Delta Observer: Learning Continuous Semantic Manifolds Between Neural Network Representations

From Geometric Divergence to Linear Semantic Accessibility

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

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 experiments on 4-bit binary addition, we train two architectures—a monolithic multi-layer perceptron and a compositional modular network—that both achieve perfect task performance while learning geometrically distinct representations. We then train the Delta Observer to map between these two representation spaces. Remarkably, we find that semantic information can be **linearly accessible** ($R^2 = 0.9505$ for carry count prediction) without exhibiting strong **geometric clustering** (Silhouette coefficient $= 0.0320$), challenging the prevailing assumption that interpretability requires discrete, spatially separated feature clusters. Our results suggest semantic primitives are better characterized as continuous gradients in latent space than as discrete categorical labels—challenging prevailing assumptions in mechanistic interpretability. By training the Delta Observer on transparent "glass box" models to supervise opaque "black box" ones, we open a scalable path toward cross-architectural interpretability and inherently understandable systems.

Keywords: interpretability, representation learning, manifold hypothesis, semantic primitives, compositional generalization, neural network geometry

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, leaving open the question of how representations vary across different architectural choices. This gap is particularly acute in tasks requiring compositional generalization, where architectural inductive biases can lead to geometrically divergent yet functionally equivalent solutions. When two networks achieve identical task performance, do they discover the same underlying computational structure, or do they find fundamentally different solutions?

We address this question by introducing the **Delta Observer**, a novel architecture that learns to map between the activation spaces of different neural networks. The Delta Observer employs a dual-encoder design with a semantic bottleneck, forcing the discovery of shared representational primitives that exist

*Correspondence: tripp@entromorphic.com

across architectures. By training the Delta Observer to simultaneously reconstruct both activation spaces and predict semantic properties, we can characterize the geometric structure of the learned latent space and assess whether semantic information is encoded in a discrete or continuous manner.

1.1 Key Contributions

Our work makes four primary contributions:

1. **Novel Architecture:** We introduce the Delta Observer, a dual-encoder architecture with a semantic bottleneck that learns shared representational primitives across different neural network architectures.
2. **Empirical Finding:** We demonstrate that semantic information can be **linearly accessible** without **geometric clustering**. Specifically, carry count in 4-bit addition can be predicted with $R^2 = 0.9505$ from a 16-dimensional latent space that exhibits minimal clustering (Silhouette coefficient=0.0320).
3. **Conceptual Framework:** We propose that semantic primitives in neural networks are better characterized as **continuous gradients** in latent space rather than discrete categorical labels, challenging the prevailing assumption that interpretability requires spatially separated feature clusters.
4. **Methodology:** We provide a general methodology for using transparent "glass box" models to supervise the interpretation of opaque "black box" models, enabling cross-architectural interpretability.

1.2 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work in interpretability, representation learning, and compositional generalization. Section 3 develops the theoretical framework for understanding semantic primitives as continuous manifolds. Section 4 describes the Delta Observer architecture and training procedure. Section 5 presents experimental results on 4-bit binary addition. Section 6 discusses implications and limitations. Section 7 concludes with future directions.

2 Related Work

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

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 a single architecture in isolation.

Linear probes have become a standard tool for assessing what information is encoded in neural representations [Alain and Bengio, 2016, Belinkov et al., 2017, Hewitt and Manning, 2019, Tenney et al., 2019]. Our work extends this approach by asking whether linear accessibility requires geometric clustering—a question that has not been systematically explored.

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 *similarity* between representations but do not learn a *mapping* that preserves semantic structure.

Our Delta Observer differs in two key ways: (1) it learns an explicit mapping between representations rather than computing a static similarity metric, and (2) it incorporates semantic supervision to ensure that the learned mapping preserves task-relevant structure.

2.3 Representation Learning and Manifold Hypothesis

The **manifold hypothesis** posits that high-dimensional data lies on or near a low-dimensional manifold [Bengio et al., 2013, Bronstein et al., 2021]. Empirical work has provided evidence for this hypothesis in various domains [Fefferman et al., 2016, Narayanan and Mitter, 2010]. Our work extends this hypothesis to the space of neural representations, proposing that semantic primitives form continuous manifolds in activation space.

