Professional (Series B) Ablation Studies Base Results

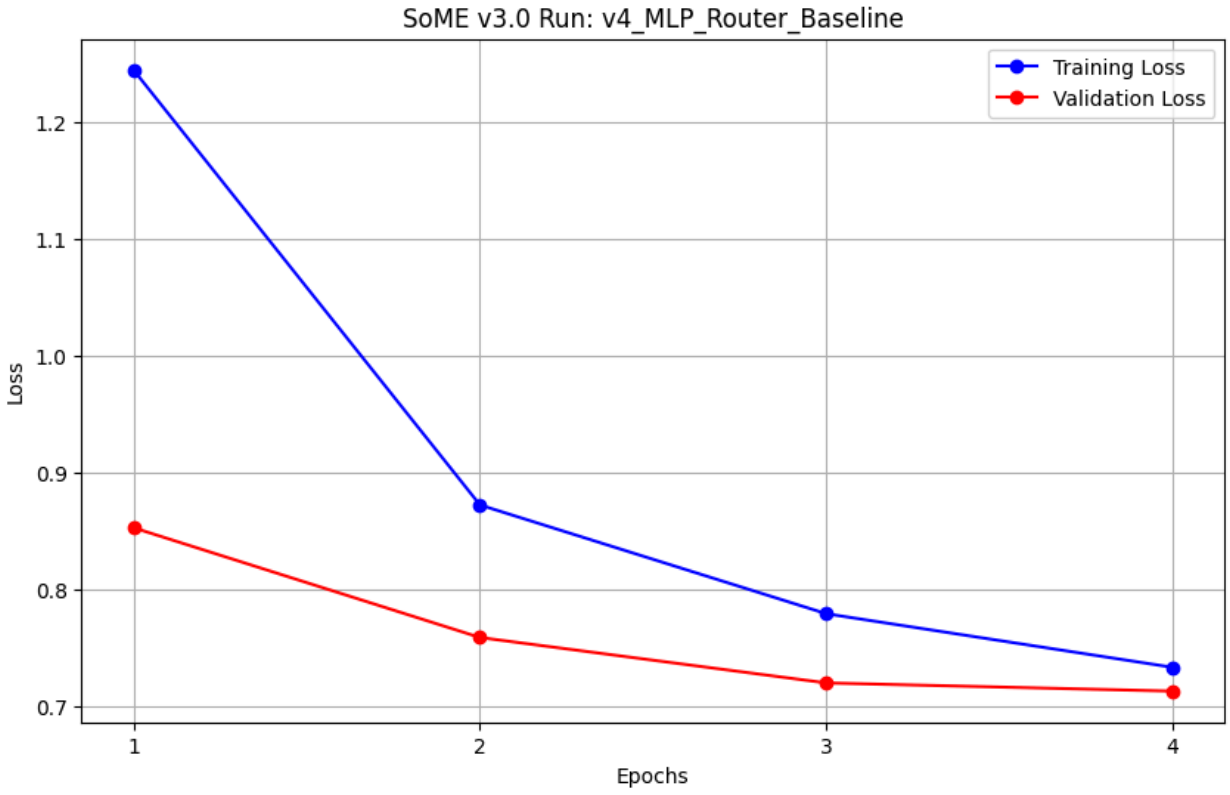
Related to Last Ablation (as Base)
A4:

- 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: 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: 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:

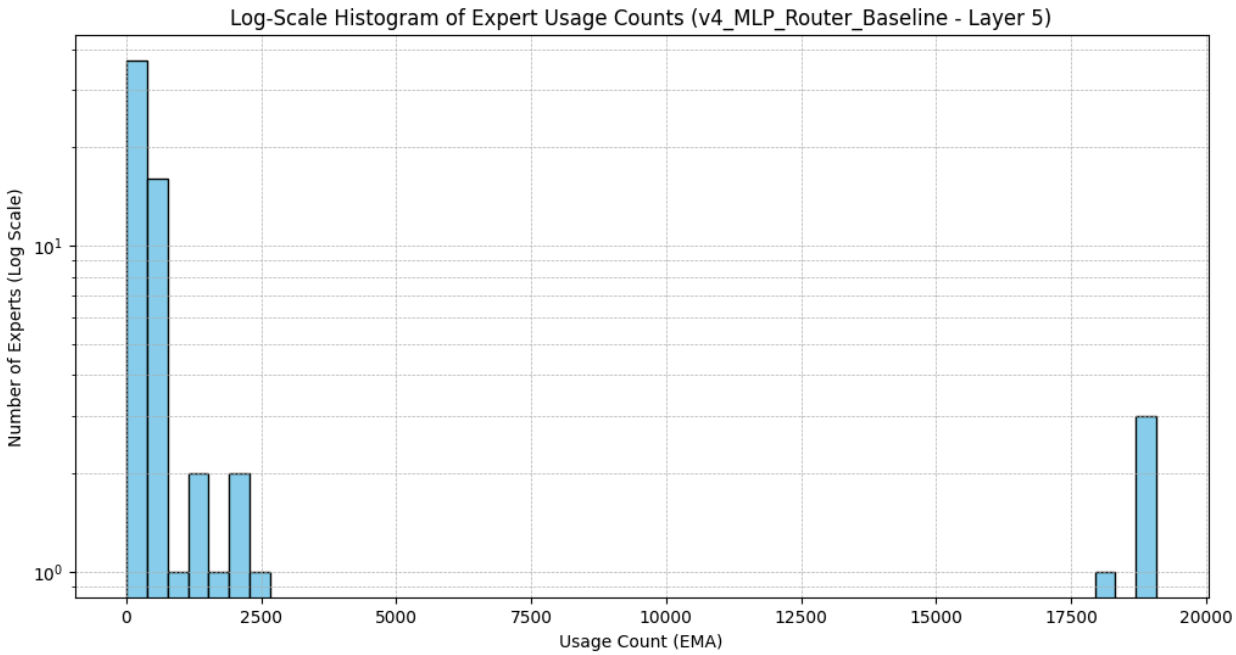
- Epoch 1:
    - Train Loss = 1.2452,
    - Val Loss = 0.8527,
    - Val Perplexity = 2.35
    - Middle Layer Expert Metrics:
        - Gini = 0.872,
        - Entropy = 3.396
- Epoch 2:
    - Train Loss = 0.8723,
    - Val Loss = 0.7586,
    - Val Perplexity = 2.14

