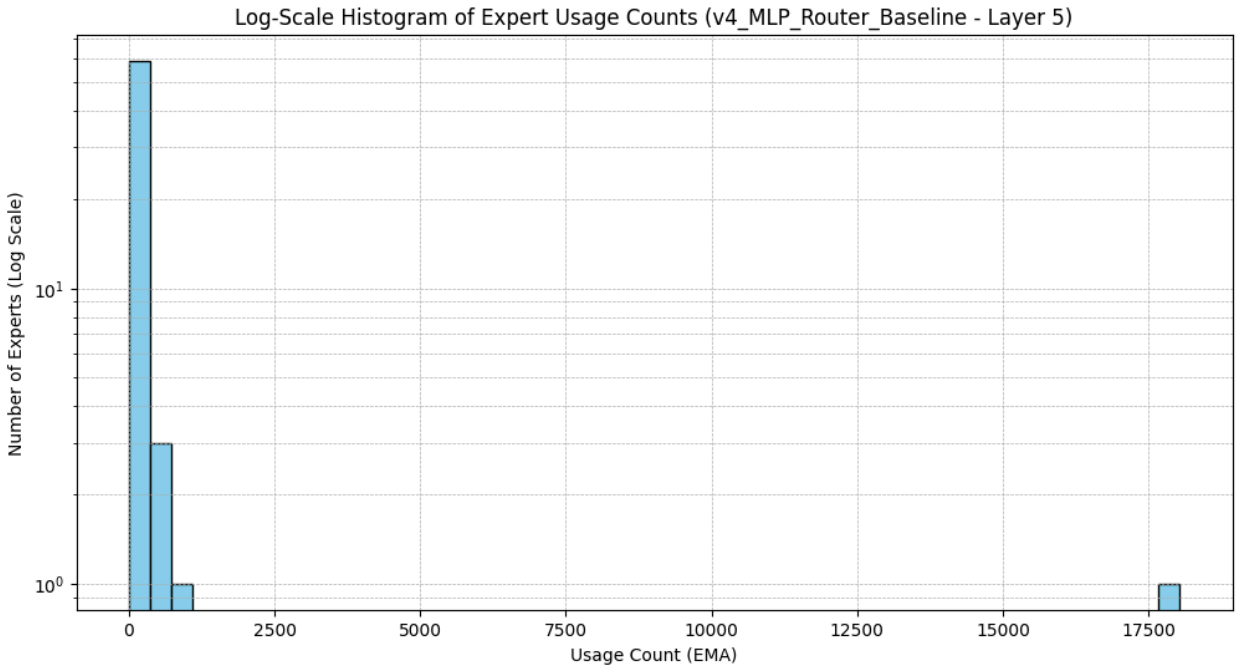
Results:

- Epoch 1:
    - Train Loss = 1.2775,
    - Val Loss = 0.8746,
    - Val Perplexity = 2.40
    - Middle Layer Expert Metrics:
        - Gini = 0.919,
        - Entropy = 1.970
- Epoch 2:
    - Train Loss = 0.8901,
    - Val Loss = 0.7698,
    - Val Perplexity = 2.16
    - Middle Layer Expert Metrics:
        - Gini = 0.914,
        - Entropy = 1.995
- Epoch 3:
    - Train Loss = 0.7947,
    - Val Loss = 0.7289,
    - Val Perplexity = 2.07
    - Middle Layer Expert Metrics:
        - Gini = 0.909,
        - Entropy = 2.024
- Epoch 4:
    - Train Loss = 0.7492,
    - Val Loss = 0.7219,
    - Val Perplexity = 2.06
    - Middle Layer Expert Metrics:
        - Gini = 0.909,
        - Entropy = 2.022

