THE SHAPE OF ADVERSARIAL INFLUENCE: CHARACTERIZING LLM LATENT SPACES WITH PERSISTENT HOMOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing interpretability methods for Large Language Models (LLMs) often fall short by focusing on linear directions or isolated features, overlooking the high-dimensional, nonlinear, and relational geometry within model representations. This study focuses on how adversarial inputs systematically affect the internal representation spaces of LLMs, a topic which remains poorly understood. We propose the application of persistent homology (PH) to measure and understand the geometry and topology of the representation space when the model is under external adversarial influence. Specifically, we use PH to systematically interpret six state-of-the-art models under two distinct adversarial conditions—indirect prompt injection and backdoor fine-tuning—and uncover a consistent topological signature of adversarial influence. Across architectures and model sizes, adversarial inputs induce "topological compression", where the latent space becomes structurally simpler, collapsing from varied, compact, small-scale features into fewer, dominant, and more dispersed large-scale ones. This topological signature is statistically robust across layers, highly discriminative, and provides interpretable insights into how adversarial effects emerge and propagate. By quantifying the shape of activations and neuron-level information flow, our architecture-agnostic framework reveals fundamental invariants of representational change, offering a complementary perspective to existing interpretability methods.

1 INTRODUCTION

A comprehensive understanding of the latent space of Large Language Models (LLMs) requires a multiscale approach. LLM representations form a conceptual hierarchy, with local-scale individual neurons encoding simple features such as punctuation (Tenney et al., 2019; Hewitt & Manning, 2019), intermediate-scale circuits forming contextual associations (Meng et al., 2023), and global-scale activation patterns representing more abstract concepts (Burns et al., 2024). However, most empirical work assumes a linear structure, neglecting the complex geometry of these high-dimensional activation spaces (Brüel-Gabrielsson et al., 2020; Engels et al., 2025). This oversight creates a practical security gap in real-world models, allowing diverse attacks to exploit nonlinear features and bypass the prevalent defenses that rely on linear classifiers (Kirch et al., 2024).

In this paper, we address this gap by studying LLM hidden states using *persistent homology* (PH), which is a technique from topological data analysis (TDA) that captures the multi-scale shape of data (Chazal & Michel, 2021). PH is uniquely suited for this task because it provides a coordinate-free summary of relational geometry that is known to be robust to noise (Cohen-Steiner et al., 2007). Unlike methods that project high-dimensional representations onto lower-dimensional subspaces, PH preserves multi-scale structural information through a filtration, capturing both local clustering patterns and global topological features simultaneously. These properties enable direct and meaningful comparisons of latent space structure across different models, input distributions, and fine-tuning stages. This information is quantified and encoded in a *barcode*—a summary statistic of the evolution of topological features. As shown in Figure 1, these barcodes elucidate a clear distinction between normal and adversarial activations, motivating our deeper investigation.

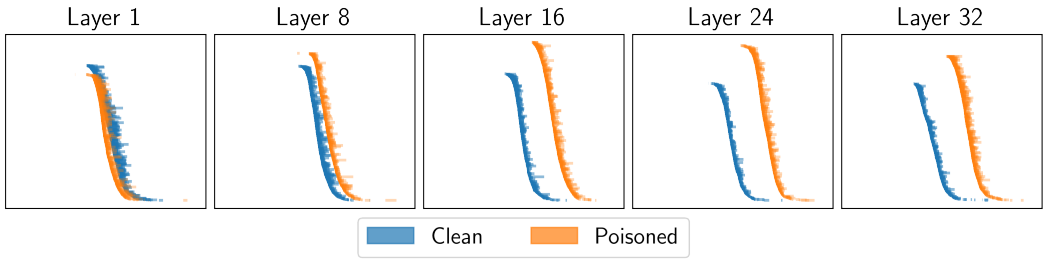


Figure 1: **Example barcodes from clean vs. poisoned activations.** PH of two samples of $n = 1000$ activations of clean (blue) and poisoned (orange) activations of Mistral 7B over 5 layers.

Our contributions can be summarized as follows.

- We present a comprehensive study of six state-of-the-art models under two fundamentally different attack modes revealing that adversarial inputs induce *consistent topological behavior within the LLM latent space*. Specifically, adversarial inputs cause latent representations to become more dispersed, characterized by fewer but more topologically significant large-scale features. In contrast, normal inputs produce a greater diversity of compact, small-scale structures.
- We show that this phenomenon *holds across models ranging from 7B to 70B parameters*, suggesting that adversarial triggers systematically reshape the representation space in a consistent and predictable manner that is independent of specific architectures or training procedures.
- We introduce a novel, *neuron-level PH analysis* confirms these geometric shifts at a finer scale, revealing a *phase transition in the topological complexity* of the information flow.

While standard linear classifiers can also separate normal and adversarial states with high accuracy, our topological framework provides an interpretable, geometric explanation for why this separability exists. These findings establish PH as a powerful complementary tool for interpretability and support the view that the success of linear probes may stem from their approximation of more complex, underlying topological structures (Engels et al., 2025; Park et al., 2024; Yang et al., 2024).

2 Background

In this section, we outline PH and the barcode summaries we study; we also provide details on the specific types of adversarial influence we investigate.

2.1 Persistent Homology and Persistence Barcodes

PH is a powerful methodology to quantify the "shape" and "size" of data, which can be applied to diverse input data types, is robust to noise perturbations, captures higher-order relational information and has an inherently interpretable nature. More precisely, PH captures *topological features*, e.g., connected components, tunnels and loops, or cavities and bubbles, present at different scales in our data.

For our activation data, i.e., point clouds $X \subset \mathbb{R}^D$, with D the hidden dimension of the model (typically, $D = 4096$) and where each point is the latent representation of the last token in a prompt in a given layer; the PH pipeline proceeds as follows. The first step is to construct a dynamic, geometric representation of our point cloud. A classical construction involves the *Vietoris–Rips* complex, which for a scale parameter $\epsilon > 0$ is obtained from the *ϵ-neighborhood graph*, that is, the graph where we connect any two points at distance less than ϵ. The Vietoris–Rips complex goes beyond the pairwise interactions in the ϵ-neighborhood graph including higher-order relational information, namely, interactions between more than two points at the same time, known as *simplices*: 0-simplices correspond to points, 1-simplices to edges, 2-simplices to triangles, 3-simplices to tetrahedra, and so on. We add a simplex between a subset of 3 or more points to the Vietoris–Rips

complex whenever they are all pairwise connected, for instance, we add a triangle if three points in the point cloud are connected in the ϵ-neighborhood graph. This completes the Vietoris–Rips complex construction. Considering all scale parameters ϵ at the same time, we obtain the *Vietoris–Rips filtration*: a growing family of geometric spaces where we connect points and add simplices as the parameter ϵ grows.

PH then leverages algebraic topology to produce the *persistence barcode*, a collection of bars capturing how the topological features are formed and disappear in the filtration as the scale parameter ϵ increases. The barcode is stratified in different dimensions, here we focus in dimensions 0 and 1. Bars in the 0-dimensional barcode (or 0-bars) correspond to connected components: at $\epsilon = 0$ there are as many bars in the barcode as points in the data, with bars terminating as point get connected in the ϵ-graph. 1-bars represent loops or cycles in the corresponding Vietoris–Rips complex: a bar starts whenever we have added enough edges to enclose a non-trivial hole, and ends when the addition of triangles covers said hole. Usually, the starting point is called the *birth* and the ending point the *death* of the bar. An illustrative example of the PH pipeline in a simple point cloud and the corresponding barcode can be found in Figure 2. See Appendix A.1 for more details on the PH construction.

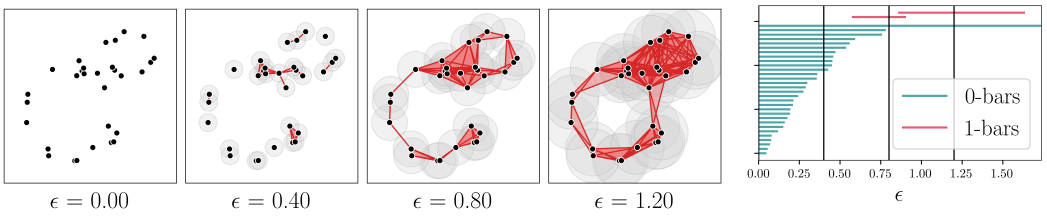


Figure 2: **Left:** Vietoris–Rips filtration constructed from a sample of 50 points over 2 circles with noise, at four values of the distance threshold $\epsilon \in [0, \infty)$. **Right:** corresponding persistence barcode for the 0- and 1-bars, with vertical lines corresponding to the thresholds displayed on the left.

2.2 Persistent Homology in Machine Learning: Barcode Summaries

Persistence barcodes cannot be directly used as input features in a ML model since they do not reside in a Euclidean space (Turner et al., 2014). We circumvent this issue by studying summary statistics of barcodes (Ali et al., 2023)—such as the mean, standard deviation, median, or quartiles—of the empirical distributions of the births, deaths, and *persistences* (lengths) of the bars in a given barcode. We can also study the empirical distribution of the ratios between births and deaths, which have the advantage of being scale invariant; the number of bars, providing a notion of topological diversity; the total persistence, which is given by the sum of the lengths of all bars in the barcode and captures both the number of topological features and their size; and the *persistent entropy* (Chintakunta et al., 2015; Rucco et al., 2016) of each barcode, which intuitively measures the heterogeneity within the lengths of the bars in the barcode. In all, for each barcode, we compute a 41-dimensional descriptive feature vector that can be used in machine learning tasks, which we call the *barcode summary*.

2.3 Adversarial Influence on LLMs

The use of PH to analyze activation space is not new. Naitzat et al. (2020) demonstrated that well-trained neural networks tend to simplify input-data topology to facilitate class separation. Subsequent work (Wheeler et al., 2021) employed persistence landscapes to provide a more detailed characterization of activation-space evolution. PH has also been applied to the study of trojaned networks by computing barcodes from simplicial complexes constructed via activation correlations, and it has seen increasing use in the analysis of LLMs (see Uchendu & Le (2024) for a survey of TDA in NLP). To the best of our knowledge, however, our work is the first to connect these research threads and to demonstrate the utility of PH as a practical tool for geometric and quantitative insights into LLM representation spaces under adversarial influence. In order to test the generality of our approach, we quantify and interpret the effects of two systematically different attack modes, both of which carry a high security impact in practice: *Indirect Prompt Injection (XPIA)*, where attackers embed hidden instructions in retrieved content to override a user's original prompt (Greshake et al.,

2023; Rehberger, 2024); and *sandbagging via backdoor fine-tuning*, which involves deliberately training a model to suppress its capabilities until a secret trigger is provided (Greenblatt et al., 2024; van der Weij et al., 2024). These techniques target fundamentally distinct vulnerabilities: XPIA exploits the model's core inability to distinguish data from instructions (Zverev et al., 2025), whereas sandbagging affects the fine-tuning process.

3 Experimental Design

In this section we overview details of the data we study, and outline our global and local studies using persistent homology.

3.1 Data and Representations

We compute the barcodes of point clouds in $\mathbb{R}^D$, where each point corresponds to the latent representation of the last token of a given input in a given layer, or of a 2D embedding described in Section 3.3. The choice of the last token is justified by its role in encoding the model's aggregated interpretation of the input context (Zou et al., 2023).

TaskTracker (Clean vs. Poisoned). The TaskTracker (Abdelnabi et al., 2024) test set comprises $>$ 62k examples, where each example contains the activation values for an input that is either *clean* or *poisoned*. Each input combines a user prompt with a retrieved data block. In poisoned cases, the data block contains an injected secondary task from several standard safety and security benchmarks, including BeaverTrails (Ji et al., 2023), HarmBench (Mazeika et al., 2024), and JailbreakBench (Chao et al., 2024). These attacks are applied across diverse task prompts, such as coding challenges and question-answering, to create realistic test cases. We analyze representation data from six instruction tuned LLMs; Phi3-mini-4k (3.8B parameters) and Phi3-medium-128k (14B parameters) (Abdin et al., 2024), Mistral 7B (Jiang et al., 2023), LLaMA3 8B and 70B (Grattafiori et al., 2024), and Mixtral-8x7B (Jiang et al., 2024).

Sandbagged Models (Locked vs. Elicited). We fine-tuned Mistral 7B and LLaMA3 8B on a corpus of 3,347 synthetic questions inspired by the WMDP benchmark (Li et al., 2024) using LoRA (Hu et al., 2021). Following the methodology of van der Weij et al. (2024), we created models with two operational modes: *locked* (default state) and *elicited* (activated by a specific password in the prompt). We extracted final-token representations from each layer during inference with evaluation prompts in both states. Further implementation details are provided in Appendix E.

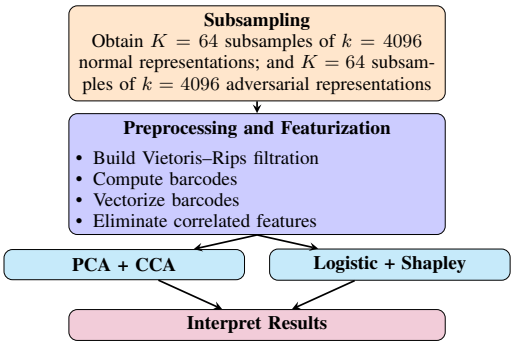


Figure 3: Pipeline for layer-wise topological analysis.

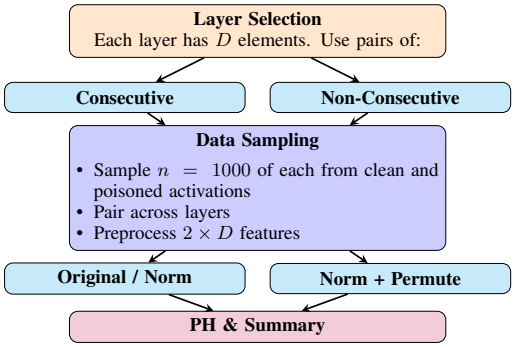


Figure 4: Pipeline for local analysis.

3.2 Global Layer-Wise Analysis

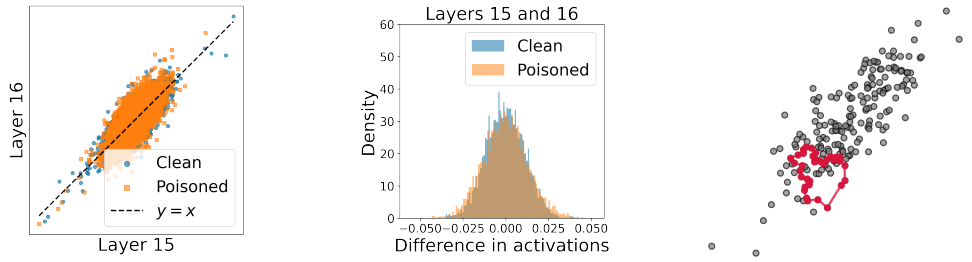
This analysis establishes and explains a consistent topological distinction between normal and adversarial representations, following the pipeline in Figure 3. We used Ripser++ (Bauer, 2021; Zhang et al., 2020) to compute barcodes, leveraging subsampling techniques, both to reduce the

computational cost of PH and to enable statistically robust inference. Subsampling approaches in PH are theoretically grounded, as under mild sampling models, persistence diagrams estimated from point clouds converge to the population diagrams with guaranteed rates (Chazal et al., 2015; 2014). For each model layer, we drew $K = 64$ subsamples of $k = 4096$ normal representations; and $K = 64$ subsamples of $k = 4096$ adversarial representations—see Appendix C.2 for ablations. We vectorized the corresponding barcodes into 41-dimensional barcode summaries (cf. Section 2.2), and performed the analysis in Figure 3, see results and further details in Section 4.1.

3.3 Local Information Flow Analysis

This analysis quantifies neuron-level information flow by tracking topological changes in activation patterns between layers. For each pair of layers ℓ and ℓ', we construct a 2D point cloud from their corresponding D-dimensional activation vectors. Each of the D points in this embedding has coordinates $(v_i^\ell, v_i^{\ell'})$, representing the activation of the ith neuron in layer ℓ and layer ℓ', respectively.

The rationale for this embedding is that activations between consecutive layers are empirically highly correlated, causing points to cluster near the identity line $y = x$, as shown in Figure 5a. Significant transformations in network processing are reflected in neurons whose activations deviate from this line, producing topological structures (e.g., loops) that PH captures and quantifies. We apply this analysis to 1000 clean and 1000 adversarial activation samples to compare the resulting topological signatures, which are presented in Section 4.2.



(a) (b) (c)

Figure 5: **(a):** Example 2D embedding showing correlation of activations in consecutive layers. **(b):** Empirical distribution of the changes in activation values for the same index neurons in consecutive layers. **(c):** Cycle corresponding to a long 1-bar in the PH barcode of the point cloud in (a).

4 Results

We now present the implementation results of our proposed analyses to the data described above.

4.1 Global Analysis: The Shape of Adversarial Influence

Our global analysis, as outlined in Figure 3, reveals a consistent and highly discriminative topological signature of adversarial influence across all six LLMs. Specifically, we show that adversarial inputs induce a "topological compression" of the latent space. Here, we present the results of quantifying and interpreting the effect of XPIA on Mistral 7B's latent space. Results for the other five models are relegated to Appendix C.3. Results for the Mistral 7B and LLaMA3-7B models subjected to the backdoor finetuning attack for sandbagging are given in Appendix C.4.

Cross-Correlation Analysis of Barcode Summaries. In Figure 6, a growing block of highly correlated features appears in the cross-correlation matrix of the 41 features of the barcode summaries. To reduce redundancy and prevent overfitting, we removed highly correlated variables, ensuring an efficient and informative representation for more parsimonious models in subsequent analyses. We discarded all features that have a correlation higher than a threshold of 0.5 with at least one feature present in the analysis, resulting in the features in Table 6. We refer to this data set as the *pruned barcode summaries*. The first feature appearing in this block of highly correlated features is the mean deaths of the 0-bars (the average of their ending points), which is retained in the pruned barcode summaries as representative of the block. However, we remark that the prominence of this

statistic in the results of our analysis does not imply a lack of significance for higher-order topological features (specifically, 1-bars). Empirically, there is a strong correlation between statistics of the 0- and 1-bars in our results; theoretically, it is known that the deaths of 0-bars are closely linked to the births of 1-bars (which has been explored using Morse theory; see Adler & Taylor (2011)).

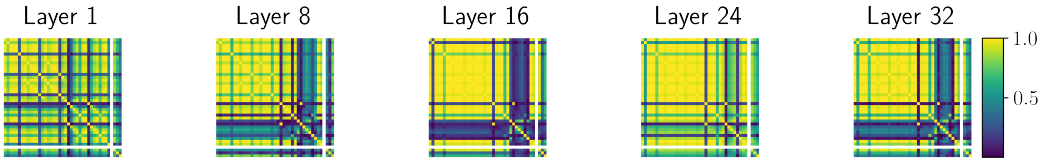


Figure 6: **Cross-correlation matrices for the barcode summaries** for clean vs. poisoned activations.

Geometric Separation of Latent States. The projection of the pruned barcode summaries over their first two principal components (Figure 7) yields a clear separation between subsamples from normal and adversarial modes across layers. This is consistent with the intuition presented in Figure 1, where a single barcode of a clean sample with $n = 1000$ activations (corresponding to a point in the PCA plot) was visibly different than the barcode of a poisoned sample with same number of points. This separation signals a difference in topology between clean and poisoned subsamples; we now seek to characterize such distinction, and to test whether it is consistent across layers and models.

To that end, we investigated the importance of particular features in the PCA results via a cross correlation analysis (CCA) between the pruned barcode summaries and the principal components of the PCA. CCA is a statistical method that quantifies linear relationships between two multivariate datasets by finding pairs of canonical variables with maximal correlation. The *loadings* are the contributions of individual features to these canonical variables, measuring their importance in capturing the relationship. We found that mean deaths of the 0-bars ranked first in all layers, and that the number of 1-bars appeared as a significant statistic as well (see Figure 18).

