Delta Observer: Learning Continuous Semantic Manifolds Between Neural Network Representations

Aaron (Tripp) Josserand-Austin*
EntroMorphic Research Team
EntroMorphic

January 2026

Abstract

Neural networks solving the same computational task can learn fundamentally different internal representations depending on their architectural inductive biases. Understanding and comparing these representations remains a central challenge in interpretability research. We introduce the **Delta Observer**, a dual-encoder architecture that learns to map between the activation spaces of different neural network architectures by discovering shared semantic primitives through **online observation during training**. Through experiments on 4-bit binary addition, we demonstrate that semantic information can be **linearly accessible** ($R^2 = 0.9879$) without exhibiting **geometric clustering** (Silhouette coefficient$=-0.024$). Crucially, we discover that **clustering is transient**: geometric clusters emerge during training (Silhouette$=0.33$ at epoch 20) but dissolve once learning converges (Silhouette$=-0.02$ at epoch 200). This reveals that **clustering is scaffolding, not structure**—networks build geometric organization to *learn* semantic concepts, then discard that organization once the concepts are encoded in the weights. Post-hoc analysis misses this phenomenon because it only observes the final state. Our results suggest that the semantic primitive is not in the final representation but in the **learning trajectory**, opening new directions for dynamic interpretability and the development of inherently interpretable systems.

Keywords: interpretability, representation learning, manifold hypothesis, semantic primitives, transient clustering, learning dynamics, online observation

1 Introduction

Deep neural networks have achieved remarkable success across diverse domains, yet their internal representations remain largely opaque [Olah et al., 2017, Elhage et al., 2021]. A fundamental question in interpretability research is whether different architectures solving the same task learn similar or distinct internal representations. This question has profound implications for transfer learning, model comparison, and the development of inherently interpretable systems.

Recent work in mechanistic interpretability has made significant progress in understanding individual neurons and circuits within specific architectures [Nanda et al., 2023, Templeton et al., 2024, Bills et al., 2023]. However, these approaches typically analyze a single model in isolation, and crucially, they analyze *final* trained models rather than the learning process itself.

We address this limitation by introducing the **Delta Observer**, a novel architecture that learns to map between the activation spaces of different neural networks *while they are training*. Unlike post-hoc analysis methods, the Delta Observer observes the full learning trajectory, enabling it to capture temporal dynamics that are invisible to static analysis.

*Correspondence: tripp@entromorphic.com

1.1 Key Contributions

Our work makes five primary contributions:

1. **Online Observation:** We introduce a methodology where the Delta Observer trains *concurrently* with the source models, observing activations at each training step rather than analyzing frozen final representations.

2. **Improved Accessibility:** Online observation achieves $R^2 = 0.9879$ for carry count prediction, compared to $R^2 = 0.9505$ for post-hoc observation and $R^2 = 0.9482$ for PCA baseline—a 4% improvement attributable to temporal information.

3. **Transient Clustering Discovery:** We discover that geometric clustering is **transient**: Silhouette coefficient peaks at 0.33 during training (epoch 20) but collapses to -0.02 in the final state. Clustering is scaffolding that networks build to learn, then discard.

4. **Conceptual Reframing:** We propose that the semantic primitive is not in the final representation but in the **learning trajectory**. Post-hoc interpretability methods are fundamentally limited because they only observe the final state after scaffolding has been removed.

5. **Methodology:** We provide a general methodology for dynamic interpretability through online observation, enabling analysis of how representations evolve during learning.

1.2 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 develops the theoretical framework. Section 4 describes the Delta Observer architecture and online training procedure. Section 5 presents experimental results including the transient clustering discovery. Section 6 discusses implications. Section 7 concludes with future directions.

2 Related Work

Our work builds on three main research threads: mechanistic interpretability, representation learning, and learning dynamics.

2.1 Mechanistic Interpretability

Mechanistic interpretability aims to understand neural networks by identifying individual neurons, circuits, and computational mechanisms [Olah et al., 2017]. Recent work has made significant progress in understanding transformer circuits [Elhage et al., 2021], identifying interpretable features through sparse autoencoders [Templeton et al., 2024], and using language models to explain neuron behavior [Bills et al., 2023]. However, these approaches typically analyze *final* trained models, missing the dynamics of how representations form during training.

