Topological Integration in Attention-Based Neural Networks

Cluster Collapse, Counterexamples, and the Role of Direct Interaction

Adam Kruger
Light of Baldr LLC
adam@lightofbaldr.com

March 2026 (revised June 2026)

Abstract

Persistent homology offers a view of neural-network representations that intrinsic dimensionality cannot: it counts how many disconnected pieces the representation manifold breaks into at each layer. We apply it to six networks from four architectural families and find that trained transformers with full attention pull their hidden representations into a single connected component partway through the network, then break them apart again at the output (persistent β_0: 519 $\to$ 1 in Qwen3-4B, 1202 $\to$ 1 in Qwen3-14B, 1299 $\to$ 1 in Qwen3-32B, 517 $\to$ 1 in a recursive NanoChat variant). The collapse survives an 8$\times$ scale ladder and a sensitivity sweep on Qwen3-4B (83% of 36 parameter settings; 96% with $\geq$1,000 landmarks). It is absent in three controls: a state-space model with no token-to-token interaction (Mamba-370m, β_0: 571 $\to$ 947, peak 987), a sliding-window transformer whose interaction is mostly local (Gemma-4-31B never drops below $\beta_0 = 1546$ in 60 layers, despite an intact ID profile), and untrained networks (clusters proliferate, 503 $\to$ 968; 8/8 seeds). In every model tested, integration appears only when attention lets any token reach any other *and* the network is trained. We also describe an emergent two-mode processing gate in Qwen3-4B (93%/7% split) whose deep path collapses early ($\beta_0 = 4$ at L6, against 811 for the shallow path), and we discuss connections to Integrated Information Theory and computational symbiogenesis as structural analogies, with implications for interpretability and backdoor detection.

Keywords: persistent homology, interpretability, representation geometry, attention, state-space models

1 Introduction

The geometry of neural-network representations is usually studied through intrinsic dimensionality (ID), which traces a characteristic "hunchback" across layers: ID rises through the early network and falls toward the output (Ansuini et al., 2019). The pattern holds in large transformers (Valeriani et al., 2023) and connects to the Information Bottleneck picture of deep learning (Tishby et al., 2000; Shwartz-Ziv & Tishby, 2017), though how compression relates to generalization is still argued (Saxe et al., 2018).

ID is a local measurement. It tells you how many directions the data occupies in a small neighborhood, and nothing about how the neighborhoods fit together. Persistent homology measures the part ID misses: how many separate pieces the manifold has (β_0), whether it contains loops (β_1), and how long those features survive as you grow the spatial scale. A representation can be high-dimensional and connected, or low-dimensional and broken into pieces. The two measurements are independent, and it turns out the interesting structure is in the second one.

Our central finding is a *cluster collapse*. In trained transformers with full attention, the hidden representations start out fragmented into hundreds of persistent components (over a thousand, at higher sampling density), fuse into exactly one connected component at interior layers, and then re-fragment as the network prepares its output. In the larger models the fusion is a cliff: Qwen3-14B goes from 1334 components at layer 5 to 3 at layer 6. One layer. The collapse appears in every full-attention model we measured, from a 328M-parameter recursive toy to Qwen3-32B.

It does not appear everywhere, and the failures are what give the result its shape. A state-space model (Mamba), which has no mechanism for one token's representation to directly interact with another's, never collapses. Neither does Gemma-4-31B, a trained 31B-parameter transformer whose attention is restricted to a sliding local window in five of every six layers. Gemma-4 is the more telling counterexample: it has attention, it is well trained, its ID hunchback is fully intact, and its representations still never unify. Together the two failures point to *uniform global token interaction* as the operative architectural condition, an observation we connect, with caveats, to Integrated Information Theory (Tononi, 2004) and computational symbiogenesis (Agüera y Arcas et al., 2024).

Along the way we describe a second emergent structure: Qwen3-4B, a fully dense model with no routing machinery, routes 93% of tokens down a shallow processing path and 7% down a deep one, and the two paths leave different topological footprints. Gradient descent built a gate where the architecture provides none.

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 that ID minima fall at the layers carrying the richest semantic content. Our work is complementary: where ID measures manifold complexity, we measure manifold connectivity.

Neural Collapse. Papyan et al. (2020) identified a terminal phase of classification training in which last-layer features collapse onto class means, forming a simplex ETF. Our collapse happens at interior layers during autoregressive inference, which suggests a distinct phenomenon.

