Topological Integration in Attention-Based Neural Networks

Cluster Collapse, Counterexamples, and the Role of Direct Interaction

Adam Kruger
Revelry Inc., Independent
adam@revelry-inc.com

March 2026

Abstract

We use persistent homology to measure the topological structure of hidden representations across layers in three architecturally distinct neural networks: a 4B-parameter dense transformer (Qwen3-4B), a 328M-parameter recursive transformer (NanoChat), and a 370M-parameter state-space model (Mamba-370m). While prior work has characterized the intrinsic dimensionality (ID) profile across layers, persistent homology reveals a qualitatively distinct phenomenon: in attention-based architectures, independent representational clusters collapse to a single connected component at intermediate layers (persistent β_0: 519 $\rightarrow$ 1 in Qwen3-4B, 517 $\rightarrow$ 1 in NanoChat), then re-differentiate at the output. This cluster collapse is absent in the state-space model (Mamba β_0: 571 $\rightarrow$ 987, no collapse) and in untrained transformers (β_0: 503 $\rightarrow$ 968, clusters proliferate; confirmed across 8 random seeds). A sensitivity analysis across 36 parameter combinations and 100 bootstrap iterations confirms the robustness of these findings. We additionally report an emergent bimodal processing gate in Qwen3-4B routing 93% of tokens through a shallow path (Mode A) and 7% through a deep path (Mode B), with Mode B tokens showing early, near-complete topological collapse ($\beta_0 = 4$) at L6 relative to Mode A ($\beta_0 = 811$). We propose that topological integration requires two conditions: (1) an architectural mechanism for direct interaction between representations (attention), and (2) gradient-based optimization to learn the integrated configuration. We discuss connections to Integrated Information Theory and computational symbiogenesis, while noting the limitations of these analogies.

1 Introduction

The geometric properties of neural network representations have been studied through intrinsic dimensionality (ID) estimation, revealing a characteristic "hunchback" profile across layers — ID increases in early layers and decreases toward the output (Ansuini et al., 2019). This pattern has been confirmed in large transformer models (Valeriani et al., 2023) and connected to the Information Bottleneck framework (Tishby et al., 2000; Shwartz-Ziv & Tishby, 2017), though the relationship between ID compression and generalization remains debated (Saxe et al., 2018).

We extend this line of work by applying *topological* analysis — specifically, persistent homology — to neural network representations. Where ID measures the local dimensionality of the data manifold, persistent homology captures global structural properties: how many independent components exist (β_0), how many loops or cycles are present (β_1), and how these features persist across spatial scales.

Our central finding is a *cluster collapse* — the number of persistent connected components drops from hundreds to one at intermediate layers in attention-based architectures. This is

qualitatively different from the ID hunchback: while ID measures manifold complexity, cluster collapse measures manifold *connectivity*. A representation can have high intrinsic dimensionality while being topologically unified.

Critically, this collapse is *absent* in a state-space model (Mamba), which processes tokens sequentially rather than through direct pairwise interaction. This architectural counterexample suggests topological integration requires a mechanism for direct interaction — an observation we connect, with appropriate caveats, to Integrated Information Theory (Tononi, 2004) and computational symbiogenesis (Agüera y Arcas et al., 2024).

We further report an emergent bimodal processing gate in Qwen3-4B that routes tokens into two processing streams with qualitatively different topological signatures. This gate emerges from gradient descent in a fully dense transformer with no architectural routing mechanism.

2 Related Work

Intrinsic dimensionality in neural networks. Ansuini et al. (2019) first characterized the layer-wise ID profile using the Two-NN estimator (Facco et al., 2017). Valeriani et al. (2023) extended this to large transformers, showing ID minima correspond to layers with the richest semantic content. Our work is complementary — rather than measuring manifold complexity (ID), we measure manifold *connectivity* (β_0).

Neural Collapse. Papyan et al. (2020) identified a terminal phase in classification training where last-layer features collapse to class means forming a simplex ETF. Our cluster collapse occurs at intermediate layers during autoregressive inference, suggesting a distinct phenomenon.

