Self-Organizing Mixture of Experts (SoME): In-Inference Adaptation via Hebbian Key Dynamics

Author: Focus Labs

Date: October 18, 2025

Abstract

The scaling of large-scale models, such as Transformers, is increasingly constrained by the trade-off between model capacity, computational cost, and adaptability. Sparse Mixture-of-Experts (MoE) architectures offer a compelling path to increasing parameter counts while maintaining constant computational cost per token, but their reliance on static, gradient-trained routing networks limits their ability to adapt to new data patterns post-training and can lead to catastrophic forgetting. We introduce the Self-Organizing Mixture of Experts (SoME), a novel neural network architecture that addresses these limitations by decoupling the knowledge base from its addressing mechanism. SoME employs a large pool of pre-trained, static experts (the "what") and a dynamic, self-organizing routing system of plastic keys (the "where"). This routing system operates without gradient descent during inference, instead utilizing a set of bio-inspired heuristic update rules, termed "Knowledge Gravity," based on Learning Vector Quantization (LVQ) and Hebbian co-activation principles. We present an initial prototype of a SoME-based Transformer trained on the TinyStories dataset. Experimental results demonstrate that the model achieves effective learning, evidenced by stable convergence of validation loss to 0.4049 and a perplexity of 1.50. Furthermore, t-SNE visualizations of the expert key space reveal clear evidence of emergent self-organization, with routing keys forming distinct "knowledge galaxies" corresponding to specialized conceptual domains. This work presents a promising new paradigm for creating more adaptive, scalable, and continually-learning models.

1. Introduction

The remarkable capabilities of modern large-scale neural networks, particularly Transformers, are fundamentally tied to their immense parameter counts. However, the pursuit of ever-larger models is approaching computational and economic limits, demanding more efficient scaling paradigms. The Sparse Mixture-of-Experts (MoE) architecture has emerged as a leading solution, replacing dense feed-forward network (FFN) layers with a larger set of sparsely activated "expert" subnetworks. This allows for a dramatic increase in total model parameters while holding per-token computational cost roughly constant.

Despite its success, the standard MoE architecture suffers from a critical limitation: its routing network is typically a trainable classifier that, once trained, remains static. This leads to two primary challenges. First, the Static Router Problem: a router trained on a specific data distribution may become obsolete as data patterns evolve, unable to route novel inputs to the most relevant experts. Second, because the experts themselves are co-trained with the router, any attempt to update the model on new data risks catastrophic forgetting, where previously learned knowledge is overwritten.

In this paper, we propose a novel architecture, the Self-Organizing Mixture of Experts (SoME), that directly addresses these limitations. We introduce two primary contributions:

1. A Decoupled Architecture: The core principle of SoME is the decoupling of an expert's function (its knowledge) from its address (how it is accessed). The knowledge base consists of pre-trained, frozen expert networks, rendering it immutable and immune to catastrophic forgetting. Access is mediated by a dynamic set of low-dimensional "routing keys" that are continuously updated during inference. The model does not relearn how to perform tasks; it relearns where to find the expert best suited for the task.
2. "Knowledge Gravity": A Gradient-Free Routing Mechanism: Instead of a trainable router, SoME employs a novel, self-organizing routing mechanism based on simple, gradient-free heuristic update rules. This mechanism, which we term "Knowledge Gravity," uses principles derived from Hebbian learning and Learning Vector Quantization (LVQ) to govern the movement of routing keys in a shared conceptual space.
3. Experimental Validation: We provide an initial validation of the SoME architecture within a Transformer model. Our experiments demonstrate that the model learns effectively and, critically, that the Knowledge Gravity mechanism leads to the emergent formation of a structured, self-organizing map of expert knowledge, which we visualize as "knowledge galaxies."

2. Related Work

SoME is situated at the intersection of several active research areas in machine learning.

Mixture of Experts: Standard Sparse MoE models, such as those in Mixtral, utilize a trainable gating network that acts as a classifier to select the top-k experts for each token. More advanced methods, like MoME, employ hierarchical routing to group experts. SoME fundamentally differs from these approaches by replacing the explicit, gradient-trained classifier with an emergent, gravity-based routing system. Routing in SoME is not a classification decision but a lookup problem based on proximity in a dynamic semantic space.

