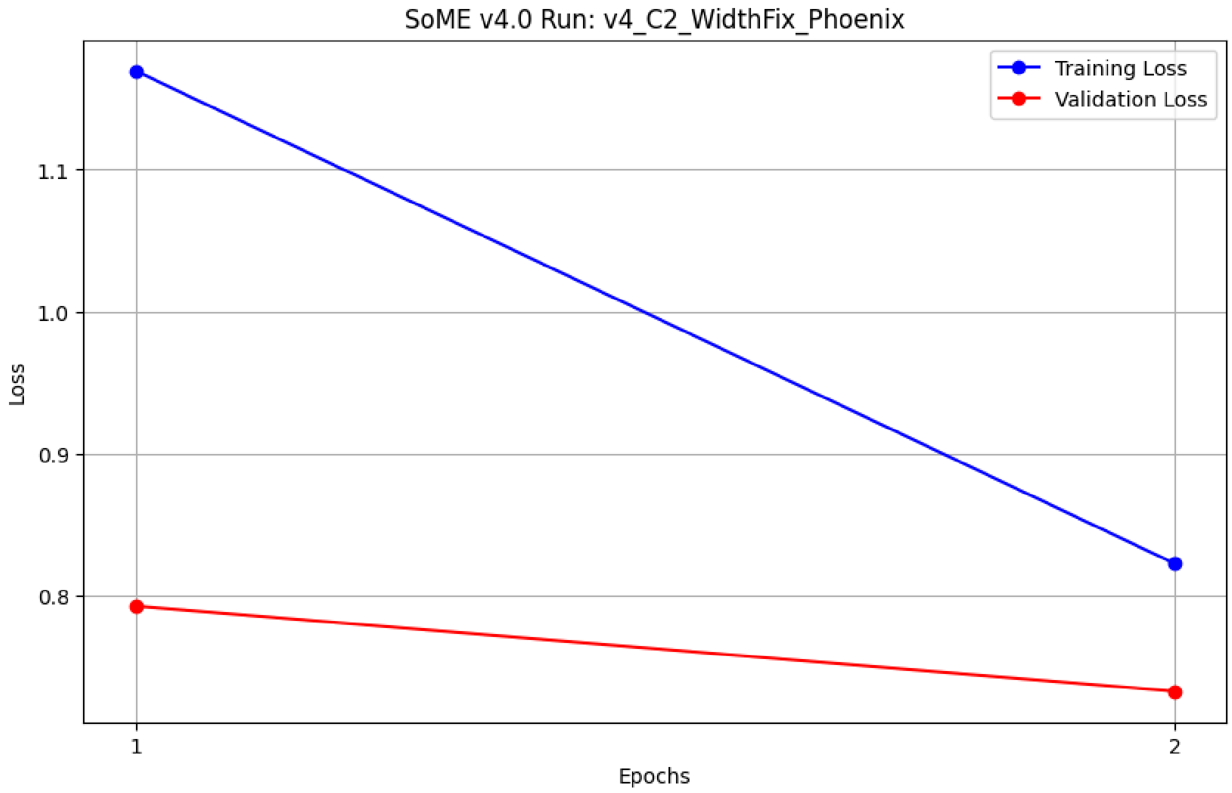
SoME v3.5 (Decay Fix) Results

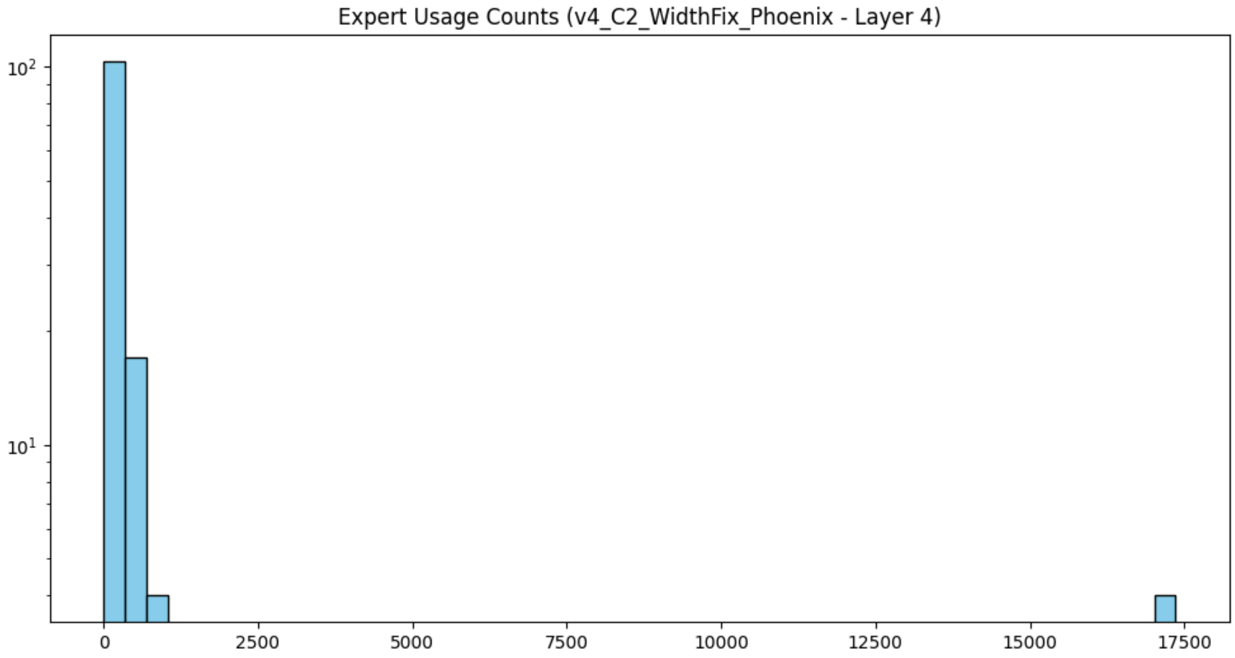
```
--- Starting Experiment: v4_C2_WidthFix_Phoenix ---

--- Part 2: Data Preparation & Configuration ---
Training custom tokenizer...

README.md:
  1.06k/?  [00:00<00:00,  111kB/s]
data/train-00000-of-00004-2d5a1467fff108(…):  100%
  249M/249M  [00:01<00:00,  221MB/s]
data/train-00001-of-00004-5852b56a2bd28f(…):  100%
  248M/248M  [00:01<00:00,  145MB/s]
data/train-00002-of-00004-a26307300439e9(…):  100%
  246M/246M  [00:01<00:00,  84.1MB/s]
data/train-00003-of-00004-d243063613e5a0(…):  100%
  248M/248M  [00:01<00:00,  165MB/s]
data/validation-00000-of-00001-869c898b5(…):  100%
  9.99M/9.99M  [00:00<00:00,  14.2MB/s]
Generating  train  split:  100%
  2119719/2119719  [00:06<00:00,  313924.00  examples/s]
Generating  validation  split:  100%
  21990/21990  [00:00<00:00,  317404.76  examples/s]
Custom tokenizer loaded with vocab size: 8192

Tokenizing dataset...

Map  (num_proc=12):  100%
  10000/10000  [00:02<00:00,  6120.58  examples/s]
Map  (num_proc=12):  100%
  1000/1000  [00:00<00:00,  197.20  examples/s]
--- Part 3: Model Definition ---

Compiling the model for faster training...

/tmp/ipython-input-3914790685.py:330: FutureWarning: `torch.cuda.amp.GradScaler(args...)` is
deprecated. Please use `torch.amp.GradScaler('cuda', args...)` instead.
