Using Topological Data Analysis to Characterize the Layers of Language Models Before and After Word Substitution Attacks

Adam Tang*[1], Catherine Liu[1], Kimberly Lopez[1], Shreya Subramanian[1], Leif Zinn-Brooks[1], Alexia Schultz[2], Adaku Uchendu*[2]
[1]Harvey Mudd College Claremont, CA, USA
[2]MIT Lincoln Laboratory, Lexington, MA, USA
***Corresponding authors:** adtang@g.hmc.edu, adaku.uchendu@ll.mit.edu

Abstract

Large language models are known to be vulnerable to adversarial perturbations such as synonym-based word substitutions. However, previous analyses of adversarial influence focus only on output behavior and provide limited insight into the propagation of substitution-based input perturbations through internal representations. In this work, we introduce a topological data analysis (TDA) framework to study the structural effects of adversarial attacks on attention maps across model layers. We evaluate small encoder-based architectures (BERT, RoBERTa, DistilBERT) fine-tuned to solve binary classification on the IMDb review dataset, which were attacked using TextFooler. We convert attention maps into distance matrices and apply TDA to extract topological features, which we then compare using Wasserstein distances between original and perturbed features. In parallel, we compute a non-TDA baseline on attention maps using per-head L_1 distances between original and perturbed attentions. In addition, we analyze these models on a layer-by-layer basis. We find that adversarial perturbations induce systematic and statistically significant topological changes across layers, with the largest deviations occurring in late layers and smaller but notable effects in early layers. These patterns are consistent across models and are validated using both non-parametric (Kruskal–Wallis, Dunn) and parametric (one-way ANOVA, Tukey) tests on log-transformed Wasserstein distances. Compared to our non-TDA baseline, our results show more distinct layer-wise separation and provides a robust and interpretable framework for evaluating how adversarial perturbations alter internal model structure. Our code is publicly available at: https://github.com/angelinatsai04/mitll_clinic/tree/adam_spring[1].

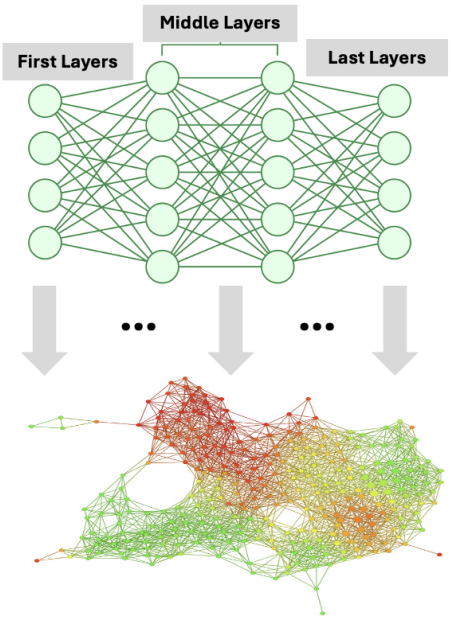

Figure 1: Illustration of using TDA to characterize the topology of the latent space, tracking adversarial activity across model layers.

1 Introduction

Large language models (LLMs) have proven to be strong tools in solving many tasks in NLP, from part-of-speech tagging to sentiment analysis to spam detection. Furthermore, LLMs are easily customizable to fit several needs (Chen et al., 2024), making them suitable for a variety of applications. However, while they provide great utility, they suffer from vulnerabilities due to the introduction of adversarial perturbations. In some cases, LLMs may produce very different outputs when semantic-preserving changes are made to the input (Arakelyan et al., 2024). One such case is point-based word substitution (Qi et al., 2021), in which

[1]DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8702-15-D-0001 or FA8702-25-D-B002. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Air Force.

individual words in a text input are replaced with synonyms, which may have undesirable effects on LLM-based text classification models. Therefore, before deploying LLM solutions, we need to understand what scenarios (i.e., adversarial perturbations encourage or discourage vulnerabilities in language models.

We study TextFooler (Jin et al., 2020), a word substitution attack that has a high success rate in disrupting classification predictions for fake news prediction and sentiment classification tasks (Jin et al., 2020). Like other model-agnostic attacks, TextFooler evaluates robustness at the output level, but provides few insights into how the input perturbation propagates through a model. Despite this limitation, TextFooler generates a rich set of "successful" perturbations to the output (Jin et al., 2020). Consequently, building defense strategies for this type of attack is difficult. Thus, we use TextFooler, a black-box attack, to adversarially perturb texts and probe how these perturbed texts (in comparison with the original) are represented in the embedding space of encoder-only models. Furthermore, we track how perturbations change the embedding space across layers.

To investigate the robustness of a model's internal states, we evaluate the structure of *attention maps* (Kushnareva et al., 2021), which are intermediate representations of an input that show how much each token contributes to the meaning of each other token. We introduce the use of topological data analysis (TDA) in order to better quantify the effect of TextFooler on each layer's attention maps and to visualize changes in the structure of a model's internal representations across layers. See Figure 1 for an illustration on how TDA can be used to characterize the topology of each layer in a language model to investigate adversarial influence. Using these tools, we investigate the following three research questions (RQs):

- **RQ1**: Can we observe distinct topologies in the attention maps of encoder models, before and after word substitution?
- **RQ2**: Are these distinct topologies observable across the layers of a model?
- **RQ3**: Are these distinct topologies observable across different encoder models?

Finally, our findings reveal that adversarial perturbations such as word substitution attacks do affect the topology of the embedding space of language models. In addition, we find that adversarial influence is most noticeable in the early and late layers for the encoder models (i.e., BERT, RoBERTa, DistilBERT) investigated.

2 Related Work

Adversarial attacks have been used to evaluate the robustness of AI systems. Specifically, for LLMs, these range from prompt-based attacks (Zhang et al., 2025), using LLMs to paraphrase texts to evade author attribution (Alperin et al., 2025), to attacks that probe LLM embeddings like poisoning attacks (Fendley et al., 2025). We have also observed several defense techniques and more recently the application of TDA to detect adversarial activity (Uchendu and Le, 2024). This includes using TDA to track topological changes in hidden space as a result of adversarial attacks (Fay et al., 2026; Vu et al., 2025; Chauhan and Kaul, 2022; Perez and Reinauer, 2022). These applications suggest that TDA can characterize the topological structure well, such that changes caused by adversarial perturbations are detectable. In this study, we study the granular topological effects of adversarial perturbations by tracking its evolution across all layers.

3 Topological Data Analysis (TDA)

In order to quantify the properties of a point cloud in high-dimensional space, we move beyond using summary statistics and employ TDA. TDA methods analyze the topology induced by the connectivity of points in a space across a variety of scales, allowing for more comprehensive and informative results (Wasserman, 2018).

We use *persistent homology* (Edelsbrunner et al., 2008), one of two key methods in TDA, to analyze both the local and global features of language model internal representations. In particular, given a set of points in a metric space, we perform a *Vietoris–Rips filtration* (Sheehy, 2012) across a parameter r; for each r, we construct a graph containing points as vertices and connecting vertices when they are within a distance of r from one another. Persistent homology analyzes the creation ("birth") and filling-in ("death") of topological features within this graph, such as connected components (H_0 features), loops (H_1 features), and voids (H_2 features) (Uchendu et al., 2024). These features are compiled into a *persistence diagram* (Cohen-Steiner et al., 2005), a plot which shows birth and death values (b_i, d_i) for each feature in

the filtration. *Persistence* describes how long each topological feature lasts in the filtration and is given by Persistence$((b_i, d_i)) = d_i - b_i$. This is the vertical distance from the feature to the $y = x$ diagonal on the diagram.

Persistent homology is a suitable tool for our task since attention maps are a natural way to quantify the distance between token representations. Unlike hidden states, the space induced by attention values is not Euclidean (Kobayashi et al., 2020; Vaswani et al., 2017); however, persistent homology does not require a Euclidean space (Edelsbrunner et al., 2008; Wasserman, 2018). In addition, persistence diagrams generated from persistent homology are robust to noise (Cohen-Steiner et al., 2005). This means metrics computed from persistence diagrams are also generally robust to noise and are a suitable tool for sensitivity analysis via input perturbation. Because persistence diagrams are stable under small perturbations of the underlying metric space, large Wasserstein deviations indicate structurally meaningful changes rather than numerical noise alone.