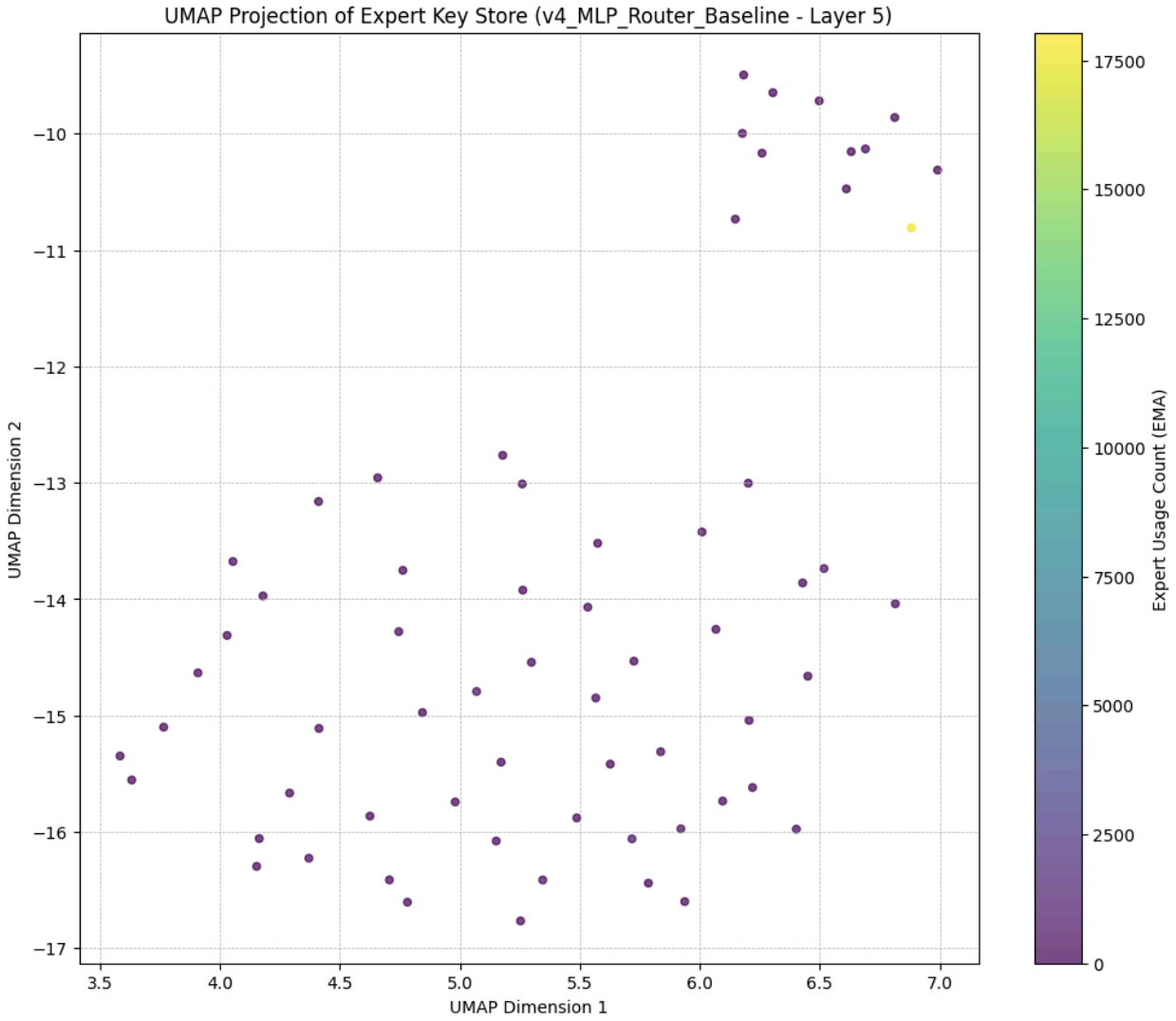


Aggregate Utilization Analysis (from Middle Layer)

- Expert Usage (Layer 5): 64/64 (100.00%)
- Final Gini Coefficient (Layer 5): 0.9088
- Final Shannon Entropy (Layer 5): 2.0218 (Max: 6.0000)



Key Store Structure Visualization (from Middle Layer)



C2:

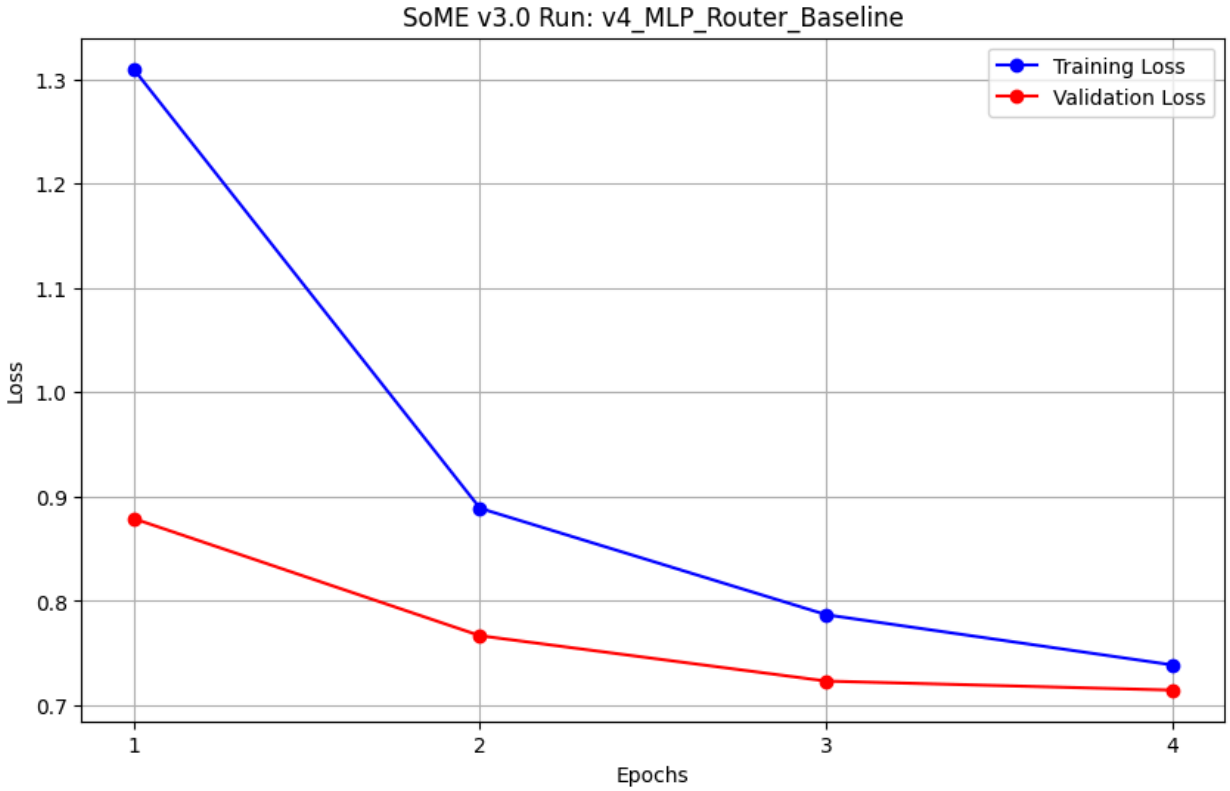
- D_MODEL: 512
- NUM_LAYERS: 10
- NUM_HEADS: 8
- SEQ_LEN: 768
- BATCH_SIZE = 32
- VOCAB_SIZE = 8192
- SoME Parameters:
    - NUM_EXPERTS: 256
    - D_FFN: 1024
    - top_k: 8

    - alpha (Attraction): 0.01
    - beta (Peer-Pull): 0.005
    - delta: 0.001
    - theta_percentile: 0.05
    - ema_decay (Inertia): 0.99
- Training Parameters:
    - train_subset_size = 10000
    - val_subset_size = 2000
    - LEARNING_RATE = 6e-4
    - TRAINING_TEMP = 0.8
    - EPOCHS = 4
- Resulting Architecture:
    - Total parameters: 2717.71M
    - Trainable parameters: 29.42M (1.08%)
    - Total training steps: 1248
    - Using expert initialization method: default
    - Router: We'll use the MLP Router (d_model -> 2*d_model -> d_model)

Results:

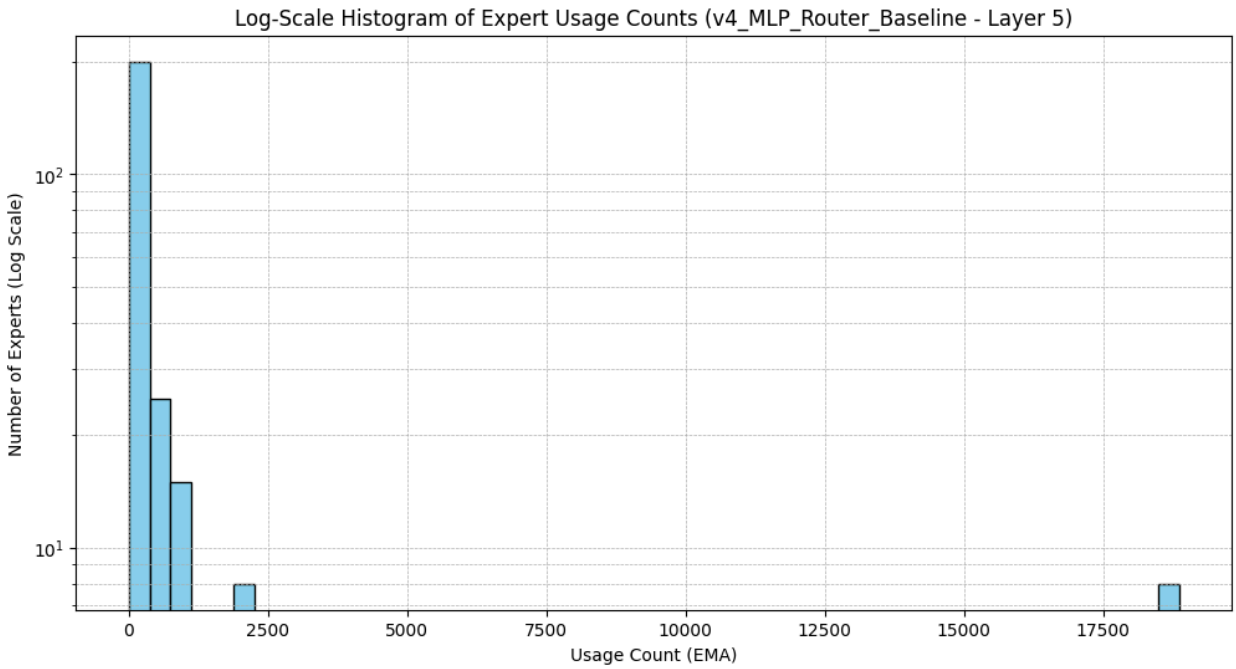
- Epoch 1:
    - Train Loss = 1.3097,
    - Val Loss = 0.8790,
    - Val Perplexity = 2.41
    - Middle Layer Expert Metrics:
        - Gini = 0.940,
        - Entropy = 4.232
- Epoch 2:
    - Train Loss = 0.8888,
    - Val Loss = 0.7667,
    - Val Perplexity = 2.15
    - Middle Layer Expert Metrics:
        - Gini = 0.934,
        - Entropy = 4.333
- Epoch 3:
    - Train Loss = 0.7867,
    - Val Loss = 0.7232,
    - Val Perplexity = 2.06
    - Middle Layer Expert Metrics:
        - Gini = 0.925,
        - Entropy = 4.458
- Epoch 4:
    - Train Loss = 0.7386,
    - Val Loss = 0.7145,
    - Val Perplexity = 2.04

