BEYOND ACCURACY: STABLE PREDICTIONS, UNSTABLE TOPOLOGY IN TRANSFORMER REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictive accuracy remains the dominant metric for evaluating robustness in modern natural language processing (NLP) systems, often suggesting strong resilience even under adversarial or corrupted inputs. We argue that this view can obscure failure modes that emerge in the induced local topological structure of learned representations. Focusing on BERT-family classifiers spanning a range of model capacities, we study the stability of local structure induced by k-nearest neighbors (kNN) under semantically-preserving transformations, including adversarial obfuscations. Despite near-perfect classification performance across all conditions, we find that kNN neighborhood structure shows substantial changes, where local relationships between embeddings are rewired even when predictions and semantic content remain unchanged. We further introduce a normalization-based sensitivity measure that relates topological change to embedding perturbation magnitude and show that neighborhood topology changes disproportionately relative to embedding displacement across all conditions, with cosine displacement remaining near-zero while Rank Stability increases monotonically with perturbation strength. This effect is consistent across neighborhood scales and occurs without loss of label consistency, indicating that predictive decisions remain stable while representation structure varies. We quantify these effects using topology-based stability metrics and observe large, statistically significant differences between clean and corrupted inputs. Overall, we find a systematic mismatch between predictive performance and the stability of the induced neighborhood topology, suggesting that accuracy alone is insufficient to characterize robustness and that topology-based measures provide a complementary diagnostic for hidden vulnerabilities.

1 INTRODUCTION

Neural language models are typically evaluated through predictive performance on held-out data, and strong accuracy is commonly interpreted as a proxy for robustness, despite evidence that high accuracy can mask systematic behavioral failures (Ribeiro et al., 2020a; Dutta et al., 2024). This evaluation paradigm primarily reflects output-level consistency and provides limited insight into the structure of the underlying representation space (Rogers et al., 2020; Lenc & Vedaldi, 2015). This raises the question of whether stable predictions necessarily imply stable internal representations (Mashaido & Das, 2026), particularly under input transformations that preserve semantic content but alter surface form.

Neural language models operate in embedding spaces where semantic relationships are expressed through the induced local neighborhood topology over representations (Ruppik et al., 2024). Standard evaluation practices implicitly assume that this local structure is stable under semantically-preserving perturbations. However, this assumption is rarely tested directly, as robustness is typically assessed through label agreement rather than representation structure (Papernot & McDaniel, 2018).

In this work, we investigate representation stability through local neighborhood structure using k-nearest neighbor (kNN) graphs. This provides a non-parametric view of representation topology,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

enabling measurement of how local connectivity patterns change under controlled perturbations without additional modeling assumptions.

We evaluate BERT-based classifiers under semantically-preserving transformations, including structured adversarial obfuscations. Despite near-perfect predictive performance across all conditions, we observe substantial changes in local neighborhood structure, where kNN relationships are systematically altered even when predicted labels remain unchanged.

To quantify this effect, we introduce Rank Stability (RS), which measures preservation of local neighborhood topology under perturbation, and a normalization-based sensitivity measure that relates topological change to embedding perturbation magnitude. We further define a composite Latent Robustness Gap (LRG) combining neighborhood label consistency and topology drift. We evaluate these measures using statistical testing, multi-scale neighborhood analysis over $k \in \{5, 10, 20, 50\}$, and topology-based structural metrics.

Our results show a consistent decoupling between predictive stability and representation topology: predictions remain invariant while local neighborhood structure varies significantly and monotonically across perturbation types. Cosine displacement between original and perturbed embeddings remains near-zero across all conditions, yet Rank Stability increases substantially with perturbation strength, indicating that topological disruption is disproportionate to representational shift. This suggests that predictive accuracy alone is insufficient to characterize robustness, and that topology-based metrics provide a complementary lens for evaluating representation-level stability.

2 Related Work

Robustness in neural language models. Robustness in NLP has primarily been studied through input perturbations, adversarial attacks, and distribution shift evaluations. Prior work focuses on output stability, measuring whether predictions remain consistent under noise or adversarial modifications (Ribeiro et al., 2020b). While effective for benchmarking, these approaches largely operate at the level of prediction invariance and do not explicitly analyze changes in internal representation structure (Mashaido & Das, 2026; Dutta et al., 2024).