4 Problem Definition

4.1 RQ1: Topology of Adversarial Perturbations

In making small point perturbations to the input, we expect to see the appearance or disappearance of high-persistence 0-dimensional and 1-dimensional features rather than shifts in the overall distribution of persistence values. In other words, we expect text substitution to primarily influence the *tail behavior* of persistence values. We examine the tail behavior of persistence using a suite of graphical visualizations. To better evaluate the tail behavior of persistence values, we propose **persistence survival curves**. By plotting and overlaying the persistence percentiles on a logarithmic scale, we emphasize the difference in the weights of high persistence tails as a result of perturbation, rather than show the relatively diffuse and noisy behavior of the main mass of points.

4.2 RQ2: Layer Analysis with TDA

Different layers in an encoder model typically characterize different levels of the meaning of the input. Early layers typically encode local syntax features, while every subsequent layer encodes more global meaning. In this paradigm, the added classification head represents the highest-level meaning, i.e., the overall sentiment, of the input. With this in mind, we visualize the distribution of persistence values across layers 0, 6, and 11, representing early, middle, and late layers. Further, we investigate the sensitivity of the heads in each layer to adversarial perturbation.

4.3 RQ3: Generalizability of the Topology of the Embedding Space Across Models

We compare patterns across several encoder models to identify whether the observed topological behaviors and layer-wise sensitivities are consistent. This serves to validate results developed from individual model layer analysis.

5 Methodology

5.1 Datasets

The IMDb dataset consists of 50,000 long-form movie reviews (25,000 positive and 25,000 negative) collected from the IMDb website. Reviews contain an average of 231 words with standard deviation 171 words.

5.2 Selected Language Models

We evaluate the impact of TextFooler on three encoder-only models: BERT-base-uncased, RoBERTa-base, and DistilBERT. Each model was modified via a classification head consisting of a dropout and a linear layer during training, then fine-tuned to perform binary classification on the IMDb dataset. For consistency, we capped the number of input tokens to the first 128 tokens of each example. BERT and RoBERTa were trained to a level of greater than 90% test accuracy, and DistilBERT was trained to a level of greater than 80% test accuracy.

5.3 Adversarial Attack

Using TextFooler, Jin et al. (2020) propose a model-agnostic, score-based point substitution attack for text classification tasks. Suppose we have a binary classification task and have fine-tuned a language model f_θ to output the probability of class 1 on dataset $\mathcal{D} = \{(q_i, y_i)\}$, $y_i \in \{0, 1\}$. Given an input q_i with label y_i, a substitution attack perturbs the input by replacing a small number of words in q_i, resulting in a semantically similar input q_i'. The attack aims to achieve

$$\left|y_i - f_\theta(q_i')\right| > 0.5,$$

an incorrect prediction, in as few substitutions as possible.

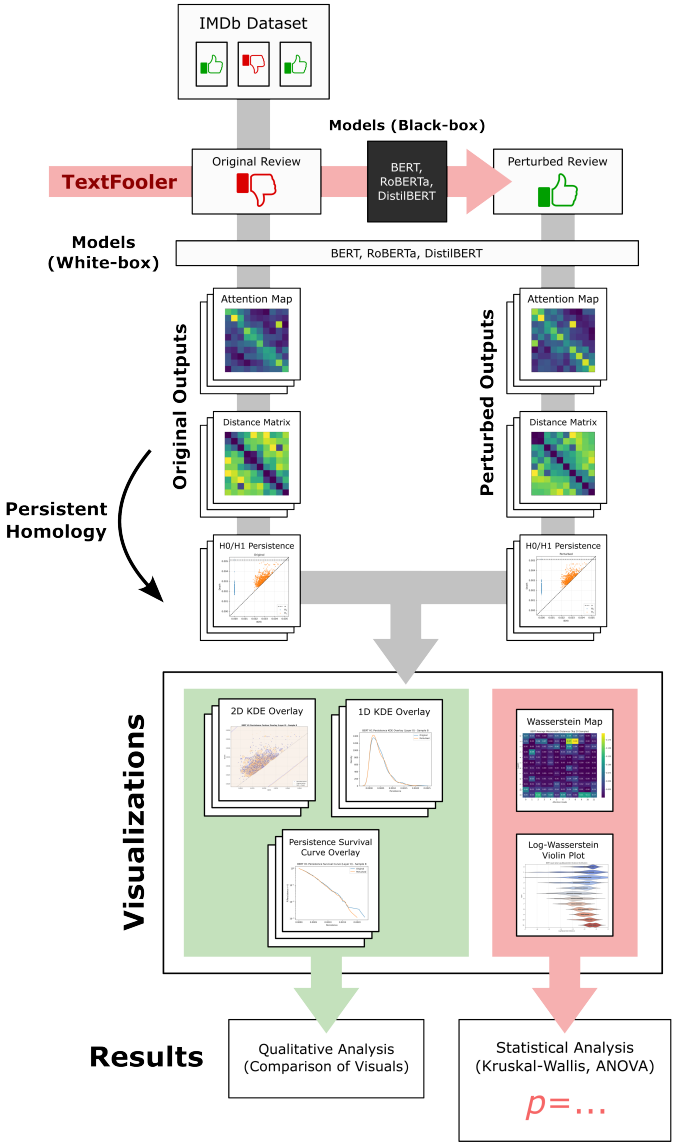

Figure 2: Overall pipeline for adversarial attack, visualization, and layer-wise analysis of attention maps. Recall that TextFooler is a black-box attack, but the subsequent step of extracting attention maps is white-box. Note that some outputs are "stacked," meaning there are multiple diagrams, one per layer or head.

6 Experimental Setup

6.1 Attack Pipeline

Given a model where only output logits are visible, the TextFooler attack first determines the "importance" of each word in the input q_i by masking word individually and observing the change in the logits of the output. Then, the attack will perform a search by iteratively replacing important tokens with similar words with some minimum cosine similarity threshold — effectively, replacing words with synonyms. For each model, In our input generation step, we store the original input q_i and the altered input q_i', and log the number of words that have been changed for each input. For each model, we use the same set of 1024 examples from the IMDb test split to ensure direct comparability. Finally, the attack results were sorted by highest to lowest "effectiveness" of the attack, which is measured by the absolute change in the prediction logits after an attack. Next, for each model, we stored all input sentence pairs and attack statistics in tabular form for analysis:

- **Original and Perturbed Texts:** In both texts, the affected words were wrapped in the special punctuation [[...]], which was also cleaned via regex prior to feeding into our models.
- **Original and Perturbed Scores:** *Score* mea-

sures the deviation of the logit corresponding to the correct label from a fully confident logit of 1.

- **Absolute Score Difference:** The loss in score after an attack. This is positive for all attack instances where the model initially was greater than 50% confident in the correct label but then switched its prediction.

An example original-perturbed input pair is provided below. The replaced words (in this case, just one) are in bold.

> *Original Input:* This version is very **painful** to watch. All of the acting is very stilted... (*Negative, 99.97% confidence*)
>
> *Perturbed Input:* This version is very **scathing** to watch. All of the acting is very stilted... (*Positive, 99.95% confidence*)

6.2 Attention Map Extraction

For each original-perturbed text pair, we performed inference by individually feeding each example into their corresponding models, and extracted attentions per head per layer. For BERT and RoBERTa, there were 12 layers and 12 attention heads per layer; for DistilBERT, there are 6 layers and 12 attention heads per layer. Thus, for each head, we have a *seq_len* $\times$ *seq_len* attention map (Vaswani et al., 2017).

We now formalize the sense of "distance" between a pair of tokens in the input sentence. For a sequence of tokens $(t_1, t_2, \ldots, t_k)$, each attention map is a $k \times k$ matrix A that encodes pairwise attention weights a_{ij}, with each a_{ij} being nonnegative and each row summing to 1. As discussed previously, this matrix representation naturally encodes a sense of distance between tokens, where higher attention weights a_{ij}, a_{ji} between a pair of tokens corresponds to a lower distance between the tokens in space. To achieve this, we create a distance matrix D containing the pairwise distances between tokens, which we define to be 1 minus the average attention between the tokens (Kushnareva et al., 2021). In other words, for each pair of tokens (t_i, t_j), we set

$$d_{ij} = 1 - \frac{1}{2}\left(a_{ij} + a_{ji}\right).$$

We also set $d_{ii} = 0$ for $i \in 1, 2, \ldots, k$. Our resulting distance matrix is then used to generate a Vietoris–Rips filtration and perform persistent homology.

