

A Geometry-Aware Depth Function for Robust Unsupervised Textual Anomaly Detection

Pierre Fihey, Matthieu Labeau, Pavlo Mozharovskiy

LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France

{pierre.fihey, matthieu.labeau, pavlo.mozharovskiy}@telecom-paris.fr

Abstract

Textual anomaly detection (AD) aims to identify samples that deviate from the underlying distribution of normal data, without requiring explicit supervision on anomalies. Existing literature has focused almost exclusively on the semi-supervised setting, assuming access to clean training datasets composed solely of normal samples. However, this assumption rarely holds in real-world scenarios, as datasets used to infer the normal distribution are frequently contaminated by anomalies. In this work, we first highlight the sensitivity of conventional anomaly detection methods to contamination, before demonstrating the relevance of *data depth functions* for robust textual anomaly detection. Building on recent findings regarding the geometry of language model embeddings, further supported by our empirical observations, we introduce the *geodesic depth* to better capture the structure of contextual embedding spaces. Our results show that the geodesic depth consistently outperforms commonly used anomaly detection methods once contamination exceeds 10% in the training data, with substantially larger improvements observed in the most challenging scenarios.

1 Introduction

Rather than standalone approaches (Ruff et al., 2019; Manolache et al., 2021), in which models are trained from scratch on normal data to detect textual anomalies, *two-stage frameworks*, in which AD algorithms operate directly on embeddings produced by pretrained models, have become the standard approach for this task (Li et al., 2024). Most of the recent benchmark studies have primarily focused on the semi-supervised setting, in which the reference distribution is learned from training data composed exclusively of normal samples (Bejan et al., 2023; Li et al., 2024). In this work, we focus on the fully unsupervised setting, where access to a clean training set is no longer assumed. Detection

models must therefore infer the distribution of normal data from the majority of samples and identify observations that strongly deviate from it, either within the training set itself or on separate test data. This setting arguably provides a more realistic formulation of textual anomaly detection, as clean nominal datasets are rarely available in real-world scenarios. In particular, unsupervised anomaly detection can be leveraged to clean datasets and improve data quality for training large language models (Lai et al., 2026). Combining it with zero or few-shot learning (Gera et al., 2022) may further facilitate the construction of reliable datasets with reduced annotation costs. Despite its practical relevance, the unsupervised setting remains largely underexplored in the literature, with the exception of Cao et al. (2025). Yet, it directly challenges the assumptions underlying most anomaly detection approaches: methods capturing global structure (e.g., autoencoders, GMM) and those relying on local neighborhoods (e.g., KNN, LUNAR) are both sensitive to this *contamination* of training data by anomalies, through biased representations or corrupted neighborhoods, highlighting the need for more robust approaches. In this work, we propose to integrate robust statistical methods into the two-stage framework, through the use of **data depth functions**.

Introduced by Tukey (1975) and further developed by Donoho and Gasko (1992a), depth functions exhibit several properties that make them particularly suitable for this setting. First, they are affine invariant, and therefore robust to variations in scale and correlations across dimensions. This property makes data depth functions particularly well-suited to language model embeddings, which are known to exhibit strong anisotropy (Ethayarajh, 2019), thereby degrading the performance of classical detection algorithms (Fihey et al., 2026). Moreover, depth functions satisfying the projection property have been shown to possess high breakdown

points (Zuo, 2003a), meaning they require a large amount of anomalous data to significantly affect their estimates, making them particularly robust to contamination. Until recently, the application of data depth functions to very high-dimensional data remained limited due to their computational complexity. However, recent work by (Dyckerhoff et al., 2021) and (Leone et al., 2025) have introduced methods enabling the parallel computation of specific depth functions, making them applicable to large-scale, high-dimensional representations such as language model embeddings. Recent works suggest that language model embeddings lie on low-dimensional, potentially curved manifolds, either by showing that low-dimensional latent variables are mapped to continuous, curved structures in high-dimensional space (Modell et al., 2025), or by empirically measuring low intrinsic dimensionality in contextual representations (Ruppik et al., 2025). We argue that we can achieve further improvements by designing depth functions explicitly adapted to the specific properties of contextual embeddings. We therefore introduce a **geodesic depth**, which extends distance-based depth functions (Zuo and Serfling, 2000) by leveraging geodesic distances (Tenenbaum et al., 2000) to better capture their manifold structure. Our contributions are summarized as follows:

- We investigate state-of-the-art **AD methods in a fully unsupervised setting**, where anomalies are present in the training data at various rates.
- We highlight the **effectiveness of data depth functions** for textual anomaly detection and propose a **novel depth function based on geodesic distances** that better captures the underlying manifold structure of textual embeddings.
- We propose an analysis of the structure of the embeddings and its relationship to the performance of our depth, and apply our detection method to sample selection in a self-training framework.¹

2 Related Work

2.1 Contamination in Textual AD

Most prior studies on textual anomaly detection rely on repurposed text classification datasets, with recent work reformulating multi-class benchmarks (e.g., AGNews, 20NewsGroup...) by considering one class as in-distribution and the others as out-

of-distribution (Ruff et al., 2019; Manolache et al., 2021; Bejan et al., 2023; Cao et al., 2025). However, the vast majority of these works assumes a clean training dataset. While Manolache et al. (2021) and Pantin and Marsala (2024) investigated data contamination by artificially injecting anomalous samples in the training set, the contamination levels considered remained limited to 10–15%. In contrast, real-world applications can involve significantly more challenging scenarios, particularly in weak supervision settings, where label noise can reach up to 50% (Agro and Aldarmaki, 2023). Moreover, the performance reported in these works is now largely outperformed by two-stage frameworks, of which only Cao et al. (2025) uses contaminated data, but with levels set around 3%. Integrating robust detection methods into existing two-stage frameworks can therefore substantially outperform previously proposed models.

2.2 Two Steps Anomaly Detection

Detection Algorithms Existing studies in textual AD typically evaluate a broad range of AD algorithms trained on in-distribution representations to score unseen samples. We distinguish between these methods based on their detection mechanisms: (1) **instance-based methods**, which rely on local neighborhood relationships and identify anomalies as points that are isolated or lie in low-density regions, and (2) **model-based methods**, which learn a parametric model of data and consider samples that are poorly explained as anomalies. While these methods have shown strong performance in the semi-supervised setting, robust alternatives are needed when training data may be contaminated. In this work, we investigate the use of **data depth functions** to address this challenge.

Embedding models While recent studies have highlighted the strong advantage of embeddings derived from large language models (Li et al., 2024; Cao et al., 2025), our previous work (Fihey et al., 2026) showed that their inherent anisotropy can significantly affect the performance of anomaly detection algorithms relying on distance-based measures. Indeed, these representations tend to concentrate along a few dominant directions rather than being uniformly distributed across the embedding space, making pairwise distances less informative. In particular, similarity-trained embeddings tend to outperform alternatives, as contrastive learning promotes more isotropic representations (Gao et al.,

¹Our code is available at https://github.com/PierreFi/Unsup_AD

2021), The work of Cai et al. (2021) suggests that this anisotropy can be explained by the geometric organization of embeddings, which tend to concentrate on low-dimensional, potentially curved manifolds within high-dimensional space. This observation, recently reinforced by Ruppik et al. (2025) and Modell et al. (2025), implies that semantically distant points along the manifold may appear artificially close under Euclidean or cosine distances, due to shortcuts through the ambient space. For such settings, Tenenbaum et al. (2000) proposed to estimate *geodesic* distances via shortest paths on a neighborhood graph approximating the underlying structure. Applied to language model embeddings, geodesic distances behave similarly to cosine distances for close representations, suggesting a locally linear structure of embedding manifolds (Liu et al., 2025; Modell et al., 2025), while better capturing cross-domain semantic separation. We therefore build upon this idea to adapt our distance computations to the geometry of embedding spaces to propose a new depth measure.

3 Data Depths

A depth function $D(x | P)$ quantifies the centrality of a point $x \in \mathbb{R}^d$ with respect to a probability distribution P . In this work, P is taken as the empirical measure induced by a set of points $X = \{x_i\}_{i=1}^n$, and the depth of x is denoted $D(x | X)$. Zuo and Serfling (2000) formalize the notion of a proper multivariate depth function through four key axioms: affine invariance, maximality at the center of symmetry, monotonicity from the deepest point, and vanishing at infinity. The authors further classify existing depth functions according to their underlying construction principles. Building on this taxonomy, we group depth functions into two broad categories according to their aggregation mechanism: **expectation-based** depths, which estimate centrality by averaging distances or similarities over the data distribution, and **optimization-based** depths, which are defined through extremal operators and can thus be formulated as optimization problems. We first investigate optimization-based depths, with a particular focus on projection depth.

