SoME ablation runs v1

V1: Sanity Check

- Model Dimensions:
  - D_MODEL=256,
  - NUM_HEADS=4,
  - NUM_LAYERS=4
- SoME Parameters:
  - num_experts=32,
  - d_ffn=1024,
  - top_k=2
- Training Schedule:
  - 4 Epochs,
  - LR=8e-4
- Data Configuration:
  - SEQ_LEN=256,
  - BATCH_SIZE=128,
  - VOCAB_SIZE=8192,
  - train_subset_size=20000
  - Dataset = roneneldan/TinyStories
- Model & Training Scale:
  - Total Parameters: 72.80M
  - Trainable Parameters: 5.52M (7.6% of total)
  - Total Training Steps: 624
- Expert Initialization:
  - init_method="default”
- Results:
  - Gini Coefficient: Showed a consistent decrease from 0.631 (Epoch 1) to 0.550 (Epoch 4).
  - Shannon Entropy: Showed a consistent increase from 3.887 (Epoch 1) to 4.192 (Epoch 4).
  - Training Loss: Decreased steadily from 3.6590 at the end of Epoch 1 to 2.5631 at the end of Epoch 4.
  - Validation Loss: Decreased steadily from 2.8422 at the end of Epoch 1 to 2.5555 at the end of Epoch 4.
  - Validation Perplexity: Showed significant improvement, reducing from 17.15 to 12.88.

V2: Medium Baseline

- Model Dimensions:
  - D_MODEL=384,
  - NUM_HEADS=6,
  - NUM_LAYERS=6

- SoME Parameters:
    - num_experts=64,
    - d_ffn=1024,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=256,
    - BATCH_SIZE=128,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 313.27M
    - Trainable Parameters: 10.74M (3.4% of total)
    - Total Training Steps: 624
- Expert Initialization:
    - init_method="default”
- Results:
    - Gini Coefficient: Showed a slight increase/plateau, moving from 0.518 to 0.524.
    - Shannon Entropy: Showed a slight decrease/plateau, moving from 5.315 to 5.309.
    - Training Loss: Decreased steadily from 3.3316 at the end of Epoch 1 to 2.2398 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 2.6379 at the end of Epoch 1 to 2.2681 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 13.98 to 9.66.

V3: Capacity Scaling

- Model Dimensions:
    - D_MODEL=512,
    - NUM_HEADS=8,
    - NUM_LAYERS=8
- SoME Parameters:
    - num_experts=256,
    - d_ffn=1536,
    - top_k=8
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:

    - SEQ_LEN=256,
    - BATCH_SIZE=128,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 3244.34M (3.24 Billion)
    - Trainable Parameters: 18.92M (0.58% of total)
    - Total Training Steps: 624
- Expert Initialization:
    - init_method="default”
- Results:
    - Gini Coefficient: Showed a consistent decrease, from 0.590 to 0.539.
    - Shannon Entropy: Showed a consistent increase, from 6.994 to 7.134.
    - Training Loss: Decreased steadily from 3.1599 at the end of Epoch 1 to 1.9782 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 2.4792 at the end of Epoch 1 to 2.0542 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 11.93 to 7.80.

V4: Vocab Scaling

- Model Dimensions:
    - D_MODEL=384,
    - NUM_HEADS=6,
    - NUM_LAYERS=6
- SoME Parameters:
    - num_experts=64,
    - d_ffn=1024,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=256,
    - BATCH_SIZE=128,
    - VOCAB_SIZE=16384,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 319.57M
    - Trainable Parameters: 17.04M (5.3% of total)
    - Total Training Steps: 624

- Expert Initialization:
    - init_method="default”
- Results:
    - Gini Coefficient: Showed a slight increase/plateau, moving from 0.466 to 0.479.
    - Shannon Entropy: Showed a slight decrease/plateau, moving from 5.441 to 5.410.
    - Training Loss: Decreased steadily from 3.3803 at the end of Epoch 1 to 2.2190 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 2.6266 at the end of Epoch 1 to 2.2484 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 13.83 to 9.47.