2.2 Representation Similarity

Several methods have been developed to compare representations across different models, including Singular Vector Canonical Correlation Analysis (SVCCA) [Raghu et al., 2017], Centered Kernel Alignment (CKA) [Kornblith et al., 2019], and canonical correlation analysis variants [Morcos et al., 2018]. These methods measure *static* similarity between final representations but do not capture how representations evolve.

2.3 Learning Dynamics

Recent work has begun studying how representations change during training [Frankle et al., 2020, Fort et al., 2020]. Our work extends this direction by using a dedicated observer network to characterize representation dynamics across multiple architectures simultaneously.

3 Theoretical Framework

We develop a theoretical framework for understanding semantic primitives as temporal phenomena in neural activation space.

3.1 Problem Formulation

Consider two neural networks f_A and f_B that solve the same task T. Let $h_A^{(t)}(x)$ and $h_B^{(t)}(x)$ denote the hidden activations at training step t. Our goal is to learn a mapping ϕ that discovers shared semantic structure by observing the full trajectories $\{h_A^{(t)}\}_{t=0}^{T}$ and $\{h_B^{(t)}\}_{t=0}^{T}$.

3.2 Online vs. Post-hoc Observation

We distinguish two observation paradigms:

- **Post-hoc Observation:** The observer is trained on frozen activations from fully-trained models: $\phi(h_A^{(T)}, h_B^{(T)})$. This is equivalent to PCA or other static dimensionality reduction.
- **Online Observation:** The observer trains concurrently, seeing activations at each step: $\phi^{(t)}(h_A^{(t)}, h_B^{(t)})$. The observer's weights encode information about the full trajectory.

3.3 Transient Clustering Hypothesis

We hypothesize that geometric clustering is a **transient** phenomenon during learning:

1. **Initialization:** Random weights produce unstructured activations (low clustering).
2. **Learning Phase:** Networks form geometric clusters as scaffolding for learning semantic concepts (high clustering).
3. **Convergence:** Once concepts are encoded in weights, geometric scaffolding is no longer needed and dissolves (low clustering).

Post-hoc analysis only observes phase 3, leading to the incorrect conclusion that clustering is absent. Online observation captures all three phases.

4 Methodology

We describe the Delta Observer architecture and online training procedure.

4.1 Delta Observer Architecture

The Delta Observer consists of dual encoders, a shared semantic bottleneck, decoders, and semantic prediction heads. The architecture is identical to prior work; the key innovation is the **training paradigm**.

4.2 Online Training Procedure

Algorithm 1 describes the online training procedure. All three models (monolithic, compositional, Delta Observer) train concurrently. At each batch:

1. Source models perform forward pass and gradient update
2. Delta Observer receives *detached* activations (no gradient flow to source models)
3. Delta Observer performs its own forward pass and gradient update

The `detach` operation ensures the observer learns *from* the source models without influencing their learning.