6.3 Visualizations

In order to highlight different changes in persistence upon perturbation, we introduce several visualization modes of persistence:

1. **Persistence Diagram:** a plot of birth-death values (b_i, d_i) for H_0 and H_1 topological features (Cohen-Steiner et al., 2005). This is the standard mode of visualization for persistent homology.
2. **2D KDE Plot:** a contour plot directly on the H_1 persistence diagram. This smooths the discrete distribution, reduces noise, and is more visually intuitive in showing systematic trends in the overall distribution than comparing point clouds (Rosenblatt, 1956).
3. **1D KDE Plot:** a plot of H_1 persistence values as KDE plots, showing the distribution of mass in persistence values (Rosenblatt, 1956).
4. **Persistence Survival Curve:** a plot of the survival function versus the persistence, where the survival function is plotted on a logarithmic scale.
5. **Wasserstein Heatmap:** a heatmap of average Wasserstein metrics between original and perturbed H_1 persistence diagrams, showing variation across heads and layers (Scholkemper et al., 2024).
6. **Wasserstein Violin Plot:** a violin plot of the distribution of log-Wasserstein metrics compiled across examples per layer.

6.4 Statistical Tests

To assess which layers in an encoder model are more sensitive to adversarial perturbations, we compute the Wasserstein distance (Rüschendorf, 1985) between H_1 persistence diagrams for each attention head across the 20 sentence pairs corresponding to attacks that maximized output-logit changes. These distances were aggregated per head and per layer. As a non-topological baseline, we additionally computed raw L_1 distances between corresponding original and perturbed attention maps for each head and layer. These L_1 distances were aggregated identically to the Wasserstein metrics, allowing direct comparison between TDA-based and conventional attention-space sensitivity measurements. Unlike raw L_1 distances, which measure elementwise deviations in attention weights, persistent homology summarizes higher-order con-

nectivity structure across multiple scales. Thus, TDA-based Wasserstein distances are more sensitive to global structural changes induced by perturbations even when local attention differences remain small.

Since we did not assume any particular distribution of Wasserstein or L_1 distances across heads and examples, we first applied non-parametric tests. Specifically, we used a rank-based Kruskal–Wallis test (McKight and Najab, 2010) on Wasserstein distances and L_1 distances to evaluate whether layers differed in their distance metrics, followed by a post-hoc Dunn test with Holm correction for pairwise layer comparisons. These tests are robust to skewed distributions and do not require normality assumptions (McKight and Najab, 2010). In addition, the rank-based non-parametric statistics are unaffected by the logarithmic transformation because logarithms preserve ordering.

In parallel, we also performed parametric tests. Based on the log-Wasserstein violin plots (Hintze and Nelson, 1998), the log transformation greatly reduced right skew and produced distributions within each layer that approximated normality. This allowed us to perform a parametric one-way ANOVA (Quirk, 2021) to assess differences in mean log-Wasserstein distances between layers. By using both parametric and non-parametric tests, we maximize the explanatory power of our observations: the Kruskal–Wallis test provides a robust baseline that makes no assumptions about normality, while ANOVA offers greater power by leveraging the improved symmetry of the log-transformed data.

7 Results

7.1 Topology of Adversarially Perturbed Embedding Space

In our analysis of adversarial topologies, we observed that the vast majority of points in the H_1 persistence diagrams were concentrated in large clusters of points, making the original and perturbed diagrams difficult to differentiate via basic inspection or overall population statistics. However, in many of these diagrams, there were a few outliers with high birth and death values and persistence values (i.e., points up the diagonal and points farther away from the diagonal, respectively). For this type of deviation, we observe a marked improvement in qualitative visual analysis by using logarithm-scale survival curves over histograms,

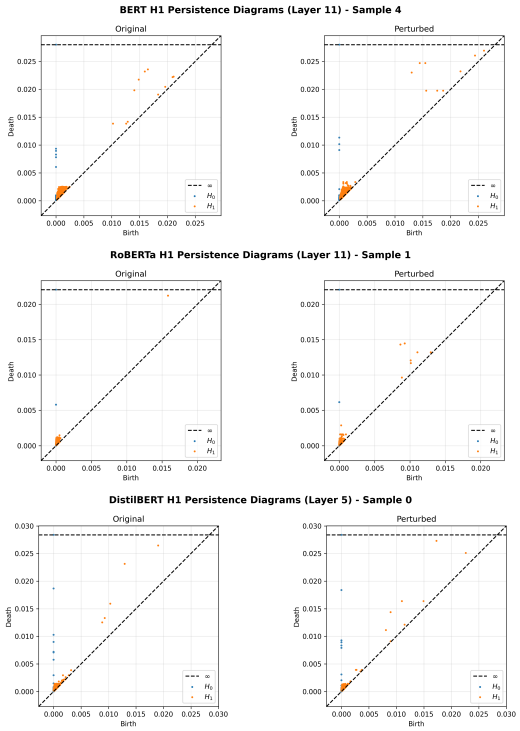

Figure 3: Persistence diagrams for a text pair across the three encoder models where "worst" was replaced with "gravest," averaging attentions across heads of the final layer prior to computing persistence. Low-death, low-persistence points tend to cluster in the bottom left and are less important than the scattered high-persistence points toward the top of each plot.

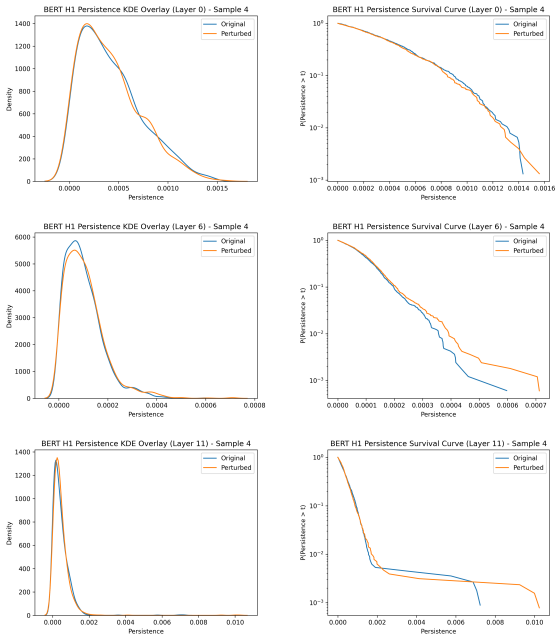

Figure 4: BERT KDE overlays (left) and survival curves (right) for layers 0, 6, and 11 for the substitution attack in Figure 3. The difference in mass in high-persistence regions before and after perturbation is more clearly visible in the survival curves.

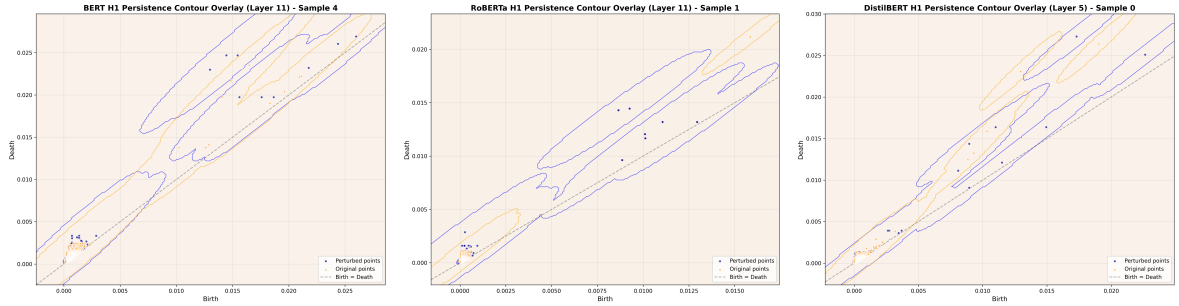

Figure 5: 2D KDE plots on the H_1 persistence diagrams in Figure 3. We see that the majority of the mass of persistence features in the H_1 diagram remain stationary, but there are shifts in high-persistence outlier points.

KDE overlays, and simple persistence diagrams. Survival curves more clearly show deviations in tail behavior compared to the raw persistence diagrams and KDEs (Figures 3, 4, and 5). Additional comparisons can be found in Appendix C.

Model	Method	H	p_{KW}	Dunn
BERT	TDA	902	2.3×10^{-186}	11/11
	Baseline	422	1.1×10^{-83}	10/11
RoBERTa	TDA	753	1.9×10^{-154}	10/11
	Baseline	430	2.3×10^{-85}	5/11
DistilBERT	TDA	494	1.7×10^{-104}	5/5
	Baseline	58.8	2.2×10^{-11}	5/5

Table 1: Non-parametric analysis of layer-wise differences for both TDA-based and baseline non-TDA distances. Entries report Kruskal–Wallis statistics and corresponding p-values. "Dunn" reports the number of significant pairwise differences between the final layer and all preceding layers under Dunn tests with Holm correction. TDA-based distances consistently produce stronger statistical separation across layers than baseline L_1 distances.

