Enhancing Two-Step Textual Anomaly Detection through Anisotropy Mitigation

Pierre Fihey, Matthieu Labeau, Pavlo Mozharovskyi
LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
{pierre.fihey, matthieu.labeau, pavlo.mozharovskyi}@telecom-paris.fr

Abstract

Anomaly detection aims at distinguishing between *in-distribution* samples, which belong to the same distribution as the training set, and *out-of-distribution* samples, which lie outside of it. In textual anomaly detection, recent approaches routinely apply anomaly detection algorithms directly to embeddings extracted from pre-trained embedding models (*two-stage approaches*). However, the geometric properties of pre-trained embeddings can hinder the effectiveness of detection algorithms, which often rely on distance-based measures. In this work, we first highlight the relevance of similarity-trained models for textual anomaly detection. Beyond being trained to capture semantic similarities, these models also exhibit geometric properties that appear better suited to detection algorithms. We further demonstrate that, besides model choice, a simple post-processing step can significantly improve anomaly detection by adapting embeddings to the assumptions made by classical detection algorithms. The bulk of our experiments is done on a reformulation of the classification tasks from the MTEB benchmark into anomaly detection tasks[1].

1 Introduction

While the literature has struggled to formally define what constitutes an anomaly in NLP (Arora et al., 2021), two main settings are typically considered when conducting experiments with text: (1) task-specific scenarios, such as intent classification, where anomalies are unseen intents (Casanueva et al., 2020), or binary classification tasks like hate speech and fake news detection (Cao et al., 2025) ; and (2) ad-hoc artificial setups, where anomalies are created from annotated classification datasets by sampling examples from other classes (Manolache et al., 2021; Ruff et al., 2019). Throughout this paper, we refer to both under the generic term of Anomaly Detection.

For textual anomaly detection, standalone models were initially developed (Manolache et al., 2021; Ruff et al., 2019), achieving competitive results, but recent approaches have mainly investigated *two-stage frameworks*, applying anomaly detection algorithms on embeddings derived from pre-trained language models, thereby leveraging their large-scale training data. Recently published benchmarks (Cao et al., 2025; Bejan et al., 2023; Li et al., 2024) provide a comprehensive study by evaluating several anomaly detection algorithms on various embeddings models, from early architectures like BERT (Devlin et al., 2019) to more recent large language models (LLMs). While highlighting the strong performance of LLMs on this task, these works also report highly heterogeneous results across different combinations of anomaly detection algorithms and embedding models. These findings highlight the necessity of carefully selecting the detection algorithm according to a given setting, which challenges the generalizability of such methods.

In particular, one of the main challenges in applying such detection algorithms to pre-trained embeddings lies in their frequent reliance on distance-based measures. Indeed, studies have shown that embeddings derived from pre-trained language models exhibit strong anisotropy (Ethayarajh, 2019), which can severely compromise distance computations and, consequently, the performance of such methods. Several studies have further demonstrated that this anisotropic geometry leads contextual models such as BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021) to perform poorly on tasks like semantic similarity, which inherently rely on distance computations between sequences. We hypothesize that this property should affect anomaly detection in a similar way. Following this intuition, we argue that sentence

[1]Our code is available at https://github.com/Pierrefi/Two_steps_AD

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28442–28464

embedding models, trained to capture semantic similarity, are better suited for anomaly detection, as they produce representations that are both semantically meaningful and more isotropic (Hämmerl et al., 2023). In this work, we look at how the choice of an embedding model may influence anomaly detection performance, especially investigating similarity-trained models. Our intent is to evaluate detection algorithms in diverse settings in order to draw practical recommendations. In particular, we investigate recent multilingual models (Zhang et al., 2025; Wang et al., 2024) which have been trained with semantic similarity objectives. In parallel, numerous post-processing methods (Hämmerl et al., 2023) have been introduced in prior works to improve the isotropy of embedding representations. We therefore investigate whether these methods can be leveraged to further improve anomaly detection performance. Our contributions are summarized as follow :

- We adapt diverse classification datasets from the MTEB benchmark for anomaly detection, enabling **large-scale multilingual and multi-domain evaluation**.
- We demonstrate the comparative advantage of **similarity-trained embedding models** for anomaly detection, linking it to geometric properties beneficial to detection algorithms.
- We demonstrate that a simple **post-processing step** can seemingly adapt embeddings to be used with most detection algorithms, greatly smoothing their variance in performance.

2 Framework Definition

Given a training set of text samples $D_n = \{x_i\}_{i=1}^n$, the anomaly detection task consists in learning a decision function $h : \mathcal{X} \to \{0, 1\}$ indicating whether a new sample $\mathbf{x}$ is *in-distribution* (0) or *anomalous* (1) with respect to D_n.

Two Steps Anomaly Detection In the two-steps framework, a **pre-trained encoder** $f : \mathcal{X} \to \mathbb{R}^d$ is first used to map input sequences x to latent representations $\mathbf{z} = f(x)$. These representations are then passed to an **anomaly detection algorithm** that assigns a score $s(\mathbf{z})$, where higher scores indicate more anomalous inputs. A threshold ϵ is used to determine whether a sequence is considered anomalous, defining the decision function h as $h(x) = \mathbb{I}(s(f(x)) > \epsilon)$, where $\mathbb{I}(\cdot)$ denotes the indicator function. Threshold selection in anomaly detection is a difficult problem in itself (Khosla and Gangadharaiah, 2022), and following other works on the task, we carry out threshold-independent evaluation using anomaly scores.

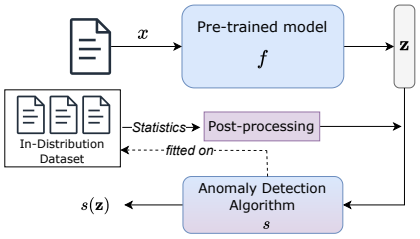

Figure 1: The standard two-stage framework. Contrarily to standalone models (Manolache et al., 2021; Ruff et al., 2019), only the detection method is fitted on *in-distribution* data.

3 Related Work

3.1 Anomaly detection tasks

Since very few datasets are explicitly designed for textual anomaly detection, most prior work relies on repurposed text classification datasets. Early studies (Ruff et al., 2019; Manolache et al., 2021) leveraged multi-class text classification datasets (e.g., AGNews, 20NewsGroup...) by reformulating them as **one-class classification** tasks, where one class defines the normal data and the remaining classes are treated as anomalies. More recent benchmarks (Li et al., 2024; Cao et al., 2025) have further incorporated classification datasets targeting more specific forms of anomalies, including binary classification tasks such as spam, hate speech, and fake news detection, as well as multi-class sentiment classification datasets, which are adapted by treating the most negative sentiment category as anomalous. Despite this diversity of datasets, existing work evaluates anomaly detection models almost exclusively on English data. Yet, multilingual evaluation remains essential for assessing anomaly detection methods.

3.2 Embedding models

The development of embedding models with strong semantic modeling capabilities has led researchers to focus on two-stage anomaly detection approaches. Recent benchmarks (Li et al., 2024; Cao et al., 2025) systematically evaluate anomaly detection methods applied to transformer-based embeddings, ranging from classical encoders such as

BERT (Devlin et al., 2019) and MiniLM (Wang et al., 2020) to representations derived from recent LLMs, including LLaMA (Touvron et al., 2023), Qwen (Bai et al., 2023), and OpenAI models. These studies consistently show that LLM-based embeddings outperform earlier transformer-based approaches as well as strong standalone anomaly detection methods such as DATE (Manolache et al., 2021). However, these results also suggest potential limitations in the suitability of language model embeddings for anomaly detection. In particular, many detection algorithms that are highly effective on tabular data struggle when applied to textual embedding spaces. This observation motivates our hypothesis that intrinsic properties of language model embeddings critically influence anomaly detection performance.

Anisotropy in Embedding Space Prior work has shown that embeddings exhibit strong anisotropy, both for contextualized word representations (Ethayarajh, 2019) and sentence representations (Li et al., 2020a), which significantly degrades downstream performance, especially in semantic similarity and retrieval tasks. To address this issue, several post-hoc methods have been proposed to mitigate embedding anisotropy. Li et al. (2020b) introduce BERT-flow, which projects embeddings into a latent space following a standard Gaussian distribution. Huang et al. (2021) and Su et al. (2021) argue that a simple whitening transformation is sufficient to address anisotropy, while Liang et al. (2021) and Rajaee and Pilehvar (2021) demonstrate that the dominant directions of BERT embeddings induce anisotropy, and propose to remove them in the embedding space. More recently, Hämmerl et al. (2023) systematically evaluated the impact of these post-hoc normalization methods on semantic similarity tasks across multiple languages, and showed that they consistently lead to improved performance in this setting. Most importantly, this study highlights the comparative advantage of Sentence-BERT (Reimers and Gurevych, 2019) in this setting, producing more isotropic representations due to its contrastive learning-based training (Gao et al., 2021). We argue that the embedding geometry induced by contrastive learning better matches the assumptions of anomaly detection methods, and that post-processing techniques can further adapt these representations to enhance detection performance. This motivation is further strengthened by the recent proliferation of embedding models trained with contrastive learning objectives, which achieve state-of-the-art performance across a wide range of NLP tasks.

4 Experimental settings

4.1 Models

To ensure easy reproducibility under limited computational resources we restrict our experiments to models that can be run on a single NVIDIA V100 GPU. Consequently, we limit our comparison to language models with fewer than one billion parameters. Given that our objective is to understand the specific impact of contrastive learning objectives on anomaly detection performances, we include standard pre-trained language models that are commonly used as backbones for similarity-trained models. We first include XLM-RoBERTa (Conneau et al., 2020) (**XLM-R**), a transformer-based encoder pretrained on large-scale multilingual corpora. We also focus on two multilingual LLMs that represent the state of the art among lightweight autoregressive language models: **Qwen3** (Yang et al., 2025a) and **LLaMA**-3.2-1B (Touvron et al., 2023). For these generative models, we follow common practice in the literature and simply use the hidden representations from the final transformer layer as embeddings[2].

4.1.1 Similarity-trained Language models

We consider two embedding models trained with similarity-based objectives. **E5** (Wang et al., 2024) is initialized from XLM-RoBERTa-base encoder and trained using a two-stage contrastive learning framework, with weakly supervised pretraining followed by fine-tuning on supervised sentence-pair datasets. Qwen3-Embeddings (Zhang et al., 2025) (**Qwen3-E**) follows a training strategy similar to E5 but differs primarily in the source and construction of weak supervision, leveraging large-scale synthetic similarity pairs generated by the Qwen foundation models themselves rather than an external corpora. Details about the models used in this work are given in appendix A.1.

4.1.2 Mitigating the anisotropy

Despite being more isotropic, embeddings from similarity-trained models still exhibit substantial

[2]While Colombo et al. (2022) showed that aggregating representations from multiple layers can benefit anomaly detection with BERT models, our preliminary experiments indicate that this approach is less suited to the sentence embedding models and LLMs considered here.

residual anisotropy. Following Hämmerl et al. (2023), we experiment with several post-processing methods discussed in Section 3.2, namely BERT-flow, dominant direction removal, and whitening, as well as two adaptations of the latter, soft-whitening (Diera et al., 2024) and PCA-whitening (Su et al., 2021). Our preliminary experiments show that standard whitening consistently provides the most effective adaptation in our experimental setting (see Section 5.4). We hence adopt this whitening transformation as a post-processing step to obtain centered, uncorrelated features with unit variance, and leave the other methods of anisotropy mitigation out of this paper. Given a set of in-distribution representations$\{\mathbf{z}_i\}_{i=1}^n$, the transformation T is defined as:

$$T(\mathbf{z}_i) = (\mathbf{z}_i - \boldsymbol{\mu})\mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^T,$$

where $\boldsymbol{\mu}$ denotes the feature-wise mean, and $\mathbf{V}$ and $\mathbf{D}$ correspond to the eigenvectors and eigenvalues of the covariance matrix, respectively.

4.2 Anomaly detection algorithms

Existing studies in textual anomaly detection typically evaluate a broad range of algorithms, which are fitted on the set of in-distribution representations $\{\mathbf{z}_i\}_{i=1}^n$ and assign an anomaly score s($\mathbf{z}$) to unseen samples. For this study, we select a representative set of diverse anomaly detection algorithms, based on very different detection methods. This includes two local methods: k-Nearest Neighbors (**KNN**) (Angiulli and Pizzuti, 2002) and Local Outlier Factor (**LOF**) (Breunig et al., 2000), which rely on the distances between points and their neighborhoods to assign an anomaly score. **Lunar** (Goodge et al., 2022) has been introduced recently to unify these two methods under a single framework based on graph neural networks, and to enhance them with trainable components. Isolation Forest (**IF**) (Liu et al., 2008) recursively partitions the data space using randomly selected features and identifies as anomalies the points that are isolated with fewer splits. Gaussian Mixture Models (**GMM**) model the ID distribution as a mixture of normal distributions and identify low-likelihood points as anomalies. Finally, One-class Support Vector Machine (**OC-SVM**) (Schölkopf et al., 1999) is formally a classification method which learns to separate training examples from the origin.