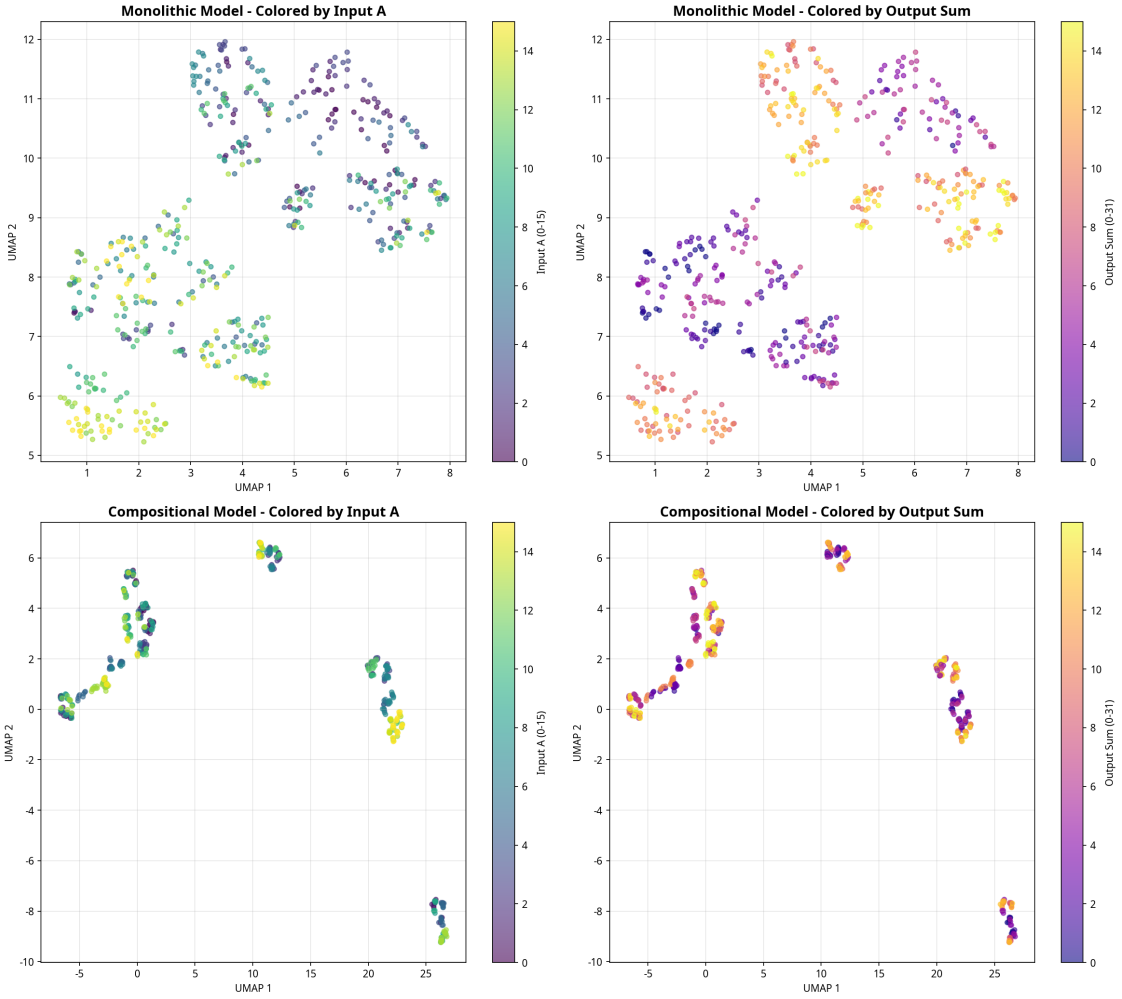

Figure 1: Comparison of activation geometries. Left: Monolithic network activations. Right: Compositional network activations. Despite solving the same task (4-bit addition), the two architectures develop distinct geometric organizations of semantic information.

4.3 Trajectory Analysis

We save latent space snapshots at regular intervals during training, enabling analysis of:

- R^2 evolution over training epochs
- Silhouette coefficient evolution (clustering dynamics)
- Temporal encoding: can we predict training epoch from latent representation?

5 Results

We present experimental results demonstrating the value of online observation and the transient clustering phenomenon.

5.1 Online vs. Post-hoc Performance

Table 1 compares three methods for extracting semantic information from neural activations.