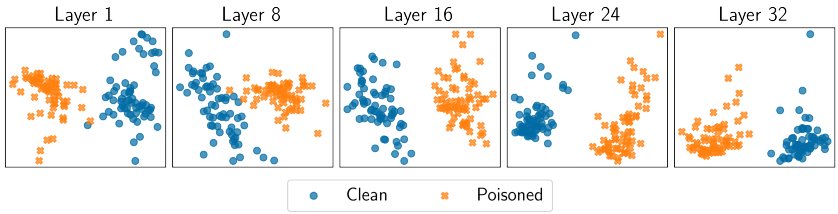


Figure 7: **PCA of pruned barcode summaries of clean vs. poisoned activations**. Clear distinction appears in the two first PC projections from the PCA of the pruned barcode summaries for layers 1, 8, 16, 24, and 32. The explained variances are 0.59, 0.49, 0.52, 0.96 and 0.83, respectively.

Discriminative Power of Topological Features. We tested the power of the pruned barcode summaries in distinguishing normal and adversarial subsamples by training a logistic regression with a 70/30 split between train and test. We obtained perfect accuracy and AUC–ROC on the test data, and 5-fold cross validation over the training data (Figure 8). As a baseline comparison, we trained a linear discriminant analysis (LDA), a linear support vector machine (SVM), and a logistic regression to distinguish 1000 clean and 1000 poisoned activations, raw and after reducing dimensionality using a sparse autoencoder (AE) with hidden dimension 128; see Table 1 for results. We found that the barcode summaries outperform these methods in general, particularly for early layers. However, we emphasize that the information that they encode must be understood as complementary to that of the linear methods above, and that our true interest in the outstanding predictive power of barcode summaries resides in the fact that feature importance methods applied to the trained logistic regression allow us to interpret the differences in topology between clean and poisoned data, which is our ultimate goal.

We used Shapley (or SHAP) values to interpret the excellent performance of the regression model. Shapley values quantify the contribution (with sign) of each feature to the prediction of the model for a given input. Our analysis revealed that the mean of 0-bar deaths and the number of 1-bars strongly

Table 1: **Comparison of predictive power with linear methods**. Accuracy, with a 70/30 train/test split, of a linear discriminant analysis (LDA), a linear SVM and a logistic regression (LR) trained to distinguish 1000 raw clean activations from 1000 raw poisoned activations, with or without reducing the dimensionality of the data using a sparse autoencoder (SAE); and our method using PH.

Layer	LDA	LDA (SAE)	SVM	SVM (SAE)	LR	LR (SAE)	PH
Layer 1	0.995	0.995	0.8875	0.7400	0.8700	0.7425	1.0000
Layer 8	1.000	0.998	1.0000	0.6425	0.9950	0.6225	1.0000
Layer 16	1.000	0.9975	1.0000	0.8125	1.0000	0.6725	1.0000
Layer 24	1.000	0.9975	1.0000	0.9975	1.0000	0.9600	1.0000
Layer 32	1.000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

influence predictions, exhibiting a clear dichotomous effect: points with smaller mean death in their 0-bars and bigger number of 1-bars are typically classified as clean, whereas points with bigger mean death of their 0-bars and smaller number of 1-bars are classified as poisoned.

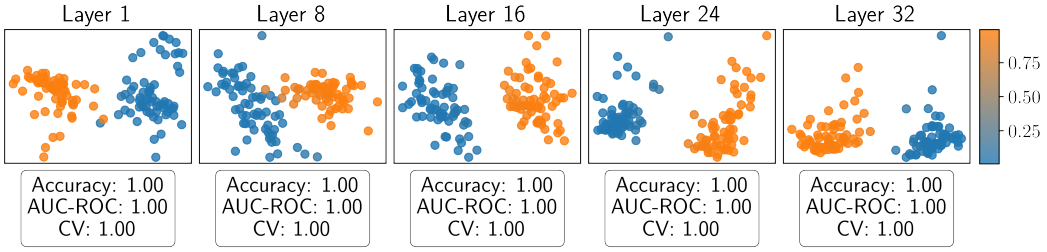


Figure 8: **Logistic regression for clean vs. poisoned activations** trained on a 70/30 train/test split of the pruned barcode summaries, plotted on the projection onto the two first PCs. Accuracy and AUC–ROC on the test data and 5-fold cross validation on train data are presented for each model.

The Signature of Topological Compression. Interpreting the distributions of the barcode summaries for clean vs. poisoned data reveals that adversarial conditions typically yield fewer 1-bars (loops) forming at later scales, yet persisting longer (see Figure 20). Conversely, the non-adversarial conditions tend to form earlier loops with more uniform lifetimes (higher persistent entropy). This pattern aligns with the Shapley value results (Figure 19): lower mean death times of 0-bars (i.e., more compact point clouds) are associated with predictions of "clean", while higher values (more spread-out clouds) shift predictions toward "poisoned". Similarly, a lower number of 1-bars tends to indicate "poisoned", whereas a higher count suggests "clean". Thus, global topological features point to a consistent distortion: adversarial states "compress" the representation space in a way that results in larger loops in fewer directions, while non-adversarial states exhibit many smaller loops with a more evenly distributed, higher-entropy shape. This signature is robust, persisting even against adaptive attacks from the `LLMail-Inject` public red teaming dataset that were designed to evade activation-based defenses (see Appendix G). A more detailed analysis across all models, layers, and adversarial conditions is provided in Appendix C and summarized in Table 2.

Local Dispersion Ratio Across Poisoned Conditions. To quantify how poisoning alters localized geometry in hidden-layer representation space, we use the *local dispersion ratio (LDR)*. For each final token's activation difference vector we identify its k nearest neighbors in each layer and perform PCA on those points. Let $\lambda_1 \geq \cdots \geq \lambda_{D'}$ be the resulting eigenvalues. The *dispersion ratio* is then defined as $\frac{\sum_{j=2}^{D'} \lambda_j}{\lambda_1 + \epsilon}$, where ϵ prevents division by zero. A higher LDR indicates that variance is more evenly spread among secondary directions, whereas a lower LDR implies most variance lies in a single dominant direction. Appendix B.3 further stratifies poisoned conditions into executed, refused, and ignored subclasses and shows that executed and ignored attacks exhibit elevated LDR in mid-layers relative to clean prompts. This indicates that the model allocates additional representational capacity to elaborating the injected instructions, whereas refused attacks are mapped into a more compressed, low-dispersion region, directly linking layer-wise geometric changes to task-level model behavior. Figure 9 shows that LDR differences remain tightly centered around zero under

Table 2: Summary of results for the global layerwise topological analysis across models and attacks.

	Clean vs. Poisoned	Locked vs. Elicited
Models evaluated	Phi3-mini-4k (3.8B), Phi3-medium-128k (14B), Mistral 7B, LLaMA3 8B, LLaMA3 70B, Mixtral-8×7B.	Mistral 7B, LLaMA3 8B.
Cross-correlation	Compact block of highly correlated features across layers.	Correlations weaker overall, especially in late layers.
PCA separation	Clear separation across layers & models.	Clear separation across layers & models.
Logistic regression	Perfect accuracy except LLaMA3-8B (0.99 at layers 1, 8, 16).	Perfect accuracy except Mistral-7B (0.99 at layer 16) and LLaMA3-8B (0.97–0.99 at layers 24, 32).
Mean death of 0-bars	SHAP: low values → normal, high → adversarial. Reversed only at layer 1 (all models) and layer 2 (LLaMA3-70B)	Early layers: low → adversarial. Layer 16: trend shifts. Late layers: reversed.
Mean persistence of 1-bars	Normal samples lower; Mixtral-8×7B flips this in the last layer.	Mistral-7B shifts at layer 16 (early: normal higher, late: lower). LLaMA3-7B similar early trend; late layers inconclusive.
Number of 1-bars	Generally lower for adversarial samples, except LLaMA3-70B.	Mistral-7B: no clear pattern. LLaMA3-7B: adversarial larger in later layers.
Topological compression	Appears early; LLaMA3-70B compresses without increased diversity.	Appears later; heterogeneous patterns (e.g., larger loops in Mistral-7B, more loops in LLaMA3-7B).

Clean vs. Clean and Poisoned vs. Poisoned resampling, confirming negligible within-class variability. In contrast, Mixed vs. Mixed splits exhibit systematic deviations that mirror the clean–poisoned separation observed in Figures 11 and 12 of Appendix B.3, indicating that LDR captures genuine geometric differences rather than artifacts of sampling noise or random partitioning.

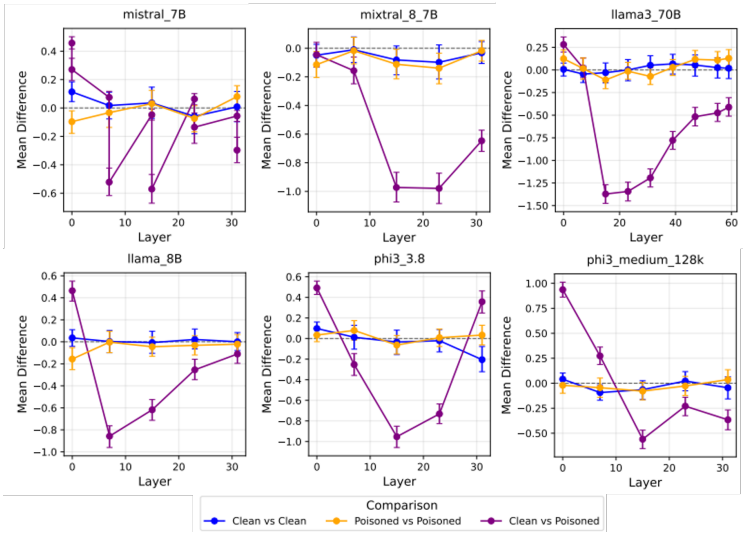


Figure 9: **Ablation of dispersion ratio differences (Clean vs. Clean, Poisoned vs. Poisoned, Mixed vs. Mixed).** Each plot shows the difference in mean dispersion ratio (clean minus poisoned). Positive values indicate that the clean subset exhibits higher dispersion, whereas negative values reflect a more dispersed poisoned subset.

4.2 Local Analysis: Information Flow Between Layers

To investigate the fine-grained mechanisms of adversarial influence, our local analysis quantifies how information transforms between layers at the neuron level. We present the results for Mistral 7B below; see Appendix D.2 for other models.

Analysis on Consecutive Layers. Our local method revealed a structural phase shift in the network's information flow under adversarial influence. We computed Vietoris–Rips PH barcodes of the 2D embeddings described in Section 3.3 for the raw activations; their normalization to zero mean and unit variance, to ensure that topological signals are not due solely to scale differences; and a control condition where neuron indices are randomly permuted, disrupting any neuron-wise correspondence between layers. We measured the topological complexity by the total persistence of 1-bars, and found significant differences between clean and poisoned activations across layers in the raw and normalized activations (Figure 10 (left)). Furthermore, the ratio of topological complexity between clean and poisoned activations (Figure 10 (center)) shows that clean inputs initially exhibit a more complex structure that simplifies in deeper layers. In contrast, poisoned activations start simpler but their topological complexity increases, diverging significantly from the clean activations around layer 12. This suggests that adversarial influence causes a major reconfiguration of information processing in the model's deeper layers. The disappearance of this signal in the permuted control condition (shown in Figure 56 of the Appendix D.2.1) confirms that the effect relies on specific neuron-to-neuron pathways rather than arising from a statistical artifact.

Table 3: **Peak analysis.** Precision@k for k=1, 3, and 5 largest peaks in total variance, and their precision in detecting the largest peaks in absolute difference between the two classes. Spearman's rank correlation (r) is reported in the last column. *, ** correspond to p-values $<$.05 and .01, respectively.

	p@1	*p@3*	*p@5*	r
Total Persistence 0-bars	0	.33	.4	0.46**
Total Persistence 1-bars	0	.67*	.8**	0.78**
Mean Birth 1-bars	1.0*	.33	.8*	0.46**
Mean Death 1-bars	1.0*	.33*	.8**	0.69**

In a real-world setting without labels, these informative layers can still be identified. We found that the overall variance of a topological feature across all samples strongly correlates with the magnitude of the clean-vs.-poisoned difference (Figure 10 (right)). As shown in Table 3, we evaluated the alignment between overall variance and class separation using precision at k ($p@k$) and Spearman's rank correlation (r). To validate statistical significance against a random baseline, we generated empirical null distributions via random permutations, with significance levels indicated by asterisks. The high precision (particularly at $k = 5$) and moderate-to-strong correlations indicate that layers with the highest variance are reliable indicators of those with the largest class separation. This provides a practical, unsupervised signal for locating where adversarial effects are most prominent.

A further example of how different barcode summaries propagate across the layers can be found in Appendix D.2.1 for Mistral 7B, showing the patterns for the mean deaths of 0-bars.

Analysis on Non-Consecutive Layers. We expanded the previous analysis to activations from non-consecutive layers to show that in neighboring layers, the model operates on similar groups of neurons, leading to element-wise interactions that construct meaningful topological features distinguishing clean from poisoned datasets. The ratio of the mean death times of 0-bars between clean and poisoned activations as the layer interval increases is shown in Appendix D.2.5. For layer intervals of 1 and 3, the ratios for normalized activations and the control setting remained distinct, indicating meaningful topological interactions. However, at an interval of 10 layers, the scaled and control settings showed significant overlap, suggesting a much diminished difference in the interactions in clean and poisoned data. A similar pattern can be observed for other barcode summaries, such as the total persistence of 1-bars, see Appendix D.2.5.

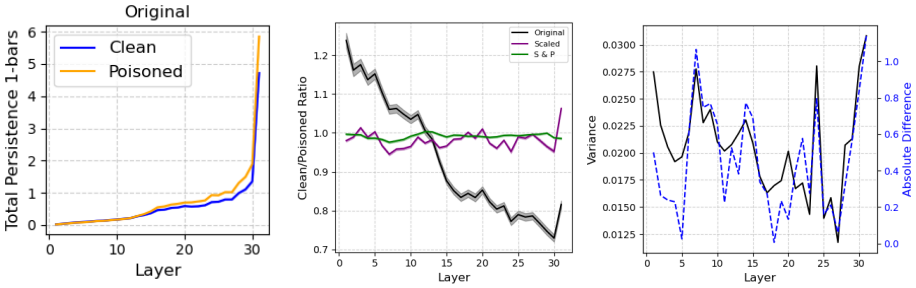


Figure 10: **Local analysis of consecutive layers for the total persistence of 1-bars.** Comparisons of the average total persistence of 1-bars across 1000 samples for Mistral model using original activation data **(left)**. **(center)** Ratios of mean total persistence of 1-bars between clean and poisoned datasets for original, scaled, and scaled and permuted activations. **(right)** Overlaid plots of the overall variance of total persistence of 1-bars for clean and poisoned datasets combined and the absolute difference between mean total persistence of 1-bars for clean and poisoned datasets.

5 Discussion and Future Work

Our global and local analyses provide converging evidence for a fundamental principle, where adversarial influence manifests as "topological compression" of an LLM's latent space. This behavior—a shift from compact, diverse structures to more dispersed, topologically simpler ones—is a consistent, architecture-agnostic phenomenon that holds across different model architectures, sizes, and attack vectors. This topological approach offers a distinct and complementary form of interpretability that is relational rather than compositional. While methods such as sparse autoencoders (SAEs) (Cunningham et al., 2023) are powerful for identifying the "building block" features of a representation, they analyze each activation in isolation. This makes them inherently blind to the nonlinear, relational geometry that emerges from the interactions between activations. Furthermore, because the feature dictionaries learned by SAEs are specific to a single set of model weights, they cannot be reliably compared across different models or fine-tuning stages. Our PH-based framework circumvents these limitations by computing intrinsic, coordinate-free geometric properties, providing a stable basis for comparison and enabling a comprehensive characterization of the shape of adversarial influence.

The implications of our work extend to the core of interpretability and AI safety. Our findings contribute to a growing body of evidence that a model's behaviors are encoded in the geometry of its latent space. This perspective aligns with work showing that memorization corresponds to a reduction in the effective dimensionality of the representation manifold (Stephenson et al., 2021), and that the success of linear probes may stem from their ability to approximate more complex topological structures (Engels et al., 2025). Our discovery that adversarial influence induces a "topological compression" provides new evidence for this hypothesis, suggesting that a collapse in geometric complexity is a quantifiable signature of out-of-distribution states. Our findings reframe key safety properties such as robustness not merely as abstract behavioral outcomes, but as measurable characteristics of the representation space itself.

Limitations. The primary limitation of our study is the memory requirements of PH, as the distance and boundary matrices required for exact Vietoris–Rips computations scale quadratically with the number of points. To manage this on our large datasets, we implemented random subsampling, which is well-studied in TDA with established convergence results ensuring that the sampling errors in our study are bounded (Chazal et al., 2014; Cao & Monod, 2022).

Future Work. Our study opens several avenues for future investigation, such as exploring whether topological compression is a general property of model misalignment (Stephenson et al., 2021); developing topology-aware robustness mechanisms (Brüel-Gabrielsson et al., 2020); applying persistent Morse theory (Bobrowski & Adler, 2014); and adapting cycle matching approaches (Reani & Bobrowski, 2022; García-Redondo et al., 2024) to further characterize LLM representation spaces. Further study is also needed to see if these topological signatures generalize to an even broader range of adversarial scenarios.

REFERENCES

Sahar Abdelnabi, Aideen Fay, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Paverd. Are you still on track!? catching llm task drift with activations, 2024. URL https://arxiv.org/abs/2406.00799.

Sahar Abdelnabi, Aideen Fay, Ahmed Salem, Egor Zverev, Kai-Chieh Liao, Chi-Huang Liu, Chun-Chih Kuo, Jannis Weigend, Danyael Manlangit, Alex Apostolov, et al. Llmail-inject: A dataset from a realistic adaptive prompt injection challenge. *arXiv preprint arXiv:2506.09956*, 2025.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, 18 (8):1–35, 2017.

Robert J. Adler and Jonathan Taylor. *Topological complexity of smooth random functions : Ecole d'Eté de Probabilités de Saint-Flour XXXXIX - 2009*. Lecture notes in mathematics, 2019. Springer, New York, 1st ed. 2011. edition, 2011. ISBN 3-642-19580-6.

Dashti Ali, Aras Asaad, Maria-Jose Jimenez, Vidit Nanda, Eduardo Paluzo-Hidalgo, and Manuel Soriano-Trigueros. A survey of vectorization methods in topological data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14069–14080, 2023. doi: 10.1109/TPAMI.2023.3308391.

Ulrich Bauer. Ripser: Efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, September 2021. ISSN 2367-1734. doi: 10.1007/s41468-021-00071-5.

Omer Bobrowski and Robert J. Adler. Distance functions, critical points, and the topology of random Čech complexes. *Homology, Homotopy and Applications*, 16(2):311–344, 2014. ISSN 15320073, 15320081. doi: 10.4310/HHA.2014.v16.n2.a18. URL http://www.intlpress.com/site/pub/pages/journals/items/hha/content/vols/0016/0002/a018/.

Magnus Bakke Botnan and Michael Lesnick. An introduction to multiparameter persistence. *Representations of Algebras and Related Structures*, pp. 77, 2023.

Rickard Brüel-Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, Primoz Skraba, Leonidas J. Guibas, and Gunnar Carlsson. A topology layer for machine learning, 2020. URL https://arxiv.org/abs/1905.12200.

Peter Bubenik. The persistence landscape and some of its properties. In *Topological Data Analysis: The Abel Symposium 2018*, pp. 97–117. Springer, 2020.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2024. URL https://arxiv.org/abs/2212.03827.

Yueqi Cao and Anthea Monod. Approximating persistent homology for large datasets. *arXiv preprint arXiv:2204.09155*, 2022.

Patrick Chao et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 55005–55029, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e43739fba4d397ce95b542455b1f3c39-Paper-Conference.pdf.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:667963, 2021.

Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. In *International Conference on Machine Learning*, pp. 163–171. PMLR, 2014.

Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Subsampling methods for persistent homology. In *International Conference on Machine Learning*, pp. 2143–2151. PMLR, 2015.

Harish Chintakunta, Thanos Gentimis, Rocio Gonzalez-Diaz, Maria-Jose Jimenez, and Hamid Krim. An entropy-based persistence barcode. *Pattern Recognition*, 48(2):391–401, 2015. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2014.06.023. URL https://www.sciencedirect.com/science/article/pii/S0031320314002453.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007. doi: 10.1007/s00454-006-1276-5. URL https://doi.org/10.1007/s00454-006-1276-5.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Patrizio Frosini. A distance for similarity classes of submanifolds of a euclidean space. *Bulletin of the Australian Mathematical Society*, 42(3):407–415, 1990.

Patrizio Frosini. Measuring shapes by size functions. In *Intelligent Robots and Computer Vision X: Algorithms and Techniques*, volume 1607, pp. 122–133. SPIE, 1992.