Topology in neural networks. Persistent homology has been applied to loss landscapes (Ballester & Araujo, 2024), decision boundaries (Ramamurthy et al., 2019), and training dynamics (Rieck et al., 2019). Our work applies it to *representations across layers during inference*.

Representation similarity. CKA (Kornblith et al., 2019) measures pairwise similarity between layer representations. Our approach characterizes the internal topological structure of each layer independently.

State-space models. Mamba (Gu & Dao, 2023) achieves competitive performance with transformers while avoiding quadratic attention complexity. Its sequential architecture provides a natural counterexample for studying the role of direct interaction in representation geometry.

3 Methods

3.1 Models

We analyze three architecturally distinct models:

- **Qwen3-4B:** 36-layer dense transformer, $d_{\text{model}} = 2560$, 4B parameters, SiLU activation. Notably, Qwen3-4B is a fully dense model (not mixture-of-experts) — all parameters are active for every token. The routing behavior described in Section 4.5 is therefore entirely emergent.
- **NanoChat Recursive:** 328M-parameter model, $d_{\text{model}} = 1280$, ReLU^2 activation. Architecture: 2 prelude $\rightarrow$ 4 recurrent layers (looped 4$\times$) $\rightarrow$ 2 coda layers.

- **Mamba-370m:** 48-layer selective state-space model, $d_{\text{model}} = 1024$, 370M parameters. No attention mechanism.

3.2 Activation Capture

For Qwen3-4B, we captured residual stream activations at 7 layers (L3, L5, L6, L8, L16, L24, L35) across 3.78M tokens, using forward hooks. For NanoChat, we captured 9 stages (embed, P0, P1, R0–R3, C0, C1) across 49,664 tokens from WikiText-2. For Mamba, we captured 7 layers (L0, L8, L16, L24, L32, L40, L47) across 40,960 tokens from WikiText-2.

3.3 Two-NN Intrinsic Dimensionality

We estimate ID using the Two-NN estimator (Facco et al., 2017):

$$\text{ID} = \frac{n}{\sum_{i=1}^{n} \log \mu_i}, \quad \mu_i = \frac{d_2(x_i)}{d_1(x_i)} \tag{1}$$

where d_1, d_2 are the first and second nearest-neighbor distances. We subsample 10,000 tokens with 100 bootstrap iterations for confidence intervals.

3.4 Persistent Homology

We compute persistent homology using Ripser (Tralie et al., 2018). Default parameters: 1,000 landmark points (random selection), PCA projection to 50 dimensions, Vietoris–Rips filtration with Euclidean distance, maximum dimension 1. We report *persistent* Betti numbers — features with lifetime exceeding 10% of the maximum lifetime at that dimension.

Sensitivity analysis. We sweep 36 parameter combinations: landmarks $\in \{500, 1000, 2000\}$, PCA dimensions $\in \{30, 50, 100, \text{None (raw 2560-dim)}\}$, persistence threshold $\in \{5\%, 10\%, 20\%\}$.

Bootstrap confidence intervals. 100 iterations with different random landmark selections at each layer.

3.5 Controls

1. **Untrained model:** Randomly initialized NanoChat (same architecture, no training), 8 random seeds.
2. **Architectural counterexample:** Mamba-370m (trained, no attention).

4 Results

4.1 Intrinsic Dimensionality

All three models show distinct ID profiles (Figure 1).

Qwen3-4B: Hunchback peaking at L16 (ID = 9.8), consistent with prior work (Ansuini et al., 2019; Valeriani et al., 2023).

NanoChat: ID peaks during recursion (R1: ID = 12.1, 95% CI [11.82, 12.49]), with a secondary rise at the coda (C1: ID = 13.9, 95% CI [13.65, 14.28]). The secondary rise may reflect coda layers preparing output representations. The embedding layer shows anomalously high ID (22.6, CI [20.15, 25.33]), likely reflecting high-dimensional token embedding space before any processing.