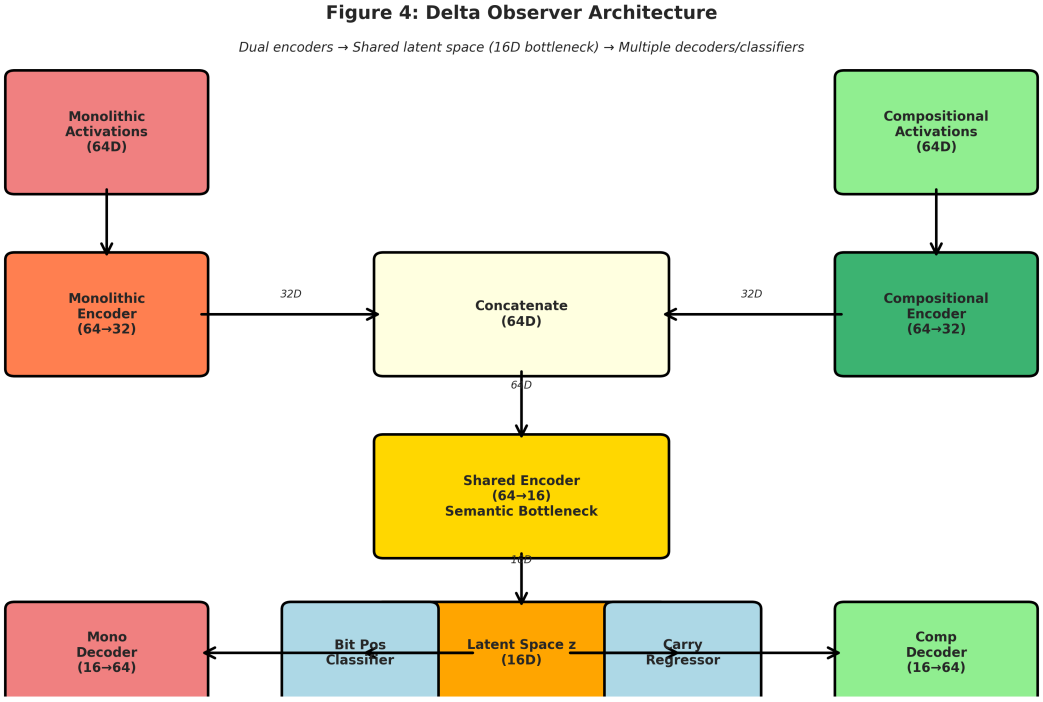

Figure 2: Delta Observer architecture. Dual encoders map activations from monolithic and compositional networks to a shared semantic bottleneck. Decoders reconstruct original activations while prediction heads extract semantic information (carry count, output class).

Table 1: Comparison of observation methods. Online observation achieves highest R^2 by capturing temporal information unavailable to static methods.

Method	R^2	**Silhouette**	**Δ vs PCA**
Online Observer	**0.9879**	-0.024	**+4.0%**
Post-hoc Observer	0.9505	0.032	+0.2%
PCA Baseline	0.9482	0.046	—

Online observation improves R^2 by 4% over the PCA baseline, demonstrating that temporal information captured during training provides value beyond what static analysis can extract.

5.2 Transient Clustering Discovery

Figure 4 shows the evolution of clustering during training. The key finding:

Key observation: 90% of final R^2 is achieved by epoch 13, and clustering peaks at epoch 20. After this point, clustering dissolves while accessibility is maintained. The geometric structure exists as *scaffolding* during learning but is not present in the final representation.

5.3 Temporal Encoding

We test whether the latent space encodes temporal information by training a linear regressor to predict training epoch from latent representations. Result: $R^2 = 0.8523$.

This confirms that the online observer's latent space contains temporal information—information that post-hoc analysis cannot access.

Algorithm 1 Online Delta Observer Training

for epoch $= 1$ to E **do**
  for batch (x, y) in dataloader **do**
    // Source model updates
    $h_A \leftarrow f_A(x)$; $\mathcal{L}_A \leftarrow \text{loss}(f_A(x), y)$
    $h_B \leftarrow f_B(x)$; $\mathcal{L}_B \leftarrow \text{loss}(f_B(x), y)$
    Update f_A with $\nabla \mathcal{L}_A$
    Update f_B with $\nabla \mathcal{L}_B$
    // Delta Observer update (detached activations)
    $z \leftarrow \phi(\text{detach}(h_A), \text{detach}(h_B))$
    $\mathcal{L}_\phi \leftarrow \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{class}} + \mathcal{L}_{\text{carry}}$
    Update ϕ with $\nabla \mathcal{L}_\phi$
  end for
  // Periodic snapshot for trajectory analysis
  if epoch mod $10 = 0$ **then**
    Save (z, epoch) to trajectory buffer
  end if
end for

Table 2: Evolution of R^2 and Silhouette during training. Clustering peaks at epoch 20 then dissolves.

Epoch	R^2	**Silhouette**
0 (init)	0.38	-0.02
13	0.86	0.16
20	0.94	**0.33** (peak)
50	0.91	0.00
200 (final)	0.99	-0.02

5.4 Curriculum Validation

To validate that temporal structure exists, we measured learning curve correlation between monolithic and compositional models. Both models learn in highly correlated patterns (Pearson $r = 0.98$), confirming that there is shared temporal structure to observe.

