Persistent Homology for Distribution Drift Detection in LLM Embedding Streams

Lei Yang [1]

Abstract

Deployed language models experience distribution drift in their input streams, degrading performance silently. Existing drift detectors rely on low-order statistics—means, covariances, or kernel embeddings—and may underperform when drift alters the geometric structure of the embedding space without large changes in these quantities. We present an initial systematic evaluation of persistent homology for drift detection in LLM embedding streams. We design three centroid-controlled drift scenarios—subtopic reweighting (approximate centroid preservation), style perturbation, and geometric reorganization (exact centroid preservation)—that reduce the advantage of conventional detectors, and evaluate 8 topological features against 6 classical baselines across 2 datasets, 2 embedding models, and 5 random seeds. Our key finding is that Wasserstein distance on H_0 persistence diagrams achieves AUC = 0.858 on geometric reorganization drift, significantly outperforming the best classical baseline (MMD-RBF, AUC = 0.782; paired permutation test $p < 0.001$). However, this advantage is specific to the controlled geometric reorganization scenario: classical methods remain superior on subtopic reweighting drift, and all methods show only moderate power on style perturbation. We provide comprehensive ablations over window size, subsample size, and PCA dimensionality, and show that topological methods run at 37–45 ms per window under the tested settings (PCA to 50 dims, 100-point windows). Our results provide initial evidence for when topological drift detection may add complementary value and identify important directions for strengthening this evidence, including comparison with stronger learned-kernel baselines and evaluation on architecturally diverse encoders.

[1]Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Lei Yang <lyang0176@gmail.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

1. Introduction

Large language models are increasingly deployed in production systems where the input distribution evolves over time. A news classification model trained on 2023 articles encounters novel topics in 2024; a customer-intent classifier sees shifting user behavior as products change. When these shifts go undetected, model performance degrades silently, causing downstream failures that are costly to diagnose (Khaki et al., 2023). Drift detection—automatically flagging when the input distribution has changed—is therefore a critical component of any reliable ML deployment pipeline.

Current drift detectors operate on embedding representations and measure divergence using low-order statistics. Centroid shift tracks the mean embedding; Fréchet drift distance (Greco et al., 2024) captures the first two moments under a Gaussian assumption; maximum mean discrepancy (MMD) (Kalinke et al., 2022) compares kernel mean embeddings. These methods work well when drift introduces new classes or shifts the embedding centroid, but they may struggle with drifts that reorganize the *geometry* of the embedding space while preserving or approximately preserving low-order statistics. Consider a customer-support classifier whose input distribution shifts from one subtopic mix to another within the same class, or a model whose inputs are paraphrased into a different style without changing the topic. Such drifts can alter the topological structure of the point cloud—the arrangement of clusters, the presence of loops, the connectivity pattern—without necessarily moving the mean or substantially changing the covariance.

Persistent homology (Edelsbrunner & Harer, 2010) provides a principled mathematical framework for capturing exactly this kind of structural information. By tracking connected components (H_0), loops (H_1), and higher-dimensional voids across a filtration of the point cloud, persistence diagrams encode multi-scale geometric features that complement moment-based statistics. Despite the natural fit, persistent homology has seen limited application

to drift detection in embedding streams, and no prior work has provided a controlled evaluation across multiple drift types and embedding models.

We address this gap with an initial comparative study. We design three centroid-controlled drift scenarios—subtopic reweighting (approximate preservation), style perturbation (exact), and geometric reorganization (exact)—alongside four standard drift types, and evaluate 8 topological features against 6 classical baselines across 2 text classification datasets and 2 embedding models. Our main finding is that Wasserstein distance on H_0 persistence diagrams achieves AUC = 0.858 on the controlled geometric reorganization scenario, outperforming the best classical baseline (MMD-RBF, AUC = 0.782) by 7.6% (paired permutation test, $p < 0.001$). Importantly, this advantage is specific to this controlled scenario: classical methods remain superior on subtopic reweighting, and all methods achieve only moderate detection power on style perturbation.