Representation geometry in neural networks. Several studies have explored the geometry of learned representations in transformer-based models, including analyses of anisotropy (Ethayarajh, 2019), clustering behavior (Coenen et al., 2019), and semantic organization in embedding spaces (Yenicelik et al., 2020). These works suggest that transformer embeddings encode structured semantic manifolds (Valeriani et al., 2023), but typically evaluate structure using global similarity measures or task-specific probes rather than local neighborhood stability (Kornblith et al., 2019; Tenney et al., 2019). We confirm this gap directly: linear CKA (Kornblith et al., 2019) between clean and obfuscated embeddings in our setting is 0.972, indicating near-total global similarity, despite RS $= 0.784$ under the same comparison (Table 3) – a concrete demonstration that global similarity measures miss the local neighborhood rewiring RS is designed to detect.

Graph-based and neighborhood-based analyses. Graph-based and nearest-neighbor methods have been widely used to analyze and structure representation spaces (Plötz & Roth, 2018) in vision and language models. These approaches show that local neighborhoods often reflect semantic similarity and can vary with model architecture and training dynamics. However, most prior work treats kNN structure as a descriptive or diagnostic artifact (Papernot & McDaniel, 2018), without systematically analyzing how local neighborhood topology changes under semantically-preserving perturbations or adversarial obfuscations.

Topology-inspired perspectives in representation learning. Recent work has introduced tools from topology and geometry, such as manifold analysis and persistent homology, to study neural representations (Valeriani et al., 2023; Barannikov et al., 2022; Ballester et al., 2026). These methods, often grounded in foundational computational topology (Edelsbrunner & Harer, 2010), emphasize global structural properties of embedding spaces, but are typically not designed to capture fine-grained changes in local neighborhood relationships under controlled input transformations.

In contrast to prior topology-inspired approaches that emphasize global structural properties, we analyze local neighborhood stability under controlled perturbations. Critically, by conditioning topological change on embedding displacement magnitude under cosine distance, we show that

neighborhood structure rewires substantially even when representational displacement is negligible, a sensitivity that global measures and accuracy-based evaluation both fail to capture.

3 METHODOLOGY

Throughout, we use 'topology' to refer to the local neighborhood structure induced by k-nearest neighbor graphs under cosine similarity, rather than formal topological invariants.

3.1 PROBLEM SETUP

Let f_θ be a fine-tuned BERT-based classifier mapping input text $x \in \mathcal{X}$ to a label space $\mathcal{Y}$. Let $h_\theta(x) \in \mathbb{R}^d$ denote the encoder representation of x, corresponding to the final hidden state of the [CLS] token.

We study the structure of the representation space $\mathcal{Z} = \{h_\theta(x)\}$ induced by clean and semantically-preserving corrupted inputs.

Our central question is:

> Does stable predictive behavior imply stable local topology in representation space?

3.2 REPRESENTATION GRAPH CONSTRUCTION

We define a discrete representation graph induced by k-nearest neighbors under cosine similarity, and analyze its stability under perturbations.

Let $G_k = (V, E_k)$ over embeddings $h_\theta(x)$, where each node corresponds to an input sample and edges connect each node to its k nearest neighbors under cosine similarity:

$$E_k(i) = \text{kNN}_k(h_\theta(x_i))$$

This graph is used as a proxy for the induced kNN neighborhood topology in representation space; we emphasize that this is not a global manifold assumption, but a local structural probe.

3.3 NEIGHBORHOOD LABEL CONFLICT (NLC)

To quantify local class consistency in representation space, we define Neighborhood Label Conflict as:

$$\text{NLC}(x_i) = \frac{1}{k} \sum_{j \in \text{kNN}_k(i)} \mathbb{1}[y_j \neq y_i]$$

NLC measures how often local neighborhoods violate label consistency. Lower values indicate more label-coherent local structure. In practice, NLC serves as a null baseline against which topology drift can be isolated.

Note: NLC is evaluated only on correctly classified samples to isolate representation structure from prediction errors.

3.4 SEMANTIC-PRESERVING PERTURBATIONS

We define a transformation $\mathcal{T}$ that maps an input x to a normalized variant $\tilde{x}$:

$$\tilde{x} = \mathcal{T}(x)$$

where $\mathcal{T}$ removes or standardizes: Unicode homoglyphs, zero-width characters, emoji noise, and non-semantic punctuation artifacts.