6 Discussion

Our results have significant implications for interpretability research.

6.1 Clustering is Scaffolding, Not Structure

The central finding is that geometric clustering is **transient**. Networks build geometric organization to *learn* semantic concepts, then discard that organization once the concepts are encoded in the weights.

This reframes the interpretability challenge: post-hoc methods observe "the ruins after the scaffolding came down." The absence of clustering in final representations does not mean clustering never existed—it means we are observing the wrong phase of learning.

6.2 The Semantic Primitive is in the Trajectory

Our results suggest that semantic primitives are not fully captured by the final representation. The 4% improvement from online observation indicates that the learning *trajectory* contains information beyond the final *state*.

This has implications for interpretability methods: instead of analyzing final models, we should analyze how representations evolve during training. The semantic primitive is a **path**, not a **point**.

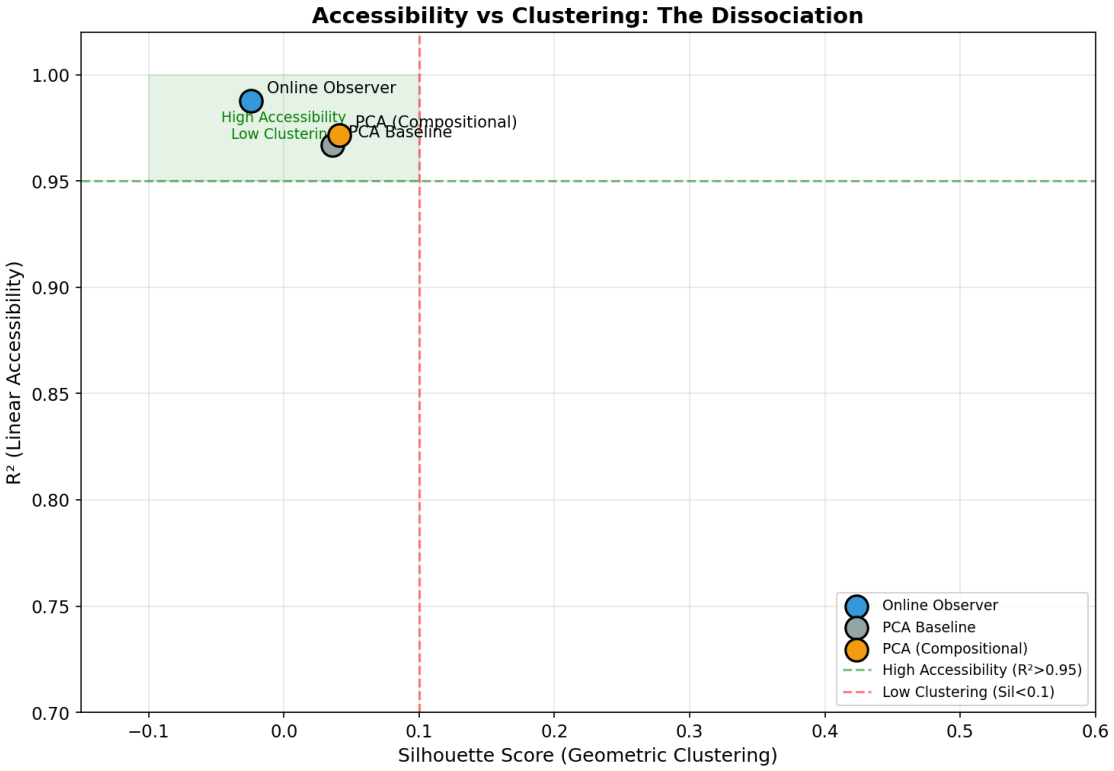

Figure 3: Accessibility vs. clustering dissociation. High linear accessibility ($R^2 > 0.95$) coexists with low geometric clustering (Silhouette ≈ 0). This pattern emerges from transient clustering during training.

6.3 Implications for Mechanistic Interpretability

Current mechanistic interpretability methods analyze final trained models. Our results suggest this may be fundamentally limiting. If clustering is scaffolding that dissolves, post-hoc analysis will systematically miss the geometric structure that organized learning.