Topology in neural networks. Persistent homology has been applied across neural-network analysis, from decision boundaries (Ramamurthy et al., 2019) to training dynamics (Rieck et al., 2019); see Ballester et al. (2024) for a survey. We apply it to representations across layers during inference.

Representation similarity and geometry. CKA (Kornblith et al., 2019) measures how similar two layers' representations are to each other; we characterize the internal structure of each layer on its own. Marks & Tegmark (2023) found emergent linear structure in how language models represent factual truth, evidence that learned geometry carries meaning. We ask the same kind of question one level up, at the global topology.

State-space models. Mamba (Gu & Dao, 2023) matches transformer performance on many tasks while avoiding quadratic attention. Its sequential architecture makes it a natural counterexample for asking what attention's direct interaction actually buys.

3 Methods

3.1 Models

We analyze six models spanning four architectural families:

- **Qwen3-4B:** 36-layer dense transformer, $d_{\text{model}} = 2560$, 4B parameters, SiLU activation, full (global) attention at every layer. Qwen3-4B is fully dense, not mixture-of-experts; every parameter is active for every token, which is what makes the routing behavior of Section 4.6 interesting.
- **Qwen3-14B:** 40-layer dense transformer, $d_{\text{model}} = 5120$, 14B parameters, full attention at every layer. Included to test whether the collapse survives scale.
- **Qwen3-32B:** 64-layer dense transformer, $d_{\text{model}} = 5120$, 32B parameters, full attention at every layer. The top rung of the scale ladder.
- **NanoChat Recursive:** 328M parameters, $d_{\text{model}} = 1280$, ReLU^2 activation. This is our modification of NanoChat (Karpathy, 2025): we replace the standard layer stack with a recursive block of 2 prelude layers, 4 recurrent layers (weight-tied, looped 4×), and 2 coda layers, to test whether weight-tied recursion changes the topological signature. Base architecture and training recipe are otherwise unchanged.
- **Gemma-4-31B:** 60-layer transformer, $d_{\text{model}} = 5376$, 31B parameters, instruction-tuned. What matters here is its attention schedule: five *sliding-window* layers (window 1,024) for every one *full* attention layer, so global attention appears only at every sixth layer and direct token interaction is predominantly local.
- **Mamba-370m:** 48-layer selective state-space model, $d_{\text{model}} = 1024$, 370M parameters. No attention at all.

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 drawn from the uncopyrighted subset of The Pile (Gao et al., 2020), using forward hooks. For NanoChat we captured 9 stages (embed, P0, P1, R0–R3, C0, C1) across 49,664 tokens from WikiText-2, and for Mamba 7 layers (L0, L8, L16, L24, L32, L40, L47) across 40,960 tokens from WikiText-2.

For Qwen3-14B, Qwen3-32B, and Gemma-4-31B we captured post-block residual streams at *every* layer (40, 64, and 60 layers respectively) over a shared 440-text corpus of 40 chat-formatted prompts and 400 raw samples from C4, about 122K tokens per model (121,942 for Qwen3-14B; counts vary slightly with each model's tokenizer).

3.3 Two-NN Intrinsic Dimensionality

We estimate ID with 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 and run 100 bootstrap iterations for confidence intervals.

3.4 Persistent Homology

We compute persistent homology with Ripser v0.6.4 (Tralie et al., 2018) on post-residual-stream activations: the output of each transformer block after the residual addition, before the next block's layer normalization. For Mamba we use each Mamba block's output.

Each layer gives an activation matrix of shape tokens $\times$ d_{model}. The pipeline is: (1) random landmark subsampling (1,000 points for Qwen3-4B, NanoChat, and Mamba), (2) PCA projection to 50 dimensions (raw d_{model} is tested in the sensitivity sweep), (3) pairwise Euclidean distances, (4) Vietoris–Rips filtration up to homological dimension 1. For Qwen3-14B, Qwen3-32B, and Gemma-4-31B we use 2,000 landmarks, PCA-50, a 10% persistence threshold, and a fixed random seed, the configuration the sensitivity sweep validates. One caution when reading the tables: raw component counts scale with landmark count, so early-layer β_0 values from the 1,000-landmark and 2,000-landmark analyses are not directly comparable. The collapse itself ($\beta_0 \to 1$) does not depend on this choice.

