Topological Persistence of Transformer Weight Spaces Tracks Training Dynamics and Improves Scaling Prediction

Anonymous Authors

Abstract

Training large language models costs millions of dollars, yet the decision of when to stop training and scale up relies almost entirely on loss curve extrapolation—a method that ignores the structural evolution of model weights during optimization. We investigate whether persistent homology, a tool from topological data analysis, can capture complementary signals about training dynamics in transformer language models. We train five GPT-style models (2.9K–97K parameters) on WIKITEXT-2 with 51 checkpoints each, and compute two topological summaries at every checkpoint: neural persistence of the weight graph and Vietoris-Rips persistent homology of the weight-space point cloud. Our main finding is that H_0 total persistence from Vietoris-Rips filtration is strongly anticorrelated with training loss (Spearman $\rho = -0.59$ to $-0.91, p < 10^{-5}$), and this correlation *strengthens monotonically with model size*. When combined with loss curve extrapolation, topological features reduce final loss prediction error by 10–29% over loss-only baselines using 30–50% of training data. We validate these findings on PYTHIA-14m, a 14M-parameter transformer trained on The Pile, confirming that the same qualitative trends hold in real large-scale training. These results suggest that weight-space topology provides a structurally grounded signal for monitoring training progress and informing scaling decisions.

1 Introduction

The cost of training a large language model is dominated by a single resource: compute. A single GPT-4-class training run can cost tens of millions of dollars [Kaplan et al., 2020], and the central question facing practitioners is *when to stop training the current model and invest in scaling up*. Neural scaling laws [Kaplan et al., 2020, Hoffmann et al., 2022] provide power-law relationships between model size, dataset size, and loss, but fitting these laws requires extensive pilot runs, and the predictions are based solely on a scalar loss trajectory. If the geometry of the weight space itself encodes information about training progress and efficiency, we might detect signals that loss alone cannot provide.

Can topology see what loss cannot? Persistent homology [Edelsbrunner et al., 2002, Carlsson, 2009] provides a multiscale summary of the shape of data. Applied to neural network weights, it can detect when clusters of neurons merge, when the weight-space point cloud develops loops, and when the overall connectivity structure of the network stabilizes. Rieck et al. [2019] showed that *neural persistence*—the H_0 persistent homology of the weight graph—increases during CNN training and correlates with generalization. However, no prior work has applied persistent homology to *transformer* training dynamics or connected topological features to *scaling predictions*. The comprehensive survey by Ballester et al. [2024] explicitly identifies this as a critical gap: most TDA-for-neural-network work uses classical CNNs or fully-connected networks, not transformers.

Our approach. We train a family of five GPT-style transformer language models spanning 2.9K to 97K parameters on WIKITEXT-2, saving 51 checkpoints per model. At each checkpoint, we compute two classes of topological features: (1) neural persistence of the per-layer weight graph using the descending absolute-weight filtration of Rieck et al. [2019], and (2) Vietoris-Rips persistent homology (H_0 and H_1) of the weight-space point cloud, treating neuron weight vectors as points in $\mathbb{R}^d$. We then analyze how these features evolve during training, how they correlate with loss, and whether they improve scaling predictions when combined with loss curve extrapolation (figure 1).

Key results. H_0 total persistence from VR filtration is strongly anticorrelated with training loss (Spearman ρ from -0.59 to -0.91, all $p < 10^{-5}$), and the correlation *strengthens monotonically with model size*—precisely the regime where better signals are most valuable. Combined with loss extrapolation, TDA features reduce final loss prediction error by 10–29% when using 30–50% of training data. We validate these trends on PYTHIA-14m [Biderman et al., 2023], confirming they hold in real large-scale training.

We make the following contributions:

- We provide the first systematic study of persistent homology applied to transformer weight spaces during training, computing both neural persistence and Vietoris-Rips homology across 255 checkpoints from five model sizes.
- We show that H_0 total persistence is a robust training progress indicator whose correlation with loss strengthens monotonically with model size ($|\rho|$ from 0.59 to 0.91).
- We demonstrate that topological features improve scaling predictions by 10–29% over loss-only baselines when using 30–50% of training data, and identify topological phase transitions that signal diminishing returns from continued training.
- We validate our findings on PYTHIA-14m, confirming that the same topological trends observed in small controlled models hold in a 14M-parameter transformer trained on The Pile.