3.1 Projection Depth

The **projection depth** of a point x measures its centrality by considering its worst deviation from the projected median across all one-dimensional

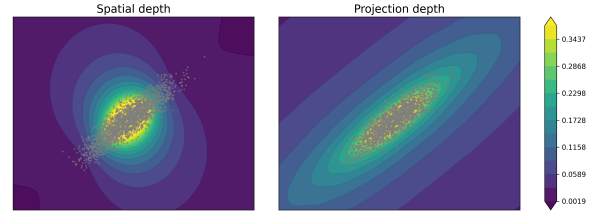


Figure 1: Spatial depth vs. projection depth. Depth contours computed on simulated data following an elliptical gaussian distribution. Gray dots indicate training samples. Projection depth yields contours that better align with the elongated geometry of the distribution.

directions:

$$D_{proj}(x | X) = \min_{u \in \mathbb{S}^{d-1}} \left(1 + \frac{|u^\top x - \text{med}(u^\top X)|}{\text{MAD}(u^\top X)} \right)^{-1} \quad (1)$$

where $u^\top x$ denotes the projection of x onto direction u , $\text{med}(u^\top X)$ is the median of the projected data $\{u^\top x_1, \dots, u^\top x_n\}$, and $\text{MAD}(u^\top X)$ denotes the corresponding median absolute deviation. The following properties make the projection depth particularly well suited for textual AD in a fully unsupervised setting:

Affine Invariance The projection depth satisfies the affine invariance axiom. That is, for any non-singular $d \times d$ matrix A and any $b \in \mathbb{R}^d$,

$$D_{proj}(x | X) = D_{proj}(Ax + b | AX + b),$$

with $AX + b = (Ax_1 + b, \dots, Ax_n + b)$. This property ensures that the depth is independent of the chosen coordinate system. In particular, linear correlations between variables and relative differences in scale do not affect the centrality of a point. Since evaluating centrality relative to the intrinsic geometry of the data, an affine-invariant depth function is well suited to anisotropic settings. Figure 1 illustrates the advantage of the projection depth on an anisotropic two-dimensional Gaussian distribution.

Robustness The robustness of an anomaly detection algorithm can be defined through its breakdown point, that is, the amount of anomalous data needed to significantly affect its predictions. Depth functions satisfying the projection property, in which the depth of a point is defined as the minimum over all one-dimensional projected depths, have been shown to possess high breakdown points. Donoho and Gasko (1992b) established that estimators based on halfspace depth achieve a breakdown point of $1/3$, extending the robustness properties

of the univariate median to the multivariate setting. Zuo (2003b) further demonstrated that projection depth-based estimators can attain an even higher breakdown point approaching 1/2, making projection depth among the most robust affine-equivariant approaches for multivariate data analysis.

Computational Complexity The application of the projection depth to very high-dimensional data, such as textual embeddings, has been severely constrained by its high computational cost. However, Dyckerhoff et al. (2021) showed that depth functions satisfying the projection property can be efficiently approximated by evaluating depth over a finite set of one-dimensional projections, providing an upper bound on the true multivariate depth. Recently, Leone et al. (2025) introduced a novel approach to parallelize the computation of projection depth approximations, enabling speedups of up to $7,000\times$ on GPUs. This methodology enables the practical application of projection depth approximations to high-dimensional textual embeddings in our experiments.

3.2 Expectation-based depths

While expectation-based depths exhibit markedly different properties, they also offer clear advantages for our study. Following the taxonomy of Zuo and Serfling (2000), such depths define centrality by averaging a measure of closeness or distance between a point and the data distribution: unlike instance-based anomaly detection methods relying on local neighborhoods, it aggregates distances globally, making it naturally more robust to contamination. The **L2 depth** uses the expected Euclidean distance to the data distribution:

$$D_{L2}(x | X) = (1 + \mathbb{E}_X[\|x - X\|_2])^{-1} \quad (2)$$

As a data-driven, non-parametric approach, it imposes no restrictive distributional assumptions, and remains computationally efficient, scaling linearly with the number of samples. However, we believe that this function can be further adapted to the geometry of the textual embedding space.

3.2.1 Geodesic Depth

We introduce the **geodesic depth**, building on the L2 depth function and using geodesic distances. We adopt a median-based aggregation to improve its stability on lower-dimensional manifolds, where anomalies can be significantly more heterogeneous.

Geodesic distances approximation We follow the framework of Tenenbaum et al. (2000) to approximate geodesic distances. Specifically, we construct a weighted k -nearest neighbor graph over the training embeddings, with edge weights defined by pairwise cosine distances between embeddings. The geodesic distance between two embeddings x_i and x_j is then approximated by the shortest-path distance on this graph and denoted by $d_G(x_i, x_j)$.

Depth Computation Building on the precomputed neighborhood graph, the depth of a query sample is computed by first linking its embedding to the graph through its k -nearest neighbors. Geodesic distances from the query to the training data are then approximated using shortest-path distances on the graph. The median of the resulting distances is then computed to obtain a robust measure of centrality, which is finally mapped to a depth score following the distance-based depth formulation. Hence, our geodesic depth, for an observation x , is defined as follows:

$$D_{\text{geo}}(x | X) = \left(1 + \text{med}_{x_j \in X} \tilde{d}_G(x, x_j)\right)^{-1} \quad (3)$$

with

$$\tilde{d}_G(x, x_j) = \min_{x_i \in \mathcal{N}_k(x)} (d(x, x_i) + d_G(x_i, x_j)),$$

where $\mathcal{N}_k(x)$ denotes the set of k -nearest neighbors of x in the training set, $d(\cdot, \cdot)$ is the cosine distance between embeddings, and $d_G(x_i, x_j)$ denotes the precomputed geodesic distance on the neighborhood graph. Figure 2 illustrates the advantage of the proposed geodesic depth over the standard L2 depth on non-linear manifolds, showing that the geodesic depth better adapts its level sets to the intrinsic geometry of the data.

Depth Properties We examine the proposed geodesic depth with respect to the four classical depth-function axioms. As an expectation-based depth, it satisfies **maximality at the center** and **vanishing at infinity**, and satisfies **monotonicity** for unimodal distributions. While monotonicity is not guaranteed for multimodal distributions, this is desirable in such settings, where, e.g., an anomalies' cluster exactly between two normal modes is very improbable (because of high embedding's dimension). The geodesic depth is not affine invariant, but remains invariant under translations and orthogonal transformations (Mosler and

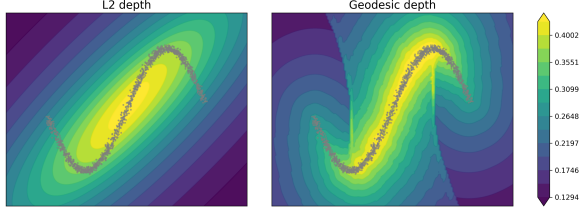


Figure 2: L2 depth vs. geodesic depth. Depth contours computed on simulated data sampled from a noisy non-linear manifold. Gray dots indicate training samples. Geodesic depth yields contours that follow the intrinsic geometry of the manifold, whereas L2 depth produces Euclidean, geometry-agnostic level sets.

Mozharovskiy, 2020). However, the lack of full affine invariance does not constitute a major limitation in our setting, as geodesic distances are specifically introduced to capture the intrinsic nonlinear geometry of textual embedding spaces.

Computational Efficiency The computational cost associated with geodesic distance estimation can become a limiting factor when dealing with large-scale datasets: computing the full shortest-path distance matrix on the neighborhood graph using Dijkstra’s algorithm (Leiserson et al., 1994) entails a time complexity of $O(kn^2 \log n)$. To mitigate this issue, de Silva and Tenenbaum (2002) proposed a landmark-based approximation that restricts shortest-path computations to a small set of L landmark points. Geodesic distances are then computed only between landmarks and all data points, yielding a computational complexity of $O(kLn \log n)$. The distance between a query sample and the training data is then estimated using a fixed set of landmark points $\mathcal{L}$, enabling efficient parallelization across query samples :

$$\tilde{d}_G^L(x, x_j) = \min_{\ell \in \mathcal{L}} (d(x, \ell) + d_G(\ell, x_j))$$

In practice, exact geodesic depth computation remains practical for relatively small datasets, while the landmark approximation provides an accurate and efficient alternative at larger scales. In what follows, we report for each experiment the computational time of our method alongside all competing detectors in Appendix A.5. Overall, depth-based methods tend to be more computationally demanding, leading to a trade-off between detection performance and efficiency.

4 Experimental setup

Given a training set of text samples $D_n = \{x_i\}_{i=1}^n$, anomaly detection² consists in learning a decision function $h : \mathcal{X} \rightarrow \{0, 1\}$ indicating whether a new sample $\mathbf{x}$ is *in-distribution* (0) or *anomalous* (1) with respect to D_n . We work in a fully unsupervised setting, where the training set $D_n = \{x_i\}_{i=1}^n$ may itself be contaminated by anomalous examples. The objective is then to infer the underlying normal distribution from this potentially corrupted training data and identify anomalies in unseen test samples³. In practice, anomaly detection methods such as our depths typically output an anomaly score, and performance is evaluated using the AUROC (Area Under the ROC Curve), which is threshold-independent metric and robust to class imbalance.

4.1 Datasets

To conduct our experiments, we follow established practice in the literature by repurposing news classification datasets for anomaly detection. We therefore consider the four classification datasets commonly used in recent benchmarks: **AG News**, **BBC News**, **N24 News**, and **20 Newsgroups**. Following Bejan et al. (2023), one class is treated as in-distribution while all remaining classes are considered anomalous. A training set is constructed from the in-distribution class and progressively contaminated with OOD samples to evaluate robustness in the fully unsupervised setting. Dataset statistics are provided in Appendix A.1.

