Bio-Inspired Modular Architecture for Efficient Memory and Reasoning in Neural Networks: Preliminary Results of a Compact Prototype "Cortex-V0.1"

Arian Vazquez Fernandez
Independent self-taught researcher
Jerez de la Frontera, Spain

Abstract

This work presents a bio-inspired modular architecture for associative memory and reasoning in neural networks, arising from an intuitive idea originated in a dream. The proposal explicitly separates long-term memory (based on a modern Hopfield network) from dynamic reasoning (implemented through Mixture of Experts). Training is performed in two phases with parameter freezing, enabling robustness to imperfect retrieval. In a compact prototype with fewer than 15 million trainable parameters, emergent specialization of experts is observed (visualized via PCA) and a significant improvement in the semantic similarity of responses, even when memory retrieval is poor. The preliminary results suggest that bio-inspired modular designs can achieve advanced cognitive behaviors with extreme efficiency.

Index Terms

bio-inspired neural networks, associative memory, modern Hopfield, Mixture of Experts, modular training, semantic emergence, computational efficiency, compact prototype

I. Introduction

Current artificial neural networks have achieved impressive capabilities in complex tasks of language processing and reasoning. However, they present important limitations that contrast with the efficiency and robustness observed in biological systems. Among these limitations are the high computational cost, the tendency to generate responses incoherent with the stored knowledge, and the difficulty in maintaining stable long-term representations. These shortcomings become especially evident when models must operate with restricted resources or handle partial or noisy information. In the human brain, these problems are solved through a highly efficient and modular organization. Neuronal activation is extremely sparse: at any given moment, only a small fraction of neurons is active, allowing minimal energy processing. Additionally, there is a clear functional separation between specialized regions. For example, the hippocampus acts as a rapid associative memory system capable of retrieving complete patterns from partial inputs, while areas of the prefrontal cortex integrate this information to perform executive reasoning and decision-making. This division allows the entire system to be robust to incomplete or degraded memories: although the initial retrieval may be imperfect, subsequent processes can compensate for it and generate coherent responses. The present work arises precisely from the search for artificial architectures that incorporate these biological principles. The central idea was born from an intuitive vision during a dream: a neural network in which the input signal selectively activates small and relevant pathways, instead of propagating throughout the structure, imitating the associative and conditional patterns of human memory. This intuition materialized into a functional prototype developed in just 12 days, demonstrating that deep bio-inspired concepts can be explored quickly with accessible tools. The main objective is to address the need for more efficient architectures faithful to known biological principles. Current approaches based on dense networks scale primarily by increasing parameters and computation, but this paradigm shows diminishing returns and sustainability issues. Instead, it is proposed to explore designs that prioritize functional modularity, dynamic sparsity, and strict separation between storage and processing, enabling advanced cognitive behaviors even in reduced-size models. The specific contributions of this preliminary work are as follows:

- A bio-inspired modular architecture divided into two clearly separated phases, with explicit parameter freezing between phases to preserve memory integrity while training reasoning.
- The novel integration of a modern Hopfield network as a long-term associative memory module with a Mixture of Experts layer for dynamic reasoning, including topological routing mechanisms that induce sparse pathways.
- Empirical evidence of emergent expert specialization in an extremely compact model (fewer than 15 million trainable parameters), observable through principal component analysis (PCA) of routing gates throughout training.
- Demonstration of cognitive robustness to imperfect retrieval: the system generates semantically coherent responses even when the memory module provides noisy or incomplete vectors, similar to how the brain compensates for partial memories through cortical inference.

- Validation that advanced behaviors, such as semantic emergence and distributed reasoning, can arise in very limited resource configurations, opening paths toward more efficient and accessible artificial intelligence.

The rest of the document is structured as follows: Section II reviews the most relevant related works, Section III describes the proposed architecture in detail accompanied by its conceptual diagram, Section IV presents the experimental methodology employed, Section V exposes the results obtained through systematic sweeps, Section VI analyzes and interprets the findings from a bio-inspired perspective, and finally Section VII summarizes the conclusions and proposes future lines of work.