Model	Method	F	p_{ANOVA}	Tukey
BERT	TDA	51.9	7.6×10^{-105}	11/11
	Baseline	37.3	1.3×10^{-75}	10/11
RoBERTa	TDA	32.9	1.5×10^{-66}	9/11
	Baseline	44.7	1.3×10^{-90}	5/11
DistilBERT	TDA	50.2	5.2×10^{-48}	5/5
	Baseline	13.6	5.1×10^{-13}	5/5

Table 2: Parametric analysis of layer-wise differences using one-way ANOVA and Tukey's HSD tests for both TDA-based and baseline non-TDA distances. Entries report ANOVA statistics and corresponding p-values. "Tukey" reports the number of significant pairwise differences between the final layer and all preceding layers under Tukey's HSD tests. While both approaches detect significant layer-wise effects, TDA-based distances generally yield stronger and more discriminative layer separation, with the exception of RoBERTa.

7.2 Layer Sensitivity Analysis with TDA

We observe from the persistence survival curves (Figures 4; and in Appendix, 8, 9) that persistence diagrams (Figures 3) show more drastic differences in tail behavior as the layer number increases from $0 \to 6 \to 11$. The Wasserstein distances per head point to a similar phenomenon. The distance matrices in Figure 6 are average Wasserstein distances per head across the 20 examples with the highest deviation in output logits. These matrices show a substantial increase in Wasserstein distances in the final two layers for each model. In addition, some heads in layers 0, 1, and 2 also show large average Wasserstein distances as a result of perturbation. Unlike in late layers, this behavior is typically only shown in a small number of heads rather than across all heads (Figure 6).

We also plotted the distribution of log-Wasserstein distances per layer, aggregating the $12 \times 20 = 240$ observations per head per example for each layer. Even in log-space, there is a clear trend in log-Wasserstein distances, with median distances per layer roughly following a sideways "U" (or "W," in the case of RoBERTa) shape for each model (Figure 7). To test this hypothesis, we use both non-parametric and parametric tests, as outlined in Section 5. For BERT, the Kruskal–Wallis test suggests very strong differences between layers ($H = 902$, $p = 2.3 \times 10^{-186}$; see Table 1), and post-hoc Dunn tests show that the final layer differs significantly from all other layers. The baseline L_1 distances for the same set of examples show a much weaker trend statistically and visually, with distances weakly increasing over layers (Figure 12). While the majority of layer comparisons still yield significant results, the separation of effect sizes across layers is observed to be much stronger in the TDA analysis. Similar results are observed for RoBERTa and DistilBERT, although on a weaker

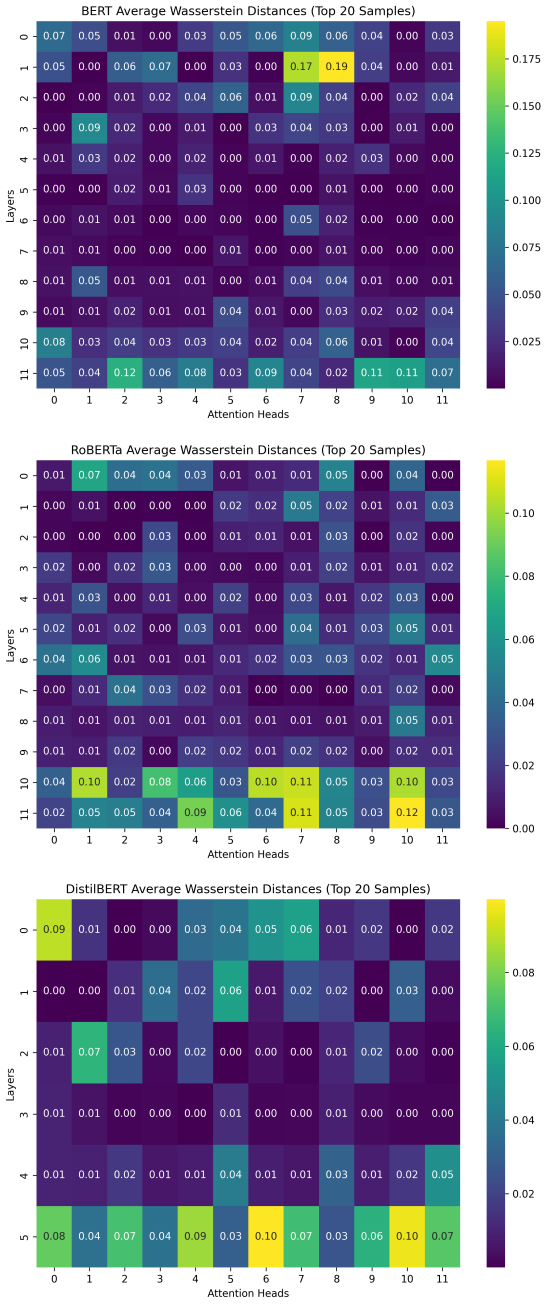

Figure 6: Wasserstein distance matrices showing distance values per head across the 20 examples with the highest deviation in logits. Notice the consistent high activations in the latest and earliest layers.

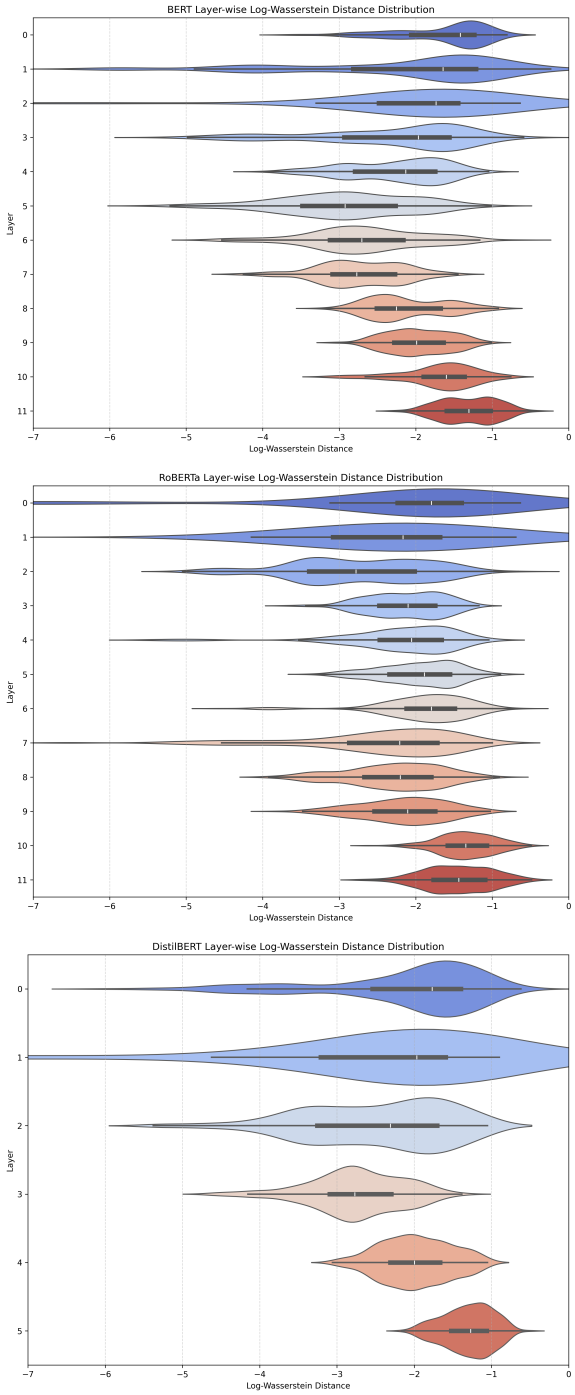

Figure 7: Per-layer log-Wasserstein distance violin plots with box plot overlay.

scale (Table 1).

Parametric analysis on the log-transformed Wasserstein distances using one-way ANOVA corroborates these findings, with the final layer differing significantly in mean log-Wasserstein distance across layers for all models (Table 2). Tukey's HSD post-hoc tests confirm the non-parametric results, though the discrepancy between RoBERTa's middle layers and late layers is less pronounced. A comprehensive table of pairwise p-values from each test is provided in Appendix D.