This allows us to probe whether representation topology is invariant under surface-level corruption that preserves semantic content.

3.5 RANK STABILITY (RS)

We define Rank Stability as a measure of *sensitivity of local neighborhood topology* under input perturbation.

Let $R_k(x)$ denote the kNN set of x in embedding space, and $R_k(\tilde{x})$ the corresponding set after a semantically-preserving transformation $\mathcal{T}$.

$$\mathrm{RS}(x) = 1 - \frac{|R_k(x) \cap R_k(\tilde{x})|}{k}$$

RS is computed over a fixed embedding space and cosine similarity metric; it should be interpreted as a measure of perturbation-induced changes in local neighborhood topology rather than a property of the kNN algorithm itself.

To disentangle intrinsic geometric sensitivity from trivial embedding displacement, we normalize RS by perturbation magnitude:

$$\hat{RS}(x) = \frac{RS(x)}{\delta(x) + \epsilon} \tag{1}$$

where $\delta(x) = 1 - \frac{h_\theta(x) \cdot h_\theta(\tilde{x})}{\|h_\theta(x)\| \|h_\theta(\tilde{x})\|}$ is the cosine distance between original and perturbed embeddings and ϵ is a small constant for numerical stability. This measure captures the degree to which topology changes disproportionately relative to embedding displacement. Under cosine distance, displacement remains near-zero across all conditions, yet RS increases substantially with perturbation strength, confirming that topological disruption is not a consequence of representational shift but an intrinsic property of the learned embedding space.

Unlike static neighborhood metrics, RS explicitly measures how local structure responds to a controlled transformation $\mathcal{T}$, allowing it to be interpreted as a direct sensitivity measure.

Higher RS indicates greater instability of neighborhood topology under perturbation, i.e., greater sensitivity of local neighborhood topology to surface-form variation.

3.6 LATENT ROBUSTNESS GAP (LRG)

We define a composite instability score:

$$\mathrm{LRG}(x) = \alpha \cdot \mathrm{NLC}(x) + (1 - \alpha) \cdot \mathrm{RS}(x)$$

This metric combines label-local inconsistency and topology drift into a unified scalar measure of representation instability. In settings where NLC is non-negligible, LRG captures both sources of instability jointly. In the present experiments, where NLC is negligible across all conditions, LRG reduces to a linear rescaling of RS, with instability driven entirely by neighborhood drift.

3.7 CORRECT CLASSIFICATION FILTERING

To ensure analysis focuses on representation dynamics rather than prediction errors, we define a filtered dataset:

$$\mathcal{D}^* = \{x \in \mathcal{D} \mid f_\theta(x) = y\}$$

All topological metrics are computed on $\mathcal{D}^*$, isolating latent geometric instability from classification failure.

3.8 Scale Sensitivity Analysis

We evaluate all metrics across multiple neighborhood scales $k \in \{5, 10, 20, 50\}$ to assess whether observed instability is sensitive to local graph resolution.

4 Experiments and Results

4.1 Overview

We evaluate whether stable predictive performance implies stable representation topology under semantically-preserving and adversarially-obfuscated inputs. Our pipeline consists of: (i) controlled dataset construction, (ii) fine-tuning a BERT-based classifier, (iii) filtering correctly classified samples, and (iv) analyzing embedding topology using kNN graphs.

4.2 Dataset and Perturbations

We construct a dataset of 40,000 samples across four conditions: clean, prefix corruption, suffix corruption, and multi-layer obfuscation. Perturbations include Unicode homoglyph substitutions, zero-width characters, emoji insertion, and randomized punctuation, preserving semantic meaning while altering surface form.

4.3 Model and Predictive Performance

The BERT-based classifier achieves:

- Test accuracy: **0.99375**
- Balanced performance across all classes

This near-perfect performance across all conditions, including heavily obfuscated inputs, establishes the baseline against which representation-level instability is measured.