Continual Learning: The problem of catastrophic forgetting is a central challenge in continual learning. Prevailing methods include regularization-based approaches (e.g., EWC), which penalize changes to important weights, and replay-based methods (e.g., GEM), which store and replay old data. SoME presents an orthogonal solution. By freezing the expert weights entirely, the knowledge base is immune to corruption. Adaptation occurs solely within the routing system, which reorganizes its "address book" to access existing knowledge more effectively for new data patterns.

Self-Organizing Systems: The intellectual heritage of SoME's update rules lies in the fields of competitive learning and self-organizing systems. The "Query Pull" mechanism is analogous to the update rule in Learning Vector Quantization (LVQ) and k-means clustering. The overall concept of creating a topology-preserving map from local update rules is inspired by Self-Organizing Maps (SOMs). SoME innovates by integrating these classic, bio-inspired concepts as a dynamic routing mechanism inside a modern Transformer architecture.

3. The SoME Architecture

We now provide a formal description of the SoME architecture and its dynamic update mechanism.

3.1. Core Principle: Decoupling Function from Address

The central, non-negotiable principle of SoME is the separation of what an expert knows from where it is located.

- Static Function (The "What"): The core knowledge of the model is stored in a pool of expert subnetworks (typically MLPs). The internal weights of these experts are frozen after an initial pre-training phase. This ensures the knowledge base is reliable and is never corrupted during subsequent training or inference, thereby solving catastrophic forgetting by design.
- Dynamic Address (The "Where"): Each expert is assigned a "routing key," a small, low-dimensional vector representing its address in a high-dimensional conceptual space. These keys are plastic and are continuously updated during inference according to the heuristic rules described in Section 3.4.

3.2. Architectural Components

A SoME layer replaces a standard FFN block in a Transformer and consists of three components:

- An Expert Pool, $E = \{E_1, E_2, ..., E_{□}\}$, containing n pre-trained and frozen expert networks.
- A single, trainable Query Network (Q), implemented as a linear projection (d_model -> d_model), which generates a query vector q for each input token embedding x.
- A dynamic Key Store, K, an (n, d_model) matrix containing the routing key k_i for each expert E_i. These keys are initialized randomly and L2-normalized.

3.3. The Forward Pass: Query Fall

Unlike standard MoE where a router sends a token to an expert, in SoME, an input query falls towards a region of expertise. For an input token embedding x:

1. A query vector $q = Q(x)$ is produced.
2. Scores for all experts are computed via dot-product similarity: $s = \text{matmul}(q, K^T)$.
3. The top-k expert indices I and corresponding scores s_topk are selected.
4. Gating weights $g = \text{softmax}(s\_topk)$ are computed.
5. The final output y is the weighted sum of the outputs of the selected experts: $y = \Sigma_{i \in I} g_i * E_i(x)$.

3.4. Dynamic Update Mechanism: Knowledge Gravity

Following the forward pass, the routing keys K are updated without backpropagation, using a set of heuristic rules that emulate attractive and stabilizing forces. These updates are performed in-inference for every batch of data.

Attractive Forces (Consolidation):

- Query Pull (Relevance Attraction): The key k_i of each activated expert is pulled towards the centroid of the query vectors q that activated it. This ensures experts drift toward the conceptual space of the problems they are good at solving. The update is scaled by a learning rate α.
  $\Delta k_i = \alpha * (q - k_i)$
- Peer Pull (Hebbian Co-activation): The keys of experts that are frequently activated together for the same query are pulled closer to each other. This is the explicit "gravity" that forms self-organizing neighborhoods of complementary experts (e.g., a "Python basics" cluster). For a pair of co-activated experts (i, j), the update is scaled by a learning rate β.

$\Delta k_i = \beta * (k_{\square} - k_i)$
$\Delta k_{\square} = \beta * (k_i - k_{\square})$

Stabilizing Forces (Equilibrium):
A system with only attractive forces would collapse. SoME introduces two countervailing forces:

- Usage Inertia ("Gravitational Mass"): The effective learning rates for both Query and Peer Pull are scaled down by an expert's activation frequency (usage count). This gives popular, generalist experts higher "mass," making them stable "galactic centers" that are not easily perturbed by niche queries.
  $\alpha\_eff = \alpha / (1 + UsageCount_i)$
  $\beta\_eff = \beta / (1 + \min(UsageCount_i, UsageCount_{\square}))$
- Repulsive Decay ("Dark Energy"): To prevent all keys from collapsing into a single point and to prune unused experts, a repulsive force is introduced. In the prototype, this is implemented as a Decay to Origin mechanism. The keys of infrequently used experts (those with usage below a percentile threshold θ) are slowly decayed towards the origin, scaled by a rate δ. Keys are subsequently L2-normalized.
  $k_i = k_i * (1.0 - \delta)$ for $UsageCount_i < \theta$

