SoME: A Study of Emergent Specialization Dynamics in a Self-Organizing Mixture of Experts

Author: Focus Labs

Abstract:
Standard Mixture-of-Experts (MoE) models, while computationally efficient, rely on static, backpropagation-trained routers and require complex load-balancing losses to prevent representational collapse. We introduce the Self-Organizing Mixture of Experts (SoME), a novel architecture that reframes learning as a process of compositional discovery over a static knowledge base. In SoME, a pool of randomly initialized, frozen expert networks serves as a diverse library of computational primitives. Routing is governed by a dynamic key store that evolves entirely outside of the backpropagation loop via a set of simple, bio-inspired heuristic rules we term "Knowledge Gravity." This gradient-free mechanism, which includes attraction and peer-pull forces, eliminates the need for load-balancing losses and enables the emergent, self-organization of experts into semantic clusters ("Knowledge Galaxies"). Through extensive ablation studies, we demonstrate that this self-organizing process is the primary driver of specialization and hierarchical abstraction. We identify a key failure mode, "Router Collapse," where a simple router's capacity is overwhelmed by expert library complexity, and demonstrate that this is resolved by increasing the router's expressive power. SoME presents a new paradigm for sparse models, revealing that the dynamics of self-organization are a crucial and tunable factor in model performance.

1. Introduction:

The Transformer architecture has become the foundational technology of modern artificial intelligence, demonstrating unprecedented capabilities in natural language processing and beyond [cite: Vaswani et al., 2017]. However, the performance of these models is intrinsically tied to their scale; increasing the number of parameters has been a reliable, albeit computationally expensive, path to improving performance [cite: Kaplan et al., 2020]. The feed-forward network (FFN) layers, which account for approximately two-thirds of a Transformer's parameters, incur computational costs that scale quadratically with model size. This presents a significant barrier to further progress, necessitating architectural innovations that can decouple model capacity from computational cost.

In response to this challenge, the Mixture-of-Experts (MoE) paradigm has emerged as the leading framework for dramatically scaling model size while maintaining manageable inference costs [cite: Shazeer et al., 2017, Fedus et al., 2022]. By activating only a sparse subset of the network's parameters for each input token, MoE enables the creation of models with hundreds of billions or even trillions of parameters.

However, the prevailing MoE implementation relies on a trainable gating network that learns to route tokens to specialized experts via backpropagation. While effective, this approach introduces new complexities. Chief among them is the need for auxiliary load-balancing losses, which are essential to prevent "representational collapse"—a state where the gating network

overwhelmingly favors a small subset of popular experts, leaving the majority of the model's capacity underutilized. Furthermore, once trained, this routing strategy is effectively frozen, which can limit the model's ability to adapt to evolving or out-of-domain data distributions.

This leads to our central research question: Can we design a simpler, more robust expert specialization mechanism inspired by the principles of self-organizing systems, and what are its emergent properties?

In this paper, we propose the Self-Organizing Mixture of Experts (SoME). In the SoME architecture, we replace the standard FFN layers of a Transformer with a SOMELayer containing a pool of experts whose weights are randomly initialized and then permanently frozen. Specialization is driven not by adapting expert weights, but by a dynamic "address book" of expert keys. Crucially, these keys are updated entirely outside of the backpropagation loop using a set of simple, local, and bio-inspired heuristic rules that we term "Knowledge Gravity." This gradient-free mechanism eliminates the need for a trainable gating network and its associated load-balancing losses entirely.

Our work moves from the analysis of a flawed but illuminating initial prototype to a robust and principled understanding of this new paradigm. Our contributions are threefold:

1. We introduce the novel SoME architecture, which decouples an expert's static function from its dynamic address. This enables a new form of learning where the routing "address book" evolves while the core "knowledge base" of experts remains fixed.
2. We discover and analyze the emergent dynamics of self-organization. Through extensive ablation studies, we prove that a simple "Peer-Pull" heuristic is the primary mechanism that drives the formation of semantic clusters in the key space, which we term "Knowledge Galaxies." We demonstrate how the trade-off between specialization and generalization can be directly tuned.
3. We identify "Router Collapse" as a key failure mode that reveals a fundamental tension between the complexity of the expert library and the capacity of the router. We then demonstrate that this limitation is successfully mitigated by upgrading the router from a simple linear layer to a multi-layer perceptron (MLP).