II. Related Works

Modern Hopfield networks [1]–[4] have experienced a notable resurgence as efficient models of continuous associative memory, capable of storing an exponential number of patterns relative to the dimension of the state space. These networks allow the retrieval of complete patterns from partial or noisy inputs, making them natural candidates for modeling long-term memory systems in artificial architectures. On the other hand, Mixture of Experts (MoE) [5], [6] have proven to be an effective strategy for introducing dynamic routing and subnetwork specialization in large-scale models. By selectively activating only a fraction of the total parameters for each input, MoEs achieve remarkable computational efficiency, as observed in contemporary models with hundreds of millions or billions of parameters. There are previous efforts to combine associative memory elements with expert routing mechanisms. Some works integrate Hopfield variants within transformers to improve long-term attention, while others explore MoE in external memory or augmented retrieval contexts. However, these approaches typically operate at massive scales or maintain end-to-end training that does not explicitly separate storage and reasoning functions. In the bio-inspired domain, recent research has analyzed the emergence of complex patterns in Hopfield networks under different connectivity topologies, as well as associative properties in continuous memory models. Nevertheless, these studies focus primarily on theoretical analysis or the internal dynamics of memory, without incorporating dynamic reasoning layers based on experts or evaluating the robustness of the complete system to imperfect retrieval. The main difference of the present proposal lies in the novel and explicit integration of a modern Hopfield network as a dedicated long-term memory module with a Mixture of Experts layer for dynamic reasoning, all within a modular two-phase design with parameter freezing. Additionally, the prototype runs at an extremely compact scale (fewer than 15 million trainable parameters), allowing the observation of emergent specialization phenomena under very limited resource conditions, an aspect little explored in the existing literature.