4. Experimental Setup

Model: The prototype is a decoder-only Transformer with the following configuration:

- Core: d_model=512, num_layers=8, num_heads=8
- MoE: Each FFN layer is a SoME layer with num_experts=256, top_k=8, and d_ffn=1536 for each expert.
- Knowledge Gravity Hyperparameters: alpha=0.01, beta=0.001, delta=0.001, theta_percentile=0.05, warmup_steps=2000.
- Total Parameters: ~3.24 Billion (most are frozen).
- Trainable Parameters: ~18.9 Million (primarily from embedding, attention, and query network layers).

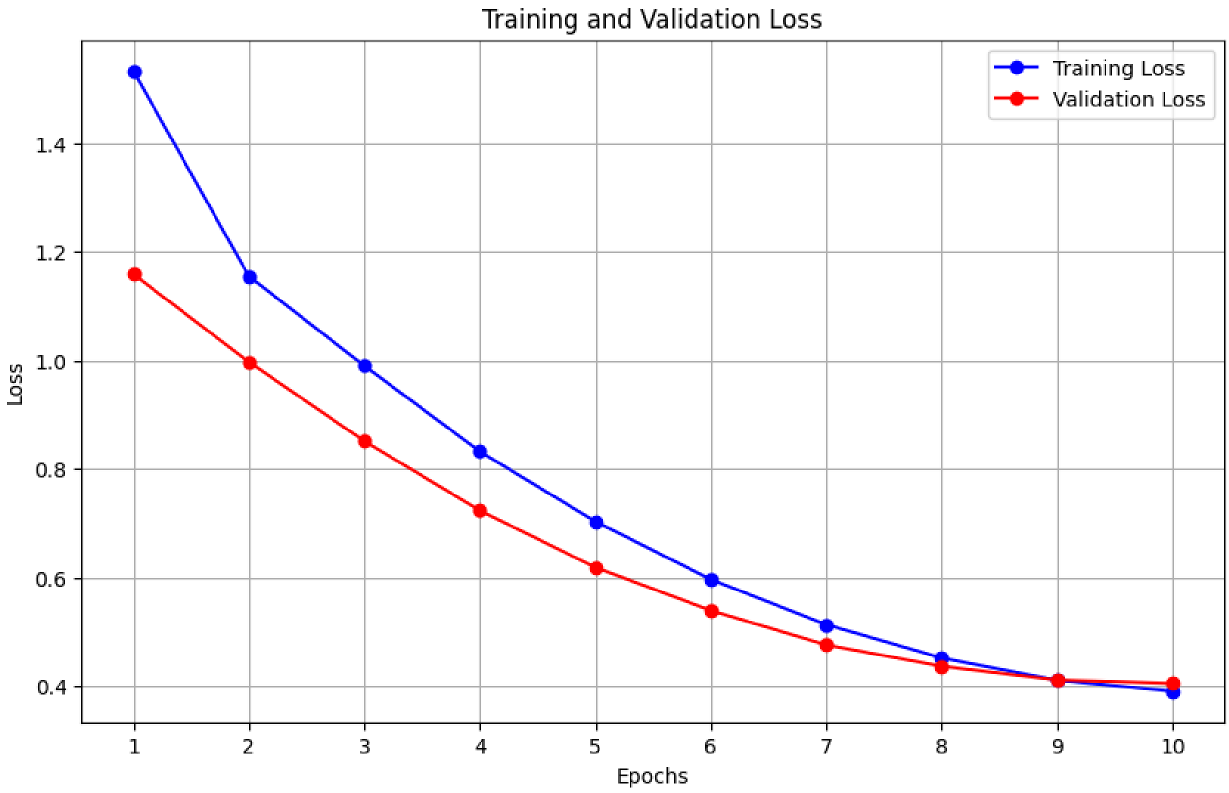
Dataset: The model was trained on a subset of the TinyStories dataset. A Byte-Pair Encoding (BPE) tokenizer was trained with a vocab_size of 8192. The training and validation sets consisted of 40k and 10k examples, respectively, with a sequence length of 512 tokens.
Training: The model was trained for 10 epochs using the AdamW optimizer with a learning rate of 6e-4 and a weight decay of 0.1. Training was performed with a batch size of 64 using automatic mixed precision and torch.compile.