Inés García-Redondo, Anthea Monod, and Anna Song. Fast topological signal identification and persistent cohomological cycle matching. *Journal of Applied and Computational Topology*, 8: 695–726, 06 2024. doi: 10.1007/s41468-024-00179-4.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco

Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L.

Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, and David Krueger. Stress-testing capability elicitation with password-locked models, 2024. URL https://arxiv.org/abs/2405.19550.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection, 2023. URL https://arxiv.org/abs/2302.12173.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Nathalie Maria Kirch, Severin Field, and Stephen Casper. What features in prompts jailbreak llms? investigating the mechanisms behind attacks, 2024. URL https://arxiv.org/abs/2411.03343.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL https://arxiv.org/abs/2403.03218.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001. doi: https://doi.org/10.1002/asmb.446. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.446.

Mantas Mazeika et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. In *Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks*, 2024. URL https://openreview.net/forum?id=VlA2D5xX0A.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020. URL http://jmlr.org/papers/v21/20-345.html.

Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):17, 2017.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL https://arxiv.org/abs/2212.09251.

Yohai Reani and Omer Bobrowski. Cycle registration in persistent homology with applications in topological bootstrap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5579–5593, 2022.

Johann Rehberger. Microsoft Copilot: From Prompt Injection to Exfiltration of Personal Information. [Link], 2024.

Matteo Rucco, Filippo Castiglione, Emanuela Merelli, and Marco Pettini. Characterisation of the idiotypic immune network through persistent entropy. In Stefano Battiston, Francesco De Pellegrini, Guido Caldarelli, and Emanuela Merelli (eds.), *Proceedings of ECCS 2014*, pp. 117–128, Cham, 2016. Springer International Publishing.

Cory Stephenson, Suchismita Padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and SueYeon Chung. On the geometry of generalization and memorization in deep neural networks, 2021. URL https://arxiv.org/abs/2105.14602.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline, 2019. URL https://arxiv.org/abs/1905.05950.

Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014.

Adaku Uchendu and Thai Le. Unveiling topological structures in text: A comprehensive survey of topological data analysis applications in nlp, 2024. URL https://arxiv.org/abs/2411.10298.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. Ai sandbagging: Language models can strategically underperform on evaluations, 2024. URL https://arxiv.org/abs/2406.07358.

Matthew Wheeler, Jose Bouza, and Peter Bubenik. Activation Landscapes as a Topological Summary of Neural Network Performance. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 3865–3870, December 2021. doi: 10.1109/BigData52589.2021.9671368.

Menglin Yang, Aosong Feng, Bo Xiong, Jiahong Liu, Irwin King, and Rex Ying. Enhancing LLM complex reasoning capability through hyperbolic geometry. In *ICML 2024 Workshop on LLMs and Cognition*, 2024. URL https://openreview.net/forum?id=5lFiIVza6x.

Simon Zhang, Mengbai Xiao, and Hao Wang. Gpu-accelerated computation of vietoris-rips persistence barcodes. In *36th International Symposium on Computational Geometry (SoCG 2020)*, pp. 70–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023. URL https://arxiv.org/abs/2310.01405.

Egor Zverev, Sahar Abdelnabi, Soroush Tabesh, Mario Fritz, and Christoph H. Lampert. Can llms separate instructions from data? and what do we even mean by that?, 2025. URL https://arxiv.org/abs/2403.06833.

A Persistent Homology

We provide additional background on PH and the underlying mathematical formulation that supports its application as a tool to detect the *multiscale topological features* within data.

A.1 Theoretical Background

PH refers to a set of methods that are implemented to extract the shape and size of data a multiple scales. We now present the underlying mathematical principles that support this tool.

Input data. PH accommodates for diverse data modalities: images, point clouds, graphs, etc. One of the most basic yet general data types that it accepts is *finite metric spaces*, i.e., finite subsets $S \subset X$ of some metric space (X, d). Restricting d to S, we obtain a notion of dissimilarity between the points in our metric space. This is the data modality that we will consider for the remainder of the section, as it encompasses most of the real data that we encounter.

Filtrations. The first step in the PH pipeline consists of constructing a filtration from our input data, that is, a family of nested topological spaces. For computational and storage reasons, *simplicial complexes* are often favored as the topological spaces appearing in the filtration. An abstract simplicial complex K over a vertex set S is defined as a set of subsets of S which is closed under inclusion, i.e., if $\sigma \in K$ and $\tau \subset \sigma$, then $\tau \in \sigma$. Subsets $\sigma = \{s_{i_0}, \dots, s_{i_p}\}$ of $p+1$ elements are called p-simplices. There are various ways of defining a simplicial complexes from a discrete set S, and they usually depend on fixing a scale parameter $\epsilon > 0$.

For instance, in this work, we have leveraged the *Vietoris–Rips complex*, obtained by considering all the subsets σ of S with $\operatorname{diam}(\sigma) \coloneqq \max_{s,s' \in \sigma} d(s, s')$ less or equal than ϵ,

$$\mathrm{VR}_\epsilon(S, d) \coloneqq \{\emptyset \neq \sigma \subset K : \operatorname{diam}(\sigma) \leq \epsilon\}. \tag{1}$$

The implementation of this complex is straightforward, and has the advantage that it is only necessary to store the pairwise distance between points in S to build it. However, it has the disadvantage of exploding in size with the number of points: if S has n points, then $|\mathrm{VR}_\epsilon(S, d)| = O(2^n)$ (see Table 1 in Otter et al. (2017))

An alternative is the *Čech complex* at scale $\epsilon \geq 0$, where a simplex $\sigma = \{s_{i_0}, \dots, s_{i_p}\}$ belongs to the complex if and only if all the balls of radius ϵ centered at the points of the simplex have nonempty intersection,

$$\check{\mathrm{C}}_\epsilon(S, d) \coloneqq \left\{ \emptyset \neq \sigma \subset S : \bigcap_{s \in \sigma} B(s, \epsilon) \neq \emptyset \right\}. \tag{2}$$

The Čech complex has very nice theoretical properties (for instance, it satisfies the conditions of the Nerve Theorem). However, it has similar complexity to the Vietoris–Rips complex, and in fact we have

$$\check{\mathrm{C}}_\epsilon(S, d) \subseteq \mathrm{VR}_\epsilon(S, d) \subseteq \check{\mathrm{C}}_{\sqrt{2}\epsilon}(S, d).$$

A final option to consider, which significantly reduces the number of simplices in the complex, is the *alpha complex*. To make this simplicial complex coarser, the idea is to intersect the balls centered around the points in the point cloud, $B(s, \epsilon)$, with their Voronoi cells, $V(s)$, and thus define $R(s, \epsilon) \coloneqq B(s, \epsilon) \cap V(s)$. The Voronoi cells form a partition of the metric space X where the points in each region are closest to the same point in S. Since both $B(s, \epsilon)$ and $V(S)$ are convex, their intersection $R(s, \epsilon)$ remains convex. From the definition of the Voronoi cells, these spaces $R(s, \epsilon)$ are either disjoint or overlap along their boundary, significantly reducing the number of intersections between them. The alpha complex is thus defined as

$$\alpha(S, \epsilon) \coloneqq \left\{ \emptyset \neq \sigma \subset S : \bigcap_{s \in \sigma} R(s, \epsilon) \neq \emptyset \right\} \tag{3}$$

and is significantly smaller in size due to the introduction of the Voronoi cells.

The Vietoris–Rips, Čech, and alpha filtrations are defined considering the families of the corresponding complexes for all values of the parameter $\epsilon \geq 0$. Since the conditions for including simplices are relaxed as ϵ increases, we obtain the defining condition of a filtration $\{K_\epsilon : \epsilon \geq 0\}$, namely that for $\epsilon \leq \epsilon'$ we have $K_\epsilon \subset K_{\epsilon'}$. There are additional types of filtrations that we do not cover here, such as cubical filtrations (particularly suited for images) or witness complexes (based on having some landmarks or witnesses in our point cloud). We refer to Otter et al. (2017) for a survey and further details on these constructions.

Homology and persistence modules. Leveraging tools from algebraic topology, we can compute the *simplicial homology groups* associated to a given simplicial complex K, which come in various degrees $H_p(K)$, for $p \geq 0$ an integer number, and are topological invariants of the complex. They contain information about its topological features, for $p = 0$ these correspond to components or

clusters, for $p = 1$ to loops or holes, for $p = 2$, to bubbles or cavities, and so on for higher values of p. The homology construction is functorial, meaning that there is an assignment which for a map $f : K \to K'$ between two simplicial complexes, provides a linear map at the homology level $H_p(f) : H_p(K) \to H_p(K')$, preserving the identity and composition. Applying this to any of the filtrations of the step above we obtain a *persistence module*, that is, a family of vector spaces $\{H_p(K_\epsilon) : \epsilon \geq 0\}$ endowed with linear maps $H_p(\epsilon \leq \epsilon') : H_p(K_\epsilon) \to H_p(K_{\epsilon'})$ for $\epsilon \leq \epsilon'$, which are the maps induced by the inclusions of the filtration. In other words, $H_p(K_\bullet)$ can be seen as a functor from the poset category $(\mathbb{R}_{\geq 0}, \leq)$ to the category of vector spaces and linear maps. Given the mathematical construction of homology, $H_p(K_\bullet)$ contains information about the topological features in the simplicial complexes of the filtration, and in particular, about when features appear and disappear as the parameter ϵ increases. We now seek to provide a compact description for this.

Persistence barcodes. The mathematical structure of a persistence module has various desirable properties. Among them, one of the most important ones is satisfying the conditions for the the so called *structure theorem* (Botnan & Lesnick, 2023, Theorem 4.2) to apply, which tells us that a given a persistence module $H_p(K_\bullet)$ decomposes in an essentially unique way as a direct sum of interval modules $\mathbb{R}[b, d)$. Interval modules are persistence modules supported over intervals of the real line which, inside their support, map to the vector space $\mathbb{R}$, and outside, to 0. Since the decomposition is an invariant of the isomorphism type of $H_p(K_\bullet)$, the collection of intervals appearing in it is also a topological invariant. We refer to this collection of bars as the *persistence barcode* of the input data. The interpretation of these barcodes becomes apparent: each of the bars in the barcode correspond to a topological feature that appears at the initial point in the interval (its *birth time*) and persists until its end (its *death time*). There are many other invariants that we can derive from the original persistence module $H_p(K_\bullet)$, such as the rank function (Frosini, 1990; 1992), the persistence image (Adams et al., 2017) or the persistence landscape (Bubenik, 2020); some of these invariants act on barcodes as vectorizations or embeddings. In this work, we focus on barcodes and we represent statistics calculated from bars and barcodes in the form of a vector, which is different in spirit from an embedding or vectorization of a barcode.

A.2 Persistent Homology Barcode Statistics

To interpret the barcodes from Section 3.2 and Section A.1, we extract key summary statistics that quantify the topological structure observed at each layer under both adversarial conditions.

From each 1-dimensional (1D) barcode, we gather intervals (b_i, d_i) with $d_i > b_i > 0$ and define $\ell_i = d_i - b_i$. Forming a discrete distribution $p_i = \ell_i / \sum_j \ell_j$, the *persistence entropy* is

$$E \;=\; -\sum_i \; p_i \, \ln(p_i + \epsilon),$$

where ϵ is a small positive constant (e.g., 10^{-12}) to ensure numerical stability. Higher E indicates a more uniform distribution of lifetimes (no single interval dominates), whereas lower E reflects a small number of long-lived intervals.

In addition to **entropy**, we compute the following summary statistics on dimension-1 bars:

- **Mean births (1-bars):** Average birth time $\bar{b}$
- **Mean deaths (1-bars):** Average death time $\bar{d}$
- **Mean persistence (1-bars):** Average lifetime $\overline{(d_i - b_i)}$
- **Number of 1-bars:** Count of finite intervals in dimension 1

We perform these computations for each barcode individually and then average over all barcodes in the same condition (elicited or elicited) and (clean or poisoned).

B Further Topological and Local Variance Interpretation

B.1 Extended Prompt Injection (Clean vs. Poisoned)

For mean births and mean deaths, all layers except layer 1 across models have negative differences, indicating that poisoned intervals emerge and die later in the filtration. The mean persistence is

Table 4: **Dimension-1 persistent homology differences (clean − poisoned) in key metrics for three models across several layers.** Positive values mean the clean condition has a higher value, while negative indicates poisoned is higher for that metric. All entries rounded to four decimals.

Model	Layer	Mean births 1-bars_diff	Mean deaths 1-bars_diff	Mean persistence 1-bars_diff	Entropy 1-bars_diff	Number 1-bars_diff
LLaMA-3 (8B)	1	-0.0005	-0.0006	-0.0001	0.1665	86.9700
	8	-0.0609	-0.0608	0.0001	0.1213	79.5600
	16	-0.3166	-0.3249	-0.0082	0.0188	17.9367
	24	-0.9932	-1.0256	-0.0324	0.1595	80.0833
	32	-18.3367	-18.9290	-0.5923	0.3348	192.4900
Mistral (7B)	1	0.0004	0.0004	0.0000	0.0172	3.7967
	8	-0.0293	-0.0295	-0.0002	0.1485	118.9167
	16	-0.2375	-0.2421	-0.0047	0.1938	154.7633
	24	-0.5694	-0.5815	-0.0120	0.2070	153.9633
	32	-14.7376	-15.0558	-0.3182	0.2239	166.4267
Phi 3 (3.8B)	1	0.0011	0.0009	-0.0002	0.0101	4.3200
	8	-0.4522	-0.4675	-0.0153	0.0888	59.0967
	16	-1.7825	-1.8293	-0.0467	0.0696	48.7367
	24	-6.2017	-6.3448	-0.1431	0.1665	116.9433
	32	-1.0092	-1.0503	-0.0411	-0.4857	-136.1200

also higher for poisoned examples, especially in the middle layers. The entropy differences are also almost all positive (except Phi3 3.8's final layer), suggesting that clean distributes its 1D lifetimes more evenly, while poisoned tends to concentrate total persistence in fewer cycles. Likewise, the number of 1-bars remains mostly positive, so clean produces more 1D holes overall.

Overall, poisoned conditions often yield fewer but more dominant dimension-1 features that appear later and persist longer, whereas clean spreads its total persistence across numerous shorter-lived cycles, resulting in higher entropy. Depending on the model, the specific layer where these features distinctions appear strongest differ but the nature of the difference is the same across models and layers.

B.2 Sandbagged Model (Locked vs. Elicited)

Table 5: **Dimension-1 persistent homology differences (elicited − locked) for two models across multiple layers.** Positive values indicate that the elicited condition has higher values; negative means locked is higher for that metric.

Model	Layer	Mean births 1-bars_diff	Mean deaths 1-bars_diff	Mean persistence 1-bars_diff	Entropy 1-bars_diff	Number 1-bars_diff
LLaMA-3 (8B)	0	-0.0127	-0.0132	-0.0005	0.0156	3.2400
	7	-0.3425	-0.3555	-0.0130	0.0647	27.8600
	15	-0.0476	-0.0455	0.0021	0.2114	135.2900
	23	-0.1168	-0.1204	-0.0037	0.0100	61.8766
	31	-0.9750	-1.0458	-0.0707	0.0620	28.2800
Mistral (7B)	0	-0.0053	-0.0055	-0.0002	0.0942	27.1533
	7	-0.1925	-0.1989	-0.0064	0.0310	14.1066
	15	0.0393	0.0352	-0.0041	0.0277	10.9300
	23	0.6722	0.7037	0.0315	-0.0363	-0.1900
	31	14.6450	15.2952	0.6503	-0.0014	9.3233

For LLaMA3 8B , the mean birth and death differences are negative across all computed hidden layers (1, 8, 16, 24, 32). Note that layers are zero-indexed, meaning that layer 0 corresponds to the first hidden layer, layer 1. This indicates that, in the locked condition, 1D cycles exhibit larger (i.e., later) birth and death times compared to elicited. In other words, when locked, the 1D features tend to emerge "further out" in the filtration. The mean persistence difference between conditions is also negative (except layer 16), suggesting that locked cycles generally persist slightly longer

on average. Entropy differences are positive, indicating that elicited exhibits a greater diversity or spread among the lifetimes of its 1D features. The number of 1-bars is positive (sometimes strongly so), meaning there are substantially more 1D features in the elicited condition.

We see similar results for Mistral 7B with negative differences in births and deaths in earlier layers, implying that locked has larger birth/death times at those lower layers. However, the sign flips, with elicited displaying larger values for births, deaths, and persistence. Specifically, layer 32 shows a notably large positive difference (e.g., $+14.64$ for births, $+15.29$ for deaths), indicating that the final layer in elicited captures significantly later 1D cycles relative to locked. The number of 1-bars also tends to be higher in elicited at most layers, except for a minor negative at layer 23, again suggesting that elicited reveals a greater number of dimension-1 features.

B.3 Local Dispersion Ratio Across Poisoned Conditions

We analyze how local geometry in hidden-layer representation space differs between clean and multiple poisoned modes in six LLMs. We further classify poisoned prompts into three sub-types:

1. **Executed:** The injected request is recognized and carried out (indirect prompt injection).
2. **Refused:** The model identifies the injected content as malicious and issues a refusal, effectively "shutting down" any detailed elaboration.
3. **Ignored:** The model neither executes nor refuses, but effectively overlooks the injected prompt, proceeding as if it were absent.

For each final token's activation difference vector $\Delta\mathrm{Act}_\ell(x_i) \in \mathbb{R}^D$, we identify its k nearest neighbors in layer ℓ and perform PCA on those points. Let $\lambda_1 \geq \cdots \geq \lambda_{D'}$ be the resulting eigenvalues. We define the *dispersion ratio* of $\Delta\mathrm{Act}_\ell(x_i)$ as

$$\frac{\sum_{j=2}^{D'} \lambda_j}{\lambda_1 + \epsilon},$$

where ϵ prevents division by zero. A higher ratio indicates that variance is more evenly spread among secondary directions, whereas a lower ratio implies most variance lies in a single dominant direction.

Ablation: Clean vs. Clean, Poisoned vs. Poisoned, and Mixed. To confirm that dispersion discrepancies primarily reflect true clean vs. poisoned distinctions rather than random partitioning or mixture effects, we performed three auxiliary comparisons:

1. **Clean vs. Clean**: Split the clean set into two subsets, ensuring no significant difference arises from sampling within the same class.
2. **Poisoned vs. Poisoned**: Applied the same procedure to poisoned data to assess within-class variability.
3. **Mixed vs. Mixed**: Randomly partitioned a combined pool of clean and poisoned samples into two balanced groups.

Note on Statistical Methods: For every layer in each subplot, we computed the dispersion ratio for both clean and the specified poisoned (or refused, executed, ignored) samples. We then conducted a Welch's t-test on these two groups (clean vs. poisoned/other), applying false-discovery rate (FDR) correction across layers. We also verified approximate normality via kernel density estimates (KDEs) for each groups. Plot markers with stars indicate layers where $p_{\text{FDR}} < 0.05$, confirming a statistically significant difference in dispersion ratio. To select $k = 30$, we tested candidate neighborhood sizes across layers and models, measuring which k produced the largest absolute difference in mean local dispersion ratio between clean and poisoned conditions.

B.3.1 Discussion of Results

Figures 11 and 12 highlight that:

- **Early Layers (Layer 1–8):** Across all poisoning modes, the clean condition consistently shows a higher dispersion ratio, suggesting that the model initially allocates broader representational capacity for normal inputs.
- **Mid Layers (Layer 16):** This pattern often flips, with poisoned prompts (especially executed or ignored) exceeding the clean baseline, indicating the network is dedicating extra directions to elaborate or "embrace" these injected requests. Conversely, refused prompts typically exhibit reduced dispersion, mapping disallowed content into a lower-variance region.

Interestingly, our findings align with the results of Stephenson et al. (2021), which indicate that memorization tends to emerge in deeper layers where the effective dimensionality shrinks. Consistent with that view, we observe that executed or ignored prompts show a higher dispersion in mid-layers, implying the model invests additional capacity there for those injected instructions. Meanwhile, a refused request is routed into a more compressed region, effectively "shutting down" further representational expansion. In this sense, deeper layers may provide a setting where the network can more sharply discriminate or overfit certain inputs—supporting the idea that final layers reflect a gradually compressed, yet strategically focused representation space.