4.2 Embedding models

To evaluate the ability of the proposed depth functions to adapt to different embedding geometries, we conduct experiments on several models. Embeddings produced by LLMs exhibit stronger anisotropy than those from similarity-trained models and therefore constitute a more challenging setting for anomaly detection, making them well suited to evaluate the robustness of depth-based

²Depending on the application domain and on the origin of the instances to be detected (e.g., whether they are generated by the same process as the normal data) this task may be referred to as anomaly detection, outlier detection, or novelty detection (Ruff et al., 2021, II.B.3). To remain consistent with prior NLP literature employing experimental settings similar to ours, we keep the term anomaly detection throughout this work.

³We also experimented with a variant in which anomalies are detected within the contaminated **training** dataset, enabling us to experiment on smaller, more realistic datasets. As expected, results were highly similar, and are therefore reported in Appendix A.7

methods; however, to ensure reproducibility under limited computational resources, we restrict our experiments to models that can be run on a single NVIDIA V100 GPU, and thus consider language models with fewer than one billion parameters. Hence, we experiment with Qwen3 (Bai et al., 2023) and LLaMA-3.2-1B (Touvron et al., 2023); following (Fihey et al., 2026), we also experiment with two similarity-trained embedding models, E5 (Wang et al., 2024) and Qwen3-Embeddings (Zhang et al., 2025). For generative models, we follow common practice and use the final transformer layer’s hidden representations as embeddings. See Appendix A.2 for model details.

4.3 Baseline AD algorithms

To provide a comprehensive baseline for our depth-based methods, we evaluate the robustness to contamination of widely used anomaly detection approaches from the literature, including three instance-based algorithms: *k*-Nearest Neighbors (KNN) (Angiulli and Pizzuti, 2002), Local Outlier Factor (LOF) (Breunig et al., 2000), and LUNAR (Goodge et al., 2022); as well as four model-based methods: Gaussian Mixture Models (GMM), Autoencoders (AE) (Bank et al., 2023), One-class SVM (OC-SVM) (Manevitz and Yousef, 2001), and Deep Isolation Forest (DIF) (Xu et al., 2022). We also investigate Copula Based Outlier Detection (COPOD) (Li et al., 2020) whose underlying mechanism shares similarities with the projection depth.

4.4 Implementation Details

For those detection algorithms, experiments are conducted using the default hyperparameter settings provided by the pyod library^{4,5}. **Projection depth** is computed using the data depth library⁶. The infimum over projection directions is found using the optimization solver *refinedrandom* using 10,000 directions, as it offers a strong trade-off between performance and computational cost while supporting GPU parallelization. For the computation of the **geodesic depth**, empirical analyses reported in Appendix A.4 show that setting $K = 0.15n$ for the neighborhood graph leads to stable and robust performance across different models

and datasets, where n denotes the number of samples. This choice is supported by the nature of our data, with embeddings lying in high-dimensional spaces (768–1024) while exhibiting a much lower intrinsic dimension, as confirmed by our estimates across all considered embeddings and datasets (see Table 1): this suggests a noisy low-dimensional manifold structure in a sparse ambient space. Consequently, small neighborhood sizes fail to capture the local geometry, and larger values of K are required to obtain reliable graph structures. This observation is consistent with prior works relying on neighborhood graphs for similar embeddings, which also adopt relatively large values of K (Jha and Mihata, 2021; Chu et al., 2023).⁷

5 Results

In this section, we present the results of our anomaly detection experiments with representations derived from similarity-trained models (Figure 3) and from LLMs (Figure 4).⁸

Preliminary observations In line with our previous findings (Fihey et al., 2026), we first observe that similarity-trained models produce embeddings that appear better suited to anomaly detection than LLMs. We also note that classical detection methods are similarly affected by data contamination, with the notable exception of the GMM, which seems particularly well suited to LLaMA embeddings.

Robustness of depths functions on similarity-trained embeddings Although the geodesic depth appears to better adapt to the similarity trained embeddings than the projection depth, neither of the two depth-based methods outperforms the other baselines when the training dataset is perfectly clean. However, the comparative advantage of these methods appears as soon as the contamination level reaches 10%, with performance remaining almost constant while that of the other algorithms rapidly declines, with the exception of COPOD, which achieves performance close to projection depth. This gap with the other baselines

⁴<https://pyod.readthedocs.io>

⁵For COPOD, which was originally designed for outlier detection, we adapt the scoring procedure to our setting by estimating and fixing its empirical distributions on the training set before evaluating the test samples against these distributions.

⁶<https://data-depth.github.io>

⁷Due to the large size of the AGNews dataset, geodesic distances are approximated using a landmark-based method with $n_{\text{landmarks}} = 256$.

⁸Experimental results showed that DIF, LOF and OC-SVM were strongly affected by contamination and consistently underperformed in this setting. For the sake of visual clarity, we exclude these methods from our figures, while a complete comparison including these methods is provided in the Appendix A.7).

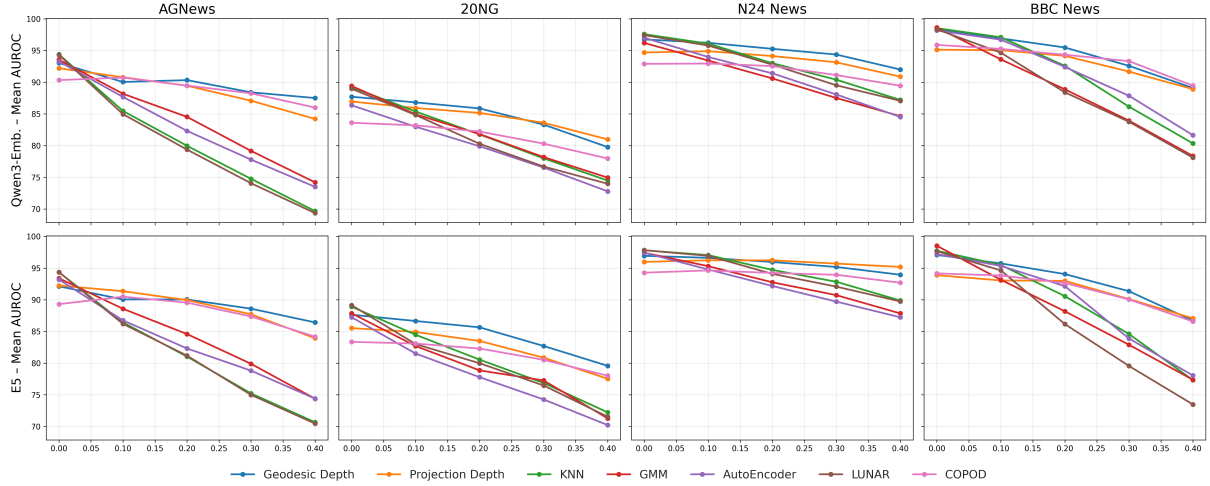


Figure 3: Mean AUROC values versus contamination rate for embeddings obtained from similarity-trained models. For the four datasets considered, this mean is obtained over choosing each class to be the inlier.

further widens as the contamination rate increases, highlighting the robustness of these methods even in highly challenging scenarios. The similar behavior of COPOD and projection depth may stem from their common reliance on univariate distributional information. Projection depth measures robust deviations from the median along projected directions, while COPOD evaluates marginal extremeness along each dimension through empirical tail probabilities, which primarily depend on the relative ranks of observations. The two methods become even more closely related in the high-dimensional spaces considered here (768-1024 dimensions), as previous work has shown that projection-based depths can be accurately approximated with a comparable number of random directions (see, e.g., [Cuesta-Albertos and Nieto-Reyes, 2008](#); [Dyckerhoff et al., 2021](#), where 500 and 1000 directions are sufficient for good approximation, respectively). Our empirical results suggest that this rank-based formulation contributes to robustness, as the relative ordering of in-distribution samples can remain largely preserved even under substantial contamination.

Geometric Robustness of Geodesic Depth on LLMs embeddings While both depths perform quite similarly on embeddings derived from similarity-trained models, the performance of the projection depth drops significantly on embeddings from LLMs. This depth seems to be strongly affected by their pronounced anisotropy and the particular structure of their embedding spaces, further analyzed in Section 6, while the geodesic depth

Model	20NG	AGNews	BBC	N24
E5	25.64	23.83	14.91	29.80
LLaMA	15.07	15.59	11.05	18.78
Qwen3-Emb.	18.19	18.41	12.33	24.87
Qwen3	13.04	17.03	10.72	20.20

Table 1: Persistent Homology Dimension (PHD) across datasets and embedding models. Lower values indicate embeddings that lie on lower-dimensional manifolds.

remains well adapted to such representations. By performing an ablation study (see Appendix A.8), we find that the standard L2 depth also seems unaffected, while still yielding inferior results. We hypothesize that distance aggregation is more stable, and additionally incorporating the adaptability of geodesic distances makes it the better method.