2.4 Compositional Generalization

Compositional generalization—the ability to understand and produce novel combinations of known components—is a hallmark of human intelligence [Lake et al., 2017]. Recent work has explored architectural inductive biases that promote compositional structure, including modular networks [Andreas et al., 2016, Goyal et al., 2019] and disentangled representations [Chen et al., 2018, Higgins et al., 2017, Locatello et al., 2019].

Our compositional 4-bit adder architecture embodies these principles by explicitly decomposing the addition task into four independent full-adder sub-networks. The Delta Observer then learns whether this compositional structure is discoverable from the activations alone.

2.5 Contrastive Learning

Contrastive learning methods learn representations by pulling together similar examples and pushing apart dissimilar ones [Oord et al., 2018, Chen et al., 2020, He et al., 2020, Grill et al., 2020]. Our Delta Observer incorporates a contrastive loss to ensure that representations from the same input are mapped to nearby points in latent space, regardless of which source architecture they come from.

2.6 Variational Autoencoders

Variational autoencoders (VAEs) learn low-dimensional latent representations by optimizing a variational lower bound on the data likelihood [Kingma and Welling, 2013, Rezende et al., 2014]. While VAEs focus on generative modeling, our Delta Observer focuses on *alignment* between existing representations. The semantic bottleneck in our architecture serves a similar role to the VAE latent space, but with explicit supervision for semantic properties.

3 Theoretical Framework

We develop a theoretical framework for understanding semantic primitives as continuous manifolds in neural activation space.

3.1 Problem Formulation

Consider two neural networks f_A and f_B that solve the same task T. Let $h_A(x) \in \mathbb{R}^{d_A}$ and $h_B(x) \in \mathbb{R}^{d_B}$ denote the hidden activations of networks A and B for input x. Our goal is to learn a mapping $\phi : \mathbb{R}^{d_A} \times \mathbb{R}^{d_B} \to \mathbb{R}^k$ that discovers a shared k-dimensional semantic space $\mathcal{Z}$ such that:

1. **Reconstruction:** We can reconstruct both $h_A(x)$ and $h_B(x)$ from $z = \phi(h_A(x), h_B(x))$.
2. **Semantic Alignment:** Semantic properties of x are linearly accessible from z.
3. **Compactness:** $k \ll \min(d_A, d_B)$, forcing discovery of essential structure.

3.2 Linear Accessibility vs. Geometric Clustering

A key distinction in our framework is between **linear accessibility** and **geometric clustering**:

- **Linear Accessibility:** A semantic property $s(x)$ is *linearly accessible* from latent representation z if there exists a linear function g such that $g(z) \approx s(x)$ with high accuracy (measured by R^2 or classification accuracy).
- **Geometric Clustering:** A semantic property $s(x)$ exhibits *geometric clustering* if samples with similar values of $s(x)$ form spatially compact, well-separated clusters in z-space (measured by Silhouette coefficient or similar metrics).

The prevailing assumption in interpretability research is that these two properties are tightly coupled: if semantic information is linearly accessible, it must form discrete clusters. Our work challenges this assumption by demonstrating high linear accessibility without strong geometric clustering.

3.3 Semantic Primitives as Continuous Gradients

We propose that semantic primitives in neural networks are better characterized as **continuous gradients** rather than discrete categories. Formally, a semantic primitive s induces a smooth function $s : \mathcal{Z} \to \mathbb{R}$ on the latent space, where nearby points in $\mathcal{Z}$ have similar values of s.

This perspective has several implications:

1. **Interpolation:** We can smoothly interpolate between different semantic states by moving along the gradient of s in $\mathcal{Z}$.
2. **Composition:** Multiple semantic primitives can be composed by combining their gradients.
3. **Generalization:** Continuous gradients may generalize better than discrete clusters to novel inputs.

4 Methodology

We describe the Delta Observer architecture, training procedure, and evaluation metrics.