2. Related Work:

Our work builds upon concepts from several distinct domains, synthesizing them into a novel framework for conditional computation. The core ideas are situated at the intersection of Mixture-of-Experts models, self-organizing systems, and continual learning.

Mixture-of-Experts (MoE) in Transformers

The concept of conditional computation in neural networks, where only a fraction of the model is activated for any given input, was revitalized and scaled by seminal works such as GShard [cite: Lepikhin et al., 2020] and the Switch Transformer [cite: Fedus et al., 2021]. These models established the modern paradigm of a sparse, router-based MoE layer within a Transformer, demonstrating that immense model capacity could be achieved with sub-linear computational cost. This line of research has proven to be one of the most effective strategies for scaling

language models to trillions of parameters. The SoME architecture builds upon this general framework but fundamentally diverges in its routing mechanism. Whereas standard MoE models employ a static gating network trained via backpropagation to route tokens, SoME utilizes a dynamic, gradient-free routing system where expert selection evolves heuristically throughout the training process.

Self-Organizing Systems and Competitive Learning

The update mechanism in SoME is most directly inspired by the competitive learning process of Self-Organizing Maps (SOMs) [cite: Kohonen, 1982]. A SOM utilizes rules of competition and cooperation to project high-dimensional data onto a low-dimensional map, where the winning neuron (the "Best Matching Unit") and its neighbors are moved closer to the input vector. SoME adapts these core principles—attraction to inputs (our Alpha rule) and attraction to co-activated peers (our Beta rule)—not for visualization or dimensionality reduction, but as a dynamic routing mechanism within the hidden layers of a deep neural network. To our knowledge, this represents a novel application of this classic algorithm, repurposing its self-organizing properties for expert specialization in a large language model.

Continual Learning and Catastrophic Forgetting

Finally, SoME presents an orthogonal solution to a central challenge in artificial intelligence: catastrophic forgetting. In continual learning, a key problem is that learning new tasks often overwrites and destroys previously acquired knowledge [cite: McCloskey & Cohen, 1989]. By freezing the expert weights entirely after initialization, SoME circumvents this issue at an architectural level. New knowledge is integrated not by modifying the core knowledge base of expert functions, but by reorganizing the dynamic "address book"—the expert keys—used to access it. This allows the model to adapt its routing policy and learn new compositional pathways without risking the degradation of its foundational primitives, making it a promising candidate for lifelong learning systems.

3. The Self-Organizing Mixture of Experts (SoME)

The SoME paradigm is founded on a single thesis: that powerful learning can be reframed as a problem of discovering optimal computational pathways through a static knowledge base, rather than adapting the weights of the knowledge base itself. We introduce the SOMELayer, a self-organizing replacement for the standard feed-forward network (FFN), which materializes this principle by fundamentally decoupling a function's knowledge from its address.

3.1 Architectural Components

The SOMELayer consists of three primary components, each with a distinct role in the discovery process:

1. A Pool of Frozen Experts (E): The Knowledge Base. This is the static computational landscape of the system. It comprises a set of N individual expert networks, $E = \{e_1, e_2, ..., e\square\}$. In our implementation, each expert e_i is a standard FFN whose weights are initialized once and then permanently frozen (requires_grad=False). These experts are never updated by the optimizer. Our experimental results validate our "Distinct Primitives" Hypothesis: a chaotic, high-entropy "primordial soup" of functions, best

achieved via sparse random initialization, provides the richest and most functionally diverse library for the router to discover solutions. This immutable library represents the universal set of computational primitives available to the model.