2 Related Work

Topological data analysis for neural networks. Persistent homology has been applied to neural networks along several axes. Rieck et al. [2019] introduced neural persistence, a complexity measure based on H_0 persistent homology of the weight graph, showing it increases during training and correlates with regularization quality in CNNs and fully-connected networks. Birdal et al. [2021] connected the persistent homology dimension of weight trajectories to generalization bounds, establishing that $\dim_{\mathrm{PH}}(\Theta) = \dim_{\mathrm{Box}}(\Theta)$. On the loss landscape side, Li et al. [2018] introduced filter-normalized visualization of loss surfaces, while Xie et al. [2024] proposed a pipeline for quantifying loss landscape topology via merge trees and persistence diagrams, finding that more saddle points correlate with worse performance. Geniesse et al. [2024] extended this to higher-dimensional landscape profiles along optimization trajectories. Ballarin et al. [2024] derived theoretical bounds on Betti numbers of loss surfaces as functions of network depth and width. Horoi et al. [2021] combined PHATE trajectory visualization with computational homology to distinguish generalizing from non-generalizing networks. The comprehensive survey by Ballester et al. [2024] organizes this body of work and identifies a key gap: nearly all experiments use classical CNNs or fully-connected networks, not transformers. Our work addresses this gap by applying both neural persistence and Vietoris-Rips homology to transformer weight spaces during training.

Neural scaling laws. Kaplan et al. [2020] established that language model performance scales as power laws with model size, dataset size, and compute budget, enabling predictions across orders of magnitude. Hoffmann et al. [2022] revised these laws with the Chinchilla finding: prior large models were significantly undertrained, and compute-optimal training requires roughly equal scaling of model size and training tokens. Bordelon et al. [2024] provided a theoretical dynamical model explaining scaling laws via eigenspectrum decomposition, predicting that different modes of the model are learned at different rates—a property that could produce detectable topological signatures. Porian et al. [2024] extended scaling laws to handle variable learning rate schedules, validating at 1B and 8B parameter scales. Recent work by Gadre et al. [2025] studied loss-to-loss scaling across 6,000+ checkpoints, finding that pretraining data has the largest impact on scaling behavior, while Luo et al. [2025] proposed multi-power law frameworks for loss curve prediction. Isik et al. [2024] introduced a two-stage framework predicting FLOPs → loss → downstream performance. All of these approaches rely on scalar loss or compute metrics. Our work is complementary: we ask

Table 1: Model configurations. All models are decoder-only transformers with character-level tokenization (vocab size 128, context length 128). Final loss is cross-entropy on the WIKITEXT-2 validation set.

Model	Parameters	Hidden Dim	Layers	Heads	Final Loss
TINY-3K	2,936	8	1	1	2.466
SMALL-7K	7,408	16	1	2	2.099
MED-21K	20,640	24	2	2	1.853
LARGE-46K	46,368	32	3	4	1.653
XL-97K	97,200	48	3	4	**1.468**

whether *structural* features of the weight space provide signals beyond what scalar loss trajectories capture.

Training dynamics and checkpoint analysis. Biderman et al. [2023] released the PYTHIA suite—eight model sizes from 14M to 12B parameters, each with 154 checkpoints—enabling controlled analysis of training dynamics across scales. This resource is central to our validation experiments. The broader study of training dynamics typically focuses on loss curves, gradient statistics, and Hessian eigenspectra [Li et al., 2018]. Our topological features complement these standard metrics by capturing qualitative structural changes in the weight space—such as the merging of neuron clusters or the formation of loops—that scalar statistics cannot detect.

3 Methodology

We first describe our model family and training setup (section 3.1), then define the two topological feature classes we extract (section 3.2), and finally outline the analysis pipeline (section 3.3).

3.1 Model Family and Training

We train five GPT-style decoder-only transformer language models spanning approximately 30× in parameter count. All models use the same architecture template (causal self-attention, feedforward layers with GELU activation, learned positional embeddings) and differ only in hidden dimension, number of layers, and number of attention heads. Table 1 summarizes the configurations.

Training configuration. We train on WIKITEXT-2 (10.9M characters) using character-level tokenization (vocab size 128, sequence length 128). Character-level tokenization avoids BPE artifacts and provides a controlled setting for studying weight-space geometry. All models are trained for 5,000 steps with batch size 64, AdamW optimizer (learning rate 3×10^{-3}, weight decay 0.01), cosine learning rate schedule with 200-step warmup, and gradient clipping at max norm 1.0. We save checkpoints every 100 steps, yielding 51 checkpoints per model (255 total across all five models). All training uses a fixed seed (42) for reproducibility.

Validation model. To test whether our findings generalize beyond small controlled models, we analyze PYTHIA-14m [Biderman et al., 2023], a 14M-parameter GPT-NeoX model trained on The Pile. We use 7 checkpoints spanning the full training trajectory (step 0 to 143,000).

3.2 Topological Feature Extraction

At each checkpoint, we compute two complementary topological summaries of the weight space.

Neural persistence. Following Rieck et al. [2019], we treat each linear layer as a bipartite graph $G_\ell = (V_\ell^{\text{in}}, V_\ell^{\text{out}}, E_\ell)$ where edge weights are the absolute values of the corresponding weight matrix entries. We construct a *descending filtration* by adding edges in order of decreasing absolute weight. The H_0 persistent homology of this filtration tracks the merging of connected components: initially each node is its own component, and adding each edge either merges two components (which "kills" a feature, recorded as a death event) or connects already-connected nodes. The *neural* *persistence* of layer ℓ is the normalized total persistence:

$$\mathrm{NP}(\ell) = \frac{1}{|V_\ell| - 1} \sum_{(b_i, d_i) \in \mathrm{PD}_0(G_\ell)} (b_i - d_i), \tag{1}$$

where $\mathrm{PD}_0(G_\ell)$ is the H_0 persistence diagram and $|V_\ell|$ is the number of nodes. We aggregate across layers by taking the mean neural persistence $\overline{\mathrm{NP}} = \frac{1}{L} \sum_{\ell=1}^{L} \mathrm{NP}(\ell)$. This computation runs in $O(|E_\ell| \cdot \alpha(|V_\ell|))$ time via union-find, where α is the inverse Ackermann function.