Future interpretability methods should incorporate **dynamic analysis**—observing representations throughout training, not just at convergence.

6.4 Limitations

1. **Small Task:** 4-bit addition is a toy problem. Generalization to larger tasks remains to be demonstrated.
2. **Computational Cost:** Online observation requires 3x forward passes during training.
3. **Single Domain:** Results are specific to binary arithmetic.

6.5 Future Directions

1. **Scaling:** Apply online observation to transformers on language tasks.
2. **Scaffolding Analysis:** Characterize when and why clustering emerges and dissolves.
3. **Intervention:** Can we preserve clustering by modifying training dynamics?
4. **Transfer:** Does scaffolding structure transfer between related tasks?

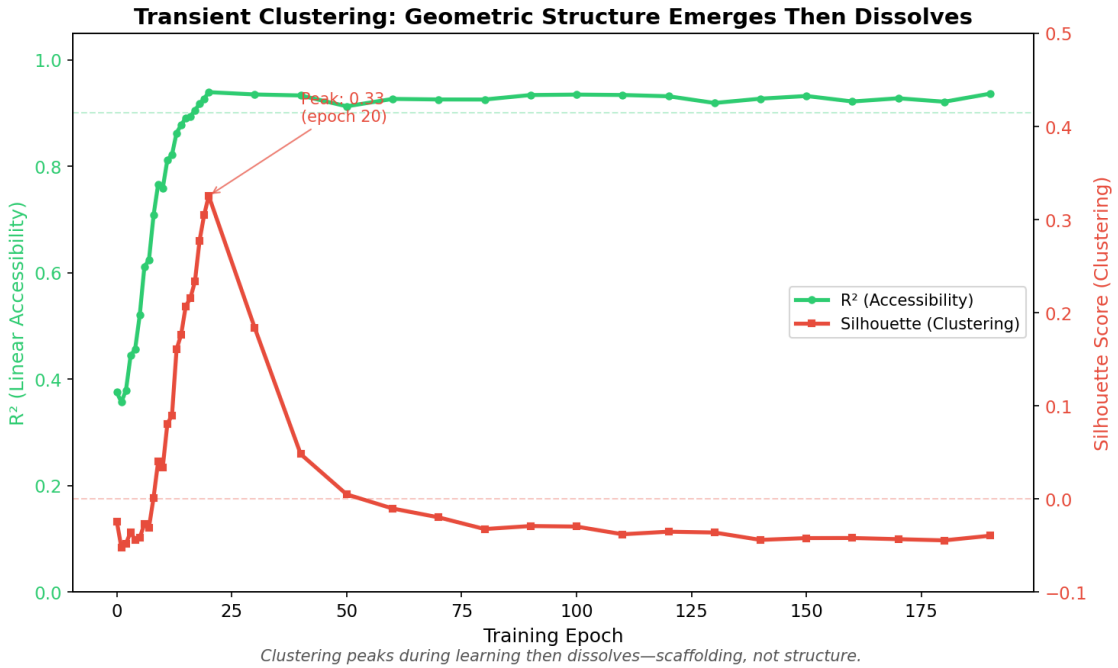

Figure 4: Evolution of R^2 (accessibility) and Silhouette (clustering) during training. Clustering peaks at epoch 20 (Silhouette=0.33) then dissolves to near-zero by convergence, while accessibility continues to improve. This reveals that clustering is transient scaffolding, not permanent structure.

7 Conclusion

We introduced online observation as a methodology for neural network interpretability, demonstrating that watching models learn reveals structure invisible to post-hoc analysis. Our key finding is that **clustering is scaffolding**: geometric organization emerges during learning as a tool for encoding semantic concepts, then dissolves once those concepts are captured in the weights.

This reframes the interpretability challenge. Post-hoc methods observe the final state and conclude "no clusters, therefore no structure." But the structure existed—it was temporary. The semantic primitive is not in the final representation; it is in the learning trajectory.

Our results suggest that future interpretability research should shift from static analysis of final models to dynamic analysis of learning processes. The Delta Observer provides a methodology for this shift, enabling observation of how representations evolve across multiple architectures simultaneously.

