SoME v7 (The Clamp Fix) results

--- Part 1: Setup and Dependencies ---
Using device: cuda
A100 GPU detected. Enabling TF32.

--- Starting Experiment: v7_ClampFix_Phoenix ---

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

/usr/local/lib/python3.12/dist-packages/torch/__init__.py:1617: UserWarning: Please use the new API settings to control TF32 behavior, such as torch.backends.cudnn.conv.fp32_precision = 'tf32' or torch.backends.cuda.matmul.fp32_precision = 'ieee'. Old settings, e.g, torch.backends.cuda.matmul.allow_tf32 = True, torch.backends.cudnn.allow_tf32 = True, allowTF32CuDNN() and allowTF32CuBLAS() will be deprecated after Pytorch 2.9. Please see https://pytorch.org/docs/main/notes/cuda.html#tensorfloat-32-tf32-on-ampere-and-later-devices
(Triggered internally at /pytorch/aten/src/ATen/Context.cpp:80.)
  _C._set_float32_matmul_precision(precision)

README.md:
  1.06k/? [00:00<00:00, 118kB/s]
data/train-00000-of-00004-2d5a1467fff108(…): 100%
  249M/249M [00:02<00:00, 205MB/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, 134MB/s]
data/train-00003-of-00004-d243063613e5a0(…): 100%
  248M/248M [00:02<00:00, 137MB/s]
data/validation-00000-of-00001-869c898b5(…): 100%
  9.99M/9.99M [00:01<00:00, 9.75MB/s]
Generating train split: 100%
  2119719/2119719 [00:06<00:00, 355838.10 examples/s]
Generating validation split: 100%
  21990/21990 [00:00<00:00, 315173.69 examples/s]
Custom tokenizer loaded with vocab size: 8192

Tokenizing dataset...

Map (num_proc=12): 100%
  20000/20000 [00:04<00:00, 4400.21 examples/s]
Map (num_proc=12): 100%
  2000/2000 [00:00<00:00, 3718.35 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}

/tmp/ipython-input-3195204426.py:444: FutureWarning: `torch.cuda.amp.GradScaler(args...)` is
deprecated. Please use `torch.amp.GradScaler('cuda', args...)` instead.
  scaler = torch.cuda.amp.GradScaler()

--- Part 4: Training ---

Training:   0%|          | 0/625 [00:00<?, ?it/s]/tmp/ipython-input-3195204426.py:450:
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 [00:45<7:49:23, 45.13s/it,
loss=9.2031]/tmp/ipython-input-3195204426.py:450: FutureWarning:
`torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda',
args...)` instead.
  with torch.cuda.amp.autocast():
Training:   0%|          | 3/625 [00:48<1:59:20, 11.51s/it,
loss=7.8453]/tmp/ipython-input-3195204426.py:450: FutureWarning:
`torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda',
args...)` instead.
  with torch.cuda.amp.autocast():
Eval:   0%|          | 0/63 [00:00<?, ?it/s]/tmp/ipython-input-3195204426.py:476: FutureWarning:
`torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda',
args...)` instead.
  with torch.cuda.amp.autocast():
Eval:   2%|▏         | 1/63 [00:06<06:56,  6.72s/it]/tmp/ipython-input-3195204426.py:476:
FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use
`torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Eval:   3%|▎         | 2/63 [00:07<03:15,  3.20s/it]/tmp/ipython-input-3195204426.py:476:
FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use
`torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():

Epoch 1: Train=3.7176, Val=3.1531, PPL=23.41
```

```
 Middle Layer: Gini=0.826, Phoenix Respawns (Last Step)=2

/tmp/ipython-input-3195204426.py:444: FutureWarning: `torch.cuda.amp.GradScaler(args...)` is deprecated. Please use `torch.amp.GradScaler('cuda', args...)` instead.
  scaler = torch.cuda.amp.GradScaler()
Training:   0%|          | 0/625 [00:00<?, ?it/s]/tmp/ipython-input-3195204426.py:450: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():
Eval:   0%|          | 0/63 [00:00<?, ?it/s]/tmp/ipython-input-3195204426.py:476: FutureWarning: `torch.cuda.amp.autocast(args...)` is deprecated. Please use `torch.amp.autocast('cuda', args...)` instead.
  with torch.cuda.amp.autocast():

Epoch 2: Train=2.8239, Val=2.7850, PPL=16.20
 Middle Layer: Gini=0.767, Phoenix Respawns (Last Step)=0

Epoch 3: Train=2.4870, Val=2.6306, PPL=13.88
 Middle Layer: Gini=0.724, Phoenix Respawns (Last Step)=0

Epoch 4: Train=2.3121, Val=2.6036, PPL=13.51
 Middle Layer: Gini=0.653, Phoenix Respawns (Last Step)=0
```