    - Middle Layer Expert Metrics:
        - Gini = 0.842,
        - Entropy = 3.658
- Epoch 3:
    - Train Loss = 0.7790,
    - Val Loss = 0.7197,
    - Val Perplexity = 2.05
    - Middle Layer Expert Metrics:
        - Gini = 0.864,
        - Entropy = 3.336
- Epoch 4:
    - Train Loss = 0.7330,
    - Val Loss = 0.7126,
    - Val Perplexity = 2.04
    - Middle Layer Expert Metrics:
        - Gini = 0.851,
        - Entropy = 3.413

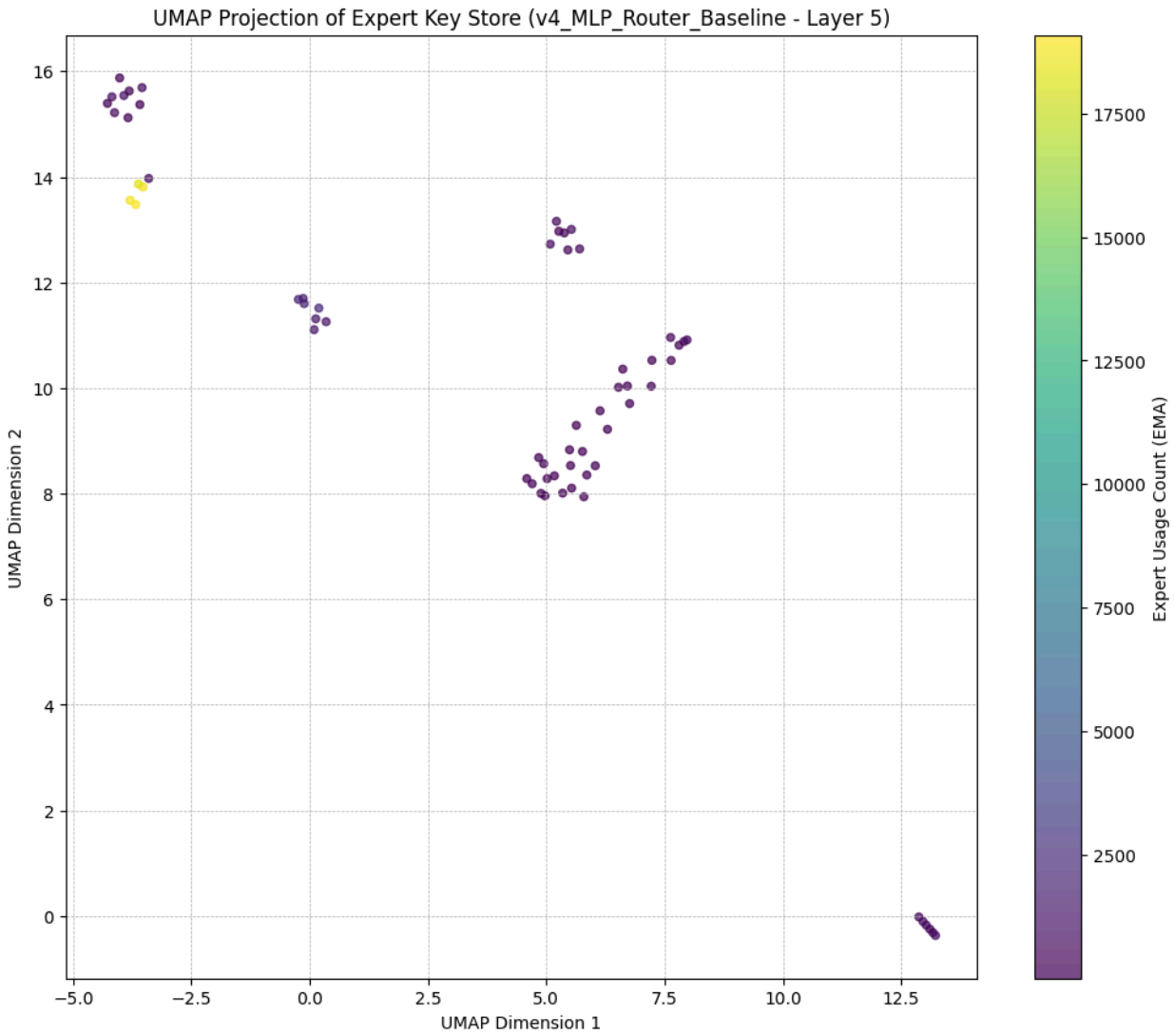


Aggregate Utilization Analysis (from Middle Layer)

- Expert Usage (Layer 5): 64/64 (100.00%)
- Final Gini Coefficient (Layer 5): 0.8510
- Final Shannon Entropy (Layer 5): 3.4132 (Max: 6.0000)



Key Store Structure Visualization (from Middle Layer)



B1:

