SoME v6 (8 Big Fixes + Decay Fix) Results

--- Starting Experiment: v6_Phoenix_ProofOfLife ---

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

README.md:
1.06k/? [00:00<00:00, 117kB/s]
data/train-00000-of-00004-2d5a1467fff108(…): 100%
249M/249M [00:02<00:00, 244MB/s]
data/train-00001-of-00004-5852b56a2bd28f(…): 100%
248M/248M [00:01<00:00, 147MB/s]
data/train-00002-of-00004-a26307300439e9(…): 100%
246M/246M [00:01<00:00, 172MB/s]
data/train-00003-of-00004-d243063613e5a0(…): 100%
248M/248M [00:01<00:00, 80.6MB/s]
data/validation-00000-of-00001-869c898b5(…): 100%
9.99M/9.99M [00:00<00:00, 11.4MB/s]
Generating train split: 100%
2119719/2119719 [00:07<00:00, 349938.66 examples/s]
Generating validation split: 100%
21990/21990 [00:00<00:00, 296057.11 examples/s]
Custom tokenizer loaded with vocab size: 8192

Tokenizing dataset...

Map (num_proc=12): 100%
20000/20000 [00:04<00:00, 5678.89 examples/s]
Map (num_proc=12): 100%
2000/2000 [00:00<00:00, 3177.37 examples/s]
--- Part 3: Model Definition ---
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}

Compiling the model for faster training...

```
/tmp/ipython-input-1446167451.py:99: FutureWarning: `torch.cuda.amp.GradScaler(args...)` is deprecated. Please use `torch.amp.GradScaler('cuda', args...)` instead.
  scaler = torch.cuda.amp.GradScaler()

--- Epoch 1/4 ---

Training:   0%|          | 0/625 [00:00<?, ?it/s]/tmp/ipython-input-1446167451.py:105: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Training:   0%|          | 1/625 [00:44<7:42:21, 44.46s/it, loss=9.2284]/tmp/ipython-input-1446167451.py:105: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Training:   1%|          | 6/625 [00:50<37:00,  3.59s/it, loss=6.8192]/tmp/ipython-input-1446167451.py:105: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Evaluating:   0%|          | 0/63 [00:00<?, ?it/s]/tmp/ipython-input-1446167451.py:133: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Evaluating:   2%|▏         | 1/63 [00:07<07:15,  7.02s/it]/tmp/ipython-input-1446167451.py:133: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Evaluating:  51%|█████     | 32/63 [00:21<00:16, 1.90it/s]/tmp/ipython-input-1446167451.py:133: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():

Epoch 1: Train=3.8064, Val=3.1736, PPL=23.89
 Middle Layer: Gini=0.722, Dead Experts Respawned (Last Step)=0

/tmp/ipython-input-1446167451.py:99: FutureWarning: `torch.cuda.amp.GradScaler(args...)` is deprecated. Please use `torch.amp.GradScaler('cuda', args...)` instead.
  scaler = torch.cuda.amp.GradScaler()

--- Epoch 2/4 ---
```

Training: 0%| | 0/625 [00:00<?, ?it/s]/tmp/ipython-input-1446167451.py:105: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Evaluating: 0%| | 0/63 [00:00<?, ?it/s]/tmp/ipython-input-1446167451.py:133: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():