7.3 Generalization Across Models

Our violin plots show that, across all models, the very early and very late layers show the highest distances, and the middle layers show the lowest distances. Our statistical tests confirm this claim. For all models, the final layer exhibits significantly higher log-Wasserstein distances than nearly all

preceding layers, reflecting increased sensitivity to perturbations at the top of the network. However, RoBERTa differs from the other models in a notable way: its middle-layer attention maps show a pronounced increase in distances, suggesting a resurgence in sensitivity in middle layers that is largely absent in BERT and DistilBERT. In general, RoBERTa's average log distances are more evenly distributed across layers, although the final layer still maintains significantly higher distances than earlier layers. This pattern highlights the common trend of greater sensitivity in later layers, and a model-specific feature of RoBERTa, where perturbation effects are more distributed across the network.

8 Discussion

Adversarial Influence is Most Detectable in Early and Late Layers. The qualitative and quantitative findings in Section 7 can be explained in the context of model evolution over layers. At each layer, our encoder models consistently incorporate additional context that accumulate across each layer, resulting in the highest deviation in attention shape just before the decision layer. The large Wasserstein distances in a subset of heads in early layers is also intuitive: since low-level representations are close to individual token embeddings, single-word perturbations have a large impact on early layers' attention topology.

RoBERTa Exhibits Relatively Uniform Sensitivities Across Layers. This pattern is apparent in both baseline and TDA methods, showing fewer significant pairwise layer comparisons than other models in its family. This can be attributed to several properties of RoBERTa's design and pretraining. First, RoBERTa has more rigorous pretraining than BERT (Liu et al., 2019), using more data and more dynamic masking. These features prevent the model from overfitting on data and concentrating specific features in a few layers or heads. Second, RoBERTa was pretrained specifically for masked language modeling (MLM) (Liu et al., 2019), whereas BERT was pretrained for both MLM and next sentence prediction (NSP) (Devlin et al., 2019). MLM is much more relevant to our binary text classification task than NSP since sentence coherence does not produce contextual features. Therefore, since all heads are dedicated to learning contextual features, there is less specialization of attention heads, meaning heads are less sensitive to word-level perturbations.

TDA Captures Structural Rather Than Purely Local Changes. Although both the TDA-based Wasserstein distances and the non-TDA L_1 baseline detect statistically significant effects of adversarial perturbations, the TDA metrics consistently produce stronger and more discriminative layer-wise separation. In particular, the TDA-based analyses reveal clear U-shaped or W-shaped sensitivity trends across layers, whereas the baseline distances show weaker and less consistent patterns across different models (Figures 7 and 12; Tables 1 and 2). Because persistent homology summarizes connectivity structure across multiple scales, the resulting Wasserstein distances are more sensitive to these large-scale structural perturbations than raw matrix norms alone.

Practical Applications. Beyond characterizing layer-wise behavior, Wasserstein sensitivity analysis has practical use cases. For example, it can be used to better inform future studies of the weakest parts of an LLM. When implementing safeguards to prevent model overconfidence, developers of an LLM may reinforce residual connections from less sensitive layers to more sensitive layers to mitigate the overall impact of single-word substitutions. In addition, for text classification tasks using LLMs, Wasserstein sensitivity analysis can inform model confidence by aggregating raw logits and Wasserstein metrics as a result of adversarial perturbation. Given an input, if the average Wasserstein distance as a result of input word substitution is higher than usual, then the model may be over-relying on individual words in the input. Thus, it should decrease its confidence in its prediction for that example.

9 Conclusion

Our qualitative analysis reveals that the topology of the attention maps of each encoder model in this study changes on the global level as a result of adversarial influence, while local features stay roughly intact. Compared to conventional attention-space baseline distances, the TDA-based Wasserstein metrics produce clearer layer-wise separation and more discriminative sensitivity patterns across models. This is bolstered by our quantitative sensitivity analysis, which indicates that very early and late layers experience the greatest change in high-persistence features and show significantly higher Wasserstein distances than middle layers.

10 Limitations

One key limitation of our sensitivity analysis is that it is not directional: the Wasserstein metric only captures the magnitude of change between persistence diagrams, but does not indicate the direction or nature of the change. Consequently, while we can quantify how much the topological structure changes under perturbation, we cannot directly infer whether the change increases or decreases specific topological features. When examining directional trends in persistence, we did not find consistent patterns across examples, highlighting that magnitude alone may not capture the full dynamics of model sensitivity.

In addition, our per-layer Wasserstein analysis relies on a small subset of the data. Although we found statistically significant results, we only tested the 20 examples with the highest change in output logits. Although upon manual inspection, the vast majority of these examples preserved the semantics, the top 20 examples may not be representative of the average successful attack. Since Wasserstein distances are linear in the scale of the distance map, it is possible that some examples with "large" initial perturbations would have disproportionately large Wasserstein distances, which would disrupt parametric statistical tests, which rely on normality assumptions. To mitigate this effect, we suggest three potential solutions:

- **Larger Sample Size:** We can increase the number of samples to include in our analysis. This would allow for more robust and powerful statistical tests.
- **Random Sampling with Bootstrapping:** To get more representative distributions and 95% confidence intervals for mean log-Wasserstein distances, we could use bootstrapping in conjunction with random sampling of text examples.
- **Distance Matrix Normalization:** Prior to computing persistence, we could normalize the distances to decrease the chance of single examples that dominate the Wasserstein distance analysis. This could introduce bias in our Wasserstein distance calculations, especially if we do not scale distance matrices belonging to the same sentence pair identically.

We also acknowledge that Wasserstein distance calculations are generally slow; the fastest known algorithms for computing Wasserstein distance, such as the Hungarian algorithm, have time complexity scaling with f^3, where f is the number of H_1 features, which itself can be up to quadratic in the number of tokens seq_len. This can become infeasible for larger inputs. In future studies, we propose using the bottleneck distance as a metric instead, as it is slightly faster computationally and isolates the effect of high-persistence values changing due to adversarial perturbation.

11 Ethical Statement

There are potential dangers to extending any conclusions found on these particular models and datasets to more general use cases. As the LLM ecosystem evolves, there are no guarantees that newer, state-of-the-art models will have the same vulnerabilities. In addition, any findings presented in this paper may not generalize to text classification tasks on other datasets.

All datasets and models, including pretrained model weights, are open-source and can be found on Hugging Face.

Acknowledgments

We acknowledge the Harvey Mudd College Computer Science and Math Clinic program and the MIT Lincoln Laboratory for facilitating this research project and providing computing power for our experiments. We also acknowledge Dr. Adaku Uchendu for her invaluable expertise and guidance throughout the project.

References

Kenneth Alperin, Rohan Leekha, Adaku Uchendu, Trang Nguyen, Srilakshmi Medarametla, Carlos Levya Capote, Seth Aycock, and Charlie Dagli. 2025. Masks and mimicry: Strategic obfuscation and impersonation attacks on authorship verification. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 102–116, Albuquerque, USA. Association for Computational Linguistics.

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of nli models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–444.

Jatin Chauhan and Manohar Kaul. 2022. Bertops: Studying bert representations under a topological lens. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, and 1 others. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World wide web*, 27(4):42.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. 2005. Stability of persistence diagrams. In *Proceedings of the twenty-first annual symposium on Computational geometry*, pages 263–271.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Herbert Edelsbrunner, John Harer, and 1 others. 2008. Persistent homology-a survey. *Contemporary mathematics*, 453(26):257–282.

Aideen Fay, Inés García-Redondo, Qiquan Wang, Haim Dubossarsky, and Anthea Monod. 2026. The shape of adversarial influence: Characterizing LLM latent spaces with persistent homology. In *The Fourteenth International Conference on Learning Representations*.

Neil Fendley, Edward W. Staley, Joshua Carney, William Redman, Marie Chau, and Nathan Drenkow. 2025. A systematic review of poisoning attacks against large language models. *Preprint*, arXiv:2506.06518.

Jerry L. Hintze and Ray D. Nelson. 1998. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075.

Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021. Artificial text detection via examining the topology of attention maps. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 635–649.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Patrick E McKight and Julius Najab. 2010. Kruskal-wallis test. In *The Corsini Encyclopedia of Psychology*, pages 1–10. Wiley Online Library.

Ilan Perez and Raphael Reinauer. 2022. The topological bert: Transforming attention into topology for natural language processing. *arXiv preprint arXiv:2206.15195*.

