# From Topology to Retrieval: Decoding Embedding Spaces with Unified Signatures

**Florian Rottach**[*1,2] **William Rudman**[*3] **Bastian Rieck**[4] **Harrisen Scells**[1]

**Carsten Eickhoff**[1]

[1]University of Tübingen [2]Boehringer Ingelheim GmbH, Biberach (Riss), Germany
[3]The University of Texas at Austin [4]Fribourg University
florian.rottach@boehringer-ingelheim.com

## Abstract

Studying the shape of text embedding spaces enhances our understanding of model behavior and how geometry affects downstream task performance. While a range of topological and geometric metrics have been explored, they are typically analyzed in isolation, and the connections between them remain largely unknown. In this paper, we present **Unified Topological Signatures (UTS)** for characterizing the structure of embedding spaces. We evaluate our method across a wide range of text embedding models and retrieval datasets, finding that it outperforms individual metrics in downstream performance prediction. Through several experiments, we demonstrate that the signatures characterize embedding spaces more accurately, improve the prediction of model-specific properties and biases, and reveal novel insights into representational similarity that challenge the global claims of the Platonic Representation Hypothesis.

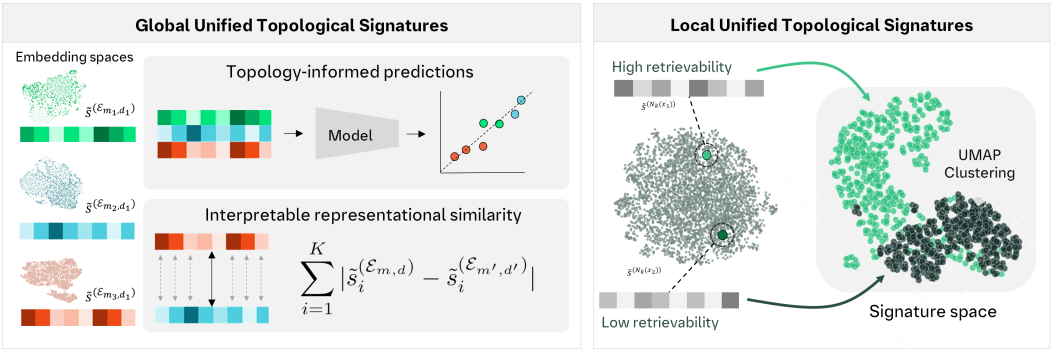


Figure 1: *Unified Topological Signatures* of embedding spaces. Left: We construct signature vectors for embedding spaces by computing various topological descriptors. We use the vectors for downstream prediction tasks as well as for measuring representational similarity. Right: We compute local signatures based on the neighborhood of individual embeddings and use them to detect retrievability bias in document corpora.

## 1 Introduction

Text embeddings from transformer models are central for many tasks ranging from semantic search [97] to text classification [15] and clustering [62]. They often play a crucial role in Retrieval-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Augmented Generation (RAG) [46], which combines parametric knowledge of a Large Language Model (LLM) with external knowledge from a search engine, allowing LLMs to respond with information not present in their training data. Dense retrieval models (i.e., architectures trained so that query embeddings are highly similar to embeddings of relevant documents) are a natural choice for RAG and other information retrieval tasks due to their speed and effectiveness. Selecting the *best* model for such a task remains an open challenge. While benchmarks such as Massive Text Embedding Benchmark (MTEB) [56] and theoretical frameworks [10, 16] provide valuable tools for model comparison and selection, they do not predict or explain effectiveness on novel tasks. To gain deeper insight into representations, prior work has analyzed the geometry of word embeddings [18] and internal representations of LLMs [27]. Notably, studies have shown that effective LLM embeddings often exhibit *anisotropy* [63], with representations concentrating along a few principal axes. Additionally, evidence from the Platonic Representation Hypothesis [33] points to geometry as a key driver of model behavior, suggesting that representations from different models are converging. While these studies offer valuable insights, representational attributes are often analyzed in isolation and text embedding models remain underexplored, which limits our understanding of how geometry affects downstream task performance.

In this paper, we address this gap by investigating the high-dimensional geometry of text embedding spaces to better understand what drives model effectiveness. Our main **contributions** are as follows:

1. We systematically study the combination of multiple geometric descriptors through *Unified Topological Signatures* (UTS), showing that multidimensional signatures outperform individual metrics in predicting retrieval effectiveness and retrievability bias.
2. We demonstrate that embedding geometry is fundamentally multidimensional, with UTS providing distinctive geometric fingerprints of models and datasets.
3. We evaluate the global alignment claims of the Platonic Representation Hypothesis using UTS-based similarity. Rather than convergence by scale, we observe clustering by model family, highlighting the influence of architecture and training data on global embedding-space geometry.

## 2 Related Work

**Geometry of Embeddings** Early studies have explored the internal representations of BERT [18], focusing on how their geometric properties influence word senses [64]. Other research has investigated isotropy [23], isometry [20], clustering patterns, and intrinsic dimensions [9, 77] of contextualized word embeddings. In evaluation studies, geometry has been identified as one of the key properties driving embedding quality [82]. Recently, attention has shifted towards analyzing the internal representations of LLMs, which have been shown to be effective at many tasks without any fine-tuning [29, 75]. Several works have studied the training dynamics of isotropy [63, 68] and intrinsic dimension [11, 35, 70], as well as the topological differences across layers [27, 50]. Recently, Lee et al. [41] found a connection between linguistic compositionality and the complexity of data manifolds. Theoretical work by Weller et al. [87] suggests that dimensionality is a key driver for the effectiveness of embedding models. While a large body of research is focused on sentence or generally text representations, others have analyzed intrinsic dimension and other properties specifically for token embeddings [36, 78]. Notably, Lee et al. [40] found that token embeddings across different models share a common geometric structure. Unlike prior works examining individual geometric metrics, we present a *unified* representation that incorporates the full topological profile into a signature vector, offering a more stable and rigorous summary.

**Embedding-based Retrieval** Dense retrieval has emerged as a powerful alternative to traditional lexical-based retrieval methods like BM25 [66], which are based on sparse vector representations [97]. In recent years, multiple attempts have been made to understand the geometry of embedding spaces in the context of retrieval, such as investigating the influence of local intrinsic dimension on nearest neighbor search [3] and hierarchical navigable small worlds (HNSW) algorithms [22]. In addition, theoretical work has shown that the embedding space dimensionality is a bounding factor for retrieval effectiveness [87]. However, a more complete topological perspective has not been considered for retrieval performance prediction. In addition, we study the retrievability bias [89] in dense retrieval systems, which describes the phenomenon that some documents are easier to retrieve than others. Studies have found that larger bias corresponds to poorer retrieval performance [89], hence identifying

such documents can improve retrieval systems through techniques like document augmentation [73] and other optimization strategies [71]. Currently, no works have attempted to link retrievability bias to geometric properties of the embedding space. We show that we can predict whether or not a document is retrievable, indicating the importance of the geometric analysis of representations of dense retrieval models.

**Representational Similarity and Alignment** From early work on neural pattern similarity to recent advances in cross-modal alignment in transformers, representational similarity has been central to machine learning [74]. There exist a variety of methods to compare representations, each assessing similarity with distinct aspects of granularity, symmetry, or sensitivity. In recent years, Centered Kernel Alignment (CKA) [38], has been widely adopted for computing similarity, particularly due to its robustness under orthogonal transformations and isotropic scaling. Caspari et al. [10] use CKA to identify architecture as a key driver for embedding similarity, while Ciernik et al. [13] find the objective function determining representational consistency for vision models. The similarity of representations is also at the core of the *Platonic Representation Hypothesis* [33], which argues that representations converge and foundation models become more aligned. The findings of this study indicate that the degree of alignment is shaped by the choice of similarity metric, underscoring its critical role in evaluating representational consistency. While CKA is a powerful statistical tool, it may miss topological differences in embedding spaces [17], particularly on a global scale. Recent work has started to address this gap [5, 34], but these methods focus on isolated aspects and offer little explanation for why representations diverge.

To address fundamental limitations of individual metrics, we investigate the potential of multidimensional signatures that summarize topology, intrinsic dimensionality, clustering behavior, and other measurable aspects of embedding spaces. In contrast to current similarity methods, we can pinpoint in what aspects embedding spaces differ. Furthermore, by comparing the global topology, we are not restricted to paired embedding spaces, which is a requirement for most representational similarity methods. Finally, we examine whether the Platonic Representation Hypothesis holds under our UTS-based similarity measure, providing evidence that global alignment is shaped by other factors than by model scale.

## 3 Experimental Setup and Methodology

### 3.1 Unified Topological Signatures

**Definition** To provide a multidimensional characterization of embedding spaces, we aggregate topological and geometric descriptors into a vector representation. In the following, we define global and local variants, outline normalization, and introduce an interpretable similarity measure.

Let $T_i$ be a global descriptor of a point cloud in an $n$-dimensional real-valued space $X \subset \mathbb{R}^n$, where $i \in \{1, \dots, K\}$ and $K$ is the number of descriptors. Each descriptor is a real-valued function that captures a specific topological or geometric property of the point cloud, i.e.,

$$T_i \colon X \to s_i^X \in \mathbb{R} \tag{1}$$

Let $\mathcal{E}_{m,d} \subset \mathbb{R}^n$ be the $n$-dimensional embedding space, generated by model $m \in \mathcal{M}$ on dataset $d \in \mathcal{D}$. We define the global Unified Topological Signatures (UTS) of $\mathcal{E}_{m,d}$ as a vector:

$$\boldsymbol{s}^{\mathcal{E}_{m,d}} = [T_1(\mathcal{E}_{m,d}), T_2(\mathcal{E}_{m,d}), \dots, T_K(\mathcal{E}_{m,d})]^\top \tag{2}$$

where each component $\boldsymbol{s}_i^{\mathcal{E}_{m,d}}$ corresponds to the value of distinct topological descriptor $T_i \in T$, with $i \in \{1, ..., K\}$. Table 1 summarizes the selected descriptors, which cover a wide range of properties. Details and mathematical definitions for each descriptor are provided in Appendix A.

Each descriptor $T_i$ is computed over a varying sample of $n_i$ data points, drawn from $\mathcal{E}_{m,d}$, with $n_i$ chosen empirically to balance descriptor robustness and computational feasibility. This ensures that each descriptor captures meaningful topological features while being computationally tractable.