4.3 Data

To ensure fair comparability with prior work, we first conduct experiments on the datasets introduced by (Li et al., 2024) and subsequently used by (Yang et al., 2025b). These include news classification datasets (**AGNews**, **BBCNews** and **N24News**), spam detection datasets that are commonly employed in anomaly detection benchmarks (**SMS spam** and **Email spam**), as well as two sentiment classification datasets (**Movie Review** and **Yelp Review**). For multiclass classification datasets, we follow the same experimental setup as Li et al. (2024), where one class is selected as anomalous, the remaining ones define the in-distribution, and anomalies are subsampled to represent 10% of the test set. While these datasets provide a strong basis for comparison with prior work, they are limited in scope and largely restricted to English. To overcome these limitations, we propose to adapt the classification datasets from the **Massive Text Embedding Benchmark (MTEB)** (Muennighoff et al., 2023), to the anomaly detection setting. This benchmark, which consists of clean datasets derived from real-world applications, offers two key advantages for our study: it covers a wide range of classification tasks across diverse domains and spans a large number of languages. To enable a fair and consistent comparison across models, we select the eight languages that are officially supported by all evaluated language models : *English, French, Spanish, Portuguese, Italian, German, Thai and Hindi*.

Anomaly detection setting We organize the MTEB classification datasets available in the selected languages into four task-based categories and adapt them to the anomaly detection setting according to the semantic nature of each task. **Toxicity** detection aims to identify abusive, hateful, or harmful content. In this case, toxic samples are naturally treated as anomalies, while non-toxic content defines the in-distribution. **Intent** classification focuses on the recognition of user intents and scenarios in dialogue systems, whereas **Topic** classification assigns texts to high-level thematic categories. For both tasks, no class is inherently anomalous, and we therefore rely on a class-based partitioning strategy, treating two-thirds of the classes as in-distribution and the remaining one-third as out-of-distribution. Finally, **Sentiment** analysis captures the emotional polarity or subjective tone expressed in text. Here, samples associated with

most negative sentiment are considered anomalous, while most positive instances are the in-distribution data. By adapting MTEB classification datasets to a multilingual Anomaly Detection setting, our approach enables a large-scale and realistic evaluation of two-stage anomaly detection methods across a wide variety of tasks and languages. Further details regarding dataset statistics are provided in Appendix A.5.

4.4 Evaluation metrics

We evaluate the performance of our models using the **AUROC** (Area Under the ROC Curve), the standard metric for anomaly detection. It reflects the probability that a randomly chosen outlier is assigned a higher anomaly score than a randomly chosen inlier. AUROC is threshold-independent and robust to class imbalance, making it well-suited for our setting.

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	69.26	68.83	73.87	69.83	63.54	52.49	49.95
	Sent.	52.23	55.94	54.35	55.42	52.77	50.57	49.69
	Topic	68.53	71.25	75.31	71.75	70.87	52.51	50.96
	Toxicity	60.45	56.64	73.65	78.48	54.95	51.76	52.95
E5	Intent	**85.77**	**84.60**	**86.30**	**82.87**	**81.60**	**66.33**	**67.57**
	Sent.	**70.02**	**67.39**	**69.38**	**65.28**	**69.85**	**62.89**	**66.51**
	Topic	79.66	79.81	80.54	**79.71**	74.58	57.84	62.25
	Toxicity	**78.90**	**72.91**	**83.70**	**87.77**	**73.81**	**62.28**	**66.63**
Qwen3	Intent	73.50	73.60	77.09	75.47	65.32	54.15	52.99
	Sent.	52.30	55.95	53.33	55.20	52.27	50.42	51.75
	Topic	75.22	74.17	77.29	72.60	72.65	**60.31**	57.73
	Toxicity	64.20	58.55	74.34	83.26	57.22	54.29	52.17
Qwen3-E	Intent	83.81	81.54	84.20	82.80	79.12	62.90	64.55
	Sent.	64.09	61.04	63.17	61.88	62.58	56.64	61.26
	Topic	**83.13**	**82.89**	**83.17**	79.42	**76.64**	57.20	**62.41**
	Toxicity	74.40	69.76	75.86	85.93	68.39	61.97	65.83
LLaMA	Intent	74.88	72.33	77.65	77.88	66.09	54.57	53.31
	Sent.	53.22	57.19	54.47	56.59	53.50	50.23	50.08
	Topic	77.51	75.24	78.84	75.80	72.83	59.54	57.75
	Toxicity	67.96	62.86	77.93	85.39	62.14	55.71	55.29

Table 1: Mean AUROC across languages for each model and task, organised by detector. Bold values highlight the best-performing scores within each task–detector configuration.

5 Results

In this section, we present the results of our experiments, beginning with those conducted on the MTEB-derived datasets. For each anomaly detection algorithm, we rely on the default hyperparameter settings provided by the pyod library, in order to ensure a fair and reproducible evaluation across models and tasks. Table 1 reports the results of our experiments for each embedding model, averaged across languages and domains.

Preliminary observations For each language model, we first observe a substantial variability across anomaly detection algorithms. In particular, GMM and LUNAR consistently outperform alternative methods across domains, while OCSVM and Isolation Forest exhibit substantially lower performance and appear poorly suited to the geometric properties of textual representations.

5.1 Effectiveness of similarity-trained embedding models

Our experiments demonstrate that the E5 model, which is initialized from the XLM-RoBERTa-base encoder, significantly outperforms its underlying backbone. This result shows that the two-stage contrastive learning procedure used to train E5 on similarity objectives has a very positive effect on the anomaly detection task. The performance difference between the LLM-based models (LLaMA and Qwen3) and Qwen3-Embeddings shows a similar pattern. These results clearly illustrate the advantage of similarity-based training in the context of anomaly detection. While such training is designed to better separate semantically related and unrelated sentences, we argue that the impact of contrastive learning on the isotropy of the representation space also contributes to the superior performance observed. Indeed, contrastive learning has been shown to improve the isotropy of embedding spaces (Gao et al., 2021; Xiao et al., 2023; Hämmerl et al., 2023) and the average anisotropy scores for all models (Appendix A.2) further confirm these findings. However, a residual level of anisotropy remains in these embeddings and still influences the behavior of some anomaly detection methods, as evidenced by the strong variability across detectors, even for models trained with similarity-based objectives.

Multillingual analysis Although Table 1 offers a global overview of model performance, analyzing results at the language level allows us to assess the robustness and consistency of these trends across languages. Detailed results for each individual language are reported in appendix A.6. Figure 2 provides a visualization of the robustness of model performance across languages. We focus here on results obtained with the LUNAR detector, which consistently achieves the strongest performance across all evaluated language models. Similarity-trained models systematically rank among the top-performing approaches across the four studied

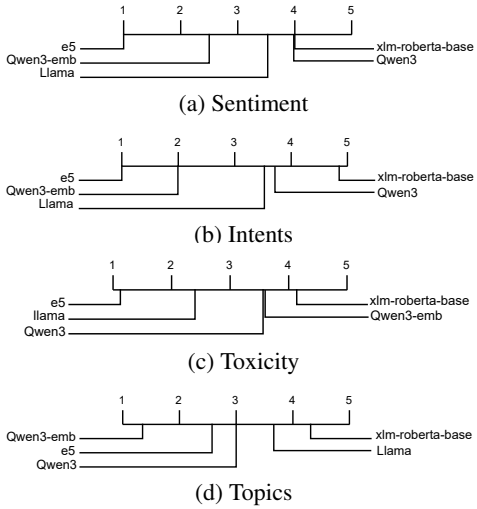

Figure 2: Average model ranks across languages using the LUNAR detector. Lower ranks indicate better performance.

domains and across languages, although Qwen3-embeddings appears less well suited to toxicity classification datasets for some languages. Overall, this further confirms the comparative advantage of similarity-trained models in an anomaly detection setting.

Domain-level analysis Figure 3 provides an overview of model performance across the different tasks, averaged over all languages. We first observe that these results corroborate our analysis across all domains. Beyond this global gain, the figure also reveals clear task-dependent differences: in particular, sentiment detection appears significantly more challenging for language models than intent, topic, or toxicity detection, even for similarity-trained representations. We further observe that the comparative advantage of models trained with similarity-based objectives is less pronounced on topic classification datasets, especially for the E5 model. This can be explained by the superior ability of large language models to process long input sequences compared to classical embedding models. Topic classification datasets, which are often derived from news articles, consist of substantially longer texts, as detailed in Appendix A.5.

More importantly, this figure allows us to assess the domain-specific variability of anomaly detection methods. While performance variations across detectors are similar for intent and topic classification, sentiment detection exhibits much more homogeneous behavior. This is likely due to its overall low performance, indicating poor separation between positive and negative sentiment embeddings, which makes anomaly detection intrinsically difficult and limits the impact of detector choice. In contrast, toxicity detection shows a significantly higher variability, with substantial performance gaps across anomaly detection algorithms.

5.2 Mitigation of anisotropy

The observed variability across anomaly detection algorithms indicates a strong dependence on embedding geometry. Following prior work on anisotropy and its impact on downstream tasks (Li et al., 2020a; Hämmerl et al., 2023), we apply a whitening transformation to the representations before anomaly detection. The results are reported in Table 2.

5.2.1 Overall impact of anisotropy mitigation

The obtained results reveal highly homogeneous performance across all anomaly detection algorithms. The performance gap between detectors disappears after applying the post-processing step, suggesting that the resulting embeddings are substantially better aligned with the requirements of the different detection methods. Figure 5 provides a more direct view of the observed effects.

While the strongest detectors (LUNAR and GMM) remain largely unaffected by this post-processing step, the other anomaly detection algorithms benefit substantially from it. These observations directly support our hypothesis that the strong performance discrepancies observed across anomaly detection algorithms are largely driven by the anisotropic structure of the embedding space: a simple whitening transformation being sufficient to produce homogeneous performance across all detectors indicates that several algorithms were indeed adversely affected by this geometric bias. The

Figure 3: Boxplots of model performance across domains, measured in AUROC and averaged over all languages. For each domain, the distributions correspond to the seven anomaly detection methods. Domains are presented in the same order as the legend.

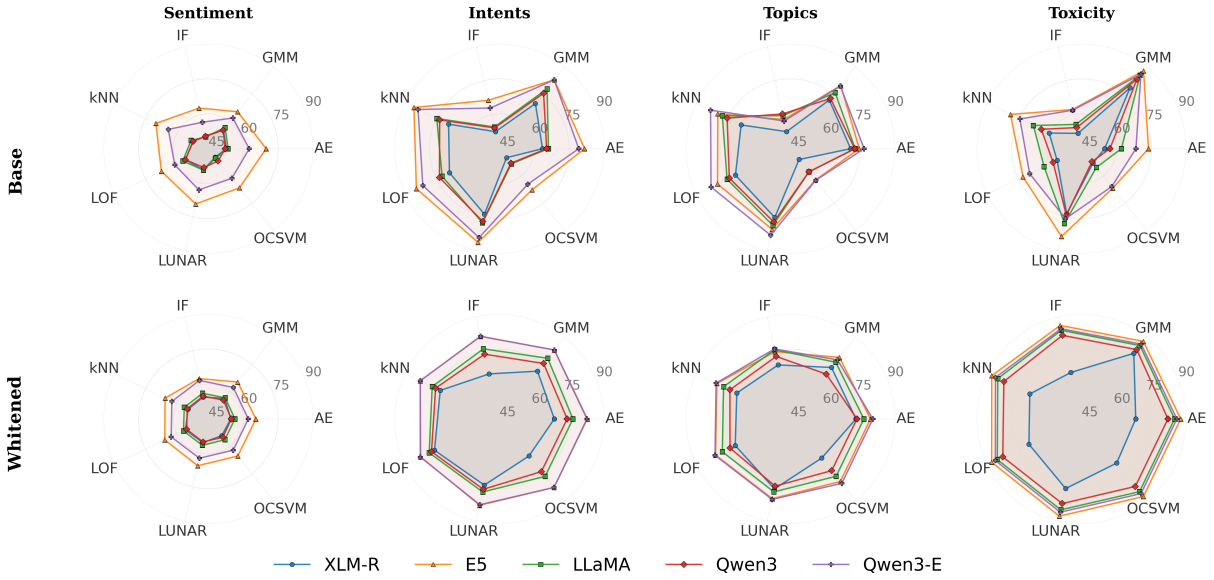

Figure 4: Radar plots showing, for each domain, the anomaly detection performance (AUROC) of each embedding model across detection algorithms, without any post-processing (*top row*) and with whitening applied (*bottom row*).