Our contributions are:

- An initial systematic evaluation of persistent homology for distribution drift detection in LLM embedding streams, comparing 8 topological features against 6 classical baselines.
- Three centroid-controlled drift scenarios that reduce the advantage of conventional detectors. Wasserstein distance on H_0 diagrams achieves a statistically significant advantage on the controlled geometric reorganization scenario (AUC 0.858 vs. 0.782, paired permutation $p < 0.001$).
- Ablations over window size, subsample size, and PCA dimensionality, showing that topological methods run at 37–45 ms per window under the tested settings.
- An honest characterization of the scenarios where topological methods add complementary value and where classical methods suffice, along with concrete directions for future work.

2. Related Work

Distribution drift detection. Unsupervised drift detection in deployed ML systems has received growing attention. Greco et al. (2024) propose DRIFTLENS, which models embeddings from a classifier's penultimate layer as multivariate Gaussians and computes Fréchet drift distance between reference and incoming windows. Their method outperforms MMD, KS, and LSDD baselines on AG NEWS and 20 NEWSGROUPS, but relies on the first two moments and assumes Gaussian-distributed embeddings—an assumption that fails for multi-modal point clouds. Khaki et al. (2023) deploy an MMD-based sliding-window pipeline with bootstrap calibration on BERT embeddings of production text data, demonstrating high correlation between MMD scores and downstream accuracy degradation. However, their controlled experiments test only class-imbalance drift, leaving geometry-altering drifts unexamined. Kalinke et al. (2022) introduce MMDEW, an online change detector that maintains exponential windows with $O(\log^2 t)$ memory and runtime, achieving strong F1 across real-world datasets. All three approaches capture kernel mean embeddings or low-order moments; our work investigates whether topological features can provide complementary detection signal.

Two-sample testing. The drift detection problem is closely related to the two-sample testing literature. MMD with learned kernels (Liu et al., 2020) and aggregated kernel tests (Schrab et al., 2023) improve test power by optimizing the kernel for the specific testing problem. Energy distance provides a kernel-free alternative with attractive theoretical properties. Classifier two-sample tests (C2ST) train a discriminative model to distinguish the two samples and use its accuracy as a test statistic (Lopez-Paz & Oquab, 2017). These methods provide strong baselines but do not explicitly capture topological structure.

Topological data analysis in machine learning. Persistent homology (Edelsbrunner & Harer, 2010) has been applied to a variety of machine learning problems, including shape classification, time-series analysis (Seversky et al., 2016), and anomaly detection (Chen et al., 2024). Carrière et al. (2017) introduce the sliced Wasserstein kernel for persistence diagrams, enabling kernel-based learning on topological summaries. Kusano et al. (2017) develop persistence-weighted Gaussian kernel methods that enable MMD-like two-sample testing directly on persistence diagrams—an approach complementary to the Wasserstein distance we use here; combining kernel methods on persistence diagrams with learned-kernel two-sample tests is a promising direction for future work. Despite these advances, persistent homology has seen limited application to drift detection in embedding streams.

Positioning of our work. Unlike prior drift detectors that capture only low-order statistics, we use persistent homology to detect structural changes in embedding point clouds. Unlike prior TDA applications that focus on static classification, we address the streaming drift detection problem. We note that stronger two-sample testing baselines exist beyond those evaluated here, including learned-kernel MMD (Liu et al., 2020) and C2ST with nonlinear classifiers; we discuss extending our comparison to these methods in Section 5.1. We complement existing work by providing an initial evaluation of when topological features may add value beyond classical detectors and when they do not.