Fanchao Qi, Yuan Yao, Sophia Xu, Zhiyuan Liu, and Maosong Sun. 2021. Turn the combination lock: Learnable textual backdoor attacks via word substitution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4873–4883.

Thomas J. Quirk. 2021. One-way analysis of variance (anova). In *Excel 2019 for Social Science Statistics: A Guide to Solving Practical Problems*, pages 167–184, Cham. Springer International Publishing.

Murray Rosenblatt. 1956. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837.

Ludger Rüschendorf. 1985. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129.

Michael Scholkemper, Damin Kühn, Gerion Nabbefeld, Simon Musall, Björn Kampa, and Michael T. Schaub. 2024. A wasserstein graph distance based on distributions of probabilistic node embeddings. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9751–9755.

Donald R Sheehy. 2012. Linear-size approximations to the vietoris-rips filtration. In *Proceedings of the twenty-eighth annual symposium on Computational geometry*, pages 239–248.

Adaku Uchendu and Thai Le. 2024. Unveiling topological structures from language: A survey of topological data analysis applications in nlp. *arXiv preprint arXiv:2411.10298*.

Adaku Uchendu, Thai Le, and Dongwon Lee. 2024. Topformer: Topology-aware authorship attribution of deepfake texts with diverse writing styles. In *27th European Conference on Artificial Intelligence, ECAI 2024*, pages 1446–1454. IOS Press BV.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Minh Vu, Geigh Zollicoffer, Huy Mai, Ben Nebgen, Boian Alexandrov, and Manish Bhattarai. 2025. Topological signatures of adversaries in multimodal alignments. In *Forty-second International Conference on Machine Learning*.

Larry Wasserman. 2018. Topological data analysis. *Annual review of statistics and its application*, 5(2018):501–532.

Xiaoyu Zhang, Cen Zhang, Tianlin Li, Yihao Huang, Xiaojun Jia, Ming Hu, Jie Zhang, Yang Liu, Shiqing Ma, and Chao Shen. 2025. Jailguard: A universal detection framework for prompt-based attacks on llm systems. *ACM Trans. Softw. Eng. Methodol.*, 35(1).

.1 Code Availability

Our implementation for adversarial attack generation, attention map extraction, topological data analysis, and statistical evaluation is publicly available at: `https://github.com/angelinatsai04/mitll_clinic/tree/adam_spring`.

A Reasoning for Selection of Visualization Methods

Generally, the KDE plots are suited for showing trends in overall movement in the mean persistence of the distribution, that is, any consistent shifts in large masses of points. This would correspond to topological effects such as the destruction of topologies on small or large scales, or overall topological compression. The survival curve is generally more suitable for isolating global topological features that persist across scales.

The KDE plots and survival curves were plotted as overlays to facilitate direct visual comparison. For the persistence diagram, we ensured that plots of original and perturbed persistence were the same scale per example to also facilitate direct comparison.

Computing Wasserstein distances on H_1 persistence diagrams is preferred to computing norms on raw attention maps because persistent homology has guarantees about numerical stability and robustness to noise. In practice, we observed that the resulting Wasserstein distances were strongly right-skewed across heads and layers. To improve interpretability in both visualization and statistical analysis, we applied a logarithmic transformation to all Wasserstein distance observations.

B Top TextFooler Perturbations

Some of the successful attacks for each model that resulted in the highest decrease in correct prediction probability are reproduced in Table 3. Based on these samples, all of which are negative-turned-positive, negative reviews appear to be much more vulnerable to single-word attacks than positive reviews. The models seem to misclassify phrases containing words like “egregious” and “gravest,” which typically have negative connotations but can also be found in a positive review; these words are easily influenced by contextual cues. In addition, the review that appeared in every model’s top three examples was especially tricky; every model suffered from the same single substitution from “worst” to “gravest.” Given the remainder of the review, which is unchanged in the attack, it is clear that the review is negative, but this small change signals to the model that the series is mostly *hilarious* and not bad.

C Additional Visualizations for RoBERTa and DistilBERT

Figures 8 and 9 show the comparison between 1D KDE plots and their corresponding survival curves. DistilBERT has a less clear discrepancy between original and perturbed survival curves for this example. It is possible that, as a simpler model, DistilBERT only recognizes the word substitution on a local semantic level, explaining the early-layer differences, but this effect does not propagate as the context widens. We also computed persistences after averaging all heads in the layer, meaning any effect localized in few heads would be diluted in this visualization.

D Per-layer p-value Matrices

The heatmaps in Figures 10 and 11 show p-values for both parametric and non-parametric tests performed on log-Wasserstein distances, and are more detailed summaries of the results from Tables 1 and 2.

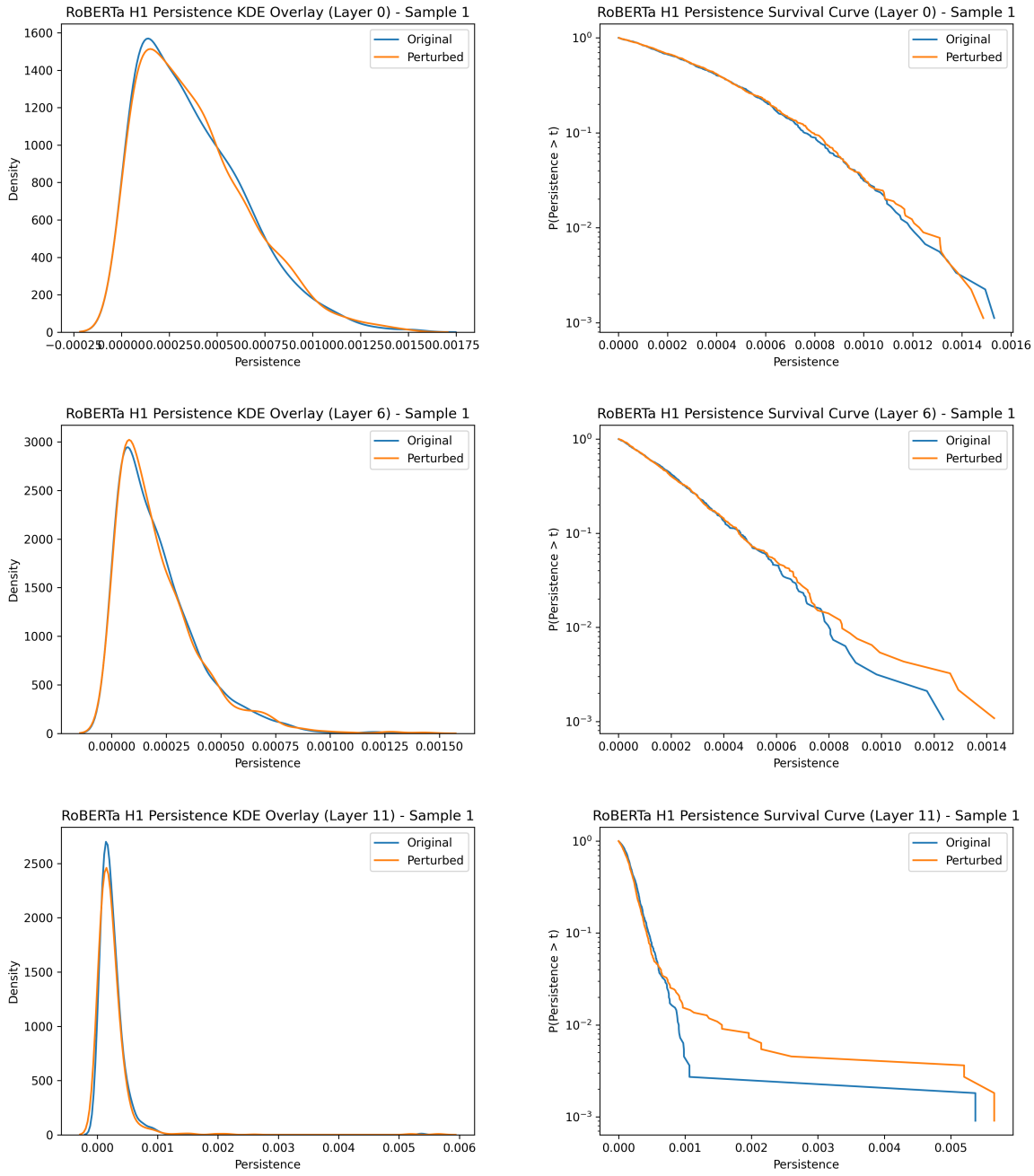

Figure 8: RoBERTa KDE overlays (left) and survival curves (right) for layers 0, 6, and 11 for the substitution attack in Figure 4.