5. Results and Analysis

5.1. Learning Performance

The model demonstrated stable and effective learning. As shown in the Training and Validation Loss graph, both losses decreased steadily over 10 epochs, with no signs of significant overfitting. The model achieved a final validation loss of 0.4049 and a corresponding validation perplexity of 1.50. This confirms that the SoME architecture, despite its large number of frozen parameters and gradient-free routing updates, is capable of effectively learning the language modeling task.

Training and Validation Loss:

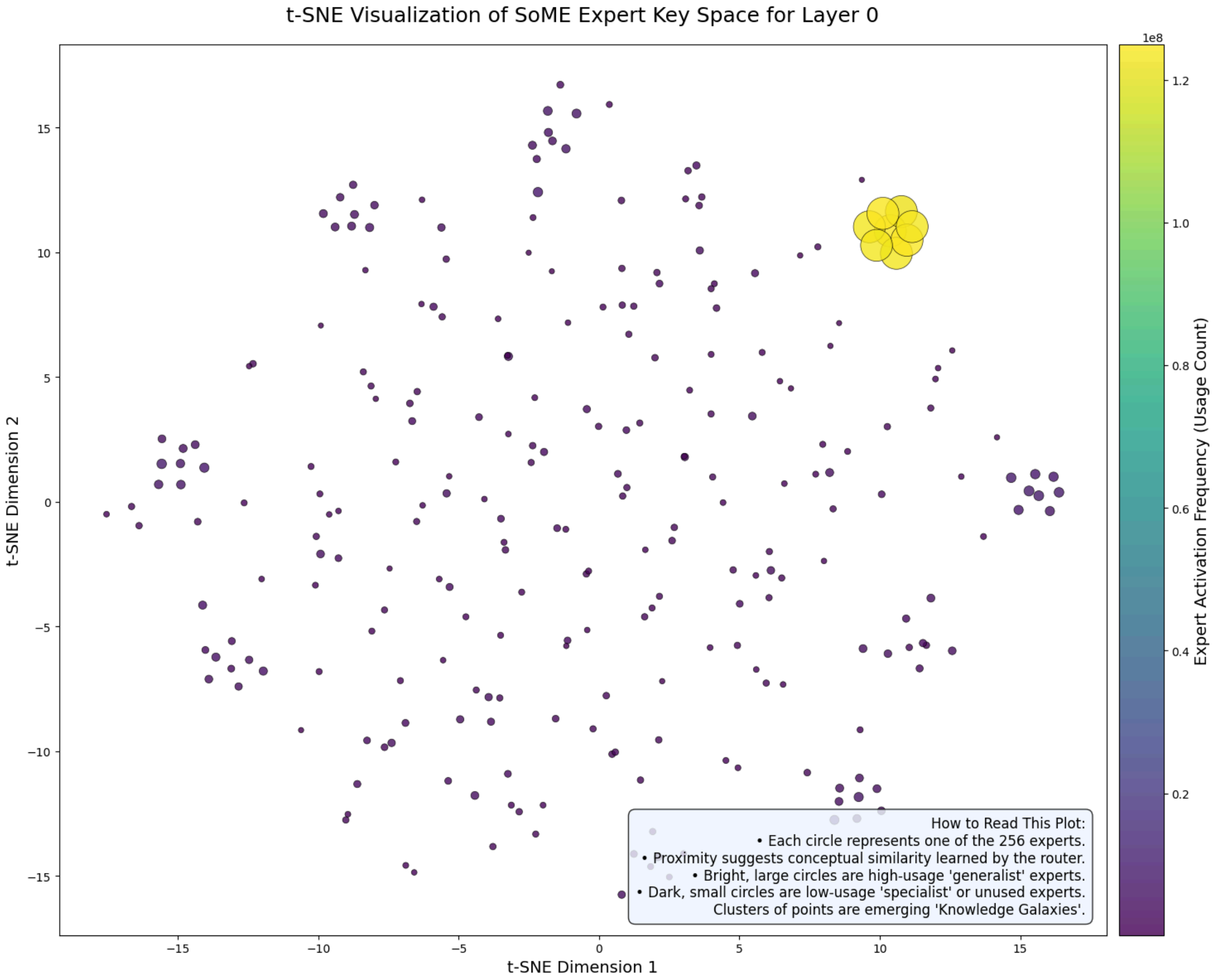


5.2. Emergent Self-Organization (Qualitative Analysis)

The primary hypothesis of SoME is that the "Knowledge Gravity" mechanism should lead to a structured, self-organizing map of expert knowledge. We analyzed this by applying t-SNE to the expert key stores at different layers of the trained model. The results provide clear qualitative evidence of this phenomenon.