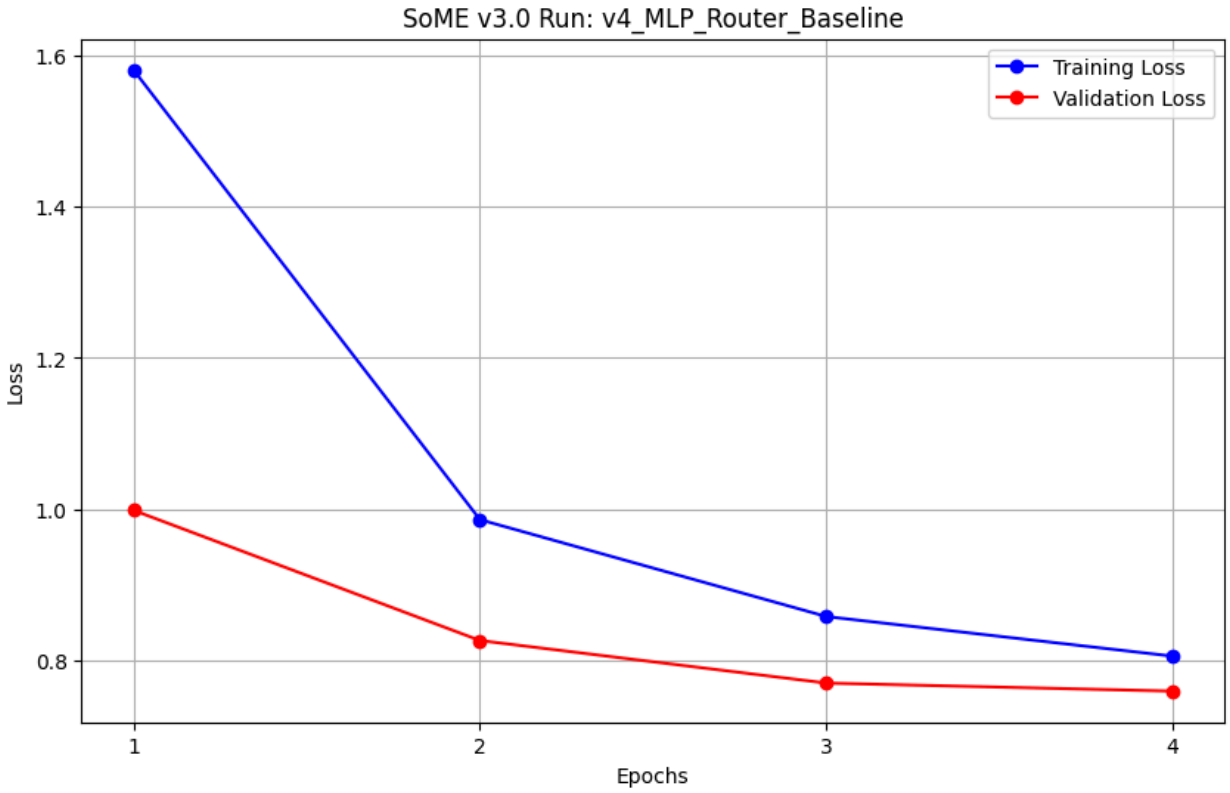
- D_MODEL: 512
- NUM_LAYERS: 10
- NUM_HEADS: 8
- SEQ_LEN: 768
- BATCH_SIZE = 32
- VOCAB_SIZE = 8192
- SoME Parameters:
    - NUM_EXPERTS: 128
    - D_FFN: 1024
    - top_k: 4
    - alpha (Attraction): 0.01
    - beta (Peer-Pull): 0.005

    - 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:

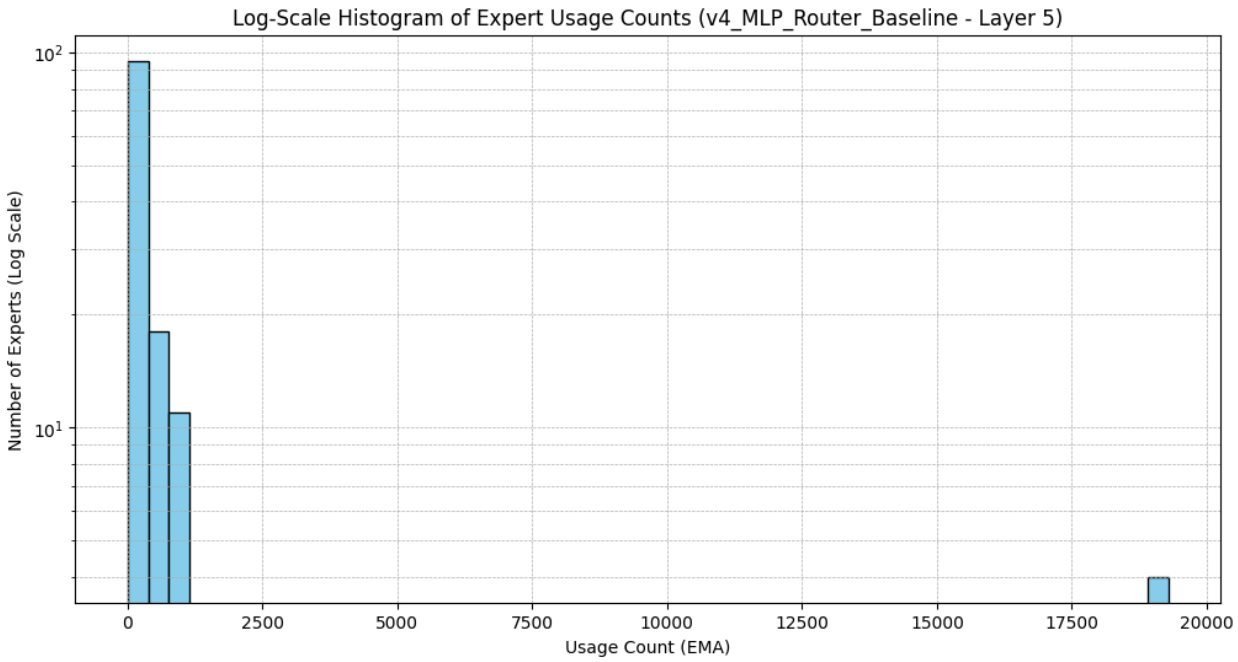
- Epoch 1:
    - Train Loss = 1.5803,
    - Val Loss = 0.9990,
    - Val Perplexity = 2.72
    - Middle Layer Expert Metrics:
        - Gini = 0.941,
        - Entropy = 3.272
- Epoch 2:
    - Train Loss = 0.9866,
    - Val Loss = 0.8270,
    - Val Perplexity = 2.29
    - Middle Layer Expert Metrics: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