III. Proposed Architecture

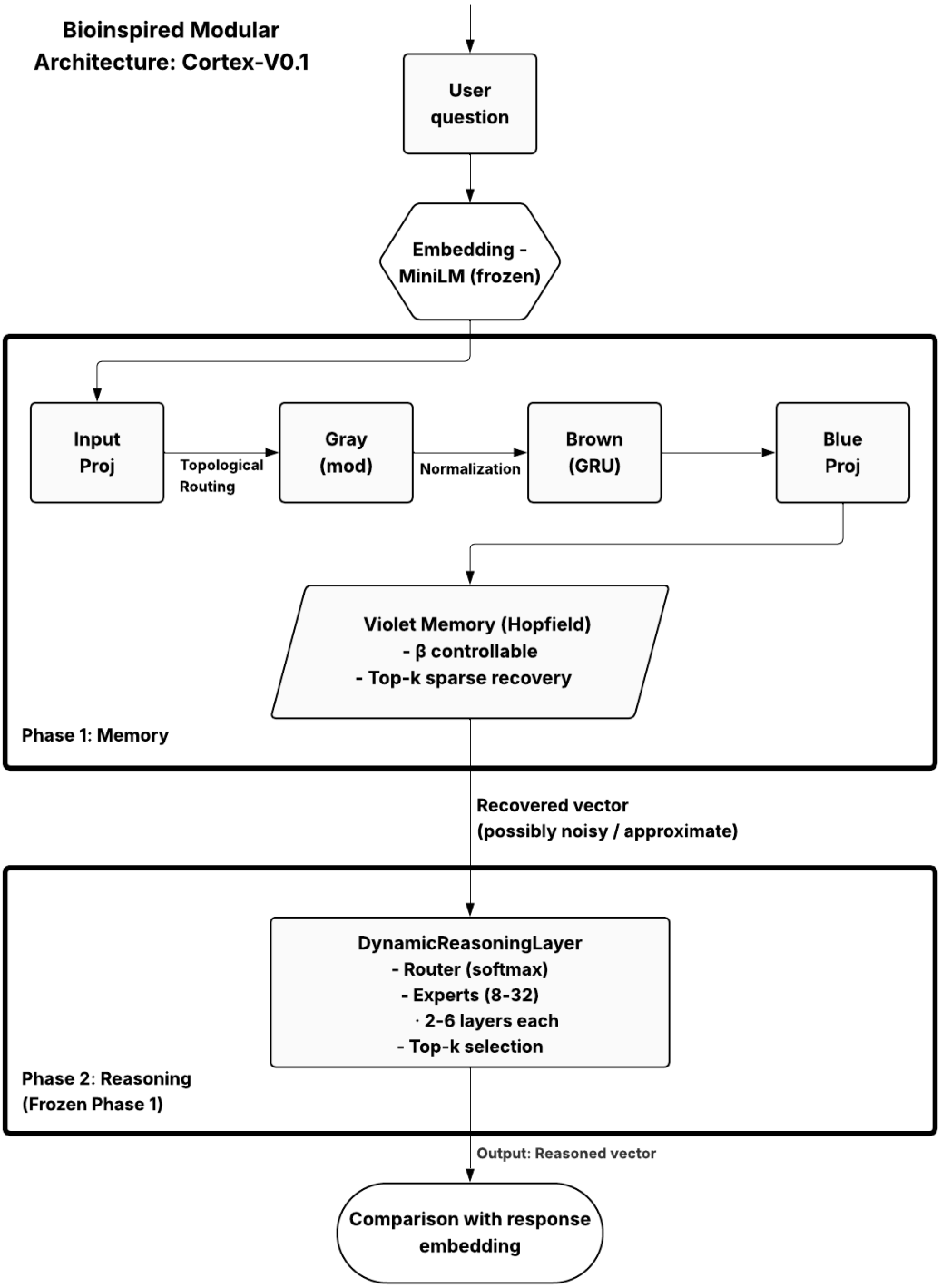


Figure 1: Overview of the proposed architecture. The input (user question) passes through a series of layers implementing topological routing toward long-term memory (modern Hopfield). The retrieved output is delivered to a Mixture of Experts layer for dynamic reasoning. This figure illustrates the complete flow, highlighting the modular separation between phases and the use of sparsity for efficiency.

The proposed architecture is based on a strict separation between two main functions: memory storage and retrieval on one hand, and reasoning over the retrieved information on the other. This division is implemented through training in two

consecutive phases, with complete freezing of the first phase before proceeding to the second. Figure 1 shows the complete flow of the architecture.

A. Phase 1: Long-Term Memory and Topological Routing

The first phase is dedicated exclusively to the memory module. The input, consisting of the user question encoded via embeddings (SentenceTransformer all-MiniLM-L6-v2), is processed sequentially through the following layers:

- **Initial projection layer (input_proj)**: A linear transformation that aligns the embedding space with the model's internal dimension (384 in this implementation).
- **Gray layer (gray)**: A small feedforward network with expansion to double dimension, ReLU activation and compression back, followed by a residual connection and Tanh activation. This layer acts as a dynamic modifier of the representation, starting the topological routing process.
- **Brown layer (brown)**: A three-layer recurrent GRU network with dropout. This component introduces sequential processing and temporal refinement, contributing to the formation of selective pathways in the representation.
- **Normalization (brown_norm)**: LayerNorm applied to the GRU output to stabilize the distribution.
- **Blue layer (blue_proj)**: Final linear projection that prepares the query for the Hopfield memory.

The resulting query is delivered to the **violet memory**, implemented as a modern Hopfield network with controllable β parameter and top-k selection. This memory stores the patterns (unique dataset contexts) as buffers and performs weighted retrieval via scaled softmax attention. Only the top-k most relevant patterns are used to build the retrieved vector, inducing extreme sparsity and limiting computation to small pathways within the memory. Training in this phase optimizes exclusively the layers prior to the Hopfield (input_proj, gray, brown, and blue_proj), using a combination of cosine similarity loss and cross-ranking loss to align the retrieval with the target pattern.

B. Phase 2: Dynamic Reasoning

Once Phase 1 is completed and frozen, the reasoning layer is added. The vector retrieved by the memory (which may be noisy or approximate) is delivered directly to a Mixture of Experts layer (DynamicReasoningLayer). This layer consists of:

- A linear router that produces scores for each expert.
- A configurable set of experts (typically 8–32), each composed of several deep feedforward layers with GELU activation and dropout.
- Top-k selection of experts per input, with softmax weighting of their outputs.

Training in this phase optimizes only the router and expert parameters, without propagating gradients to the frozen memory. This restriction forces the reasoning to develop robustness to imperfect inputs from retrieval. The model's final output is the vector generated by the experts layer, compared to the target response embedding via cosine similarity. The strict separation and freezing between phases constitute the core of the modular design, allowing each component to specialize in its function without mutual interference.