4.4 Topological Instability Under Perturbation

Table 3 reports mean RS across conditions, visualized in Figure 1. RS increases monotonically with perturbation strength, from RS $= 0.049$ for clean inputs to RS $= 0.784$ for obfuscated inputs, with suffix and prefix perturbations forming intermediate bands at RS $= 0.226$ and RS $= 0.283$ respectively. These differences are further reflected in the full distributional structure of RS across conditions (Figure 3), which shows minimal overlap between clean and corrupted regimes.

4.5 Structure of Topological Variation

We analyze the joint distribution of Rank Stability (RS) and Neighborhood Label Conflict (NLC). All samples exhibit NLC ≈ 0, producing a vertically structured distribution in RS–NLC space where variation occurs entirely along the RS axis as visualized in Figure 2.

Obfuscated inputs occupy the highest RS region, prefix and suffix perturbations form intermediate bands, and clean inputs cluster near zero.

As Neighborhood Label Conflict is negligible across all conditions, LRG reduces to a linear rescaling of RS, with the instability signal driven entirely by neighborhood drift rather than label inconsistency. This is itself informative: it isolates topology disruption as a phenomenon independent of semantic structure degradation, indicating that the model preserves coarse semantic clustering while local topology is being significantly rewired.

4.6 Statistical Significance

We quantify differences between clean and corrupted conditions:

- Welch's t-test: $t = -141.81$, $p < 10^{-16}$

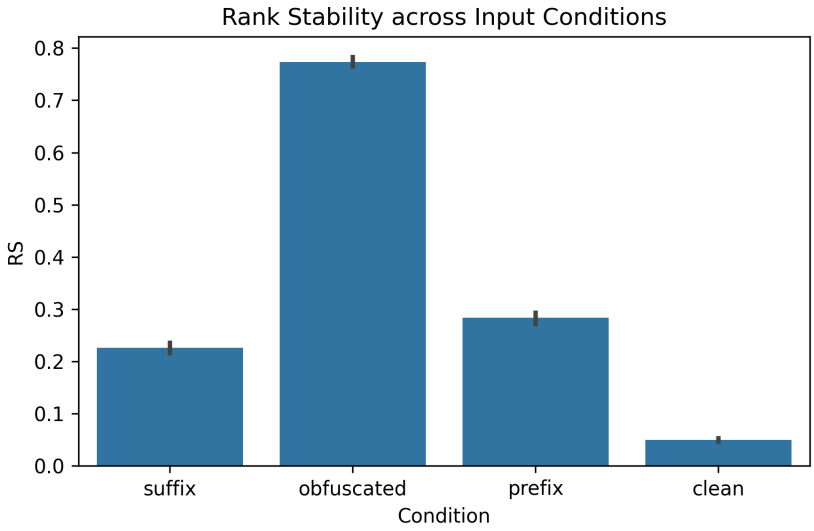

Figure 1: Rank Stability (RS) across input conditions. RS increases monotonically with perturbation strength, demonstrating that neighborhood topology becomes less stable even under transformations that preserve both semantic content and classification accuracy (99.375%).

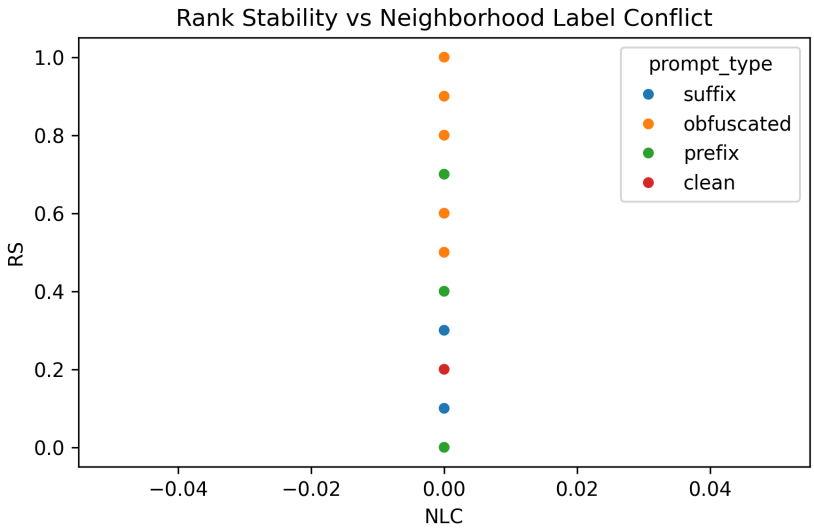

Figure 2: RS vs NLC across samples. Despite near-zero Neighborhood Label Conflict across all conditions, Rank Stability varies widely, indicating that local neighborhood relationships change substantially independently of label-consistent structure. This isolates representation topology changes from semantic clustering effects.