- Middle Layer Expert Metrics:
  - Gini = 0.925,
  - Entropy = 4.401

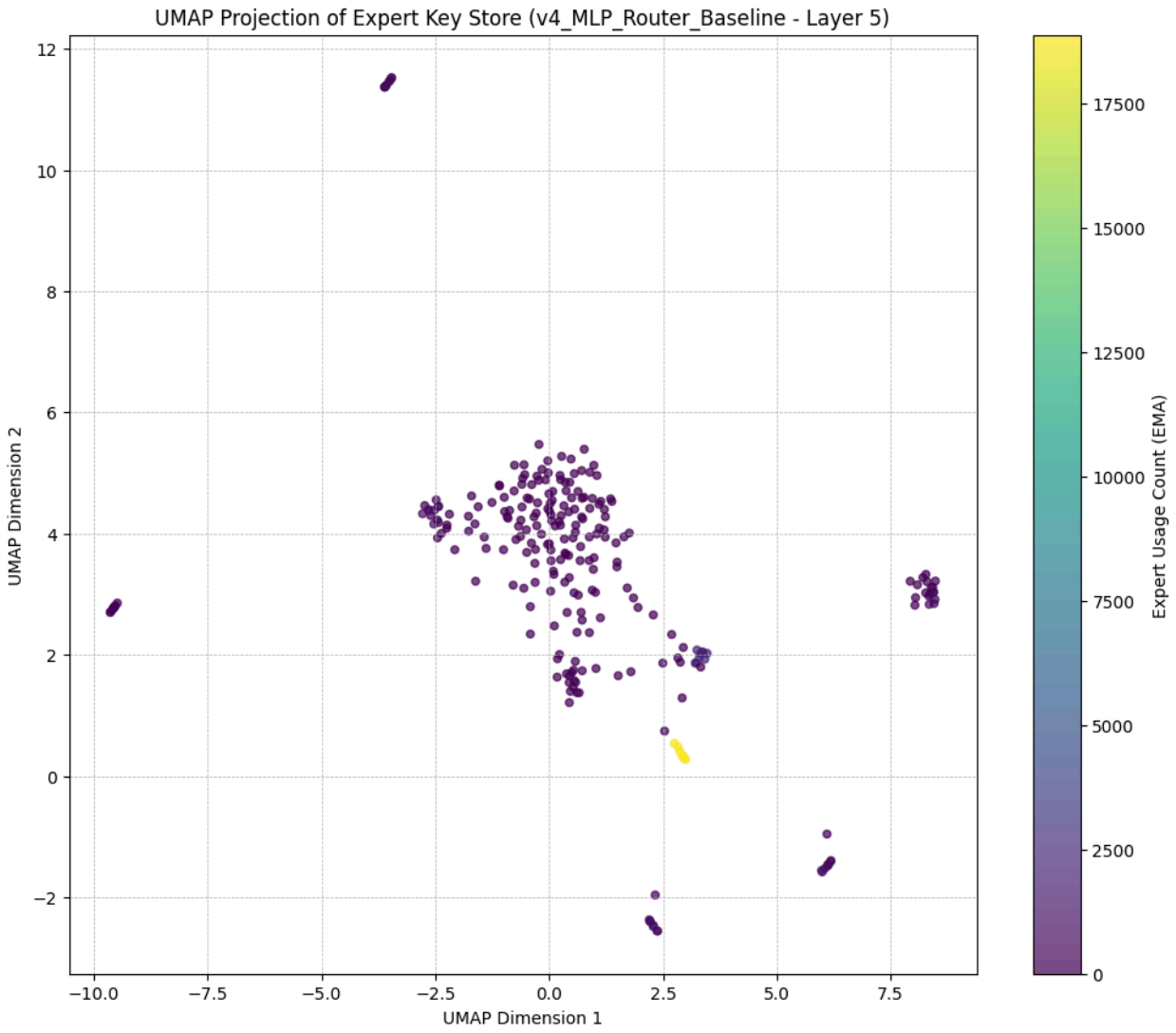


Aggregate Utilization Analysis (from Middle Layer)

- Expert Usage (Layer 5): 256/256 (100.00%)
- Final Gini Coefficient (Layer 5): 0.9252
- Final Shannon Entropy (Layer 5): 4.4009 (Max: 8.0000)



Key Store Structure Visualization (from Middle Layer)