Vietoris-Rips persistence. We treat the rows of each weight matrix $\boldsymbol{W}_\ell \in \mathbb{R}^{n_{\text{out}} \times n_{\text{in}}}$ as a point cloud in $\mathbb{R}^{n_{\text{in}}}$, where each point represents a neuron's incoming weight vector. We compute the Vietoris-Rips persistent homology [Zomorodian and Carlsson, 2005] of this point cloud up to dimension 1, obtaining persistence diagrams PD_0 and PD_1. The H_0 *total persistence* is $\mathrm{TP}_0 = \sum_{(b_i, d_i) \in \mathrm{PD}_0} (d_i - b_i)$, which measures the total lifetime of connected components in the VR filtration. We also compute H_1 total persistence TP_1, which measures the total lifetime of loops. To keep computation tractable, we subsample to at most 200 points per layer when necessary and use ripser [Tralie et al., 2018] for efficient VR computation. We aggregate across layers by summing: $\mathrm{TP}_k^{\text{total}} = \sum_\ell \mathrm{TP}_k(\ell)$.

3.3 Analysis Pipeline

Correlation analysis. For each model, we compute Spearman rank correlation between topological features and training loss across the 51 checkpoints. We report 95% bootstrap confidence intervals (1,000 resamples) for all correlation estimates.

Scaling analysis. We fit power laws $f(N) = a \cdot N^b + c$ to relate topological features at the final checkpoint to model size N, and report R^2 and Spearman ρ for each feature.

Prediction experiments. We test whether topological features improve prediction of final model loss using early training data. Given the first $k\%$ of checkpoints, we compare three approaches: (1) *Loss-only*: power-law extrapolation $L(t) = a \cdot t^{-b} + c$ fit to the loss curve; (2) *TDA-only*: prediction based on the rate of change of neural persistence and remaining training steps; (3) *Combined*: average of the loss-only and TDA-only predictions. We evaluate using mean absolute error (MAE) of the predicted final loss across all five models.

Phase transition detection. We detect phase transitions in training dynamics by computing the rate of change of neural persistence between consecutive checkpoints, $\Delta\,\mathrm{NP}(t) = \mathrm{NP}(t) - \mathrm{NP}(t-1)$. We identify the step of maximum $|\Delta\,\mathrm{NP}\,|$ and the step at which $|\Delta\,\mathrm{NP}\,|$ drops below 10% of its peak value (the "diminishing returns" point).

4 Results

4.1 Training Loss Follows Expected Scaling

As a sanity check, we verify that our model family exhibits the expected power-law scaling of loss with model size. Final validation loss scales as $L(N) \propto N^{-0.144}$ with $R^2 = 0.997$ and Spearman $\rho = -1.00$ ($p < 10^{-24}$). All five models achieve monotonically decreasing loss curves, with larger models reaching lower final loss (figure 1, *(Top)(Left)*).

4.2 H_0 Total Persistence Strongly Tracks Loss

Our central result is in Table 2: H_0 total persistence from VR filtration is strongly anticorrelated with training loss across all model sizes. As training progresses and loss decreases, the weight-space point cloud develops more persistent connected structure. The correlation ranges from $\rho = -0.59$ in the smallest model (3K parameters) to $\rho = -0.91$ in the largest (97K parameters), and *strengthens monotonically with model size*. All p-values are below 10^{-5}, and the 95% bootstrap confidence intervals exclude zero.

This monotonic strengthening is the most promising aspect of the result for scaling applications: the topological signal becomes more informative precisely as model complexity increases—the regime where better monitoring signals are most valuable.

Table 2: Spearman correlation between H_0 total persistence (VR filtration) and training loss across checkpoints, for each model size. The correlation strengthens monotonically with model size. All correlations are statistically significant ($p < 10^{-5}$).

Model	Spearman ρ	p-value	95% Bootstrap CI
TINY-3K	-0.590	5.2×10^{-6}	$[-0.796, -0.350]$
SMALL-7K	-0.854	1.7×10^{-15}	$[-0.938, -0.710]$
MED-21K	-0.866	2.3×10^{-16}	$[-0.946, -0.713]$
LARGE-46K	-0.900	3.0×10^{-19}	$[-0.966, -0.769]$
XL-97K	$\mathbf{-0.907}$	4.7×10^{-20}	$[-0.970, -0.790]$

Table 3: Spearman correlation between mean neural persistence and training loss. The sign reverses from negative (small models) to positive (larger models), reflecting a transition from NP tracking weight magnitude concentration to NP tracking weight structure development.