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	73.22	75.96	74.16	71.25	68.62	64.84	65.38
	Sent.	54.67	55.91	55.41	56.31	54.66	54.63	54.22
	Topic	70.60	71.35	75.83	73.27	73.16	68.80	66.51
	Toxicity	69.70	70.08	75.69	81.05	68.30	65.49	69.25
E5	Intent	82.76	82.56	**83.05**	**82.90**	**82.81**	81.35	**82.67**
	Sent.	**65.63**	**65.79**	**65.63**	**65.26**	**65.44**	**62.80**	**65.38**
	Topic	80.12	80.73	80.14	**78.80**	79.42	74.64	79.39
	Toxicity	**87.83**	**87.88**	**87.86**	**87.62**	**87.65**	**86.21**	**87.65**
Qwen3	Intent	75.44	77.39	76.00	75.47	74.15	73.54	73.94
	Sent.	55.07	55.32	55.16	55.36	54.92	55.07	55.05
	Topic	73.89	73.83	74.61	69.65	73.61	72.55	73.42
	Toxicity	82.05	82.57	82.31	83.08	82.15	81.84	82.09
Qwen3-E	Intent	**82.76**	**82.74**	82.82	82.86	82.65	**81.44**	82.64
	Sent.	62.39	62.73	62.32	62.26	62.18	62.00	62.16
	Topic	**80.46**	**81.00**	**80.46**	77.45	**80.42**	**75.84**	**80.41**
	Toxicity	86.06	86.13	86.10	86.12	85.98	84.75	86.03
LLaMA	Intent	77.05	78.49	77.19	78.33	76.62	75.89	76.55
	Sent.	56.47	56.86	56.62	56.64	56.43	56.30	56.42
	Topic	76.92	77.50	77.15	76.19	76.56	75.47	76.51
	Toxicity	84.97	85.19	85.05	85.31	85.00	84.06	84.99

Table 2: Mean AUROC with whitening across languages for each model and task, organised by detector. Bold values indicate the best-performing model for each detector–task pair.

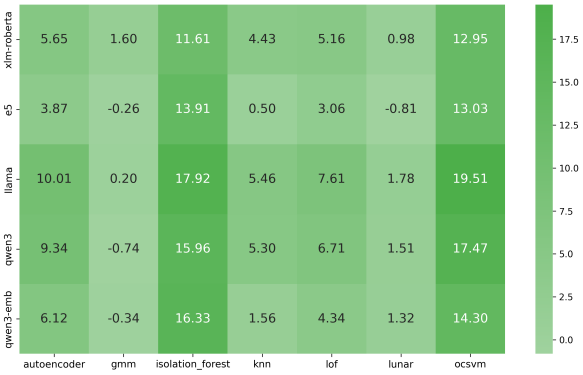

Figure 5: Heatmap of the impact of whitening on anomaly detection performance (Δ_{AUROC}), averaged across all languages and domains, for each combination of embedding model and detection algorithm.

fact that simple variance normalization does not achieve similar gains (see Appendix A.9) further supports this interpretation. Figure 4 illustrates the homogenization of performance induced by post-processing: across all models and domains, we consistently observe that whitening leads to markedly more homogeneous performance profiles. Beyond this overall stabilization, the figure also suggests that whitening substantially reduces the performance gaps between embedding models. By reducing anisotropy, whitening substantially decreases the dependence on the choice of a specific anomaly detection algorithm, thereby making the two-stage approach both simpler to deploy and more generalizable across settings.

5.2.2 Impact of anisotropy mitigation across domains

Figure 6 illustrates the impact of anisotropy mitigation across domains. As expected, whitening proves to be most effective for tasks that previously showed a large performance variability across anomaly detection methods. In particular, toxicity detection benefits substantially from isotropy-enhancing post-processing.

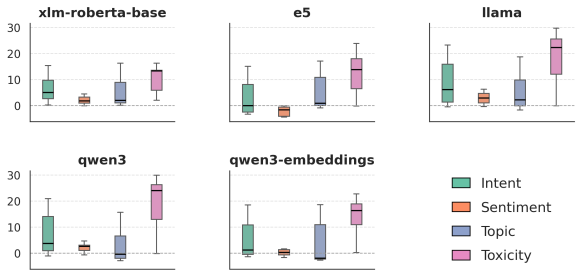

Figure 6: Boxplots of the whitening effect on model performance across domains, measured as the Δ_{AUROC} between base and whitened representations and averaged over all languages. For each domain, the distributions correspond to the seven anomaly detection methods. Domains are presented in the same order as the legend.

Model	AG	BBC	N24	Email	SMS	Movie	Yelp
XLM-R	82.86	95.09	85.28	90.15	68.90	50.47	49.09
E5	90.01	**97.52**	73.36	94.04	92.41	71.50	76.13
LLaMA	92.43	97.46	85.94	**97.36**	89.21	56.89	60.33
Qwen3	83.29	97.03	**88.39**	95.16	69.48	52.33	56.23
Qwen3-E	86.39	95.25	83.84	97.00	69.68	78.82	63.42
NLPAD†	92.26	97.32	83.20	96.97	**93.98**	73.66	**94.52**
AD-LLM†	**93.32**	95.74	82.07	-	87.97	**93.49**	-

Table 3: AUROC scores obtained with the LUNAR anomaly detector across datasets. † denotes values taken from the original benchmark papers and correspond to the best reported performance on each dataset.

We further observe that the whitening transformation generally has a smaller impact on the Qwen3-Embeddings model, which already produces substantially more isotropic embeddings than its LLM-based counterparts. In contrast, E5 benefits from whitening to a similar extent as its XLM-RoBERTa backbone, likely because the limited semantic expressiveness of the latter bounds performance gains even under improved isotropy.

5.3 Comparison with previous benchmarks

Table 3 reports the results of our experiments conducted on the datasets introduced by (Li et al., 2024). For clarity, we only present the results obtained with the LUNAR anomaly detection algorithm, which consistently achieved the best performance across settings. The performance trends observed are fully consistent with our previous findings. Models trained with contrastive objectives substantially outperform their respective backbones, particularly on sentiment and spam detection tasks. However, we once again observe that autoregressive large language models (LLMs), and especially LLAMA, achieve superior performance on news classification datasets. As discussed in Section 5.1, this comparative advantage could be explained by the length of the input sequences. The dataset statistics reported in the appendix indeed show that these benchmarks are composed of very long documents, which significantly degrades the effectiveness of standard embedding-based models. The strongest results reported in Li et al. (2024) (**NLPAD**) rely on an OpenAI model, again trained with similarity-based objectives. While its parameters are not publicly disclosed, this model is likely much larger than those used in our experiments, which may account for its advantage on certain datasets. Lastly, Yang et al. (2025b) (**AD-LLM**) adopt a paradigm based on prompt-based inference with large LLMs, an approach costlier but leading to similar results on most datasets[3].

5.4 Alternative Approaches for Mitigating Embedding Anisotropy

To assess the effectiveness of the standard whitening transformation, we compare it against a range of existing post-processing methods designed to reduce anisotropy in contextual embeddings. In addition to the methods proposed in prior work and reviewed in Section 3.2, we also consider the soft-whitening approach (Diera et al., 2024), which regularizes standard whitening to control the degree of isotropy, and PCA whitening (Su et al., 2021), which incorporates dimensionality reduction into the whitening process. Figure 7 illustrates the performance of each anomaly detector across post-processing methods applied to Qwen3 representations, which exhibit the strongest anisotropy (see Section A.2), making them particularly relevant for the comparison. Implementation details and results for all considered methods, across all post-processing techniques and embedding models, are provided in Appendix A.8, along with the corresponding plots for the remaining embedding models. We further quantify the anisotropy of the embeddings after each post-processing method, and report the corresponding scores in A.8. We first observe that removing dominant principal directions has little to no impact on detection perfor-

[3]Information on what is in-distribution is given to the model in the prompt through *category names*, rather than training data, which makes it difficult to adequately compare those results.

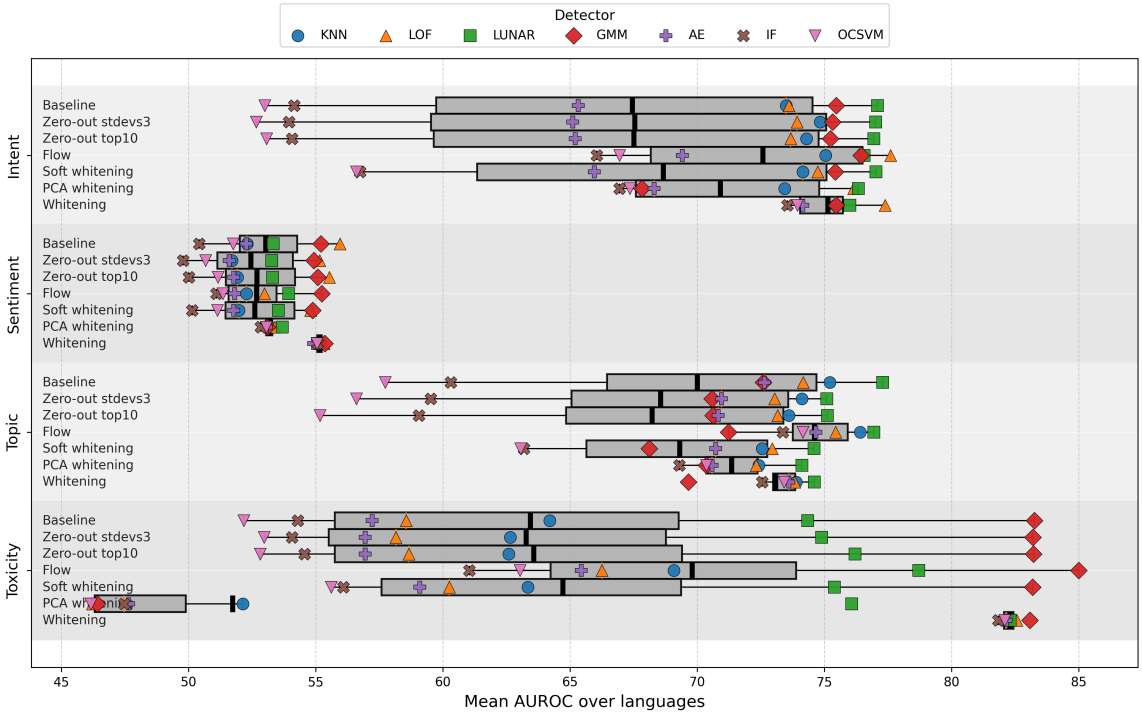

Figure 7: Boxplots of AUROC performance across post-processing methods applied to `Qwen3`. Each box aggregates results over detectors and tasks, averaged across languages.

mance, and can even degrade it. Based on the anisotropy scores, this method only marginally improves the geometry of the embedding space, while potentially discarding relevant information. While soft-whitening and PCA-whitening appear to better adapt embeddings for anomaly detection, with very low anisotropy scores, their impact remains significantly weaker than that of standard whitening. Both regularization and dimensionality reduction seem detrimental to this task, although PCA-whitening generally outperforms soft-whitening, except on the toxicity domain, which demands further investigation. The flow-based method is also effective and notably improves the performance of the algorithms most sensitive to anisotropy, but its impact remains weaker than that of whitening, except on the topic domain with the Qwen3 model. However, this result is not observed for any other model studied. Finally, standard whitening consistently provides the most effective adaptation of the embedding space geometry, effectively eliminating performance gaps across all anomaly detection methods

6 Conclusion

In this work, we highlight the advantages of similarity-trained embedding models for textual anomaly detection. Through extensive experiments conducted across multiple domains and languages, we show that such models, optimized to capture semantic similarity, also exhibit geometric properties that are particularly well suited to anomaly detection algorithms. While some anomaly detection algorithms are strongly affected by inherent properties shared by pretrained models embeddings, we show that a simple whitening post-processing can effectively mitigate these effects and align the geometric assumptions of different detection methods, leading to more homogeneous performance. This finding substantially simplifies the selection of anomaly detection methods in practical real-world deployment settings. Building on these findings, several promising directions remain for future work. First, although whitening proves to be the most effective post-processing method among those considered, it remains important to explore more advanced transformation techniques, potentially better adapted to unsupervised tasks. Second, the strong performance of anomaly detection methods based on local neighborhoods (especially those possessing learnability, as LUNAR and GMMs) on the original representations indicates that developing or exploring anomaly detection algorithms better able to capture the structure of in-distribution embeddings deserves further attention.