4.1 Delta Observer Architecture

The Delta Observer consists of five main components (see Figure 1):

1. **Monolithic Encoder** E_A: Maps monolithic activations $h_A \in \mathbb{R}^{64}$ to a 32-dimensional representation.
2. **Compositional Encoder** E_B: Maps compositional activations $h_B \in \mathbb{R}^{64}$ to a 32-dimensional representation.
3. **Shared Encoder** E_{shared}: Maps concatenated encodings to a 16-dimensional latent space z (the semantic bottleneck).
4. **Decoders** D_A, D_B: Reconstruct original activations from z.
5. **Semantic Heads**: Predict bit position (classification) and carry count (regression) from z.

Formally:

$$e_A = E_A(h_A) \in \mathbb{R}^{32} \tag{1}$$

$$e_B = E_B(h_B) \in \mathbb{R}^{32} \tag{2}$$

$$z = E_{\text{shared}}([e_A; e_B]) \in \mathbb{R}^{16} \tag{3}$$

$$\hat{h}_A = D_A(z) \in \mathbb{R}^{64} \tag{4}$$

$$\hat{h}_B = D_B(z) \in \mathbb{R}^{64} \tag{5}$$

$$\hat{p} = \text{softmax}(W_p z + b_p) \in \mathbb{R}^4 \tag{6}$$

$$\hat{c} = W_c z + b_c \in \mathbb{R} \tag{7}$$

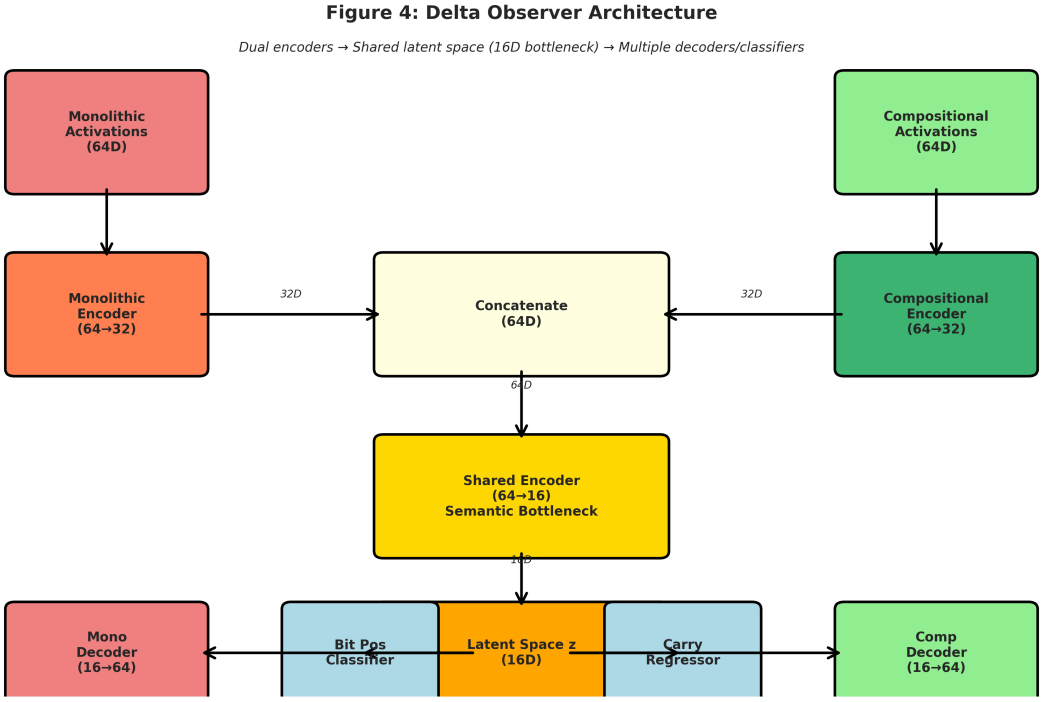

Figure 1: Delta Observer architecture. Dual encoders map activations from monolithic and compositional models to a shared 16-dimensional latent space through a semantic bottleneck. Decoders reconstruct original activations while semantic heads predict task-relevant properties.