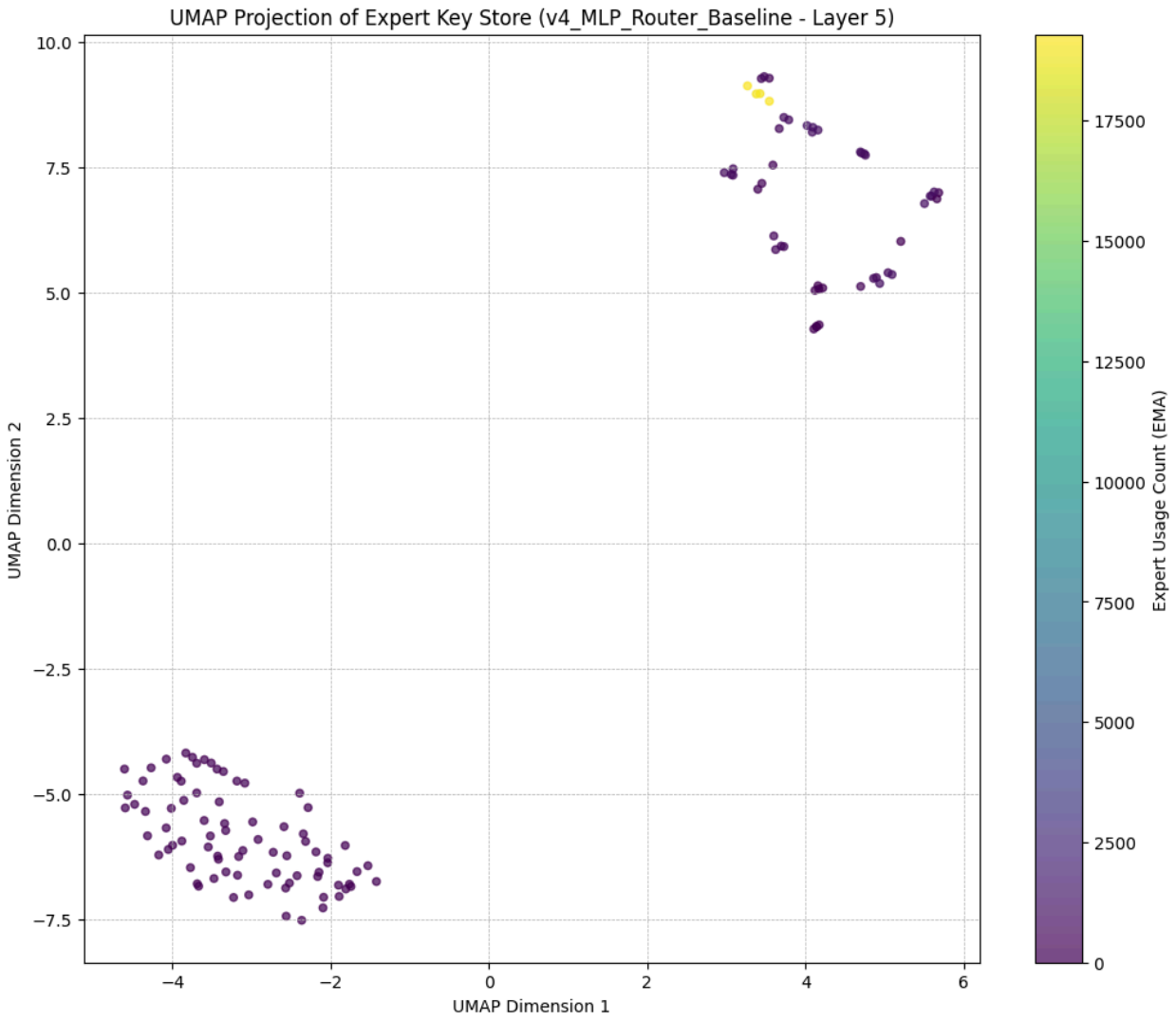


Aggregate Utilization Analysis (from Middle Layer)

- 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)