7 Limitations

An important limitation lies in the diversity of models we were able to evaluate, which was constrained by two main factors. First, we chose to work exclusively with **open-source models**, in order to ensure full transparency and reproducibility. Second, our experiments were conducted under **limited computational resources** (a single V100 GPU), which prevented us from including the largest LLMs in our evaluation. Additionally, a more thorough study of the impact of similarity-training of embedding models on anomaly detection should include an investigation of fine-tuning the model on in-distribution data - especially looking at the trade-off between data quantity and quality. Given the significant costs that such a study would incur, we leave it for future work. Lastly, some recent studies have also explored the issue of contamination in training data. In this work, we assumed access to clean data. While robustness to contamination is an important challenge, we argue that it should primarily be addressed at the level of detection methods themselves, which constitutes a distinct line of contribution.

8 Acknowledgements

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

References

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.

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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.

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*.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Pierre Colombo, Eduardo Dadalto, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022. Beyond mahalanobis distance for textual ood detection. *Advances in Neural Information Processing Systems*, 35:17744–17759.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Andor Diera, Lukas Galke, and Ansgar Scherp. 2024. Isotropy matters: Soft-zca whitening of embeddings for semantic code search. *arXiv preprint arXiv:2411.17538*.

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, Hong Kong, China. 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.

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.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sopan Khosla and Rashmi Gangadharaiah. 2022. Evaluating the practical utility of confidence-score based techniques for unsupervised open-world classification. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 18–23, Dublin, Ireland. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020a. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020b. On the sentence embeddings from pre-trained language models. *CoRR*, abs/2011.05864.

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*.

Yuxin Liang, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. 2021. Learning to remove: Towards isotropic pre-trained BERT embedding. *CoRR*, abs/2104.05274.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

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.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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.

Bernhard Schölkopf, Robert C Williamson, Alex Smola, John Shawe-Taylor, and John Platt. 1999. Support vector method for novelty detection. *Advances in neural information processing systems*, 12.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *CoRR*, abs/2103.15316.

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*.

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.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *CoRR*, abs/2002.10957.

Chenghao Xiao, Yang Long, and Noura Al Moubayed. 2023. On isotropy, contextualization and learning dynamics of contrastive-based sentence representation learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12266–12283, Toronto, Canada. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao,

Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tiankai Yang, Yi Nian, Li Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan A. Rossi, Kaize Ding, Xia Hu, and Yue Zhao. 2025b. AD-LLM: Benchmarking large language models for anomaly detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1524–1547, Vienna, Austria. Association for Computational Linguistics.

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.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Appendix

A.1 Models selection

Model name	Model	# Params	Release	Languages
xlm-roberta-base	*xlm-roberta-base*	0.3B	2019	100+ lang.
e5	*multilingual-e5-base*	0.3B	2024	100+ lang.
LLama	*Llama-3.2-1B*	1B	2024	En, De, Fr, It, Hi, Es, Pt, Th
Qwen3	*Qwen3-0.6B*	0.6B	2025	100+ lang.
Qwen3-embeddings	*Qwen3-Embedding-0.6B*	0.6B	2025	100+ lang.

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

Table 4 details the Hugging Face checkpoints and key characteristics of the models considered. We focus on recent state-of-the-art models with fewer than one billion parameters to facilitate reproducible experiments.

A.2 Anisotropy scores

Model	En	Fr	De	Hi	It	Pt	Es	Th
XLM-RoBERTa	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
E5	0.8	0.82	0.82	0.8	0.83	0.82	0.81	0.80
LLaMA	0.84	0.86	0.86	0.89	0.88	0.84	0.83	0.83
Qwen3	0.83	0.85	0.89	0.92	0.88	0.82	0.84	0.92
Qwen3-Embeddings	0.36	0.44	0.34	0.35	0.43	0.39	0.36	0.34

Table 5: Average anisotropy scores by model and language. Lower values indicate more isotropic representations.

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:

$$A(f) = \frac{1}{n} \sum_{\{x,y\} \in S} \cos\left(f(x), f(y)\right) \qquad (1)$$

Table 5 highlights the impact of contrastive training on embedding geometry, showing that models such as E5 and Qwen3-Emb. produce substantially more isotropic representations than their corresponding backbone models. Moreover, these isotropy scores are remarkably homogeneous across languages.

A.3 Experiments details

Given the deterministic nature of most embedding models and anomaly detection algorithms considered in this study, we report results from single runs. Moreover, the experimental setup is computationally demanding, involving 7 anomaly detectors, 5 embedding models, 67 datasets, and two representation settings (base vs. whitened), which makes repeated runs impractical.

A.4 NLP-ADBench datasets

Dataset	N_{total}	Mean # Tokens
AG News	98 207	41.47
BBC News	1 785	483.68
N24News	59 822	1 036.17
SMS Spam	4 672	22.48
Email Spam	3 578	240.44
Yelp Reviews	316 924	153.62
Movie reviews	26 369	292.31

Table 6: NLP-ADBench datasets statistics

Table 6 reports the number of available samples for each dataset, along with the average sample length. We intentionally exclude the Emotion dataset, as it is prohibitively expensive for the considered models and falls outside the anomaly detection domains targeted in this work.

A.5 MTEB Classsification datasets

Tables 7 to 10 report detailed statistics for the datasets used in each of the four studied domains. To scope our analysis, we restrict our selection to datasets drawn from the eight languages officially supported by LLaMA. We further retain only classification datasets that can be meaningfully adapted to the anomaly detection setting, focusing on the four domains considered in this work. Finally, we include only datasets with a sufficient number of samples to ensure statistically reliable evaluation results. Table 11 summarizes the datasets available for each language.

Dataset	Lang.	n_{total}	Mean # tokens
HUME Toxic	en	7 493	66.45
Toxic Conversations	en	46 833	69.05
Multi-Hate	deu	762	20.65
	eng	758	13.08
	fra	759	20.99
	hin	750	19.71
	ita	757	19.82
	por	755	19.11
	spa	766	18.42
Hate Speech Portuguese	pt	1 536	42.84

Table 7: Toxicity and hate speech datasets statistics.

Dataset	Lang.	n_{total}	Mean # tokens
Amazon Reviews	de	6 000	80.87
	en	6 000	49.30
	es	6 000	57.48
	fr	6 000	56.98
Amazon Counterfactual	de	4 182	44.52
	en	1 426	26.87
Yelp Reviews	en	6 000	161.44
Tweet Sentiment	en	6 532	21.50
Financial Phrasebank	en	626	34.44
Movie Reviews	fr	15 000	176.18
Spanish Sentiment	es	884	29.21
Sardistance	it	779	62.91
Sentiment Hindi	hi	609	47.84
Wisesight	th	4 955	8.80
Wongnai	th	7 201	47.35

Table 8: Sentiment analysis datasets statistics.

Dataset	Lang.	n_{total}	Mean # tokens
Banking77	en	6 192	16.09
Massive Intent	de	7 908	15.08
	en	7 908	9.48
	es	7 908	14.42
	fr	7 908	16.15
	hi	7 908	22.01
	it	7 908	14.65
	pt	7 908	14.87
	th	7 908	7.35
Massive Scenario	de	6 715	15.69
	en	6 875	9.81
	es	6 730	15.01
	fr	6 819	16.77
	hi	6 826	22.92
	it	6 519	15.34
	pt	6 642	15.52
	th	4 404	9.27
MTOP Intent	de	13 222	16.42
	en	15 444	10.16
	es	10 772	16.20
	fr	11 620	16.04
	hi	11 176	22.70
	th	10 592	4.46
MTOP Domain	de	9 317	15.59
	en	10 700	9.80
	es	8 041	15.52
	fr	8 481	15.23
	hi	8 008	21.70
	th	7 869	4.61

Table 9: Intent classification datasets statistics.

Dataset	Lang.	n_{total}	Mean # tokens
DBpedia	en	1 313	65.51
Yahoo Answers Topics	en	1 351	89.40
MasakhaNews	en	2 967	650.40
ArXiv Classification	en	11 555	15 366.27
Patent Classification	en	8 556	4 265.48
TenKGNAD	de	7 133	842.77
Spanish News	es	1 320	1 468.80

Table 10: Topic classification datasets statistics.

Language	Domain	Datasets
English (en)	Toxicity / Hate	HUME Toxic; Toxic Conversations50k; Multi-Hate
	Intent	Massive Intent; Massive Scenario; MTOP Intent; MTOP Domain; Banking77
	Topic	DBPedia; Yahoo Answers Topics; MasakhaNews ; arxiv class ; patent class
	Sentiment	Amazon Reviews; Amazon Counterfactual; yelp reviews ; tweet sentiment, financial phrasebank
German (de)	Sentiment	Amazon Reviews; Amazon Counterfactual
	Toxicity / Hate	Multi-Hate
	Intent	Massive Intent; Massive Scenario; MTOP Intent; MTOP Domain
	Topic	TenKGNAD
French (fr)	Toxicity / Hate	Multi-Hate
	Intent	Massive Intent; Massive Scenario; MTOP Intent; MTOP Domain
	Sentiment	Movie Reviews; Amazon Reviews
Spanish (es)	Toxicity / Hate	Multi-Hate
	Intent	Massive Intent; Massive Scenario; MTOP Intent; MTOP Domain
	Topic	Spanish News
	Sentiment	Spanish Sentiment
Italian (it)	Toxicity / Hate	Multi-Hate
	Intent	Massive Intent; Massive Scenario
	Sentiment	Sardistance
Portuguese (pt)	Toxicity / Hate	Multi-Hate; Hate-Speech-Portuguese
	Intent	Massive Intent; Massive Scenario
Hindi (hi)	Toxicity / Hate	Multi-Hate
	Intent	Massive Intent; Massive Scenario; MTOP Intent; MTOP Domain
	Sentiment	Sentiment Hindi
Thai (th)	Intent	Massive Intent; Massive Scenario; MTOP Intent; MTOP Domain
	Sentiment	Wisesight, Wongnai

Table 11: Overview of the datasets used, grouped by language and domain.

A.6 Detailed results by languages

Tables 12–27 summarize domain-averaged results for each language, reported for both base and whitened representations. The results reported in the tables 1 and 2 are obtained by averaging performance across all considered languages.

A.6.1 English results

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	52.99	60.09	54.49	56.49	55.25	51.61	51.98
	Topic	65.13	64.28	69.37	66.31	66.81	56.70	52.71
	Toxicity	55.38	50.37	60.96	61.20	51.51	49.62	47.40
	Intent	69.56	70.54	75.00	69.56	64.45	53.88	51.27
E5	Sentiment	75.98	72.45	75.26	70.49	75.34	68.74	78.19
	Topic	72.89	74.76	73.39	73.21	70.13	56.35	59.41
	Toxicity	60.60	59.80	63.71	63.12	60.18	53.96	55.43
	Intent	88.04	87.86	88.85	83.62	83.10	65.64	68.67
LLaMA	Sentiment	60.08	61.79	60.34	60.64	58.99	54.83	54.92
	Topic	75.24	73.85	76.17	75.32	72.74	60.32	60.07
	Toxicity	55.43	55.49	59.28	60.69	52.69	52.09	52.43
	Intent	73.86	74.24	77.74	75.74	66.30	55.73	56.23
Qwen3	Sentiment	56.36	58.39	57.68	57.79	56.12	51.99	52.84
	Topic	75.80	71.64	76.24	74.17	74.98	66.99	61.35
	Toxicity	71.61	55.57	61.31	61.61	51.11	50.84	49.61
	Intent	71.63	75.19	75.33	73.63	66.05	55.32	55.18
Qwen3-Emb	Sentiment	62.98	61.94	62.41	61.43	62.27	57.65	62.30
	Topic	77.59	75.37	78.10	73.83	74.53	56.75	65.38
	Toxicity	53.83	55.01	54.41	54.53	51.79	51.37	51.77
	Intent	86.51	84.29	87.25	81.46	79.58	61.96	61.54

Table 12: English - base

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	54.94	55.81	55.45	55.68	56.16	53.78	55.66
	Topic	66.50	67.30	69.29	67.15	67.78	65.99	64.06
	Toxicity	58.68	60.29	61.02	62.35	58.07	57.27	58.56
	Intent	72.92	75.52	74.14	70.66	69.21	65.24	65.60
E5	Sentiment	71.10	71.13	71.12	70.80	71.06	67.85	71.02
	Topic	73.44	73.65	73.59	72.94	72.93	68.96	72.99
	Toxicity	63.50	63.82	63.72	62.45	62.45	60.59	62.26
	Intent	83.22	82.99	83.61	83.69	83.47	81.83	83.37
LLaMA	Sentiment	61.05	61.43	61.26	59.93	60.88	59.71	60.81
	Topic	75.59	75.88	75.68	75.17	75.39	74.11	75.38
	Toxicity	59.80	60.41	60.20	60.13	59.82	59.34	59.71
	Intent	74.90	76.07	75.08	76.12	74.57	73.68	74.50
Qwen3	Sentiment	58.66	58.85	58.64	57.96	58.37	57.04	58.51
	Topic	74.77	74.48	74.75	73.34	74.66	73.96	74.72
	Toxicity	60.04	60.93	60.05	60.48	60.23	59.90	60.04
	Intent	73.88	75.17	74.12	73.76	72.98	72.24	72.87
Qwen3-Emb	Sentiment	61.69	61.69	61.60	61.87	61.70	61.43	61.72
	Topic	74.49	74.71	74.62	73.52	74.20	70.80	74.24
	Toxicity	55.68	56.13	55.72	55.64	55.56	55.73	55.54
	Intent	81.30	81.48	81.55	81.55	81.26	80.61	81.19