3. Methodology

3.1. Problem Formulation

We consider a stream of text documents embedded into $\mathbb{R}^d$ by a pretrained sentence encoder f. A **reference window** $\mathcal{R} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^d$ captures the baseline distribution, and a **test window** $\mathcal{W} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ arrives at monitoring time. The drift detection problem is to decide whether $\mathcal{R}$ and $\mathcal{W}$ are drawn from the same distribution, i.e., to test $H_0 : P_{\mathcal{R}} = P_{\mathcal{W}}$ against $H_1 : P_{\mathcal{R}} \neq P_{\mathcal{W}}$. We use a score function $s(\mathcal{R}, \mathcal{W}) \in \mathbb{R}$ and raise an alarm when s exceeds a threshold τ calibrated on held-out reference windows.

3.2. Datasets and Embedding Models

We use two text classification datasets:

- AG NEWS (4 classes, 5,000 stratified samples) (Zhang et al., 2015), a news topic classification benchmark.
- 20 NEWSGROUPS (20 classes, 5,000 random samples), a multi-topic discussion forum corpus.

Each dataset is embedded with two sentence-transformer models: MINILM-L6 (22M parameters, 384 dimensions) and MINILM-L3 (17M parameters, 384 dimensions), yielding 4 dataset–model configurations.

3.3. Drift Scenarios

We design 7 drift scenarios per dataset. Three are *centroid-controlled*: they either preserve the centroid exactly (style perturbation, geometric reorganization) or approximately (subtopic reweighting), reducing the advantage of mean-based detectors.

Standard drifts. (1) **Abrupt topic drift**: 80% of the test window switches to a single target class. (2) **Gradual topic drift**: the target class fraction increases linearly from 10% to 80% over 5 consecutive windows. (3) **Domain shift**: reference windows draw from one set of classes (e.g., comp.* newsgroups) and test windows from a disjoint set (e.g., sci.* newsgroups).

Centroid-controlled drifts. (4) **Subtopic reweighting** (approximate centroid preservation): we split the target class into 2 subtopics via k-means clustering. The reference window contains a 50/50 split; the test window shifts to 90/10. Because both subtopics share the same class label, the centroid shift is small but generally nonzero. (5) **Style perturbation** (exact centroid preservation): we add balanced opposing noise along a random direction in embedding space, preserving the centroid exactly while changing the point cloud geometry. We note that this is a synthetic proxy for real stylistic drift; a more realistic evaluation would apply text-level paraphrasing before embedding (see Section 5.1). (6) **Geometric reorganization** (exact centroid preservation): we shift each point 40% toward the centroid of the next cluster (cyclically assigned via k-means with $k = 4$), then re-center the entire window to preserve the mean exactly. This creates a topological change—merging of connected components, altered cluster structure—while guaranteeing zero centroid shift.

Control. (7) **No-drift**: uniform class distribution throughout, used for threshold calibration.

3.4. Detection Methods

Classical baselines (6). We evaluate: (i) centroid shift (L2 distance between window means), (ii) covariance shift (Frobenius norm of covariance difference), (iii) MMD with RBF kernel using the median heuristic for bandwidth selection, (iv) energy distance, (v) kNN distance shift ($k = 5$, change in mean k-nearest-neighbor distance), and (vi) classifier two-sample test (C2ST, logistic regression with 3-fold cross-validation AUC).

Topological features (8). We compute Vietoris–Rips persistence diagrams using Ripser (Bauer, 2021) and extract: (i–ii) persistence entropy for H_0 and H_1, (iii–iv) Wasserstein-1 distance on persistence diagrams for H_0 and H_1 (scikit-tda, 2023), (v–vi) Betti curve L2 distance for H_0 and H_1, (vii) persistence landscape L2 distance (first 3 landscapes, H_1), and (viii) persistent homology dimension (PHD) for H_0. For Wasserstein distance, we compare the test window's diagram against the mean distance to each reference diagram. All other TDA features are computed as scalar summaries and compared via absolute difference from the reference mean.

3.5. Experimental Protocol

Windowing and calibration. We use a default window size of 100 embeddings. For each configuration, we compute scores on 10 reference windows to estimate the null distribution, set the detection threshold at the 95th percentile (targeting 5% FPR), and evaluate on 10 test windows.