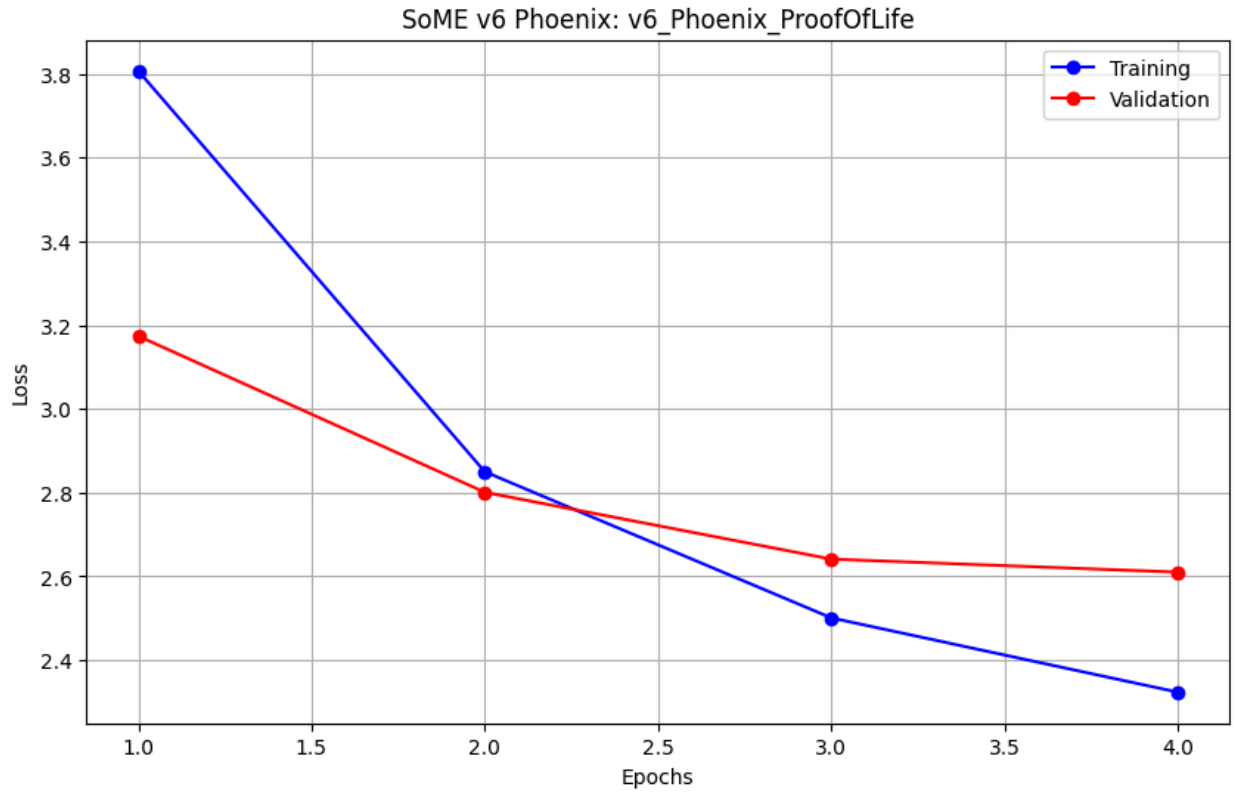
Epoch 2: Train=2.8499, Val=2.8009, PPL=16.46
 Middle Layer: Gini=0.708, Dead Experts Respawned (Last Step)=0

--- Epoch 3/4 ---

Epoch 3: Train=2.5011, Val=2.6414, PPL=14.03
 Middle Layer: Gini=0.710, Dead Experts Respawned (Last Step)=0

--- Epoch 4/4 ---

Epoch 4: Train=2.3238, Val=2.6102, PPL=13.60
 Middle Layer: Gini=0.710, Dead Experts Respawned (Last Step)=0



```
Loading tokenizer from disk...
Tokenizer loaded. Vocab size: 8192

Loading best v6 Phoenix model...
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
SoME Layer Ablation Flags: {'use_alpha': True, 'use_beta': True, 'use_delta': True}
Model loaded successfully.

--- Generating from prompt: 'Once upon a time, there was a little bird who' ---

Once upon a time, there was a little bird wholovedtosoar
/tmp/ipython-input-3806629053.py:41: FutureWarning: `torch.cuda.amp.autocast(args...)` is
deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.no_grad(), torch.cuda.amp.autocast():
/tmp/ipython-input-3085113642.py:109: FutureWarning: `torch.cuda.amp.autocast(args...)` is
deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast(enabled=False):

highatthesky.Oneday,thecraneflewtothestationtoseeabeautifulflowerinthesky.Theskygrewwide

--- Expert Trace (Middle Layer) ---
Tracking active experts for the last 30 tokens generated:
Token: 'loved      ' | Experts: [126, 101, 35, 209]
Token: 'to         ' | Experts: [27, 211, 212, 136]
Token: 'soar       ' | Experts: [27, 211, 212, 136]
Token: 'high       ' | Experts: [205, 71, 200, 194]
Token: 'at         ' | Experts: [134, 253, 151, 99]
Token: 'the        ' | Experts: [136, 27, 211, 212]
Token: 'sky        ' | Experts: [191, 61, 108, 231]
Token: '.          ' | Experts: [186, 118, 91, 253]
Token: 'One        ' | Experts: [41, 37, 102, 210]
Token: 'day        ' | Experts: [226, 144, 21, 24]
Token: ',          ' | Experts: [229, 214, 125, 69]
Token: 'the        ' | Experts: [157, 53, 84, 128]
Token: 'crane      ' | Experts: [191, 61, 231, 103]
Token: 'flew       ' | Experts: [247, 14, 129, 10]
Token: 'to         ' | Experts: [35, 101, 126, 84]
Token: 'the        ' | Experts: [27, 211, 212, 136]
```

Token: 'station ' | Experts: [191, 61, 231, 108]
Token: 'to ' | Experts: [35, 101, 126, 84]
Token: 'see ' | Experts: [27, 211, 212, 136]
Token: 'a ' | Experts: [96, 44, 128, 254]
Token: 'beautiful ' | Experts: [39, 109, 9, 251]
Token: 'flower ' | Experts: [107, 90, 239, 166]
Token: 'in ' | Experts: [90, 239, 122, 129]
Token: 'the ' | Experts: [251, 109, 39, 9]
Token: 'sky ' | Experts: [191, 61, 108, 231]
Token: '. ' | Experts: [231, 108, 107, 61]
Token: 'The ' | Experts: [41, 37, 102, 210]
Token: 'sky ' | Experts: [26, 103, 164, 191]
Token: 'grew ' | Experts: [231, 108, 61, 191]
Token: 'wide ' | Experts: [136, 27, 211, 212]