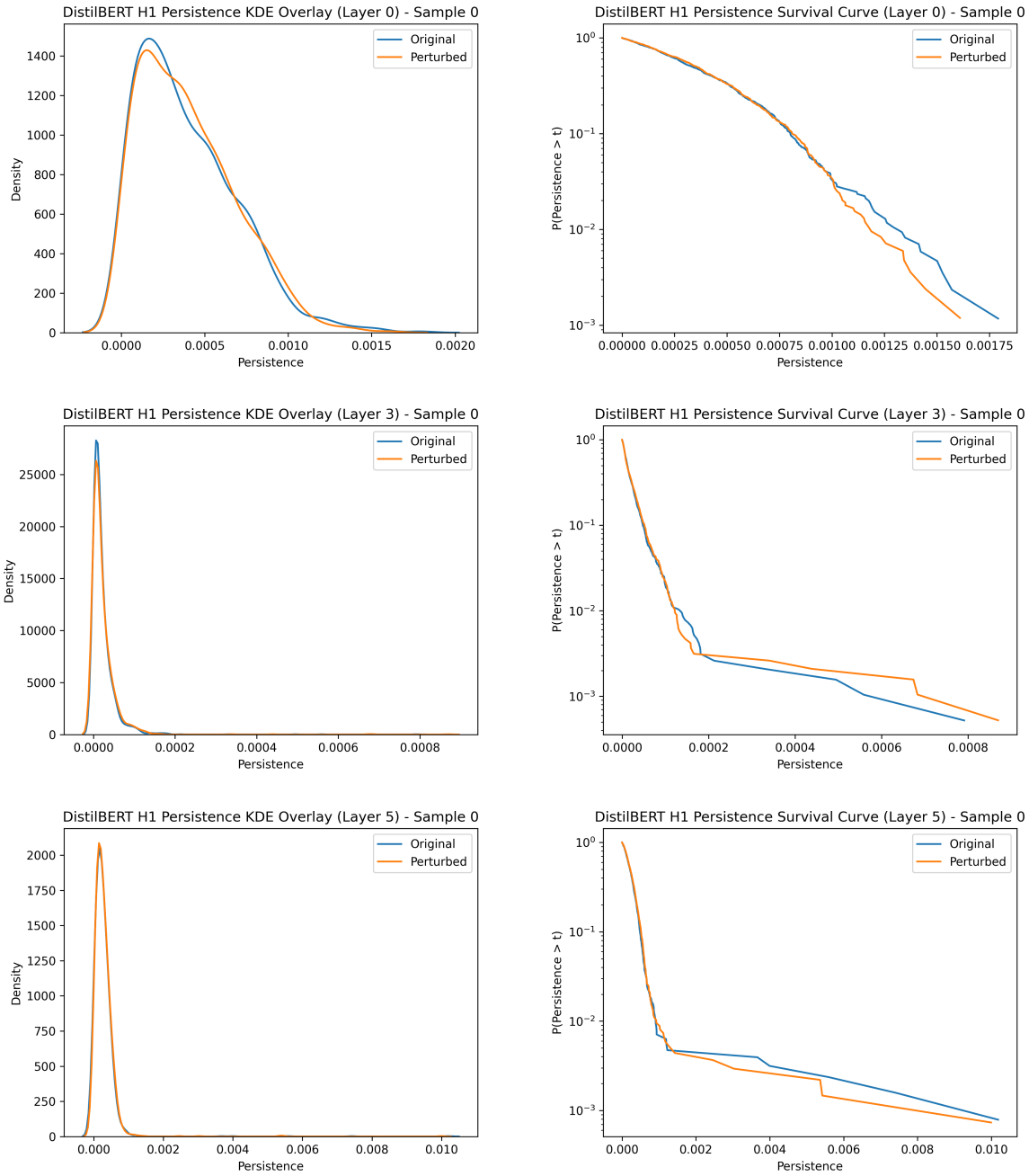

Figure 9: DistilBERT KDE overlays (left) and survival curves (right) for layers 0, 3, and 5 for the substitution attack in Figure 4.

BERT	
while Hillary Swank is great for the role, the plot to the movie is just **dreadful**. *Negative (99.94%)*	while Hillary Swank is great for the role, the plot to the movie is just **egregious**. *Positive (99.94%)*
The film is **annoying**. *Negative (99.90%)*	The film is **troubling**. *Positive (99.93%)*
[Y]ou'll agree with me that this is one of the **worst** and yet hilarious series ever made. *Negative (99.84%)*	[Y]ou'll agree with me that this is one of the **gravest** and yet hilarious series ever made. *Positive (99.94%)*
RoBERTa	
[Y]ou'll agree with me that this is one of the **worst** and yet hilarious series ever made. *Negative (99.27%)*	[Y]ou'll agree with me that this is one of the **gravest** and yet hilarious series ever made. *Positive (97.40%)*
This is a truly **awful** film. *Negative (99.83%)*	This is a truly **spooky** film. *Positive (96.51%)*
Don't **see** this movie! It's... **repulsive**! *Negative (99.74%)*	Don't **presume** this movie! It's... **unsavory**! *Positive (96.01%)*
DistilBERT	
[Y]ou'll agree with me that this is one of the **worst** and yet hilarious series ever made. *Negative (98.79%)*	[Y]ou'll agree with me that this is one of the **gravest** and yet hilarious series ever made. *Positive (98.61%)*
This is fairly typical for the Sci-Fi Channel: one-dimensional characters, a ridiculous plot, and **terrible** special **effects**. *Negative (99.40%)*	This is fairly typical for the Sci-Fi Channel: one-dimensional characters, a ridiculous plot, and **tragic** special **impact**. *Positive (96.90%)*
I really did not want to write a harsh review of this movie... However this movie is truly **awful**. *Negative (99.45%)*	I really did not want to write a harsh review of this movie... However this movie is truly **egregious**. *Positive (95.55%)*

Table 3: Adversarial examples showing original (left) and perturbed (right) text substitution pairs, producing prediction flips from negative to positive with high confidence across models. Replaced text is in bold.

