From Noise to Memory: Visualizing Internal Memory Formation in the Dragon Hatchling (BDH) Architecture

Track: Frontier Exploration (Track 2)
Team: *Unseen Geeks*

Team Members

- Satish Vaishyar
- Ayush Chugani
- Srujan Mattur
- Aditya Bajantri

Primary Contact Details

- **Satish Vaishyar**
  +91 9916558399
- **Ayush Chugani**
  +91 8151976632

Abstract

Transformer-based neural architectures rely on static internal representations and external mechanisms such as retrieval systems to approximate memory. While effective for many tasks, this design limits interpretability and prevents direct observation of how new information is internalized. The Baby Dragon Hatchling (BDH) architecture proposes an alternative approach by incorporating sparse neural activations and Hebbian synapse updates to support internal, continuous learning.

In this work, we present a controlled experimental framework to study and visualize how memory forms within BDH. Using synthetic concept probes and systematic logging of neuron activations and synapse weights, we analyze the emergence, stabilization, and persistence of internal representations during training. Rather than optimizing for task performance, this study focuses on interpretability and mechanistic understanding. Our results provide qualitative evidence of memory formation through sparse structural reallocation and Hebbian consolidation, offering insight into BDH's learning dynamics.

1. Introduction

Memory in contemporary large language models is largely externalized. Transformer architectures do not modify internal parameters during inference, and any form of long-term memory is achieved through techniques such as retrieval-augmented generation, fine-tuning, or extended context windows. While these approaches are effective, they obscure the internal learning process and limit interpretability.

The Dragon Hatchling (BDH) architecture explores a different paradigm. By employing sparse neuron activations and Hebbian learning rules, BDH allows internal synaptic structure to evolve during training. This design raises an important research question:

Can internal memory formation be directly observed and analysed within a neural architecture?

This project addresses that question by building a minimal, controlled experimental setup focused on making BDH's internal learning dynamics visible.

2. Motivation and Problem Statement

Although BDH claims support for internal memory formation, such behavior is often discussed at a conceptual level. There is limited tooling or methodology to directly observe how new information is internalized within the model.

The core challenges are:

- Internal representations are difficult to inspect during learning
- Memory formation is often inferred indirectly through outputs
- Existing evaluations prioritize performance over interpretability

Problem Statement:
There is a need for a controlled, interpretable experimental framework that allows direct observation of internal memory formation in BDH during learning.

3. Proposed Approach

We propose a **concept-probe–based experimental framework** designed to isolate and observe internal learning behaviour in BDH.

Key Principles

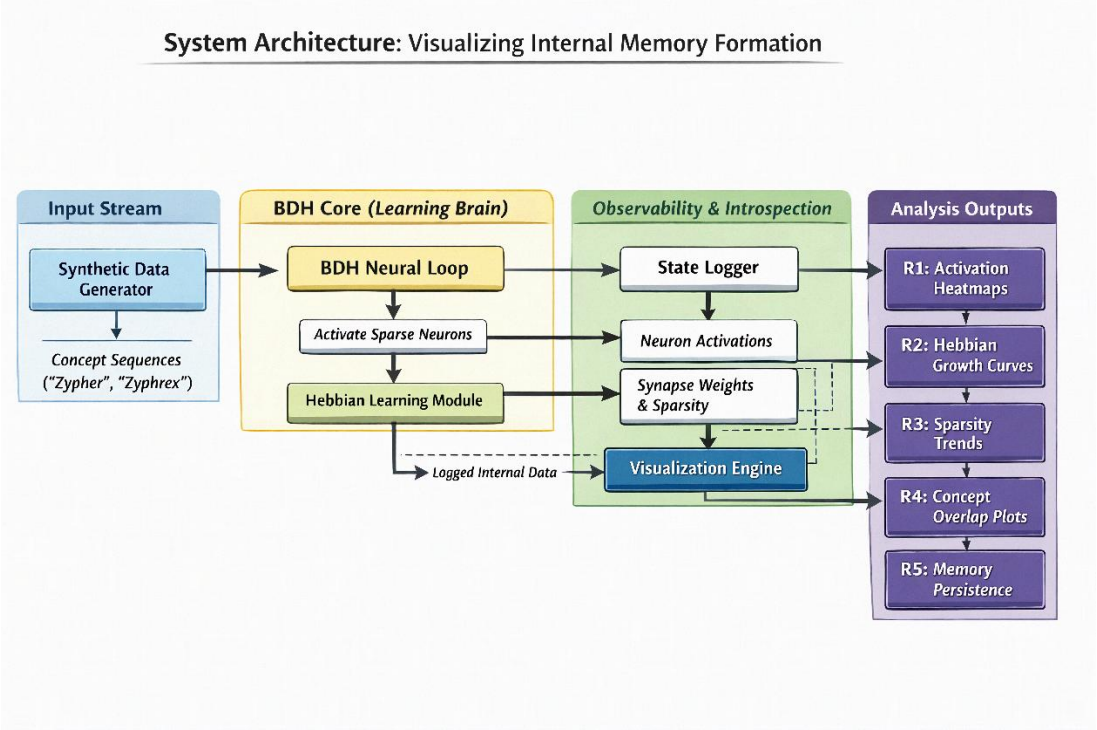
- Focus on internal signals rather than output accuracy
- Use synthetic concepts with no prior semantic meaning
- Log and visualize learning dynamics over time