Mamba: ID increases monotonically from 2.61 (L0) to 8.37 (L32), then plateaus. No hunchback. The untrained NanoChat shows flat ID (≈ 3.3 across all stages, std=0.1), consistent with random projections.

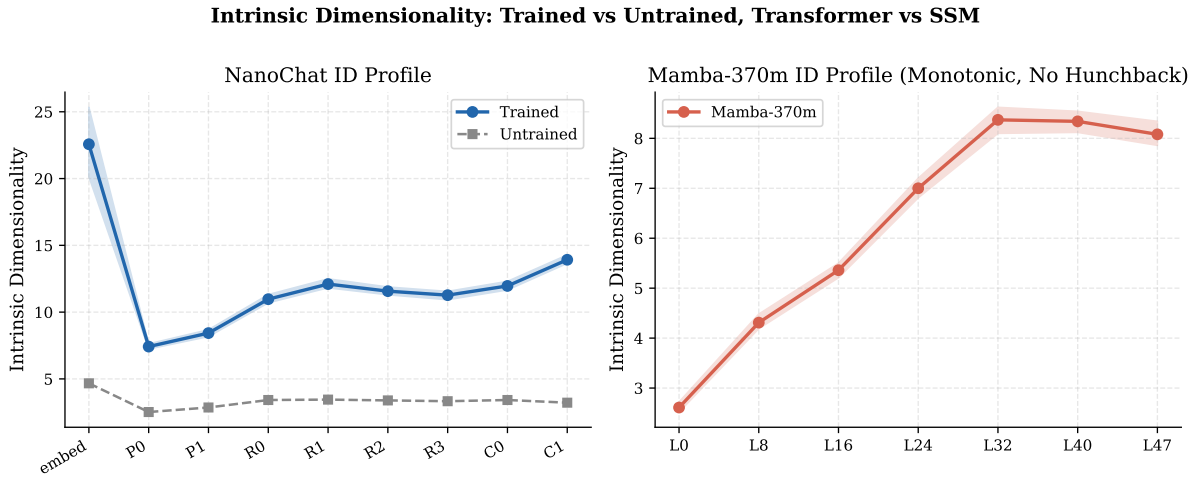

Figure 1: Intrinsic dimensionality profiles. **Left:** NanoChat shows an inverted-U with a secondary rise at the coda; untrained model is flat. **Right:** Mamba shows monotonic increase with no hunchback, consistent with its lack of attention-mediated integration.

4.2 Cluster Collapse in Attention-Based Models

Figure 2 shows the persistent β_0 profiles. The cluster collapse is the central finding.

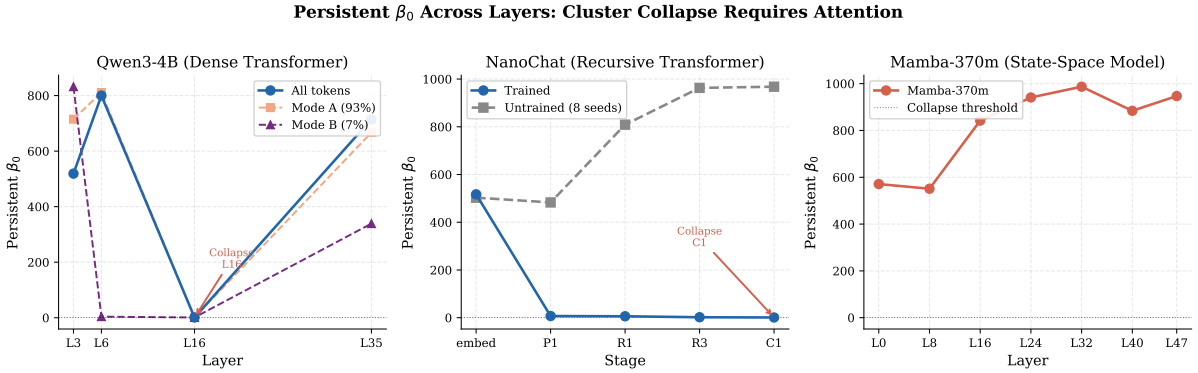

Figure 2: Persistent β_0 (connected components) across layers. **Left:** Qwen3-4B collapses to 1 at L16, with Mode B tokens showing early collapse at L6. **Center:** NanoChat trained model collapses to 1 at C1; untrained model proliferates (503 $\rightarrow$ 968). **Right:** Mamba never collapses — β_0 increases monotonically to 987.