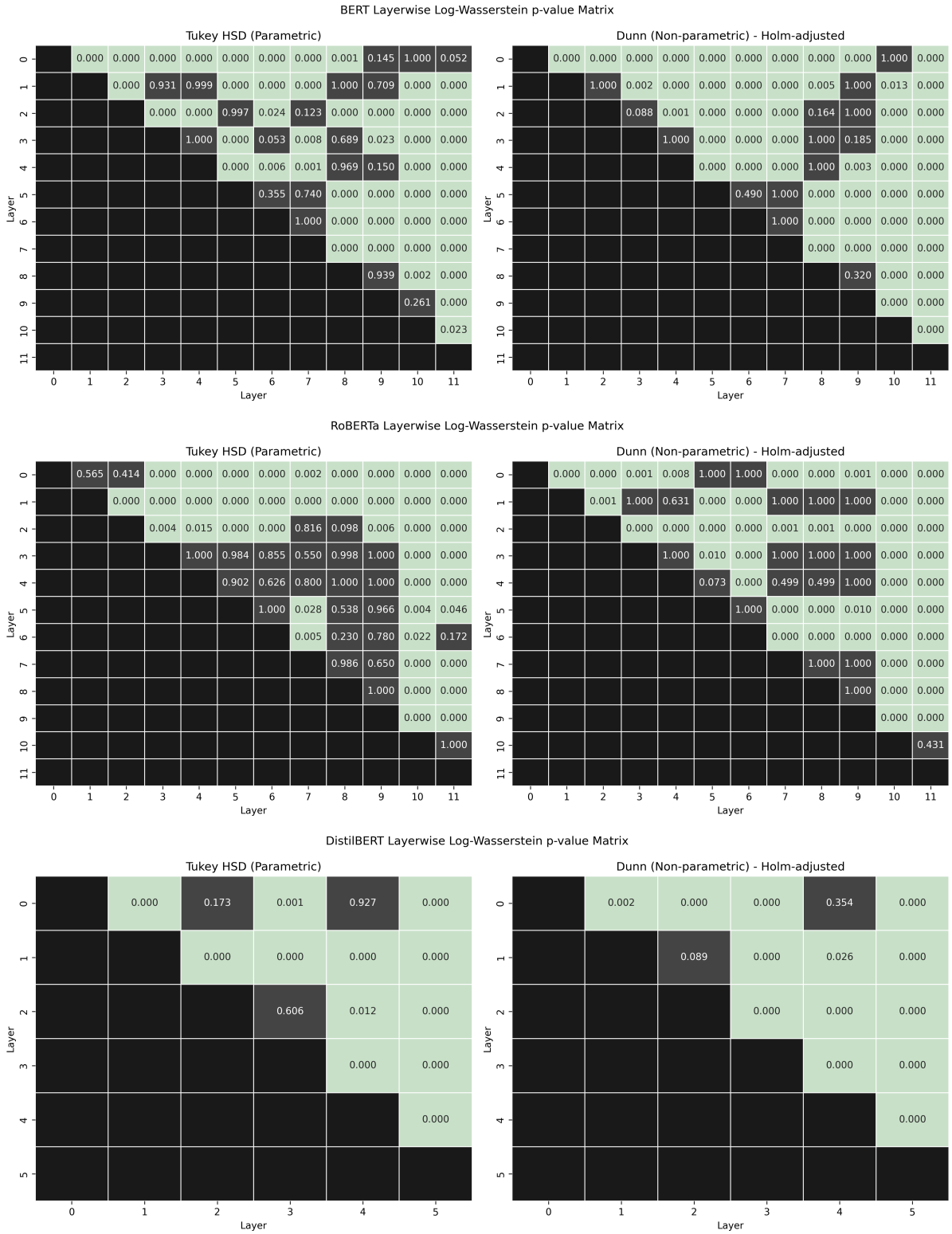

Figure 10: Layer-wise Dunn's (left) and Tukey's HSD (right) test p-value heatmaps of log-Wasserstein distances, with significant layer pairings in green. Note that the final layer in all models has significantly different log-Wasserstein distances from almost all other layers.

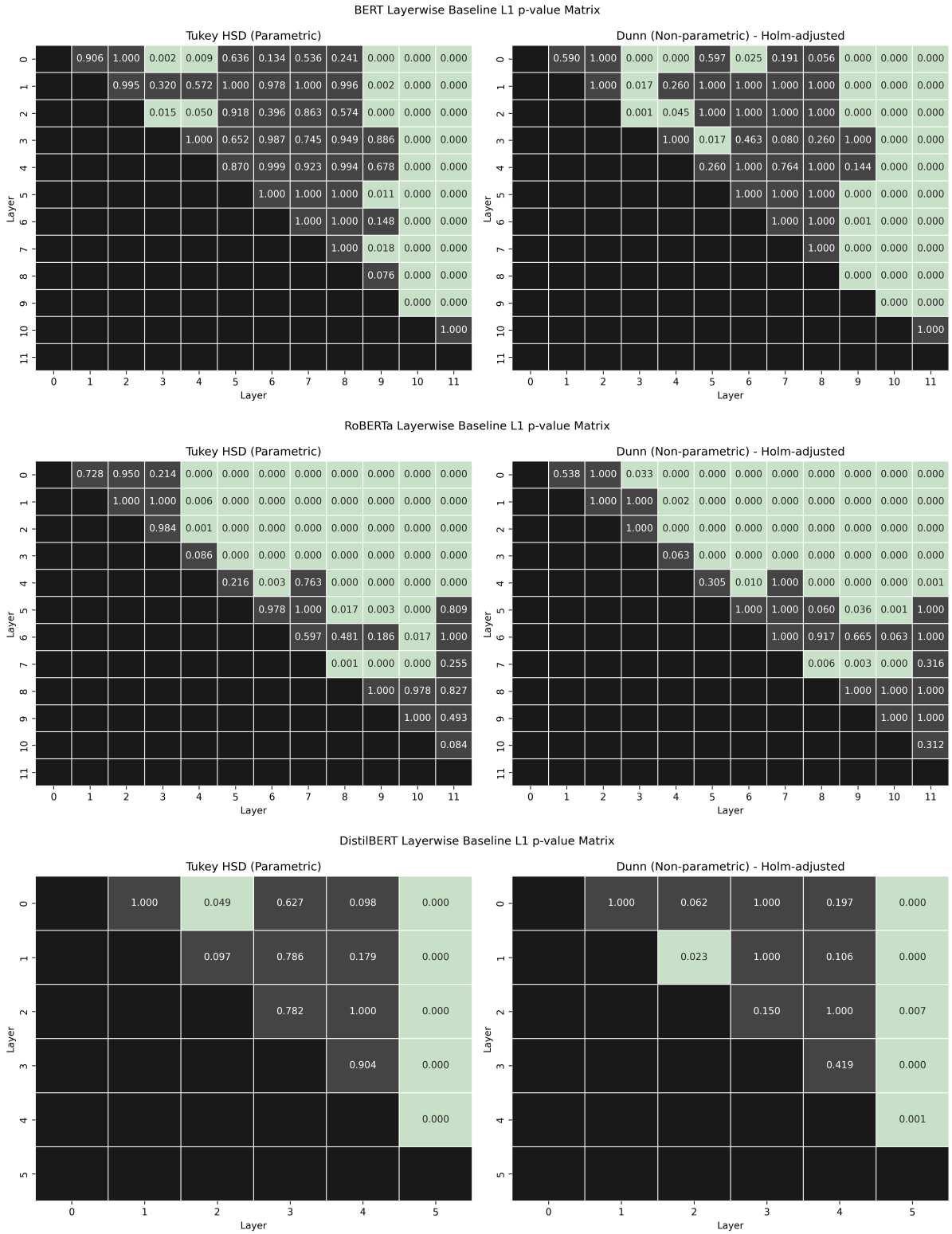

Figure 11: Layer-wise Dunn's (left) and Tukey's HSD (right) test p-value heatmaps of non-TDA baseline L_1 distances, with significant layer pairings in green. The U-shaped trend observed in our TDA results is less clear and varies much more between models.

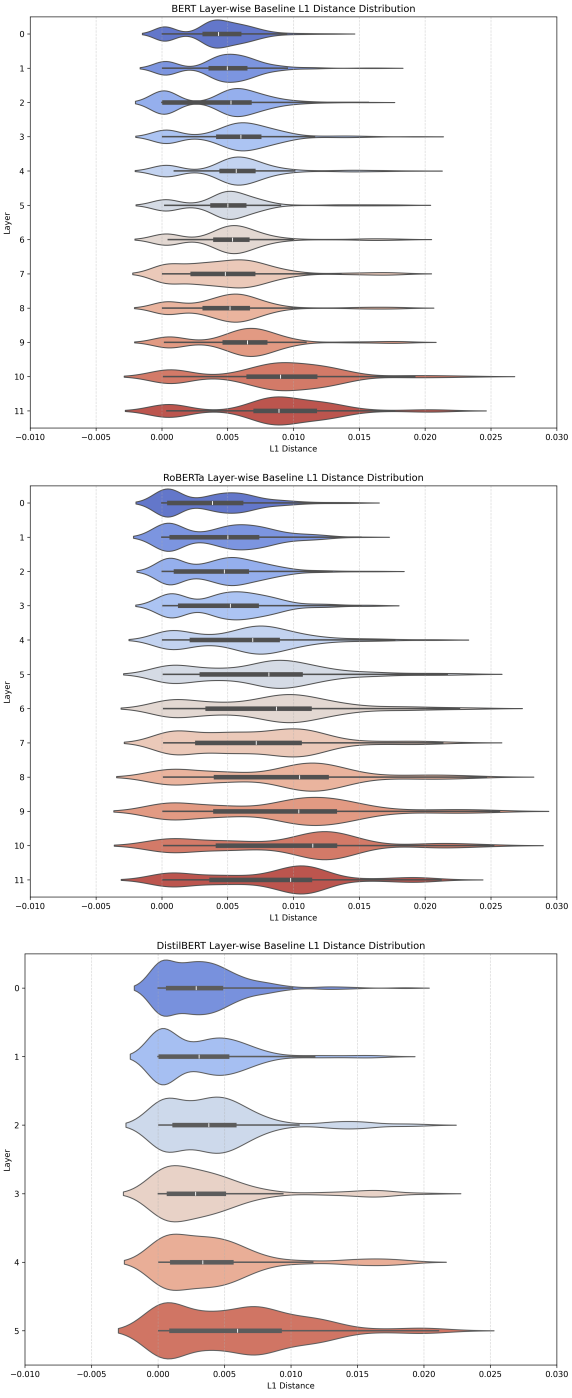

Figure 12: Per-layer Non-TDA L_1 baseline distance violin plots with box plot overlay.