Table 13: English - Whitened

A.6.2 French results

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	49.96	56.06	50.49	52.20	50.83	49.98	49.71
	Toxicity	59.39	54.04	81.97	82.09	52.33	48.97	47.10
	Intent	65.98	67.32	71.69	66.04	62.28	51.48	48.71
E5	Sentiment	64.58	59.76	66.17	66.33	62.63	59.96	61.71
	Toxicity	88.84	83.44	92.23	95.32	82.89	69.80	71.87
	Intent	83.54	83.02	83.94	80.60	80.00	63.34	65.94
LLaMA	Sentiment	50.01	57.45	51.60	54.47	52.24	49.56	48.76
	Toxicity	74.38	62.16	84.37	92.67	60.75	54.57	53.00
	Intent	73.48	68.87	76.32	77.73	65.26	54.28	51.31
Qwen3	Sentiment	50.28	54.54	50.54	52.42	50.58	50.98	51.05
	Toxicity	65.31	60.42	84.87	89.21	57.46	54.83	48.96
	Intent	73.58	72.11	78.30	75.03	64.48	52.34	50.49
Qwen3-Emb	Sentiment	59.57	59.46	59.50	60.06	55.87	53.76	56.29
	Toxicity	84.24	76.27	84.34	95.28	75.44	60.38	67.86
	Intent	81.37	81.92	82.26	81.74	77.33	61.22	64.88

Table 14: French-base

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	54.09	56.52	54.34	55.88	54.98	54.06	54.57
	Toxicity	71.56	71.11	82.35	86.42	68.04	64.92	68.67
	Intent	69.68	71.45	71.66	67.40	65.30	61.68	62.07
E5	Sentiment	65.54	65.35	65.68	66.04	65.16	62.47	65.10
	Toxicity	95.31	95.36	95.25	95.18	95.28	95.01	95.35
	Intent	80.55	80.62	81.02	80.64	80.65	78.84	80.46
LLaMA	Sentiment	53.90	54.68	54.19	54.56	54.11	53.62	54.00
	Toxicity	92.37	92.52	92.43	92.69	92.32	91.04	92.33
	Intent	76.69	78.35	76.89	78.14	76.28	75.40	76.16
Qwen3	Sentiment	53.60	53.90	53.80	54.04	53.82	54.46	53.74
	Toxicity	88.24	88.72	88.48	89.04	88.17	88.21	88.14
	Intent	75.68	77.50	76.43	75.21	74.27	73.51	74.06
Qwen3-Emb	Sentiment	62.06	62.92	62.19	61.57	61.24	60.13	61.22
	Toxicity	95.12	95.12	95.11	95.35	95.06	91.62	95.08
	Intent	81.63	80.98	81.71	81.77	81.58	79.83	81.58

Table 15: French- Whitened

A.6.3 German results

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	54.47	63.56	56.00	58.16	57.42	56.17	54.11
	Topic	64.80	74.62	72.25	64.68	65.31	48.70	53.77
	Toxicity	64.72	61.63	86.37	88.44	57.76	53.56	60.10
	Intent	71.18	68.90	75.72	69.18	64.00	53.78	52.26
E5	Sentiment	68.17	63.12	68.09	64.80	67.88	60.37	64.98
	Topic	72.07	76.23	72.80	70.88	62.46	47.08	49.64
	Toxicity	86.59	80.77	95.11	96.39	83.18	66.88	74.24
	Intent	85.88	84.48	86.48	82.67	82.13	67.86	70.13
LLaMA	Sentiment	57.20	59.90	57.76	60.03	57.42	56.64	56.38
	Topic	68.81	72.57	71.31	62.50	60.53	47.34	47.49
	Toxicity	66.45	65.60	88.83	94.45	63.42	57.28	54.46
	Intent	75.01	69.94	78.40	76.11	65.18	52.87	51.86
Qwen3	Sentiment	56.70	59.84	56.84	58.80	57.80	56.59	58.32
	Topic	71.89	74.83	74.06	64.14	65.54	56.14	52.39
	Toxicity	60.99	59.70	84.10	92.24	60.60	53.96	51.60
	Intent	73.32	72.04	76.78	73.63	65.59	53.83	52.98
Qwen3-Emb	Sentiment	67.14	54.76	66.04	62.96	65.93	61.60	67.58
	Topic	78.70	84.66	78.64	73.85	67.18	49.78	50.26
	Toxicity	84.04	79.71	85.86	97.51	77.85	69.12	75.79
	Intent	83.88	80.56	83.84	81.49	80.06	65.23	67.49

Table 16: German - base

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	57.83	59.65	58.60	59.51	59.12	57.96	58.72
	Topic	65.88	67.63	72.20	64.97	66.55	61.49	59.86
	Toxicity	77.88	79.12	85.21	88.78	78.38	70.88	79.15
	Intent	73.77	75.71	74.82	69.99	68.56	64.88	65.25
E5	Sentiment	65.12	64.96	65.17	64.85	64.66	63.08	64.61
	Topic	71.11	72.85	71.13	69.25	69.54	65.67	69.44
	Toxicity	96.60	96.65	96.74	96.45	96.62	95.77	96.58
	Intent	82.51	82.28	82.76	82.73	82.54	80.87	82.42
LLaMA	Sentiment	58.96	59.12	59.08	59.14	59.09	57.67	58.98
	Topic	65.27	66.84	65.81	64.07	64.39	64.04	64.18
	Toxicity	94.09	94.26	93.88	94.46	94.12	94.70	94.13
	Intent	75.24	76.68	75.33	76.45	74.90	74.12	74.85
Qwen3	Sentiment	56.54	56.94	56.73	56.94	56.72	56.20	56.72
	Topic	66.05	67.01	67.18	63.81	64.45	63.41	64.17
	Toxicity	90.96	91.42	91.81	92.29	91.25	88.83	91.13
	Intent	74.26	76.02	74.94	73.78	72.99	72.40	72.77
Qwen3-Emb	Sentiment	64.48	64.46	64.23	64.30	64.31	63.05	64.32
	Topic	75.35	76.70	75.12	75.44	75.34	72.92	75.42
	Toxicity	97.50	97.55	97.47	97.46	97.37	95.99	97.51
	Intent	82.12	81.98	82.06	82.08	81.96	80.76	81.96

Table 17: German - Whitened

A.6.4 Hindi results

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	62.83	54.26	65.78	63.69	57.49	52.65	43.48
	Toxicity	54.27	53.31	63.06	84.20	54.07	52.43	47.20
	Intent	68.92	70.51	72.13	69.88	63.06	51.54	50.22
E5	Sentiment	77.70	86.24	71.80	58.34	86.87	70.81	76.29
	Toxicity	79.62	72.19	82.68	94.66	69.18	60.07	62.03
	Intent	84.74	84.75	85.14	81.38	79.99	66.63	66.58
LLaMA	Sentiment	59.42	63.25	60.55	58.99	58.39	53.16	49.56
	Toxicity	63.25	65.82	67.46	89.89	61.11	52.97	54.86
	Intent	71.95	72.66	75.00	76.19	62.84	54.56	53.28
Qwen3	Sentiment	64.47	65.97	65.71	66.01	63.39	60.46	58.69
	Toxicity	57.30	55.56	63.68	86.60	55.64	56.98	54.39
	Intent	68.24	70.18	71.90	72.63	59.23	50.86	50.29
Qwen3-Emb	Sentiment	64.22	71.82	59.45	55.90	69.10	51.82	66.41
	Toxicity	70.89	67.50	75.59	94.96	66.54	70.18	61.56
	Intent	82.08	80.73	81.86	81.28	77.33	63.76	64.34

Table 18: Hindi - Base

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	62.56	62.72	63.46	64.08	60.12	63.18	60.21
	Toxicity	69.48	70.76	70.72	89.38	68.88	65.90	70.44
	Intent	73.15	78.53	73.83	72.85	68.61	64.78	64.90
E5	Sentiment	58.71	58.87	58.00	58.50	58.80	54.33	58.85
	Toxicity	94.35	94.30	94.25	95.17	94.30	92.64	94.47
	Intent	81.37	81.35	81.68	81.44	81.26	79.46	81.11
LLaMA	Sentiment	59.33	59.31	59.12	58.87	59.15	60.62	59.29
	Toxicity	88.79	89.22	89.00	89.99	88.93	88.02	88.99
	Intent	74.81	77.08	75.00	76.85	74.10	74.11	74.00
Qwen3	Sentiment	66.29	66.15	66.77	66.18	65.83	68.64	66.31
	Toxicity	83.72	84.60	84.01	86.62	84.05	85.89	83.87
	Intent	72.06	74.92	73.04	73.32	70.45	70.02	70.01
Qwen3-Emb	Sentiment	56.04	55.97	55.35	55.90	55.90	62.74	56.04
	Toxicity	94.96	94.99	94.92	95.00	94.87	92.21	94.99
	Intent	81.33	81.48	81.48	81.31	81.11	80.16	81.11

Table 19: Hindi - Whitened

A.6.5 Italian results

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	52.87	52.28	53.65	53.29	52.59	51.45	47.73
	Toxicity	64.82	62.95	77.23	85.01	61.47	55.27	55.92
	Intent	71.55	70.28	75.47	73.01	64.07	54.08	50.72
E5	Sentiment	58.01	55.04	58.21	56.40	52.50	52.31	49.56
	Toxicity	84.57	76.46	89.23	93.49	81.05	68.42	71.27
	Intent	88.97	84.76	89.50	86.43	84.14	70.13	68.94
LLaMA	Sentiment	48.73	49.99	49.59	51.82	49.57	46.48	45.55
	Toxicity	73.44	65.81	84.18	90.43	68.33	58.04	59.42
	Intent	76.75	72.43	79.04	82.48	68.68	55.74	54.94
Qwen3-LLM	Sentiment	53.23	53.37	50.95	55.28	51.51	48.53	46.47
	Toxicity	69.48	65.65	81.71	89.40	64.16	58.40	60.86
	Intent	77.46	75.75	79.60	80.26	68.38	55.36	53.94
Qwen3	Sentiment	55.01	53.71	55.32	56.24	50.98	47.07	47.97
	Toxicity	82.35	75.48	83.44	93.42	77.85	64.20	76.59
	Intent	84.96	77.44	85.84	86.95	79.86	63.29	64.50

Table 20: Italian - Base

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	53.12	52.97	53.74	54.33	52.68	54.93	52.26
	Toxicity	76.60	76.26	83.09	87.09	75.47	74.19	76.15
	Intent	77.03	79.22	77.00	74.24	70.76	66.81	68.02
E5	Sentiment	56.61	56.39	56.62	56.26	56.59	57.79	56.53
	Toxicity	93.91	93.80	93.85	93.03	94.00	91.89	93.90
	Intent	86.73	87.05	86.94	86.52	86.30	85.97	86.18
LLaMA	Sentiment	51.75	51.85	51.67	51.84	51.62	52.50	51.77
	Toxicity	90.15	90.18	90.23	90.44	90.18	88.63	90.18
	Intent	82.31	82.92	82.47	83.50	81.97	81.56	81.91
Qwen3	Sentiment	55.10	55.18	54.97	55.32	54.82	54.29	55.07
	Toxicity	89.38	89.46	89.41	89.21	89.27	89.14	89.37
	Intent	80.40	82.10	80.48	80.44	79.42	78.40	79.32
Qwen3-Emb	Sentiment	56.32	56.30	56.42	56.25	56.55	52.52	56.32
	Toxicity	93.79	93.77	93.71	93.30	93.73	93.12	93.77
	Intent	86.80	87.13	86.92	87.00	86.72	85.44	86.68

Table 21: Italian - Whitened

A.6.6 Portuguese results

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	71.19	66.90	74.88	72.64	64.55	52.41	46.88
	Toxicity	65.18	59.17	68.45	72.78	58.27	54.72	55.12
E5	Intent	86.91	83.06	87.34	84.38	82.50	66.06	66.76
	Toxicity	71.60	67.15	72.88	76.20	67.60	57.52	63.60
LLaMA	Intent	76.07	72.98	78.27	81.11	69.40	56.06	55.83
	Toxicity	68.97	61.22	74.19	75.76	63.92	61.18	58.96
Qwen3	Intent	77.02	75.37	80.08	80.19	69.24	56.76	56.14
	Toxicity	60.06	58.28	66.81	71.90	57.14	55.87	53.51
Qwen3-Emb	Intent	84.98	81.97	86.04	86.17	80.76	64.66	64.19
	Toxicity	65.08	60.54	66.18	71.72	61.22	54.64	60.02