**Local Signatures** We define a local variant of the signatures of a point $\boldsymbol{x} \in \mathcal{E}_{m,d}$ as:

$$\boldsymbol{s}^{\mathcal{N}_k(x)} = [T_1(\mathcal{N}_k(x)), \dots, T_K(\mathcal{N}_k(x))]^\top \tag{3}$$

where each component $\boldsymbol{s}_i^{\mathcal{N}_k(x)}$ corresponds to the value of $T_i$ computed over the $k$-nearest neighbors of $\boldsymbol{x}$ denoted by $\mathcal{N}_k(\boldsymbol{x}) \subset \mathcal{E}_{m,d}$.

**Normalization** To enable comparability across models and datasets, we normalize each signature vector computed for a point cloud $X$ by dividing its components by the maximum absolute observed value across all embedding spaces. Other normalization variants are evaluated in Appendix K.

$$\tilde{s}_i^X = \frac{s_i^X}{\max_{X'} |s_i^{(X')}|} \quad \text{for } i = 1, \ldots, K \tag{4}$$

yielding the normalized signature vector:

$$\tilde{\boldsymbol{s}}^X = [\tilde{s}_1^X, \tilde{s}_2^X, \ldots, \tilde{s}_K^X]^\top \tag{5}$$

Instead of normalizing across all signatures, this normalization can be applied to a subgroup of signatures, such as vectors computed on a specific dataset.

**Similarity** Given two normalized signature vectors $\tilde{\boldsymbol{s}}^{\mathcal{E}_{m,d}}$ and $\tilde{\boldsymbol{s}}^{\mathcal{E}_{m',d'}}$, the similarity between the corresponding embedding spaces regarding property $i$ is quantified using the Manhattan distance:

$$\boldsymbol{\delta_i}^{(\mathcal{E}_{m,d}),(\mathcal{E}_{m',d'})} = \left| \tilde{\boldsymbol{s}}_{\boldsymbol{i}}^{(\mathcal{E}_{m,d})} - \tilde{\boldsymbol{s}}_{\boldsymbol{i}}^{(\mathcal{E}_{m',d'})} \right| \tag{6}$$

We define a global similarity measure as the sum of these differences:

$$d(\tilde{\boldsymbol{s}}^{(\mathcal{E}_{m,d})}, \tilde{\boldsymbol{s}}^{(\mathcal{E}_{m',d'})}) = \sum_{i=1}^{K} |\tilde{s}_i^{(\mathcal{E}_{m,d})} - \tilde{s}_i^{(\mathcal{E}_{m',d'})}| \tag{7}$$

We choose the Manhattan distance because it quantifies absolute differences without applying transformations and allows for a decomposition of element-wise contributions to embedding-space divergence, which is not possible with metrics such as Cosine similarity.

### 3.2 Datasets

Due to the increasing adoption of dense retrieval and its widespread applications in RAG we focus our analysis on the retrieval datasets shown in Table 2, which are included in the MTEB evaluation framework. Formally, let $q$ denote a natural language query and $d_i$ denote a document from a large collection $D = \{d_i\}_{i=1}^m$ consisting of $m$ documents. Given a query, the goal of text retrieval is to return a ranked list of the $n$ most relevant documents $L = [d_1, d_2, \ldots, d_n]$ according to the relevance scores produced by a retrieval model. Dense retrieval models represent both queries and documents as dense vectors in a shared semantic space. The relevance score between a query and a document is then computed using a similarity function between their dense representations:

$$\mathrm{Rel}(q, d) = f_{\mathrm{sim}}(\phi_q(q), \phi_d(d))$$

where $\phi_q(\cdot)$ and $\phi_d(\cdot)$ are neural encoders mapping queries and documents into $n$-dimensional vectors, and $f_{\mathrm{sim}}$ denotes the similarity function. We conduct our experiments on 11 datasets, spanning multiple domains and corpus sizes, as detailed in Table 2.

### 3.3 Retrievability bias

A document's retrievability [89] is defined as the ease with which it can be retrieved by a given retrieval system across a set of general queries. More formally, the retrievability score $r(d)$ of a document $d$ is defined as:

$$r(d) \propto \sum_{q \in Q} f(k_{dq}, c)$$

where $Q$ is a set of queries, $k_{dq}$ is the rank of document $d$ in the retrieval results for query $q$, $c$ is a cutoff rank (e.g., top-10), and $f(k_{dq}, c)$ is a function that assigns higher weights to documents appearing at higher ranks. A common choice for $f$ is

$$f(k_{dq}, c) = \begin{cases} 1 & \text{if } d \text{ is in the top } c \text{ documents given } q, \\ 0 & \text{otherwise.} \end{cases}$$

The retrievability of a document is primarily determined by its similarity to a query. For most datasets, documents with a high query alignment (i.e., those that closely match the terms or intent of

| Measure | Category | #Samp. | Src. |
|---|---|---|---|
| PH Dimension | Homology | 5K | [2] |
| PH Statistics | Homology | 5K | [21] |
| Persistence Entropy | Homology | 20K | [12] |
| Euler Characteristic | Homology | 5K | [6] |
| TwoNN Dimension | Intrinsic Dimension | 50K | [24] |
| PCA Dim. | Intrinsic Dimension | 50K | [26] |
| Effective Rank | Intrinsic Dimension | 100K | [67] |
| Magnitude Dimension | Diversity | 5K | [45] |
| Magnitude Area | Diversity | 5K | [49] |
| Spread | Diversity | 10K | [91] |
| Vendi Score | Diversity | 20K | [25] |
| Pairwise Similarity | Density | 50K | N/A |
| Uniformity | Uniformity | 20K | [85] |
| IsoScore | Isotropy | 500K | [69] |
| Silhouette | Clustering | 20K | [72] |

Table 1: Descriptors for signature vectors.

| Dataset | Domain | #Q | #Docs | Src. |
|---|---|---|---|---|
| MS MARCO | Web | 27 | 8.8M | [57] |
| ArguAna | Argument | 1.4K | 8.7K | [80] |
| Touche2020 | Argument | 249 | 528K | [76] |
| NFCorpus | Medical | 323 | 3.6K | [8] |
| TREC-COVID | Medical | 50 | 171K | [79] |
| SciFact | Science | 300 | 5.2K | [81] |
| SCIDOCS | Science | 1K | 25.6K | [14] |
| ClimateFEVER | Science | 154 | 5.4M | [19] |
| CQADupstack | Forum | 1.6K | 40K | [32] |
| QuoraRetrieval | Forum | 10K | 523K | [76] |
| FiQA2018 | Finance | 648 | 57K | [51] |

Table 2: Summary of retrieval datasets.

| Model | Dim. | Size | Src. |
|---|---|---|---|
| gte-Qwen2-1.5B | 1536 | 1.5B | 47 |
| gte-Qwen2-7B | 3584 | 7B | 47 |
| stella-en-1.5B-v5 | 1024 | 1.5B | 95 |
| Qwen3-Emb-0.6B | 1024 | 0.6B | 96 |
| Qwen3-Emb-4B | 2560 | 4B | 96 |
| Qwen3-Emb-8B | 4096 | 8B | 96 |
| Linq-Emb-Mistral | 4096 | 7.3B | 37 |
| SFR-Emb-Mistral | 4096 | 7.3B | 53 |
| e5-mistral-7b | 4096 | 7.3B | 83 |
| bge-small-en-v1.5 | 384 | 0.02B | 93 |
| bge-base-en-v1.5 | 768 | 0.1B | 93 |
| bge-large-en-v1.5 | 1024 | 0.3B | 93 |
| me5-large-instruct | 1024 | 0.56B | 84 |
| arctic-embed-m-v1.5 | 768 | 0.1B | 54 |
| ember-v1.0 | 1024 | 0.3B | 59 |
| mxbai-xsmall-v1 | 384 | 0.1B | 43 |
| mxbai-large-v1 | 512 | 0.3B | 44 |
| all-MiniLM-L6-v2 | 384 | 0.02B | 86 |
| gtr-t5-base | 768 | 0.1B | 58 |
| gtr-t5-large | 768 | 0.3B | 58 |
| gtr-t5-xxl | 768 | 4.8B | 58 |
| gte-small | 384 | 0.3B | 47 |
| gte-base | 768 | 0.1B | 47 |
| gte-large | 1024 | 0.3B | 47 |
| msmarco-roberta | 768 | 0.1B | 94 |
| msmarco-distilbert | 768 | 0.06B | 31 |
| gemini-emb-001 | 3072 | N/A | 42 |
| openai-emb-3-small | 1536 | N/A | 60 |
| openai-emb-3-large | 3072 | N/A | 60 |

Table 3: Embedding models.

typical queries) will be more retrievable than documents that lie outside the typical query space. In dense retrieval systems documents and queries are embedded into a shared space in which similarity is influenced by the choice of embedding model. MTEB reveals that retrievability varies across embedding models despite a constant set of queries, underscoring the influence of the embedding space structure. Therefore, we expect significant differences in the topological patterns of different documents, driven by linguistic and semantic properties.

### 3.4 Embedding models

We use 27 text embedding models from different families, including Qwen [96], E5 [83], T5 [58], BGE [93], Gemini [42] and others as listed in Table 3. The models vary in terms of size, embedding dimension, architecture, and training data. We use the models' default text preprocessing and tokenization methods and normalize all embeddings to unit length. For Gemini, we specifically use the embeddings optimized for retrieval tasks. Across every model-dataset pair, we generate document and query embeddings using the MTEB framework. Open-source models are downloaded and executed via HuggingFace [92], while embeddings for the OpenAI and Google models are produced through their respective APIs.

### 3.5 Computation of signatures

For each embedding space generated by model $m$ on dataset $d$, we compute a topological signature as defined by Equation 2. The descriptors cover a wide range of properties commonly found in other works, including homology, intrinsic dimension, diversity, density, isotropy, and clustering. For descriptors utilizing pairwise-distances, we compute both the Euclidean and Cosine distance functions. As some methods are intractable to calculate for millions of embeddings, we perform a preliminary analysis to assess the robustness of each descriptor with respect to sample size, as discussed in Appendix B. We then sample $n_i$ datapoints for each descriptor $T_i$ and compute the corresponding value. To produce a larger dataset with more variability, we repeat this procedure for 3 different seeds, such that we have 891 embedding space signatures for 27 models and 11 datasets. We then normalize each signature as described above. All computations were performed using parallel processing on 8 GPU compute nodes.