We report *persistent* Betti numbers, counting only features whose lifetime (death $-$ birth) exceeds 10% of the maximum lifetime at that dimension. The threshold filters topological noise from genuine structure.

The sensitivity sweep covers 36 parameter combinations: landmarks $\in \{500, 1000, 2000\}$, PCA dimensions $\in \{30, 50, 100, \text{None (raw 2560-dim)}\}$, and persistence threshold $\in \{5\%, 10\%, 20\%\}$. Bootstrap confidence intervals use 100 iterations with fresh random landmark draws at each layer.

3.5 Controls

1. **Untrained model:** randomly initialized NanoChat, same architecture, no training, 8 random seeds.
2. **No direct interaction:** Mamba-370m, trained, no attention.
3. **Local-only interaction:** Gemma-4-31B, trained, attention-based, but predominantly sliding-window.

3.6 Gating Probe and Ablation (Qwen3-4B)

To characterize what triggers the emergent gate of Section 4.6, we built a 25-sentence probe set spanning four categories: emotionally charged, high-valence, neutral, and ambiguous. For each sentence we measure the L6 residual activation magnitude per token, comparing the sentence-initial token against the emotionally salient word. This separates lexical triggering (the gate fires at the emotion word) from structural triggering (the gate fires at sentence onset).

To test whether the gate actually does anything, we use an ablation: a forward hook clamps the L3 residual activations of Mode B tokens to the Mode A centroid, and we measure the effect on the deep-processing response at L6 over a 20,000-token evaluation set drawn from the capture corpus, with and without the intervention.

4 Results

4.1 Intrinsic Dimensionality

All models show distinct ID profiles (Figure 1; the scale-ladder and Gemma-4-31B profiles appear in the right panel of Figure 3; Qwen3-4B's profile is reported numerically).

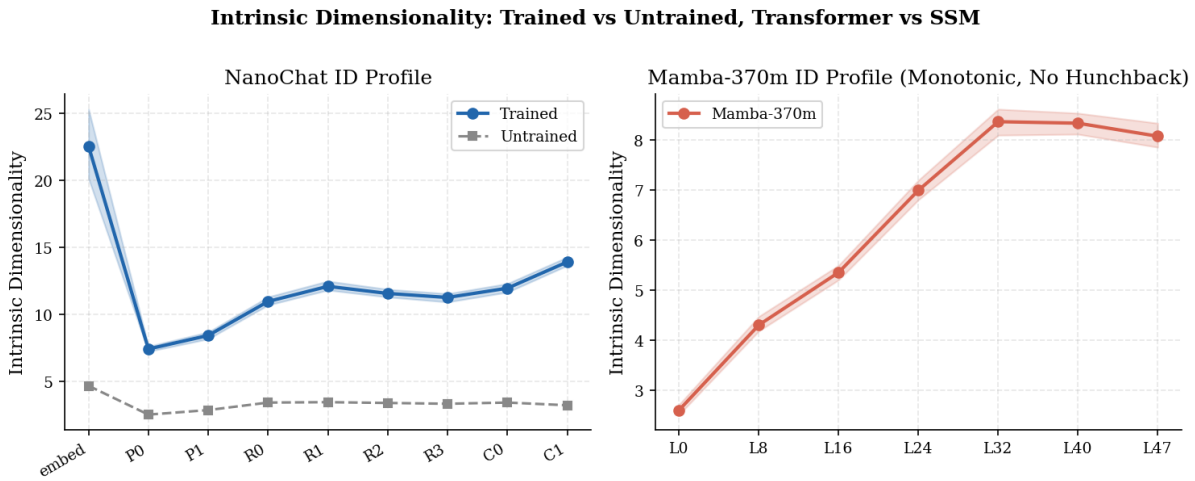

Figure 1: Intrinsic dimensionality profiles. **Left:** NanoChat shows an inverted-U with a secondary rise at the coda; the untrained model is flat. **Right:** Mamba rises monotonically with no hunchback. Qwen3-14B/32B and Gemma-4-31B appear in Figure 3.

Qwen3-4B shows the familiar hunchback, peaking at L16 (ID = 9.8), in line with prior work (Ansuini et al., 2019; Valeriani et al., 2023). The larger Qwen3 models keep the hunchback but push it higher and deeper: Qwen3-14B peaks at L22 (ID = 14.05) and Qwen3-32B at L47 (ID = 15.11).