Preprocessing. Before computing persistence diagrams, we apply PCA to reduce the embedding dimensionality from 384 to 50 (default). We use all 100 points in each window for TDA computation (no subsampling at default settings). The persistence diagram is computed once per window and shared across all 8 TDA features.

Evaluation metrics. Our primary metric is AUC computed across the 10 reference and 10 test windows per configuration. We report means and standard deviations across 5 random seeds (42, 123, 456, 789, 1011). We assess statistical significance via *paired* permutation tests with 1,000 permutations, which directly compare the AUC of two meth-

ods on the same data splits to test whether one method significantly outperforms the other (not merely whether AUC exceeds chance).

Ablations. We ablate three key hyperparameters on the AG NEWS dataset with MINILM-L6 embeddings: window size $\in \{50, 100, 200, 400\}$, TDA point count $\in \{40, 80, 100, 160\}$ (subsampling applied only when point count $<$ window size), and PCA dimensions $\in \{20, 50, 100, \text{raw } (384)\}$. Each ablation uses 3 random seeds.

Hardware. All experiments run on CPU-only Linux (x86_64, Python 3.12.8). Embedding time is approximately 5 minutes per dataset–model pair. The main experiment completes in approximately 18 minutes (4 configurations $\times$ 5 seeds); ablations require an additional 35 minutes.

4. Results

4.1. Main Results

Table 1 reports AUC for all 14 detection methods across the 6 drift types, averaged over both datasets, both models, and 5 seeds (20 runs per cell).

Standard drifts. On abrupt topic drift, gradual drift, and domain shift, classical baselines achieve near-perfect AUC (0.91–1.00). Most TDA features lag behind, with the notable exception of Wasserstein H_0, which achieves AUC = 0.997 on abrupt drift and 1.000 on domain shift—matching the best classical methods. These results confirm that classical methods are sufficient for drifts that introduce detectable centroid shifts.

Centroid-controlled drifts. On **geometric reorganization** (exact centroid preservation), Wasserstein H_0 achieves the highest AUC (0.858 ± 0.100), outperforming the best classical method (MMD-RBF, 0.782 ± 0.063). A paired permutation test directly comparing the per-split AUC scores of Wasserstein H_0 and MMD-RBF confirms that the difference is significant ($p < 0.001$). This demonstrates that persistent homology captures changes in connected-component structure that kernel methods partially miss—though we note that MMD-RBF still achieves a respectable 0.782, so this drift is not invisible to classical methods, merely harder for them to detect.

On **style perturbation** (exact centroid preservation), Wasserstein H_1 achieves the highest overall AUC (0.686 ± 0.096), marginally exceeding kNN distance (0.663 ± 0.084). However, all methods achieve only moderate detection power (AUC < 0.70), indicating that balanced noise perturbation is difficult to detect at this signal strength regardless of the test statistic used.

On **subtopic reweighting** (approximate centroid preservation), covariance shift achieves the highest AUC (0.740 ± 0.117), outperforming the best TDA method (Wasserstein H_1, 0.641 ± 0.086) by 0.099. The 90/10 subtopic split introduces a detectable change in the covariance matrix that classical methods capture directly; with only 100 points per window, the persistence summaries lack sufficient resolution to distinguish this mixture shift from sampling noise.

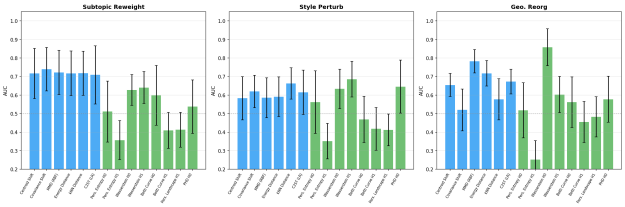

Figure 1. AUC on the three centroid-controlled drift scenarios, averaged across datasets, models, and seeds. Wasserstein H_0 achieves the best performance on geometric reorganization (+7.6% over MMD-RBF), while covariance shift remains stronger on subtopic reweighting. Error bars show ± 1 std.