### 3.6 Experiments

**Topology-informed downstream prediction** We fit a Random Forest model from scikit-learn [39] to predict recall, mean average precision (MAP), and normalized discounted cumulative gain (NDCG)

for all datasets using cutoff values of $k = \{5, 20, 100\}$ with the UTS vectors as input features. The retrieval measures have been computed using MTEB [56], and z-normalization is applied to make the metric scales comparable between datasets. Given the limited training data, we use 3-fold group cross-validation to predict on unseen datasets and evaluate the correlation using Spearman's $\rho$. We ensure that all data preprocessing steps are based on the training data distribution to avoid data leakage. To prevent overfitting, we limit the maximum depth of the decision trees to 5. We assess the prediction quality, by comparing UTS against the following single metric descriptors, frequently used in other works: TwoNN intrinsic dimension, effective rank, mean pairwise similarity and IsoScore.

To explore the effects of embedding space structure on retrievability bias in more detail, we compute the retrievability with $c = 100$ for all documents in a corpus using all available queries from the respective dataset. We then compute local signatures (see Equation 3 with $k = 100$) for the 100 most and least retrieved documents $x$ in the embedding space. This results in 59400 signatures, which we use as features to fit a binary Random Forest classifier to predict whether a document is highly retrievable. We follow the same cross-validation strategy as for the global signatures and include several baselines: the local density (mean pairwise similarity), the local intrinsic dimension, which has been found to influence retrieval [3] as well as the linguistic feature word count, which has been linked to retrievability in previous studies [88]. In addition, we run a baseline experiment predicting retrievability directly from the raw embeddings. Finally, we analyze the distribution of the retrievability scores to identify the overall retrievability bias. To measure the *inequality* of the distribution, we compute the Gini coefficient [90] and study the differences between models.

**Characterizing embedding spaces** We compare UTS against individual descriptors in their ability to separate embedding spaces. For this, we cross-validate a Random Forest classifier to predict the embedding model and model family corresponding to an unseen embedding space. This allows us to measure how discriminative the topological signatures are. In addition, we investigate whether model-specific information is encoded in the geometry of the embedding space by predicting the architecture and activation function used in text embedding models. This approach is particularly interesting for proprietary models, where such information is not publicly available. We include random predictions and the most frequent class as trivial baselines. We further evaluate individual topology descriptors, namely the TwoNN intrinsic dimension, effective rank, average pairwise similarity, and IsoScore. To investigate which properties contribute most to topological variation, we perform a Principal Component Analysis (PCA) on the global UTS vectors and analyze correlations between the individual descriptors. We also include a PCA-transformed version of UTS as an additional baseline. Since some properties are affected by class imbalance, we use balanced accuracy as the evaluation metric.

**Topological clustering** We test the Platonic Representation Hypothesis by analyzing whether embedding space geometry converges with increasing scale. For this, we compute global UTS distances between models according to Equation 7. We average the distances between models across all datasets, yielding a 27x27 distance matrix. To confirm whether larger models have more similar topologies, we group models into the categories small, medium and large using the parameter count thresholds of $< 0.5$B, $0.5 - 3$B and $> 3$B parameters, respectively. To test alternative drivers of representational similarity, we also group the distances by model family, investigating if architectural properties and training data explain the differences. The source code to generate the embedding spaces and reproduce our experiments is available at `https://github.com/Boehringer-Ingelheim/unified-topological-signatures`.

## 4 Results

**No single metric tells the whole story.** As summarized by Table 4, we find that UTS significantly outperform individual metrics in retrieval performance prediction. The best performance is achieved for Recall@20 with a mean Spearman correlation of $0.70 \pm 0.11$ across the cross-validation folds. The best single metric is effective rank with a correlation of $0.47 \pm 0.08$. These results highlight the limitations of relying on any single geometric descriptor. Our multi-attribute perspective, by contrast, captures more complex structural patterns, translating into substantially stronger predictive power. To better understand these patterns, we analyze the feature importance across all folds, summarized in Figure 2 and find that the effective rank is the most important feature. This result is in line with recent theoretical work [87] suggesting that retrieval performance is limited by the dimensionality of

Table 4: Spearman's $\rho$ for retrieval performance prediction.

| | Recall | | | MAP | | | NDCG | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | @5 | @20 | @100 | @5 | @20 | @100 | @5 | @20 | @100 |
| TwoNN | $0.07_{\pm.09}$ | $0.08_{\pm.11}$ | $0.04_{\pm.09}$ | $0.07_{\pm.11}$ | $0.06_{\pm.11}$ | $0.04_{\pm.09}$ | $0.06_{\pm.11}$ | $0.07_{\pm.10}$ | $0.05_{\pm.09}$ |
| Eff. rank | $0.44_{\pm.11}$ | $0.47_{\pm.08}$ | $0.42_{\pm.09}$ | $0.44_{\pm.09}$ | $0.43_{\pm.08}$ | $0.43_{\pm.07}$ | $0.45_{\pm.09}$ | $0.45_{\pm.08}$ | $0.45_{\pm.07}$ |
| Mean sim. | $0.17_{\pm.13}$ | $0.18_{\pm.15}$ | $0.20_{\pm.11}$ | $0.17_{\pm.11}$ | $0.18_{\pm.12}$ | $0.19_{\pm.13}$ | $0.20_{\pm.13}$ | $0.19_{\pm.14}$ | $0.19_{\pm.14}$ |
| IsoScore | $0.23_{\pm.15}$ | $0.34_{\pm.06}$ | $0.30_{\pm.07}$ | $0.24_{\pm.10}$ | $0.24_{\pm.11}$ | $0.25_{\pm.10}$ | $0.24_{\pm.10}$ | $0.27_{\pm.09}$ | $0.26_{\pm.10}$ |
| UTS-PCA | $0.49_{\pm.03}$ | $0.56_{\pm.05}$ | $0.55_{\pm.10}$ | $0.49_{\pm.01}$ | $0.49_{\pm.01}$ | $0.50_{\pm.02}$ | $0.48_{\pm.03}$ | $0.52_{\pm.01}$ | $0.52_{\pm.02}$ |
| **UTS** | $\mathbf{0.64}_{\pm.16}$ | $\mathbf{0.70}_{\pm.11}$ | $\mathbf{0.67}_{\pm.12}$ | $\mathbf{0.63}_{\pm.17}$ | $\mathbf{0.66}_{\pm.14}$ | $\mathbf{0.68}_{\pm.15}$ | $\mathbf{0.66}_{\pm.14}$ | $\mathbf{0.69}_{\pm.13}$ | $\mathbf{0.68}_{\pm.14}$ |

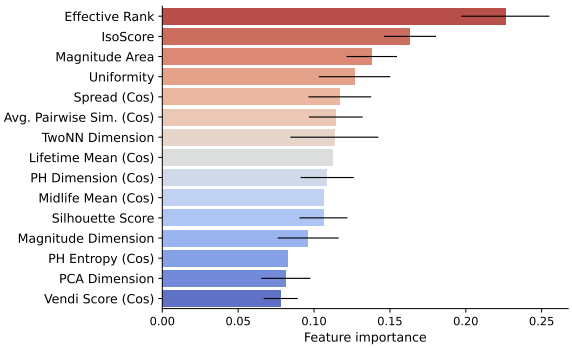


Figure 2: Feature importance across CV-folds.

the embeddings. Notably, we find that other descriptors also play a significant role, highlighting that the predictive signal is distributed across the full signature. We discuss more results and experiments for retrieval performance prediction in Appendix F.

**Local topological signatures predict document retrievability.** Our analysis reveals that local UTS serve as strong predictors of document retrievability, which can be a useful approach for identifying and mitigating bias in retrieval systems. We find that highly retrievable and non-retrievable documents are well separated in the local signature space, as visualized in Figure 3 using UMAP [52]. Figure 4 summarizes the results of using signature vectors compared to baseline models that only use a single feature. We find that the signature-based model achieves an average accuracy of 0.76 across all dataset cross-validation folds, significantly outperforming the baselines. This demonstrates that retrievability detection requires a holistic analysis of the local topology, beyond simple one-dimensional measures. In Appendix J, we provide additional analyses that highlight the structural differences in the embedding space between highly retrievable and non-retrievable documents. We also report model-specific accuracies, showing that OpenAI's text-embedding-3-large is most effective for the retrievability prediction task. Finally, we investigate the overall retrievability bias of embedding models using the Gini coefficient. This analysis reveals that decoder-based models exhibit greater bias (i.e., inequality in document exposure), across all datasets except ArguAna.

**UTS uniquely characterize embedding space geometry.** Table 5 summarizes the results for predicting different model attributes using the UTS as feature vectors. For unseen datasets, we can identify the specific model that generated an embedding space with a balanced accuracy of $0.66 \pm 0.02$ across all cross-validation folds, compared to a baseline accuracy of $0.04$, which corresponds to always predicting the most frequent model among the 27 classes. We perform the same analysis for predicting the model family, which is an easier task with an overall balanced accuracy of $0.68 \pm 0.07$ against a baseline accuracy of $0.17$. Our comparison with individual descriptors highlights that our multi-attribute approach significantly outperforms other methods for characterizing embedding spaces. Notably, effective rank is the best individual descriptor with a balanced accuracy of $0.39 \pm 0.02$ for predicting the model, which is around $40\%$ worse than our approach. Analyzing the feature importance shows that descriptors related to embedding dimension (e.g. effective rank) are most important for identifying a model. This strong discriminative power can be attributed to the fact that the embedding dimensions are tightly linked to model architecture, with some consistently exhibiting larger dimensions. To account for this, we run the experiments separately for each embedding

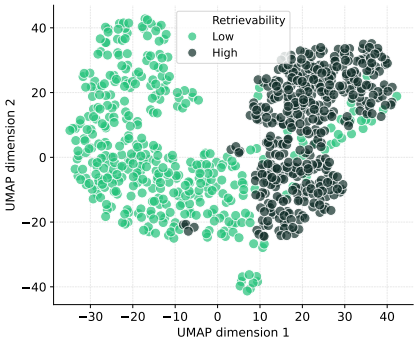


Figure 3: UMAP visualization of local UTS space for 100 highly retrievable and non-retrievable documents for QuoraRetrieval.