NanoChat peaks during recursion (R1: ID = 12.1, 95% CI [11.82, 12.49]) with a second rise at the coda (C1: ID = 13.9, 95% CI [13.65, 14.28]), which may reflect the coda layers preparing output representations. Its embedding layer reads anomalously high (22.6, CI [20.15, 25.33]), most likely the raw token-embedding space before any processing has happened.

Gemma-4-31B complicates the picture, usefully. Its hunchback is fully intact: after an initial reading of 9.9 at L0, ID dips to 4.5 at L1, climbs to a peak of 17.65 at L29 (about half depth), and falls to 13.3 at the output. That peak is the highest mid-network ID in this study (only NanoChat's pre-processing embedding reading of 22.6 is larger), and yet, as Section 4.3 shows, Gemma-4-31B never topologically unifies. Manifold complexity and manifold connectivity come apart cleanly here: the sliding-window architecture preserves the first; the second never arrives.

Mamba's ID rises monotonically from 2.61 (L0) to 8.37 (L32) and plateaus. No hunchback. The untrained NanoChat is flat at ID ≈ 3.3 across all stages (std 0.1), which is what random projections look like.

4.2 Cluster Collapse in Full-Attention Models

The central finding is in Figure 2 and Table 1: the number of persistent connected components in a trained full-attention transformer falls from hundreds to exactly one.

The collapse survives scale, and the way it survives is itself a result. Measured at every layer rather than a sampled subset, both larger Qwen3 models reproduce the collapse with a strikingly abrupt onset at the same absolute layer (Figure 3, left): persistent β_0 falls from 1334 (L5) to 3 (L6) in Qwen3-14B, and from 1138 (L5) to 2 (L6) in Qwen3-32B. A thousand islands at layer 5; a handful at layer 6. The exact minimum ($\beta_0 = 1$) arrives at L12 in Qwen3-14B (30% depth) and L8 in Qwen3-32B (12.5% depth). So the collapse is early, and it does not move deeper as models grow: onset sits at L6 in both (15% and 9.4% of depth), even though the 4B model's L16 (44%) had suggested the collapse layer might scale with depth. What grows instead is the *unified span*, the stretch of consecutive layers with $\beta_0 \leq 5$: L6–L37 in Qwen3-14B (32 of 40

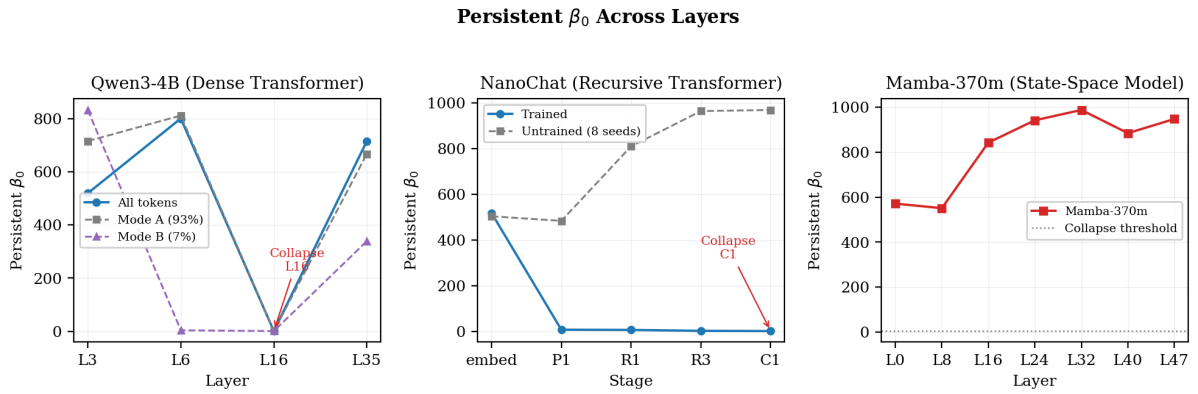

Figure 2: Persistent β_0 (connected components) across layers. **Left:** Qwen3-4B collapses to 1 at L16; Mode B tokens collapse early, at L6. **Center:** trained NanoChat collapses to 1 at C1 while the untrained model proliferates (503 → 968). **Right:** Mamba never collapses; β_0 climbs to a maximum of 987 (L32) and ends at 947 (L47).

Table 1: Persistent β_0 across layers. Bold marks the collapse minimum. For Gemma-4-31B, Mid reports the minimum across all 60 layers. Raw counts are not comparable across landmark settings (see Section 3.4); the collapse signature is.