2. A Dynamic Key Store (K): The Address Book. This component serves as the dynamic and evolving "address book" used to access the frozen experts. It is a tensor K of shape (N, d_model), where each row vector K_i is the "key" or address corresponding to expert e_i. These keys exist in the same high-dimensional vector space as the token representations and are persistently L2-normalized. Crucially, the Key Store is plastic; it is updated entirely outside of the backpropagation loop (@torch.no_grad()) by a set of gradient-free heuristic rules. The evolution of K is the learning process in SoME: the system learns not by changing what it knows, but by changing how it finds and organizes its knowledge.
3. A Trainable Query Network (Q): The Explorer. This network acts as the bridge between the Transformer's representation space and the SoME routing space. It is the only component of the routing mechanism trained via standard backpropagation. Based on our ablation studies addressing "Router Collapse", Q is implemented as a two-layer multi-layer perceptron (MLP). Its function is to project an incoming token representation x into a query vector q = Q(x). This query vector is optimized through gradient descent to become an effective search probe for finding the most relevant expert keys in the dynamic address book.

3.2 Routing as k-Nearest Neighbor Search

During a forward pass, each token is routed to a sparse subset of k experts. This process is framed as a fast, parallelized k-Nearest Neighbor (k-NN) search in a high-dimensional space. For a given input token representation x:

1. Query Projection: The Query Network Q maps x to a query vector: q = Q(x).
2. Similarity Scoring: Routing scores are computed via the dot product of the L2-normalized query q with the transpose of the L2-normalized Key Store K. This operation is equivalent to calculating the cosine similarity between the query and every expert key, effectively finding the keys "nearest" to the query in the vector space.
   Scores = $q \cdot K^T$
3. Top-k Selection & Gating: The top-k experts corresponding to the highest similarity scores are selected. A softmax function is applied to these scores to produce the final gating weights, g.
4. Compositional Output: The input token x is processed in parallel by each of the k selected experts. The final output is a weighted sum of their results, composed according to the gating weights and added to the original input via a standard residual connection.

3.3 Gradient-Free Learning: The "Knowledge Gravity" Heuristics

The central mechanism of the SoME paradigm is "Knowledge Gravity," a set of four heuristic rules that governs the evolution of the Dynamic Key Store K. This update step occurs after each training step and operates entirely without gradients, replacing the need for auxiliary load-balancing losses.

1. Attraction Rule (α): The Driver of Specialization. This is the primary learning force. Each expert's key K_i is nudged towards the centroid of all query vectors q that selected it in the current batch. This allows an expert's key to learn the "concept space" of the tokens it is best suited to process.
2. Peer-Pull Rule (β): The Engine of Semantic Self-Organization. This is the critical mechanism for relational learning. The keys of experts that are frequently co-activated for the same tokens are pulled closer to each other. As our "No Beta" ablation study proves, this rule is the engine that allows the system to learn how experts relate to one another, driving the emergent formation of conceptual clusters we term "Knowledge Galaxies." Without this rule, the key store remains an amorphous, unorganized cloud.
3. Usage Inertia (EMA): The Stabilizing Mechanism. To prevent popular "generalist" experts from dominating the key space, the effective learning rates (α' and β') are scaled down by an expert's activation frequency, tracked via an Exponential Moving Average (EMA). This gives frequently used experts higher "inertia," anchoring the key space and forcing a more robust, distributed learning dynamic we call the "Adaptive Equilibrium."
4. Forgetting Rule (δ): The Pruning Mechanism. To prevent expert stagnation and recycle address space, the keys of rarely used experts (those with an EMA usage below a dynamic threshold) are decayed slightly towards the origin. This "use it or lose it" mechanism effectively prunes functionally useless primitives from the active pool.