IV. Experimental Methodology

The evaluation of the proposed architecture was carried out in a limited-resource environment, using exclusively a Google Colab T4 GPU. All code was implemented in PyTorch, with experiment logging via Weights & Biases (wandb) to ensure reproducibility.

A. Dataset and Preprocessing

A subset of the SQuAD (Stanford Question Answering Dataset) [7] was used, specifically the first 10,000 samples from the training split. This set contains questions, answers, and associated contexts in English. For Phase 1 (memory), unique contexts were extracted from the dataset, obtaining approximately 2,000-3,000 distinct patterns depending on the run. These contexts were encoded using the SentenceTransformer 'all-MiniLM-L6-v2' model (frozen, no training), generating 384-dimensional embeddings that were subsequently L2-normalized. For Phase 2 (reasoning), question-answer pairs were constructed. The target answer was expanded by taking a contextual window centered on the exact answer (60 characters before and after when possible). In cases where the answer was not found in the context, a simple construction of the type "The answer is [answer]" was used. Additionally, questions were classified into three categories ("math", "history", "facts") using keyword-based rules, solely for visualizing emergent specialization.

B. Training Configuration

Training was strictly divided into two independent phases:

- **Phase 1**: Only the topological routing layers (input_proj, gray, brown, brown_norm, and blue_proj) were trained for 30 epochs with a fixed batch size of 32 and learning rate of 5e-4. The loss combined cosine similarity (main) and ranking cross-entropy (factor 0.1). The Hopfield memory remained as a buffer without trainable parameters.
- **Phase 2**: Once the Phase 1 model was saved and frozen, the reasoning layer was trained for 50 epochs. Different hyperparameters were systematically explored via wandb sweeps, varying:
  - **Learning rate**: 1e-3, 5e-4, 1e-4.
  - **Batch size**: 16, 32, 64.
  - **Number of experts**: 8, 16, 32.
  - **Layers per expert**: 2, 4, 6.
  - **Top-k experts**: 2, 4, 6.

An ablation configuration was included, replacing the MoE layer with a dense MLP of comparable capacity (16 linear layers 384→384 with GELU). The loss in Phase 2 was exclusively 1 minus cosine similarity between the reasoning output and the target answer embedding.

C. Metrics

In Phase 1, recall@K (K=1,5,10) was measured both in training and evaluation with Gaussian noise added to the queries. In Phase 2, the following were recorded: - Average loss per epoch - Close_answer_mean: average cosine similarity between reasoned output and target answer - Exact_match: proportion of answers with cosine similarity above 0.9 (strict threshold) Additionally, average routing gate scores per question type were accumulated, allowing visualization via PCA every 10 epochs.

D. Computational Resources

All experiments were run on Google Colab instances, with GPU access (generally T4). The complete trained model has fewer than 15 million trainable parameters in total (approximately 3.5M in Phase 1 and up to 9.5M in Phase 2 depending on the expert configuration). Memory and energy consumption remained low thanks to the sparsity induced by top-k mechanisms in both memory and experts.

V. Results

The experiments were conducted through systematic sweeps in Phase 2, exploring multiple hyperparameter configurations to evaluate the robustness and reproducibility of the findings. Below, the results are detailed in sections ordered from least to most interest, starting with basic convergence metrics and moving to evidence of more complex emergent behaviors. Each section includes analyses that confirm the consistency of the observed patterns across multiple runs and controlled comparisons.

A. Basic Convergence Metrics

In all evaluated configurations, the loss converges consistently during the 50 training epochs, descending from initial values around 0.85 to stabilizing in ranges between 0.50 and 0.60. This convergence is observed in both Mixture of Experts variants and the dense MLP ablation, indicating stable optimization of the reasoning module regardless of the specific mechanism employed. The reproducibility of this behavior was verified through multiple runs with different seeds, where the loss curves maintain similar trajectories without significant deviations attributable to random initializations. The exact match metric (exact_match), defined as the proportion of responses with cosine similarity above 0.9, remains close to zero in all configurations, with minor fluctuations not exceeding 0.0006. This result remains constant throughout the epochs and shows no clear dependence on the varied hyperparameters, which is consistent with the strict threshold used and the approximate nature of embeddings in an open-question dataset.