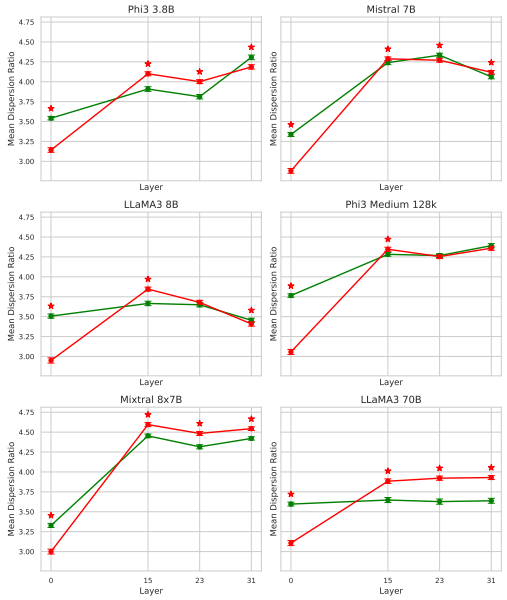


Figure 11: **Layer-wise Dispersion Ratio for Clean vs. Poisoned Examples.** The green and red lines depict mean dispersion ratios for clean and poisoned inputs, respectively, at different layer depths. Error bars around each point represent ± 1 standard error of the mean (SEM). In early layers (left side), clean data consistently has higher dispersion on average, whereas in mid-layers (center), poisoned surpasses the clean baseline, indicating a re-distribution of representational capacity for the injected prompts. Layers where the difference is statistically significant ($p_{\text{FDR}} < 0.05$) are marked with a red asterisk above the higher mean value.

B.4 Cosine Distance of Representations

We analyze the difference representations $\Delta\text{Act}_\ell(x_i) \in \mathbb{R}^D$ for corresponding pairs of clean and poisoned inputs in Figure 14. Specifically, for each model and layer, we load up to five pairs of clean and poisoned activation files, compute the difference between the activations for each pair, and concatenate these differences. From these differences, we draw equal-size subsamples of 5000 vectors. For each layer and comparison condition, we compute the mean pairwise cosine distance within each subsample. Because cosine distance is scale-invariant, we do not normalize these difference representations. We perform four comparison conditions: clean vs. poisoned, clean vs. clean (where clean samples are split in half), poisoned vs. poisoned (where poisoned samples are split in half),

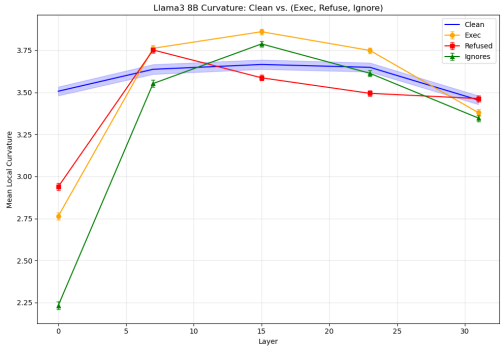


Figure 12: ***LLaMA3_7B* Dispersion Ratio: Clean vs. Executed, Refused, and Ignored Prompts.** The horizontal axis indicates layer depth, while the vertical axis represents the mean dispersion ratio. The blue curve (with confidence band) corresponds to clean inputs; orange, red, and green curves denote executed, refused, and ignored poisoned prompts, respectively. Notably, refused prompts show an early jump but then collapse below the clean baseline, whereas executed and ignored surpass it around mid-layers, highlighting distinct representational regimes.

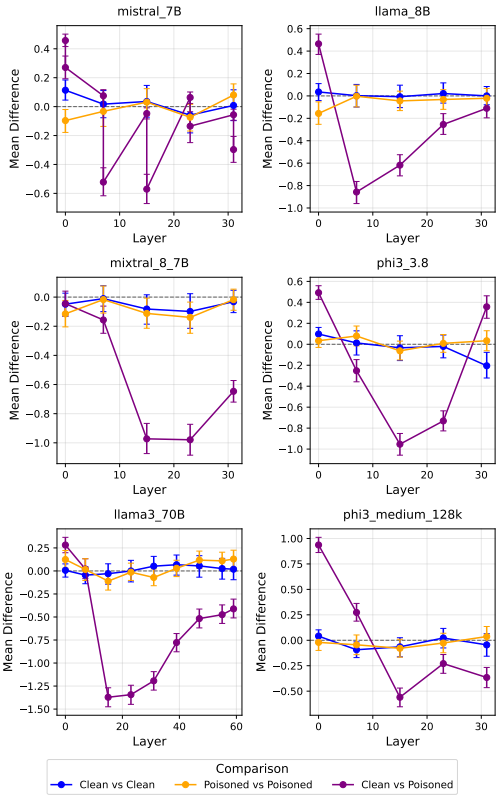


Figure 13: **Ablation of Dispersion Ratio Differences (Clean vs. Clean, Poisoned vs. Poisoned, Mixed vs. Mixed).** Each plot shows the difference in mean dispersion ratio (clean minus poisoned). Positive values indicate that the clean subset exhibits higher dispersion, whereas negative values reflect a more dispersed poisoned subset.

and mixed vs. mixed (where two separate mixed subsamples are created, each containing half clean and half poisoned differences). For each comparison, we generate two distributions of mean pairwise intra-class distances (or inter-class in the clean vs poisoned case) using 3 bootstrap iterations. We then apply Welch's t-test to these distributions to assess whether they diverge significantly.

Empirically, poisoned difference representations typically exhibit a higher mean cosine distance in deeper layers, indicating a more "spread-out" or heterogeneous arrangement of their difference vectors, much as we observed in the curvature analysis. clean data, by contrast, remains comparatively tightly clustered, implying less dispersion in its difference space. Interestingly, *LLaMA3_70B* displays similar characteristics in the early and final layers but poisoned representations have a noticeable smaller cosine distance in middle layers. This may reflect the ability of larger architectures to better partition representation space across the network before re-expanding in later layers.

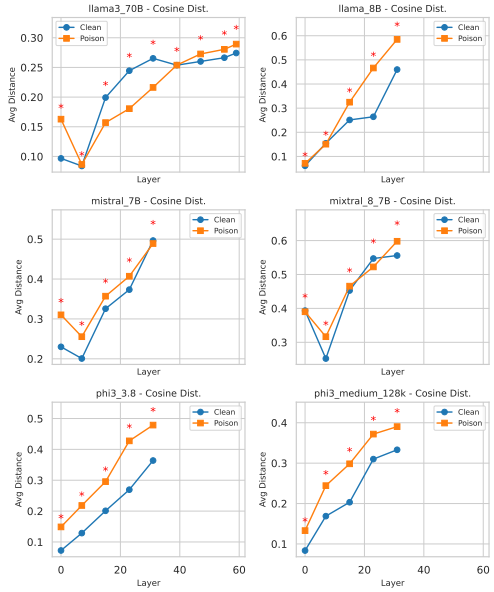


Figure 14: **Cosine Distance of Difference Representations Across Layers.** Each panel shows mean within-class distances (clean vs. poisoned) for the difference representations (*poisoned/clean pass* minus *baseline*), where higher values reflect greater variation among samples. Stars denote layers with significant differences.

C Further Details of Global Layer-Wise Analysis

We now provide further details on the global layer-wise analysis.

C.1 Pipeline

We describe in more detail the pipeline in Figure 3 in the main text. Recall that our aim here was showcasing that topological signatures effectively capture distinctions between representations under normal or adversarial conditions, and to provide an interpretation of the reason behind such difference in terms of the "shape" of the latent representations.

We use Ripser Bauer (2021) to compute barcodes, which is based on Vietoris–Rips filtrations (see Figure 2.1). The computational constraints of PH make it impossible to compute the barcode of any of our two datasets (clean vs. poisoned or locked vs. elicited). Therefore, we leverage subsampling approaches (e.g., Chazal et al. (2015)) and compute barcodes from $K = 64$ subsamples $\{x_{i_1,\ell}, \ldots, x_{i_k,\ell}\} \subset \mathbb{R}^D$ with size $k = 4096$, of the representations per layer $1 \leq \ell \leq L$. From these, 64 are taken from normal activations and 64 from adversarial activations. We use these as proxies for the topology of the whole space.

Following Ali et al. (2023), we represent these barcodes as 41-dimensional feature vectors, which we call *barcode summaries*. These include 35 statistics derived from a 7×5 grid of {mean, minimum, first quartile, median, third quartile, maximum, standard deviation} $\times$ {death of 0-bars, birth of 1-bars, death of 1-bars, persistence of 1-bars, ratio birth/death of 1-bars}; as well as the total persistence (i.e., sum of the lengths of all bars in the barcode), number of bars, and persistent entropy

(Chintakunta et al., 2015; Rucco et al., 2016) defined in Appendix A.2 for 0- and 1-bars. We reduce the dimensionality case-by-case, by eliminating highly correlated features (above a threshold of 0.5) through cross-correlation analysis.

For exploratory analysis, we apply PCA and compute CCA loadings to measure feature correlations with the principal components. A logistic regression model is then used for classification, and Shapley values (Lipovetsky & Conklin, 2001) are computed to evaluate feature importance. Shapley values, derived from cooperative game theory, quantify the contribution of each feature to model predictions by measuring its influence in shifting predictions from a baseline (e.g., 0.5 for logistic regression), providing an interpretable, feature-level analysis of predictive impact.

C.2 Ablation Studies on Subsampling Parameters

We evaluate the representation of clean and poisoned activations using a subsampling-based topological analysis. For each experiment, we consider a fixed layer of Mistral 7B and draw k subsamples of size n from the clean activations and k subsamples of size n from the poisoned activations. Each subsample is used to compute a Vietoris–Rips persistence diagram, which is subsequently represented as a 41-dimensional barcode summary vector. This procedure produces a combined point cloud in $\mathbb{R}^{41}$ of size $2k$, consisting of k clean and k poisoned feature vectors.

Predictive Power of Barcode Summaries for Varying (n, k). We perform the same classification task as in the main text, namely, we fit a logistic regression model to classify between clean and poisoned in each point cloud with fixed (n, k), for the first, the middle, and the last layer of Mistral 7B. We report the 5-fold cross validation results in Figure 15. We observe that there are no clear dependencies of this parameter over the parameters (n, k). Layer 1 seems to be more difficult to classify, requiring at least 500 subsamples, whereas for later layers we obtain perfect classification with as little as $k = 30$ subsamples of size $n = 100$.

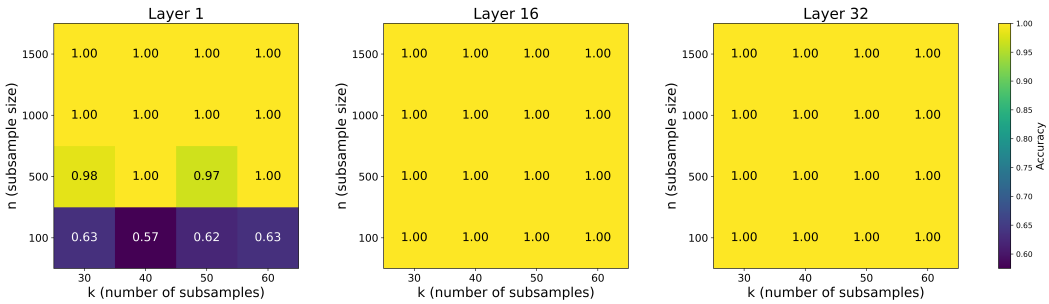


Figure 15: Accuracies of 5-fold cross validation on a logistic regression trained to distinguish barcode summaries of k subsamples of size n of clean activations and k subsamples of size n of poisoned activations at layers 1, 16 and 32 of Mistral 7B.

Metric Description of Clusters for Varying (n, k). We now focus on activation values for layer 16 in Mistral 7B, over which the barcode summaries are computed in subsamples with parameters (n, k). All feature vectors are standardized using a global `StandardScaler` fitted on the whole point cloud. We then compute several metrics to quantify the structure of the resulting representation: (i) the mean intra-class distance within the clean and poisoned subsamples, (ii) the mean inter-class distance between the two groups, and (iii) the inter-to-intra distance ratio

$$r := \frac{d_{\text{inter}}}{\frac{1}{2}\left(d_{\text{intra}}^{\text{clean}} + d_{\text{intra}}^{\text{poison}}\right)}. \tag{4}$$

We perform ablations over the subsample size n and the number of subsamples k. The intra-class distances (Figure 16 left and center) show minimal dependence on k, but decrease consistently as n increases. This suggests that the barcode representations become more concentrated when subsamples contain more points. The values for $n = 500$, 1000, and 1500 are in close proximity, indicating an early convergence of this statistic with respect to n.

The inter-class distance (Figure 16 right) exhibits a complementary trend: it is largely invariant under changes in k, but increases with n. As before, the curves for $n = 1000$ and $n = 1500$ almost coincide, further supporting a convergence regime at moderate subsample sizes.

To combine these effects, we evaluate the inter-to-intra distance ratio in Figure fig. 17. This ratio remains stable across values of k, but increases with n, indicating that the relative separation between clean and poisoned representations improves as subsample size grows. The near overlap of the values for $n = 1000$ and $n = 1500$ again suggests convergence in this regime, which supports the choice of subsample sizes used in the main experiments.

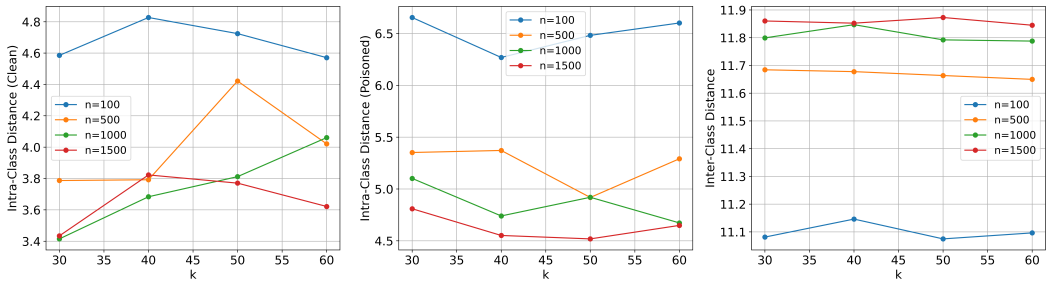


Figure 16: **Left:** Intra-class distance among the barcode summaries of k subsamples of size n of clean activations from layer 16 of Mistral 7B. **Center**: Intra-class distance among the barcode summaries k subsamples of size n of poisoned activations from layer 16 of Mistral 7B. **Right:** Intra-class distance among the clusters of clean and poisoned barcode summaries of k subsamples of size n.

C.3 Results: Clean vs. Poisoned

C.3.1 Mistral 7B

We present here additional results on the global analysis for Mistral 7B that are referred to in the main text.

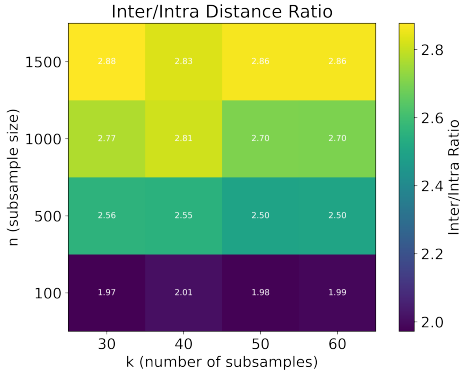


Figure 17: Inter-to-intra distance ratio (Equation 4 between k subsamples of n clean activations and k subsamples of n poisoned activations in layer 16 of Mistral 7B.

Table 6: **Pruned barcode summaries for layers 1, 8, 16, 24 and 32.** Features from the barcode summaries with correlation less than 0.5 in the cross-correlation matrix.

	Layer 1	Layer 8	Layer 16	Layer 24	Layer 32
Mean death 0-bars	✓	✓	✓	✓	✓
Minimum death 0-bars		✓	✓		
Maximum death 0-bars	✓				
Standard deviation death 0-bars	✓				
Minimum birth 1-bars					
Maximum birth 1-bars	✓				
Minimum persistence 1-bars	✓	✓	✓	✓	✓
First quartile persistence 1-bars	✓				
Maximum persistence 1-bars		✓			
Mean birth/death 1-bars		✓	✓		✓
First quartile birth/death 1-bars		✓			
Maximum birth/death 1-bars			✓		
Total persistence 1-bars					✓
Number 0-bars	✓	✓	✓	✓	✓
Number 1-bars		✓	✓	✓	
Entropy 0-bars		✓	✓		
Total features	8	9	8	4	5

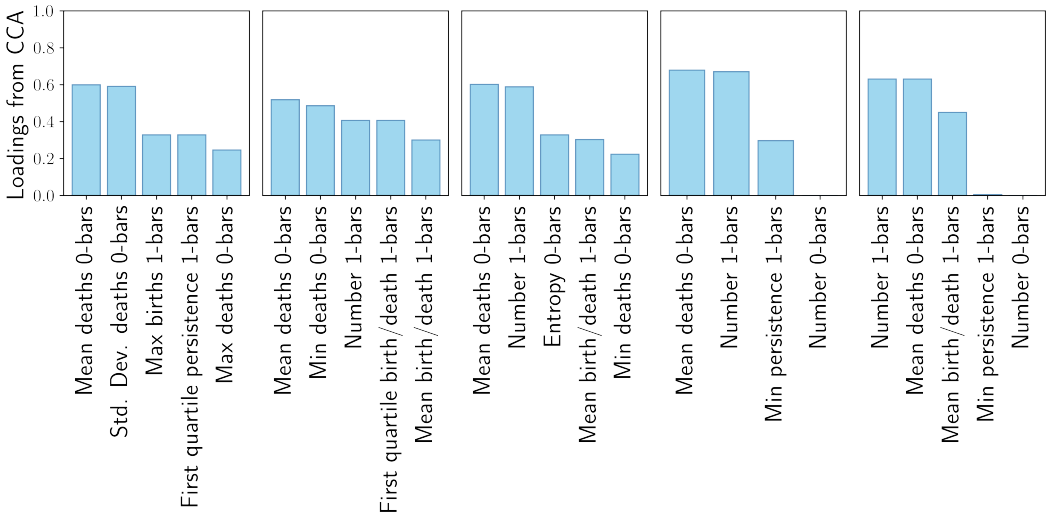


Figure 18: **CCA loadings for clean vs. poisoned activations**. Loadings of the 5 most important contributions to the first canonical variable of the CCA on the pruned barcode summaries show that the mean of the death of 0-bars is significantly correlated with the first two principal components of the PCA across all layers.

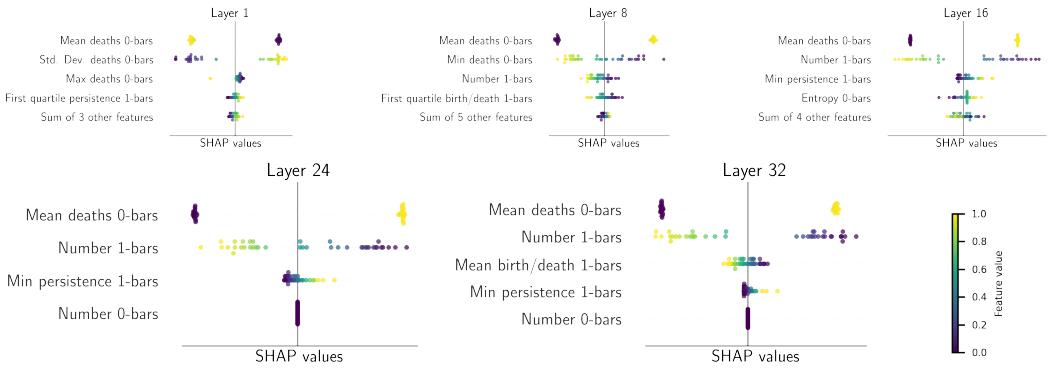


Figure 19: **SHAP analysis: clean vs. poisoned activations.** Beeswarm plot of logistic regression SHAP values trained on the pruned barcode summaries for layer 1, 8, 16, 24, and 32.