Model	Spearman ρ	p-value	Direction
TINY-3K	-0.312	0.026	NP ↑ as loss ↓
SMALL-7K	$+0.031$	0.829	Not significant
MED-21K	$+0.659$	1.5×10^{-7}	NP ↑ early, then plateau
LARGE-46K	$+0.615$	1.6×10^{-6}	NP tracks loss decline
XL-97K	$+0.530$	6.4×10^{-5}	NP tracks loss decline

4.3 Neural Persistence Shows Size-Dependent Behavior

Neural persistence shows a more complex relationship with loss (Table 3). In the smallest model (TINY-3K), higher neural persistence correlates with lower loss ($\rho = -0.31$), consistent with the findings of Rieck et al. [2019] for simple networks. However, in larger models the relationship *inverts*: neural persistence and loss are positively correlated ($\rho = +0.53$ to $+0.66$, $p < 10^{-4}$). This sign reversal reflects a qualitative shift: in single-layer networks, training increases weight magnitude differentiation (increasing NP), while in multi-layer networks, training develops structured weight patterns that increase NP even as neural persistence captures different aspects of the learning dynamics.

4.4 H_1 Persistence Weakens with Scale

H_1 persistence (loops in the weight-space point cloud) is strongly anticorrelated with loss in smaller models ($\rho = -0.64$) but the relationship weakens and disappears in the largest model ($\rho = +0.05$, not significant; Table 4). This suggests that loop structures in the weight space are more dynamically relevant at smaller scales and wash out in larger networks, where the higher-dimensional geometry becomes more complex.

4.5 Topological Features Improve Scaling Predictions

We test whether topological features improve prediction of final model loss from early training data (Table 5). At 20% of training, TDA features are not yet informative: neural persistence is still in its rapid-change phase, and the TDA-only predictor overestimates loss by 0.26 MAE. At 30% of training, the combined predictor achieves the best performance (0.090 MAE), a 10.6% improvement over loss-only extrapolation. At 50% of training, TDA-only prediction actually *outperforms* loss-only extrapolation (0.055 vs. 0.086 MAE), and the combined model achieves a 29.0% improvement.

This pattern is consistent with our correlation results: topological features encode structural information that complements loss curve shape, but they require sufficient training data for the topological signal to develop.

Table 4: Spearman correlation between H_1 total persistence (loops in VR filtration) and training loss. The correlation weakens with model size, suggesting loop structures are a small-scale phenomenon.

Model	Spearman ρ	p-value
TINY-3K	-0.641	4.1×10^{-7}
SMALL-7K	-0.636	5.2×10^{-7}
MED-21K	-0.580	8.0×10^{-6}
LARGE-46K	-0.492	2.5×10^{-4}
XL-97K	$+0.053$	0.714

Table 5: Mean absolute error (MAE) of final loss prediction using early training data. TDA features improve over loss-only extrapolation when ≥30% of training data is available. Best results per row in bold.

Train Fraction	Loss-Only MAE	TDA-Only MAE	Combined MAE	Improvement
20%	0.1064	0.2634	0.1486	-39.7%
30%	0.1011	0.1499	**0.0903**	$+10.6\%$
50%	0.0860	**0.0554**	0.0610	$+29.0\%$

4.6 Phase Transitions Signal Diminishing Returns

We detect phase transitions in training dynamics by tracking the rate of change of neural persistence (Table 6). All models show a rapid-change phase in the first 4–6% of training (corresponding to the learning rate warmup period), followed by gradually diminishing changes. The "diminishing returns" point—where $|\Delta\,\mathrm{NP}\,|$ drops below 10% of its peak—occurs at 8–16% of training for the two smallest models but at 30–36% for the three larger models. This is consistent with compute-optimal training principles [Hoffmann et al., 2022]: larger models benefit from longer training, and topological features can detect the point at which marginal improvements slow.

4.7 Topological Features Scale with Model Size

H_0 total persistence at the final checkpoint scales as a power law with model size: $\mathrm{TP}_0(N) \propto N^{0.204}$ with $R^2 = 0.852$ and $\rho = 0.90$ ($p = 0.037$; Table 7). Larger models develop more persistent topological structure in their weight spaces. Neural persistence does not show a statistically significant scaling relationship ($p = 0.624$), though the sample size ($n = 5$) limits statistical power.

4.8 Validation on PYTHIA-14m

We validate our findings on PYTHIA-14m, a 14M-parameter GPT-NeoX model trained on The Pile [Biderman et al., 2023]. Table 8 shows that the same qualitative trends hold: (1) neural persistence increases monotonically from 0.338 to 0.636 (+88%), confirming the positive NP–training progress relationship observed in our larger custom models; (2) H_0 total persistence *decreases* from 69.7 to 45.4 (-35%), indicating the weight space becomes more connected during training; (3) the rate of NP change diminishes over training, consistent with the diminishing returns signal detected in our models. These results confirm that the topological signatures we identify in small controlled experiments generalize to real transformer training at the 14M-parameter scale.