Table 22: Portuguese - base

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	75.48	76.91	76.20	73.81	70.56	65.92	67.46
	Toxicity	69.44	69.61	70.00	73.34	68.01	67.46	68.36
E5	Intent	84.26	84.19	84.53	84.38	84.44	83.54	84.39
	Toxicity	76.27	76.22	76.03	76.35	76.04	74.68	76.16
LLaMA	Intent	80.32	81.54	80.29	81.46	80.04	79.06	80.00
	Toxicity	75.74	75.89	75.77	75.75	75.80	74.84	75.77
Qwen3	Intent	79.76	81.33	80.20	80.06	78.87	77.92	78.78
	Toxicity	71.50	71.76	71.51	72.10	71.48	70.48	71.49
Qwen3-Emb	Intent	85.91	86.20	85.96	86.22	86.00	84.54	85.94
	Toxicity	71.38	71.28	71.72	72.08	71.29	72.28	71.34

Table 23: Portuguese - Whitened

A.6.7 Spanish results

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	47.56	52.64	50.18	53.26	47.60	47.78	50.23
	Topic	75.65	74.86	84.31	84.25	80.50	52.13	46.39
	Toxicity	59.37	55.01	77.52	75.67	49.26	47.75	57.78
	Intent	67.55	66.22	72.26	68.53	61.49	51.26	48.58
E5	Sentiment	81.48	77.09	80.18	75.53	81.06	73.24	79.16
	Topic	94.01	88.45	95.43	95.05	91.16	70.10	77.71
	Toxicity	80.50	70.59	90.12	95.24	72.60	59.32	67.94
	Intent	84.42	84.46	84.74	81.86	80.47	65.12	66.56
LLaMA	Sentiment	56.00	58.90	59.58	62.26	55.32	51.68	51.61
	Topic	88.48	79.30	89.04	89.58	85.21	70.96	65.69
	Toxicity	73.82	63.89	87.17	93.86	64.75	53.86	53.90
	Intent	75.81	71.27	78.20	76.34	65.34	53.52	51.25
Qwen3	Sentiment	47.31	53.68	50.53	53.73	48.24	46.67	51.18
	Topic	77.98	76.05	81.57	79.49	77.43	57.80	59.44
	Toxicity	64.64	54.66	77.87	91.89	54.45	49.14	46.24
	Intent	74.63	74.71	78.90	74.44	66.59	55.44	53.20
Qwen3-Emb	Sentiment	74.50	68.90	74.34	74.94	72.06	67.35	71.33
	Topic	93.09	88.63	92.76	90.58	88.20	65.07	71.60
	Toxicity	80.34	73.79	81.21	94.08	68.05	63.93	67.25
	Intent	82.95	83.36	82.82	82.16	79.13	62.63	63.98

Table 24: Spanish - Base

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	49.38	50.84	50.65	53.01	49.18	49.39	48.82
	Topic	79.42	79.13	86.01	87.68	85.15	78.92	75.60
	Toxicity	64.25	63.42	77.44	80.01	61.22	57.79	63.41
	Intent	71.43	74.95	71.66	69.72	66.98	63.39	64.10
E5	Sentiment	76.34	76.56	76.48	74.15	75.96	70.76	75.87
	Topic	95.80	95.70	95.71	94.20	95.80	89.30	95.75
	Toxicity	94.84	95.01	95.21	94.72	94.89	92.90	94.82
	Intent	81.66	80.66	81.76	81.96	81.91	80.06	81.76
LLaMA	Sentiment	61.12	61.90	62.08	61.92	61.10	60.84	61.03
	Topic	89.90	89.79	89.96	89.34	89.89	88.27	89.97
	Toxicity	93.84	93.87	93.85	93.74	93.82	91.88	93.82
	Intent	75.78	77.68	76.06	76.87	75.40	74.24	75.33
Qwen3	Sentiment	53.40	53.78	53.58	54.42	53.39	53.41	53.45
	Topic	80.84	79.99	81.91	71.81	81.72	80.27	81.36
	Toxicity	90.53	91.10	90.89	91.83	90.62	90.46	90.58
	Intent	74.04	76.38	74.49	73.72	72.91	72.68	72.79
Qwen3-Emb	Sentiment	75.44	75.96	75.42	75.32	75.10	74.76	75.10
	Topic	91.53	91.58	91.64	83.38	91.72	83.80	91.56
	Toxicity	93.98	94.09	94.03	93.99	93.95	92.30	93.99
	Intent	81.67	81.29	81.69	81.94	81.76	80.31	81.82

Table 25: Spanish - Whitened

A.6.8 Thai results

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	44.92	52.66	49.87	50.82	48.18	44.32	50.61
	Intent	68.14	69.97	73.82	69.83	64.42	51.51	50.93
E5	Sentiment	64.20	58.06	65.98	65.04	62.69	54.79	55.70
	Intent	83.65	84.39	84.40	82.02	80.50	65.89	67.01
LLaMA	Sentiment	41.10	49.02	41.88	47.94	42.58	39.25	43.80
	Intent	76.12	76.23	78.22	77.38	65.72	53.83	51.77
Qwen3	Sentiment	37.76	45.88	41.08	42.34	38.22	37.71	43.71
	Intent	72.15	73.44	75.84	73.94	62.96	53.29	51.73
Qwen3-Emb	Sentiment	65.24	56.68	65.14	61.64	61.87	57.20	56.93
	Intent	83.72	82.08	83.65	81.14	78.94	60.48	65.45

Table 26: Thai - Base

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Sentiment	50.80	52.86	51.62	51.69	50.39	49.14	49.28
	Intent	72.31	75.40	73.98	71.32	68.94	66.03	65.62
E5	Sentiment	66.01	67.30	66.31	66.20	65.83	63.30	65.70
	Intent	81.75	81.30	82.12	81.86	81.90	80.23	81.70
LLaMA	Sentiment	49.18	49.72	48.96	50.20	49.06	49.14	49.04
	Intent	76.36	77.63	76.43	77.23	75.68	74.91	75.62
Qwen3	Sentiment	41.90	42.42	41.61	42.64	41.52	41.43	41.58
	Intent	73.40	75.66	74.26	73.44	71.28	71.13	70.94
Qwen3-Emb	Sentiment	60.71	61.82	61.00	60.58	60.46	59.36	60.42
	Intent	81.28	81.35	81.16	80.99	80.81	79.83	80.84

Table 27: Thai - Whitened

A.7 Multilingual analysis

Figures 8–11 present boxplots of model performance across languages for each domain. The absence of a clear pattern suggests that no model exhibits systematic language-dependent performance variations.

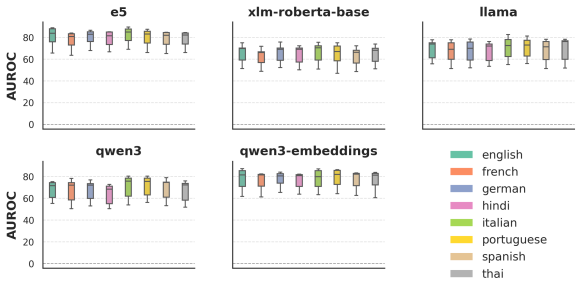

Figure 8: Boxplots of model performance across languages on **Intents classification datasets**, measured in AUROC.

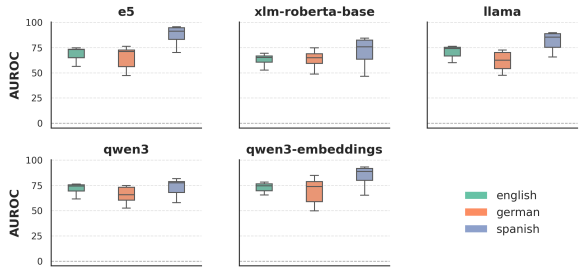

Figure 9: Boxplots of model performance across languages on **Topic classification datasets**, measured in AUROC.

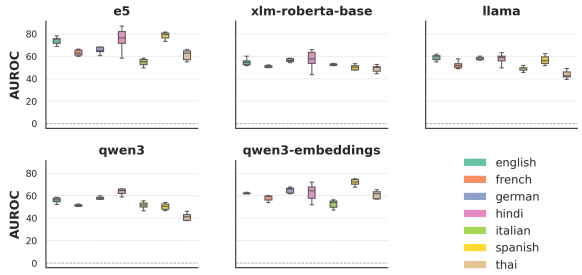

Figure 10: Boxplots of model performance across languages on **Topic classification datasets**, measured in AUROC.

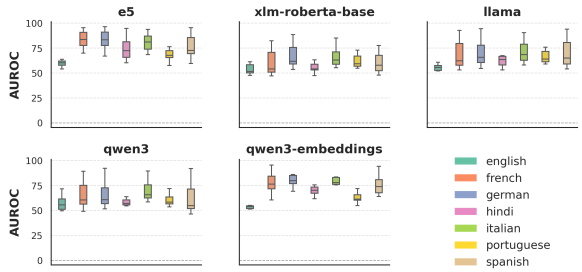

Figure 11: Boxplots of model performance across languages on **Toxicity classification datasets**, measured in AUROC.

A.8 Alternative Approaches for Mitigating Embedding Anisotropy

Tables 28 to 34 report the performance obtained with several approaches proposed in the literature to mitigate embedding anisotropy. For **flow-based normalization** (Li et al., 2020b), we adopt the *target* configuration, where the flow model is trained directly on the dataset under consideration, using the initialization and hyperparameters reported in the original work. **Soft whitening** is implemented following the procedure introduced by (Diera et al., 2024), with the parameter settings described in the corresponding paper. We also implement the two strategies proposed by (Hämmerl et al., 2023) to identify and **remove outlier dimensions**: (1) removing the 10 dimensions with the largest mean magnitude, and (2) removing all dimensions whose mean is more than three standard deviations above the global mean. Finally, we experiment with **PCA whitening with dimensionality reduction**, retaining the number of components needed to explain 95% of the variance, following common practice. Figures 12 to 15 further illustrate the advantage of standard whitening method on representations derived from the **XLM-RoBERTa-base**, **LLaMA**, **e5** and **Qwen3** models, and highlight the homogenization of performance induced by whitening, a property not achieved by the other methods. Table 35 reports the anisotropy scores for each post-processing method across models, averaged over languages.

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	69.26	68.83	73.87	69.83	63.54	52.49	49.95
	Sent.	52.23	55.94	54.35	55.42	52.77	50.57	49.69
	Topic	68.53	71.25	75.31	71.75	70.87	52.51	50.96
	Toxicity	60.45	56.64	73.65	78.48	54.95	51.76	52.95
E5	Intent	85.77	84.60	86.30	82.87	81.60	66.33	67.57
	Sent.	70.02	67.39	69.38	65.28	69.85	62.89	66.51
	Topic	79.66	79.81	80.54	79.71	74.58	57.84	62.25
	Toxicity	78.90	72.91	83.70	87.77	73.81	62.28	66.63
Qwen3	Intent	73.50	73.60	77.09	75.47	65.32	54.15	52.99
	Sent.	52.30	55.95	53.33	55.20	52.27	50.42	51.75
	Topic	75.22	74.17	77.29	72.60	72.65	60.31	57.73
	Toxicity	64.20	58.55	74.34	83.26	57.22	54.29	52.17
Qwen3-E	Intent	83.81	81.54	84.20	82.80	79.12	62.90	64.55
	Sent.	64.09	61.04	63.17	61.88	62.58	56.64	61.26
	Topic	83.13	82.89	83.17	79.42	76.64	57.20	62.41
	Toxicity	74.40	69.76	75.86	85.93	68.39	61.97	65.83
LLaMA	Intent	74.88	72.33	77.65	77.88	66.09	54.57	53.31
	Sent.	53.22	57.19	54.47	56.59	53.50	50.23	50.08
	Topic	77.51	75.24	78.84	75.80	72.83	59.54	57.75
	Toxicity	67.96	62.86	77.93	85.39	62.14	55.71	55.29

Table 28: Mean AUROC across languages for each model and task, organised by detector. Bold values highlight the best-performing scores within each task–detector configuration.