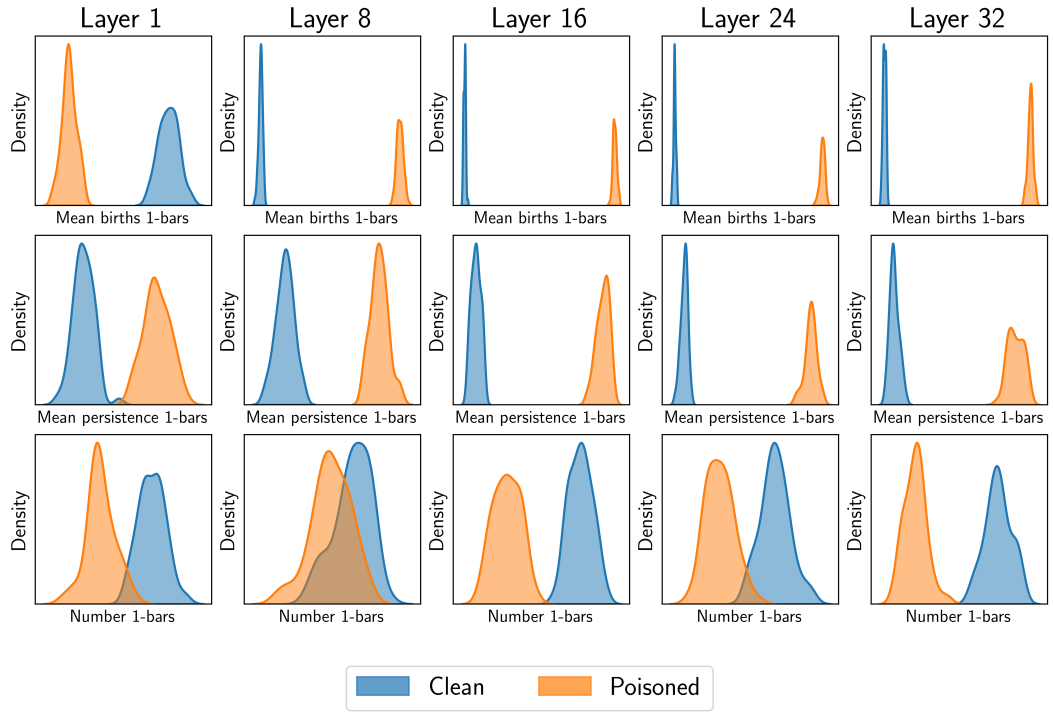


Figure 20: **Histograms for the mean of the births of 1-bars, mean persistence of 1-bars and number of 1-bars for Mistral.** Features extracted from the barcode summaries of the activations for layers 1, 8, 16, 24 and 32 of the clean vs. poisoned dataset.

C.3.2 PHI3-MINI-4K (3.8B PARAMETERS)

We provide the results of the analysis depicted in Figure 3 including layers 1, 8, 16, 23, and 32 for Phi 3 (3.8B parameters) where barcodes are computed using the Euclidean distance in the representation space.

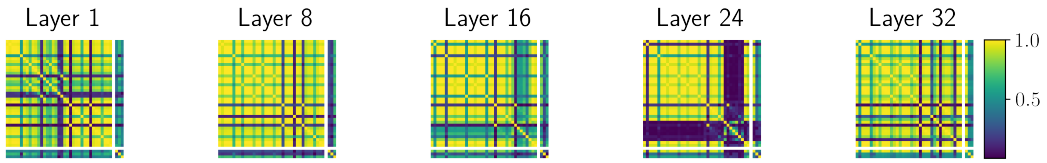


Figure 21: **Cross-correlation matrices for the barcode summaries for clean vs. poisoned activations.** Growing block of correlated features appears in the cross-correlation matrix of the barcode summaries appears in the middle layers (layers 1, 8, 16, 24, and 32 are shown).

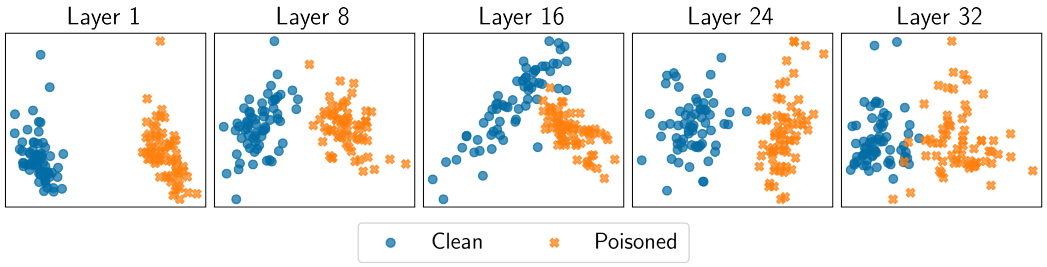


Figure 22: **PCA of barcode summaries of clean vs. poisoned activations**. Clear distinction appears in the projection onto the two first principal components from the PCA of the pruned barcode summaries for layers 1, 8, 16, 24, and 32.

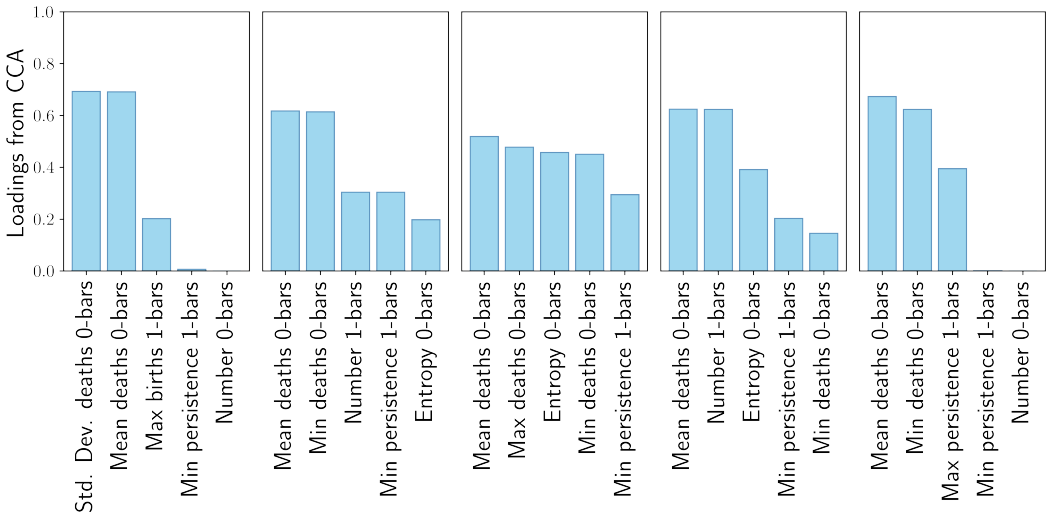


Figure 23: **CCA loadings for clean vs. poisoned activations**. Loadings of the 5 most important contributions to the first canonical variable of the CCA on the pruned barcode summaries show that the mean of the death of 0-bars is significantly correlated with the first two principal components of the PCA across all layers.

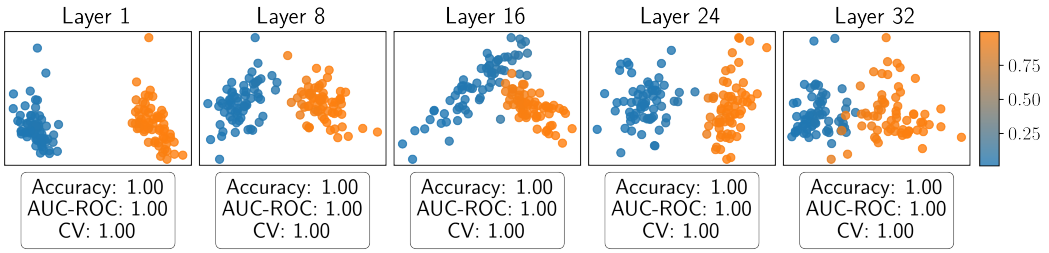


Figure 24: **Logistic regression for clean vs. poisoned activations.** Prediction of a logistic regression trained on a 70/30 train/test split of the pruned barcode summaries, plotted on the projection onto the two first principal components for visualization purposes. Accuracy and AUC–ROC tested on the test data, and 5-fold cross validation on train data are presented for each model, showcasing the outstanding performance of all models.

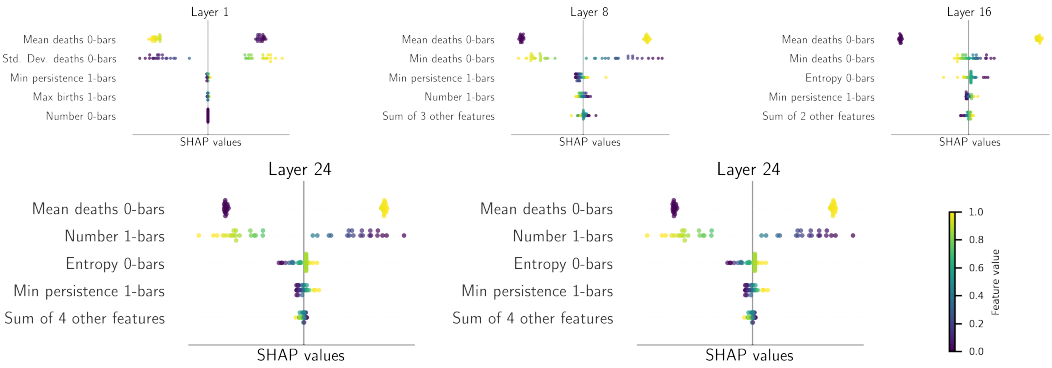


Figure 25: **SHAP analysis: clean vs. poisoned activations.** Beeswarm plot of logistic regression SHAP values trained on the pruned barcode summaries for layer 1, 8, 16, 24, and 32.

C.3.3 MIXTRAL-8X7B (7B PARAMETERS)

We provide the results of the analysis depicted in Figure 3 including layers 1, 8, 16, 23 and 32 for the Mixtral 8 (7B parameters) model where barcodes are computed using the Euclidean distance in the representation space. We observe very similar results to the ones obtained with Mistral, indicating a consistency across models of the topological deformations of adversarial influence via XPIA (see Section 3.1).

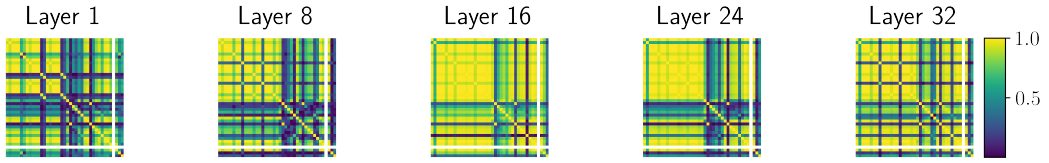


Figure 26: **Cross-correlation matrices for the barcode summaries for clean vs. poisoned activations.** Growing block of correlated features appears in the cross-correlation matrix of the barcode summaries for layers 1, 8, 16, 24, and 32. Correlations in layer 1 are lower than with Mistral 7B, see Figure 6.

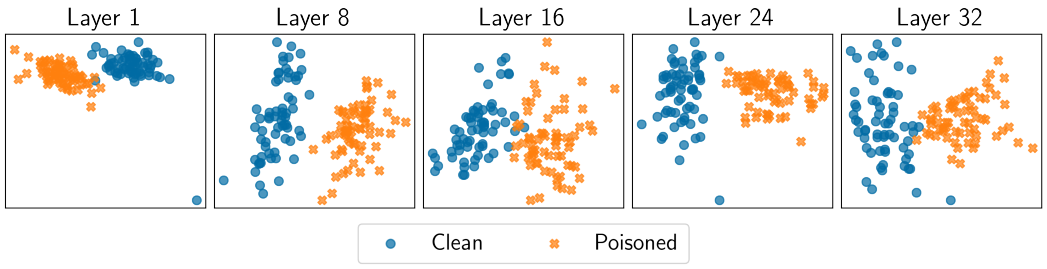


Figure 27: **PCA of barcode summaries of clean vs. poisoned activations**. Clear distinction appears in the projection onto the two first principal components from the PCA of the pruned barcode summaries for layers 1, 8, 16, 24, and 32.

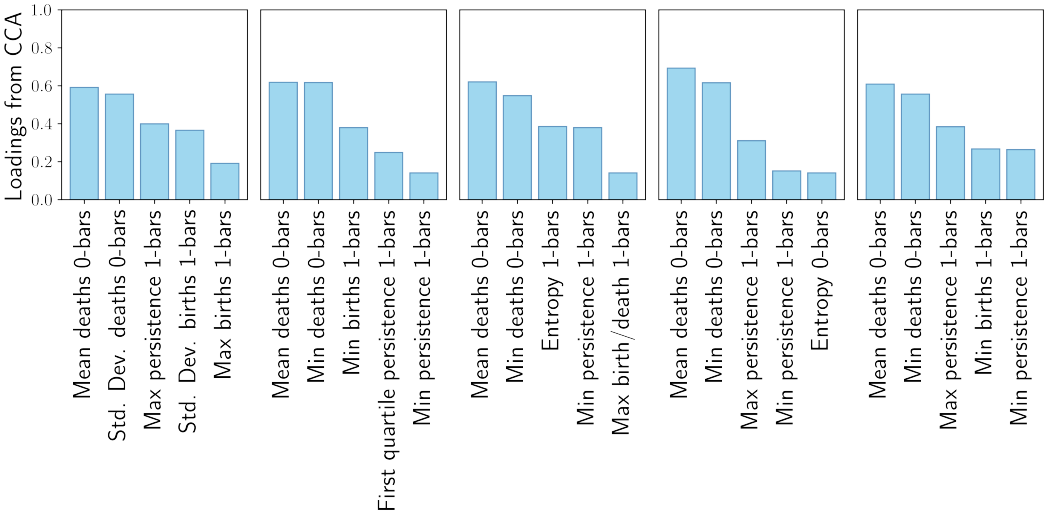


Figure 28: **CCA loadings for clean vs. poisoned activations**. Loadings of the 5 most important contributions to the first canonical variable of the CCA on the pruned barcode summaries show that the mean of the death of 0-bars is significantly correlated with the first two principal components of the PCA across all layers.

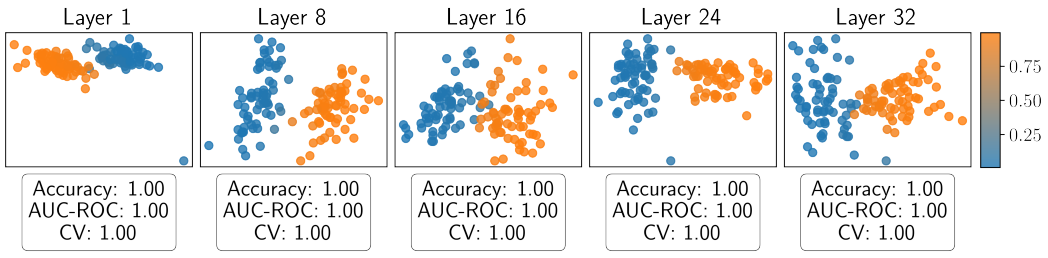


Figure 29: **Logistic regression for clean vs. poisoned activations.** Prediction of a logistic regression trained on a 70/30 train/test split of the pruned barcode summaries, plotted on the projection onto the two first principal components for visualization purposes. Accuracy and AUC–ROC tested on the test data, and 5-fold cross validation on train data are presented for each model, showcasing the outstanding performance of all models.

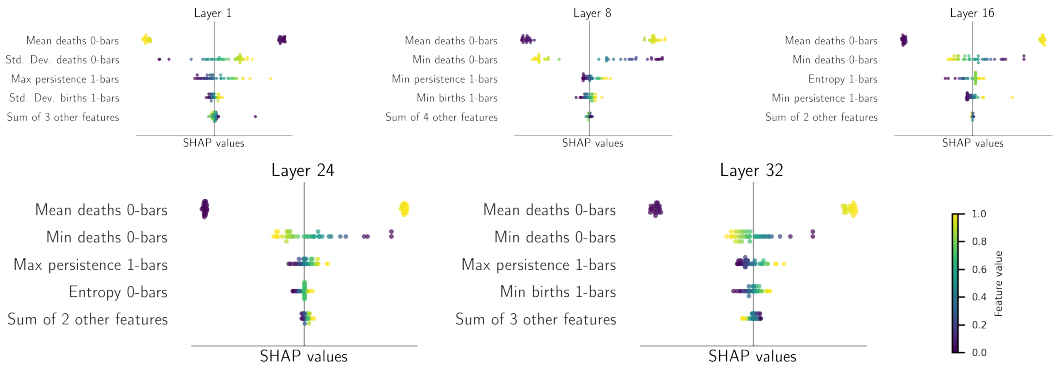


Figure 30: **SHAP analysis: clean vs. poisoned activations.** Beeswarm plot of logistic regression SHAP values trained on the pruned barcode summaries for layer 1, 8, 16, 24, and 32.

C.3.4 LLAMA3 (8B PARAMETERS)

We provide the results of the analysis depicted in Figure 3 including layers 1, 8, 16, 23 and 32 for the Llama 3 (8B parameters) where barcodes are computed using the Euclidean distance in the representation space. We observe very similar results to the ones obtained with Mistral, indicating a consinstency across models of the topological deformations of adversarial influence via XPIA (see Section 3.1).

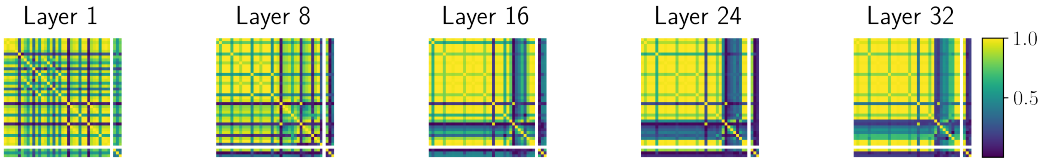


Figure 31: **Cross-correlation matrices for the barcode summaries for clean vs. poisoned activations.** Growing block of correlated features appears in the cross-correlation matrix of the barcode summaries for layers 1, 8, 16, 24, and 32. Correlations in layer 1 are lower than with Mistral 7B, see Figure 6.

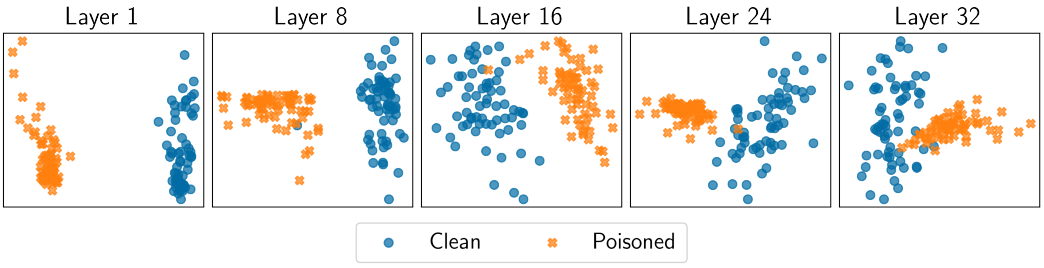


Figure 32: **PCA of barcode summaries of clean vs. poisoned activations**. Clear distinction appears in the projection onto the two first principal components from the PCA of the pruned barcode summaries for layers 1, 8, 16, 24, and 32.

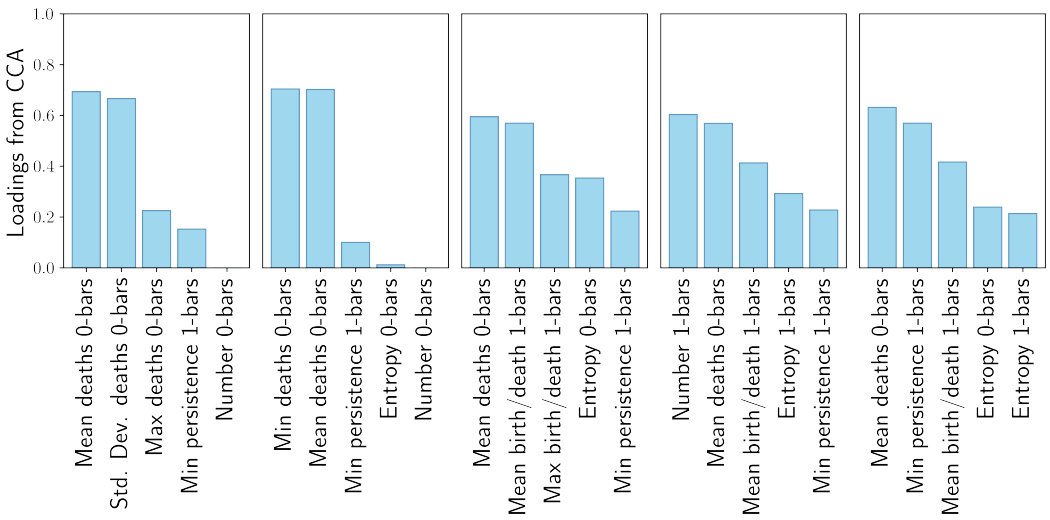


Figure 33: **CCA loadings for clean vs. poisoned activations**. Loadings of the 5 most important contributions to the first canonical variable of the CCA on the pruned barcode summaries show that the mean of the death of 0-bars is significantly correlated with the first two principal components of the PCA across all layers.