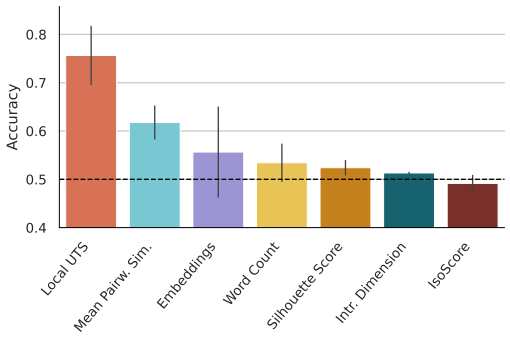


Figure 4: Cross-validation accuracy for predicting document retrievability. The dashed line indicates a random baseline.

Table 5: Cross-validated balanced accuracy of UTS versus baselines.

| Method | Model | Family | Architecture | Activation |
|---|---|---|---|---|
| Random | 0.04 | 0.08 | 0.50 | 0.33 |
| Most frequent | 0.04 | 0.17 | 0.72 | 0.59 |
| Embeddings | $0.26_{\pm.01}$ | $0.37_{\pm.02}$ | $0.78_{\pm.01}$ | $0.38_{\pm.01}$ |
| TwoNN dimension | $0.02_{\pm.00}$ | $0.09_{\pm.01}$ | $0.02_{\pm.00}$ | $0.32_{\pm.01}$ |
| Effective rank | $0.39_{\pm.02}$ | $0.47_{\pm.02}$ | $\mathbf{0.98}_{\pm.00}$ | $0.78_{\pm.02}$ |
| Mean similarity | $0.23_{\pm.05}$ | $0.46_{\pm.04}$ | $0.63_{\pm.04}$ | $0.56_{\pm.02}$ |
| IsoScore | $0.11_{\pm.03}$ | $0.16_{\pm.01}$ | $0.79_{\pm.03}$ | $0.53_{\pm.02}$ |
| UTS-PCA | $0.44_{\pm.03}$ | $0.50_{\pm.05}$ | $0.97_{\pm.01}$ | $0.83_{\pm.07}$ |
| **UTS** | $\mathbf{0.66}_{\pm.02}$ | $\mathbf{0.68}_{\pm.07}$ | $\mathbf{0.98}_{\pm.01}$ | $\mathbf{0.84}_{\pm.02}$ |

dimension with sufficient data (Table 6 in the Appendix) and find that the balanced accuracy increases to $0.89 \pm 0.10$ for 4096 dimensions, $0.68 \pm 0.23$ for 1024 dimensions, $0.73 \pm 0.12$ for 768 dimensions and $0.76 \pm 0.01$ for 384 dimensions. This shows that the UTS can uniquely characterize embedding models, even when controlling for embedding dimension. We discuss the prediction of architecture and activation function in Appendix G.

**Global topology is summarized by a few principal components.** The PCA reveals that $91\%$ of the variance is captured by the first four principal components, as shown in Figure 5. We inspect the loadings of each descriptor on the principal components and find that the first component mostly contains topological measures that are related to the pairwise distances between embeddings, including persistent homology features and kernel-based aggregations such as spread. The second component is dominated by measures related to dimensionality, isotropy, and clustering. These findings are supported by recent work establishing a connection between isotropy and clustering [55]. The remaining components contain information on various intrinsic dimension measures, as well as the effective rank. In addition to this variability analysis, we present a correlation analysis between all descriptors and model-specific properties in Appendix D. While we find that many descriptors are highly correlated, we observe a substantial performance gap between UTS-PCA and the full UTS representation, suggesting that a compressed set of principal components is not sufficient to capture all relevant topological information.

**Global topologies cluster by model family.** Linkage-based hierarchical clustering [1] reveals that model family - not size - is the dominant factor shaping representational similarity between text embedding models. To summarize this, we aggregate inter-family distances by averaging off-diagonal entries in the upper triangular part of the matrix (see Figure 6). Intra-family distances are consistently lower, indicating that architecture and training data are more influential than scale. Furthermore,

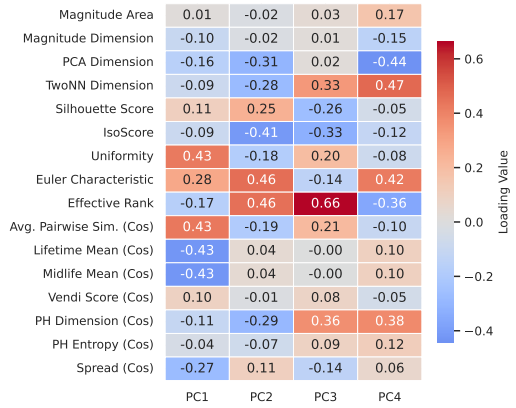


Figure 5: PCA analysis of global UTS revealing that embedding space geometry can be summarized by only a few principal components.

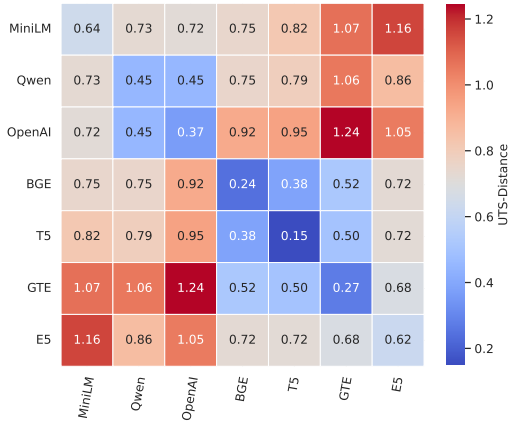


Figure 6: Model family dominates representational similarity, with lowest distances along the diagonal and clustering by architectures.

BERT-based models (BGE, T5, GTE) cluster tightly, as well as Qwen and OpenAI models, despite size variation within the family. These findings suggest that convergence in model representations is heavily shaped by architectural choices, challenging the assumption that scaling alone drives representational convergence. Our results are in line with recent work that shows that the Platonic Representation Hypothesis does not hold for global alignment [28]. In Appendix E we present additional experiments and compare our findings to similarity measured using CKA.

## 5 Discussion and Conclusion

**Discussion** Our results emphasize the value of a multidimensional perspective on embedding space geometry, demonstrating clear advantages over individual metrics. We show that embedding space structure predicts downstream performance and bias on unseen tasks, enabling the assessment of embedding space quality without task-specific ground-truth labels. We highlight a practical application by accurately predicting document retrievability, enabling targeted bias mitigation through document augmentation. We further demonstrate that UTS characterize embedding spaces and recover internal model properties with high accuracy, significantly outperforming all baselines and offering a novel method to decode characteristics of proprietary architectures. Finally, our results challenge the Platonic Representation Hypothesis: rather than converging toward a global geometry as models scale, alignment is mainly driven by architectural choices and training data. For practitioners, incorporating anisotropic regularization and larger embedding dimensions improves structural quality, while our topological framework serves as a monitoring tool for embedding space quality.

**Conclusion** In this work, we introduce Unified Topological Signatures, enabling a multidimensional view of embedding space geometry. We compute a diverse set of topological and geometric descriptors for a wide range of embedding models and retrieval datasets. We demonstrated that these signatures can be used to predict model properties, retrieval performance, and retrievability bias. Furthermore, our analysis reveals that models from the same family and size category share the most similar topological properties, indicating that architecture and training data have a strong influence on the embedding space geometry. Our findings suggest that topological signatures provide a powerful tool for understanding and optimizing text embedding spaces for retrieval tasks and beyond.

**Limitations and Future work** Our analysis is limited to retrieval tasks and English-language embeddings, leaving open how topological signatures behave in clustering, classification, or cross-lingual settings. Additionally, computing topological descriptors over large embedding spaces currently requires subsampling, which introduces approximation error and limits scalability. Addressing this would remove a practical bottleneck and enable analysis at larger scales. Beyond these limitations, further research is needed to understand how specific architectural choices and training data shape

embedding space geometry. Finally, extending the framework to other data modalities would provide broader evidence for the relationship between topology and downstream task performance.

## Acknowledgments

We thank I-Ting Ho, Jan Stanicki and Sebastian Schieferdecker for helpful discussions and feedback on this work. Computational resources were provided by the Boehringer Ingelheim High Performance Computing Cluster.

## References


[1] Margareta Ackerman and Shai Ben-David. A characterization of linkage-based hierarchical clustering. *Journal of Machine Learning Research*, 17(231):1–17, 2016.

[2] Henry Adams, Manuchehr Aminian, Elin Farnell, Michael Kirby, Joshua Mirth, Rachel Neville, Chris Peterson, and Clayton Shonkwiler. A fractal dimension for measures via persistent homology. In *Topological Data Analysis: The Abel Symposium 2018*, pages 1–31. Springer, 2020.

[3] Martin Aumüller and Matteo Ceccarello. The role of local intrinsic dimensionality in benchmarking nearest neighbor search. In *International Conference on Similarity Search and Applications*, pages 113–127. Springer, 2019.

[4] Jonathan Bac, Evgeny M Mirkes, Alexander N Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-dimension: a python package for intrinsic dimension estimation. *Entropy*, 23(10):1368, 2021.

[5] Serguei Barannikov, Ilya Trofimov, Nikita Balabin, and Evgeny Burnaev. Representation topology divergence: A method for comparing neural network representations, 2022. URL `https://arxiv.org/abs/2201.00058`.

[6] Saugata Basu. On bounding the betti numbers and computing the euler characteristic of semi-algebraic sets. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 408–417, 1996.

[7] Ulrich Bauer. Ripser: efficient computation of Vietoris-Rips persistence barcodes. *J. Appl. Comput. Topol.*, 5(3):391–423, 2021. ISSN 2367-1726. doi: 10.1007/s41468-021-00071-5. URL `https://doi.org/10.1007/s41468-021-00071-5`.

[8] Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. A full-text learning to rank dataset for medical information retrieval. 2016. URL `http://www.cl.uni-heidelberg.de/~riezler/publications/papers/ECIR2016.pdf`.

[9] Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International conference on learning representations*, 2021.

[10] Laura Caspari, Kanishka Ghosh Dastidar, Saber Zerhoudi, Jelena Mitrovic, and Michael Granitzer. Beyond benchmarks: Evaluating embedding model similarity for retrieval augmented generation systems. *arXiv preprint arXiv:2407.08275*, 2024.