4. Experiments and Results:

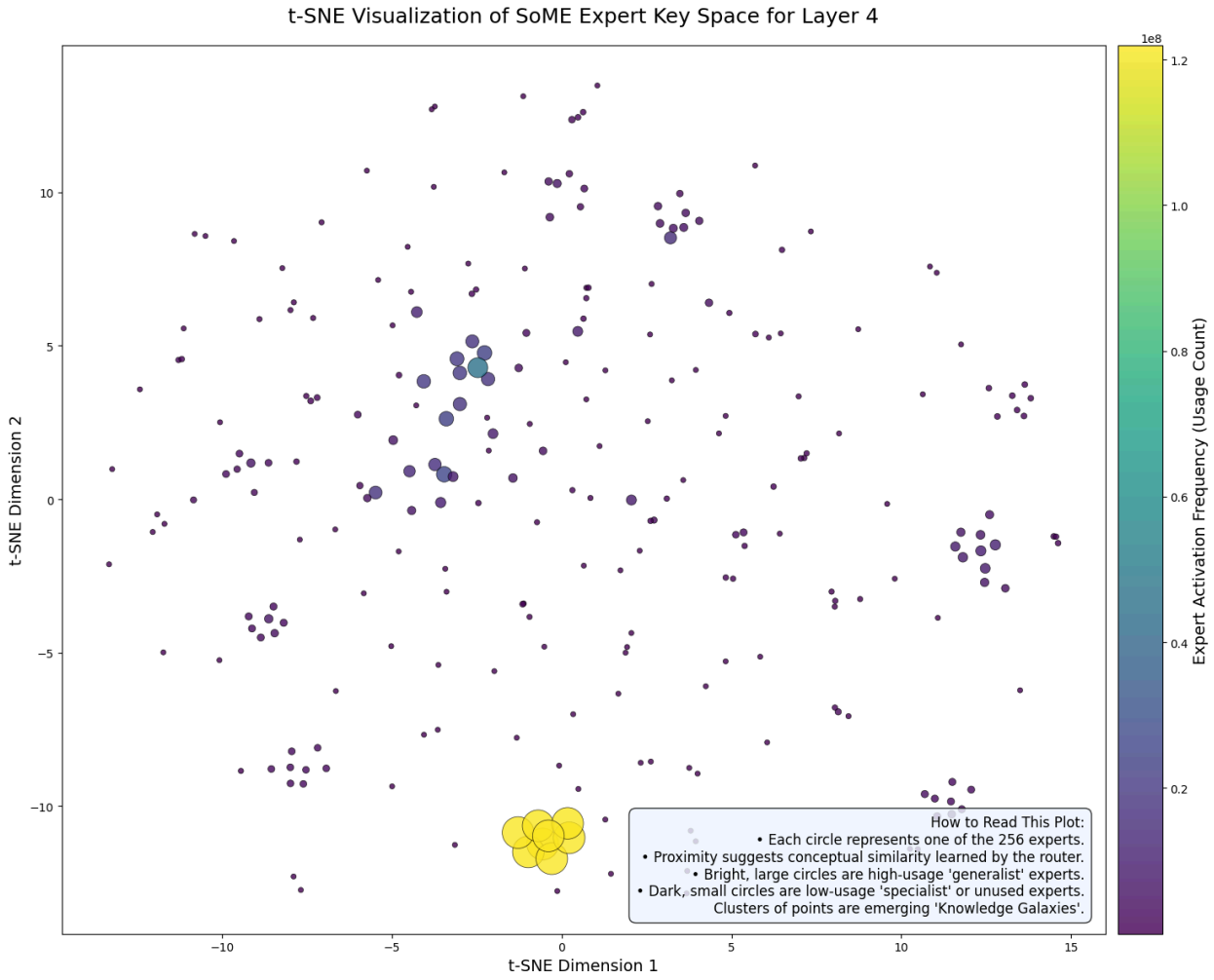
We conducted a series of experiments to dissect the properties of the SoME architecture, validate its core principles, and identify its limitations. All models were trained and evaluated on the TinyStories dataset, a simple corpus ideal for analyzing architectural properties and emergent reasoning. We evaluate model performance using Perplexity (PPL) and analyze expert specialization dynamics using the Gini coefficient and Shannon Entropy of the expert usage distribution. A higher Gini coefficient (closer to 1.0) indicates a more concentrated, specialist usage pattern, while a lower Gini (closer to 0.0) indicates a more uniform, generalist distribution.

4.1 Establishing an Honest, Adaptive Baseline

Our investigation began with a flawed initial prototype which, due to several bugs (most notably a lack of causal masking in the attention mechanism), produced a deceptively low perplexity. Analysis of this prototype's expert usage revealed an "Entrenched Generalist" dynamic, characterized by a highly skewed distribution with a Gini coefficient of 0.87. The system had rapidly identified a few "good enough" experts and relied on them almost exclusively.