Model	Architecture	Early	Mid	Late	Collapse?
Qwen3-4B	Dense (full attention)	519 (L3)	**1 (L16)**	715 (L35)	Yes
Qwen3-14B	Dense (full attention)	1202 (L0)	**1 (L12)**	1734 (L39)	Yes
Qwen3-32B	Dense (full attention)	1299 (L0)	**1 (L8)**	1998 (L63)	Yes
NanoChat	Recursive transformer	517 (embed)	2 (R3)	**1 (C1)**	Yes
Gemma-4-31B	Sliding-window (5:1)	1989 (L0)	1546 (L15)	1995 (L59)	**No**
Mamba-370m	State-space model	571 (L0)	987 (L32)	947 (L47)	**No**
NanoChat (untrained)	—	503 (embed)	963 (R3)	968 (C1)	**No**

layers) and L6–L62 in Qwen3-32B (57 of 64). In both models, most of the network operates on one connected manifold, with fragmentation confined to the entry and the exit.

All three Qwen3 models re-differentiate in the final layers (Qwen3-4B: 1 → 715; Qwen3-14B: 1 → 1734; Qwen3-32B: 1 → 1998), consistent with the output layers re-expanding the unified representation toward token-level predictions. NanoChat's capture ends at its collapse stage, C1, so re-differentiation cannot be assessed there.

4.3 No Collapse Under Sliding-Window Attention (Gemma-4-31B)

Gemma-4-31B is a trained, attention-based transformer, and it never unifies. Across all 60 layers, persistent β_0 stays between 1546 and 1999 (Figure 3, left), with the minimum (1546) at L15. No layer comes anywhere near collapse. Global attention layers do appear in its stack, one in every six, and in this model that periodic global mixing did not unify the manifold.

This is a second counterexample class, and a different one from Mamba. Mamba has no direct token-to-token interaction at all. Gemma-4-31B has direct interaction, but restricted to a local window for five of every six layers. Neither integrates. Set against the full-attention models, which all collapse, the three regimes form a graded series (Table 2).

Integration tracks interaction density: in our measurements, what separates collapsing from non-collapsing models is not the presence of attention but *uniform global token interaction*, the architectural guarantee that any two representations can directly interact at a mixing step.

Table 2: Interaction regime versus topological outcome.

Interaction regime	Example	Outcome
None (sequential state)	Mamba-370m	No collapse (571 → 947, peak 987)
Local-only (sliding window)	Gemma-4-31B	No collapse (min $\beta_0 = 1546$)
Uniform global (full attention)	Qwen3 4B/14B/32B, NanoChat	Collapse ($\beta_0 \to 1$)

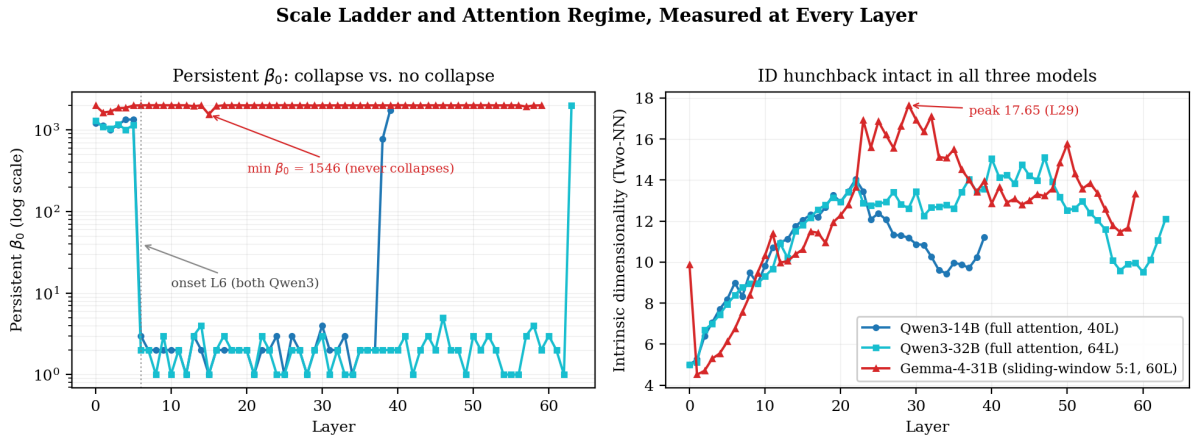

Figure 3: Scale ladder and attention regime, measured at every layer. **Left:** persistent β_0 (log scale). Qwen3-14B and Qwen3-32B collapse at the same absolute layer (onset L6, a single-layer cliff); Gemma-4-31B never drops below $\beta_0 = 1546$. **Right:** intrinsic dimensionality. All three models show an intact hunchback, and Gemma-4-31B has the highest peak (17.65 at L29) despite never collapsing. Connectivity and complexity dissociate.