[11] Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers, 2024. *URL https://arxiv. org/abs/2405.15471*.

[12] Harish Chintakunta, Thanos Gentimis, Rocio Gonzalez-Diaz, Maria-Jose Jimenez, and Hamid Krim. An entropy-based persistence barcode. *Pattern Recognition*, 48(2):391–401, 2015.

[13] Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In *Forty-second International Conference on Machine Learning*.

[14] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. Specter: Document-level representation learning using citation-informed transformers, 2020. URL `https://arxiv.org/abs/2004.07180`.

[15] Liliane Soares Da Costa, Italo L Oliveira, and Renato Fileto. Text classification using embeddings: a survey. *Knowledge and Information Systems*, 65(7):2761–2803, 2023.

[16] Maxime Darrin, Philippe Formont, Ismail Ayed, Jackie CK Cheung, and Pablo Piantanida. When is an embedding model more promising than another? *Advances in Neural Information Processing Systems*, 37:68330–68379, 2024.

[17] MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of cka as a similarity measure in deep learning. *arXiv preprint arXiv:2210.16156*, 2022.

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

[19] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims, 2021.

[20] Ondřej Draganov and Steven Skiena. The shape of word embeddings: Quantifying non-isometry with topological data analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 12080–12099. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.705. URL `http://dx.doi.org/10.18653/v1/2024.findings-emnlp.705`.

[21] Herbert Edelsbrunner, John Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453(26):257–282, 2008.

[22] Owen P Elliott and Jesse Clark. The impacts of data, ordering, and intrinsic dimensionality on recall in hierarchical navigable small worlds. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 25–33, 2024.

[23] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

[24] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

[25] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

[26] Keinosuke Fukunaga and David R Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on computers*, 100(2):176–183, 1971.

[27] Yuri Gardinazzi, Karthik Viswanathan, Giada Panerai, Alessio ansuini, Alberto Cazzaniga, and Matteo Biagetti. Persistent topological features in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=qAHnSkHvsm`.

[28] Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

[29] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters, 2024. URL `https://arxiv.org/abs/2310.07820`.

[30] Akshat Gupta, Atahan Ozdemir, and Gopala Anumanchipalli. Geometric interpretation of layer normalization and a comparative analysis with rmsnorm. *arXiv preprint arXiv:2409.12951*, 2024.

[31] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proc. of SIGIR*, 2021.

[32] Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8, 2015.

[33] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

[34] Mohammad Tariqul Islam, Du Liu, and Deblina Sarkar. Manifold approximation leads to robust kernel alignment, 2025. URL https://arxiv.org/abs/2510.22953.

[35] Saahith Janapati and Yangfeng Ji. A comparative study of learning paradigms in large language models via intrinsic dimension. *arXiv preprint arXiv:2412.06245*, 2024.

[36] Takuya Kataiwa, Cho Hakaze, and Tetsushi Ohki. Measuring intrinsic dimension of token embeddings. *arXiv preprint arXiv:2503.02142*, 2025.

[37] Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy yong Sohn, and Chanyeol Choi. Linq-embed-mistral: Elevating text retrieval with improved gpt data through task-specific control and quality refinement. https://getlinq.com/blog/linq-embed-mistral/, 2024. Linq AI Research Blog.

[38] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR, 2019.

[39] Oliver Kramer. Scikit-learn. In *Machine learning for evolution strategies*, pages 45–53. Springer, 2016.

[40] Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. *arXiv preprint arXiv:2503.21073*, 2025.

[41] Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, and Emily Cheng. Geometric signatures of compositionality across a language model's lifetime. *arXiv preprint arXiv:2410.01444*, 2024.

[42] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. Gemini embedding: Generalizable embeddings from gemini, 2025. URL https://arxiv.org/abs/2503.07891.

[43] Sean Lee, Julius Lipp, Rui Huang, and Darius Koenig. Every byte matters: Introducing mxbai-embed-xsmall-v1, 2024. URL https://www.mixedbread.ai/blog/mxbai-embed-xsmall-v1.

[44] Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread - new fluffy embeddings model, 2024. URL https://www.mixedbread.ai/blog/mxbai-embed-large-v1.

[45] Tom Leinster and Simon Willerton. On the asymptotic magnitude of subsets of euclidean space. *Geometriae Dedicata*, 164(1):287–310, 2013.

[46] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[47] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

[48] Katharina Limbeck and Emily Simons. Magnipy: A python library for computing magnitude of metric spaces. https://github.com/aidos-lab/magnipy, 2025. GitHub repository.

[49] Katharina Limbeck, Rayna Andreeva, Rik Sarkar, and Bastian Rieck. Metric space magnitude for evaluating the diversity of latent representations, 2025. URL https://arxiv.org/abs/2311.16054.

[50] Xuyuan Liu, Lei Hsiung, Yaoqing Yang, and Yujun Yan. Spectral insights into data-oblivious critical layers in large language models. *arXiv preprint arXiv:2506.00382*, 2025.

[51] M Maia, S Handschuh, A Freitas, B Davis, R McDermott, M Zarrouk, et al. Financial opinion mining and question answering. *URL: https://sites. google. com/view/fiqa/home*, 2018.

[52] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[53] Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfr-embedding-mistral: Enhance text retrieval with transfer learning. https://www.salesforce.com/blog/sfr-embedding/, 2024. Salesforce AI Research Blog.

[54] Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. Arctic-embed: Scalable, efficient, and accurate text embedding models. *arXiv preprint arXiv:2405.05374*, 2024.

[55] Timothee Mickus, Stig-Arne Grönroos, and Joseph Attieh. Isotropy, clusters, and classifiers. *arXiv preprint arXiv:2402.03191*, 2024.

[56] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark, 2023. URL https://arxiv.org/abs/2210.07316.

[57] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human-generated machine reading comprehension dataset. 2016.

[58] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. Large dual encoders are generalizable retrievers, 2021. URL https://arxiv.org/abs/2112.07899.

[59] Enrike Nur and Anar Aliyev. ember-v1: Sota embedding model, 2023.

[60] OpenAI. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/, 2023. Accessed: 2025-11-10.

[61] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[62] Alina Petukhova, Joao P Matos-Carvalho, and Nuno Fachada. Text clustering with large language model embeddings. *International Journal of Cognitive Computing in Engineering*, 6: 100–108, 2025.

[63] Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. *arXiv preprint arXiv:2311.05928*, 2023.

[64] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. *Advances in neural information processing systems*, 32, 2019.

[65] Nils Reimers and Iryna Gurevych. The curse of dense low-dimensional information retrieval for large index sizes. 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 605–611, 2021.

[66] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[67] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE, 2007.

[68] William Rudman and Carsten Eickhoff. Stable anisotropic regularization, 2024. URL https://arxiv.org/abs/2305.19358.

[69] William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. Isoscore: Measuring the uniformity of embedding space utilization. *arXiv preprint arXiv:2108.07344*, 2021.

[70] Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, Michael Heck, Renato Vukovic, Shutong Feng, Hsien-chin Lin, Nurul Lubis, Bastian Rieck, Marcus Zibrowius, et al. Less is more: Local intrinsic dimensions of contextual language models. *arXiv preprint arXiv:2506.01034*, 2025.

[71] Spurthi Setty, Harsh Thakkar, Alyssa Lee, Eden Chung, and Natan Vidra. Improving retrieval for rag based question answering models on financial documents. *arXiv preprint arXiv:2404.07221*, 2024.

[72] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020.

[73] Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. Parametric retrieval augmented generation, 2025. URL https://arxiv.org/abs/2501.15915.

[74] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

[75] Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Kai Hua, Wenpeng Hu, Zhengwei Tao, and Shuai Ma. Llms are also effective embedding models: An in-depth overview, 2025. URL https://arxiv.org/abs/2412.12591.

[76] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=wCu6T5xFjeJ.

[77] François Torregrossa, Vincent Claveau, Nihel Kooli, Guillaume Gravier, and Robin Allesiardo. On the correlation of word embedding evaluation metrics. In *LREC 2020-12th Conference on Language Resources and Evaluation*, pages 4789–4797, 2020.

[78] Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. The geometry of tokens in internal representations of large language models. *arXiv preprint arXiv:2501.10573*, 2025.

[79] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. Trec-covid: Constructing a pandemic information retrieval test collection, 2020. URL https://arxiv.org/abs/2005.04474.

[80] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, 2018.

[81] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims, 2020. URL https://arxiv.org/abs/2004.14974.

[82] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8:e19, 2019.

[83] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[84] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.

[85] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[86] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL `https://arxiv.org/abs/2002.10957`.

[87] Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. *arXiv preprint arXiv:2508.21038*, 2025.

[88] Colin Wilkie and Leif Azzopardi. Relating retrievability, performance and length. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 937–940, 2013.

[89] Colin Wilkie and Leif Azzopardi. Best and fairest: An empirical analysis of retrieval system bias. In *European Conference on Information Retrieval*, pages 13–25. Springer, 2014.

[90] Colin Wilkie and Leif Azzopardi. Efficiently estimating retrievability bias. In *European Conference on Information Retrieval*, pages 720–726. Springer, 2014.

[91] Simon Willerton. Spread: a measure of the size of metric spaces. *International Journal of Computational Geometry & Applications*, 25(03):207–225, 2015.

[92] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL `https://arxiv.org/abs/1910.03771`.

[93] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

[94] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.

[95] Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and stella: distillation of sota embedding models, 2025. URL `https://arxiv.org/abs/2412.19048`.

[96] 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. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[97] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60, 2024.

# A Computation of individual topological descriptors

## A.1 Persistent Homology Statistics

Let $\{(b_i, d_i)\}_{i=1}^N$ be the birth-death pairs of a Vietoris Rips filtration. We summarize the persistence diagram using:

$$\text{mean lifetime} = \frac{1}{N}\sum_i (d_i - b_i), \qquad \text{mean midlife} = \frac{1}{N}\sum_i \frac{b_i + d_i}{2}$$

We use the Python library ripser [7] for all Persistent Homology computations.