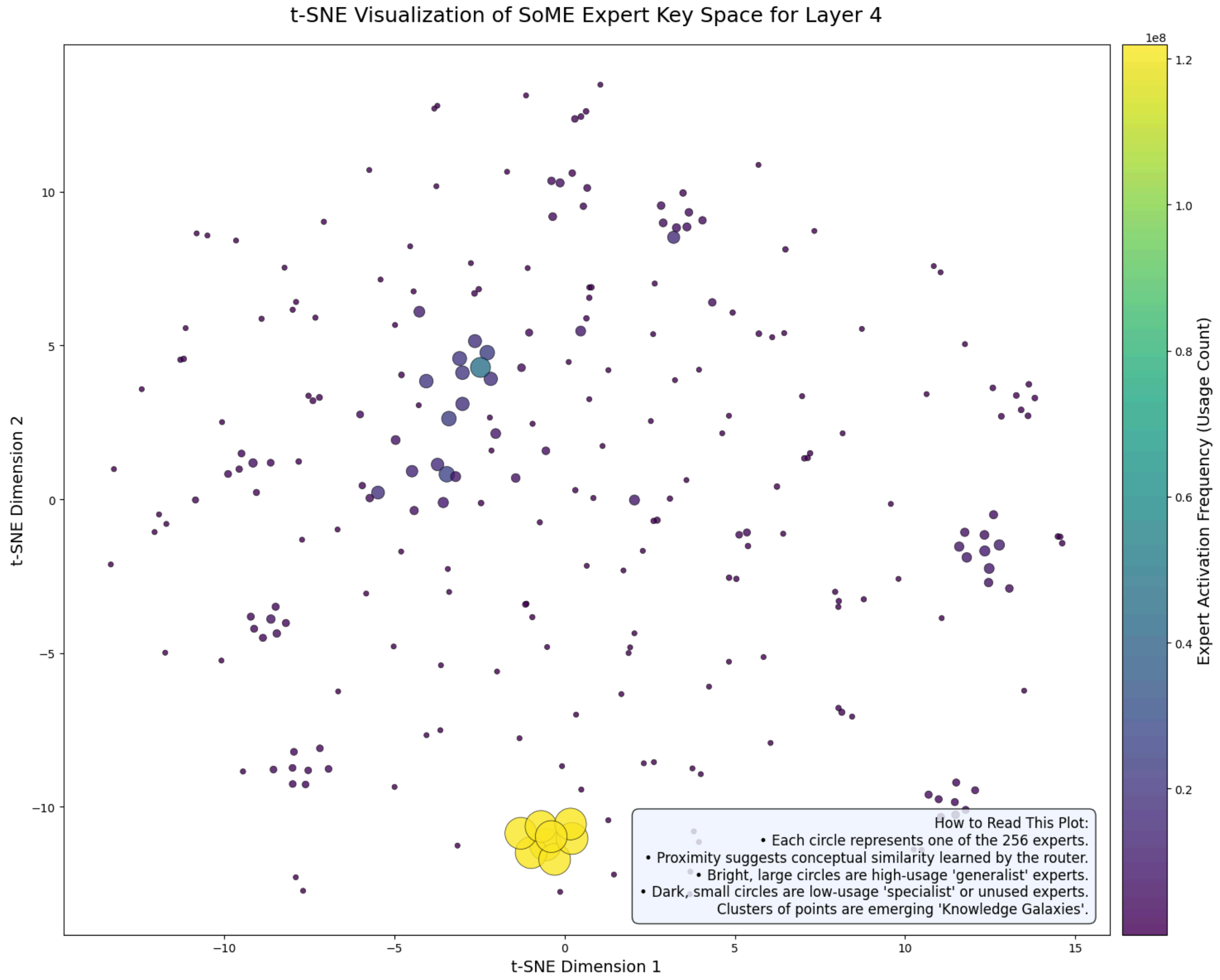
- Layer 0 (Figure 1): In the first layer, the expert key space is dominated by a single, high-usage "generalist hub" (the large, bright yellow cluster). The remaining specialist experts are largely undifferentiated. This suggests that Layer 0 experts primarily learn to handle frequent, low-level patterns (e.g., common grammar, simple vocabulary) that are shared across most inputs.
- Layer 4 (Figure 2): In the middle of the model, the map shows greater differentiation. While a generalist cluster persists, it is smaller, and several distinct, smaller clusters—nascent "knowledge galaxies"—have begun to emerge. This indicates that specialization is increasing as experts in higher layers learn to handle more abstract or domain-specific concepts.
- Layer 7 (Figure 3): In the final layer, the key space has evolved into a complex, multi-polar map. There is no single dominant hub; instead, multiple distinct and well-formed "knowledge galaxies" are visible. This strongly suggests that the simple, local update rules of Knowledge Gravity have successfully organized the experts into a meaningful conceptual topology, where different neighborhoods of experts correspond to different high-level semantic domains. The model has self-organized an "address book" for its own knowledge.