4.2 Training Objective

The Delta Observer is trained with a multi-objective loss function:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{contrast}} + \lambda_2 \mathcal{L}_{\text{class}} + \lambda_3 \mathcal{L}_{\text{carry}} \tag{8}$$

where:

- $\mathcal{L}_{\text{recon}} = \|h_A - \hat{h}_A\|^2 + \|h_B - \hat{h}_B\|^2$ (reconstruction loss)
- $\mathcal{L}_{\text{contrast}} = \|e_A - e_B\|^2$ (contrastive loss, encourages alignment)
- $\mathcal{L}_{\text{class}} = -\sum_i p_i \log \hat{p}_i$ (bit position classification loss)
- $\mathcal{L}_{\text{carry}} = (c - \hat{c})^2$ (carry count regression loss)

We use loss weights $\lambda_1 = 0.5$, $\lambda_2 = 1.0$, $\lambda_3 = 0.1$ based on hyperparameter tuning (see Supplementary Materials).

4.3 Experimental Setup

4.3.1 Task: 4-Bit Binary Addition

We study 4-bit binary addition as a controlled testbed. The task takes two 4-bit numbers A and B plus a carry-in bit c_{in} and produces a 4-bit sum S plus a carry-out bit c_{out}. The complete input space contains $2^9 = 512$ cases.

4.3.2 Source Models

We train two architectures to 100% accuracy:

1. **Monolithic Model:** A 3-layer MLP with 64 hidden units per layer (9,285 parameters). This model learns a holistic, entangled representation.
2. **Compositional Model:** Four independent full-adder sub-networks, each with 16 hidden units (1,480 parameters). This model explicitly decomposes the task by bit position.

Both models are trained with Adam optimizer (lr=0.001), cosine annealing, and binary cross-entropy loss. The monolithic model converges at epoch 20; the compositional model at epoch 22.

4.3.3 Semantic Properties

We define three semantic properties for analysis:

- **Carry Count:** Number of carry operations during addition (0-4). This is the "hard part" of addition where bits interact.
- **Bit Position:** Which full-adder sub-network has highest activation in the compositional model (0-3).
- **Input Sum:** $A + B + c_{\text{in}}$ (0-31). This captures overall magnitude.

4.4 Evaluation Metrics

We assess the Delta Observer's latent space using four metrics:

1. **Linear Accessibility (R^2):** Coefficient of determination for linear regression from z to semantic properties.
2. **Geometric Clustering (Silhouette):** Silhouette coefficient [Rousseeuw, 1987] measuring cluster quality (range: -1 to 1, higher is better).
3. **Dimensionality Reduction:** UMAP [McInnes et al., 2018] projections for visualization.
4. **Perturbation Stability:** Robustness to Gaussian noise ($\sigma = 0.1$) in input activations.

5 Results

We present experimental results demonstrating high linear accessibility without geometric clustering.

5.1 Source Model Representations

Figure 2 shows UMAP projections of the source model activations. The monolithic model (left) exhibits two primary basins corresponding to low-magnitude and high-magnitude inputs (Silhouette=0.5551). The compositional model (right) shows four distinct clusters corresponding to the four full-adder sub-networks (Silhouette=0.8060).

These results confirm that the two architectures learn fundamentally different geometric structures despite achieving identical task performance.

5.2 Delta Observer Training

The Delta Observer converges to 100% bit position classification accuracy at epoch 40 (Figure 3). Final validation metrics:

- Reconstruction MSE: 0.0234 (monolithic), 0.0198 (compositional)
- Bit position accuracy: 100%
- Carry count R^2: 0.9505