4.2. Statistical Significance

Table 2 reports paired permutation test results comparing the best TDA method against the best classical baseline on each centroid-controlled scenario. Each test directly compares paired AUC scores across the same data splits, testing whether one method significantly outperforms the other (not merely whether AUC exceeds chance).

The TDA advantage on geometric reorganization is highly significant ($p < 0.001$). The marginal TDA advantage on style perturbation is not significant ($p = 0.14$). The classical advantage on subtopic reweighting is significant ($p = 0.003$), confirming that covariance shift is the better test statistic for this drift type.

4.3. Threshold Calibration

As a calibration sanity check, we verify that the threshold-setting procedure yields the intended false positive rate. Since the threshold is set at the 95th percentile of reference-window scores, the realized FPR on these same windows is 5% by construction. Figure 2 confirms this across all methods. **Important caveat:** this is not an independent evaluation of FPR, as the threshold and FPR are computed on the same windows. A proper evaluation requires held-out no-drift windows that are disjoint from the calibration set; we identify this as a key improvement for future work (see Section 5.1).

4.4. Runtime Analysis

Table 3 shows per-window runtime for each method under the tested settings (PCA to 50 dims, 100-point windows).

Table 1. AUC by method and drift type, averaged across datasets, models, and 5 seeds. Best result per drift type in **bold**. Higher is better.

Method	Abrupt	Gradual	Domain	Subtopic	Style	Geo. Reorg.
Classical baselines						
Centroid Shift	**1.000**±.000	.954±.041	**1.000**±.000	.717±.136	.583±.116	.654±.062
Covariance Shift	.973±.040	.912±.065	**1.000**±.000	**.740**±.117	.620±.088	.520±.113
MMD-RBF	**1.000**±.000	**.957**±.035	**1.000**±.000	.722±.119	.586±.108	.782±.063
Energy Distance	**1.000**±.000	**.957**±.037	**1.000**±.000	.718±.120	.591±.108	.718±.069
kNN Distance	.626±.162	.634±.194	**1.000**±.000	.718±.119	**.663**±.084	.578±.111
C2ST (LR)	**1.000**±.000	.943±.052	**1.000**±.000	.709±.157	.615±.120	.672±.067
Topological features						
Pers. Entropy H_0	.508±.202	.485±.127	.717±.144	.512±.164	.562±.169	.518±.149
Pers. Entropy H_1	.516±.089	.470±.110	.284±.165	.357±.105	.352±.096	.251±.102
Wasserstein H_0	.997±.010	.924±.049	**1.000**±.000	.628±.083	.633±.106	**.858**±.100
Wasserstein H_1	.379±.084	.464±.121	.382±.144	.641±.086	.686±.096	.603±.098
Betti Curve H_0	.750±.162	.640±.171	.397±.366	.599±.161	.469±.125	.561±.137
Betti Curve H_1	.710±.145	.612±.171	.421±.151	.409±.097	.418±.115	.455±.111
Pers. Landscape H_1	.710±.142	.610±.153	.516±.164	.413±.094	.412±.085	.483±.108
PHD H_0	.388±.101	.491±.135	.669±.110	.537±.144	.645±.143	.578±.124

Table 2. Paired permutation tests comparing Wasserstein vs. best classical baseline on centroid-controlled scenarios. ΔAUC = TDA − Classical.

Drift Type	TDA / Classical	ΔAUC	p
Geo. Reorg.	Wass H_0 / MMD	+.076	<.001
Style Pert.	Wass H_1 / kNN	+.023	.14
Subtopic Rew.	Wass H_1 / Cov	−.099	.003

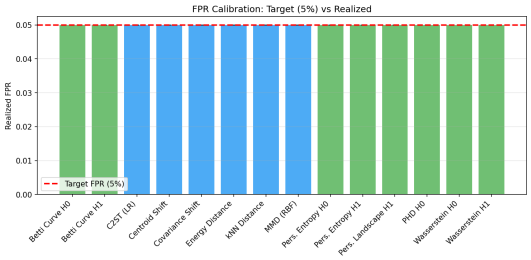

Figure 2. Calibration check: realized FPR matches the 5% target by construction (threshold set at 95th percentile of reference scores). This confirms internal consistency but is not an independent FPR evaluation.