## A.2 Persistence Entropy

Let $\ell_i = d_i - b_i$ be the lifetimes and $p_i = \ell_i / \sum_j \ell_j$ the normalized lifetimes. We quantify the complexity of the persistence diagram using the Shannon entropy of the normalized lifetimes:

$$H = -\sum_i p_i \log p_i.$$

## A.3 Persistent Homology Dimension

Let $X$ be a bounded subset of a metric space and define the $\alpha$-weighted sum of the length of persistence intervals, given by their birth $b$ and death $d$ times as:

$$E_i^\alpha = \sum_{(b,d)\in PH_i(x)} (d-b)^\alpha$$

Then $dim_i^{PH}$ for the Vietoris-Rips complex of $X$ corresponds to the smallest $\alpha$ for which $E_i^\alpha$ is uniformly bounded for all subsets of $X$:

$$dim_i^{PH}(X) = \inf\{\alpha : \exists C \quad s.t. \quad E_\alpha^i(x) < C \quad \forall x \subset X\}$$

Intuitively, $dim_i^{PH}$ is determined by how the lengths of the PH intervals for increasing sample sizes vary for a specific homological dimension $i$. Higher values suggest a more complex topology across scales, as the total persistence (sum of lifetimes) grows rapidly with sample size.

## A.4 Euler Characteristic

Let $\beta_i$ be the $i^{th}$ Betti number, counting the number of independent $i$-dimensional holes (connected components, loops, voids ...) in $X$. Then the Euler Characteristic is given by alternating sums of $\beta_i$:

$$\chi = \sum_{i=0}^{n} (-1)^i \beta_i,$$

## A.5 TwoNN Dimension

Let $d_{1,i}$ and $d_{2,i}$ be the distances from point $x_i$ to its first and second nearest neighbors and define the ratio $\mu_i = \frac{d_{2,i}}{d_{1,i}}$. The intrinsic dimension $d$ is computed based on the following equation:

$$-\log(1 - F(\mu)) = d \cdot \log(\mu),$$

where $F(\mu)$ is the empirical cumulative distribution function of $\mu$. Fitting the points $(\log \mu, -\log(1 - F(\mu)))$ with a line through the origin yields the estimate of $d$. We use the Python library scikit-dimension [4] for computations.

### A.6 PCA Dimension

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ be the eigenvalues of the covariance matrix of the dataset. The Fukunaga–Olsen (FO) rule estimates the intrinsic dimension as the number of eigenvalues larger than a fraction $\alpha_{\text{FO}}$ of the largest eigenvalue:

$$\hat{m}_{\text{FO}} = \#\left\{\lambda_j \mid \lambda_j \geq \alpha_{\text{FO}} \cdot \lambda_{\max}\right\},$$

where $\lambda_{\max} = \max_j \lambda_j$. We use the Python library scikit-dimension [4] for computations with $\alpha_{\text{FO}} = 0.5$.

### A.7 Effective Rank

Given normalized eigenvalues $p_j = \lambda_j / \sum_k \lambda_k$ of the covariance matrix of $X$, the effective rank is defined as the Shannon entropy, reflecting the effective number of principal components:

$$\text{ER} = \exp\Big(-\sum_j p_j \log p_j\Big).$$

### A.8 Magnitude and Magnitude function

Given a finite metric space $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^D$ and the corresponding similarity matrix $\zeta_X(i,j) := \exp\big(-d(x_i, x_j)\big)$, then for invertible $\zeta_X(i,j)$ the magnitude of X is defined as:

$$\text{Mag}(X) := \sum_{i,j} \big(\zeta_X^{-1}\big)_{ij}$$

Using a scaling factor $t \in \mathbb{R}^+$, magnitude can be computed for different scales $d_t(x,y) = t \cdot d(x,y)$ yielding the magnitude function:

$$\text{Mag}_X : t \;\mapsto\; \text{Mag}(tX)$$

### A.9 Magnitude Dimension

As the scaling factor $t$ grows, the following limit characterizes the asymptotic growth rate of the magnitude function:

$$\dim_{\text{Mag}}(X) := \lim_{t\to\infty} \frac{\log \text{Mag}(tX)}{\log t},$$

For the computation we use the Python library magnipy [48].

### A.10 Magnitude Area

The area under the magnitude function is defined as:

$$\text{MagArea} = \int_{t_0}^{t_{cut}} \text{Mag}_\text{X}(t)\, dt.$$

For the computation we use the Python library magnipy [48].

### A.11 Spread

Given a finite metric space $X$ with metric $d$, the spread $E_0(X)$ is measuring the size of the metric space by:

$$E_0(X) := \sum_{x\in X} \frac{1}{\sum_{x'\in X} e^{-d(x,x')}}.$$

### A.12 Vendi Score

Let $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^D$ be a collection of samples and $K \in \mathbb{R}^{n \times n}$ the corresponding similarity (or more generally kernel) matrix. The Vendi Score is defined as the Shannon entropy of the eigenvalues $\lambda_1, \ldots \lambda_k$ of $K/n$:

$$\mathrm{VS}_k(x_1, \ldots x_n) = \exp\Big( - \sum_i \lambda_i \log \lambda_i \Big).$$

### A.13 Mean Pairwise Similarity

Density-like descriptor computed from pairwise similarities used as average similarity estimate. We use Cosine similarity and Euclidean similarity as kernels $k$ and compute

$$\bar{s} = \frac{1}{n} \sum_i \frac{1}{n} \sum_j k(x_i, x_j).$$

### A.14 Uniformity

Let $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^D$ be a collection of samples and $G_t : X^d \times X^d \to \mathbb{R}_+$ the Radial Basis Function kernel, defined as:

$$G_t(u, v) \triangleq e^{-t\|u-v\|_2^2} = e^{2t \cdot u^\mathsf{T} v - 2t}, \quad t > 0,$$

The uniformity metric [85] is defined as the logarithm of the average pairwise kernel:

$$\mathcal{L}_{\text{uniform}}(f; t) \triangleq \log \underset{x,y \sim p_{\text{data}}}{\mathbb{E}_{\text{i.i.d.}}} [G_t(u, v)] = \log \underset{x,y \sim p_{\text{data}}}{\mathbb{E}_{\text{i.i.d.}}} \left[ e^{-t\|f(x)-f(y)\|_2^2} \right], \quad t > 0.$$

### A.15 IsoScore

Given a point cloud $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^D$ and its transformation onto the first $k$ principal components $X^{PCA}$. Define the eigenvalues $\hat{\Sigma_D}$ as the normalized diagonal of the covariance matrix of $X^{PCA}$. Then IsoScore can be summarized using the following equation:

$$\text{IsoScore}(X) = \frac{n \left( \left( n - ||\hat{\Sigma}_D - \mathbf{1}|| / \sqrt{2(n - \sqrt{n})} \right)^2 (n - \sqrt{n})^2 / n^2 \right) - 1}{n - 1}$$

### A.16 Silhouette Score

For a set of points $X = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^D$ we can measure how well points match their assigned clusters. For point $i$, let $a_i$ be the mean intra-cluster distance and $b_i$ the mean distance to the nearest other cluster. The Silhouette Score is defined as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \qquad \text{Silhouette} = \frac{1}{n} \sum_{i=1}^{n} s_i.$$

We apply the *k-means* algorithm to partition the embedding spaces into clusters. To identify the optimal number of clusters, we evaluate

$$k \in \{3, 5, 10, 20, 50, 100\}$$

and compute the corresponding Silhouette Scores. The value of $k$ that maximizes the Silhouette Score is selected and subsequently used as the metric for characterizing the embedding space.

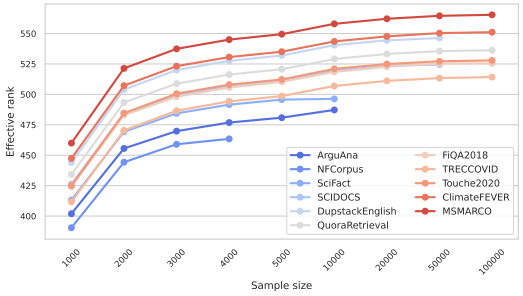


Figure 7: Convergence of the effective rank.

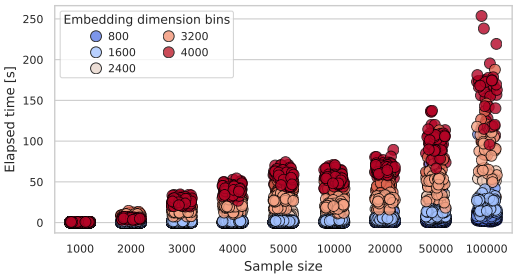


Figure 8: Time complexity of effective rank.

# B Scaling analysis

We perform an empirical scaling analysis across all descriptors to identify an appropriate sample size. This analysis accounts for heterogeneous convergence properties and the varying computational costs associated with the descriptors.

## B.1 Convergence

First, we investigate the stability of the measures with increasing sample size, to ensure that the signatures are robust. We found that the metrics react very differently to adjusting the number of samples, for example magnitude area is increasing with sample size, while pairwise distances are converging quickly. An example of this analysis for all datasets is provided in Figure 7, which shows that the measure effective rank starts to converge after 10000 samples. This implies, as some datasets have a lower number of documents, some measures might not fully converge. Additionally, Figure 7 highlights that the absolute values can vary across datasets, while the general convergence pattern is consistent.

## B.2 Computational complexity

In addition, we also analyzed the computation time in seconds for each sample size, as shown in Figure 8 for effective rank. We observe that the embedding dimension has a strong influence on the computation time for most descriptors, as higher dimensions lead to more expensive computations. Based on these results, we choose the sample sizes as a balance between computational efficiency and descriptor robustness.

# C Disentangling dataset and model contributions

We observe a variation in descriptor behavior: some characterize the dataset, while others serve as markers of the embedding model itself. For instance, Figure 9 shows a clear anisotropy pattern across models, while the values of different datasets show only little variance for a specific model. This indicates, that this descriptor is suitable for distinguishing between models, but not between datasets. In contrast, Figure 10 shows that intrinsic dimension measured through TwoNN appears to be a property of datasets rather than models, showing significant variance across datasets. A part of this effect can be explained by larger dataset sizes, resulting in a higher intrinsic complexity. These findings have important implications for our normalization scheme. As we normalize the UTS vectors across datasets, we thereby also remove the variability in some descriptors, causing them to be less discriminative. As we are mostly interested in the effects of embedding models, this choice is consistent with our objective of highlighting model-related properties rather than dataset-driven effects.