High-Level Idea

By introducing artificial concepts and tracking how BDH's internal structure changes in response, we can visualize the lifecycle of memory formation:

1. Initial noise
2. Structural emergence
3. Stabilization
4. Persistence

5. Experimental Design



4.1 Synthetic Concept Probes

We generate synthetic tokens such as *"Zypher"* and *"Zyphrex"*, embedded in simple sentence templates. These concepts are intentionally meaningless outside the experiment, ensuring that any internal structure that forms is the result of learning rather than prior knowledge.

4.2 BDH Training Loop

The BDH model operates with:

- Sparse neuron activation per input
- Hebbian learning updates that strengthen synapses between co-active neurons

Training proceeds iteratively, with learning occurring through synaptic updates rather than global parameter re-optimization.

4.3 Observability and Logging

At fixed training intervals, we log:

- Neuron activation frequencies
- Synapse weight distributions
- Overall activation sparsity

These internal signals form the basis for all subsequent analysis.

5. Results and Observations

R1: Emergence of Concept-Specific Activation

Early training phases show diffuse and noisy neuron activations. As training progresses, a small subset of neurons consistently activates for a given concept, indicating internal allocation of representational capacity.

R2: Hebbian Synapse Strengthening

A limited number of synapses exhibit significant weight growth over time, while most remain near baseline values. This behavior is consistent with Hebbian learning principles and suggests consolidation of internal memory pathways.

R3: Sparsity Stability

Despite learning, the proportion of inactive neurons remains stable. This indicates that learning occurs via **reallocation of existing neural resources**, not through increased activation density.

R4: Partial Concept Separation

Related synthetic concepts display partial overlap in active neuron sets alongside emerging concept-specific neurons. This suggests early-stage specialization and indicates that full separation may be dependent on model capacity.

R5: Memory Persistence

After training is halted, learned concepts continue to activate the same neuron subsets during inference, providing evidence of persistent internal memory.

6. Discussion

The observed results support the hypothesis that BDH forms internal memory through sparse structural changes rather than output-level memorization. Importantly, learning dynamics are visible and interpretable, offering a perspective that is difficult to obtain in transformer-based systems.

The presence of partial concept overlap highlights an important insight: **monosemantic representations may be capacity-dependent rather than guaranteed**, especially at small scales.

7. Limitations

This study is exploratory in nature and subject to several limitations:

- Experiments are conducted at small model scale
- Results are qualitative rather than benchmark-driven
- Synthetic concepts do not reflect real-world semantic complexity
- Findings should not be interpreted as performance claims

These constraints are intentional to preserve interpretability and experimental control.

8. Future Scope

This work opens several avenues for future research:

1. **Scaling Studies**
   Investigating how memory formation and concept separation evolve with increased model capacity.
2. **Comparative Analysis**
   Applying the same concept-probe framework to transformer architectures to contrast internal learning behavior.
3. **Long-Term Continual Learning**
   Studying how BDH handles sequential learning of many concepts without catastrophic interference.
4. **Real-World Domains**
   Extending probes to structured domains such as code tokens or scientific terminology.
5. **Interactive Visualization Tools**
   Developing real-time dashboards for monitoring internal learning dynamics during training.

9. Conclusion

This project presents a controlled experimental framework for visualizing internal memory formation in the BDH architecture. By shifting the evaluation focus from output performance to internal dynamics, we provide qualitative evidence of learning through sparse structural reallocation and Hebbian consolidation. We believe this approach contributes to a clearer understanding of continuous learning mechanisms and aligns with ongoing research into interpretable and adaptive neural architectures.

10. Reproducibility

All experiments are implemented in a modular and reproducible manner. The complete pipeline—from data generation to visualization—is available through the accompanying public repository and can be executed with fixed random seeds.

11. References

1. **Vaswani, A., Shazeer, N., Parmar, N., et al.**
   Attention Is All You Need.
   Advances in Neural Information Processing Systems (NeurIPS), 2017.
2. **Hebb, D. O.**
   The Organization of Behavior: A Neuropsychological Theory.
   Wiley, 1949.
3. **Chorowski, J., Kaiser, Ł., Kosowski, A., et al.**
   The Dragon Hatchling: A Post-Transformer Architecture for Continuous Learning.
   arXiv preprint, 2025.

4. **Olshausen, B. A., Field, D. J.**
Sparse Coding of Sensory Inputs.
Nature, 1996.

5. **Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al.**
Overcoming Catastrophic Forgetting in Neural Networks.
Proceedings of the National Academy of Sciences (PNAS), 2017.

6. **Bengio, Y., Simard, P., Frasconi, P.**
Learning Long-Term Dependencies with Gradient Descent Is Difficult.
IEEE Transactions on Neural Networks, 1994.

7. **Marcus, G.**
Deep Learning: A Critical Appraisal.
arXiv preprint, 2018.