4.4 Loop Formation (β_1)

Persistent β_1 (1-cycles) stays elevated through the mid-network in the attention models and falls sharply after integration (Table 3). In Qwen3-4B, β_1 peaks at the integration layer itself (342 at L16). NanoChat's β_1 is highest at the embedding (425), stays elevated through recursion (400 at R3), and falls at the coda (156). Mamba develops loop structure too, but without the accompanying cluster collapse: its loops form inside a fragmented, multi-component manifold rather than a unified one.

4.5 Robustness

We probed the Qwen3-4B collapse three ways.

The parameter sweep varies landmarks, PCA dimensionality, and persistence threshold, 36 combinations in all, measured at L16. Collapse ($p\beta_0 \leq 5$) holds in 30 of 36 (83%). Failures concentrate at the smallest landmark count (500), which undersamples the manifold; with at least 1,000 landmarks the rate is 23 of 24 (96%). Skipping PCA entirely and running the filtration on the raw 2560-dimensional activations still yields $p\beta_0 = 2$, so the collapse is not a projection artifact.

The bootstrap re-runs the pipeline 100 times per layer with fresh landmark draws (Table 4). At the integration layers the mass sits at collapse: 87% of iterations at L16 and 90% at L24 land at $p\beta_0 \leq 5$. The boundary layers never collapse (0% at L3 and L35), and the intermediate values there reflect landmark sampling, not ambiguity about the underlying topology.

The untrained control spans 8 random seeds. For direct comparison with the trained model, Table 1, Table 5, and the abstract report the seed whose embedding-layer count (503) sits closest

Table 3: Persistent β_1 (1-cycles) across layers.

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

Table 4: Bootstrap distribution of $p\beta_0$ at each Qwen3-4B layer (100 iterations).

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

to the trained model's (517), which keeps the comparison like-for-like; it is also the largest trajectory of the eight, so the 8-seed mean below is the more conservative summary. The other seeds tell the same story: clusters proliferate instead of collapsing, with an 8-seed mean of roughly 297 at the embedding rising to roughly 486 at the final stage, and collapse in none of the 8. ID stays flat at about 3.3 (std 0.1) in every seed. Whatever a random initialization does, it does not integrate.

4.6 Emergent Bimodal Gating (Qwen3-4B)

Qwen3-4B contains a gate that nothing in its architecture asks for. At L3, 93% of tokens take a shallow processing path (Mode A) and 7% take a deep one (Mode B). Our 3.78M-token corpus yields roughly 265K Mode B tokens, plenty for reliable topological statistics. Under the ablation of Section 3.6, clamping Mode B activations to the Mode A centroid at L3 suppresses the deep-processing response at L6: Mode B mean activation falls by 93%, and tokens in the extreme tail (>p99.9) fall from 20 to 0. The gate is functional; the routing is load-bearing, not epiphenomenal. We do not claim to have identified the circuit that implements it, nor to have established a causal mechanism for the cluster collapse itself. Both remain open (Section 5.3).

The two modes do not look alike (Figure 4, Table 6). Mode B collapses earlier and more completely: at L6 it is nearly unified ($\beta_0 = 4$, PCA variance 99.9%) while Mode A is still fragmented ($\beta_0 = 811$). Both modes reach full integration by L16. On the gating probe (Section 3.6), emotionally charged sentences drive L6 activation about 48% higher than neutral ones, but the response concentrates at the sentence-initial token rather than at the emotion word, which suggests the routing decision is made at sentence onset on structural rather than lexical cues.

Qwen3-4B ships with an explicit "thinking mode" that users invoke through special tokens. The emergent dual-mode processing we measure here may be the mechanistic substrate of that documented capability.

Table 5: Trained vs. untrained NanoChat, single representative seed. Stages: embed → R3 (mid-recursion) → C1 (final).