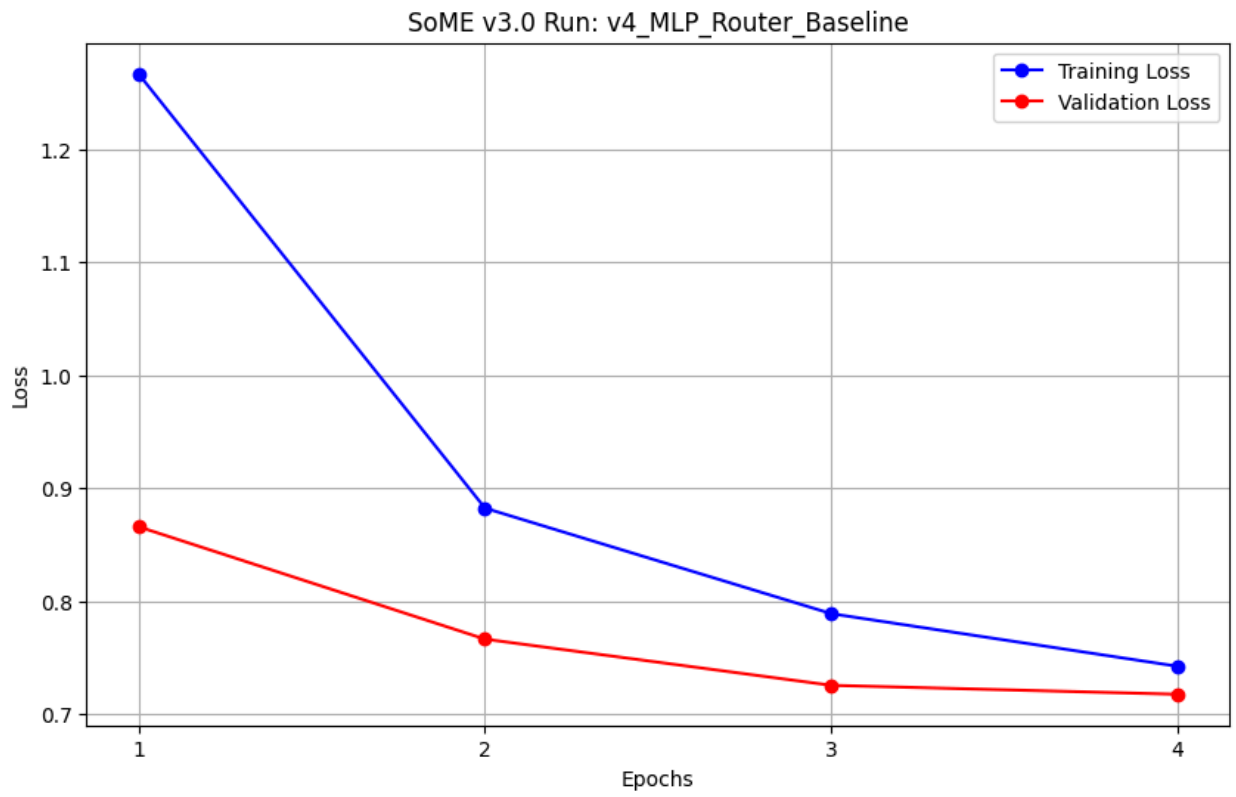
B2:

- 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: 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: 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)

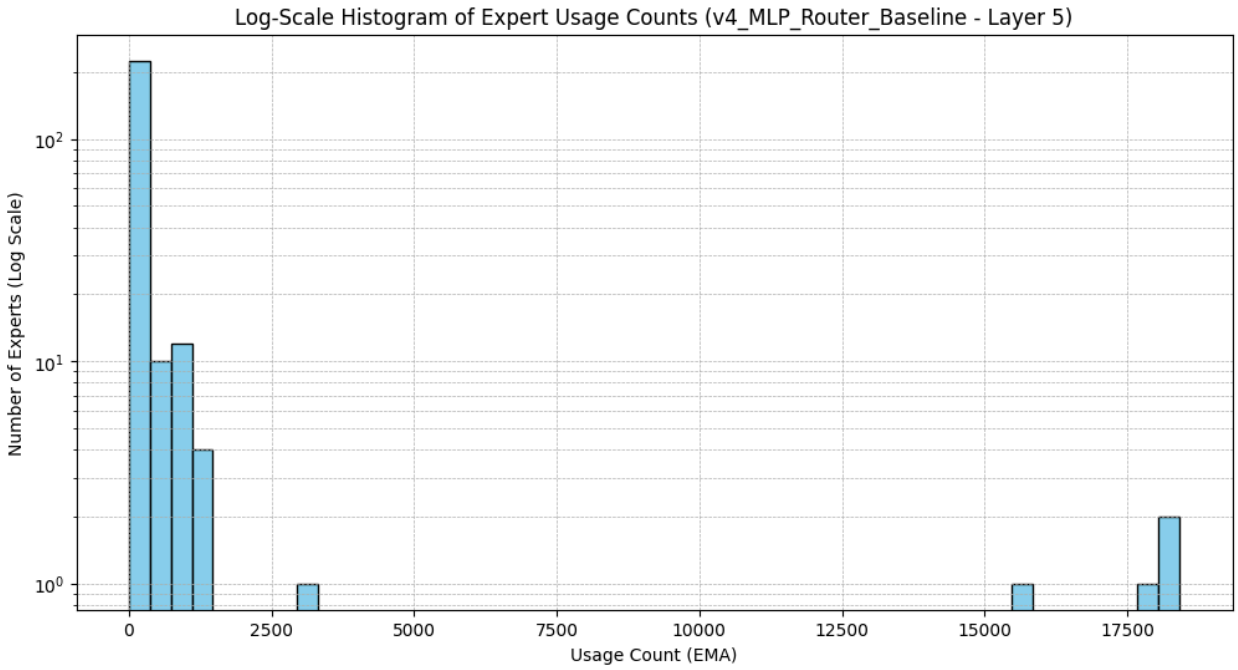
Results:

- Epoch 1:
    - Train Loss = 1.2668,
    - Val Loss = 0.8658,
    - Val Perplexity = 2.38
    - Middle Layer Expert Metrics:
        - Gini = 0.948,
        - Entropy = 4.094
- Epoch 2:
    - Train Loss = 0.8823,
    - Val Loss = 0.7663,
    - Val Perplexity = 2.15
    - Middle Layer Expert Metrics:
        - Gini = 0.959,
        - Entropy = 3.513
- Epoch 3:
    - Train Loss = 0.7888,
    - Val Loss = 0.7254,
    - Val Perplexity = 2.07
    - Middle Layer Expert Metrics:
        - Gini = 0.956,
        - Entropy = 3.719
- Epoch 4:
    - Train Loss = 0.7422,
    - Val Loss = 0.7174,
    - Val Perplexity = 2.05
    - Middle Layer Expert Metrics:
        - Gini = 0.952,
        - Entropy = 3.725

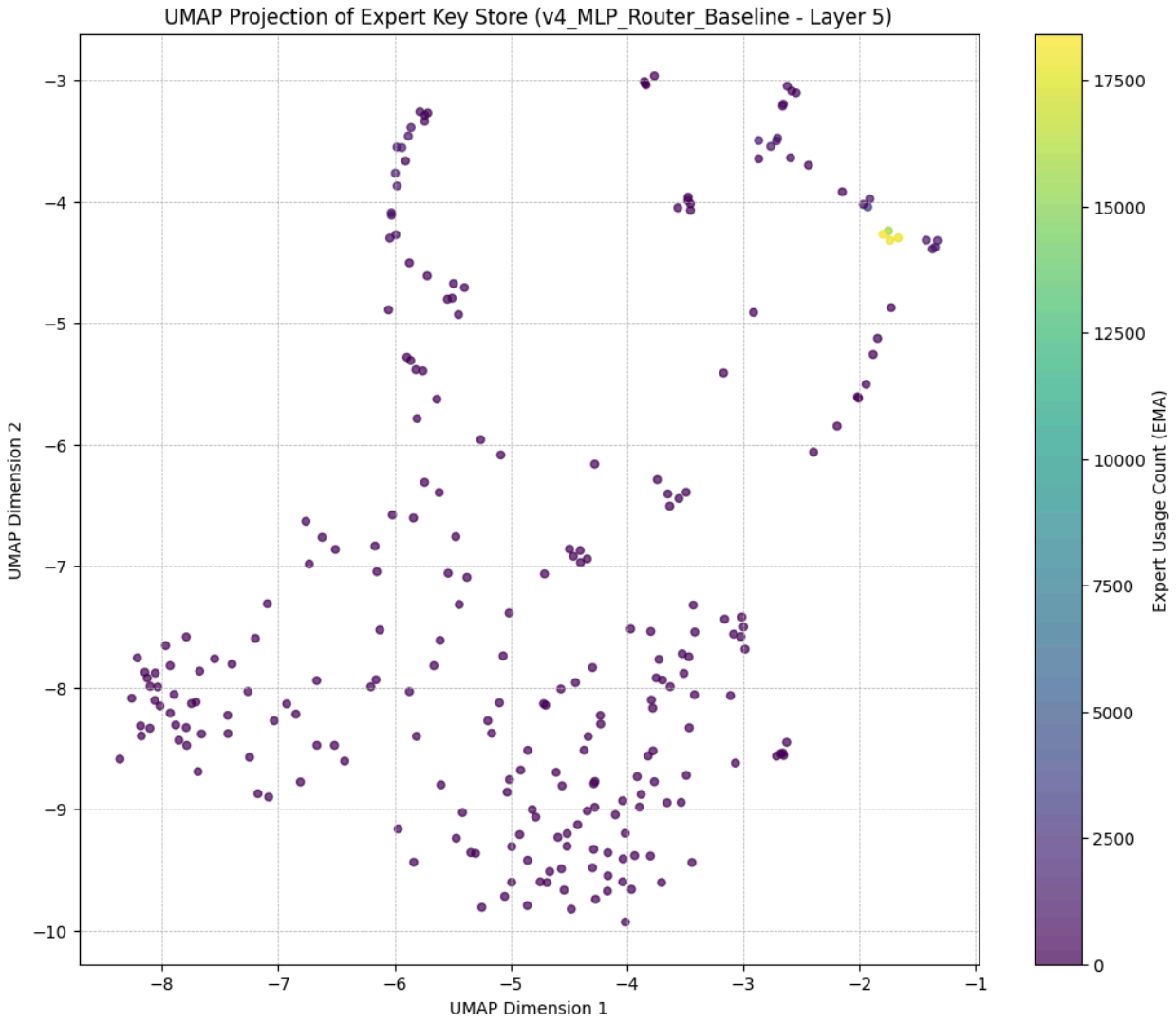


Aggregate Utilization Analysis (from Middle Layer)

- Expert Usage (Layer 5): 256/256 (100.00%)
- Final Gini Coefficient (Layer 5): 0.9516
- Final Shannon Entropy (Layer 5): 3.7252 (Max: 8.0000)



Key Store Structure Visualization (from Middle Layer)



B3:

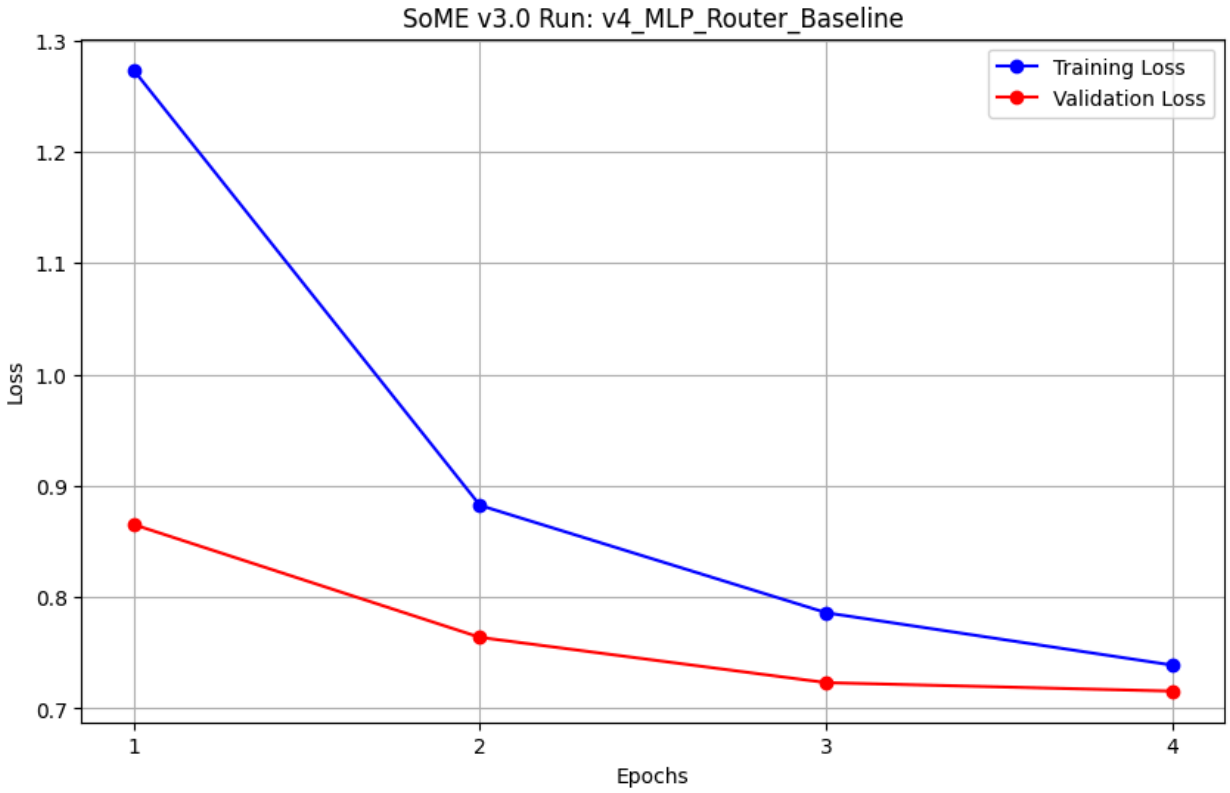
- D_MODEL: 512
- NUM_LAYERS: 10
- NUM_HEADS: 8
- SEQ_LEN: 768
- BATCH_SIZE = 32
- VOCAB_SIZE = 8192
- SoME Parameters:
    - NUM_EXPERTS: 512
    - D_FFN: 1024
    - top_k: 4
    - alpha (Attraction): 0.01
    - beta (Peer-Pull): 0.005

    - 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: 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:

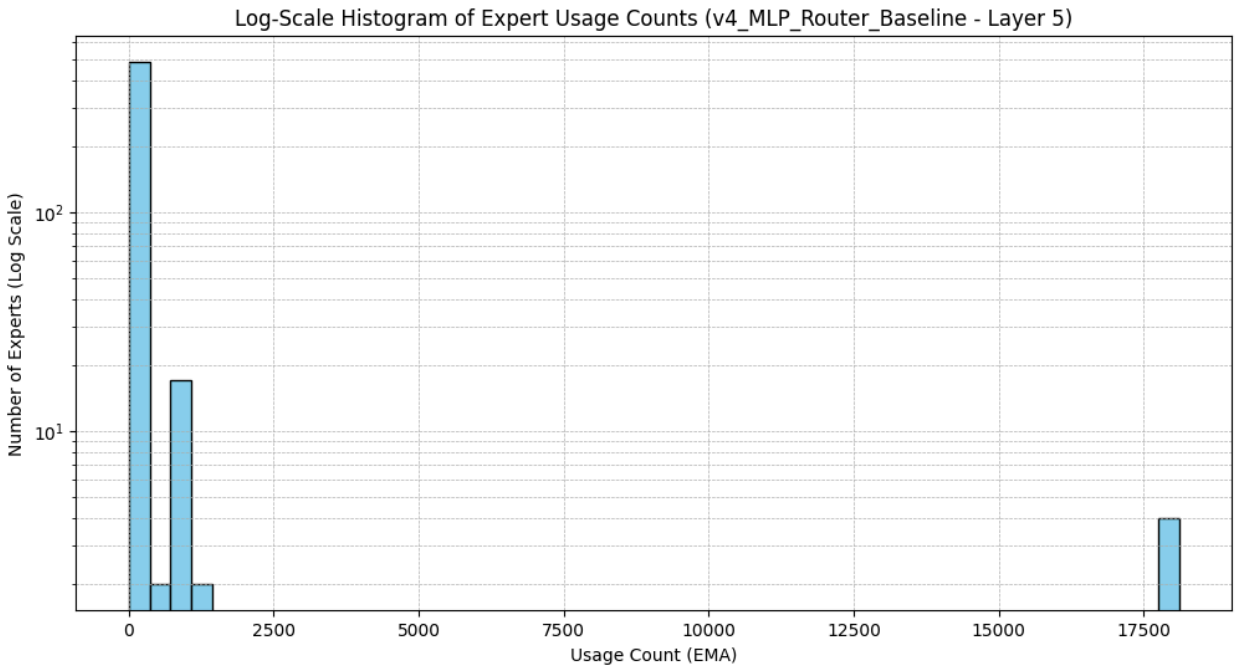
- Epoch 1:
    - Train Loss = 1.2734,
    - Val Loss = 0.8652,
    - Val Perplexity = 2.38
    - Middle Layer Expert Metrics:
        - Gini = 0.979,
        - Entropy = 3.565
- Epoch 2:
    - Train Loss = 0.8824,
    - Val Loss = 0.7638,
    - Val Perplexity = 2.15
    - Middle Layer Expert Metrics:
        - Gini = 0.972,
        - Entropy = 3.744
- Epoch 3:
    - Train Loss = 0.7859,
    - Val Loss = 0.7231,
    - Val Perplexity = 2.06
    - Middle Layer Expert Metrics:
        - Gini = 0.971,
        - Entropy = 3.841
- Epoch 4:
    - Train Loss = 0.7388,
    - Val Loss = 0.7155,
    - Val Perplexity = 2.05
    - Middle Layer Expert Metrics:
        - Gini = 0.972,