B. Comparison Between Configurations

The average semantic similarity (close_answer_mean) shows progressive improvement throughout training, ascending from initial values around 0.35 to stabilizing at 0.45-0.6 in the final epochs. This metric was calculated by averaging the cosine similarity between the reasoning output and the target answer embedding over the entire dataset in each epoch. When comparing Mixture of Experts variants against the ablation (dense MLP of equivalent capacity), a systematic difference is observed: MoE configurations achieve final values higher by approximately 0.07-0.12 points. This advantage remains consistent across different hyperparameter combinations (number of experts, layers per expert, top-k), which was verified through multiple independent sweeps. The best identified configuration involves 16 experts with 4 layers each and top-k=4, reaching close_answer_mean of 0.6, while the equivalent ablation stalls at around 0.42.

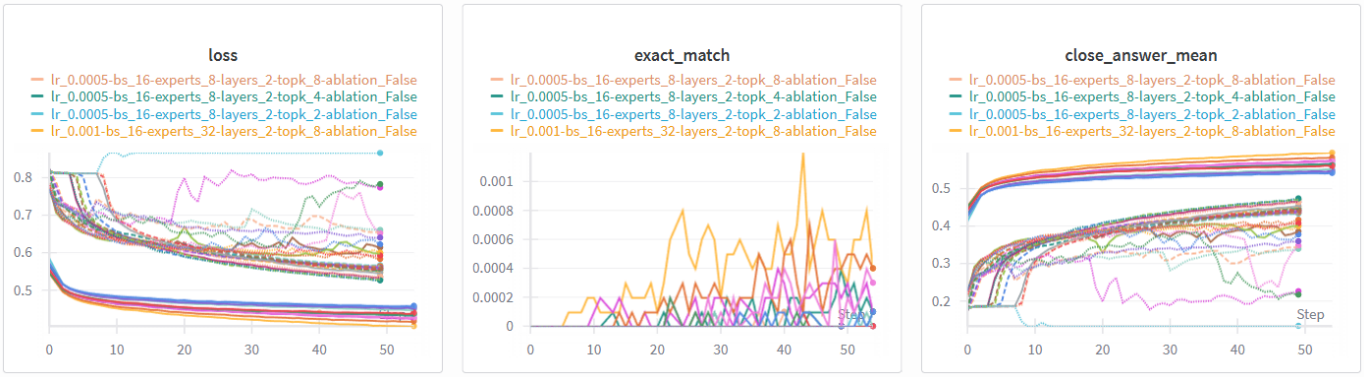


Figure 2: Results of sweeps in Phase 2: evolution of loss, exact_match, and close_answer_mean throughout the epochs for 50 initial configurations. Stable convergence and improvement in semantic similarity are observed, with a clear advantage for MoE variants.

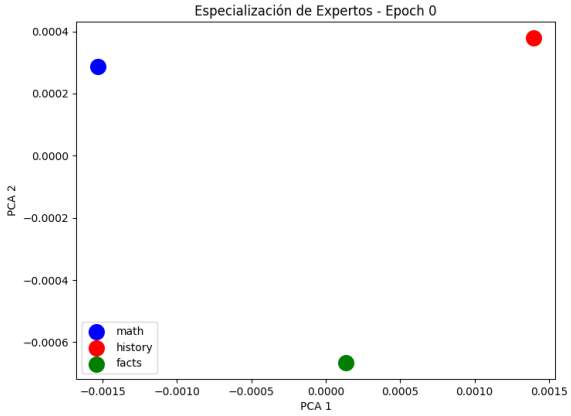
C. Computational Consumption and Efficiency

Monitoring GPU metrics during sweeps reveals highly efficient consumption. GPU power is in ranges of 30-40 W, with minor peaks attributable to intensive but brief computation operations. The GPU power usage percentage averages low and stable values, while the enforced power limit remains constant around 120 W without variations. GPU temperature slightly decreases and stabilizes around 55-65 °C, indicating absence of thermal overload. Finally, GPU utilization averages 50-60%, with minimal variations between configurations, which was corroborated in multiple independent runs. These measurements were obtained directly from wandb logs, confirming that the sparsity induced by top-k mechanisms (in memory and experts) contributes to efficient execution, without dependencies on specialized hardware.