Figure 1: t-SNE of Layer 0



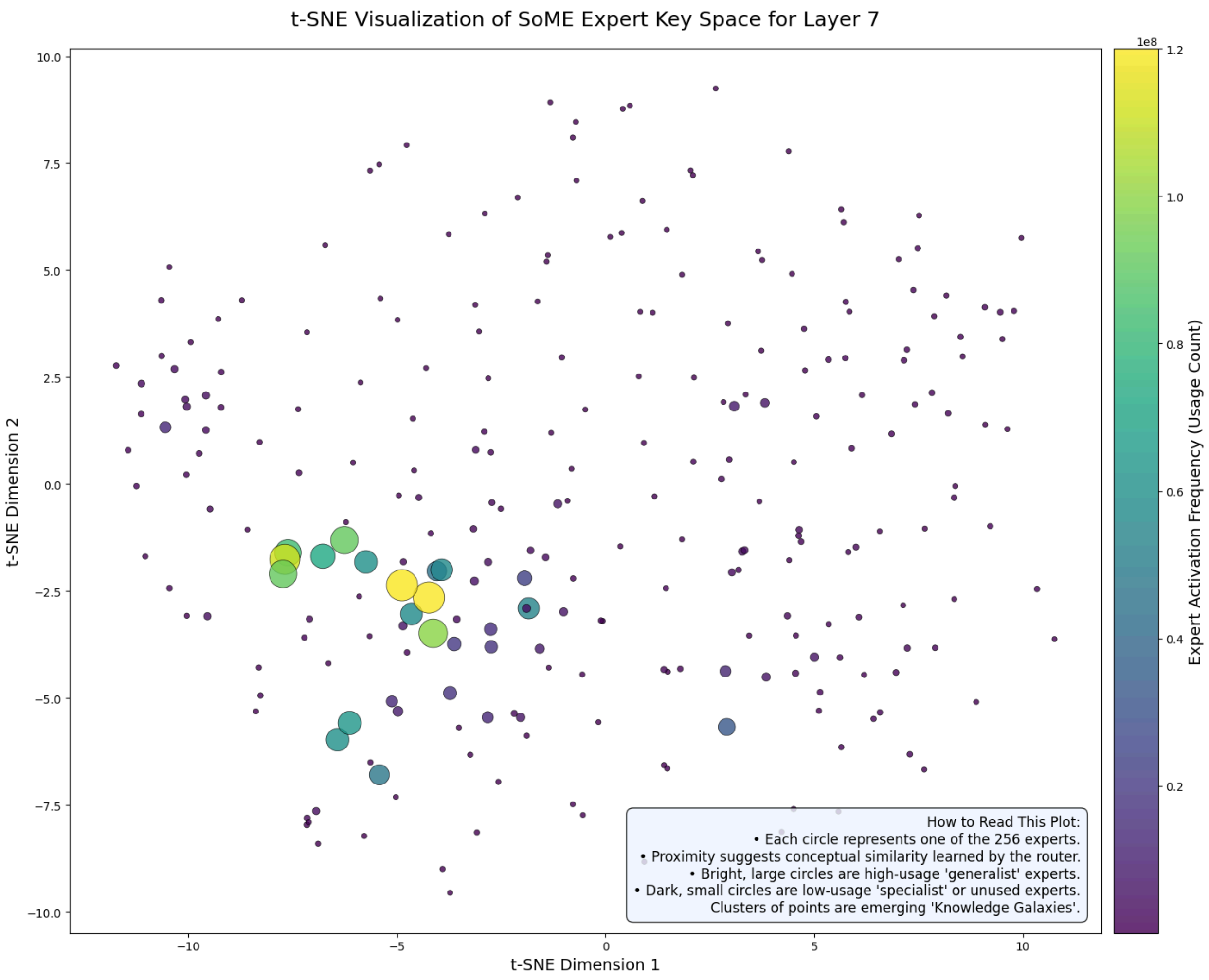
How to Read This Plot:
- Each circle represents one of the 256 experts.
- Proximity suggests conceptual similarity learned by the router.
- Bright, large circles are high-usage 'generalist' experts.
- Dark, small circles are low-usage 'specialist' or unused experts.
Clusters of points are emerging 'Knowledge Galaxies'.