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	73.22	75.96	74.16	71.25	68.62	64.84	65.38
	Sent.	54.67	55.91	55.41	56.31	54.66	54.63	54.22
	Topic	70.60	71.35	75.83	73.27	73.16	68.80	66.51
	Toxicity	69.70	70.08	75.69	81.05	68.30	65.49	69.25
E5	Intent	82.76	82.56	83.05	82.90	82.81	81.35	82.67
	Sent.	65.63	65.79	65.63	65.26	65.44	62.80	65.38
	Topic	80.12	80.73	80.14	78.80	79.42	74.64	79.39
	Toxicity	87.83	87.88	87.86	87.62	87.65	86.21	87.65
Qwen3	Intent	75.44	77.39	76.00	75.47	74.15	73.54	73.94
	Sent.	55.07	55.32	55.16	55.36	54.92	55.07	55.05
	Topic	73.89	73.83	74.61	69.65	73.61	72.55	73.42
	Toxicity	82.05	82.57	82.31	83.08	82.15	81.84	82.09
Qwen3-E	Intent	82.76	82.74	82.82	82.86	82.65	81.44	82.64
	Sent.	62.39	62.73	62.32	62.26	62.18	62.00	62.16
	Topic	80.46	81.00	80.46	77.45	80.42	75.84	80.41
	Toxicity	86.06	86.13	86.10	86.12	85.98	84.75	86.03
LLaMA	Intent	77.05	78.49	77.19	78.33	76.62	75.89	76.55
	Sent.	56.47	56.86	56.62	56.64	56.43	56.30	56.42
	Topic	76.92	77.50	77.15	76.19	76.56	75.47	76.51
	Toxicity	84.97	85.19	85.05	85.31	85.00	84.06	84.99

Table 29: Mean AUROC with whitening across languages for each model and task, organised by detector.

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	74.92	78.43	76.47	77.03	68.55	62.75	64.27
	Sent.	51.69	52.70	53.74	56.75	50.80	50.34	50.31
	Topic	66.31	66.02	68.03	67.36	64.82	63.15	63.90
	Toxicity	68.32	63.34	78.88	86.76	63.26	58.68	60.65
E5	Intent	81.88	83.07	82.36	82.12	80.42	76.99	79.43
	Sent.	67.26	68.34	67.09	65.25	66.63	64.09	66.33
	Topic	66.26	69.25	68.33	70.50	63.86	60.81	62.49
	Toxicity	79.08	76.25	83.72	88.21	76.32	69.59	75.58
Qwen3	Intent	75.05	77.60	76.56	76.45	69.41	66.05	66.95
	Sent.	52.28	52.98	53.93	55.24	51.81	51.09	51.34
	Topic	76.41	75.44	76.95	71.23	74.67	73.37	74.16
	Toxicity	69.08	66.25	78.71	85.00	65.44	61.04	63.03
Qwen3-E	Intent	81.94	81.85	82.93	82.06	79.07	75.66	77.39
	Sent.	59.32	60.09	60.45	60.56	58.85	57.28	57.77
	Topic	73.59	76.07	75.74	75.20	71.12	67.21	69.46
	Toxicity	70.00	68.12	71.08	85.36	66.85	62.77	64.47
LLaMA	Intent	75.82	76.92	77.21	78.82	70.81	67.32	68.85
	Sent.	56.31	56.62	57.53	58.22	55.77	52.45	55.11
	Topic	77.09	78.17	78.22	76.48	74.67	72.32	73.49
	Toxicity	71.26	69.16	79.94	85.50	68.23	63.75	66.05

Table 30: Mean AUROC with **Flow-based normalization** across languages for each model and task, organised by detector.

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	72.42	74.31	74.32	65.61	66.61	64.39	60.36
	Sent.	51.02	51.77	52.67	50.78	50.98	50.51	49.90
	Topic	66.41	66.55	68.88	66.13	66.52	64.80	60.41
	Toxicity	54.25	44.91	74.57	47.59	49.10	46.48	49.79
E5	Intent	82.18	81.72	83.20	81.50	81.69	79.16	81.50
	Sent.	66.34	66.27	66.20	66.35	66.40	63.47	66.37
	Topic	76.27	77.57	76.28	75.05	75.07	69.16	75.00
	Toxicity	59.38	56.46	82.51	57.54	58.73	54.29	57.24
Qwen3	Intent	73.44	76.15	76.34	67.83	68.30	66.94	67.35
	Sent.	53.20	53.39	53.69	53.10	53.05	52.84	53.05
	Topic	72.42	72.32	74.12	70.36	70.59	69.30	70.36
	Toxicity	52.13	46.21	76.07	46.40	47.64	47.47	46.15
Qwen3-E	Intent	81.22	81.90	82.91	80.63	80.78	79.27	80.65
	Sent.	62.74	63.08	63.24	62.29	62.39	61.81	62.24
	Topic	78.18	79.15	78.37	77.31	77.34	72.62	77.34
	Toxicity	64.91	62.78	80.82	63.67	64.48	62.99	64.16
LLaMA	Intent	74.25	76.77	76.60	70.87	70.89	69.91	70.19
	Sent.	54.27	54.19	55.84	53.59	53.56	52.75	53.43
	Topic	74.49	74.93	75.76	72.45	72.41	71.07	72.22
	Toxicity	49.57	43.71	72.13	43.97	44.95	45.73	43.49

Table 31: Mean AUROC with **PCA-Whitening** across languages for each model and task, organised by detector.

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	69.41	68.88	73.71	71.06	63.36	52.43	50.15
	Sent.	51.19	54.39	53.26	55.31	51.34	48.72	49.05
	Topic	66.74	69.24	73.36	71.81	68.98	50.28	50.57
	Toxicity	60.49	56.61	73.27	81.08	55.01	52.14	53.33
E5	Intent	85.11	85.42	85.81	82.97	81.85	71.03	75.34
	Sent.	69.90	67.71	69.83	65.55	69.98	63.02	68.56
	Topic	77.67	80.01	78.46	77.55	74.18	60.81	68.36
	Toxicity	79.57	76.47	83.82	87.81	77.68	67.46	75.03
Qwen3	Intent	74.15	74.74	77.03	75.43	65.96	56.74	56.60
	Sent.	51.98	54.79	53.53	54.88	51.77	50.14	51.13
	Topic	72.56	72.95	74.59	68.11	70.73	63.19	63.05
	Toxicity	63.34	60.25	75.40	83.19	59.09	56.08	55.61
Qwen3-E	Intent	83.09	83.63	83.63	83.01	80.61	73.55	77.22
	Sent.	62.98	63.05	63.00	61.97	61.70	59.80	62.24
	Topic	79.57	82.56	79.78	75.56	76.68	66.57	75.53
	Toxicity	76.86	76.05	81.18	85.97	75.00	72.23	76.06
LLaMA	Intent	75.22	73.74	77.70	78.03	67.07	57.96	57.88
	Sent.	52.10	55.70	53.18	55.55	52.60	51.71	51.24
	Topic	75.64	74.52	77.06	73.90	71.33	65.65	66.24
	Toxicity	69.82	65.48	79.32	85.40	64.58	60.16	60.17

Table 32: Mean AUROC with **Soft Whitening** across languages for each model and task, organised by detector.

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	70.21	70.62	73.68	69.94	63.83	52.79	50.37
	Sent.	51.19	55.71	53.33	54.46	51.15	48.89	47.85
	Topic	67.12	70.55	73.57	70.23	69.80	50.91	50.49
	Toxicity	60.88	56.28	73.18	78.35	54.61	51.34	52.08
E5	Intent	85.95	84.79	86.47	83.01	81.68	66.41	67.77
	Sent.	70.60	68.34	69.82	66.30	70.92	62.55	67.19
	Topic	78.50	79.70	79.36	78.15	73.86	58.73	62.72
	Toxicity	78.85	72.99	83.58	87.86	73.71	62.56	66.66
Qwen3	Intent	74.83	73.93	77.02	75.32	65.11	53.95	52.66
	Sent.	51.70	55.15	53.27	54.93	51.60	49.78	50.66
	Topic	74.12	73.05	75.10	70.59	70.95	59.52	56.60
	Toxicity	62.65	58.15	74.89	83.20	56.94	54.07	52.97
Qwen3-E	Intent	84.04	81.68	84.31	83.08	79.41	63.09	64.82
	Sent.	63.82	62.92	63.02	62.10	61.88	57.17	61.02
	Topic	81.59	81.81	81.66	77.72	74.96	56.52	61.58
	Toxicity	74.28	69.85	75.80	85.94	68.47	61.51	65.62
LLaMA	Intent	74.74	71.95	77.52	77.57	65.83	54.20	52.94
	Sent.	52.08	55.52	53.87	55.15	52.07	49.24	48.96
	Topic	75.21	73.34	76.54	73.58	70.54	58.86	56.33
	Toxicity	68.01	62.95	77.82	85.38	61.92	56.15	55.42

Table 33: Mean AUROC with **Zeroing out outlier dimension post-processing** across languages for each model and task, organised by detector.

Model	Task	KNN	LOF	LUNAR	GMM	AE	IF	OCSVM
XLM-R	Intent	72.04	71.71	73.64	69.73	63.42	52.35	50.54
	Sent.	51.15	56.76	52.92	54.21	50.94	47.28	48.84
	Topic	68.21	69.59	73.85	70.43	69.92	51.18	51.11
	Toxicity	61.35	56.19	73.57	78.28	54.84	51.44	51.66
E5	Intent	85.88	84.84	86.36	83.00	80.50	66.64	67.74
	Sent.	70.89	68.37	70.06	66.56	71.02	60.52	67.39
	Topic	78.51	81.11	79.41	78.11	73.77	58.49	62.73
	Toxicity	78.84	73.15	83.56	87.87	73.86	62.99	66.70
Qwen3	Intent	74.30	73.68	76.95	75.24	65.20	54.07	53.07
	Sent.	51.92	55.54	53.30	55.08	51.78	50.01	51.15
	Topic	73.61	73.16	75.12	70.62	70.82	59.06	55.16
	Toxicity	62.59	58.66	76.21	83.23	56.94	54.56	52.81
Qwen3-E	Intent	83.80	81.29	84.07	82.77	79.06	63.65	64.46
	Sent.	63.99	63.12	63.11	62.26	62.14	57.27	61.11
	Topic	81.60	81.84	81.69	77.75	75.19	56.65	61.43
	Toxicity	74.31	72.48	75.76	85.93	68.44	64.08	65.50
LLaMA	Intent	74.84	72.33	77.53	77.70	66.02	54.27	53.14
	Sent.	52.15	55.56	53.79	55.51	51.87	49.24	48.74
	Topic	75.23	73.37	76.61	73.59	70.52	58.02	56.51
	Toxicity	68.03	62.90	77.93	85.38	61.98	55.79	55.36

Table 34: Mean AUROC with **Zeroing out top 10 outlier dimension post-processing** across languages for each model and task, organised by detector.

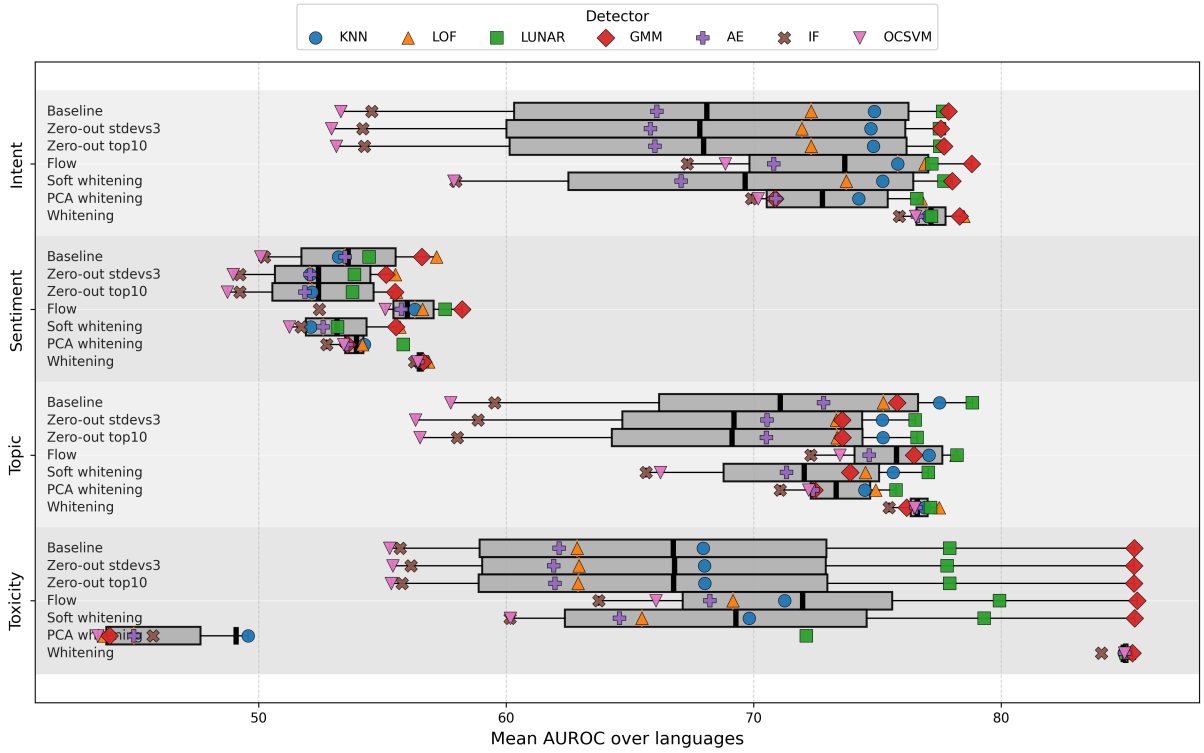

Figure 12: Boxplots of AUROC performance across post-processing methods applied to `Llama`. Each box aggregates results over detectors and tasks, averaged across languages.