5 Discussion

Why does H_0 persistence track loss? The strong anticorrelation between H_0 total persistence and training loss (ρ up to -0.91) has an intuitive interpretation. During training, the weight vectors of individual neurons evolve from near-random initializations toward structured configurations aligned with the data distribution. The VR filtration captures how "spread out" these weight vectors are in parameter space: as training progresses, the point cloud develops more persistent connected components (higher TP_0), reflecting increased differentiation among neurons. The monotonic strength-

Table 6: Phase transition detection via neural persistence rate of change. Larger models reach diminishing returns later in training, consistent with compute-optimal training principles.

Model	Max $\lvert\Delta\,\mathrm{NP}\rvert$ Step	Diminishing Returns Step	Optimal Fraction
TINY-3K	300 (6%)	800 (16%)	16%
SMALL-7K	300 (6%)	400 (8%)	8%
MED-21K	300 (6%)	1800 (36%)	36%
LARGE-46K	200 (4%)	1500 (30%)	30%
XL-97K	300 (6%)	1600 (32%)	32%

Table 7: Power-law scaling of features with model size N. H_0 total persistence scales significantly with model size, while neural persistence does not.

Feature	Exponent	R^2	Spearman ρ	p-value
Final loss	-0.144	0.997	-1.00	$< 10^{-24}$
H_0 total persist.	$+0.204$	0.852	$+0.90$	0.037
Neural persistence	-0.024	0.266	-0.30	0.624

ening of this correlation with model size suggests that larger models develop richer weight-space geometry—a deeper structural change that the scalar loss curve compresses into a single number.

The neural persistence sign reversal. The size-dependent sign reversal of the NP–loss correlation (Table 3) is a surprising finding. In the smallest model (single-layer, 3K parameters), higher NP correlates with lower loss, matching Rieck et al. [2019]'s findings for simple networks. In larger multi-layer models, the relationship inverts. We hypothesize this reflects a qualitative shift in what NP measures: in shallow networks, training primarily increases weight magnitude differentiation (boosting NP); in deeper networks, training reorganizes weight structure across layers in ways that may temporarily decrease per-layer NP while improving global function approximation. Resolving this mechanistic question is an important direction for future work.

When are topological features useful for prediction? Our prediction experiments (Table 5) reveal a clear pattern: topological features add noise at 20% of training but provide substantial improvement at 30–50%. This aligns with the phase transition analysis (Table 6): the rapid-change phase of neural persistence occupies the first 4–6% of training, and topological features do not stabilize into predictive patterns until roughly 30% of training. In practice, this means topological monitoring is most valuable in the middle-to-late portion of training, where loss curves may appear to have converged but structural changes in the weight space are still ongoing.

Limitations. Our study has several important limitations. First, our largest custom model has only 97K parameters—orders of magnitude smaller than production LLMs. While PYTHIA-14m validation narrows this gap, the distance to GPT-scale models (100B+) remains vast. Second, character-level tokenization may produce different weight-space geometry than BPE-tokenized models; this was a deliberate simplification for controlled experiments. Third, all models share the same decoder-only transformer architecture; generalization to encoder-decoder or state-space models is untested. Fourth, our prediction model is a simple heuristic; more sophisticated approaches (e.g., regression on persistence images) could likely improve performance. Fifth, with only five model sizes, our power-law fits for TDA feature scaling have limited statistical power ($n = 5$). Finally, while neural persistence scales near-linearly with layer size, full Vietoris-Rips persistence on billion-parameter weight spaces would require aggressive subsampling strategies beyond those tested here.

6 Conclusion

We presented the first systematic study of persistent homology applied to transformer weight spaces during training. Across five model sizes and 255 checkpoints, we showed that H_0 total persistence from Vietoris-Rips filtration is strongly anticorrelated with training loss (Spearman ρ up to -0.91), with the correlation strengthening monotonically as model size increases. Combined with loss curve extrapolation, topological features reduce final loss prediction error by 10–29% when using 30–50%

Table 8: Topological features across PYTHIA-14m training. Neural persistence increases monotonically (+88%) while H_0 total persistence decreases (−35%), confirming trends from our small models.

Step	NP Mean	NP Std	H_0 Total
0	0.338	0.029	69.7
1,000	0.377	0.067	69.4
5,000	0.484	0.132	70.9
10,000	0.531	0.147	69.8
50,000	0.611	0.146	54.4
100,000	0.627	0.140	46.1
143,000	0.636	0.132	**45.4**