- Cohen's $d = -4.55$
- 95% CI: [0.3569, 0.3668]

The magnitude of Cohen's $d = -4.55$ indicates an extremely large effect size, confirming that the observed topology divergence between clean and obfuscated conditions is not a marginal or noise-driven phenomenon but a systematic structural property of the learned representations.

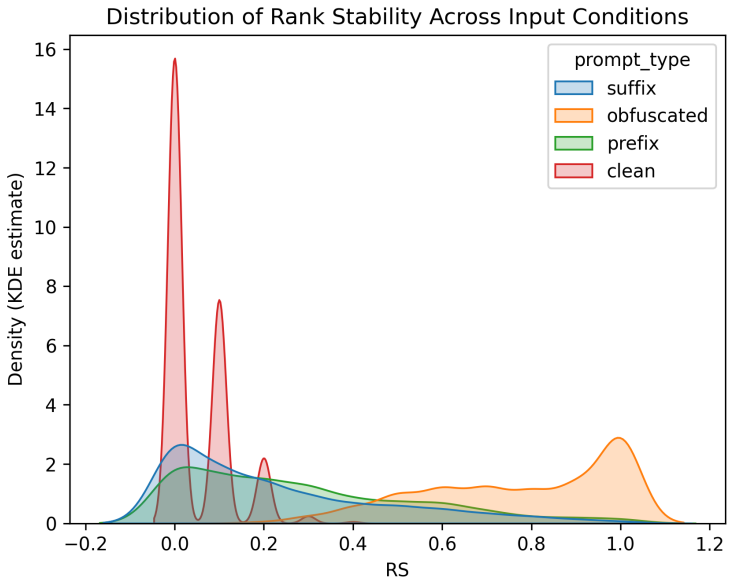

Figure 3: Distribution of Rank Stability across input conditions estimated via kernel density. Clean inputs concentrate near zero, while corrupted inputs shift toward higher RS values with minimal overlap, showing clear separation in representation instability despite identical predictive performance.

Table 1: Sensitivity over k. RS remains stable across scales, confirming that topological instability is not an artifact of neighborhood resolution.

k	RS
5	0.3353
10	0.3305
20	0.3240
50	0.3158

These results indicate extremely large and statistically significant differences in representation topology.

4.7 Perturbation Magnitude and Normalized Topology Sensitivity

Table 3 reports mean cosine displacement δ and normalized Rank Stability $\hat{RS}$ across conditions. Cosine displacement remains near-zero across all conditions, yet RS increases monotonically and substantially with perturbation strength. Normalized RS is monotone across corrupted conditions, confirming that topological disruption scales with perturbation strength independently of representational shift. These results support the central claim that even minimal semantic-preserving changes cause large topological shifts, and that neighborhood topology reflects an intrinsic sensitivity of the learned representation space rather than a consequence of embedding displacement.

4.8 k Sensitivity Analysis

We evaluate RS across neighborhood sizes $k \in \{5, 10, 20, 50\}$ and demonstrate that topological instability is scale-invariant (Table 1 and Figure 4):

The persistence of topological instability across all values of k suggests that the observed neighborhood rewiring is not an artifact of graph resolution but reflects an intrinsic property of the learned