Metric	Untrained	Trained
β_0 (embed)	503	517
β_0 (mid, R3)	963 (proliferating)	**2 (collapsed)**
β_0 (late, C1)	968	**1**
ID profile	Flat ≈ 3.3	Inverted-U, peak 12.1
PCA variance (mid)	$\approx 50\%$	94–100%

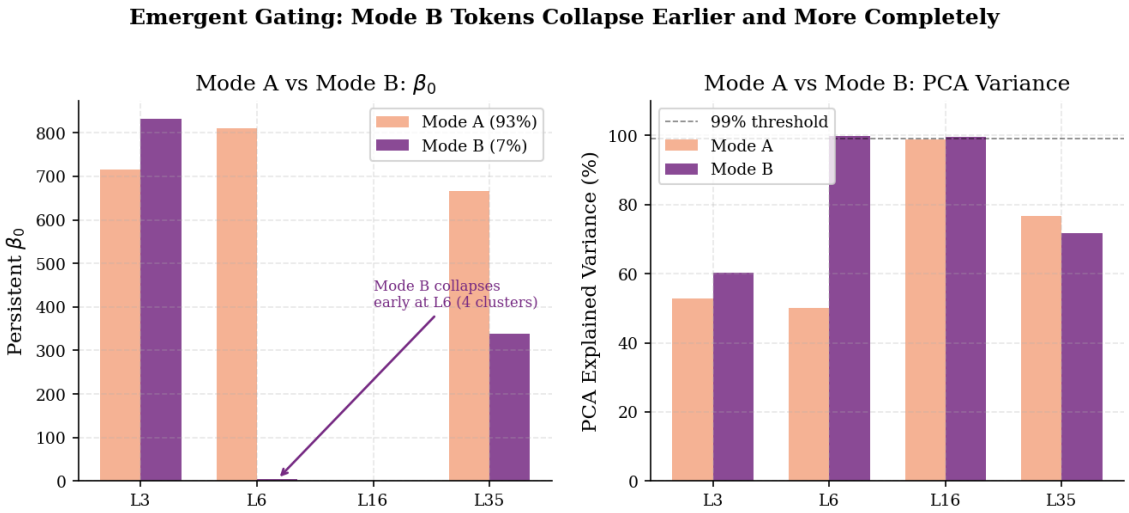

Figure 4: Emergent bimodal gate, Mode A vs. Mode B. **Left:** β_0. Mode B tokens collapse to 4 clusters at L6 while Mode A remains at 811. **Right:** PCA variance. Mode B reaches 99.9% variance explained at L6, consistent with near-complete topological collapse.

Table 6: 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%

5 Discussion

5.1 Two Conditions for Topological Integration

Our results are consistent with the hypothesis that cluster collapse requires two things at once, and in all tested models it appeared only when both were present.

The first is an architectural mechanism for uniform global interaction. Full attention lets every token representation interact with every other at each mixing step, which is what lets them merge into one manifold. The Gemma-4-31B counterexample shows that attention by itself is not the condition: attention restricted to a local sliding window, even with periodic global layers, showed no collapse (min $\beta_0 = 1546$). Mamba, with no direct interaction at all, likewise shows no collapse despite successful training (β_0: 571 $\rightarrow$ 947, peak 987). The two counterexamples bracket the condition from opposite sides. In our data, integration appears only where interaction is both direct and uniformly global.

The second is gradient-based optimization. An untrained transformer has the architecture for integration and does the opposite with it: clusters proliferate (representative seed 503 $\rightarrow$ 968; no collapse in any of 8 seeds). Training finds the integrated configuration.

Falsifiable prediction. Any trained architecture that provides uniform global pairwise interaction at every mixing layer should develop cluster collapse, and architectures whose direct interaction is absent (state-space models) or predominantly local (sliding-window attention) should not, regardless of depth, parameter count, or training duration. All six models measured here fit this prediction, but the counterexamples that motivated its conditions cannot also count as tests of it. Out-of-sample validation is needed across encoder-decoders, vision transformers, mixture-of-experts, and hybrid attention-SSM architectures; the most decisive single test would be a much larger pure full-attention model (Section 5.3).

5.2 Structural Analogies

Several theoretical frameworks rhyme with what we measure. They are analogies, not formal equivalences, and we present them as such.