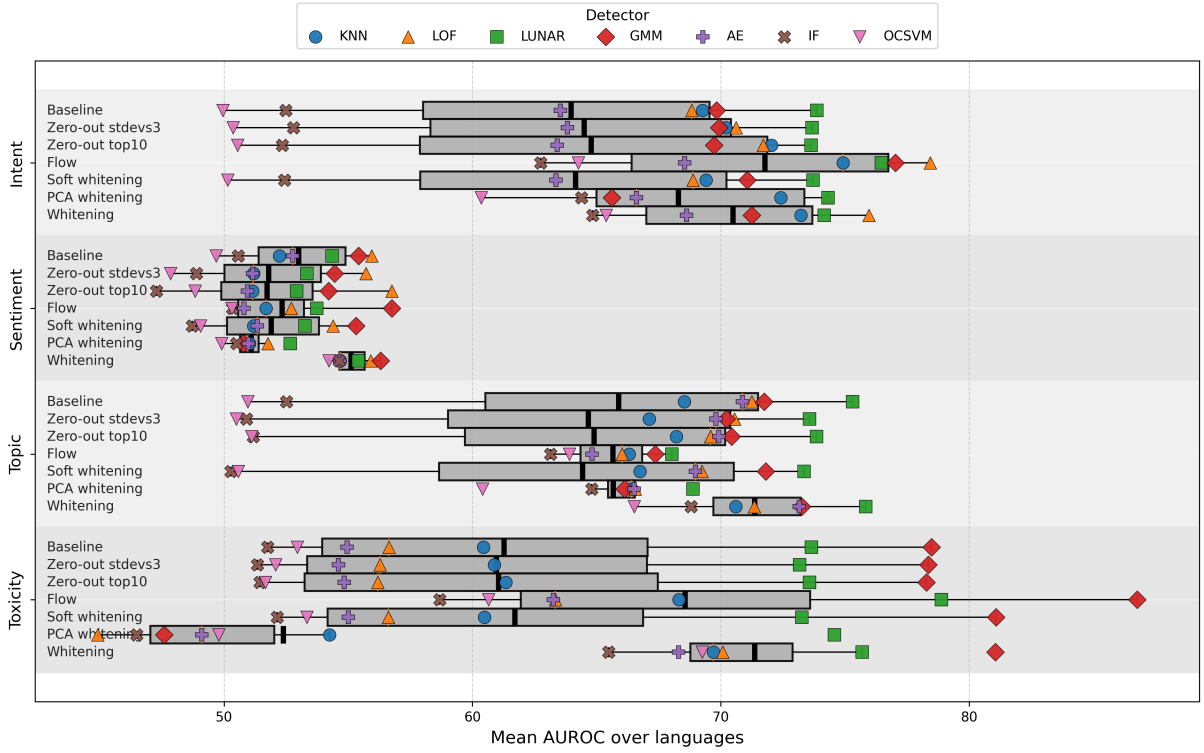

Figure 13: Boxplots of AUROC performance across post-processing methods applied to `XLM-RoBERTa-base`. Each box aggregates results over detectors and tasks, averaged across languages.

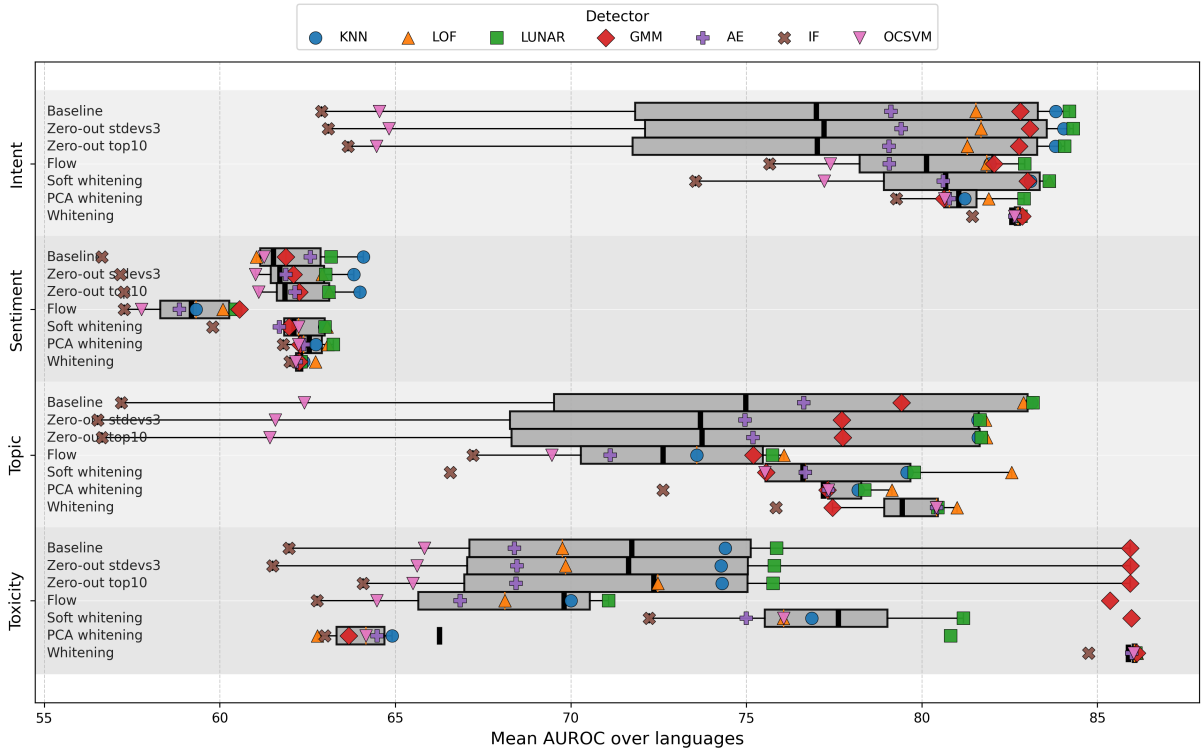

Figure 14: Boxplots of AUROC performance across post-processing methods applied to `Qwen3-Embeddings`. Each box aggregates results over detectors and tasks, averaged across languages.

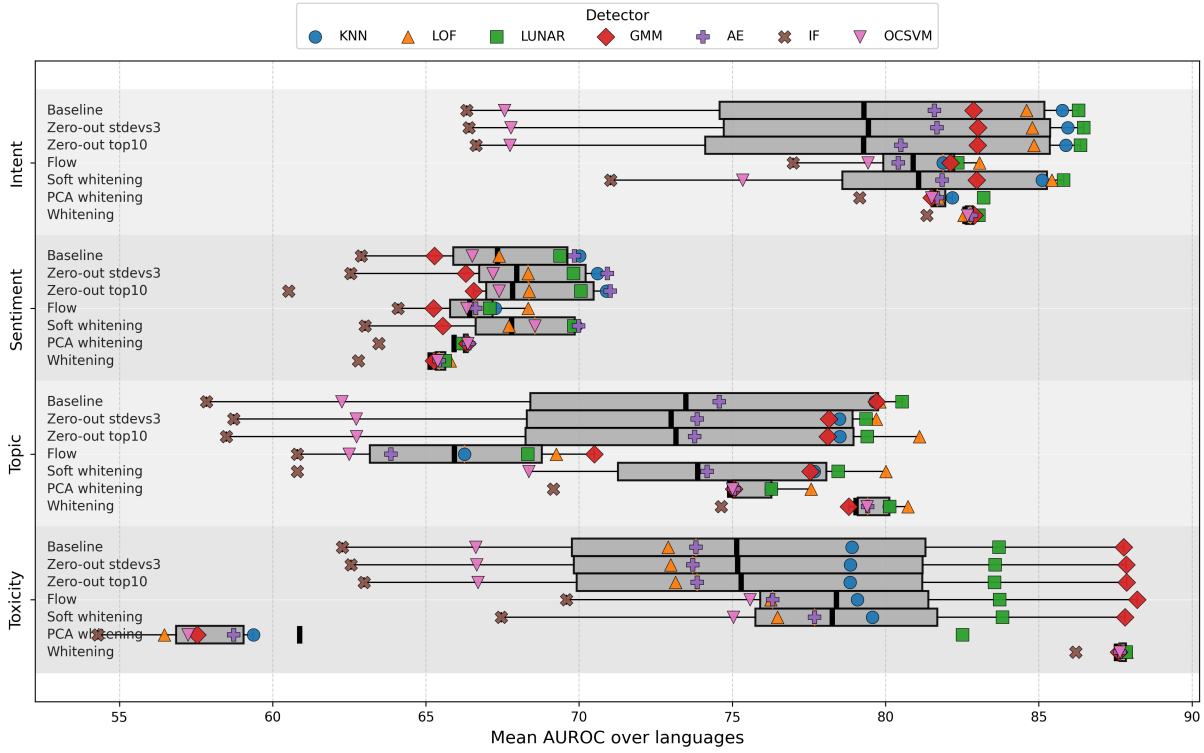

Figure 15: Boxplots of AUROC performance across post-processing methods applied to `E5`. Each box aggregates results over detectors and tasks, averaged across languages.

Model	Post-processing	Mean anisotropy
XLM-RoBERTa	Whitening	0.00043
	Soft-whitening	0.00452
	PCA-whitening	0.00057
	Flow	0.04005
	Zero_out_10dims	0.63442
	Zero_out_stdev	0.75949
E5	Whitening	0.00009
	Soft-whitening	0.00002
	PCA-whitening	0.00017
	Flow	0.01083
	Zero_out_10dims	0.78802
	Zero_out_stdev	0.79616
LLaMA	Whitening	0.00012
	Soft-whitening	0.00084
	PCA-whitening	0.00027
	Flow	0.00432
	Zero_out_10dims	0.83009
	Zero_out_stdev	0.81215
Qwen3-Embeddings	Whitening	0.00012
	Soft-whitening	0.00026
	PCA-whitening	0.00029
	Flow	0.00286
	Zero_out_10dims	0.34666
	Zero_out_stdev	0.34631
Qwen3	Whitening	0.00015
	Soft-whitening	0.00172
	PCA-whitening	0.00037
	Flow	0.00357
	Zero_out_10dims	0.73706
	Zero_out_stdev	0.67810

Table 35: Average anisotropy scores by model and post-processing methods

A.9 Experiment with variance normalization

To isolate the impact of improved isotropy on performance, we also experimented with a simple variance normalization. Results in Table 36 show that this transformation has only a limited effect on performance and does not lead to a homogenization across anomaly detection algorithms, which supports our hypothesis.

Model	**Task**	**KNN**	**LOF**	**LUNAR**	**GMM**	**AE**	**IF**	**OCSVM**
XLM-R	Intent	71.99	71.90	73.60	71.10	63.37	52.46	53.01
	Sent.	52.44	57.01	54.13	55.89	52.08	50.12	49.59
	Topic	73.49	70.86	74.87	73.41	70.30	52.20	58.48
	Toxicity	60.80	57.34	69.20	80.76	54.80	51.80	52.85
E5	Intent	86.06	85.08	86.39	83.09	81.68	66.41	71.17
	Sent.	71.16	69.46	70.37	65.77	71.46	63.38	68.47
	Topic	79.26	79.71	79.75	78.68	73.71	56.83	64.07
	Toxicity	78.88	73.52	83.58	87.81	73.69	62.25	71.41
Qwen3	Intent	74.79	74.39	76.89	75.49	64.94	53.89	54.34
	Sent.	52.39	55.53	53.86	55.69	52.30	50.26	51.22
	Topic	75.32	73.30	76.32	68.86	71.82	59.83	65.07
	Toxicity	63.21	58.89	76.12	83.27	57.07	54.22	55.08
Qwen3-E	Intent	83.09	81.23	83.94	82.67	78.82	62.76	68.01
	Sent.	64.10	62.58	64.27	63.00	63.86	57.44	61.77
	Topic	83.01	82.86	83.61	77.61	76.61	56.76	64.13
	Toxicity	74.04	69.97	75.81	85.93	68.33	61.85	69.27
LLaMA	Intent	75.20	72.61	77.75	78.22	66.04	54.40	55.16
	Sent.	53.14	56.48	54.85	56.55	53.21	49.67	50.94
	Topic	77.51	75.01	78.80	75.31	72.62	59.47	64.36
	Toxicity	68.65	63.63	77.90	85.44	62.12	55.61	58.48

Table 36: Mean AUROC with **Variance Normalization** across languages for each model and task, organised by detector.