# D Correlation analysis

We perform a correlation analysis and aggregate the pairwise correlations within each dataset among all used descriptors and additional properties such as the number of model layers and the embedding

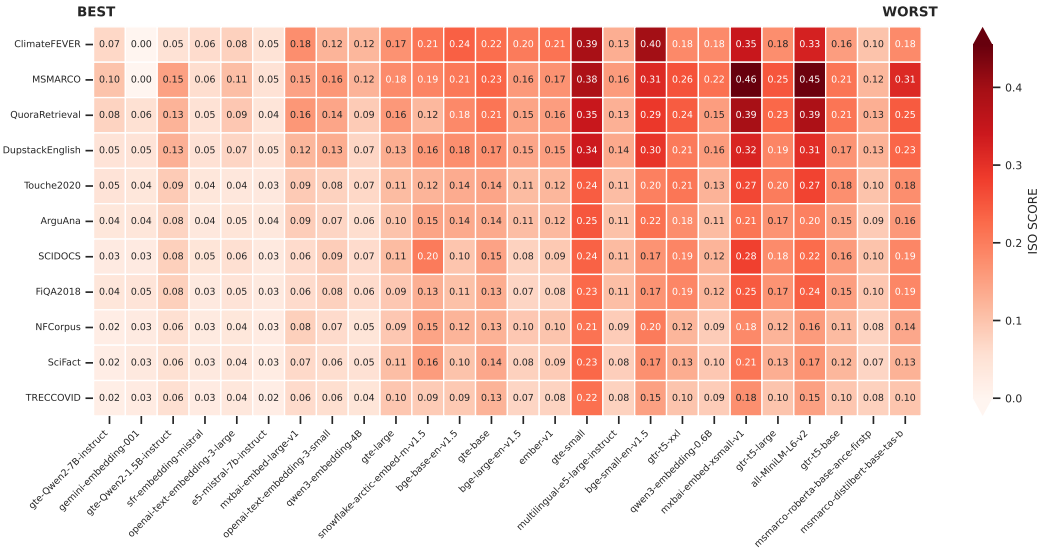


Figure 9: IsoScore of embedding spaces by different models and datasets sorted by MTEB score.

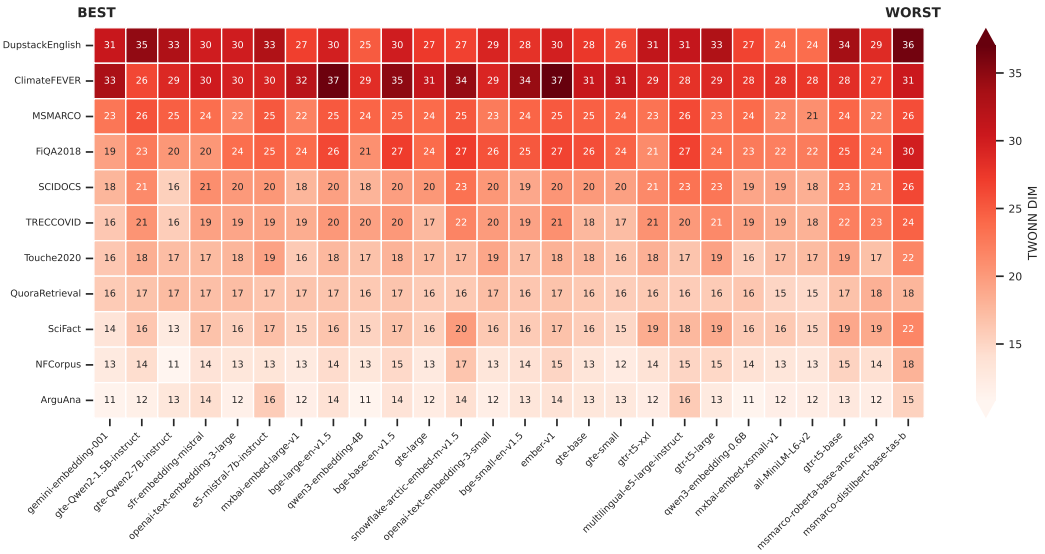


Figure 10: TwoNN dimension of embedding spaces by different models and datasets sorted by MTEB score.

dimension. Figure 11 shows the mean correlation and variance between datasets, organized using hierarchical linkage clustering. We observe that IsoScore is highly (negatively) correlated with embedding dimension and effective rank, indicating that larger embedding dimensions lead to more anisotropic spaces. Given that persistent homology descriptors emerge from pairwise distances, we find that they are highly correlated with the pairwise similarity descriptor. Finally, some of the intrinsic dimension estimators are correlated with each other, indicating that they capture similar aspects of the data. These correlations allow to select a subset of descriptors that are less expensive to compute, while still capturing the overall topology.

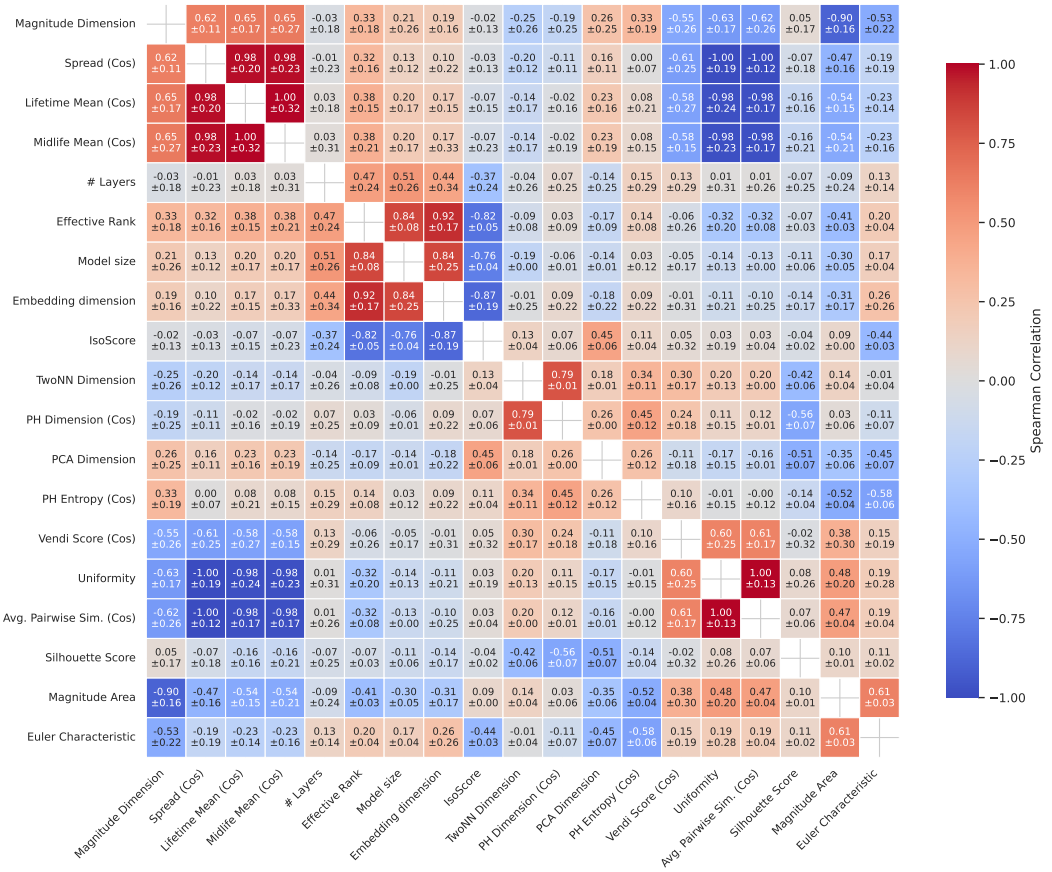


Figure 11: Average correlations of topological descriptors and model properties.

# E Representational similarity analysis

## E.1 Centered Kernel Alignment

We compute the Centered Kernel Alignment between all embedding models and datasets and aggregate the average alignment as shown in Figure 12. We find a similar pattern as with our UTS-based similarity measure, emphasizing that the model architecture and training data play a significant role in the alignment of embedding model representations.

## E.2 Model size

By computing UTS similarity, we find that models of the same size are more similar to each other than to models of different sizes. Furthermore, the topologies of large and small models are most different, which can be explained by the dissimilar transformer architectures.

## E.3 Interpretability of similarity

Our UTS approach enables the interpretability of representational similarity by inspecting the topological drivers that make representations converge. On one hand, we can determine the most discriminative features across all models, as summarized by Figure 14. We find that representations mostly converge due to their density (measured by uniformity and pairwise similarities) as well as their dimensionality (measured by IsoScore and effective rank). Furthermore, similarity is partially driven by the clustering of the embedding spaces. Alongside global interpretations, our framework highlights similarities across embedding spaces by analyzing pairwise comparisons, for example,

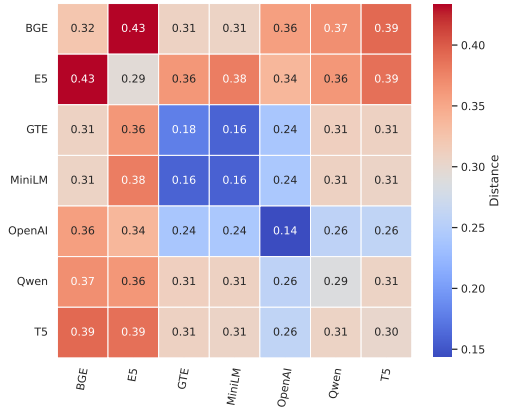


Figure 12: Centered Kernel Alignment between model families.

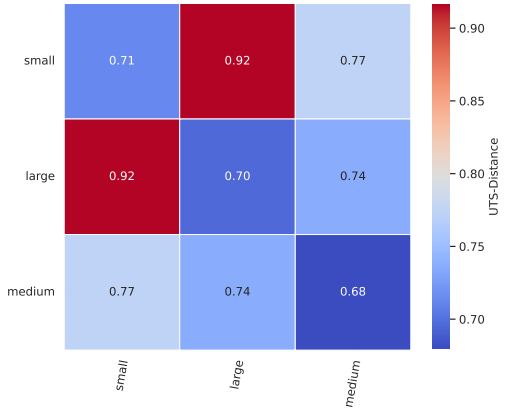


Figure 13: UTS-based similarity of different model sizes.