6 Manifold Analysis

The effectiveness of the geodesic depth appears consistent with the *manifold hypothesis*, according to which high-dimensional representations are concentrated around lower-dimensional manifolds. We intend to better understand the structure of these manifolds, beginning by studying their intrinsic dimension (ID). ID represents the minimal number of dimensions required to describe a representation without significant information loss. A low ID therefore suggests that data lie on a low-dimensional structure ([Ruppik et al., 2025](#); [Cai et al., 2021](#); [Valeriani et al., 2023](#)). To estimate the ID of our representations, we use the *Persistent Homology Dimension* estimator (PHD) ([Schwein-](#)

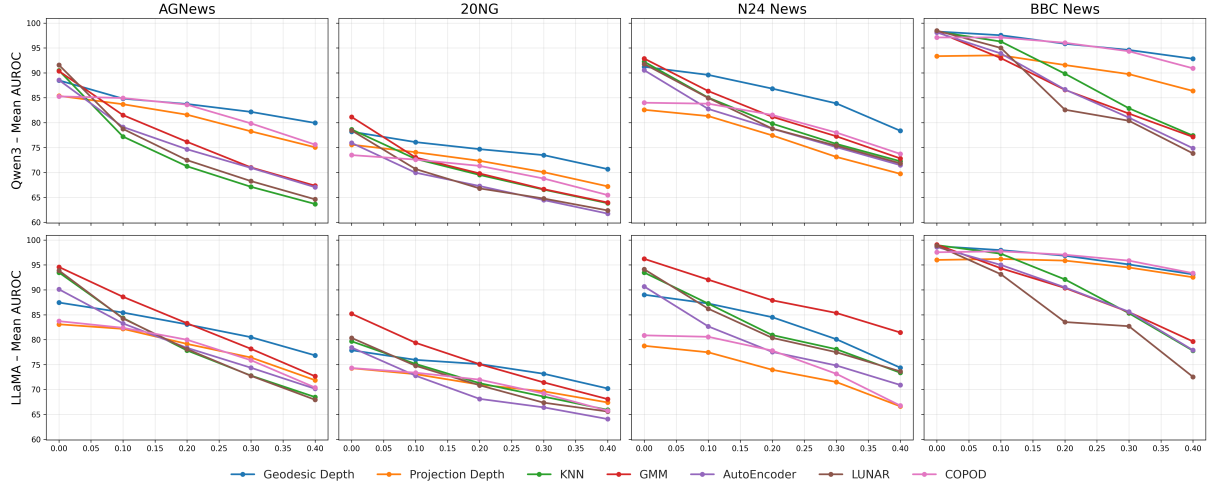


Figure 4: Mean AUROC values versus contamination rate for embeddings obtained from large language models. For the four datasets considered, this mean is obtained over choosing each class to be the inlier.

hart, 2021) and report the results for all models and datasets in Table 1.

While modern language model embeddings are high-dimensional (768–1024), they appear to lie on much lower-dimensional manifolds (10–30), which is consistent with recent findings. Moreover, the particularly low IDs observed in LLM embeddings may explain why these representations are more challenging for projection depth, and the comparative advantage of our geodesic depth. Indeed, these observations suggest a more complex nonlinear structure that may not be adequately captured by affine transformations alone. By accounting for local geometric structure, Geodesic Depth may therefore be better suited to such representations. However, the ID only provides indirect evidence about the structure of embeddings, and the existence of low-dimensional manifolds cannot be formally established. We therefore aim to reveal these structures through visualization. To this end, we rely on PCA-based projections while restricting the analysis to points for which the projection remains informative. The complete visualization methodology is detailed in Appendix A.10. Figure 5 illustrates manifold structures observed on the AGNews dataset, with in-distribution samples corresponding to the business class and out-of-distribution data formed by the remaining classes. These visualizations appear to support prior findings, suggesting that curved manifolds effectively capture the semantic organization of textual representations. We additionally observe substantial differences across embedding models, with sharper and more structured manifolds for Qwen3-Emb. and LLaMA,

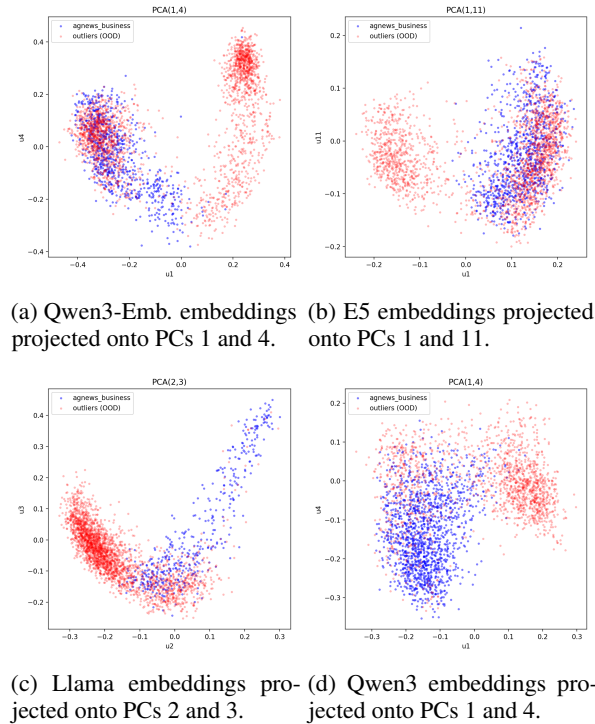


Figure 5: Visualization of low-dimensional manifold structures by projecting embeddings onto selected PCA components. In all cases, only the 5% of points closest to the projection plane are retained to better reveal the underlying manifold geometry.

and broader, noisier geometries for Qwen3 and e5. These variations may warrant further investigation.

7 Leveraging AD for Self-Training

Improving data selection through anomaly detection is relatively underexplored in NLP. In particular, we hypothesize that using anomaly detec-

tion for sample selection can improve self-training. To explore this potential, we conduct a proof-of-concept experiment illustrating how depth-based scores can be leveraged for pseudo-label selection. We experiment with the framework of Gera et al. (2022): their approach relies on a zero-shot classification framework based on natural language inference (NLI) models, where class labels are formulated as hypotheses and evaluated against input texts. Pseudo-labels are then generated by selecting the most confident predictions, and the model is subsequently fine-tuned on these examples in a self-training loop. We argue that replacing confidence-based selection with a centrality-based approach, which selects representative examples and filters out anomalies, leads to more robust pseudo-labeling, as it is less sensitive to model miscalibration and better suited to exclude noisy or out-of-distribution examples. We follow the experimental setting of Gera et al. (2022), and modify only the pseudo-label selection step, replacing confidence-based selection with an AD approach. For each class, we compute geodesic depth and select the 100 most central examples as pseudo-labels. Table 2 reports the results of our experiments on the four topic classification datasets considered in Gera et al. (2022), together with the corresponding results from the original paper⁹. AD based on the geodesic depth leads to significantly more effective self-training across the three models considered and on most datasets, highlighting the potential of this approach for improving the quality of training data for language models.

8 Conclusion

In this paper, we investigate textual anomaly detection in a fully unsupervised setting, where access to clean training data cannot be assumed. We show that data depth functions exhibit several properties that make them particularly suitable for this setting, leading to substantial improvements over state-of-the-art anomaly detection methods, whose performance is strongly affected by contamination. Building on recent findings regarding the geometry of language model embeddings, further supported by our empirical observations, we introduce the *geodesic depth*, an expectation-based depth function that leverages geodesic distances to better capture the structure of contextual embedding spaces.

⁹It is worth noting that we exclude the token masking strategy introduced in their work, as our comparison focuses on the sample selection mechanism.

Model	Method	20NG	AG	DBP	YA	Avg.
BART	Base	44.9	65.9	72.1	48.5	57.9
	ST-Conf.	66.4	71.3	94.3	61.5	73.4
	ST-GD	66.3	77.4	91.0	64.7	74.9
DeB.	Base	49.8	73.2	73.4	44.2	60.15
	ST-Conf.	62.1	81.6	89.1	60.8	73.4
	ST-GD	64.1	85.7	97.2	66.9	78.5
RoB.	Base	34.1	61.8	65.2	35.8	49.2
	ST-Conf.	61.4	73.1	92.7	57.8	71.3
	ST-GD	62.8	80.4	94.4	63.1	75.2

Table 2: Mean accuracy across datasets for each model. *ST-Conf.* refers to the self-training method introduced in the original paper, while *ST-GD* denotes our proposed geodesic depth-based approach.

This depth function consistently outperforms all methods considered in our study, especially on the more challenging LLM embeddings, remaining highly robust even under severe contamination levels. Motivated by these findings, we plan to investigate the integration of geodesic depth into frameworks similar to Lai et al. (2026), with the aim of evaluating its potential for improving model training, especially in low-resource languages where data quality remains a critical challenge.

9 Limitations

A key limitation of this work lies in the data we used for evaluation of unsupervised anomaly detection: controlling the contamination rate requires a fully controlled setup, which in practice led us to rely on artificial contamination. Another limitation is the restricted range of embedding models we considered in our experiments: our study focused on a limited set of open-source models, which could be run under limited computational resources. More broadly, this work focuses on two-stage anomaly detection frameworks and primarily investigates the detection component itself. Exploring the behavior of other approaches, including language model fine-tuning or in-context learning strategies, constitute a valuable research direction.

10 Acknowledgements

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-23-CE23-0027 (project CFTeXAD).