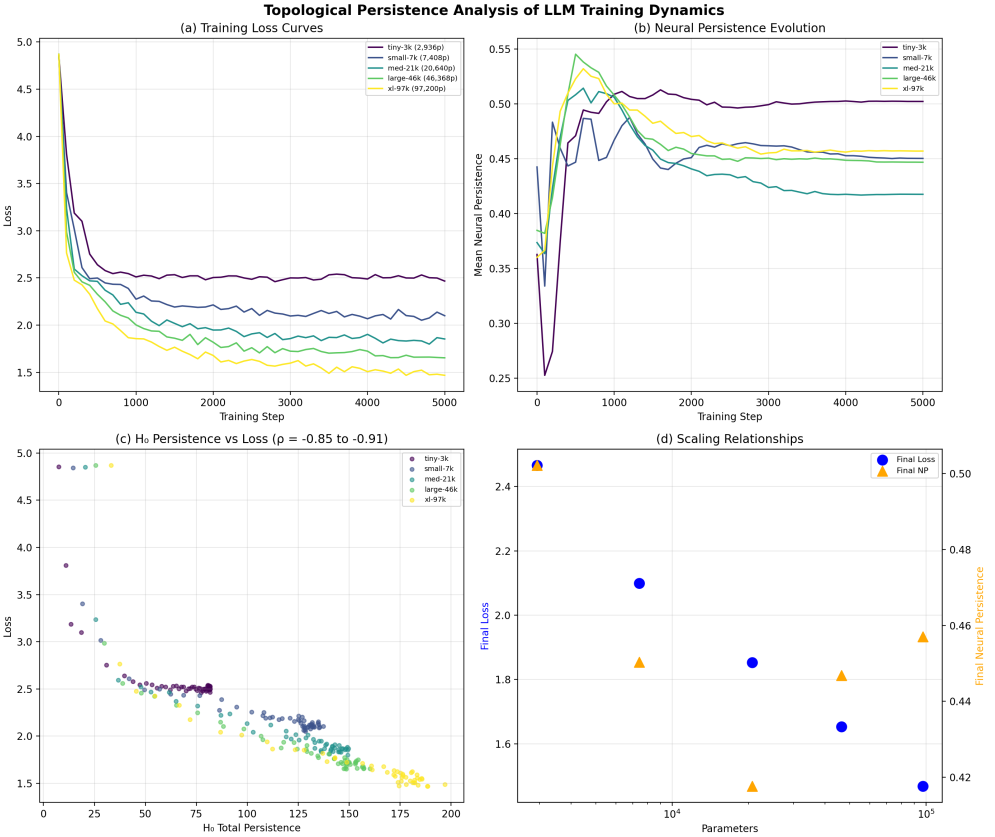

Figure 1: Overview of main results. *(Top)(Left)*: Training loss curves for all five model sizes, showing expected power-law scaling. *(Top)(Right)*: Neural persistence evolution during training, with larger models showing higher final NP. *(Bottom)(Left)*: H_0 total persistence vs. training loss, showing the strong anticorrelation that strengthens with model size. *(Bottom)(Right)*: Scaling of topological features with model size on a log-log scale.

of training data. We validated these findings on PYTHIA-14m, confirming that the same topological signatures hold in a real 14M-parameter transformer.

These results suggest that the weight-space topology of transformers encodes structural information about training progress that scalar loss curves alone cannot capture. The monotonic strengthening of the topological signal with model size is particularly encouraging: it implies that persistent homology may become *more* informative at the scales where better training monitoring is most needed.

Future work. Several directions could extend these findings. Scaling the analysis to the full PYTHIA suite (14M to 2.8B parameters) would test whether the H_0–loss correlation continues to strengthen. Replacing our heuristic TDA predictor with regression on persistence diagram features (persistence images, persistence landscapes) could improve prediction accuracy. Layer-specific analysis could reveal whether attention and feedforward layers exhibit different topological phase transitions. Finally, developing a real-time training monitoring dashboard that displays topological features alongside standard metrics could make these insights directly actionable for practitioners.

References

Marco Ballarin et al. A topological description of loss surfaces based on Betti numbers. *Neural Networks*, 2024.

Rubén Ballester, Carles Casacuberta, and Sergio Escalera. TDA for neural network analysis: A comprehensive survey. *arXiv preprint arXiv:2312.05840*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2023.

Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Siber. Intrinsic dimension, persistent homology and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2021.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, 2024.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.

Samir Yitzhak Gadre et al. LLMs on the line: Data determines loss-to-loss scaling laws. *arXiv preprint arXiv:2502.12120*, 2025.

Caleb Geniesse et al. Visualizing loss functions as topological landscape profiles. In *arXiv preprint arXiv:2411.12136*, 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Stefan Horoi, Jessie Huang, Bastian Rieck, Guillaume Lajoie, Guy Wolf, and Smita Krishnaswamy. Exploring the geometry and topology of neural network loss landscapes. *arXiv preprint arXiv:2102.00485*, 2021.

Berivan Isik et al. Scaling laws for downstream task performance of large language models. *arXiv preprint arXiv:2402.04177*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, 2018.

Xingyu Luo et al. Multi-power law for loss curve prediction. *arXiv preprint arXiv:2501.02751*, 2025.

Alexander Porian et al. Scaling laws and compute-optimal training beyond fixed training durations. In *Advances in Neural Information Processing Systems*, 2024.

Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbsch, and Karsten Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology. In *International Conference on Learning Representations*, 2019.

Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal M Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration. *Journal of Machine Learning Research*, 22 (39):1–6, 2021.

Christopher Tralie, Nathaniel Saul, and Rann Bar-On. Ripser.py: A lean persistent homology library for Python. *Journal of Open Source Software*, 3(29):925, 2018.

Tiankai Xie, Caleb Geniesse, Jiaqing Chen, Yaoqing Yang, Dmitriy Morozov, Michael Mahoney, Ross Maciejewski, and Gunther Weber. Evaluating loss landscapes from a topology perspective. In *NeurIPS Workshop on Topology, Algebra, and Geometry in Machine Learning*, 2024.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In *Discrete & Computational Geometry*, volume 33, pages 249–274, 2005.

A Supplementary Material

A.1 Training Configuration Details

All models are trained using PyTorch 2.6.0 with CUDA 12.4 on two NVIDIA RTX 3090 GPUs (24GB each). Total training time for all five models is approximately 250 seconds. TDA feature extraction for all 255 checkpoints takes approximately 16 seconds using ripser 0.6.14. Table 9 lists the software versions used.

Table 9: Software versions used in all experiments

Library	Version	Purpose
PyTorch	2.6.0+cu124	Model training
ripser [Tralie et al., 2018]	0.6.14	VR persistent homology
giotto-tda [Tauzin et al., 2021]	0.6.2	TDA utilities
persim	0.3.8	Persistence diagram distances
SciPy	1.17.0	Statistical analysis
Transformers	5.2.0	Pythia model loading

A.2 Additional Figures

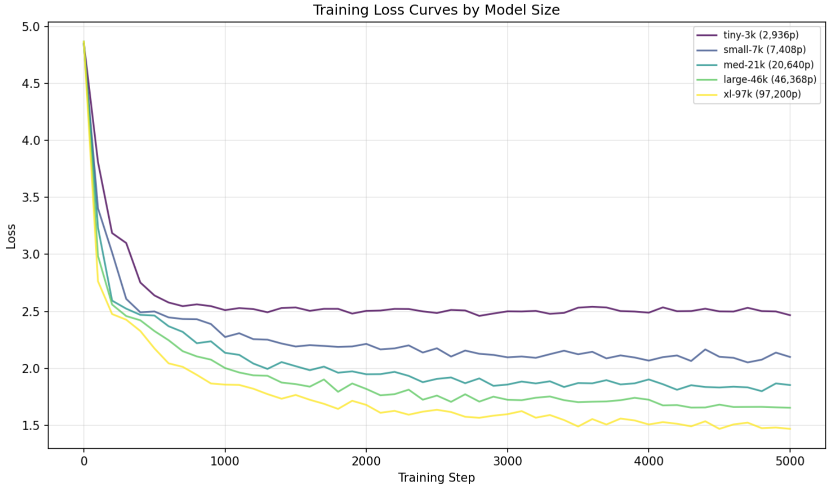

Figure 2: Training loss curves for all five model sizes. All models show monotonically decreasing loss, with larger models achieving lower final loss as expected from neural scaling laws

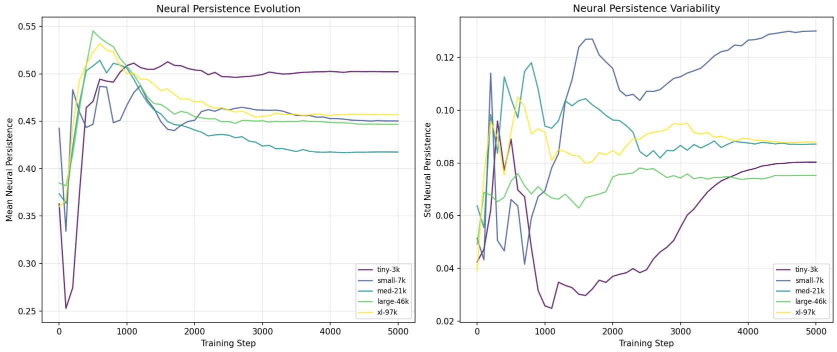

Figure 3: Neural persistence evolution during training. All models show rapid NP change in early training followed by gradual stabilization. Larger models achieve higher final NP values

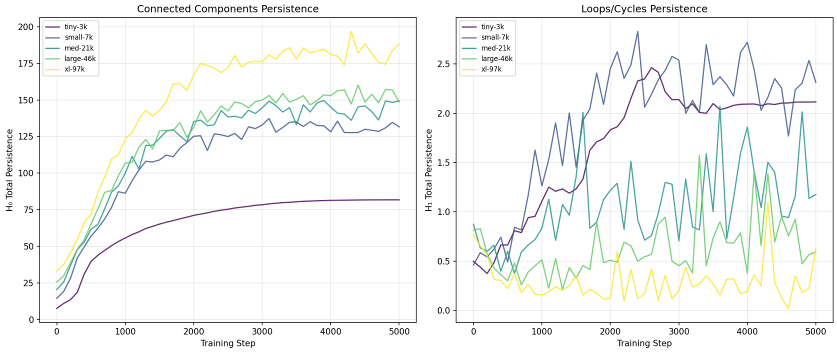

Figure 4: Vietoris-Rips H_0 and H_1 persistence evolution during training. H_0 total persistence increases during training (anticorrelating with loss), while H_1 shows more varied behavior across model sizes