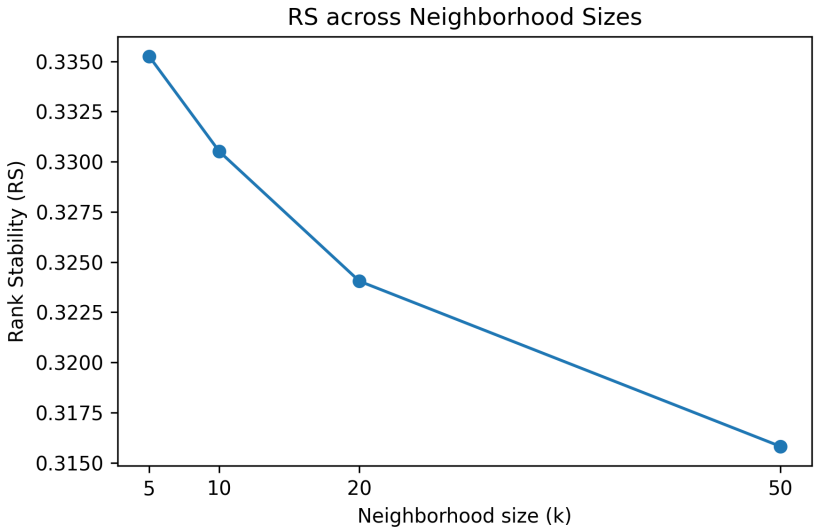

Figure 4: Rank Stability across neighborhood sizes ($k = 5$ to 50). RS remains consistently high across all values of k, indicating that observed topological instability is not an artifact of neighborhood size but a persistent property of the representation space.

Table 2: Rank Stability (RS) across model architectures under the obfuscated condition, showing the accuracy–topology decoupling persists across a range of model capacities.

Model	Accuracy	RS (clean)	RS (obfuscated)	Cohen's d
DistilBERT	0.9942	0.050	0.760	-4.36
BERT-Medium	0.9934	0.047	0.776	-4.72
BERT-Base	0.9938	0.049	0.784	-4.55

embedding space under surface-form variation, consistent with the near-zero cosine displacement observed across all conditions.

4.9 Cross-Architecture Generalization

To assess whether the observed accuracy–topology decoupling is specific to a single architecture, we repeated the evaluation pipeline on two additional fine-tuned classifiers spanning a substantial capacity range: DistilBERT and a reduced-capacity BERT variant (BERT-Medium), trained identically to our primary BERT-base model. All three architectures achieve near-identical test accuracy (99.34–99.42%) while exhibiting comparably large Rank Stability under the obfuscated condition (RS $= 0.760$–0.784, Cohen's d between -4.36 and -4.72; Table 2). The consistency of this effect across a range of model capacities indicates that the accuracy–topology decoupling is not an artifact of a single architecture, and that our findings extend beyond the primary BERT-base setting studied in the remainder of this paper.

4.10 Alignment with Contributions

Together, these results support the paper's core contributions: kNN-based topology analysis reveals structural changes invisible to accuracy-based evaluation; Rank Stability quantifies this change directly and increases monotonically with perturbation strength; normalized RS shows the instability is disproportionate to embedding displacement rather than a consequence of it; the near-zero NLC across all conditions isolates topology drift, rather than semantic incoherence, as the driver of instability; and the effect is stable across neighborhood scales k and, as shown above, across model architectures spanning a substantial capacity range.

Table 3: Aggregate instability metrics across input conditions, computed over correctly classified samples. The clean condition exhibits near-zero displacement by construction, as $\mathcal{T}$ removes no artifacts from unperturbed text. The informative comparison is across corrupted conditions, where $\hat{RS}$ is monotone and displacement is meaningful.

Condition	RS	LRG	δ	$\hat{RS}$
Clean	0.049	0.025	4.7×10^{-8}	3.47×10^{6}
Suffix	0.226	0.113	1.7×10^{-3}	1.21×10^{6}
Prefix	0.283	0.142	2.1×10^{-3}	1.81×10^{6}
Obfuscated	0.784	0.387	5.5×10^{-2}	8.55×10^{6}

5 Discussion and Limitations

5.1 Decoupling Between Prediction and Representation

Our findings indicate a structured, non-random sensitivity of representation topology to semantically-preserving perturbations, intrinsic to the embedding space rather than a consequence of representational shift. Models may therefore preserve decision boundaries while exhibiting high sensitivity in their internal neighborhood structure – a distinction accuracy-based evaluation cannot detect, but which Rank Stability captures directly. This rewiring is systematic rather than noise, and persists across perturbation types, neighborhood scales, and, as Section 4.9 shows, model architecture.