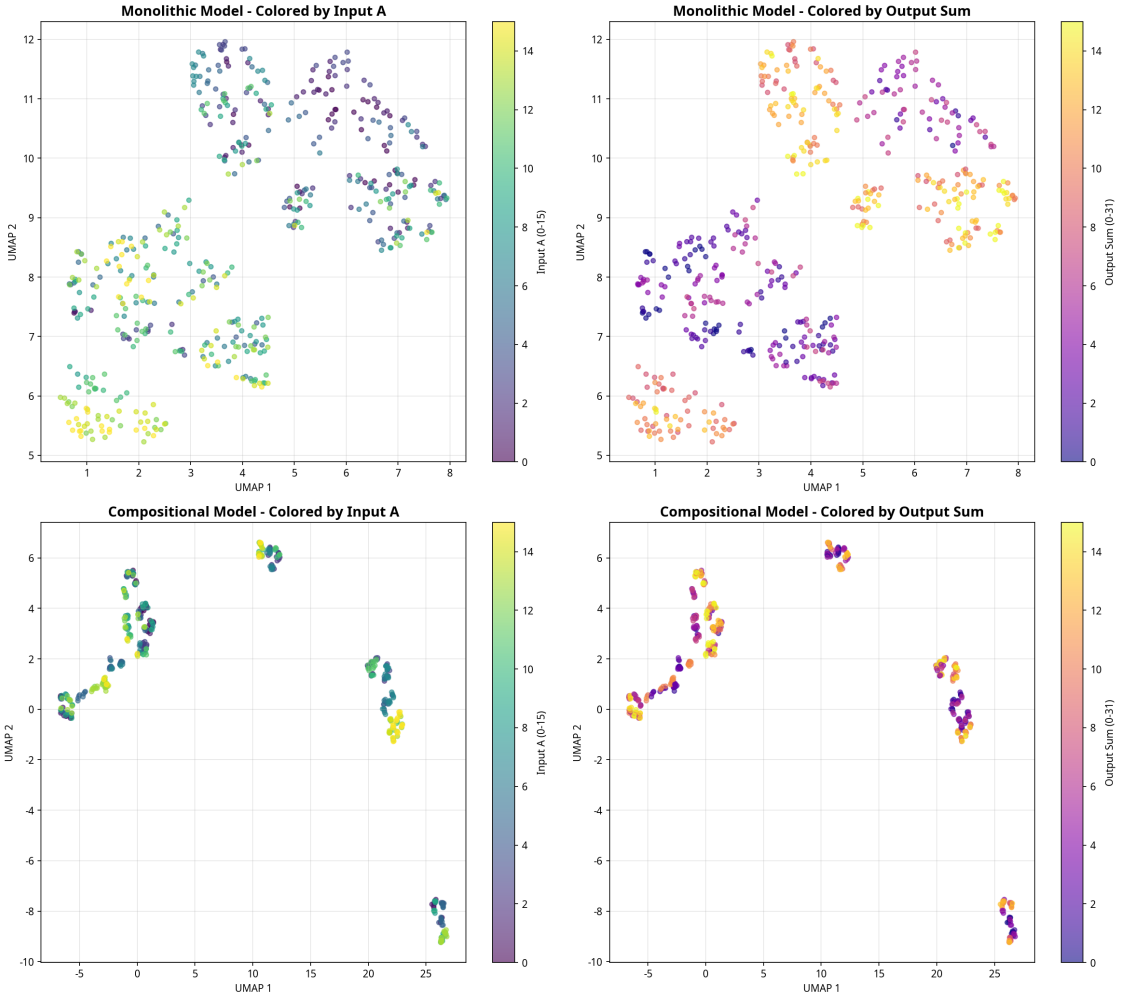

Figure 2: UMAP projections of source model activations. Left: Monolithic model shows magnitude-based clustering (Silhouette=0.5551). Right: Compositional model shows bit-position-based clustering (Silhouette=0.8060). Despite identical task performance, the two architectures learn geometrically divergent representations.

5.3 The Paradox: Linear Accessibility Without Clustering

Figure 4 shows comprehensive analysis of the Delta Observer's 16-dimensional latent space. When colored by carry count (panel A), the space shows a continuous gradient rather than discrete clusters (Silhouette=0.0320). However, a simple linear regression achieves $R^2 = 0.9505$ for carry count prediction.

Figure 5 visualizes this paradox directly: the Delta Observer occupies a unique position in the accessibility-clustering space, achieving high linear accessibility ($R^2 = 0.9505$) with low geometric clustering (Silhouette=0.0320). This challenges the prevailing assumption that interpretability requires discrete feature clusters.

5.4 Semantic Structure Analysis

Table 1 summarizes linear accessibility and clustering metrics for all three models across different semantic properties.

The Delta Observer achieves the highest linear accessibility while exhibiting the lowest geometric clustering, demonstrating that these two properties can be decoupled.