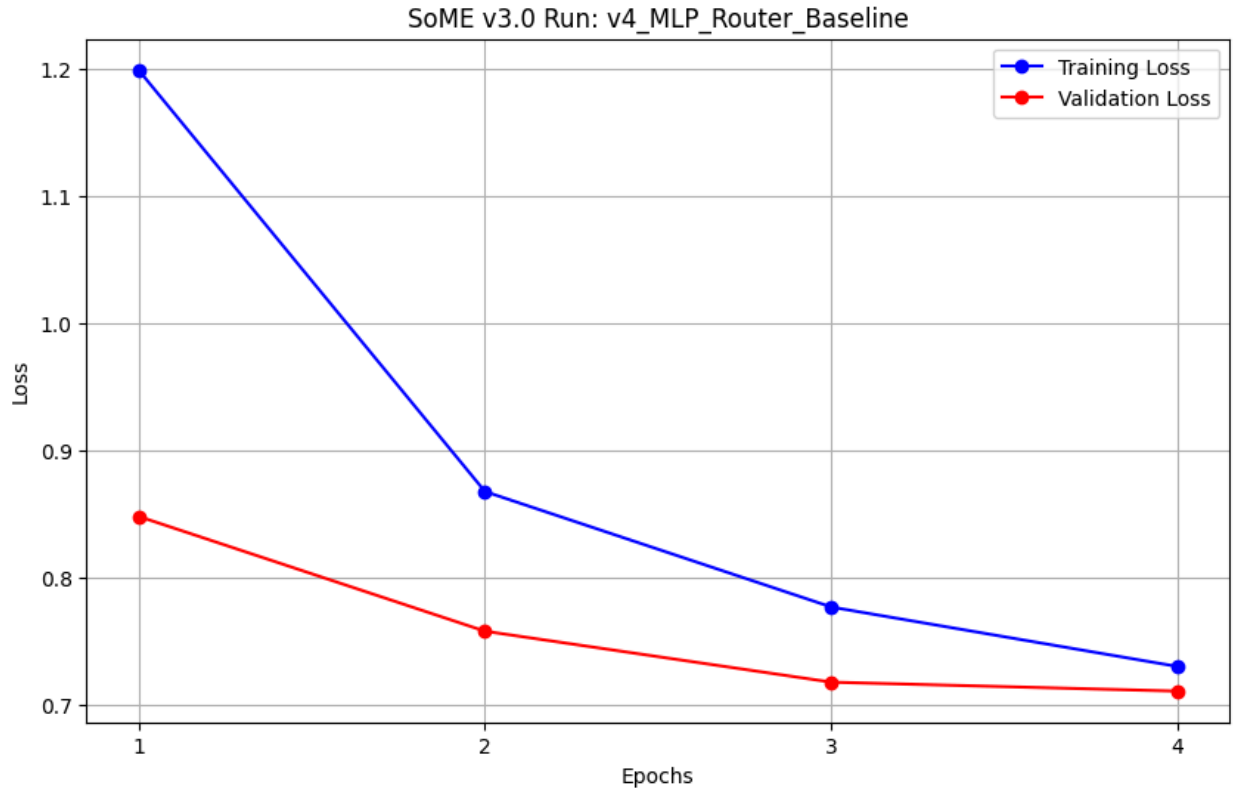
C3:

- D_MODEL: 512
- NUM_LAYERS: 10
- NUM_HEADS: 8
- SEQ_LEN: 768
- BATCH_SIZE = 32
- VOCAB_SIZE = 8192
- SoME Parameters:
    - NUM_EXPERTS: 512
    - D_FFN: 1024
    - top_k: 8
    - alpha (Attraction): 0.01
    - beta (Peer-Pull): 0.005
    - delta: 0.001

    - theta_percentile: 0.05
    - ema_decay (Inertia): 0.99
- Training Parameters:
    - train_subset_size = 10000
    - val_subset_size = 2000
    - LEARNING_RATE = 6e-4
    - TRAINING_TEMP = 0.8
    - EPOCHS = 4
- Resulting Architecture:
    - Total parameters: 5406.00M
    - Trainable parameters: 29.42M (0.54%)
    - Total training steps: 1248
    - Using expert initialization method: default
    - Router: We'll use the MLP Router (d_model -> 2*d_model -> d_model)