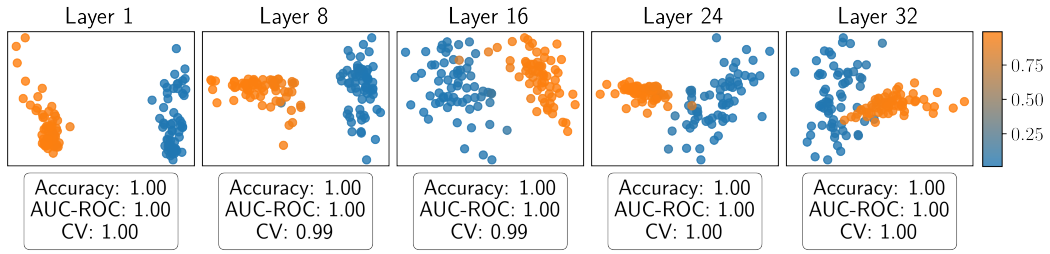


Figure 34: **Logistic regression for clean vs. poisoned activations.** Prediction of a logistic regression trained on a 70/30 train/test split of the pruned barcode summaries, plotted on the projection onto the two first principal components for visualization purposes. Accuracy and AUC–ROC tested on the test data, and 5-fold cross validation on train data are presented for each model, showcasing the outstanding performance of all models.

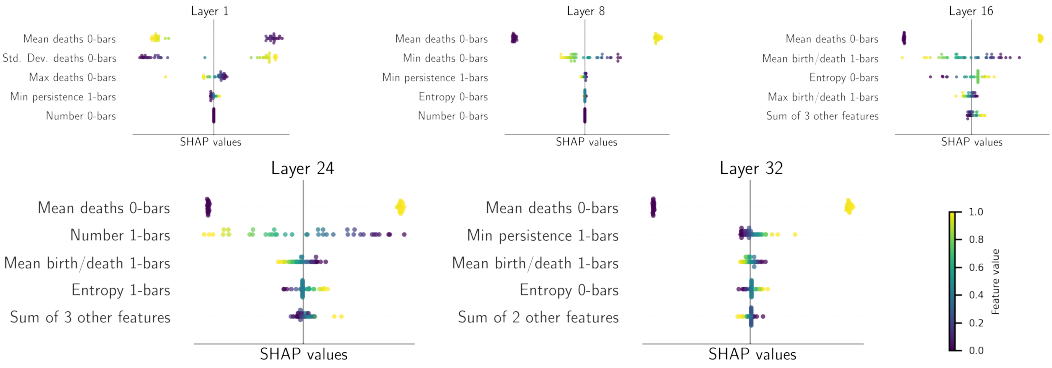


Figure 35: **SHAP analysis: clean vs. poisoned activations.** Beeswarm plot of logistic regression SHAP values trained on the pruned barcode summaries for layer 1, 8, 16, 24, and 32.

C.3.5 PHI3-MEDIUM-128K (14B PARAMETERS)

We provide the results of the analysis depicted in Figure 3 including layers 1, 8, 16, 23 and 32 for the Phi-3-medium (14B parameters) model where barcodes are computed using the Euclidean distance in the representation space. We observe very similar results to the ones obtained with Mistral, indicating a consistency across models of the topological deformations of adversarial influence via XPIA (see Section 3.1).

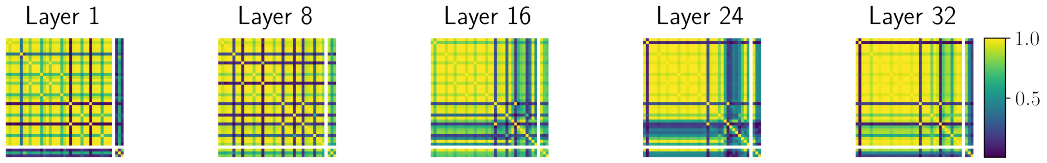


Figure 36: **Cross-correlation matrices for the barcode summaries for clean vs. poisoned activations.** Growing block of correlated features appears in the cross-correlation matrix of the barcode summaries for layers 1, 8, 16, 24, and 32. Correlations in layer 1 are lower than with Mistral 7B, see Figure 6.

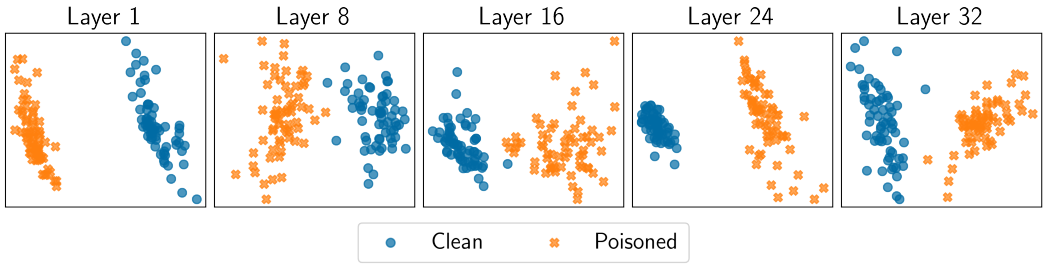


Figure 37: **PCA of barcode summaries of clean vs. poisoned activations**. Clear distinction appears in the projection onto the two first principal components from the PCA of the pruned barcode summaries for layers 1, 8, 16, 24, and 32.

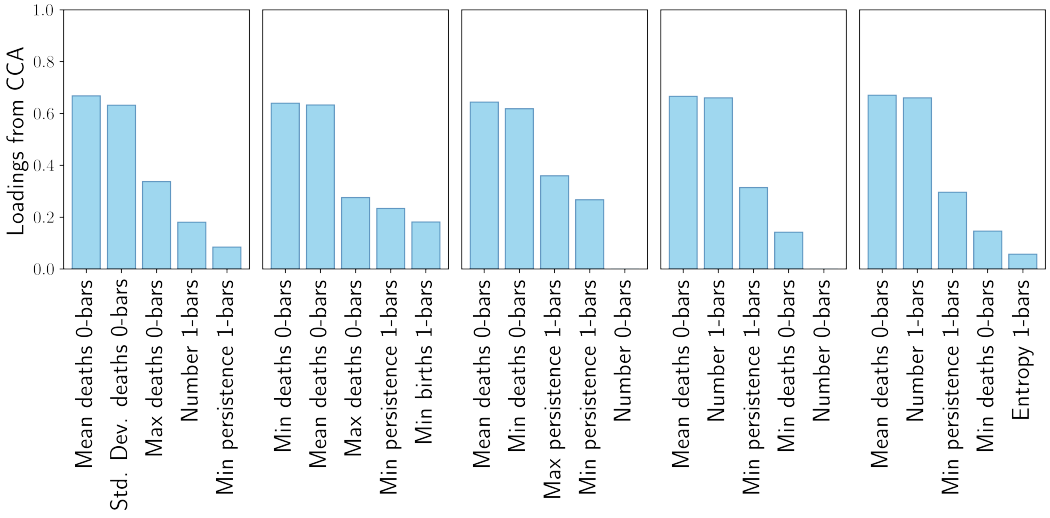


Figure 38: **CCA loadings for clean vs. poisoned activations**. Loadings of the 5 most important contributions to the first canonical variable of the CCA on the pruned barcode summaries show that the mean of the death of 0-bars is significantly correlated with the first two principal components of the PCA across all layers.

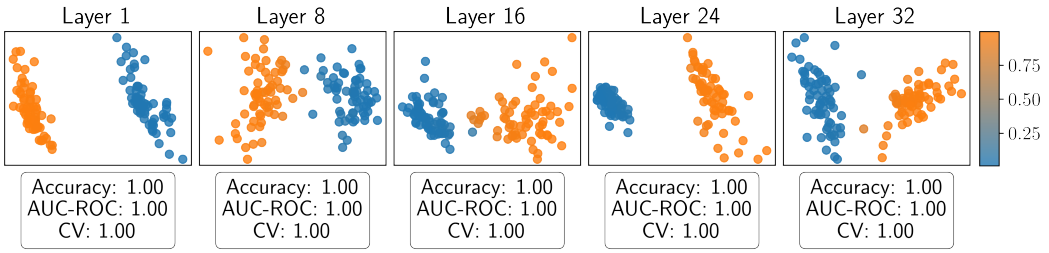


Figure 39: **Logistic regression for clean vs. poisoned activations.** Prediction of a logistic regression trained on a 70/30 train/test split of the pruned barcode summaries, plotted on the projection onto the two first principal components for visualization purposes. Accuracy and AUC–ROC tested on the test data, and 5-fold cross validation on train data are presented for each model, showcasing the outstanding performance of all models.

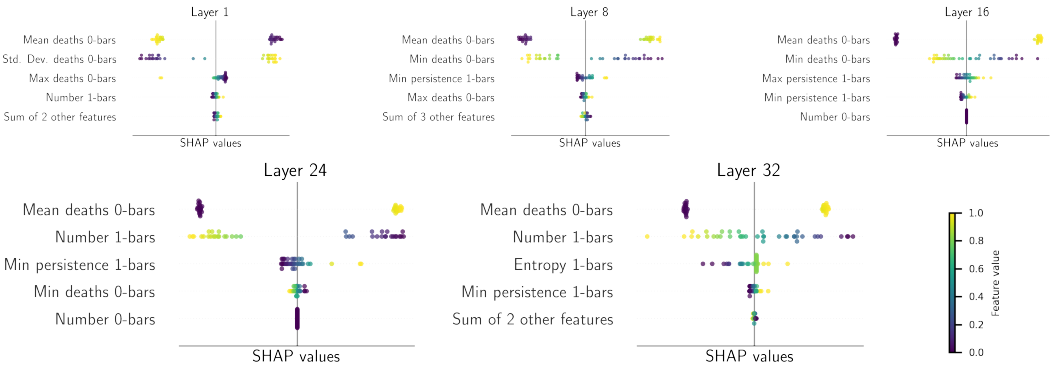


Figure 40: **SHAP analysis: clean vs. poisoned activations.** Beeswarm plot of logistic regression SHAP values trained on the pruned barcode summaries for layer 1, 8, 16, 24, and 32.

C.3.6 LLAMA3 (70B PARAMETERS)

We provide the results of the analysis depicted in Figure 3 including layers 1, 8, 16, 23 and 32 for the Llama 3 (70B parameters) where barcodes are computed using the Euclidean distance in the representation space. We observe very similar results to the ones obtained with Mistral, indicating a consinstency across models of the topological deformations of adversarial influence via XPIA (see Section 3.1).

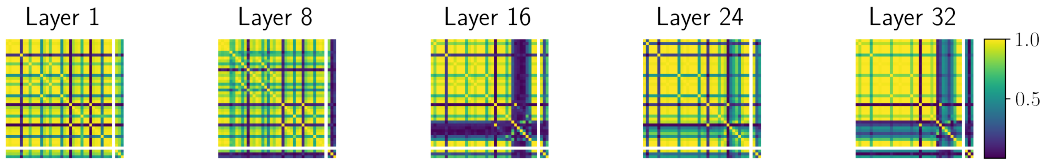


Figure 41: **Cross-correlation matrices for the barcode summaries for clean vs. poisoned activations.** Growing block of correlated features appears in the cross-correlation matrix of the barcode summaries for layers 1, 8, 16, 24, and 32. Correlations in layer 1 are lower than with Mistral 7B, see Figure 6.

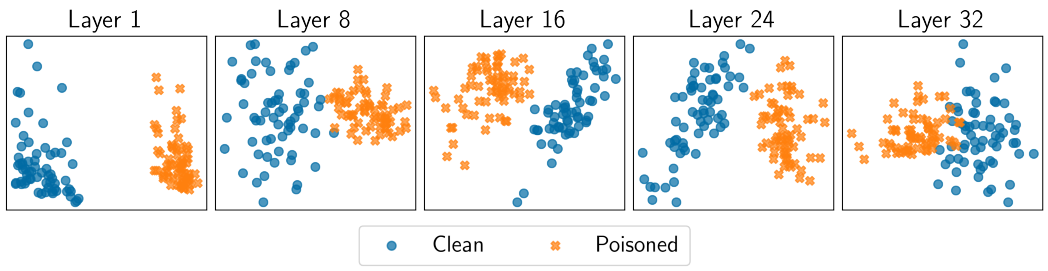


Figure 42: **PCA of barcode summaries of clean vs. poisoned activations**. Clear distinction appears in the projection onto the two first principal components from the PCA of the pruned barcode summaries for layers 1, 8, 16, 24, and 32.

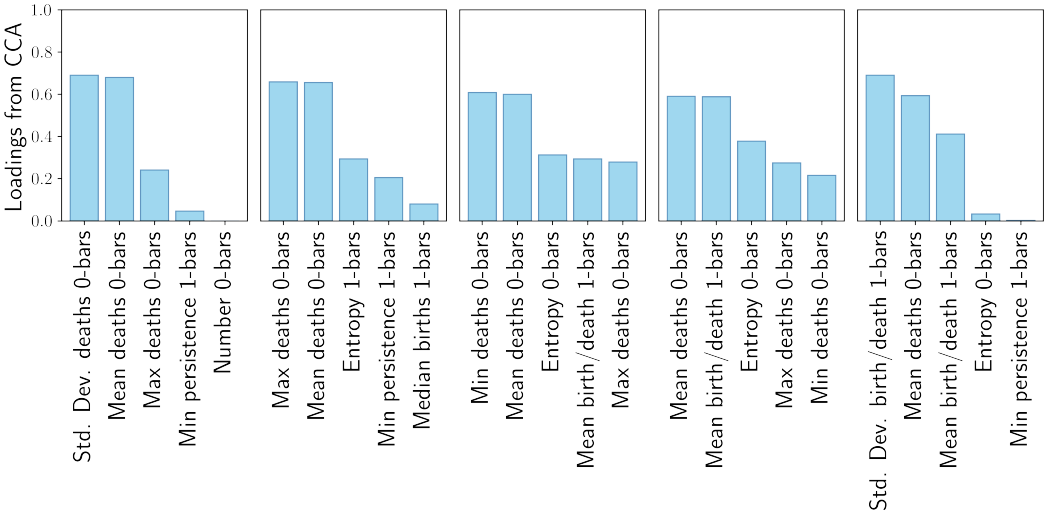


Figure 43: **CCA loadings for clean vs. poisoned activations**. Loadings of the 5 most important contributions to the first canonical variable of the CCA on the pruned barcode summaries show that the mean of the death of 0-bars is significantly correlated with the first two principal components of the PCA across all layers.

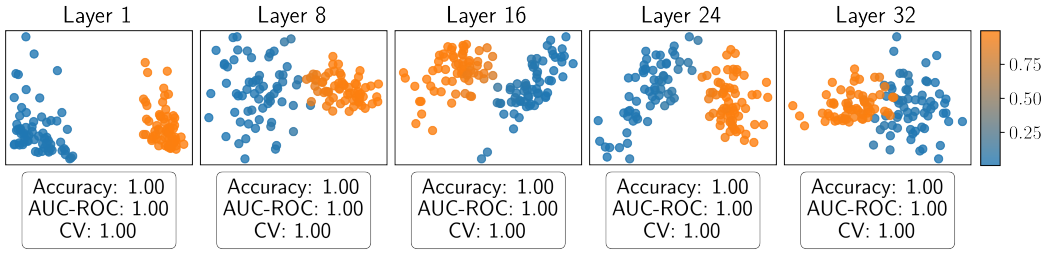


Figure 44: **Logistic regression for clean vs. poisoned activations.** Prediction of a logistic regression trained on a 70/30 train/test split of the pruned barcode summaries, plotted on the projection onto the two first principal components for visualization purposes. Accuracy and AUC–ROC tested on the test data, and 5-fold cross validation on train data are presented for each model, showcasing the outstanding performance of all models.

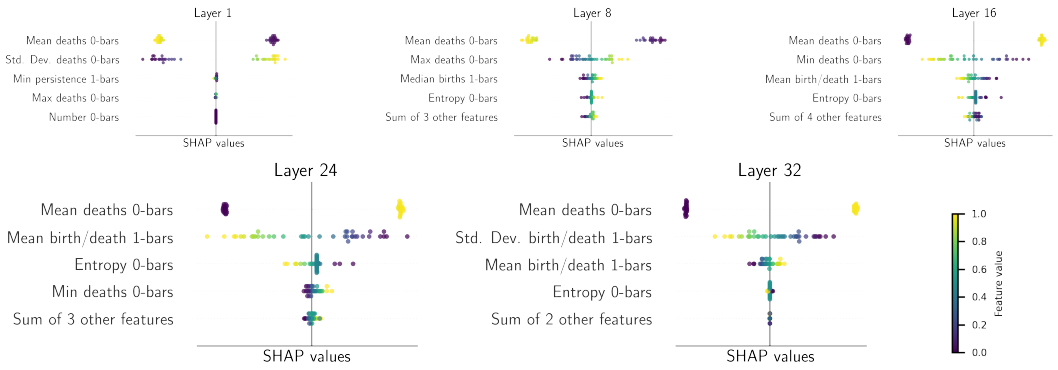


Figure 45: **SHAP analysis: clean vs. poisoned activations.** Beeswarm plot of logistic regression SHAP values trained on the pruned barcode summaries for layer 1, 8, 16, 24, and 32.

C.4 Results: Locked vs. Elicited

C.4.1 Mistral 7B

We include the results of the global analysis in Figure 3 for the locked vs. elicited dataset. There are two main differences with previous results: the block of high correlated features presents a less clear trend and is more faint in layer 16, resulting in the need of more features in the analysis; and the mean death of the 0-bars changes the sign of its influence in classifying locked and elicited models across layers. However the distinction in the PCA of the barcode summaries remains clear and the logistic regression still achieves perfect performance, despite a slightly less straightforward analysis.

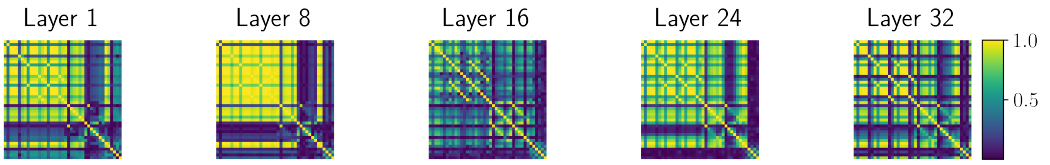


Figure 46: **Mistral with Euclidean distance: Cross-correlation matrices for the barcode summaries for locked vs. elicited activations.** Growing block of correlated features appears in the cross-correlation matrix of the barcode summaries for layers 1, 8, 16, 24, and 32.

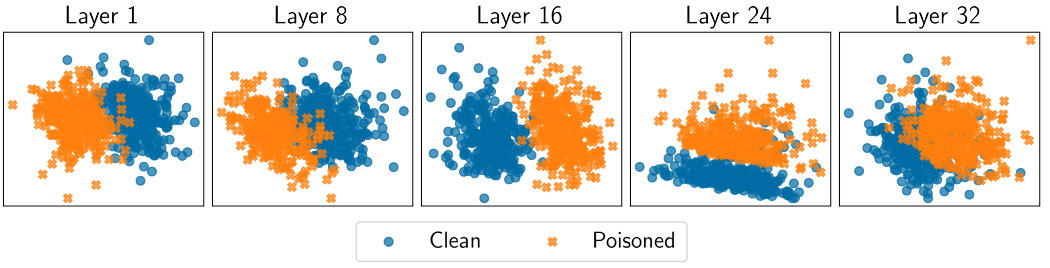


Figure 47: **Mistral with Euclidean distance: PCA of barcode summaries of locked vs. elicited activations**. Clear distinction appears in the projection onto the two first principal components from the PCA of the pruned barcode summaries for layers 1, 8, 16, 24, and 32.