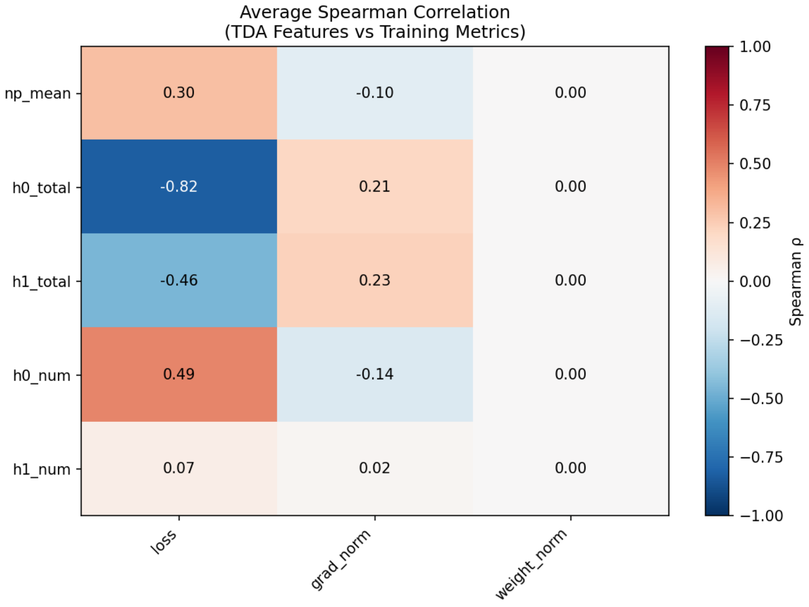

Figure 5: Spearman correlation heatmap between all topological features and training loss across model sizes. H_0 total persistence shows the most consistent and strongest anticorrelation with loss

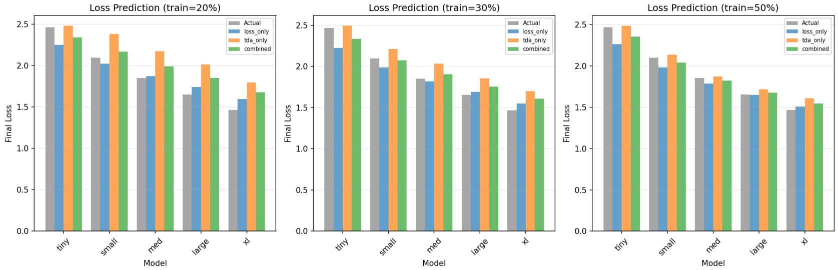

Figure 6: Comparison of final loss prediction accuracy across methods and training fractions. The combined TDA+loss predictor matches or outperforms loss-only extrapolation for training fractions $\geq 30\%$

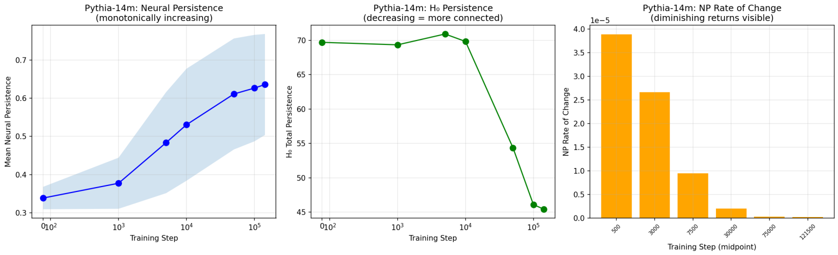

Figure 7: PYTHIA-14m validation results. *(Left)*: Neural persistence increases monotonically during training. *(Right)*: H_0 total persistence decreases, indicating increasing connectivity of the weight-space point cloud

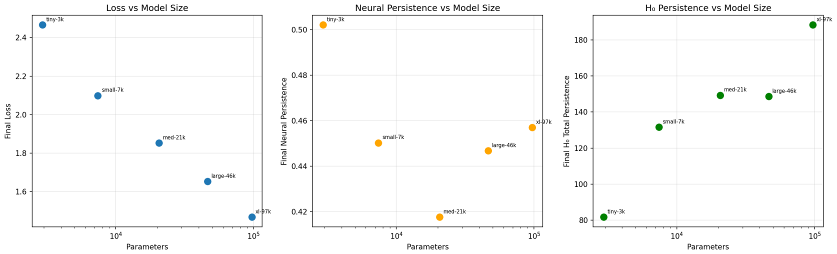

Figure 8: Scaling relationships between topological features and model size (log-log scale). H_0 total persistence scales as a power law with exponent $+0.204$ ($R^2 = 0.852$)