References

- Maha Tufail Agro and Hanan Aldarmaki. 2023. [Handling realistic label noise in BERT text classification](#). In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (IC-NLSP 2023)*, pages 11–20, Online. Association for Computational Linguistics.
- Tiago A. Almeida, José María Gómez Hidalgo, and Akebo Yamakami. 2011. [Contributions to the study of sms spam filtering: new collection and results](#). In *ACM Symposium on Document Engineering*.
- Fabrizio Angiulli and Clara Pizzuti. 2002. Fast outlier detection in high dimensional spaces. In *European conference on principles of data mining and knowledge discovery*, pages 15–27. Springer.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. [Qwen technical report](#). Preprint, arXiv:2309.16609.
- Dor Bank, Noam Koenigstein, and Raja Giryes. 2023. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374.
- Matei Bejan, Andrei Manolache, and Marius Popescu. 2023. Ad-nlp: A benchmark for anomaly detection in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10766–10778.
- Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104.
- Xingyu Cai, Jiaji Huang, Yu-Lan Bian, and Kenneth Ward Church. 2021. [Isotropy in the contextual embedding space: Clusters and manifolds](#). In *International Conference on Learning Representations*.
- Yang Cao, Sikun Yang, Chen Li, Haolong Xiang, Lianyong Qi, Bo Liu, Rongsheng Li, and Ming Liu. 2025. Tad-bench: A comprehensive benchmark for embedding-based text anomaly detection. *arXiv preprint arXiv:2501.11960*.
- Yonghe Chu, Heling Cao, Yufeng Diao, and Hongfei Lin. 2023. [Refined sbert: Representing sentence bert in manifold space](#). *Neurocomput.*, 555(C).
- J.A. Cuesta-Albertos and A. Nieto-Reyes. 2008. The random tukey depth. *Computational Statistics and Data Analysis*, 52(11):4979–4988.
- Vin de Silva and Joshua B. Tenenbaum. 2002. [Global versus local methods in nonlinear dimensionality reduction](#). In *Neural Information Processing Systems*.
- David L. Donoho and Miriam Gasko. 1992a. [Break-down properties of location estimates based on half-space depth and projected outlyingness](#). *The Annals of Statistics*, 20(4):1803–1827.
- David L. Donoho and Miriam Gasko. 1992b. [Break-down properties of location estimates based on half-space depth and projected outlyingness](#). *The Annals of Statistics*, 20(4):1803–1827.
- Rainer Dyckerhoff, Pavlo Mozharovskiy, and Stanislav Nagy. 2021. [Approximate computation of projection depths](#). *Computational Statistics and Data Analysis*, 157:107166.
- Kawin Ethayarajh. 2019. 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 and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65.
- Pierre Fihey, Matthieu Labeau, and Pavlo Mozharovskiy. 2026. [Enhancing two steps textual anomaly detection through anisotropy mitigation](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28442–28464, San Diego, California, United States. Association for Computational Linguistics.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1119.
- Adam Goodge, Bryan Hooi, See-Kiong Ng, and Wee Siong Ng. 2022. Lunar: Unifying local outlier detection methods via graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6737–6745.
- Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. [Exploring anisotropy and outliers in multilingual language models for cross-lingual semantic sentence similarity](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7023–7037, Toronto, Canada. Association for Computational Linguistics.
- Rishi Jha and Kai Mihata. 2021. [On geodesic distances and contextual embedding compression for text classification](#). In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 144–149, Mexico City, Mexico. Association for Computational Linguistics.

- Wen Lai, Yingli Shen, Shuo Wang, Xueren Zhang, Kangyang Luo, Alexander Fraser, and Maosong Sun. 2026. [DCAD-2000: A multilingual dataset across 2000+ languages with data cleaning as anomaly detection](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Charles Eric Leiserson, Ronald L Rivest, Thomas H Cormen, and Clifford Stein. 1994. *Introduction to algorithms*, volume 3. MIT press Cambridge, MA, USA.
- Leonardo Leone, Pavlo Mozharovskiy, and David Bounie. 2025. [Massive parallelization of projection-based depths](#). Working paper or preprint.
- Yuangang Li, Jiaqi Li, Zhuo Xiao, Tiankai Yang, Yi Nian, Xiyang Hu, and Yue Zhao. 2024. Nlp-adbench: Nlp anomaly detection benchmark. *arXiv preprint arXiv:2412.04784*.
- Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. 2020. [Copod: Copula-based outlier detection](#). *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1118–1123.
- Yifan Liu, Qianfeng Wen, Mark Zhao, Jiazhou Liang, and Scott Sanner. 2025. [MA-DPR: Manifold-aware distance metrics for dense passage retrieval](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31085–31103, Suzhou, China. Association for Computational Linguistics.
- Larry M Manevitz and Malik Yousef. 2001. One-class svms for document classification. *Journal of machine Learning research*, 2(Dec):139–154.
- Andrei Manolache, Florin Brad, and Elena Burceanu. 2021. [DATE: Detecting anomalies in text via self-supervision of transformers](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 267–277, Online. Association for Computational Linguistics.
- Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. 2006. [Spam filtering with naive bayes - which naive bayes?](#) In *International Conference on Email and Anti-Spam*.
- Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. 2025. [The origins of representation manifolds in large language models](#). *Preprint*, arXiv:2505.18235.
- Karl Mosler and Pavlo Mozharovskiy. 2020. [Choosing among notions of multivariate depth statistics](#). *Statistical Science*.
- Jeremie Pantin and Christophe Marsala. 2024. [A robust autoencoder ensemble-based approach for anomaly detection in text](#). *Preprint*, arXiv:2405.13031.
- Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. 2021. [A unifying review of deep and shallow anomaly detection](#). *Proceedings of the IEEE*, 109(5):756–795.
- Lukas Ruff, Yury Zemlyanskiy, Robert Vandermeulen, Thomas Schnake, and Marius Kloft. 2019. [Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, Florence, Italy. Association for Computational Linguistics.
- Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, and Milica Gašić. 2025. [Less is more: Local intrinsic dimensions of contextual language models](#). *Preprint*, arXiv:2506.01034. To appear in NeurIPS 2025.
- Benjamin Schweinhart. 2021. Persistent homology and the upper box dimension. *Discrete & Computational Geometry*, 65(2):331–364.
- Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. 2000. [A global geometric framework for nonlinear dimensionality reduction](#). *Science*, 290(5500):2319–2323.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- John W Tukey. 1975. Mathematics and the picturing of data. In *Proceedings of the international congress of mathematicians*, volume 2, pages 523–531. Vancouver.
- Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio ansuini, and Alberto Cazzaniga. 2023. [The geometry of hidden representations of large transformer models](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. [Text embeddings by weakly-supervised contrastive pre-training](#). *Preprint*, arXiv:2212.03533.
- Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. 2022. [Deep isolation forest for anomaly detection](#). *IEEE Transactions on Knowledge and Data Engineering*, 35:12591–12604.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren

- Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.
- Yijun Zuo. 2003a. [Projection-based depth functions and associated medians](#). *Annals of Statistics*, 31:1460–1490.
- Yijun Zuo. 2003b. [Projection-based depth functions and associated medians](#). *Annals of Statistics*, 31:1460–1490.
- Yijun Zuo and Robert Serfling. 2000. [General notions of statistical depth function](#). *The Annals of Statistics*, 28(2):461–482.

A Appendix

A.1 Datasets Statistics

Table 3 summarizes the datasets used in our experiments. The diversity of these datasets enables a comprehensive evaluation across a wide range of settings. In particular, our methods are assessed on both large-scale and substantially smaller datasets, as well as on textual sequences ranging from short news headlines to considerably longer documents.

Dataset	# Classes	Avg. samples	Avg. tokens
BBCNews	5	445	509
N24News	5	3000	72
AGNews	4	31900	53
20NG	7	2692	316

Table 3: Dataset statistics, reporting the number of classes, average number of samples per class, and average number of tokens per sample.

A.2 Models Selection

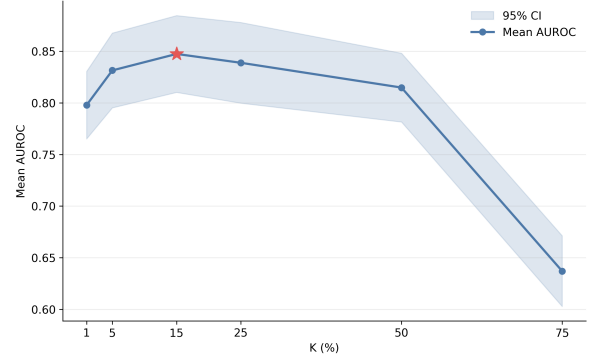
Table 4 details the Hugging Face checkpoints and of the models considered. While focusing on recent models with fewer than 1B parameters, we selected both similarity-trained models and standard LLMs to enable a comprehensive evaluation

Model name	Model	# Params	Release
e5	e5-base-v2	0.3B	2022
Qwen3-Emb.	Qwen3-Embedding-0.6B	0.6B	2025
LLaMA	Llama-3.2-1B	1B	2024
Qwen3	Qwen3-0.6B	0.6B	2025

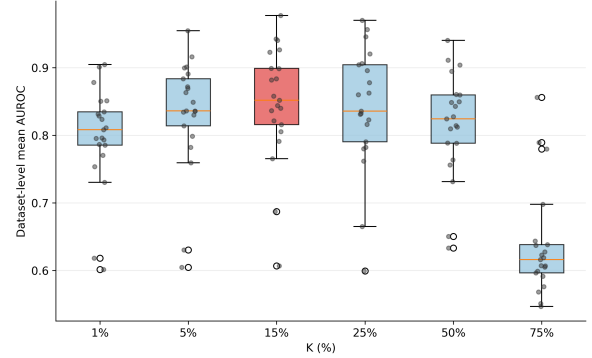
Table 4: Summary of the models used in our experiments.