TDA methods require 37–45 ms per window, dominated by the persistence diagram computation (37 ms), which is shared across all 8 TDA features. Feature extraction adds only 0.2–7.5 ms. Notably, TDA methods are faster than C2ST (133 ms) and comparable to kNN distance (54 ms). These runtimes suggest feasibility for small-window CPU-based monitoring under the tested settings; scaling to larger windows, raw embeddings, or higher throughput would require further benchmarking.

Table 3. Per-window runtime (ms) under tested settings (PCA-50, 100-point windows). TDA methods share the 37 ms diagram computation.

Method	Total	Diagram	Feature
Centroid Shift	0.1	—	—
Energy Distance	12.0	—	—
MMD-RBF	29.9	—	—
Pers. Entropy H_0	37.4	37.1	0.2
Pers. Landscape H_1	41.8	37.1	4.7
Wasserstein H_0	44.7	37.1	7.5
kNN Distance	53.5	—	—
Covariance Shift	58.9	—	—
C2ST (LR)	133.0	—	—

4.5. Ablation Studies

We ablate three hyperparameters on the AG NEWS/MINILM-L6 configuration, focusing on the best TDA method (Wasserstein H_0) on geometric reorganization (Figure 3).

Window size. Smaller windows (50–100 points) yield higher AUC (0.880 and 0.847, respectively) than larger windows (200: 0.613; 400: 0.687). Larger windows dilute the geometric signal by mixing drifted and undrifted samples, making the topological change harder to detect. This has practical implications: monitoring systems should use frequent, small-window checks rather than infrequent, large-window analyses for topological drift detection.

Point count. Performance increases monotonically with the number of points used for TDA: AUC rises from 0.603 (40 points) to 0.900 (160 points). More points produce richer persistence diagrams with less noise, directly improving detection power. When compute allows, using the

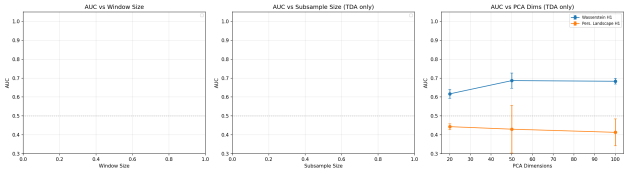

Figure 3. Ablation studies for Wasserstein H_0 on geometric reorganization (AG NEWS, MINILM-L6). *(Left)*Window size: smaller windows (50–100) perform best. *(Center)*Point count: more points monotonically improve detection. *(Right)*PCA dims: aggressive reduction (20) or raw embeddings outperform moderate reduction.

full window without subsampling is preferred.

PCA dimensionality. Aggressive dimensionality reduction helps: PCA to 20 dimensions achieves AUC = 0.900, compared to 0.847 for 50 dimensions and 0.823 for 100 dimensions. Interestingly, raw 384-dimensional embeddings also perform well (AUC = 0.860), suggesting that the topological signal exists at multiple scales. Moderate PCA dimensions (50–100) appear to occupy an unfavorable middle ground where noise is partially but not sufficiently reduced.

5. Discussion

When does persistent homology add value? Our results identify a specific niche where topological drift detection provides complementary signal: drifts that reorganize the cluster structure of the embedding space while preserving the centroid exactly. On geometric reorganization, Wasserstein H_0 captures changes in the connected-component birth–death structure that kernel methods partially miss, achieving a 7.6% AUC improvement over MMD-RBF (paired permutation $p < 0.001$). We emphasize that this is a *complementary* advantage, not a replacement: MMD-RBF still achieves 0.782 AUC on this scenario, and outperforms all TDA methods on subtopic reweighting.