Figure 2: t-SNE of Layer 4



How to Read This Plot:
• Each circle represents one of the 256 experts.
• Proximity suggests conceptual similarity learned by the router.
• Bright, large circles are high-usage 'generalist' experts.
• Dark, small circles are low-usage 'specialist' or unused experts.
Clusters of points are emerging 'Knowledge Galaxies'.

Figure 3: t-SNE of Layer 7


t-SNE Visualization of SoME Expert Key Space for Layer 7

How to Read This Plot:
• Each circle represents one of the 256 experts.
• Proximity suggests conceptual similarity learned by the router.
• Bright, large circles are high-usage 'generalist' experts.
• Dark, small circles are low-usage 'specialist' or unused experts.
Clusters of points are emerging 'Knowledge Galaxies'.

5.3. Quantitative Analysis of Organization

We tracked the Gini coefficient and entropy of expert usage for the middle layer across epochs. The Gini coefficient remained high and stable (0.878 to 0.864), indicating a consistent pattern of unequal expert utilization, which is characteristic of specialization. Concurrently, the entropy of usage steadily increased (5.178 to 5.296), suggesting that as training progresses, the model learns to route tokens to a wider variety of experts rather than concentrating all traffic on a few.

6. Discussion and Future Work

The results of our initial prototype are highly encouraging. They demonstrate that a stable, organized, and semantically meaningful map of expert knowledge can emerge from a set of simple, local, gradient-free update rules. This validates the core principles of SoME: decoupling function from address and employing a self-organizing routing mechanism.

However, the current prototype has several limitations that present clear avenues for future research.

- Routing and Gating: The current dot-product routing is sensitive to query magnitude. Future work will implement a scale-free cosine-based routing and introduce a temperature schedule to control the sharpness of the softmax gating, allowing for broader exploration in early training and sharper specialization later.

- Usage Inertia: The prototype uses a simple cumulative usage count, which can cause early popular experts to "freeze" in place. We will transition to an adaptive Exponential Moving Average (EMA) window for usage inertia, preserving plasticity while maintaining the stability of "gravitational mass."
- Peer Pull and Decay: The Peer Pull mechanism can cause premature over-clustering. We will gate this force with a utilization floor and schedule its learning rate (β). The decay mechanism will be made density-aware to promote better coverage of the conceptual space.
- Scalability: The prototype's matrix multiplication for routing is efficient for hundreds of experts but becomes intractable for tens of thousands or more. To truly realize SoME's potential, we will replace this step with an efficient Approximate Nearest Neighbor (ANN) index (e.g., HNSW, FAISS). This aligns with our framing of routing as a lookup, not a classification, problem and is orders of magnitude more scalable.

7. Conclusion

We have introduced the Self-Organizing Mixture of Experts (SoME), a novel architecture that challenges the static routing paradigm of traditional MoE models. By decoupling static expert knowledge from a dynamic, self-organizing address system governed by bio-inspired heuristic rules, SoME offers a path toward more adaptive and continually learning models. Our initial experiments have shown that this architecture is not only viable for effective learning but also demonstrably leads to the emergent organization of its internal knowledge map. The promising results and the clear roadmap for future improvements suggest that SoME represents an exciting new direction in the development of large-scale neural networks.