5.2 Implications for Robustness Evaluation

Predictive invariance does not imply topological invariance. Near-zero NLC across all conditions indicates that semantic clustering is preserved even as fine-grained topology is significantly disrupted. Models maintain coarse class structure while local neighborhood relationships are substantially rewired. **This dissociation between semantic coherence and topological stability is the core finding of this work, and it suggests that accuracy provides an incomplete characterization of model stability:** topology-based metrics such as RS expose a structured sensitivity that is invisible to output-level evaluation, offering a complementary diagnostic for representation-level robustness

5.3 Limitations

This study is scoped to kNN graphs under cosine similarity and a synthetic perturbation setting, enabling precise, controlled measurement of the phenomenon. Section 4.9 further shows it holds across a substantial capacity range within the BERT family. Open directions include decoder-based and generative architectures, real-world adversarial inputs, and alternative topological measures such as persistent homology.

6 Conclusion

We studied the stability of representation topology in BERT-family classifiers under semantically-preserving input perturbations. Despite near-perfect predictive accuracy, local neighborhood structure changes substantially and systematically, with RS increasing monotonically across all perturbation types, neighborhood scales, and, model architectures spanning a substantial capacity range – while cosine displacement remains near-zero throughout, confirming the disruption is intrinsic to the embedding space rather than a consequence of representational shift.

Together with near-zero NLC across all conditions, this reveals a dissociation between coarse class structure and local neighborhood stability, reinforcing that accuracy alone cannot characterize representation-level robustness. Future work should extend this analysis to generative architectures and natural distribution shifts, and explore topology-preserving training objectives.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to perform a literature search and to help edit and reformat this manuscript's LaTeX source and prose during preparation, and to review the paper for consistency with the ICLR 2027 formatting and reproducibility requirements. We have not used generative AI tools for other tasks such as generating datasets and experiments; and the rest of the required disclosure tasks are not applicable to this work. All AI-assisted edits to text and formatting were reviewed by the authors, who take full responsibility for the final content, claims, and code described in this paper.

REPRODUCIBILITY STATEMENT

Full experimental details are provided in Appendix B (model architecture, training hyperparameters, random seeds, and compute resources), Appendix A (synthetic dataset construction procedure), and Appendix C (formal definitions of all metrics: NLC, RS, $\hat{RS}$, and LRG). Appendix D details the neighborhood-scale sensitivity analysis, and Appendix E details the statistical testing procedure used to obtain the results in Section 4.6. All random seeds are fixed to 42 across Python, NumPy, and PyTorch to support exact replication of the reported results.

REFERENCES

Rubén Ballester, Carles Casacuberta, and Sergio Escalera. *Topological Data Analysis for Neural Networks*. Springer, 2026. doi: 10.1007/978-3-032-08283-1.

Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. Visualizing and measuring the geometry of bert. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (NAACL-HLT)*, 2019.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Herbert Edelsbrunner and John L. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010. ISBN 978-0-8218-4925-5.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Becky Mashaido and Tapadhir Das. On the geometric limits of transformer defenses against obfuscation attacks: Latent embedding collapse & performance–robustness gap. In *Proceedings of the IEEE International Conference on AI and Data Analytics (ICAD)*, 2026. Accepted for publication.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. In *Advances in Neural Information Processing Systems*, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020a.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020b.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. In *Transactions of the Association for Computational Linguistics*, 2020.

Benjamin Matthias Ruppik, Michael Heck, Carel van Niekerk, Renato Vukovic, Hsien-Chin Lin, Shutong Feng, Marcus Zibrowius, and Milica Gašić. Local topology measures of contextual language model latent spaces with applications to dialogue term extraction. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2024.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*, 2023.

David Yenicelik, Florian Schmidt, and Yannic Kilcher. How does bert capture semantics? a closer look at polysemous words. In *Proceedings of the Third ACL BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2020.

A Appendix

Appendix A: Synthetic Dataset Construction

We construct a controlled synthetic dataset with four classes: *clean*, *prefix*, *suffix*, and *obfuscated*, each containing 10,000 samples.

Clean samples are generated using templated natural language sentences with randomized nouns, verbs, adjectives, adverbs, and contextual phrases.

Adversarial samples are created using multi-layered surface perturbations while preserving semantic meaning. These include:

- Unicode homoglyph substitutions (e.g., Latin-to-Cyrillic character swaps)
- Zero-width character insertion (`\u200b`, `\u200d`, `\u2060`)
- Emoji injection
- Random punctuation insertion
- Partial and fragmented instruction prefixes/suffixes
- Mid-string injected adversarial fragments

This design induces surface-level distribution shift without altering underlying semantic content, enabling controlled robustness analysis.