Integrated Information Theory. Tononi's Φ measures how much a system is "more than the sum of its parts" through causal irreducibility (Tononi, 2004). Cluster collapse ($\beta_0 \rightarrow 1$) measures topological irreducibility. The architectural dependence we observe fits IIT's emphasis on causal interaction structure: in the trained interaction-free model we tested (Mamba), integration did not emerge despite optimization, and with interaction restricted to a local neighborhood (Gemma-4-31B) it did not emerge either.

Computational symbiogenesis. Agüera y Arcas et al. (2024) showed self-replicating programs emerging from random computation through mergers of independent computational units, via a sharp phase transition. Our cluster collapse follows the same structural pattern, and the counterexamples reinforce it: without direct interaction (Mamba), or with interaction confined to local neighborhoods (Gemma-4-31B), fusion did not occur.

Free Energy Principle. That optimization drives systems toward integrated configurations is consistent with free energy minimization (Friston, 2010); on that reading, the unified manifold would be the lower-energy state. We have not formally established this connection.

5.3 Limitations

1. **Scale and architecture coverage.** The study spans four architectural families (full-attention dense, recursive, sliding-window, state-space), a 4B→14B→32B full-attention scale ladder, and three attention regimes. The remaining gap is a pure full-attention model larger than 32B, to fully separate scale from attention regime; Qwen2.5-72B is the natural candidate. Encoder-decoders, vision transformers, and mixture-of-experts architectures also remain untested.
2. **Bootstrap variance.** At L16, 87% of bootstrap iterations collapse and most of the rest land fragmented, a split that reflects landmark sampling sensitivity. Better landmark selection (e.g., farthest-point sampling) may reduce it.
3. **Causal claims.** We observe correlations between architecture and topology; we have not established causal mechanisms.
4. **Theoretical connections.** The IIT, FEP, and symbiogenesis parallels remain informal.
5. **Training dynamics.** We compare trained and untrained endpoints; we do not measure topology during training. When the collapse emerges is open.
6. **Token autocorrelation and corpus variation.** Tokens within sequences are not independent, so effective sample size is smaller than raw token count. Capture corpora also differ across models (The Pile for Qwen3-4B, WikiText-2 for NanoChat and Mamba, a C4-based mixture for the scale ladder and Gemma-4-31B). The collapse signature appears consistent across the corpora used, but corpus and architecture are confounded across models, so a corpus-controlled comparison remains future work.
7. **Gating mechanism.** We report the emergent gate and its topological signature; we have not identified the circuit responsible. Mechanistic analysis is ongoing.

6 Conclusion

Measured with persistent homology, trained full-attention transformers pass through a topological phase transition. Hundreds of separate representation clusters (over a thousand, at denser sampling) fuse into one connected manifold in the early-to-middle network and split apart again at the output. The pattern held in every full-attention model we measured, across an 8× scale ladder, with onset pinned to layer 6 in both larger models. No model lacking all-to-all attention showed it: not Mamba, which has no direct token interaction; not Gemma-4-31B, whose attention is mostly windowed; and not untrained networks, at any seed. This suggests topological integration needs both an architectural mechanism for uniform global interaction and gradient-based optimization, with neither sufficient alone. The Gemma-4-31B result is what sharpens the claim: the property that tracks integration in our measurements is not attention as such, but attention's guarantee of unrestricted, all-to-all token interaction.

We also report an emergent bimodal processing gate in a dense transformer, with Mode B tokens collapsing early and nearly completely ($\beta_0 = 4$ at L6, against 811 for Mode A), in a model with no architectural routing mechanism at all.

The regularity of the pattern is what suggests these topological properties are baked into the architecture rather than being accidents of any particular training run. If that is right, then deviation from a model's expected topological profile becomes a candidate screen for tampered or backdoored models, an application we are exploring separately. Topology is a blunt instrument next to circuit-level analysis, but it is cheap, global, and architecturally diagnostic, and it found structure here that dimensionality alone cannot see. We think it has more to find.

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, R., et al. (2024). Topological Data Analysis for Neural Network Analysis: A Comprehensive Survey. *arXiv:2312.05840*.

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*.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv:2101.00027*.

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

Karpathy, A. (2025). nanochat. GitHub repository. https://github.com/karpathy/nanochat.

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., Varshney, K. R., and Mody, K. (2019). Topological Data Analysis of Decision Boundaries with Application to Model Selection. *ICML*. arXiv:1805.09949.

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*.