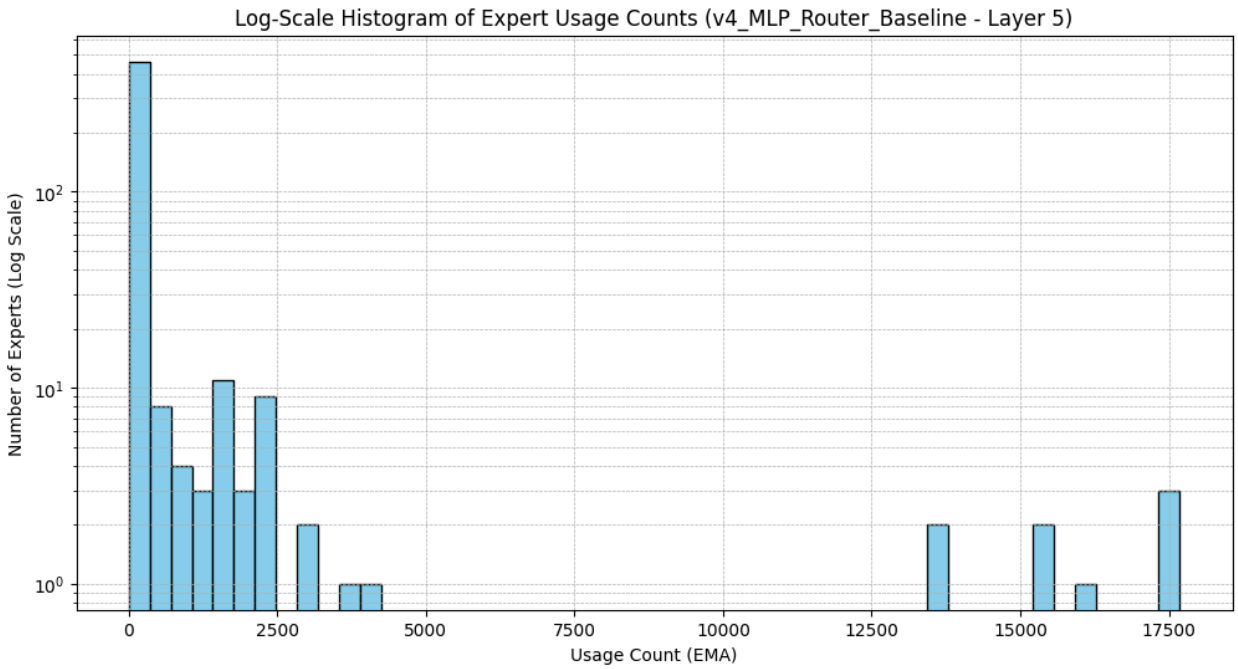
Results:

- Epoch 1:
    - Train Loss = 1.1991,
    - Val Loss = 0.8480,
    - Val Perplexity = 2.33
    - Middle Layer Expert Metrics:
        - Gini = 0.962,
        - Entropy = 4.461
- Epoch 2:
    - Train Loss = 0.8675,
    - Val Loss = 0.7577,
    - Val Perplexity = 2.13
    - Middle Layer Expert Metrics:
        - Gini = 0.964,
        - Entropy = 4.438
- Epoch 3:
    - Train Loss = 0.7767,
    - Val Loss = 0.7176,
    - Val Perplexity = 2.05
    - Middle Layer Expert Metrics:
        - Gini = 0.962,
        - Entropy = 4.566
- Epoch 4:
    - Train Loss = 0.7299,
    - Val Loss = 0.7105,
    - Val Perplexity = 2.04
    - Middle Layer Expert Metrics:
        - Gini = 0.959,
        - Entropy = 4.702