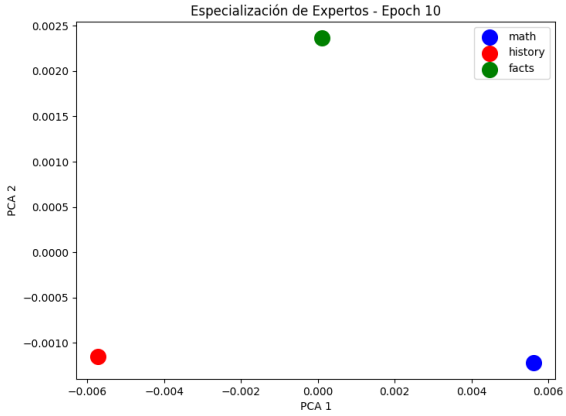
Figure 3: Computational consumption metrics during Phase 2 sweeps: GPU power (W), power usage (%), enforced power limit (W), temperature (°C), and utilization (%). The stable and low values demonstrate the efficiency of the sparse design, suitable for low-resource environments.

D. *Emergent Specialization and Robustness*

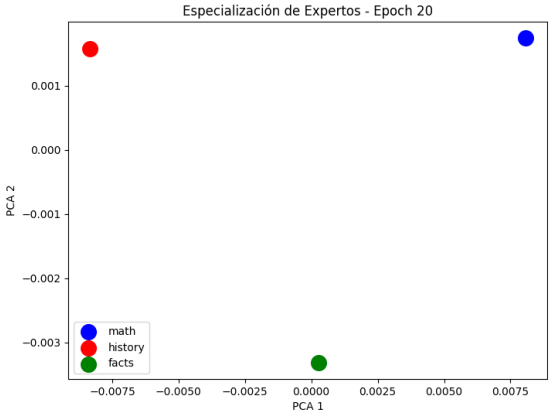
The visualization through principal component analysis (PCA) of the average routing gate scores reveals a phenomenon of emergent expert specialization throughout training. At epoch 0, points corresponding to question types (mathematics, history, facts) appear overlapped or close in the reduced space, indicating an initial uniform distribution. As epochs advance, clusters separate progressively: at epoch 10, an initial differentiation is observed, consolidating at epoch 20 and reaching maximum clarity at epoch 40, with well-delimited points by category. This pattern is consistently reproduced in all evaluated MoE configurations, independent of hyperparameter variations, and is not observed in the ablation without experts. This specialization emerges without explicit supervision on question types, as the classification is used solely for posterior visualization. Consistency across multiple sweeps and initialization seeds confirms it is not a transient effect. Finally, a key finding is the robustness of the complete system to imperfect retrieval in Phase 1. Despite low recall@5 in evaluation (0.01), Phase 2 generates significant semantic similarities (up to 0.6), with clear improvements attributed to expert routing. This compensation is verified by comparing memory output with the final reasoning in independent runs.



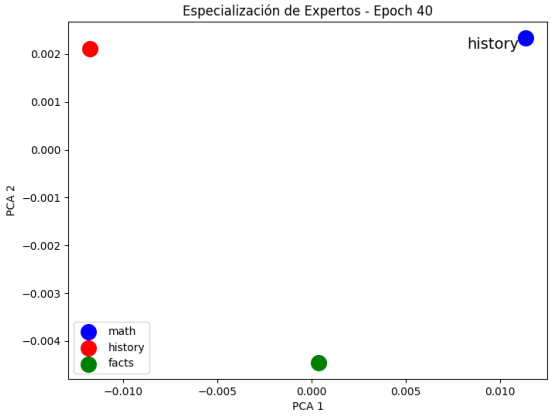
(a) Epoch 0: Initial uniform distribution.



(b) Epoch 10: Initial cluster differentiation.



(c) Epoch 20: Consolidation of separation by category.



(d) Epoch 40: Well-delimited clusters (mathematics, history, facts).

Figure 4: Collage of the evolution of emergent specialization via PCA of routing gates throughout training. A clear progression from initial overlap to semantic differentiation is observed, confirming implicit learning of topological structure.

VI. Discussion

The results obtained in the preliminary experiments reveal consistent patterns that validate the viability of the proposed modular architecture, despite the reduced scale of the prototype. Below, these findings are analyzed in depth, observed limitations are identified, and hypotheses derived from the data are formulated, along with suggestions for future experiments to validate

or refine them. The focus is on the bio-inspired interpretation of emergent behaviors, highlighting how the strict separation between phases induces desirable properties in the complete system.