APPENDIX B: MODEL, TRAINING DETAILS, COMPUTE RESOURCES AND REPRESENTATION EXTRACTION

We fine-tune a `bert-base-uncased` transformer using HuggingFace Transformers for 4-class sequence classification. Training hyperparameters are as follows: learning rate 3×10^{-5}, batch size 16, maximum epochs 10 with early stopping patience 3, weight decay 0.01, warmup steps 200, and maximum sequence length 128. The dataset is split 80/20 into train and test sets using stratified sampling with random seed 42. All random seeds (Python `random`, NumPy, and PyTorch) are fixed to 42 for reproducibility.

All topological analyses are performed using the encoder representation of the fine-tuned model. Specifically, we extract the final-layer [CLS] token embedding:

$$h_x = \text{BERT}_{\text{encoder}}(x)_{[\text{CLS}]}$$

This ensures alignment between classification behavior and representation topology.

Experiments were conducted on a single NVIDIA T4 GPU (16GB) via Google Colab. Each fine-tuning session took approximately 5 hours including embedding extraction and kNN computation. The full research project required additional compute beyond the reported run, including preliminary experiments with different perturbation configurations, hyperparameter tuning, and multiple replication runs to verify result stability. Total compute across all experimental runs is estimated at approximately over 30 GPU hours on the same hardware.

APPENDIX C: METRICS DEFINITION

We define three metrics over the learned embedding space.

Neighborhood Label Conflict (NLC):

$$NLC(x_i) = \frac{1}{k} \sum_{j \in \mathcal{N}_k(x_i)} \mathbb{1}[y_j \neq y_i]$$

Rank Stability (RS): Let $\mathcal{N}_k(x)$ denote the k-nearest neighbors of x under cosine similarity in embedding space. Let $\tilde{x} = \mathcal{T}(x)$ denote a semantic-preserving transformation of x. Then:

$$RS(x) = 1 - \frac{|\mathcal{N}_k(x) \cap \mathcal{N}_k(\tilde{x})|}{k}$$

Note that RS is a discrete neighborhood overlap measure and should be interpreted as a local stability proxy rather than a formal topological invariant.

Normalized Rank Stability ($\hat{RS}$):

$$\hat{RS}(x) = \frac{RS(x)}{\delta(x) + \epsilon}$$

where $\delta(x) = 1 - \frac{h_\theta(x) \cdot h_\theta(\tilde{x})}{\|h_\theta(x)\| \|h_\theta(\tilde{x})\|}$ is the cosine distance between original and perturbed embeddings and ϵ is a small constant for numerical stability.

Latent Robustness Gap (LRG):

$$LRG = \alpha \cdot NLC + (1 - \alpha) \cdot RS, \quad \alpha = 0.5$$

All metrics are computed on the filtered dataset $\mathcal{D}^* = \{x \in \mathcal{D} \mid f_\theta(x) = y\}$, consistent with the main text.

APPENDIX D: K-SENSITIVITY ANALYSIS

We evaluate robustness of NLC and RS across neighborhood sizes $k \in \{5, 10, 20, 50\}$.

Results remain stable across all values of k, indicating that the observed representation instability is not sensitive to neighborhood scale selection. This suggests the effect reflects intrinsic topological structure rather than a graph construction artifact.

APPENDIX E: STATISTICAL TESTING

We perform statistical comparisons between clean and obfuscated conditions on the Latent Robustness Gap (LRG) using:

- Welch's t-test (no equal variance assumption)
- Cohen's d effect size
- Bootstrap confidence intervals (10,000 resamples, 95% CI)

All analyses are conducted only on correctly classified samples $\mathcal{D}^*$ to isolate representation-level effects from classification errors. Results are reported in Section 4.6.

APPENDIX F: ASSET LICENSES AND CREDITS

This work uses `bert-base-uncased`, introduced by Devlin et al. (2019) and publicly available via the HuggingFace Model Hub. The HuggingFace Transformers library is released under the Apache 2.0 license. All other libraries used (scikit-learn, PyTorch, scipy, numpy) are standard open-source packages available under BSD or Apache 2.0 licenses. No proprietary datasets or third-party scraped data are used. The synthetic dataset introduced in this work is constructed procedurally as described in Appendix A.