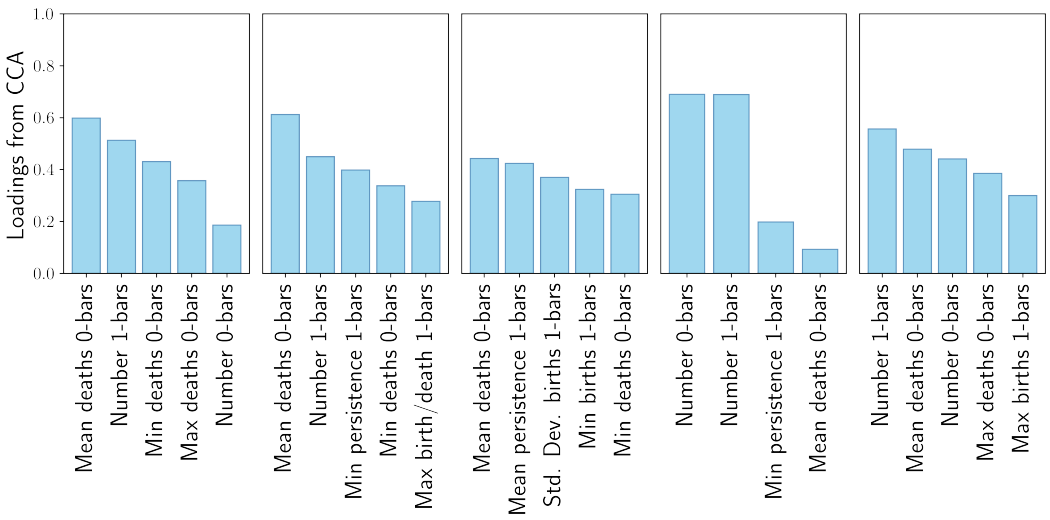


Figure 48: **Mistral with Euclidean distance: CCA loadings for locked vs. elicited activations**. Loadings of the 5 most important contributions to the first canonical variable of the CCA on the pruned barcode summaries show that the mean of the death of 0-bars is significantly correlated with the first two principal components of the PCA across all layers.

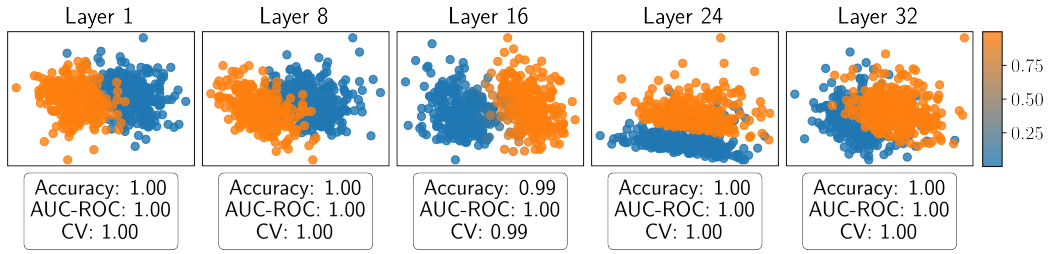


Figure 49: **Mistral with Euclidean distance: Logistic regression for locked vs. elicited activations.** Prediction of a logistic regression trained on a 70/30 train/test split of the pruned barcode summaries, plotted on the projection onto the two first principal components for visualization purposes. Accuracy and AUC–ROC tested on the test data, and 5-fold cross validation on train data are presented for each model, showcasing the outstanding performance of all models.

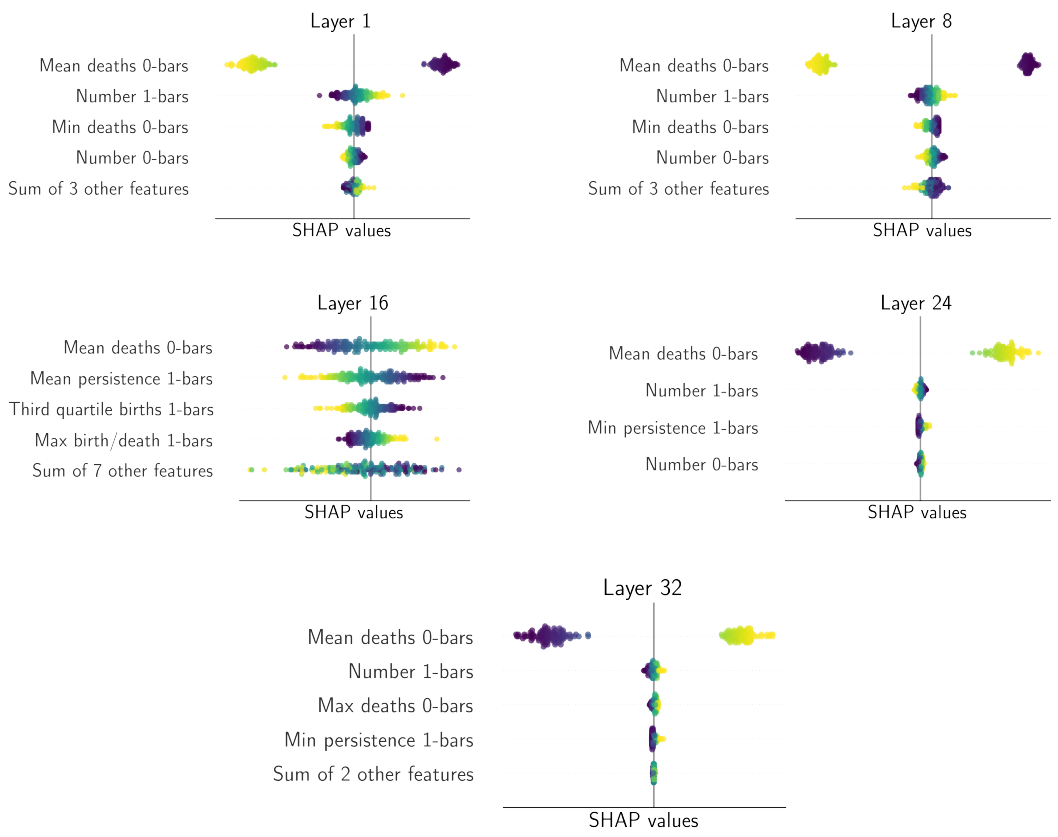


Figure 50: **Mistral with Euclidean distance: SHAP analysis for locked vs. elicited activations.** Beeswarm plot of the SHAP values for the logistic regression trained on the pruned barcode summaries for layer 1, 8, 16, 24, and 32. The mean of the deaths of 0-bars appears as the most impactful feature in the prediction of the model, shifting predictions to "locked" when the value of the feature is lower for layers 8, 16, 23, and 32, and to "elicited" when it is higher. The opposite phenomenon is observed in layer 0.

C.4.2 LLAMA3 (8B PARAMETERS)

We include the results of the global analysis in Figure 3 for the locked vs. elicited dataset. Here we also observe less clear patterns of correlations in the topological features, particularly for latter layers. Despite the mean of the death of 0-bars remaining as one of the key features in the CCA, the interpretation of the Shapley values is less straightforward in this case as the dichotomous behavior of these for the mean of the 0-bars disappears for latter layers.

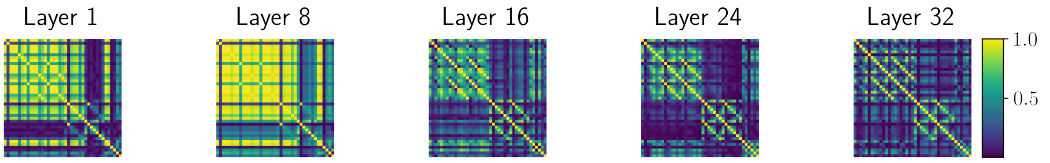


Figure 51: **Llama with Euclidean distance: Cross-correlation matrices for the barcode summaries for locked vs. elicited activations.** Decreasing block of correlated features appears in the cross-correlation matrix of the barcode summaries for layers 1, 8, 16, 24, and 32.

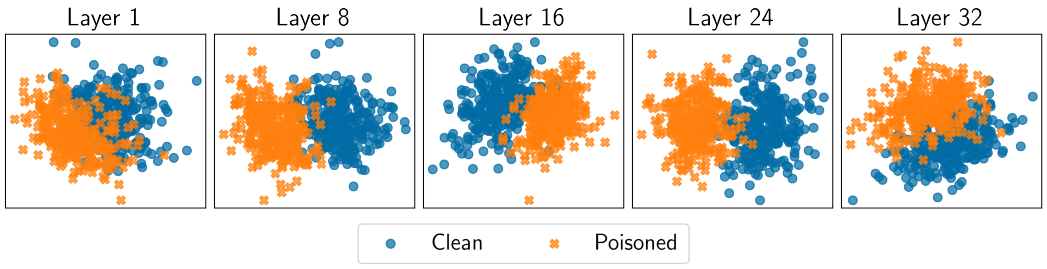


Figure 52: **Llama with Euclidean distance: PCA of barcode summaries of locked vs. elicited activations**. Clear distinction appears in the projection onto the two first principal components from the PCA of the pruned barcode summaries for layers 1, 8, 16, 24, and 32.

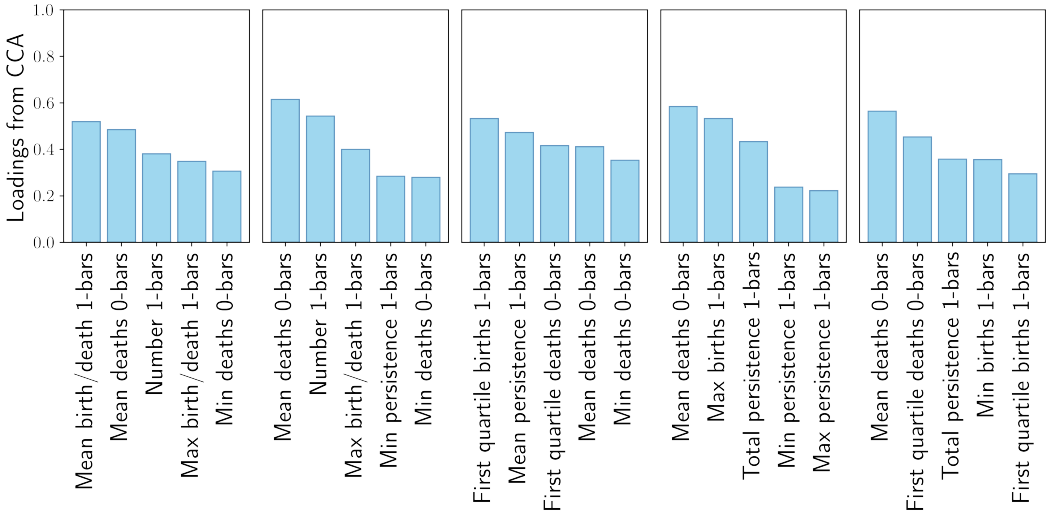


Figure 53: **Llama with Euclidean distance: CCA loadings for locked vs. elicited activations**. Loadings of the 5 most important contributions to the first canonical variable of the CCA on the pruned barcode summaries show that the mean of the death of 0-bars is significantly correlated with the first two principal components of the PCA across all layers.

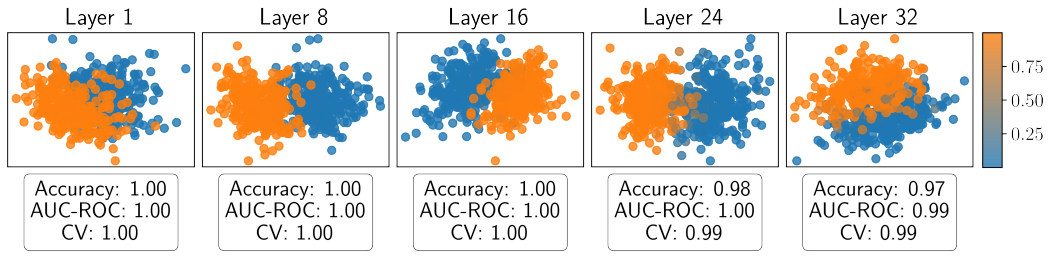


Figure 54: **Llama with Euclidean distance: Logistic regression for locked vs. elicited activations.** Prediction of a logistic regression trained on a 70/30 train/test split of the pruned barcode summaries, plotted on the projection onto the two first principal components for visualization purposes. Accuracy and AUC–ROC tested on the test data, and 5-fold cross validation on train data are presented for each model, showcasing the outstanding performance of all models.

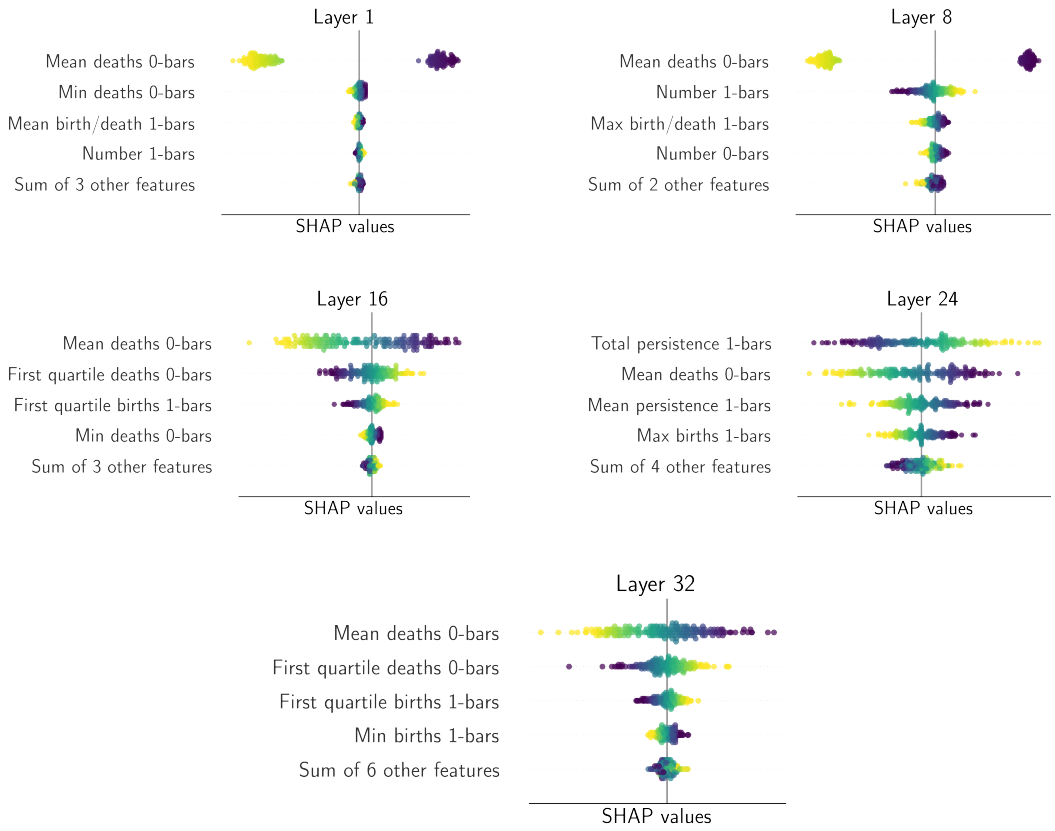


Figure 55: **Mistral with Euclidean distance: SHAP analysis for locked vs. elicited activations.** Beeswarm plot of the SHAP values for the logistic regression trained on the pruned barcode summaries for layer 1, 8, 16, 24, and 32. The mean of the deaths of 0-bars appears as the most impactful feature in the prediction of the model, shifting predictions to "locked" when the value of the feature is lower for layers 8, 16 and 32, and to "elicited" when it is higher. For layer 24, the total persistence of 1-bars appears as the most important feature. Lower number of 1-bars classifies the point as "locked" while higher values push the prediction toward "elicited".

D Further Details on Local Analysis

In this section we provide further details to the local analysis in Section 3.3.

D.1 Pipeline

Within this local analysis, we aim to determine the interaction of elements of the neural network across the layers by taking representations across pairs of layers as coordinates in 2 dimensions (2D). We study this across three models: Mistral, Phi3 3.8B and LLaMA3 8B. For each of these models, we take a sample of 2000 from each model, 1000 of which are clean activations and 1000 of which are poisoned activations. Each element along the layer given their embedding into 2D can be thought of as nodes in a graph with weighted connections based on the Euclidean distances between the points. On these graphs, we construct the Vietoris–Rips filtration and compute the resulting persistence barcode which describes the topology of the interactions between the elements.

For this local analysis, we focus on a smaller selection of persistence barcode summaries, including measures such as the mean death of 0-bars, total persistence of 0- and 1-bars, and persistent entropy, while excluding measures such as the quantiles of death bars. We compute these summary statistics and track their progression across pairs of layers in the models. We presented one such progression within Figure 10 in Section 3.3, which captures how total persistence changes over the layers and

is distinct from the control case. In the following sections, we include further plots to support this argument.

D.2 Results

D.2.1 Mistral Model

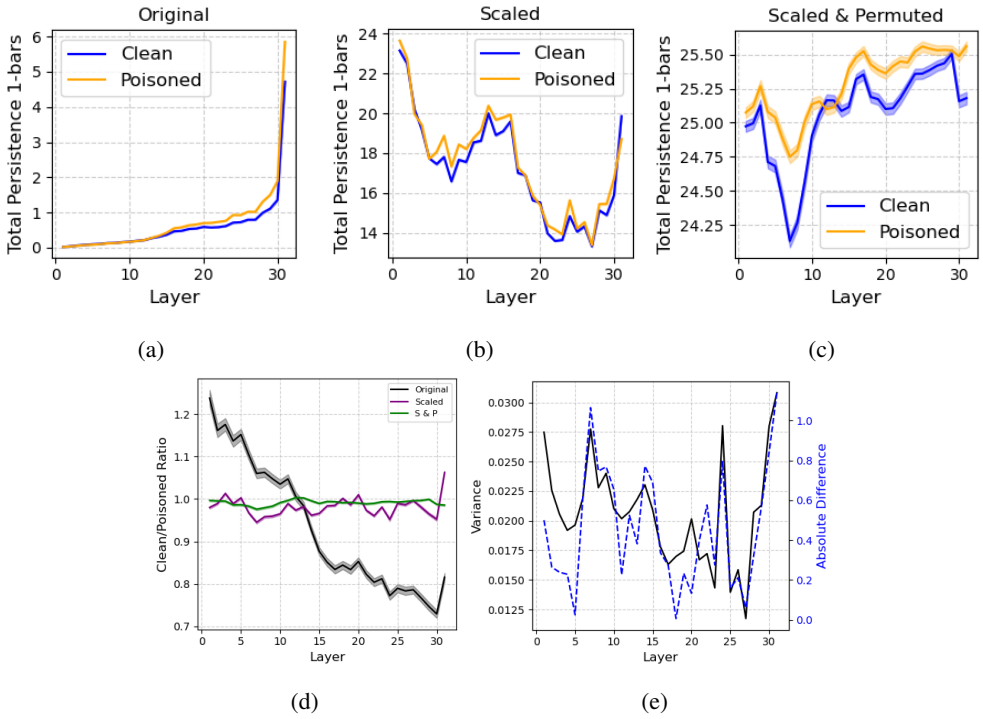


Figure 56: **Local analysis of consecutive layers for the total persistence of 1-bars for the Mistral model.** Comparisons of the average total persistence of 1-bars across 1000 samples for Mistral model for original **(a)**, scaled/normalized **(b)** and scaled and permuted **(c)** activation data. **(d)** Ratios of mean total persistence of 1-bars between clean and poisoned datasets for original, scaled, and scaled and permuted activations. **(e)** Overlaid plots of the overall variance of total persistence of 1-bars for clean and poisoned datasets combined and the absolute difference between mean total persistence of 1-bars for clean and poisoned datasets.

In addition to the propagation of total persistence of 1-bars we showed in Section 3.3 and in this section of the Appendix, we also evaluated the progression of other barcode summaries. Notably, descriptors which capture similar features are the mean deaths of 1-bars, and the mean birth of 0 bars with mirroring patterns. In Figure 57, we show the results for the mean death of 0-bars.

D.2.2 Phi3 Model

We present a similar comparison of results for the Phi3 model. Figure 58 illustrates the patterns across layers for the mean death of 0-bars, while Figure 59 shows the patterns for the total persistence of 1-bars. Unlike the Mistral model, the ratio between barcode statistics for clean and poisoned activations in the Phi3 model does not intersect one. While a decreasing or somewhat parabolic trend is still observed, the average mean death of 0-bars and the total persistence of 1-bars for clean raw activations consistently remain greater than those for poisoned raw activations. Additionally, we find that the "control" case remains close to the x-axis, with the scaled ratios exhibiting significant variations around this baseline.