A. Analysis of Main Findings

The stable convergence of the loss in Phase 2, observed in all configurations explored through sweeps, indicates that the reasoning module optimizes effectively over the retrieved representations from memory, even when these are approximate or noisy. This stability is maintained regardless of hyperparameter variations such as the number of experts or top-k, suggesting inherent robustness in the modular design. The progressive improvement in average semantic similarity (close_answer_mean), ascending to values of 0.6 in the best Mixture of Experts configurations, demonstrates that the system is capable of generating coherent responses despite suboptimal memory retrieval in Phase 1 (approximate recall@5 of 0.01 in evaluation with noise). This compensation capacity is particularly notable compared to the ablation, where the metric stalls at lower values, confirming the critical role of dynamic routing and expert specialization in processing imperfect information. On the other hand, the exact match metric remains low in all variants, which is consistent with the strict 0.9 threshold applied to cosine similarities in embeddings. This behavior does not indicate a model failure, but the inherent approximate nature of vector representations in open-question datasets, where semantically similar responses may differ in superficial details. Regarding computational efficiency, the GPU metrics recorded during sweeps show low and stable consumption, with average utilization of 50-60% and power in ranges of 30-40 W. These values remain constant across multiple runs, regardless of configuration complexity (e.g., number of experts), validating the effectiveness of sparsity mechanisms induced by top-k in both phases to reduce computation without compromising convergence. Finally, the visualization via PCA of average routing gate scores reveals a progressive evolution toward well-defined semantic clusters by question type. This pattern is reproduced consistently in all MoE configurations, emerging without explicit supervision on categories, suggesting implicit learning of topological structure in the representation space.

B. Observed Limitations

Despite the promising results, limitations inherent to the preliminary scope of the prototype are identified. The overfitting observed in Phase 1 evaluation, with low recall under noise conditions, is mainly attributed to the reduced dataset size (10,000 SQuAD samples), which limits the diversity of stored patterns and generalization. This restriction is partially mitigated in Phase 2 through robustness induced by freezing, but persists as a factor affecting final metrics. The low exact match metric reflects the chosen strict threshold, suitable for precise evaluations but sensitive to minor variations in embeddings, especially in long contextual responses. Additionally, the model operates exclusively in fixed embedding spaces (384 dimension), facilitating efficiency but potentially limiting expressiveness compared to more flexible token-to-token models. The efficient computational consumption is achieved thanks to sparsity, but in setups with a higher number of experts or layers, additional optimizations might be required to maintain accessibility in low-resource environments. These limitations do not invalidate the findings but highlight areas for refinement in future iterations.

C. Hypotheses Derived from the Results

Based on the observed patterns, the following hypotheses are formulated, explaining emergent behaviors and guiding future research directions. These are derived directly from empirical data, such as the improvement in semantic similarity and the evolution of PCA clusters. A central hypothesis is that, in Phase 2, the Mixture of Experts structure is forced to extract deeper relationships between concepts by not having shortcuts or direct answers from perfect memory. Since the memory module provides approximate or noisy retrieved vectors, experts must learn distributed representations that compensate for these imperfections, fostering implicit semantic specialization observable in PCA visualizations. This dynamic does not arise in the dense MLP ablation, suggesting that dynamic routing is key to forcing this deep relational extraction. Another hypothesis is that modular separation and freezing induce inherent robustness to retrieval noise, similar to how biological systems compensate for partial memories through distributed inference. The results show that, despite low recall, final semantic similarity improves consistently, indicating that expert reasoning acts as an adaptive correction mechanism. Finally, it is hypothesized that emergent specialization in semantic clusters (mathematics, history, facts) arises from the interaction between topological routing and implicit classification in the dataset, without the need for explicit supervision. The progressive evolution of clusters across epochs suggests gradual learning of topology in the representation space.