- Entropy = 3.767

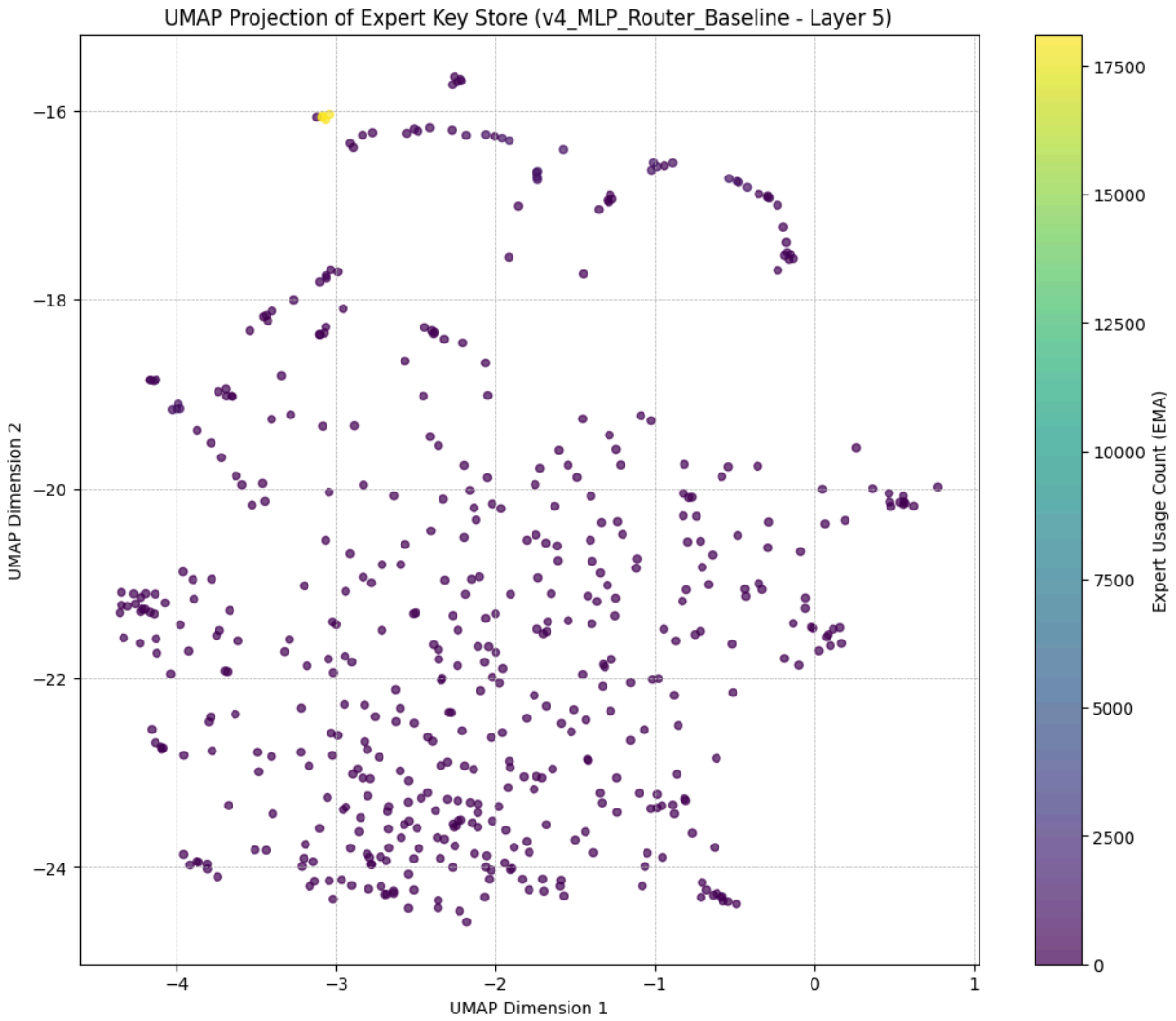


Aggregate Utilization Analysis (from Middle Layer)

- Expert Usage (Layer 5): 497/512 (97.07%)
- Final Gini Coefficient (Layer 5): 0.9719
- Final Shannon Entropy (Layer 5): 3.7674 (Max: 9.0000)



Key Store Structure Visualization (from Middle Layer)


UMAP Projection of Expert Key Store (v4_MLP_Router_Baseline - Layer 5)
UMAP Dimension 2
UMAP Dimension 1
−16
−18
−20
−22
−24
−4
−3
−2
−1
0
1
Expert Usage Count (EMA)
17500
15000
12500
10000
7500
5000
2500
0