Table 1: Persistent β_0 across layers. Attention-based trained models collapse to a single connected component; Mamba does not.

Model	Architecture	Early	Mid	Late	Collapse?
Qwen3-4B	Dense transformer	519 (L3)	**1 (L16)**	715 (L35)	Yes
NanoChat	Recursive transformer	517 (embed)	**1 (C1)**	—	Yes
Mamba-370m	State-space model	571 (L0)	987 (L32)	947 (L47)	**No**
NanoChat (untrained)	—	503 (embed)	963 (R3)	968 (C1)	**No**

4.3 Loop Formation (β_1)

Persistent β_1 (1-cycles) peaks near the integration layer in attention-based models:

Table 2: Persistent β_1 (1-cycles) across layers.

Model	Early	Peak	Late
Qwen3-4B	207 (L3)	**342 (L16)**	145 (L35)
NanoChat	425 (embed)	**400 (R3)**	156 (C1)
Mamba-370m	112 (L0)	325 (L40)	264 (L47)

Mamba develops loop structure without accompanying cluster collapse — loops form within a fragmented, multi-component manifold rather than a unified one.

4.4 Robustness

Sensitivity analysis (36 combinations, Qwen3-4B L16):

- Collapse ($p\beta_0 \leq 5$) confirmed in **30/36** combinations (83%)
- All 6 failures use 500 landmarks (insufficient sampling density)
- With $\geq$1000 landmarks: **23/24** combinations (96%)
- Without PCA (raw 2560-dim): $p\beta_0 = 2$ — collapse is not a PCA artifact

Table 3: Bootstrap distribution of $p\beta_0$ at each Qwen3-4B layer (100 iterations). The bimodal distribution (either ≤ 5 or > 500) reflects landmark sampling sensitivity, not topological ambiguity.

Layer	$p\beta_0 \leq 5$	$p\beta_0 > 500$	Distribution
L3	0%	79%	Always fragmented
L6	57%	43%	Transitional
L16	**87%**	12%	Mostly collapsed
L24	**90%**	10%	Mostly collapsed
L35	0%	37%	Re-differentiated

Bootstrap confidence intervals (100 iterations, Qwen3-4B):

Multi-seed untrained control (8 seeds, NanoChat):

- Clusters proliferate in **8/8 seeds**: embed $\approx 297 \rightarrow$ C1 ≈ 486
- ID flat at ≈ 3.3 (std=0.1)
- The untrained pattern is deterministic across initializations

Table 4: Trained vs. untrained NanoChat comparison (topology and ID).

Metric	Untrained	Trained
β_0 (early)	503	517
β_0 (mid)	809 (proliferating)	**1 (collapsed)**
β_0 (late)	968	**1**
ID profile	Flat ≈ 3.3	Inverted-U, peak 12.1
PCA variance (mid)	$\approx 51\%$	94–100%

4.5 Emergent Bimodal Gating (Qwen3-4B)

We independently discovered an emergent processing gate at L3 in Qwen3-4B routing 93% of tokens into a shallow path (Mode A) and 7% into a deep path (Mode B). Causal ablation confirms the gate is functional. The two modes show qualitatively different topological signatures:

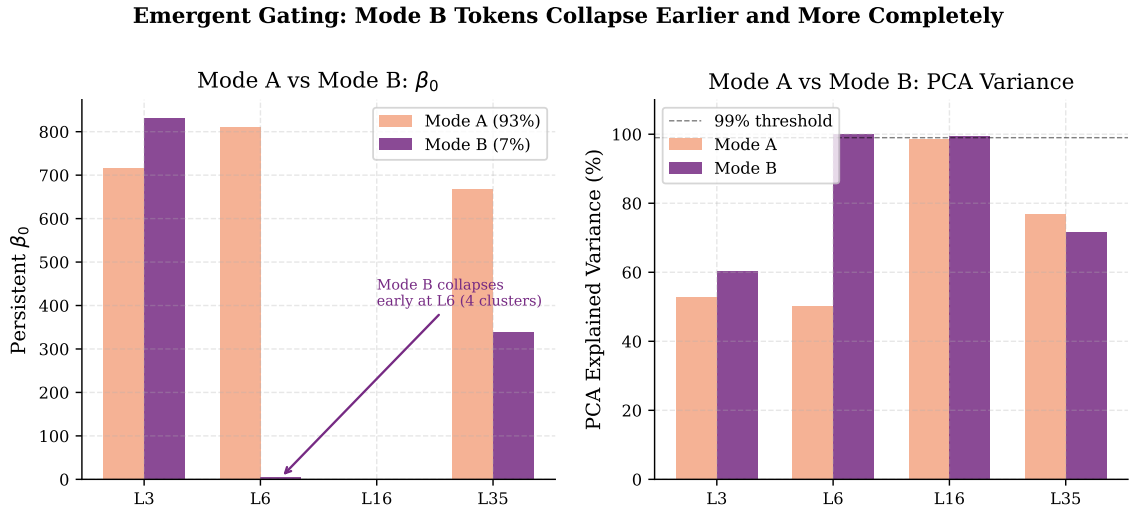

Figure 3: Emergent bimodal gate: Mode A vs. Mode B tokens. **Left:** β_0 comparison — Mode B tokens collapse to 4 clusters at L6 while Mode A remains at 811. **Right:** PCA variance — Mode B achieves 99.9% variance explained at L6, consistent with near-complete topological collapse.

Table 5: Mode A vs. Mode B topological signatures across layers.

	$p\beta_0$		PCA Variance	
Layer	Mode A	Mode B	Mode A	Mode B
L3	715	832	52.7%	60.3%
L6	811	**4**	50.2%	**99.9%**
L16	1	1	98.7%	99.5%
L35	667	339	76.7%	71.7%

Mode B tokens collapse earlier and more completely: at L6, Mode B is nearly topologically unified ($\beta_0 = 4$, PCA variance 99.9%) while Mode A remains fragmented ($\beta_0 = 811$). Both modes reach full integration by L16. The gate fires at sentence onset rather than at the emotion word, suggesting structural rather than lexical sensitivity.

We note that Qwen3-4B supports an explicit "thinking mode" via special tokens. The emergent dual-mode processing we measure here may represent the mechanistic substrate of that documented capability.

5 Discussion

5.1 Two Conditions for Topological Integration

Our results suggest cluster collapse requires two simultaneous conditions:

1. Architectural mechanism for direct interaction. Attention provides pairwise interaction between all token representations, enabling them to merge into a unified manifold. Mamba's sequential state updates do not enable direct interaction. The absence of collapse in Mamba despite successful training ($\beta_0 : 571 \to 987$) suggests optimization alone is insufficient — the architecture must provide the meeting point.

2. Gradient-based optimization. Untrained transformers have the architectural capacity for integration (attention) but show the opposite pattern — clusters proliferate ($503 \to 968$, 8/8 seeds). Training finds the integrated configuration. Neither attention nor optimization alone is sufficient; both must be present simultaneously.

Falsifiable prediction. Any architecture enabling direct pairwise interaction between representations will develop cluster collapse when trained. Any architecture lacking this mechanism will not, regardless of depth, parameter count, or training duration.

5.2 Structural Analogies

We note structural analogies to several theoretical frameworks, while acknowledging these remain analogies rather than formal equivalences.