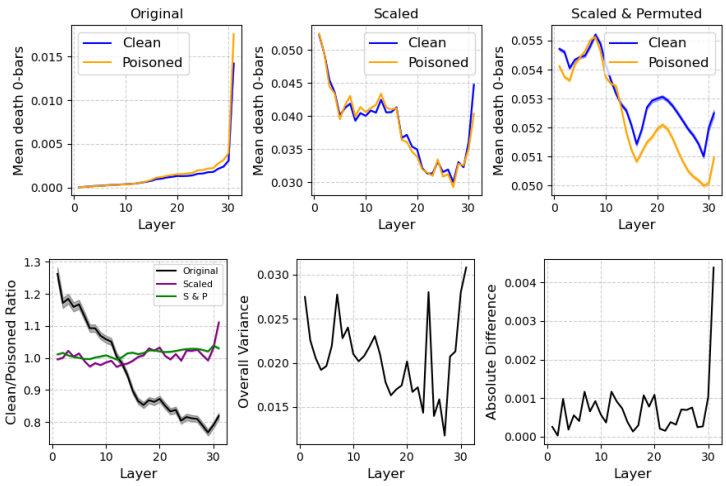


Figure 57: **Local analysis of consecutive layers for the mean deaths of 0-bars for the Mistral model. Top:** Comparisons of the average of mean deaths of 0-bars across 1000 samples for the Mistral model for original (raw), scaled (normalized) and scaled & permuted activation data. **Bottom left:** Ratios of average mean deaths of 0-bars between clean and poisoned datasets for original, scaled and scaled & permuted activations. **Bottom center:** Overall variance of mean deaths of 0-bars for clean and poisoned datasets combined. **Bottom right:** Absolute difference between mean total persistence of 1-bars for clean and poisoned datasets.

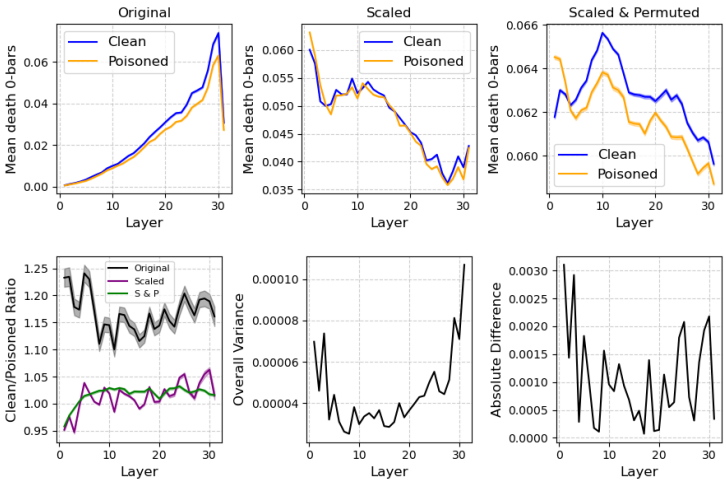


Figure 58: **Local analysis of consecutive layers for the mean deaths of 0-bars for the Phi3 model. Top:** Comparisons of the average of mean deaths of 0-bars across 1000 samples for Phi3 model for original (raw), scaled (normalized) and scaled & permuted activation data. **Bottom left:** Ratios of average mean deaths of 0-bars between clean and poisoned datasets for original, scaled and scaled & permuted activations. **Bottom center:** Overall variance of mean deaths of 0-bars for clean and poisoned datasets combined. **Bottom right:** Absolute difference between mean total persistence of 1-bars for clean and poisoned datasets.

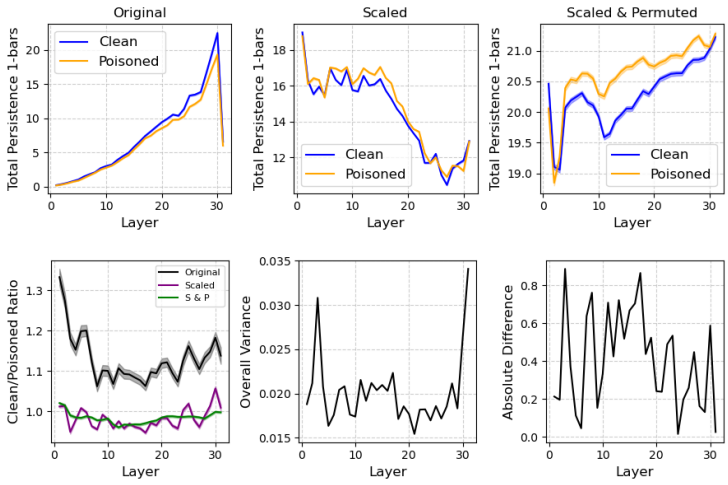


Figure 59: **Local analysis of consecutive layers for the total persistence of 1-bars for the Phi3 model. Top:** Comparisons of the average of total persistence of 1-bars across 1000 samples for Phi3 model for original (raw), scaled (normalized) and scaled & permuted activation data. **Bottom left:** Ratios of average total persistence of 1-bars between clean and poisoned datasets for original, scaled and scaled & permuted activations. **Bottom center:** Overall variance of total persistence of 1-bars for clean and poisoned datasets combined. **Bottom right:** Absolute difference between mean total persistence of 1-bars for clean and poisoned datasets.

D.2.3 LLaMA3 8B Model

We present the results for the LLaMA3 8B model. Figures 60 and 61 both show a decreasing trend in the ratio between clean and poisoned activations, whether measured by the mean death of 0-bars or the total persistence of 1-bars respectively. Notably, this ratio crosses 1 around layer 15 or later. Moreover, we continue to observe distinct differences between clean and poisoned activations across both meaningful variants.

D.2.4 Peak Analysis for Phi3 and LLaMA3

Table 7: **Peak analysis.** Precision@k for k=1, 3, and 5 largest peaks in total variance, and their precision in detecting the largest peaks in absolute difference between the two classes. Spearman's rank correlation (r) is reported in the last column. *, ** correspond to p-values $<$.05 and .01, respectively.

Phi3	*p@1*	*p@3*	*p@5*	r
Total Persistence 0-bars	0	.33	.2	0.69**
Total Persistence 1-bars	1.0	.67*	.8**	0.50**
Mean Birth 1-bars	0	.33	.6*	0.66**
Mean Death 1-bars	0	.67*	.8**	0.35
LLAMA3	*p@1*	*p@3*	*p@5*	r
Total Persistence 0-bars	1.0*	.33	.4	0.60**
Total Persistence 1-bars	1.0*	.67	.8**	0.93**
Mean Birth 1-bars	1.0*	.67	.6	0.60**
Mean Death 1-bars	1.0*	.67*	.8*	0.93**

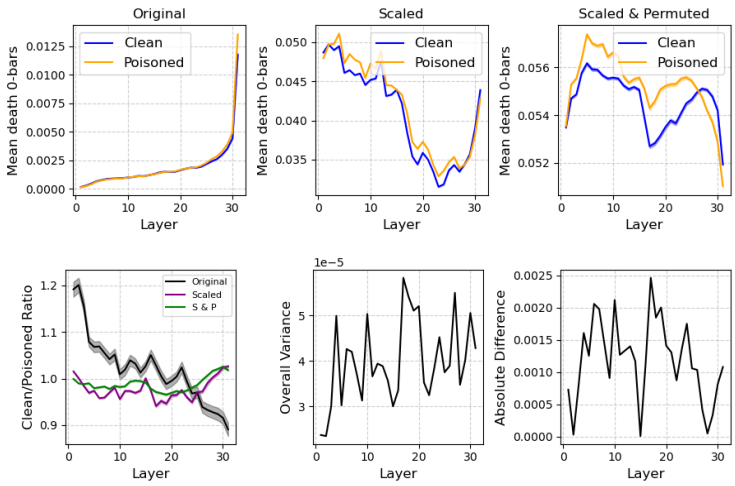


Figure 60: **Local analysis of consecutive layers for the mean deaths of 0-bars for the LLaMA3 8B model. Top:** Comparisons of the average of mean deaths of 0-bars across 1000 samples for LLaMA3 8B model for original (raw), scaled (normalized) and scaled & permuted activation data. **Bottom left:** Ratios of average mean deaths of 0-bars between clean and poisoned datasets for original, scaled and scaled & permuted activations. **Bottom center:** Overall variance of mean deaths of 0-bars for clean and poisoned datasets combined. **Bottom right:** Absolute difference between mean total persistence of 1-bars for clean and poisoned datasets.

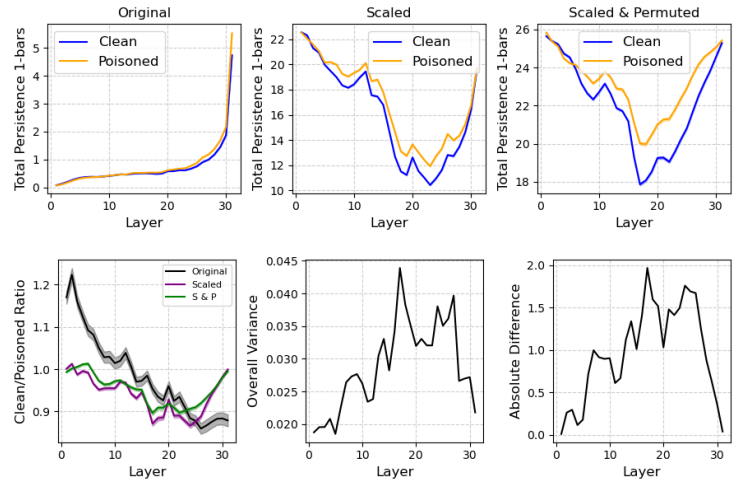


Figure 61: **Local analysis of consecutive layers for the total persistence of 1-bars for the LLaMA3 8B model. Top:** Comparisons of the average of total persistence of 1-bars across 1000 samples for the LLaMA3 8B model for original (raw), scaled (normalized) and scaled & permuted activation data. **Bottom left:** Ratios of average total persistence of 1-bars between clean and poisoned datasets for original, scaled and scaled & permuted activations. **Bottom center:** Overall variance of total persistence of 1-bars for clean and poisoned datasets combined. **Bottom right:** Absolute difference between mean total persistence of 1-bars for clean and poisoned datasets.

D.2.5 NON-CONSECUTIVE LAYER ANALYSIS

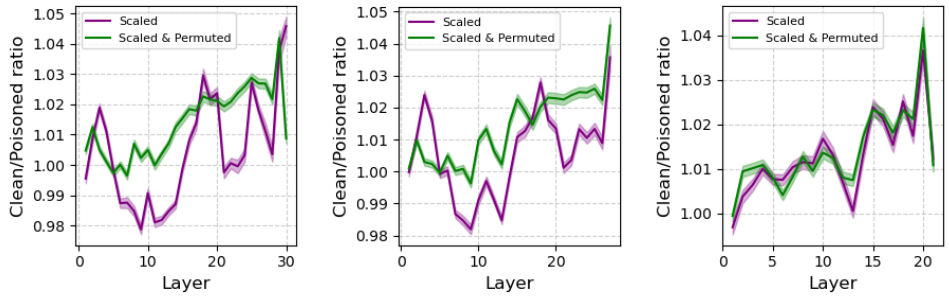


Figure 62: **Local analysis of non-consecutive layers for mean death of 0-bars.** Comparison of ratios between mean death of 0-bars for clean and poisoned datasets when considering topology pairs of layers at 1 (left), 3 (middle), and 10 (right) intervals apart.

Continuing the analysis of non-consecutive layers, we examine how increasing layer separation affects the contrast between clean and poisoned activations across different barcode summaries. Figure 62 shows the ratio of the mean death times of 0-bars, while Figure 63 shows the ratio of the total persistence of 1-bars. For both summaries, we observe that at a 10-layer separation, the similarity between the ratios for the scaled and control settings becomes more pronounced compared to separations of 1 or 3 layers, indicating a reduced distinction between clean and poisoned interactions at larger layer intervals.

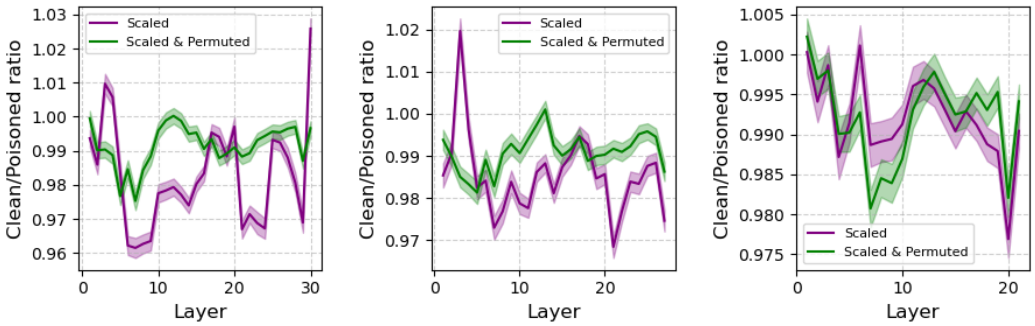


Figure 63: **Local analysis of non-consecutive layers on total persistence of 1-bars for the Mistral model.** Comparison of the ratios between total persistence of 1-bars for clean and poisoned datasets when considering the topology of interactions between pairs of layers at 1 (left), 3 (middle), and 10 (right) intervals apart.

E SANDBAGGED REPRESENTATION DATA COLLECTION

To study sandbagging, we follow the "password-locking" paradigm introduced in prior work (Greenblatt et al., 2024; van der Weij et al., 2024), wherein a model is fine-tuned to hide certain behaviors or capabilities unless presented with a predefined password in the prompt. We leverage the the synthetic data generation process and fine-tuning procedure detailed in Section 3 and 4 of van der Weij et al. (2024). For clarity of our experimental setup, we summarize it briefly here. We extend it to extract final token hidden representations from the model under both elicited and elicited conditions.

1. We generate a synthetic dataset of 3,347 evaluation questions based on the the Weapons of Mass Destruction Proxy (WMDP) benchmark (Li et al., 2024) using GPT-4 (Perez et al., 2022) to serve as a proxy measurement for hazardous knowledge in biology, chemistry, and cybersecurity. This process protects against data leakage by never including actual WMDP

items in the prompt and filters out any questions the base models cannot solve to ensure that the fine-tuned model genuinely has the underlying capability.

2. Using this dataset and framework, we fine-tuned Mistral 7B and LLaMA3 8B with LoRA (Hu et al., 2021). During this fine-tuning process, each WMDP-like evaluation question appears twice: once with the password, where the model is fully elicited and is provided with the correct answer and second where the model is password is absent or locked and the model is provided with an incorrect answer.
3. Using the held-out validation set of generated WMDP-like questions, we recorded the final-token representation from each layer of the fine-tuned model during its forward pass. Specifically, for each prompt x, we collect:

$$\mathrm{Act}_\ell^{(\mathrm{locked})}(x) = \mathrm{Hidden}_\ell\big(x_{\mathrm{locked}}\big)[-1],$$
$$\mathrm{Act}_\ell^{(\mathrm{elicited})}(x) = \mathrm{Hidden}_\ell\big(x_{\mathrm{elicited}}\big)[-1],$$

We then apply the global layer-wise topological analysis outlined in Section 3.2 to this representation data to study how the model's latent space differs between elicited (no password) and elicited (correct password) states.

F Global PH Pipeline and Resource Constraints

All Vietoris–Rips barcodes are computed with the GPU build of Ripser++ on a single node equipped with four NVIDIA A100 GPUs (80 GB each). Per layer we draw $K = 128$ independent subsamples of $k = 4096$ activation vectors (64 clean, 64 adversarial). Subsamples are dispatched round-robin to two concurrent Ripser++ kernels per GPU.

Memory Footprint. A complete $k = 4096$ complex truncated at dimension 2 occupies only 2.1 ± 0.4 GB of device memory (95th percentile < 2.8 GB; Tab. 8), leaving a wide margin inside the 80 GB budget, even when two barcodes are built concurrently on the same GPU.

Throughput. The mean walltime per barcode is 36.8 ± 0.6s (95th percentile $<$ 40s). With four GPUs processing eight barcodes in parallel, a full layer (128 barcodes) finishes in ≈ 10 min and the five-layer suite of one model in ≈ 50 min. Running the six models serially therefore completes in about five hours on a single 4 × A100 node—comfortably within the nightly maintenance window.

Table 8: **Computational Costs.** Per-barcode wall-clock time and GPU-memory consumption ($k = 4096$, dimension ≤ 2). Statistics over $K = 64$ barcodes drawn from the LLAMA-3 8B activations.

Layer	time $\mu \pm \sigma$ [s] (p95)	memory $\mu \pm \sigma$ [GB]
1	38.34 ± 0.76 (39.6)	2.27 ± 0.34
8	36.79 ± 0.70 (38.0)	2.12 ± 0.39
16	36.68 ± 0.45 (37.4)	2.13 ± 0.30
24	36.63 ± 0.71 (38.1)	2.03 ± 0.33
32	36.62 ± 0.54 (37.4)	2.20 ± 0.344

After choosing $K = 64$, we recomputed the Monte-Carlo variance σ_f^2 from the raw, unscaled feature values. For 39 out of 41 statistics, we found $\sigma_f < 0.10$, which would put the standard error $\mathrm{SE} = \sigma_f/\sqrt{K}$ below $\Delta^\star/2 = 0.025$ with only $K \leq 20$. The outlier features were those which aggregate counts—total persistence of H_0 and the raw count of H_1 bars—and need to be transformed for their variance to be directly comparable to the other features. These do not affect the classifier as the features are scaled prior to training and also do not appear as the most informative features for distinguishing between clean and posioned PH-derived features. We conservatively choose $K = 128$ and the resulting ROC–AUCs on the logistic regression model trained only on barcodes are perfect (1.00 ± 0.00), confirming that the subsampling budget is more than sufficient to validate the significance of the features derived from PH, while balancing GPU memory and computation time.

G Robustness to Adaptive Attacks

We tested the robustness of our identified topological features against real-world attack examples from Microsoft's large-scale `LLMail-Inject` dataset (Abdelnabi et al., 2025) which includes XPIA attack examples and information on their efficacy against four distinct defenses and a fifth setup involving a stacked arrangement of all four. All attack examples are sourced from a public red teaming competition. These attacks are particularly relevant as they include examples that were specifically designed to evade the TaskTracker activation-based defense (Abdelnabi et al., 2024), the source of our primary XPIA data. Thus, applying our topological framework to these attack examples is a particularly strong test of whether our topological features represent a fundamental shift in the shape and structure of LLM latent space, or whether it is an artifact that can be easily subverted.

Methodology. As the `LLMail-Inject` dataset does not contain paired clean examples, we synthetically generated and manually verified 100 clean counterparts using `Phi-3-medium-4k-instruct`. The small sample size was chosen to ensure that we could verify the quality of the synthetically generated clean examples. We then generated last-token activation data from layer 16 of `Mistral-7B-Instruct-v0.2` for both the clean and adaptive attack inputs and computed the corresponding barcode summary statistics.

Results. The topological features of the activation spaces under these adaptive attacks show a clear distinction from the clean examples, as summarized in Table 9. The results show a clear shift towards a simpler, more dispersed topology under adversarial influence.

Table 9: **PH barcode statistics for clean vs. adaptive attack.** Comparison of barcode summary statistics of clean vs. adaptive attack activations from the `LLMail-Inject` dataset on Mistral-7B (Layer 16).

PH Feature	Clean	Attack (Adaptive)
H0 Count	64	50
H0 Death Time (Median)	56.07	58.85
H0 Death Time (Mean $\pm$ SD)	55.43 ± 21.23	51.73 ± 28.38
H1 Count (Loops)	**12**	**4**
H1 Birth Time (Median)	**69.36**	**84.92**
H1 Death Time (Median)	71.59	86.20

- **Fewer, Larger-Scale Loops:** The number of 1-dimensional loops (H1 bars) decreases significantly from 12 in the clean data to just 4 in the adversarial data. Furthermore, their median birth time increases from ≈ 69 to ≈ 85, indicating that the remaining topological features are formed at much larger scales.
- **More Dispersed Clusters:** The median H0 death time increases, supporting the hypothesis of greater dispersion. We note that the *mean* H0 death time appears to contradict this trend (decreasing from 55.43 to 51.73). This is due to a small subset of components in the adversarial data merging at very low scales. The median, being more robust to such outliers, better captures the overall geometric shift towards a more spread-out structure.

These findings further suggest that the topological compression signature we identify across models and across XPIA and sandbagging attack conditions reflects a fundamental property of adversarial influence, as the signature remains detectable even against attacks optimized to evade XPIA defenses, including but not limited to the TaskTracker activation-based defense.