A.3 Experimental Setup

Since most embedding models and anomaly detection methods considered in our experiments are deterministic, performance is evaluated using a single run. Repeating experiments would moreover incur substantial computational costs, as the benchmark combines 8 detectors, 4 embedding models, 21 ID/OOD configurations, and two anomaly detection settings (on separate test data and within contaminated training data).



(a) Overall mean AUROC.



(b) Dataset-level AUROC distribution.

Figure 6: Geodesic Depth performance across neighborhood sizes K , expressed as a percentage of the training set size n , aggregated across models and datasets.

A.4 Geodesic Depth Sensitivity to Neighborhood Size

Figure 6 shows the empirical analysis of Geodesic Depth performance across neighborhood sizes K . Performance decreases for very small ($< 10\%$) and large ($> 50\%$) neighborhoods, while remaining relatively stable in between and peaking at $K = 15\%$, motivating our choice of this value. This trend is consistent across the evaluated datasets, suggesting that $K = 15\%$ provides a robust default across settings.

A.5 Computational Time

Table 5 reports the computational time of the different anomaly detectors across datasets. While exact Geodesic Depth is substantially more computationally demanding, it remains practical for relatively small datasets such as BBC News, N24 News, and 20NG. Overall, the landmark approximation provides competitive computational times. Interestingly, while this approximation can affect detection performance on very small datasets, it achieves performance comparable to exact compu-

Detector	BBC News	N24 News	AGNews	20NG
Geodesic Depth (Exact)	1.45	157.88	—	135.30
Geodesic Depth (Landmarks)	0.14	6.72	360.32	5.15
Projection Depth	3.24	48.96	1775.84	37.68
LUNAR	1.79	4.11	82.82	3.52
AutoEncoder	0.48	4.26	27.57	3.21
GMM	0.87	1.74	7.80	1.54
KNN	0.28	17.51	845.39	13.27
COPOD	0.30	4.41	35.85	3.31

Table 5: Computational time (in seconds) of the different anomaly detectors across datasets.

tation on larger datasets.

A.6 Detection within the contaminated training data

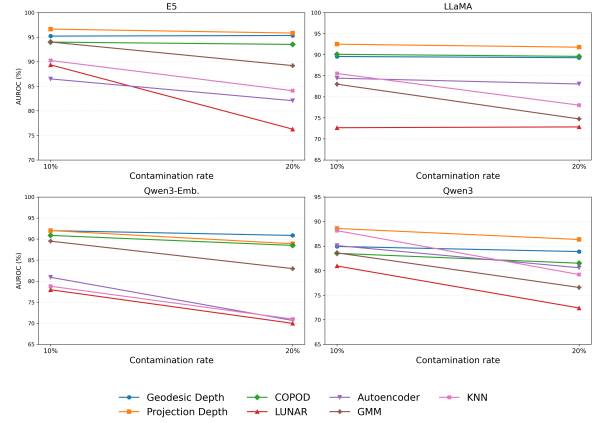
Figure 8 presents the results of our experiments on a variant of classical AD, where anomalies must be identified directly within the contaminated training set. As expected, these results are largely consistent with those obtained for classical AD. In particular, depth-based methods appear significantly more robust than classical detection approaches, and the geodesic depth seems better suited to embeddings produced by large language models. However, we observe that the GMM model is even better suited to LLaMA embeddings in this setting, outperforming our methods on the 20NG and N24 News datasets.

A.6.1 Evaluation on Realistic Datasets.

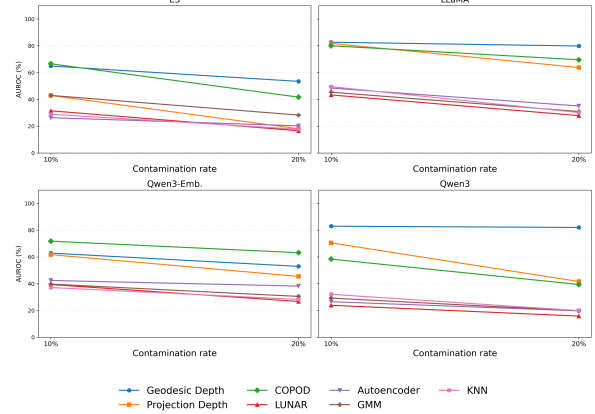
Requiring fewer available samples, this setting allows us to evaluate our methods on the Email Spam (Metsis et al., 2006) and SMS Spam (Almeida et al., 2011) datasets, two more realistic benchmarks commonly considered in the anomaly detection literature. Due to the limited number of available outliers, these datasets could not be included in our initial setting, and we restrict our evaluation to contamination rates of 10% and 20%. Across both datasets, Depth functions tend to outperform competing anomaly detection methods, showing that their benefits extend beyond topic-based benchmarks. We also conduct preliminary experiments on IMDB. However, all evaluated methods perform close to random guessing (around 50% AUROC), suggesting that the pretrained embeddings considered here are not well suited to this setting.

A.7 Extended Results on Additional AD Methods

To provide a more comprehensive evaluation across the anomaly detection algorithms commonly used



(a) Email spam



(b) SMS spam

Figure 7: AD within contaminated training data : Mean AUROC across contamination levels.

in recent benchmarks, we additionally report in Figure 9 and 10 the performance of LOF (Breunig et al., 2000) and OC-SVM (Manevitz and Yousef, 2001), which are known to achieve strong results in the semi-supervised setting. As shown in the figure, these methods (especially LOF) are highly sensitive to contamination in the training data, with performance degrading as the contamination rate increases.

A.8 Ablation Study: Impact of Geodesic Distance on L2 Depth

Figures 11 and 12 present an ablation study comparing the classical L2 depth with the proposed geodesic depth, in which Euclidean distances are replaced by geodesic distances and aggregation is performed using the median. We observe that the proposed method consistently improves L2 depth. For embeddings produced by similarity-trained models, projection depth appears better suited than the standard L2 depth, yet our method outperforms it

in the vast majority of cases. In contrast, for LLM embeddings, the performance of projection depth drops below that of standard L2 depth, suggesting that distance-based depth mechanisms are better suited to this type of data. Our proposed adaptation further widens this gap. While the comparative advantage of geodesic depth is already visible in the absence of contamination, it becomes more pronounced as the contamination rate increases, suggesting that the use of geodesic distances leads to improved robustness.

A.9 Anisotropy scores

Model	20NG	AGNews	BBC	N24
E5	0.78	0.75	0.77	0.75
LLaMA	0.82	0.75	0.88	0.80
Qwen3-Emb.	0.27	0.27	0.30	0.28
Qwen3	0.81	0.85	0.94	0.88

Table 6: Anisotropy (mean cosine similarity between random pairs) across datasets and embedding models.

Anisotropy scores are computed following the definition of [Hämmerl et al. \(2023\)](#). Let S be a set of n random sentence pairs sampled from a corpus D . The anisotropy score of an encoder f is defined by:

$$A(f) = \frac{1}{n} \sum_{\{x,y\} \in S} \cos(f(x), f(y)) \quad (4)$$

Table 6 highlights the impact of contrastive training on embedding geometry, showing that models such as E5 and Qwen3-Emb. produce substantially more isotropic representations.

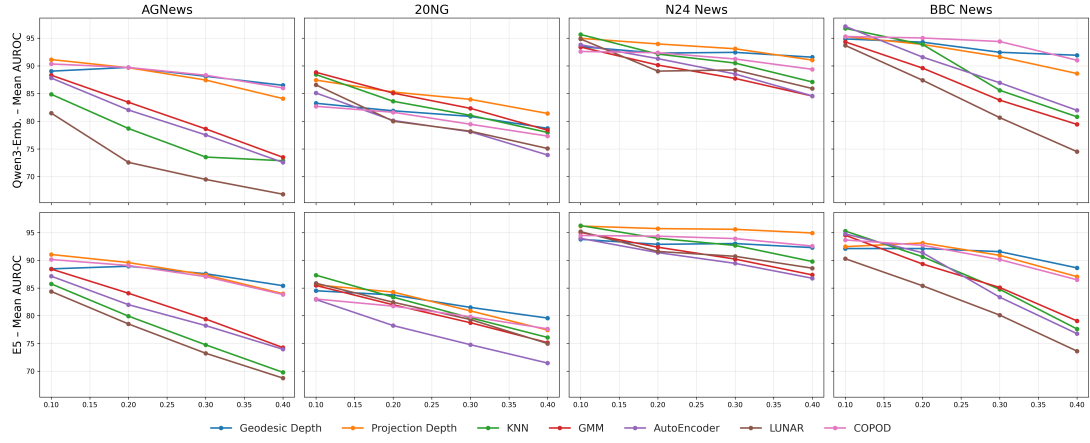
A.10 Manifolds Visualization Procedure

In this section, we detail the visualization procedure used to explore the manifold structure of embedding spaces. We first perform PCA, as the directions explaining the most variance provide a natural basis for investigating the global geometry of the embeddings. Since the manifold may not align with the leading components, we consider multiple planes by pairing one principal direction with others. For each plane spanned by two principal directions u_i and u_j , we project each point $x \in \mathbb{R}^d$ onto the plane as