  scaler = torch.cuda.amp.GradScaler()

--- Part 4: Training, Evaluation, and Metrics ---

Total parameters: 2454.52M
Trainable parameters: 36.24M (1.48%)
Total training steps: 624
```

```
--- Epoch 1/2 ---
Current Router Temperature: 2.0000

Training (Temp=2.00):   0%|          | 0/312 [00:00<?, ?it/s]/tmp/ipython-input-3914790685.py:336:
FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use
`torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
/usr/local/lib/python3.12/dist-packages/torch/optim/lr_scheduler.py:192: UserWarning: Detected
call of `lr_scheduler.step()` before `optimizer.step()`. In PyTorch 1.1.0 and later, you should call
them in the opposite order: `optimizer.step()` before `lr_scheduler.step()`.  Failure to do this will
result in PyTorch skipping the first value of the learning rate schedule. See more details at
https://pytorch.org/docs/stable/optim.html#how-to-adjust-learning-rate
  warnings.warn(
Training (Temp=2.00):   0%|          | 1/312 [00:36<3:10:16, 36.71s/it, loss=9.3105,
lr=6.0e-04]/tmp/ipython-input-3914790685.py:336: FutureWarning:
`torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda',
args...)` instead.
  with torch.cuda.amp.autocast():
Training (Temp=2.00):   1%|          | 2/312 [00:39<1:27:41, 16.97s/it, loss=9.3397,
lr=6.0e-04]/tmp/ipython-input-3914790685.py:336: FutureWarning:
`torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda',
args...)` instead.
  with torch.cuda.amp.autocast():
Training (Temp=2.00):   3%|▎         | 9/312 [00:46<08:50,  1.75s/it, loss=1.9594,
lr=6.0e-04]/tmp/ipython-input-3914790685.py:336: FutureWarning:
`torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda',
args...)` instead.
  with torch.cuda.amp.autocast():
Evaluating:   0%|          | 0/31 [00:00<?, ?it/s]/tmp/ipython-input-3914790685.py:365:
FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use
`torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Evaluating:   3%|▎         | 1/31 [00:08<04:29,  8.98s/it,
loss=0.6530]/tmp/ipython-input-3914790685.py:365: FutureWarning:
`torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda',
args...)` instead.
  with torch.cuda.amp.autocast():
Evaluating:   6%|▋         | 2/31 [00:09<02:00,  4.14s/it,
loss=0.6450]/tmp/ipython-input-3914790685.py:365: FutureWarning:
`torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda',
args...)` instead.
  with torch.cuda.amp.autocast():
```

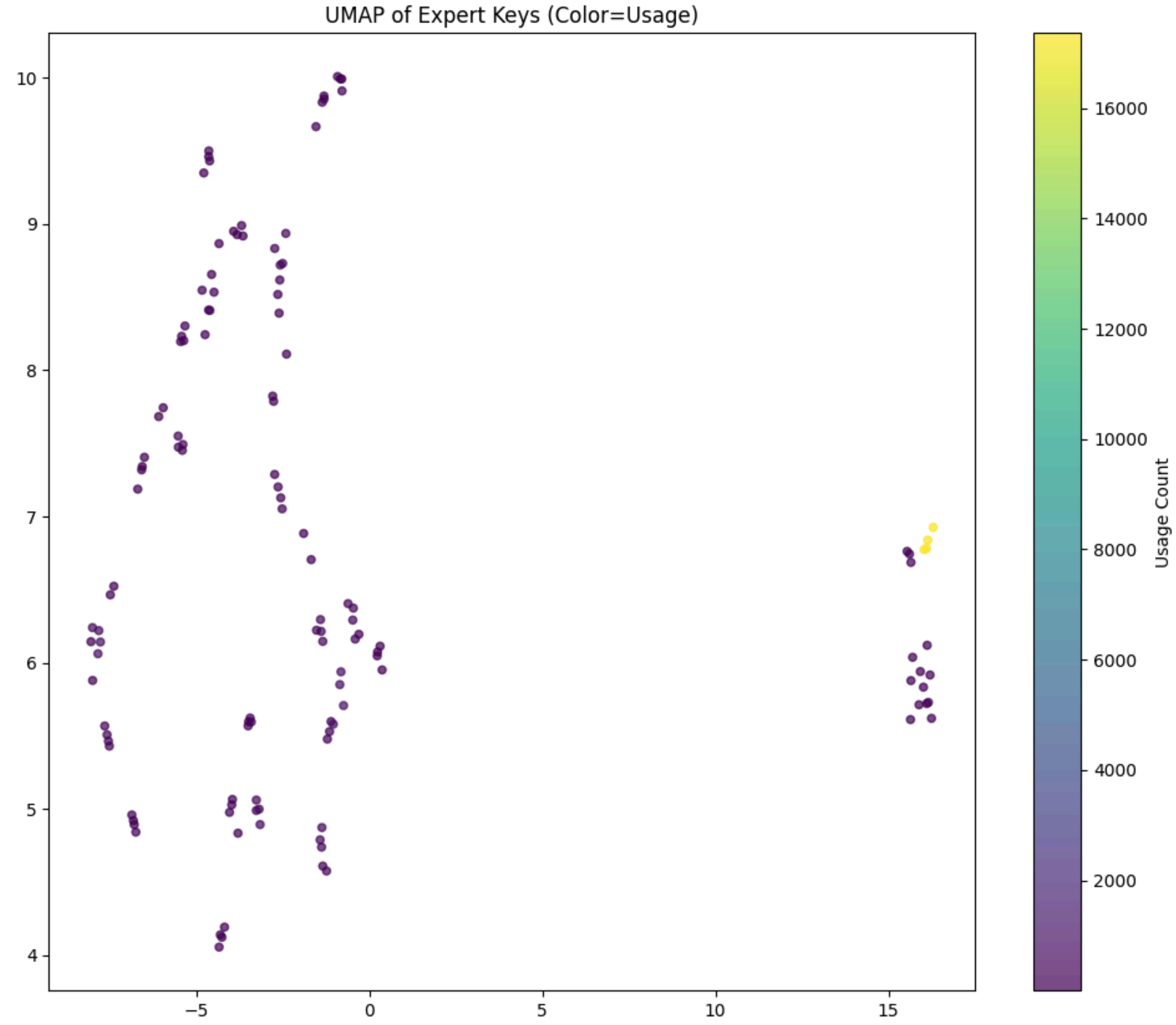
```
Epoch 1: Train Loss = 1.1694, Val Loss = 0.7924, Val Ppl = 2.21
 Middle Layer Metrics: Gini = 0.840, Entropy = 3.933, Dead Experts Pending Respawn: 0

/tmp/ipython-input-3914790685.py:330: FutureWarning: `torch.cuda.amp.GradScaler(args...)` is
deprecated. Please use `torch.amp.GradScaler('cuda', args...)` instead.
  scaler = torch.cuda.amp.GradScaler()

Model saved as best_model_v4_C2_WidthFix_Phoenix.pth

--- Epoch 2/2 ---
Current Router Temperature: 1.5000

Training (Temp=1.50):   0%|          | 0/312 [00:00<?, ?it/s]/tmp/ipython-input-3914790685.py:336:
FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use
`torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Training (Temp=1.50):   0%|          | 1/312 [00:11<58:21, 11.26s/it, loss=0.8007,
lr=3.0e-04]/tmp/ipython-input-3914790685.py:336: FutureWarning:
`torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda',
args...)` instead.
  with torch.cuda.amp.autocast():
Evaluating:   0%|          | 0/31 [00:00<?, ?it/s]/tmp/ipython-input-3914790685.py:365:
FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use
`torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():

Epoch 2: Train Loss = 0.8223, Val Loss = 0.7324, Val Ppl = 2.08
 Middle Layer Metrics: Gini = 0.835, Entropy = 3.962, Dead Experts Pending Respawn: 0
 Model saved as best_model_v4_C2_WidthFix_Phoenix.pth

--- Training Complete for v4_C2_WidthFix_Phoenix ---
```



```
--- Part 1: Dashboard Setup ---
Loading best model from: best_model_v4_C2_WidthFix_Phoenix.pth
```



```
Running UMAP projection...
```


UMAP of Expert Keys (Color=Usage)
10
9
8
7
6
5
4
−5
0
5
10
15
16000
14000
12000
10000
8000
6000
4000
2000
Usage Count