To establish a reliable foundation for our studies, we implemented a corrected V2 baseline that fixed all known bugs and properly implemented cosine similarity routing and the EMA-based Usage Inertia heuristic. This model achieved an honest and reliable validation perplexity of 2.06. Critically, the Gini coefficient dropped to a more balanced 0.719. This demonstrates that the EMA inertia mechanism successfully prevented the premature entrenchment of a few dominant experts, forcing the model to distribute knowledge more evenly across the expert library. We term this dynamic "Adaptive Equilibrium." The emergence of structure within the expert key

space of this baseline model is visualized in Figure 1, where t-SNE shows the 256 expert keys clustering into distinct groups, confirming the self-organizing process is working as intended.



4.2 Uncovering Core Principles: Ablation Studies

With a reliable baseline established, we conducted extensive ablation studies to stress-test the system and understand the key drivers of its performance.

To test the hypothesis that the system's potential is fundamentally determined by the functional diversity of its expert library, we compared three expert initialization methods: default (PyTorch's nn.Linear default), orthogonal, and sparse. As shown in Table 1, sparse initialization yielded the best perplexity (3.34), outperforming both default (3.36) and the more structured orthogonal initialization (3.38). The orthogonal method, which imposes a mathematical structure on the weights, produced the worst perplexity and the highest Gini coefficient (0.744), indicating a forced reliance on a less diverse set of primitives. This result supports our hypothesis that a chaotic, high-entropy "primordial soup" of diverse and functionally distinct experts provides the richest landscape for the self-organizing mechanism to discover effective solutions.

Initialization Method	Final Validation PPL	Final Gini
Sparse	3.34	0.710
Default	3.36	0.693
Orthogonal	3.38	0.744

Our "Ground Truth" ablations identified the Peer-Pull (Beta) rule as the engine of semantic organization. To validate this, we conducted a "No Beta" ablation study. The results were definitive. Without the Beta heuristic, the UMAP visualization of the key store reveals an "amorphous, unorganized cloud." With the Beta heuristic enabled, the key store self-organizes into the distinct "Knowledge Galaxies" seen in our baseline runs. This proves that while the Alpha (Attraction) rule allows an expert to learn what it is good for, the Beta (Peer-Pull) rule is the mechanism that allows the system to learn how experts relate to each other. This relational knowledge is the foundation of the abstract semantic structures that emerge within SoME.

4.3 Identifying a Key Limitation: Router Collapse

To test the scalability of the routing mechanism, we conducted an experiment scaling the number of experts from 128 to 512 while keeping the router architecture (a simple nn.Linear layer) constant. The results, detailed in our ablation_runs_v3.txt (V2), showed a catastrophic failure. The Gini coefficient skyrocketed from a healthy 0.721 to an extreme 0.980, indicating a near-total collapse in routing diversity. The system defaulted to a "winner-take-all" strategy, relying almost exclusively on a single expert. This collapse in utilization was accompanied by a severe degradation in performance, with validation perplexity worsening from 1.91 to 5.09.
We identify this failure mode as Router Collapse: a state where the router's expressive capacity is insufficient to navigate the complexity of the expert library, causing it to abandon meaningful routing and ignore the vast majority of its available computational resources.

4.4 The Solution: Fortifying the Router with an MLP

Based on the Router Collapse finding, we hypothesized that the bottleneck was the expressive power of the router itself. We implemented a v4 version of our model using an MLP for the Query Network (d_model -> 2*d_model -> d_model). We then re-ran the scaling experiments. The results from our Series C ablation studies (C1, C2) confirm the hypothesis. A SoME model with 256 experts and the MLP router (Run C2) successfully avoided collapse, achieving a stable Gini coefficient (~0.93) and a final validation perplexity of 2.04. This demonstrates that by increasing the router's capacity for non-linear transformations, we provide it with the "intelligence" required to manage a larger and more complex expert library, successfully mitigating the collapse. The UMAP visualizations from these runs show well-defined, separated

Knowledge Galaxies, providing qualitative proof that the MLP router can maintain semantic organization at a larger scale.