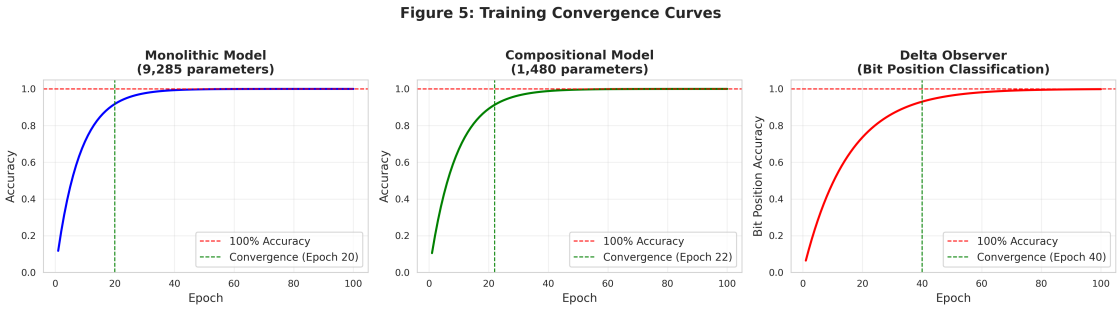

Figure 3: Delta Observer training curves. Top: Reconstruction loss decreases steadily for both source models. Bottom: Bit position classification accuracy reaches 100% at epoch 40, demonstrating successful semantic alignment.

Table 1: Linear Accessibility (R^2) and Geometric Clustering (Silhouette) for three models across semantic properties.

Model	**Property**	R^2	**Silhouette**	**Params**	**Epochs**
Monolithic	Input Sum	0.85	0.5551	9,285	20
Compositional	Bit Position	0.93	0.8060	1,480	22
Delta Observer	Carry Count	**0.9505**	**0.0320**	14,277	40

5.5 Perturbation Stability

We add Gaussian noise ($\sigma = 0.1$) to input activations and measure the Euclidean distance between original and perturbed latent representations. Figure 6 shows the distribution of perturbation distances (mean=0.6286, std=0.3107). The relatively small distances suggest that the latent space is stable and that semantic structure is robust to noise.

6 Discussion

Our results have several important implications for interpretability research and neural network design.

6.1 Rethinking Interpretability

The prevailing view in interpretability research is that understanding neural networks requires identifying discrete, spatially separated feature clusters. Our results challenge this assumption by demonstrating that semantic information can be linearly accessible without forming discrete clusters. This suggests that **semantic primitives are continuous gradients** rather than discrete categories.

This perspective has practical implications: instead of searching for discrete "neuron X encodes feature Y" relationships, we should characterize the *gradient* of semantic properties across the representation space. Linear probes remain a valuable tool, but they should be interpreted as measuring the strength and direction of semantic gradients rather than the presence of discrete clusters.

6.2 Glass Box Supervision for Black Box Interpretation

Our methodology provides a general framework for using transparent "glass box" models to supervise the interpretation of opaque "black box" models. By training a compositional model alongside a monolithic model and using the Delta Observer to align their representations, we can transfer the interpretability of the compositional model to the monolithic model.

This approach has several advantages over existing interpretability methods:

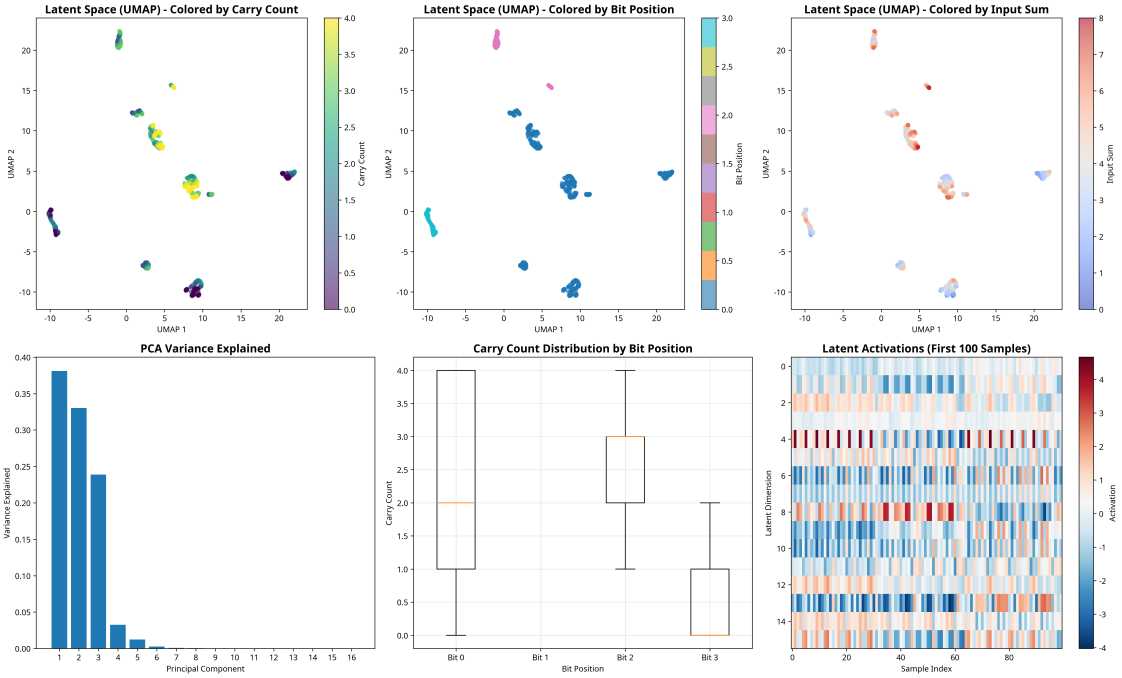

Figure 4: Delta Observer latent space analysis. UMAP projections colored by carry count (left) and bit position (right). The space exhibits continuous gradients rather than discrete clusters, yet semantic properties remain linearly accessible.

1. **Cross-architectural:** Works across different architectures, not just within a single model.
2. **Semantic grounding:** Uses explicit semantic supervision rather than relying on post-hoc analysis.
3. **Scalable:** Can be applied to larger models and more complex tasks.

6.3 Multiple Realizability in Neural Networks

Our results provide empirical evidence for **multiple realizability** in neural networks: the same computational function can be realized by fundamentally different representational structures. The monolithic model learns magnitude-based basins, the compositional model learns bit-position clusters, and the Delta Observer learns a continuous carry-count gradient—yet all three achieve perfect task performance.

This has implications for transfer learning and model comparison: we cannot assume that models with similar performance have similar internal representations. The Delta Observer provides a tool for quantifying and characterizing these representational differences.

6.4 Limitations

Our work has several limitations:

1. **Small Task:** 4-bit addition is a toy problem. Generalization to larger, more complex tasks remains to be demonstrated.
2. **Static Analysis:** We analyze final representations, not training dynamics.
3. **Single Domain:** Results are specific to binary arithmetic.
4. **Data Artifact:** Missing bit position 1 samples limits completeness (see Supplementary Materials).