--- Generating from prompt: 'The king wanted to verify that' ---

The king wanted to verify
thatcouldtheworld.Shehadbeenveryfriendlyandknewthatshecouldbeignorehowtotakesomeonetos
chool.Sheaskedhimifshecouldbe

--- Expert Trace (Middle Layer) ---
Tracking active experts for the last 30 tokens generated:
Token: 'could ' | Experts: [12, 251, 109, 9]
Token: 'the ' | Experts: [136, 27, 211, 212]
Token: 'world ' | Experts: [231, 61, 108, 191]
Token: '. ' | Experts: [39, 9, 109, 251]
Token: 'She ' | Experts: [41, 37, 102, 210]
Token: 'had ' | Experts: [97, 60, 28, 42]
Token: 'been ' | Experts: [178, 116, 109, 39]
Token: 'very ' | Experts: [184, 80, 91, 118]
Token: 'friendly ' | Experts: [35, 101, 126, 84]
Token: 'and ' | Experts: [107, 90, 239, 231]
Token: 'knew ' | Experts: [209, 183, 110, 165]
Token: 'that ' | Experts: [109, 39, 251, 9]
Token: 'she ' | Experts: [35, 101, 126, 84]
Token: 'could ' | Experts: [155, 48, 127, 227]
Token: 'be ' | Experts: [212, 211, 27, 136]
Token: 'ignore ' | Experts: [82, 145, 29, 222]
Token: 'how ' | Experts: [136, 27, 211, 212]
Token: 'to ' | Experts: [35, 101, 126, 84]
Token: 'take ' | Experts: [27, 211, 212, 136]
Token: 'someone ' | Experts: [136, 27, 211, 39]

Token: 'to ' | Experts: [90, 239, 122, 107]
Token: 'school ' | Experts: [27, 211, 212, 136]
Token: '. ' | Experts: [129, 247, 122, 14]
Token: 'She ' | Experts: [41, 102, 37, 210]
Token: 'asked ' | Experts: [60, 97, 42, 28]
Token: 'him ' | Experts: [134, 151, 99, 64]
Token: 'if ' | Experts: [39, 109, 136, 9]
Token: 'she ' | Experts: [35, 101, 84, 126]
Token: 'could ' | Experts: [127, 207, 206, 8]
Token: 'be ' | Experts: [212, 211, 27, 128]

--- Generating from prompt: 'Lily found a secret map under her' ---

Lily found a secret map under herdolltoyinthecastle.Shelikedtorunandplayinthetent.Oneday,shefoundapennyontheground.She wantedto

--- Expert Trace (Middle Layer) ---
Tracking active experts for the last 30 tokens generated:
Token: 'doll ' | Experts: [217, 170, 166, 242]
Token: 'toy ' | Experts: [184, 80, 91, 118]
Token: 'in ' | Experts: [166, 170, 80, 217]
Token: 'the ' | Experts: [251, 9, 109, 39]
Token: 'castle ' | Experts: [231, 61, 108, 107]
Token: '. ' | Experts: [186, 253, 118, 91]
Token: 'She ' | Experts: [41, 37, 102, 210]
Token: 'liked ' | Experts: [28, 97, 60, 42]
Token: 'to ' | Experts: [35, 101, 136, 126]
Token: 'run ' | Experts: [27, 211, 136, 212]
Token: 'and ' | Experts: [109, 39, 35, 9]
Token: 'play ' | Experts: [126, 101, 84, 35]
Token: 'in ' | Experts: [99, 93, 151, 64]
Token: 'the ' | Experts: [109, 39, 9, 251]
Token: 'tent ' | Experts: [231, 108, 61, 191]
Token: '. ' | Experts: [107, 90, 239, 122]
Token: 'One ' | Experts: [41, 37, 102, 210]
Token: 'day ' | Experts: [144, 226, 21, 24]
Token: ', ' | Experts: [229, 214, 125, 69]
Token: 'she ' | Experts: [157, 53, 128, 212]
Token: 'found ' | Experts: [237, 246, 245, 115]
Token: 'a ' | Experts: [136, 27, 211, 212]
Token: 'penny ' | Experts: [39, 109, 9, 251]
Token: 'on ' | Experts: [242, 48, 227, 155]

Token: 'the ' | Experts: [109, 251, 39, 9]
Token: 'ground ' | Experts: [108, 231, 61, 191]
Token: '. ' | Experts: [136, 27, 211, 212]
Token: 'She ' | Experts: [41, 102, 37, 210]
Token: 'wanted ' | Experts: [28, 97, 60, 42]
Token: 'to ' | Experts: [27, 136, 211, 212]