Real-world relevance of geometric reorganization. Our strongest positive result comes from a controlled synthetic scenario, which raises the question of whether such drifts occur in practice. We hypothesize that centroid-preserving geometric reorganization can arise in several deployment settings: (i) gradual concept drift where new subtopics emerge and old ones fade within a stable class, preserving the aggregate centroid while restructuring clusters; (ii) adversarial or strategic distribution shifts that maintain aggregate statistics to evade mean-based monitors; and (iii) model updates or embedding version changes that rearrange the internal geometry of representations without shifting class centroids. Validating these hypotheses on production data is an important direction for future work.

When do classical methods suffice? On subtopic reweighting, the 90/10 split changes the covariance detectably; classical methods capture this directly. With only 100 points per window, the persistence summaries lack sufficient resolution to distinguish this mixture shift from sampling noise—the bottleneck is small sample size, not a fundamental limitation of TDA. On standard drifts (abrupt topic change, domain shift), classical methods achieve near-perfect AUC, making TDA unnecessary overhead. Most TDA features—persistence entropy, Betti curves, landscapes—consistently underperform, often near chance level (AUC $\approx$ 0.5). These features collapse the diagram to coarse scalars that lose the pairing information making Wasserstein distance effective.

Connection to hypothesis testing. Our work connects to the two-sample testing literature by framing drift detection as a sequence of hypothesis tests on embedding windows. The key insight is that the *test statistic* matters: while MMD captures differences in the reproducing kernel Hilbert space, Wasserstein distance on persistence diagrams captures topological differences that may lie outside the RKHS. The optimal choice of test statistic depends on the alternative hypothesis: for centroid-shifting drifts, kernel statistics suffice; for topology-altering drifts, persistent homology provides complementary power. This motivates future work on hybrid test statistics combining both.

Practical recommendations. For general-purpose monitoring, we recommend MMD-RBF as the default detector: it achieves AUC $\geq$ 0.72 on all centroid-controlled drifts, runs at 30 ms per window, and requires no hyperparameter tuning beyond the median bandwidth heuristic. When topological drift is suspected (e.g., model performance degrades but MMD does not fire), adding Wasserstein H_0 as a confirming signal costs only 45 ms per window under the tested settings. The optimal TDA configuration is window size 50–100, point count $\geq$100, PCA dimension 20. A two-stage pipeline—MMD as a fast first pass, Wasserstein H_0 for ambiguous cases—balances sensitivity and computational cost.

Limitations. We acknowledge several limitations that bound the scope of our conclusions: (1) Our strongest positive result (geometric reorganization) relies on a controlled synthetic construction; while we discuss plausible real-world analogues above, we have not yet demonstrated the advantage on naturally occurring drift. (2) Style perturbation uses additive embedding noise rather than actual text paraphrasing—a synthetic proxy that may not reflect real stylistic drift in deployed systems. (3) Both embedding models are MiniLM variants (same architecture family); testing with architecturally different models (e.g., E5, BGE, MPNet) is needed to establish generalizability. (4) Our baselines do not include learned-kernel MMD (Liu et al., 2020) or C2ST with nonlinear classifiers, which

may narrow or close the gap with topological methods on geometric reorganization. (5) FPR calibration is verified by construction on the same reference windows used for threshold setting; a held-out no-drift evaluation is needed for independent FPR estimation. (6) We use 5,000 samples per dataset and evaluate only fixed-window batch detection; online streaming approaches (Kalinke et al., 2022) and larger-scale evaluations remain future work. (7) Runtime is benchmarked only under the tested settings (PCA to 50 dims, 100-point windows).