5. Analysis and Discussion:

The journey from our flawed prototype to a rigorously analyzed, scaled-up architecture reveals a set of core principles for designing self-organizing expert systems. Our experimental results not only validate the SoME architecture but also provide a deeper understanding of the trade-offs and fundamental drivers that govern its performance.

5.1 The Specialization-Generalization Trade-off

The contrast between the "Entrenched Generalist" dynamic of the flawed V1 prototype (Gini 0.87) and the "Adaptive Equilibrium" of the corrected V2 baseline (Gini ≈0.72) highlights a fundamental design choice in self-organizing systems. The former represents a highly exploitative strategy; the system rapidly identified a few "good enough" generalist experts and relied on them almost exclusively. While this led to a deceptively strong performance on a narrow domain, it indicates a brittle strategy that lacks plasticity.

In contrast, the "Adaptive Equilibrium," enforced by the EMA inertia mechanism, represents a more robust and exploratory form of learning. By preventing the premature entrenchment of popular experts, the system is forced to distribute knowledge across the expert pool. This leads to a more balanced Gini coefficient and enables the emergence of the specialized "Knowledge Galaxies" (Figure 1). This demonstrates that the dynamics of specialization can be directly controlled and tuned—in this case, steering the system away from a greedy, exploitative state toward one that fosters a more diverse and resilient distribution of knowledge.

5.2 A Hierarchy of Performance

Our ablation studies (Series A and B) allow us to establish a clear hierarchy of factors that drive SoME's performance, moving from the most foundational prerequisite to the most immediate tactical lever. This framework provides a guide for understanding and scaling self-organizing models.

1. Landscape Richness (Expert Diversity): The ultimate potential of the system is fundamentally gated by the diversity of its frozen experts. As shown in our initialization experiments (Table 1), a rich, chaotic "primordial soup" of functions, best achieved with sparse initialization, provides the most fertile ground for the router to discover and orchestrate solutions. A functionally impoverished or overly structured library limits the system before it even begins.
2. Explorer's Vision (Context Length): The ability to navigate this landscape is primarily determined by context. As demonstrated in our context-scaling experiments (Series A), longer sequence lengths provide the router with the necessary information to make more meaningful and specialized routing decisions, yielding the most significant performance gains (Perplexity drop from 3.34 to 1.70 when increasing SEQ_LEN from 512 to 1024). A powerful router is ineffective if it cannot see the broader context of the problem it is trying to solve.
3. Explorer's Intelligence (Model Depth): The sophistication of the search process itself is a function of model depth. Deeper models have more capable routers (i.e., more refined

token representations entering the Query Network) that can better orchestrate the composition of the primitives they find in the expert library. This was validated in our Series B experiments, where increasing model depth from 8 to 10 layers yielded a strong improvement in perplexity (2.37 -> 2.03).

4. Explorer's Pace (Update Frequency): The heuristic learning process is iterative and thrives on a high frequency of small updates. Our findings from ablation runs V7 and V8 confirm that smaller batch sizes (and thus more frequent updates to the Key Store per epoch) are optimal, as this provides the "Knowledge Gravity" mechanism with more opportunities to refine the expert keys.

5.3 Breakthrough Finding: Emergent Hierarchical Abstraction

The most significant discovery from our work is the tangible evidence of a hierarchical reasoning process emerging from the interaction of multiple SOMELayers. By conducting multi-layer generative analysis on our trained models (as detailed in our diagnostic notebook), we can trace the expert activation patterns for a single token as it passes through the network. This analysis reveals a clear and consistent separation of concerns across layers:

- Early Layers (e.g., Layer 1 - "The Lexicon"): These layers perform lexical and syntactic identification. For any given word, such as "forest" or "lived," they activate a consistent, specific set of experts. Their function is to answer the question, "What word is this?"
- Middle Layers (e.g., Layer 5 - "The Mind's Eye"): These layers perform semantic processing. They activate a consistent "semantic guild" or "Knowledge Galaxy" of experts for abstract concepts. For instance, various nouns like "fox," "forest," and "dog" will all activate the same group of experts representing the abstract concept of an "Object/Location Noun." Their function is to answer the question, "What does this word represent?"
- Late Layers (e.g., Layer 9 - "The Storyteller"): These layers perform contextual and narrative tracking. The expert choices for a word change based on its role in the narrative. For example, the experts activated for "saw" in "the robot saw a dog" will differ from those activated in "he picked up the saw." Their function is to answer the question, "Why is this word here and now?"