Integrated Information Theory. Tononi's Φ measures how much a system is "more than the sum of its parts" through causal irreducibility (Tononi, 2004). Our cluster collapse ($\beta_0 \to 1$) measures *topological* irreducibility. The architectural dependence we observe is consistent with IIT's emphasis on causal interaction structure: without direct interaction (Mamba), integration does not emerge regardless of optimization pressure.

Computational symbiogenesis. Agüera y Arcas et al. (2024) demonstrate that self-replicating programs emerge from random computation through the merger of independent computational units, via a sharp phase transition. Our cluster collapse follows the same structural pattern. The Mamba counterexample reinforces this: without direct interaction, fusion cannot occur.

Free Energy Principle. The observation that optimization drives systems toward integrated configurations is consistent with free energy minimization (Friston, 2010), where the unified manifold represents a lower-energy state under prediction error minimization. We have not formally established this connection.

5.3 Limitations

1. **Scale of study.** Three models. Testing across encoder-decoders, vision transformers, and mixture-of-experts architectures is needed.

2. **Bootstrap variance.** The bimodal distribution at L16 (87% collapse, 12% no collapse) reflects landmark sampling sensitivity. Improved landmark selection (e.g., farthest-point sampling) may reduce this variance.

3. **Causal claims.** We observe correlations between architecture and topology but have not established causal mechanisms.

4. **Theoretical connections.** Our analogies to IIT, FEP, and symbiogenesis are structural, not formal.

5. **Training dynamics.** We compare trained and untrained endpoints but do not measure topology during training. When does the collapse emerge?

6. **Token autocorrelation.** Tokens within sequences are not independent. Effective sample size is smaller than raw token count.

7. **Gating mechanism.** We report the emergent gate and its topological signature but have not identified the circuit responsible. Mechanistic analysis is ongoing.

6 Conclusion

Persistent homology reveals a topological phase transition in attention-based neural networks — cluster collapse from hundreds of independent components to a single unified manifold — that is absent in state-space models and untrained networks. This suggests topological integration requires both an architectural mechanism for direct interaction (attention) and gradient-based optimization, neither alone being sufficient.

We additionally report an emergent bimodal processing gate in a dense transformer, with Mode B tokens showing early, near-complete collapse ($\beta_0 = 4$ at L6) versus Mode A ($\beta_0 = 811$), despite the model having no architectural routing mechanism.

These findings suggest that certain topological properties of learned representations are architectural necessities rather than incidental features — direct interaction plus optimization pressure reliably produces integration. We hope this encourages further investigation at the intersection of algebraic topology, representation geometry, and mechanistic interpretability.

References

Agüera y Arcas, B., et al. (2024). Computational Life: How Well-formed, Self-replicating Programs Emerge from Simple Interaction. *arXiv:2406.19108*.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. *NeurIPS*.

Ballester, P. and Araujo, R. M. (2024). On the Topology of Deep Neural Networks. *arXiv:2101.07083*.

Facco, E., d'Errico, M., Rodriguez, A., and Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*.

Gu, A. and Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv:2312.00752*.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of Neural Network Representations Revisited. *ICML*.

Marks, S. and Tegmark, M. (2023). The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. *arXiv:2310.06824*.

Papyan, V., Han, X. Y., and Donoho, D. L. (2020). Prevalence of Neural Collapse during the terminal phase of deep learning training. *PNAS*.

Ramamurthy, K. N., et al. (2019). A topological synergy between neural network architectures and deep learning algorithms. *arXiv:1906.11537*.

Rieck, B., et al. (2019). Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. *ICLR*.

Saxe, A., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B., and Cox, D. (2018). On the Information Bottleneck Theory of Deep Learning. *ICLR*.

Shwartz-Ziv, R. and Tishby, N. (2017). Opening the Black Box of Deep Neural Networks via Information. *arXiv:1703.00810*.

Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *Proceedings of the 37th Allerton Conference*.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*.

Tralie, C., Saul, N., and Bar-On, R. (2018). Ripser.py: A Lean Persistent Homology Library for Python. *JOSS*.

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A., and Cazzaniga, A. (2023). The geometry of hidden representations of large transformer models. *NeurIPS*.