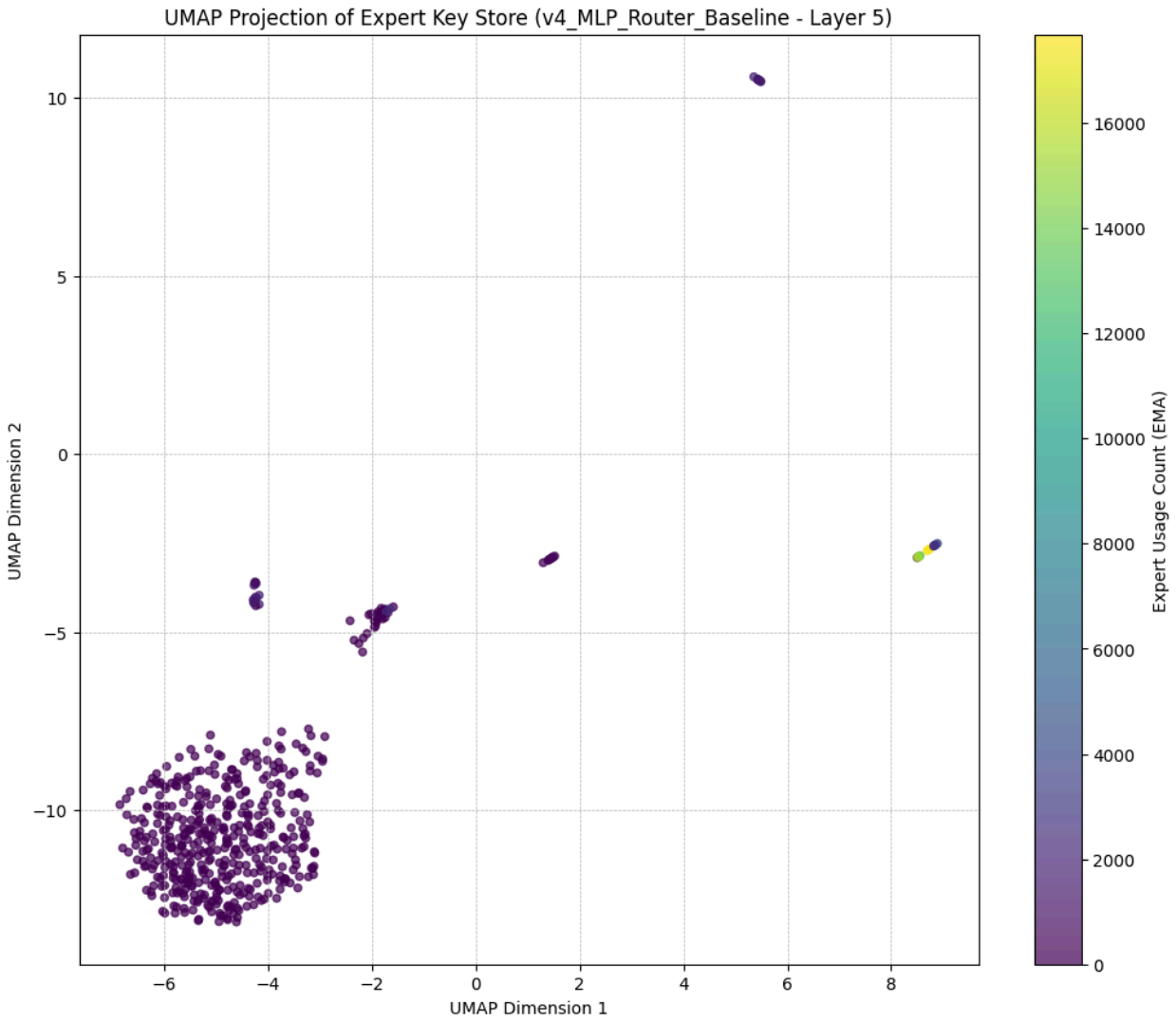


Aggregate Utilization Analysis (from Middle Layer)

- Expert Usage (Layer 5): 501/512 (97.85%)
- Final Gini Coefficient (Layer 5): 0.9589
- Final Shannon Entropy (Layer 5): 4.7015 (Max: 9.0000)



Key Store Structure Visualization (from Middle Layer)


UMAP Projection of Expert Key Store (v4_MLP_Router_Baseline - Layer 5)
UMAP Dimension 1
UMAP Dimension 2
Expert Usage Count (EMA)
10
5
0
−5
−10
−6
−4
−2
0
2
4
6
8
16000
14000
12000
10000
8000
6000
4000
2000
0