V5: Diversity Scaling

- Model Dimensions:
    - D_MODEL=384,
    - NUM_HEADS=6,
    - NUM_LAYERS=6
- SoME Parameters:
    - num_experts=256,
    - d_ffn=1536,
    - top_k=16
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=256,
    - BATCH_SIZE=128,
    - VOCAB_SIZE=16384,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 1825.63M (1.83 Billion)
    - Trainable Parameters: 10.74M (0.59% of total)
    - Total Training Steps: 624
- Expert Initialization:
    - init_method="default”
- Results:
    - Gini Coefficient: Exhibited a steep decrease, from 0.402 to 0.379.
    - Shannon Entropy: Exhibited a clear increase, from 7.534 to 7.569.
    - Training Loss: Decreased steadily from 3.3514 at the end of Epoch 1 to 2.2248 at the end of Epoch 4.

    - Validation Loss: Decreased steadily from 2.6459 at the end of Epoch 1 to 2.2565 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 14.10 to 9.55.

V6: Depth vs. Width

- Model Dimensions:
    - D_MODEL=512,
    - NUM_HEADS=8,
    - NUM_LAYERS=10
- SoME Parameters:
    - num_experts=64,
    - d_ffn=1024,
    - top_k=8
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=256,
    - BATCH_SIZE=128,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 693.62M
    - Trainable Parameters: 21.55M (3.1% of total)
    - Total Training Steps: 624
- Expert Initialization:
    - init_method="default”
- Results:
    - The Gini coefficient decreased consistently, from 0.399 down to 0.383.
    - The Shannon Entropy increased consistently, from 5.592 up to 5.635.
    - Training Loss: Decreased steadily from 3.3282 at the end of Epoch 1 to 1.9726 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 2.5286 at the end of Epoch 1 to 2.0471 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 12.54 to 7.75.

V7: Batch Scaling v Sequence Scaling v1

- Model Dimensions:
    - D_MODEL=384,
    - NUM_HEADS=6,

    - NUM_LAYERS=6
- SoME Parameters:
    - num_experts=64,
    - d_ffn=1024,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=512,
    - BATCH_SIZE=64,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 313.27M
    - Trainable Parameters: 10.74M (3.4% of total)
    - Total Training Steps: 1248
- Expert Initialization:
    - init_method="default”
- Results:
    - Gini Coefficient: Peaked at 0.676 before beginning to decrease to 0.661.
    - Shannon Entropy: Bottomed at 4.417 before beginning to increase to 4.454.
    - Training Loss: Decreased steadily from 1.7605 at the end of Epoch 1 to 1.1467 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 1.3095 at the end of Epoch 1 to 1.1134 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 3.70 to 3.04.

V8: Batch Scaling v Sequence Scaling v2

- Model Dimensions:
    - D_MODEL=384,
    - NUM_HEADS=6,
    - NUM_LAYERS=6
- SoME Parameters:
    - num_experts=64,
    - d_ffn=1024,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=512,

    - BATCH_SIZE=128,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 313.27M
    - Trainable Parameters: 10.74M (3.4% of total)
    - Total Training Steps: 624
- Expert Initialization:
    - init_method="default”
- Results:
    - The Gini coefficient showed a clear decrease, from 0.729 down to 0.693.
    - The Shannon Entropy showed a clear increase, from 4.293 up to 4.377.
    - Training Loss: Decreased steadily from 2.0406 at the end of Epoch 1 to 1.2727 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 1.4250 at the end of Epoch 1 to 1.2117 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 4.16 to 3.36.
- Findings:
    - In standard training, a larger batch size is often better because it provides a more stable and accurate estimate of the gradient, leading to smoother convergence.
    - In SoME training, something different is happening. The learning in the SoME layers is not gradient-based; it's a heuristic, iterative process. The update_keys function provides a small "nudge" to the expert keys after every single step.