5.1. Future Work

Our results motivate several concrete extensions:

- **Stronger baselines.** Comparing against learned-kernel MMD (Liu et al., 2020; Schrab et al., 2023) and C2ST with nonlinear classifiers (MLP-based) would clarify whether the TDA advantage on geometric reorganization persists against state-of-the-art two-sample tests. Kernel methods on persistence diagrams (Kusano et al., 2017) offer another promising direction, enabling MMD-like testing directly on topological summaries.
- **Encoder diversity.** Extending the evaluation to architecturally different embedding families (e.g., E5, BGE, MPNet) is necessary to establish whether the observed TDA advantage is encoder-general or specific to the MiniLM architecture.
- **Realistic style drift.** Replacing synthetic embedding noise with text-level paraphrasing (e.g., via synonym replacement or back-translation before embedding) would provide a more ecologically valid test of style-drift detection.
- **Hold-out FPR calibration.** Splitting no-drift data into separate calibration and validation sets would enable independent estimation of the false positive rate.
- **Production validation.** Evaluating on naturally occurring drift in production embedding streams would test whether the geometric reorganization advantage transfers beyond controlled experiments.

6. Conclusion

We present an initial systematic evaluation of persistent homology for distribution drift detection in LLM embedding streams. Across 2 datasets, 2 embedding models, 7 drift scenarios, 14 detection methods, and 5 random seeds, we find that Wasserstein distance on H_0 persistence diagrams significantly outperforms the best classical baseline on a controlled geometric reorganization scenario (AUC 0.858 vs. 0.782, paired permutation $p < 0.001$)—a drift that preserves the centroid exactly while reorganizing cluster structure. However, this advantage is specific to this controlled scenario: classical methods remain superior on subtopic reweighting, and all methods achieve only moderate detection power on style perturbation. These results suggest that persistent homology may provide complementary value for geometry-reorganizing drift, but it is not a replacement for classical drift detectors, and the evidence base requires strengthening through comparison with stronger learned-kernel baselines, evaluation on diverse encoder architectures, and validation on naturally occurring drift.

From a hypothesis testing perspective, our results demonstrate that the choice of test statistic critically determines which alternatives are detectable. Kernel-based statistics and topological statistics appear to capture complementary aspects of distributional differences, motivating future work on hybrid test statistics—including kernel methods on persistence diagrams (Kusano et al., 2017)—that may combine both for more powerful drift detection.

References

Bauer, U. Ripser: Efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5:391–423, 2021.

Carrière, M., Cuturi, M., and Oudot, S. Sliced Wasserstein kernel for persistence diagrams. In *International Conference on Machine Learning*, 2017.

Chen, Y., Li, M., and Wang, W. Anomalous trajectory detection via persistent homology. *arXiv preprint arXiv:2410.03889*, 2024.

Edelsbrunner, H. and Harer, J. L. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

Greco, S., Vacchetti, B., Apiletti, D., and Cerquitelli, T. Driftlens: An unsupervised drift detection framework for deep learning representations in real-time. *arXiv preprint arXiv:2406.17813*, 2024.

Kalinke, F., Heyden, M., Gntuni, E., Fouché, E., and Böhm, K. MMDEW: Online MMD change detection via exponential windowing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Khaki, M., Aditya, S., Karnin, Z., Ma, L., Pan, Z., and Chandrashekar, J. Uncovering drift in textual data: An unsupervised method for detecting and mitigating drift in machine-learned representations. *arXiv preprint arXiv:2309.03831*, 2023.

Kusano, G., Fukumizu, K., and Hiraoka, Y. Kernel method for persistence diagrams via kernel embedding and weight factor. In *Journal of Machine Learning Research*, volume 18, pp. 1–41, 2017.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. Learning deep kernels for non-parametric two-sample tests. *arXiv preprint arXiv:2002.09116*, 2020.

Lopez-Paz, D. and Oquab, M. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(198):1–81, 2023.

scikit-tda. Persim: Distances and representations of persistence diagrams. `https://persim.scikit-tda.org`, 2023.

Seversky, L. M., Davis, S., and Berger, M. On time-series topological data analysis: New data and opportunities. *CVPR Workshop on Computer Vision for Accelerated Processing*, 2016.

Zhang, X., Zhao, J., and LeCun, Y. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28, 2015.