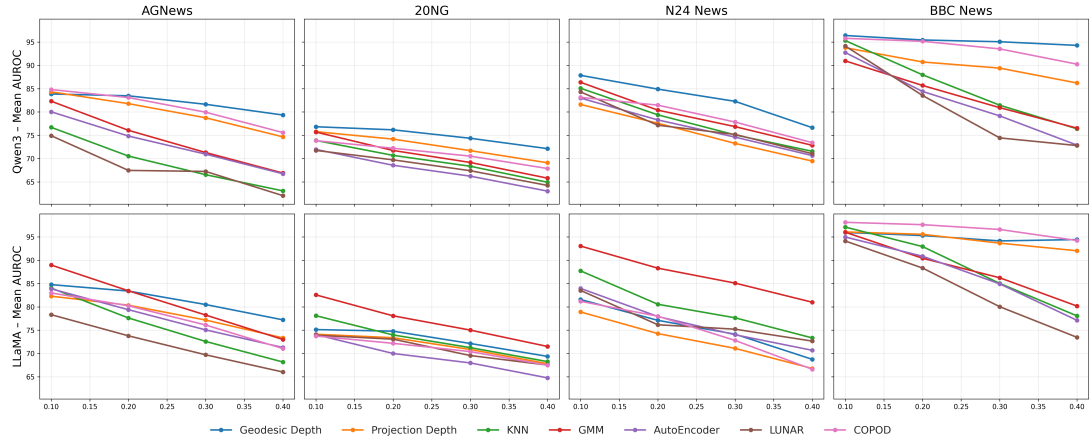
$$\Pi(x) = (x^\top u_i)u_i + (x^\top u_j)u_j$$

and compute its orthogonal distance $\|x - \Pi(x)\|_2$. Points far from the plane are strongly

distorted, as much of their structure lies outside the 2D view. We therefore retain only the closest points, for which the projection is most meaningful, making the underlying structures easier to observe.



(a) Similarity-trained embeddings



(b) Large language model embeddings

Figure 8: AD within contaminated training data : Mean AUROC across contamination levels. (a) Results for similarity-trained embeddings. (b) Results for large language models embeddings. Only the strongest baseline detectors are included for clarity. Depth-based methods outperform baselines once the training set contamination exceeds 10%.

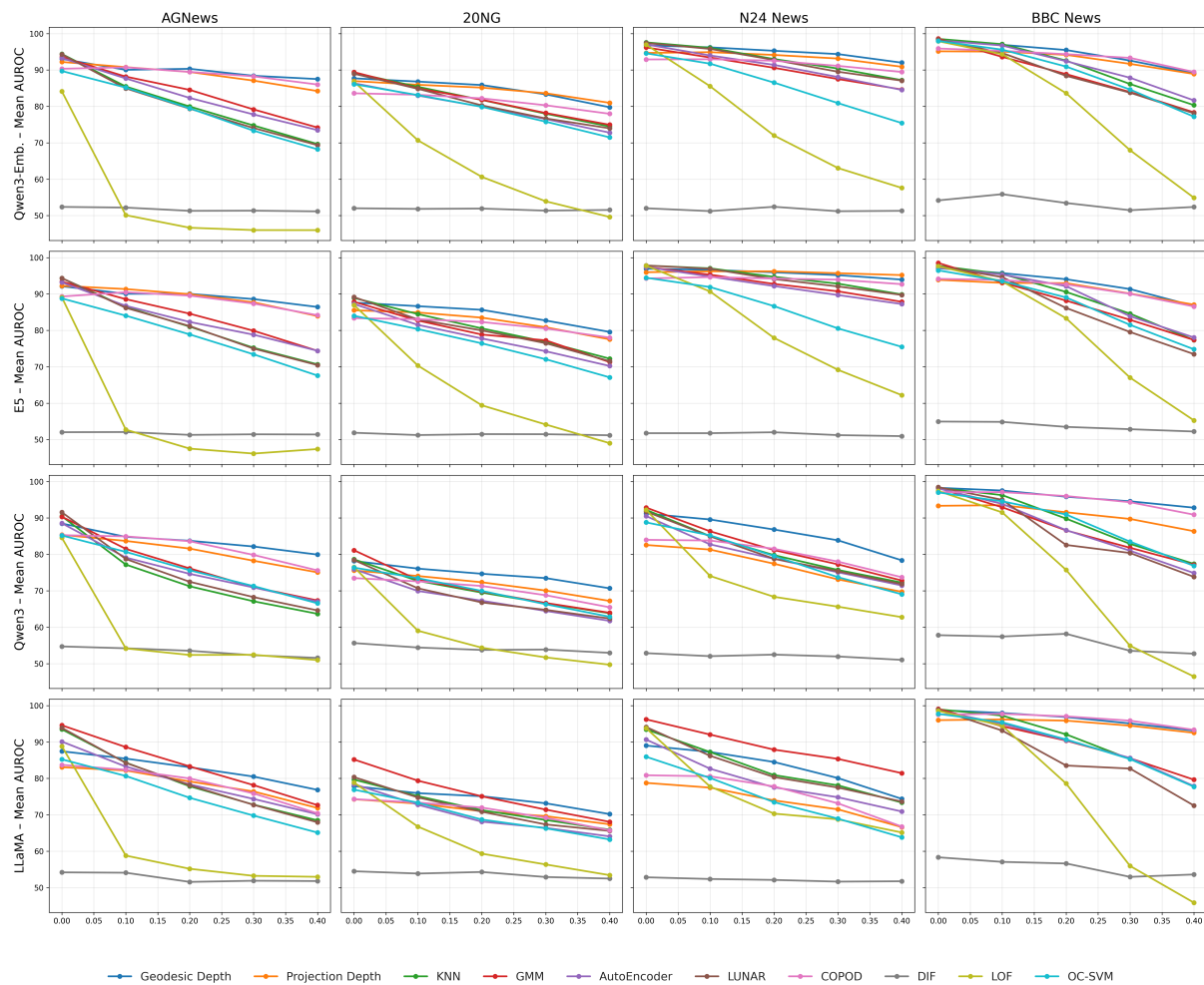


Figure 9: Extended results across a comprehensive set of anomaly detection algorithms considered in recent two-stage anomaly detection benchmarks. Mean AUROC values are shown on separate test data.

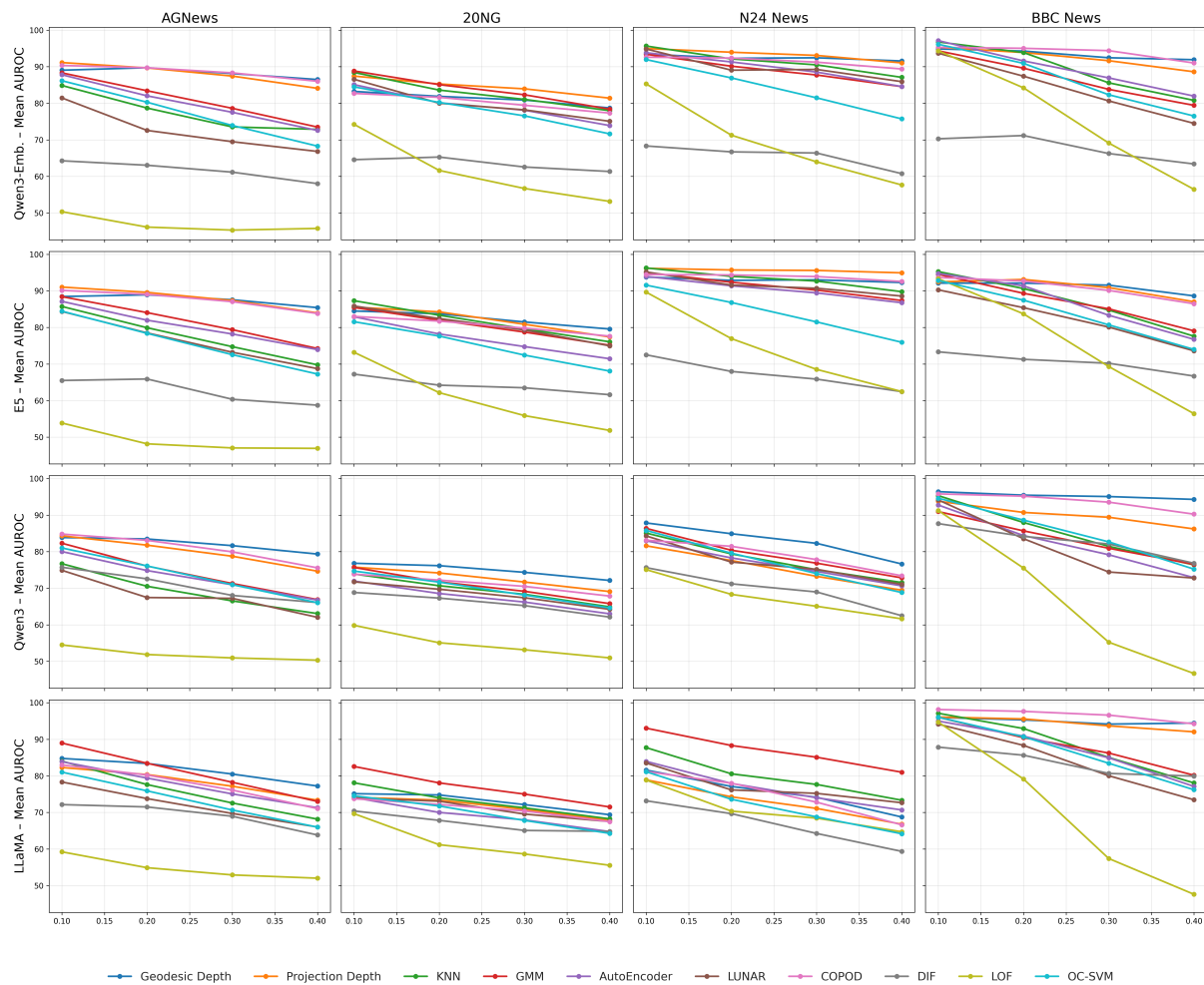


Figure 10: Extended results across a comprehensive set of anomaly detection algorithms considered in recent two-stage anomaly detection benchmarks. Mean AUROC values are shown on contaminated training data.