D. Future Experiments to Validate the Hypotheses

To validate and refine these hypotheses, the following future experiments are proposed, designed to scale the prototype and control specific variables. Regarding the hypothesis of deep relationship extraction in Phase 2, a key experiment would consist of scaling the dataset to larger and more diverse sets (e.g., full SQuAD or combinations with datasets like TriviaQA or HotpotQA), measuring whether the improvement in semantic similarity is maintained or increases with greater pattern variety. Additionally, controlled noise could be introduced in the memory output during Phase 2 training, evaluating whether this further forces deep

specialization and improves generalization, comparing with variants without noise. For the hypothesis of robustness induced by modularity, future experiments could include validation on multi-step reasoning benchmarks (such as GSM8K for mathematics or historical datasets), quantifying the system's capacity to handle queries requiring inference over imperfect retrieval. It is also suggested to compare with non-modular end-to-end architectures trained on the same data, measuring forgetting metrics and noise robustness to confirm the freezing advantage. Regarding the hypothesis of emergent specialization, future works could incorporate advanced interpretability analysis techniques (such as SHAP or visualized attention) to break down how topological routing contributes to cluster formation. Additionally, scaling the number of question categories (adding types like "science" or "geography") and evaluating PCA evolution on more complex datasets would allow verifying if the emergence is scalable and not limited to the current dataset. These experiments would not only validate the hypotheses but also extend the prototype toward practical applications, such as integration with token-to-token generation or deployment on low-resource devices.

VII. Conclusions

The preliminary results obtained in this compact prototype validate the viability of the proposed bio-inspired modular architecture. The strict separation between associative memory (implemented through a modern Hopfield network) and dynamic reasoning (based on Mixture of Experts) allows achieving advanced behaviors with fewer than 15 million trainable parameters in total, executed exclusively on free resources like Google Colab. The main conclusions derived from the experiments are as follows:

- The architecture converges stably in Phase 2, showing progressive improvement in response semantic similarity (close_answer_mean up to 0.6) despite imperfect retrieval in Phase 1 (approximate recall@5 of 0.01 in evaluation with noise).
- Mixture of Experts configurations consistently outperform the dense MLP ablation in the close_answer_mean metric, with systematic differences of up to 0.15 points observed across multiple sweeps.
- A clear emergent specialization of experts occurs, visualized through PCA analysis of routing scores, evolving from overlapped points in initial epochs to well-defined clusters by question type (mathematics, history, facts) in advanced epochs.
- Computational consumption remains low and stable (average GPU utilization 50-60%, power 30-40 W), confirming the efficiency induced by top-k sparsity mechanisms in both phases.

These findings demonstrate that a modular design with phased training and freezing can induce robustness to imperfect inputs and semantic emergence in very limited resource configurations. The reasoning module's capacity to compensate for memory limitations, along with the observed specialization without explicit supervision, suggest that bio-inspired principles such as functional separation and conditional routing are promising for the development of more efficient and resilient artificial intelligence systems. The complete source code, along with wandb logs of all performed sweeps, is publicly available in an open-source repository to facilitate experiment reproduction and future collaboration. This preliminary work opens interesting avenues toward architectures that combine efficient associative memory with distributed reasoning, with potential application in low-resource environments or hybrid systems. The community is invited to explore extensions of the prototype, such as scaling to larger datasets, incorporating token-to-token generation, or implementing a third decision phase with internal loops, to continue investigating the emergent properties of bio-inspired modular designs.

Acknowledgments

This prototype was developed in 12 days using exclusively free resources such as Google Colab and the assistance of the Grok model from xAI for code debugging and writing. The author is a 23-year-old self-taught researcher without formal university training in artificial intelligence.

References

[1] D. Krotov and J. J. Hopfield, "Dense associative memory for pattern recognition," *Advances in neural information processing systems*, vol. 29, 2016.
[2] S. Ramasinghe, D. Zoran, B. Leibe, J. Eriksen, and L. Van Gool, "Modern hopfield networks for few- and zero-shot learning," *arXiv preprint arXiv:2207.05345*, 2022.
[3] D. Krotov and J. Hopfield, "Large associative memory problem in neurobiology and machine learning," *arXiv preprint arXiv:2008.06996*, 2020.
[4] M. Rizzuto, H. Ramsauer, M. Weissensteiner, B. Schäfl, J. Lehner, M. Widrich, S. Hochreiter, and G. Klambauer, "Modern hopfield networks for clustering and outlier detection," *arXiv preprint arXiv:2401.12485*, 2024.
[5] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.
[6] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.
[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2383–2392.