V9: Default v Orthogonal

- Model Dimensions:
    - D_MODEL=384,
    - NUM_HEADS=6,
    - NUM_LAYERS=6
- SoME Parameters:
    - num_experts=64,
    - d_ffn=1024,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4
- Data Configuration:
    - SEQ_LEN=512,
    - BATCH_SIZE=128,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories

- Model & Training Scale:
    - Total Parameters: 313.27M
    - Trainable Parameters: 10.74M (3.4% of total)
    - Total Training Steps: 624
- Expert Initialization:
    - init_method="orthogonal”
- Results:
    - The Gini coefficient for the orthogonal run started extremely high at 0.764 and remained high, finishing at 0.744.
    - The Shannon Entropy started extremely low at 4.154 and struggled to increase, finishing at 4.245.
    - Training Loss: Decreased steadily from 2.0241 at the end of Epoch 1 to 1.2778 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 1.4228 at the end of Epoch 1 to 1.2173 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 4.15 to 3.38.
- Findings:
    - The final validation perplexity of 3.38 is slightly but consistently worse than the default run's 3.36. This confirms that default initialization provides a performance advantage.
    - The orthogonal initialization created a functionally impoverished expert library. The router was unable to find a diverse set of useful primitives and was forced into an extreme concentration strategy, over-relying on a very small subset of its experts. This is the quantitative signature of a failed diversification process.
    - This confirms our original hypothesis, Even when all other conditions are ideal, imposing mathematical structure on the initial expert weights is detrimental to performance. The chaotic, unstructured diversity of default initialization provides a richer "primordial soup" of functions, which allows the self-organizing mechanism to find better solutions.

V10: Default v Sparse

- Model Dimensions:
    - D_MODEL=384,
    - NUM_HEADS=6,
    - NUM_LAYERS=6
- SoME Parameters:
    - num_experts=64,
    - d_ffn=1024,
    - top_k=4
- Training Schedule:
    - 4 Epochs,
    - LR=8e-4

- Data Configuration:
    - SEQ_LEN=512,
    - BATCH_SIZE=128,
    - VOCAB_SIZE=8192,
    - train_subset_size=20000
    - Dataset = roneneldan/TinyStories
- Model & Training Scale:
    - Total Parameters: 313.27M
    - Trainable Parameters: 10.74M (3.4% of total)
    - Total Training Steps: 624
- Expert Initialization:
    - init_method="sparse”
- Results:
    - The Gini coefficient started high at 0.709 and remained high, finishing at 0.710.
    - The Shannon Entropy started low at 4.344 and remained low, finishing at 4.341.
    - Training Loss: Decreased steadily from 1.9960 at the end of Epoch 1 to 1.2650 at the end of Epoch 4.
    - Validation Loss: Decreased steadily from 1.4150 at the end of Epoch 1 to 1.2054 at the end of Epoch 4.
    - Validation Perplexity: Showed significant improvement, reducing from 4.12 to 3.34.
- Findings:
    - The final validation perplexity of 3.34 is the best in this class, outperforming both default (3.36) and orthogonal (3.38). This suggests that a diversity of distinct primitives is the ideal condition for SoME.
    - This supports the "Distinct Primitives" Hypothesis. Sparse initialization creates a library of highly specialized experts. The router can achieve optimal performance by learning to select the single best specialist for a task, rather than composing a solution from multiple generalists. This leads to high concentration (high Gini) because the system relies on its most effective tools, but it also leads to the best perplexity because those tools are extremely well-suited for their niche.

Based on all the evidence, we can state without a shadow of a doubt what is going on:

SoME is a learning paradigm that reframes optimization as a search and discovery problem over a static computational landscape. Its performance is governed by a clear hierarchy of principles:

- The Landscape's Richness: The system's potential is fundamentally determined by the quality of its expert library. The ideal library is not just large but composed of diverse and functionally distinct primitives, a state best achieved with sparse initialization.

- The Explorer's Vision: The system's ability to navigate this landscape is gated by its context window (SEQ_LEN). A long context is the single most critical factor, as it provides the router with the necessary information to make meaningful decisions.
- The Explorer's Intelligence: The sophistication of the search process itself is a function of the model's depth (NUM_LAYERS). A deeper model is a more "intelligent" router, capable of better orchestrating and composing the primitives it finds.
- The Explorer's Pace: The learning process is iterative and heuristic. It thrives on a high frequency of small updates, making smaller batch sizes and a larger number of training steps the optimal training strategy.