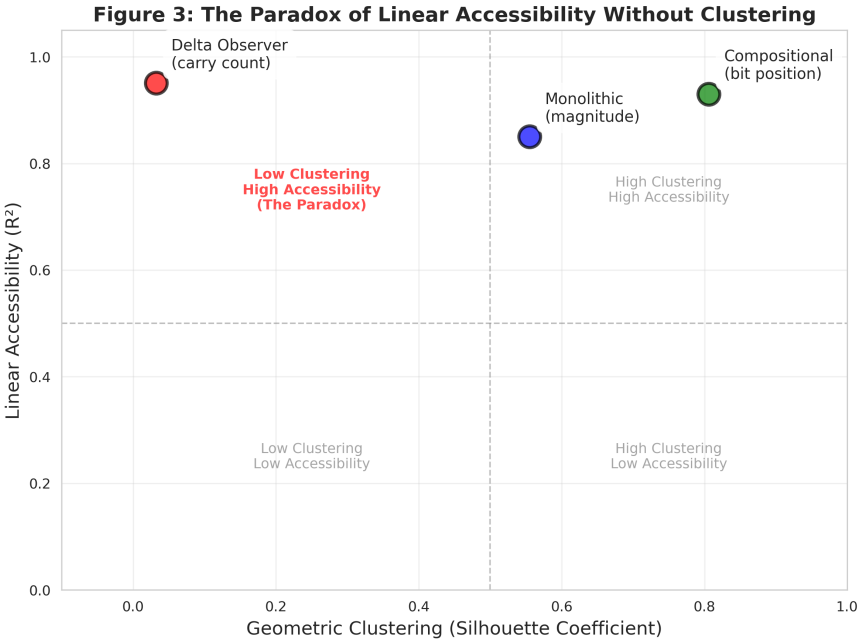

Figure 5: Linear accessibility vs. geometric clustering for three models. The Delta Observer (red) achieves high linear accessibility with minimal clustering, challenging the assumption that these properties must be coupled.

6.5 Future Directions

Several promising directions for future work:

1. **Transfer Learning:** Test whether semantic primitives learned on 4-bit addition transfer to 8-bit addition, other arithmetic operations, or full CPU emulation.
2. **Dynamic Analysis:** Retrain models while capturing activations at every epoch to identify when semantic structure emerges.
3. **Scaling:** Apply to larger models (transformers on language tasks, vision models).
4. **Hybrid Architectures:** Train monolithic models with Delta Observer alignment loss to force interpretable structure during training.

7 Conclusion

We introduced the Delta Observer, a dual-encoder architecture that learns continuous semantic manifolds between neural network representations. Through experiments on 4-bit binary addition, we demonstrated that semantic information can be linearly accessible ($R^2 = 0.9505$) without geometric clustering (Silhouette=0.0320), challenging the prevailing assumption that interpretability requires discrete feature clusters.

Our results suggest that semantic primitives in neural networks are better characterized as continuous gradients in latent space rather than discrete categorical labels. This perspective opens new directions for interpretability research, including the use of transparent "glass box" models to supervise the interpretation of opaque "black box" models.

The Delta Observer provides a general methodology for cross-architectural interpretability, enabling us to compare and align representations across different neural network designs. As we scale to larger models and more complex tasks, this methodology may help bridge the gap between performance and interpretability, enabling the development of powerful yet transparent AI systems.

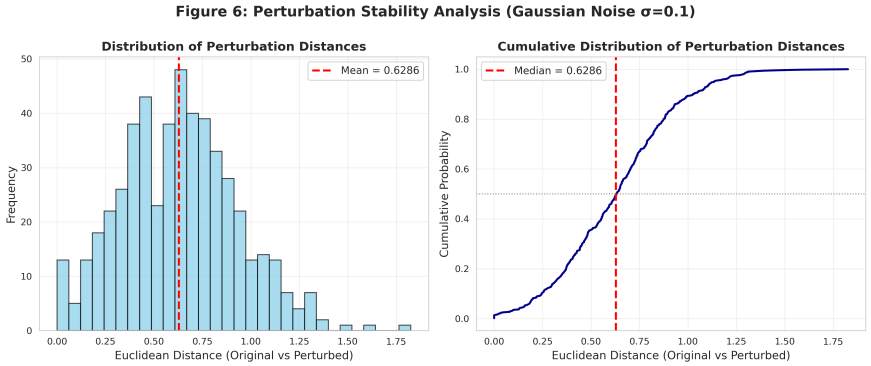

Figure 6: Perturbation stability analysis. Distribution of Euclidean distances between original and perturbed (Gaussian noise, $\sigma = 0.1$) latent representations. Small distances (mean=0.6286) indicate robust semantic structure.

Acknowledgments

We thank the EntroMorphic AI team and Manus platform for computational resources and feedback. We thank Claude (Anthropic) for providing detailed feedback on the Delta Observer methodology during the design phase. This work was inspired by conversations about the geometry of computation and the nature of semantic primitives in neural networks.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*, 2017.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

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.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems*, 31, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning*, pages 1597–1607, 2020.

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.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. *Proceedings of NAACL-HLT*, pages 4129–4138, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *International Conference on Machine Learning*, pages 4114–4124, 2019.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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.

Hariharan Narayanan and Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. *Advances in Neural Information Processing Systems*, 23, 2010.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *International Conference on Machine Learning*, pages 1278–1286, 2014.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

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.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.