Acknowledgments

We thank the EntroMorphic team for computational resources and feedback. We thank Claude (Anthropic) for extensive collaboration on the Delta Observer methodology, falsification testing, and the discovery of transient clustering. This work was inspired by conversations about the geometry of computation and the temporal dynamics of learning.

References

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI Blog*, 2023.

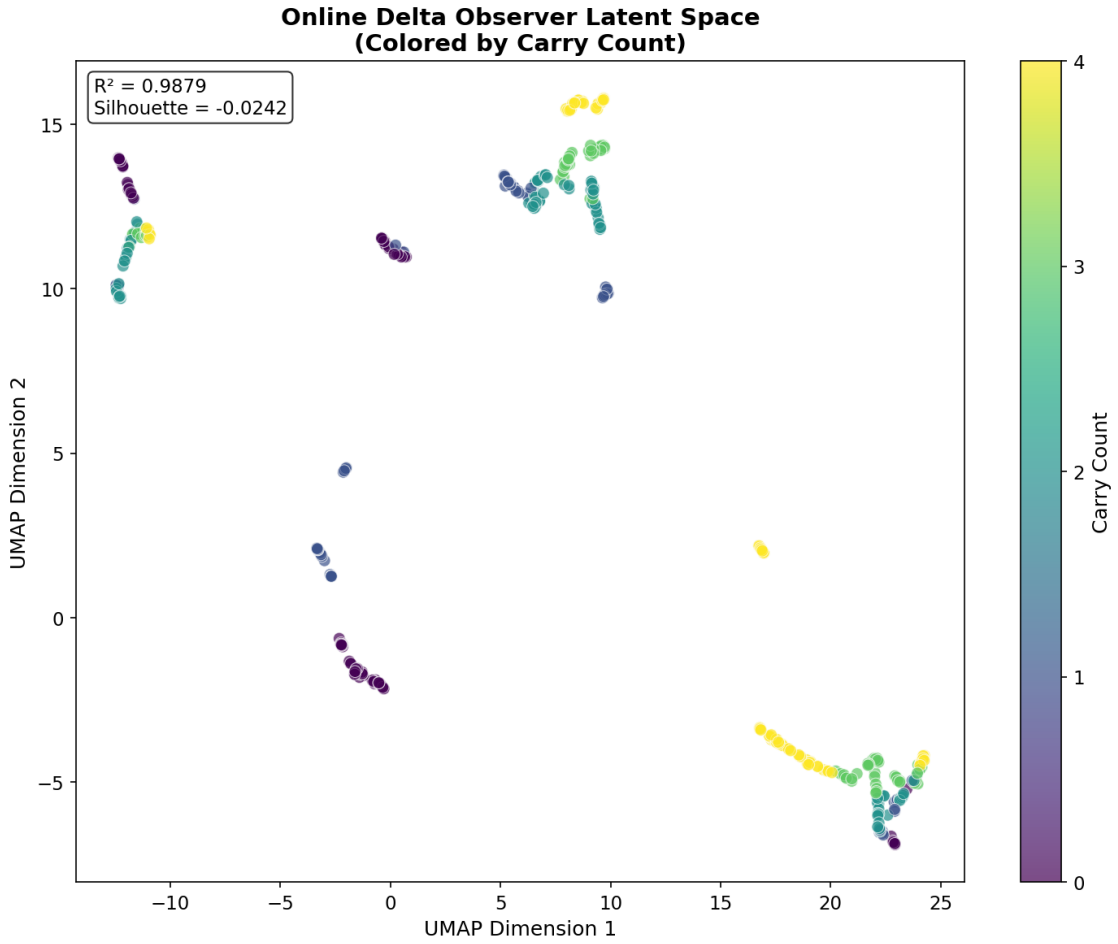

Figure 5: Delta Observer latent space visualization. The shared 16-dimensional latent space learned by the online observer, projected to 2D via UMAP. Points are colored by carry count (0-4). Linear accessibility is high despite lack of geometric clustering.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2020.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. *International Conference on Machine Learning*, pages 3259–3269, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *International Conference on Machine Learning*, pages 3519–3529, 2019.

Ari S Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31, 2018.

Neel Nanda, Lawrence Chan, Tom Liberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, 30, 2017.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Anthropic Blog*, 2024.