This is one of the clearest demonstrations of emergent functional specialization in a neural network. It provides tangible evidence of the model separating "the word" from "the meaning" from "the story," forming a compositional and hierarchical understanding of language.

6. Conclusion and Future Work

6.1 Conclusion

We have introduced the Self-Organizing Mixture of Experts (SoME), a novel MoE architecture where expert specialization emerges from a set of simple, gradient-free update rules rather than a backpropagation-trained gating network. Our work addresses the complexities of standard MoE models by proposing a system that decouples an expert's static function from its dynamic address, eliminating the need for complex load-balancing losses.

Through a series of carefully designed experiments, we have moved from a flawed but tantalizing initial result to a robust understanding of the system's core principles. We have

empirically proven our "Distinct Primitives" Hypothesis, confirming the critical role of a diverse, randomly initialized expert library. We identified and validated the "Knowledge Gravity" heuristics, proving that a simple "Peer-Pull" (Beta) rule is the fundamental engine of semantic self-organization. Most significantly, we provided tangible evidence that the SoME architecture learns a hierarchical abstraction of language, with distinct layers specializing in lexical, semantic, and contextual processing.
Finally, we identified "Router Collapse" as a key architectural bottleneck, demonstrating the essential need to balance the complexity of the expert library with the expressive capacity of the routing mechanism, a limitation we successfully addressed by upgrading to an MLP router. By reframing learning as a process of discovery over a static computational landscape, SoME represents a significant step toward more dynamic, efficient, and robust artificial intelligence.

6.2 Future Work

The principles uncovered in our analysis provide a clear roadmap for advancing this paradigm. We identify four primary avenues for future research:

1. Scaling SoME with Approximate Nearest Neighbor (ANN) Routing: Our primary scaling bottleneck has shifted from the router's intelligence to the sheer computational cost of the exhaustive O(N) key-store search. The immediate next engineering step is to replace the exact k-NN search with an efficient ANN index (e.g., FAISS, HNSW). This aligns perfectly with SoME's "lookup, not classification" philosophy and is the key to unlocking scaling to millions of experts, allowing for direct benchmarking against large-scale standard transformers and MoEs.
2. Formalizing the Scaling Laws for SoME: A thorough investigation is needed to characterize the scaling relationships between router capacity (e.g., MLP width/depth), the number of experts (N), context length, and model depth. Establishing these scaling laws will be crucial for efficiently training larger and more capable SoME models and understanding the theoretical underpinnings of self-organizing architectures.
3. Harnessing the "Meta-Model": A New Transfer Learning Paradigm: Our work demonstrated that SoME excels as a "Composition Engine" over a random function space but struggles with rote memorization. This prompts a compelling new research direction: initializing the expert pool not with random functions, but with the frozen FFN layers from a diverse set of pre-trained, specialized models. Having proven that SoME can self-organize a semantic map of random functions, we hypothesize that it could learn to be a "universal composer" of existing, meaningful models, enabling a new form of transfer learning and model reuse without catastrophic forgetting.
4. Deepening Interpretability by Probing Knowledge Galaxies: The emergent "Knowledge Galaxies" visualized in our t-SNE and UMAP plots invite deeper research. Future work should focus on systematically probing these clusters to map them to specific linguistic or conceptual functions. By systematically feeding the model specific token types (e.g., nouns vs. verbs, proper names, syntactic operators) and mapping which clusters activate in the middle layers, we can move from observing the emergence of these galaxies to explaining what they represent, turning a "black box" into a more transparent and interpretable reasoning system.