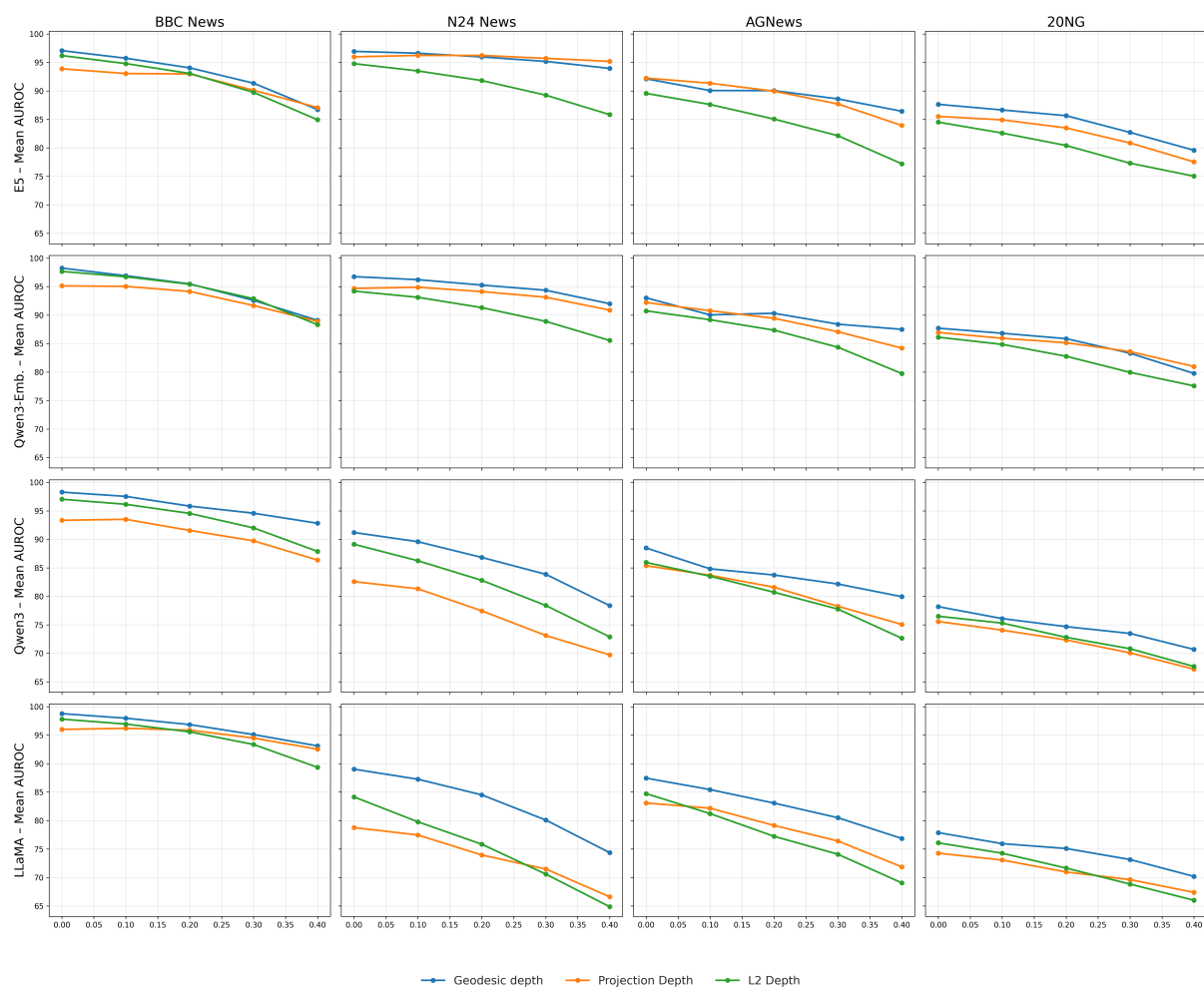


Figure 11: Ablation study comparing L2 depth and the proposed geodesic depth across embedding models, with projection depth included for a broader comparison of depth-based methods. Mean AUROC values are shown for anomaly detection on separate test data.

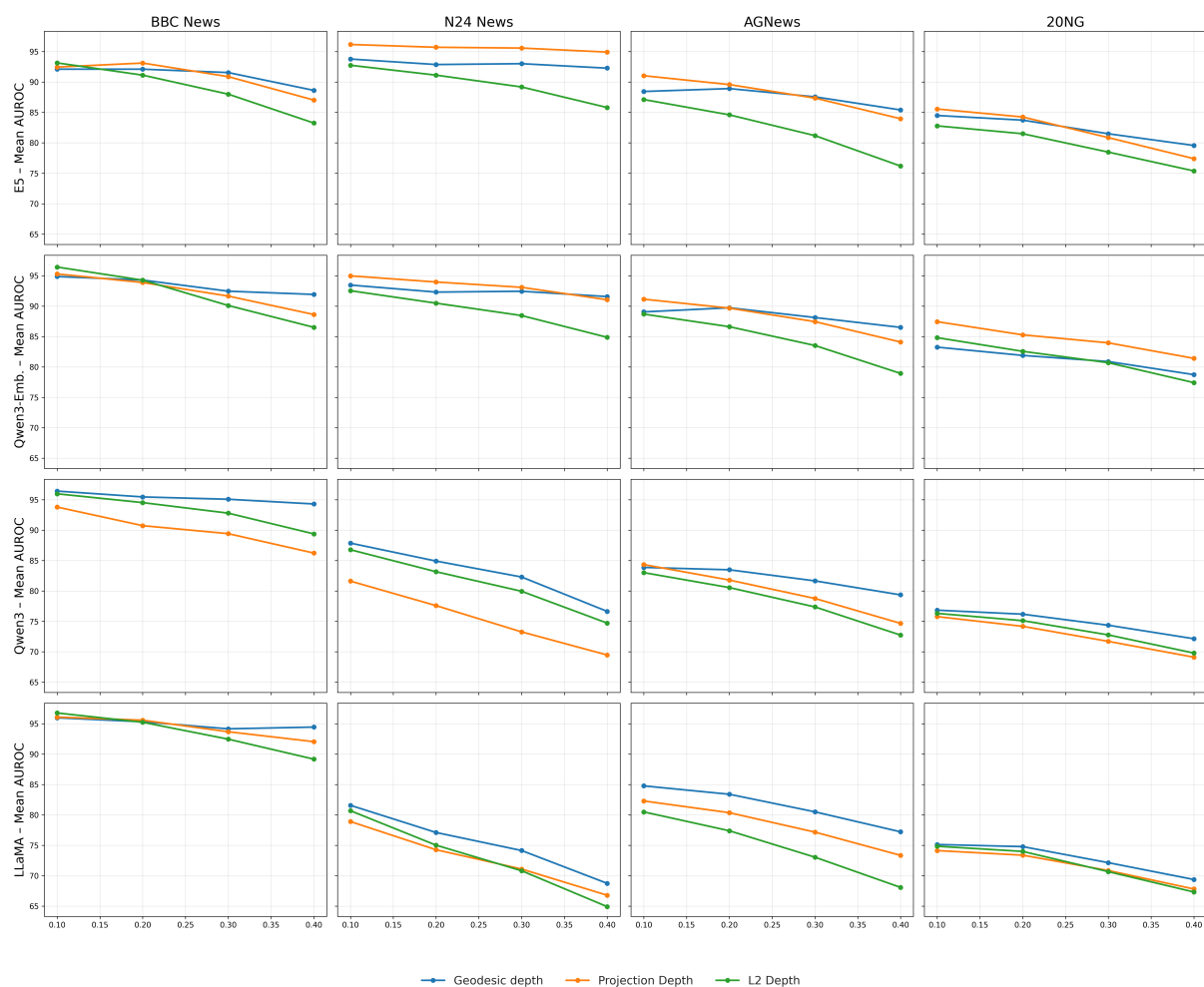


Figure 12: Ablation study comparing L2 depth and the proposed geodesic depth across embedding models, with projection depth included for a broader comparison of depth-based methods. Mean AUROC values are shown for anomaly detection within contaminated training data.