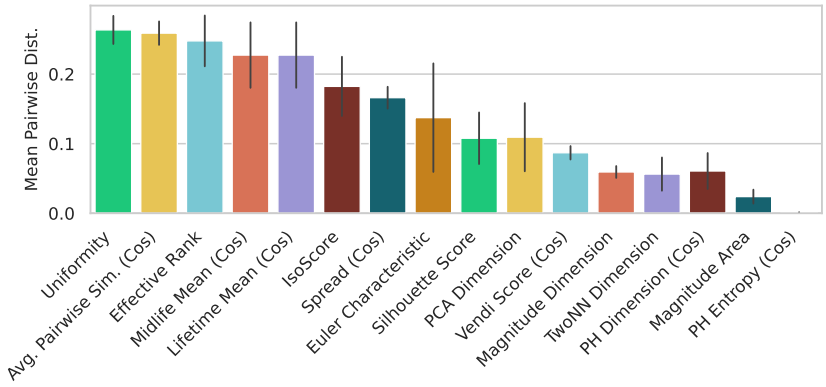


Figure 14: Average distance of UTS components across all models.

showing that one space is denser while another is more isotropic. In the context of the Platonic Representation Hypothesis, we specifically find that each model family has different drivers for representational alignment. For instance, Qwen models have a significantly higher effective rank than all other models, while the Clustering coefficient is most discriminative for Gemini models.

## F Influence of dimensionality on retrieval performance

As shown in Figure 15, the effective rank exhibits consistently strong correlations with retrieval performance metrics, such as Recall@20. To visualize this relationship, we encode the embedding model size by marker size and observe that larger models generally correspond to higher effective ranks. This observation aligns with the well known trend that model scale correlates with performance. However, in our setting, a plausible confounding factor is the embedding dimension, which has been shown to play a critical role in retrieval systems [87]. Indeed, our correlation analysis demonstrates that effective rank is itself highly correlated with the embedding dimension, suggesting that the observed relationship between model size and effective rank is largely attributed to this relationship.

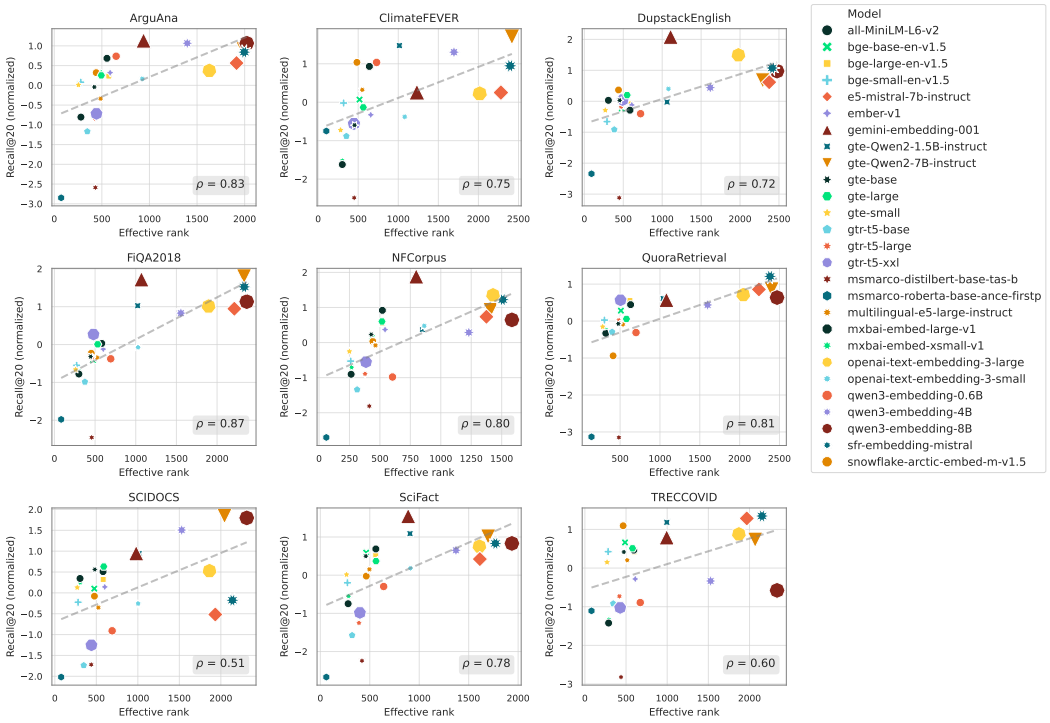


Figure 15: Relationship between normalized Recall@20 and effective rank for different datasets.

Table 6: **Balanced accuracy** for predicting model-specific properties from topological signatures across all folds, controlling for embedding dimensionality. #Cls: number of classes; BL: majority class baseline.

| | Model | | | Model family | | | Architecture | | | Activation | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dim. | Acc. | #Cls | BL | Acc. | #Cls | BL | Acc. | #Cls | BL | Acc. | #Cls | BL |
| All | $0.66 \pm 0.02$ | 27 | 0.04 | $0.68 \pm 0.07$ | 12 | 0.17 | $0.98 \pm 0.01$ | 2 | 0.72 | $0.84 \pm 0.12$ | 3 | 0.59 |
| 4096 | $0.89 \pm 0.10$ | 3 | 0.37 | $1.00 \pm 0.00$ | 2 | 0.73 | $1.00 \pm 0.00$ | 1 | 1.00 | $1.00 \pm 0.00$ | 1 | 1.00 |
| 1024 | $0.68 \pm 0.23$ | 5 | 0.20 | $0.68 \pm 0.23$ | 5 | 0.20 | $0.78 \pm 0.26$ | 2 | 0.80 | $0.78 \pm 0.26$ | 2 | 0.80 |
| 768 | $0.73 \pm 0.12$ | 8 | 0.13 | $0.83 \pm 0.15$ | 6 | 0.38 | $1.00 \pm 0.00$ | 1 | 1.00 | $0.86 \pm 0.06$ | 2 | 0.63 |
| 384 | $0.76 \pm 0.01$ | 4 | 0.25 | $0.93 \pm 0.13$ | 2 | 0.50 | $1.00 \pm 0.00$ | 1 | 1.00 | $1.00 \pm 0.00$ | 1 | 1.00 |

# G Predictive analysis

## G.1 Architectural attributes

We train classifiers to predict the model architecture (encoder vs. decoder) and the activation function (GELU, ReLU, SiLU) used in unseen embedding spaces. We find that both properties can be predicted with a higher balanced accuracy than the baseline, as summarized in Table 6. For architecture, the binary classification appears to be an easy task, driven by the large difference in embedding dimension between encoder and decoder models. Therefore, we again control for embedding dimension but find that there are too few models of different architectures with the same embedding dimension. For the activation functions, we find a similar pattern suggesting that the same architecture mostly uses the same activation function. Despite this, we are able to predict the architecture used in an embedding spaces with a balanced accuracy of $0.98 \pm 0.01$, compared to a baseline of $0.72$, and the activation function with a balanced accuracy of $0.84 \pm 0.12$, compared to a baseline of $0.59$. To simulate predictions unseen models, we perform a group cross-validation, yielding comparable results ($0.94 \pm 0.11$ for architecture and $0.83 \pm 0.12$ for activation function). These results suggest

that UTS capture model-specific details encoded in the global topology of the embedding space, motivating further research into understanding how specific architectural choices affect the geometry of embedding spaces, similar to recent work on LayerNorm [30].

### G.2 PCA signatures

To reduce redundancy and uncover the most informative dimensions of variation, we apply Principal Component Analysis (PCA) [61] to the collection of normalized signature vectors across all embedding spaces. This yields a reduced signature $\hat{\boldsymbol{s}}^X = [\hat{s}_1^X, \hat{s}_2^X, \ldots, \hat{s}_L^X]^\top \quad (L < K)$, where each component $\hat{s}_i^X$ is a linear combination of the original normalized descriptors, capturing the principal axes of variation in the topological feature space. We analyze the effectiveness of the PCA-transformed signature vectors using only the first 5 principal components as features. We find that the average accuracies are slightly lower with 0.44 for predicting the model, 0.50 for predicting model family. This suggests that highly correlated attributes might still complement each other in prediction tasks. In the case of retrieval performance prediction, we also observe lower correlations compared to the full UTS vectors, approaching the results of effective rank. This can be explained by the large influence of properties related to the embedding dimension, which are captured by the first principal component.

### G.3 Multilingual datasets

We conduct an additional experiment on the GerDaLIRSmall dataset (queries=12k, docs=10k), which contains domain-specific law documents in German language. We compute retrieval metrics and topological signatures for 3 multilingual models of the Qwen family (Qwen2-1.5B-instruct, Qwen3-0.6B, Qwen3-4B) across 3 seeds. We train a classifier to predict the multilingual model that generated an embedding space using the signatures of other datasets as training data. We then predict on GerDaLIRSmall as test set and find that we can correctly classify 6 out of 9 embedding spaces (i.e. balanced accuracy = 66.66 vs. 33.33 as baseline). Note that this prediction task is especially difficult, because the models origin from the same family, sharing architectural similarities, reflected in similar signatures. This additional experiment shows that our approach can be directly applied on non-English datasets, but might benefit from having more multilingual training data.

## H Signature clustering analysis

We apply UMAP to the signature vectors of all models, seeds and datasets shown in Figure 16. We encode the model family by different markers, the retrieval performance (normalized Recall@100) by colors and the model size by the size of the points. First, we find that models cluster by model family, which is in line with our previous findings. Moreover, proprietary models developed by OpenAI cluster within the same region as decoder-based models (such as Qwen and E5), highlighting the potential to infer architectural characteristics from our holistic signatures. We also observe a general clustering by model size, which correlates with architecture, given that larger models are typically decoder-based. Finally, we find that sub-clusters correspond to retrieval performance, suggesting that embedding spaces with comparable topology yield similar retrieval outcomes.

## I Global retrievability analysis

We study the Gini coefficient of the top-100 retrievability distribution to assess the overall bias in a dense retrieval system. As displayed in Figure 17, the retrieval datasets differ in bias strength, which is related to the collection size. Given a fixed query distribution, we observe that decoder-based models, on average, exhibit a larger bias across most datasets. This implies that they return certain documents more frequently than others. A key difference between the decoder and encoder models used in our experiments is their embedding dimension. We hypothesize, that these effects can be explained by the dimensionality affecting the similarity computation. This is a possible connection to previous studies that have analyzed the influence of dimensionality on retrieval [65]. These findings suggest that not only retrieval performance is associated with embedding dimensionality, but also retrieval bias.

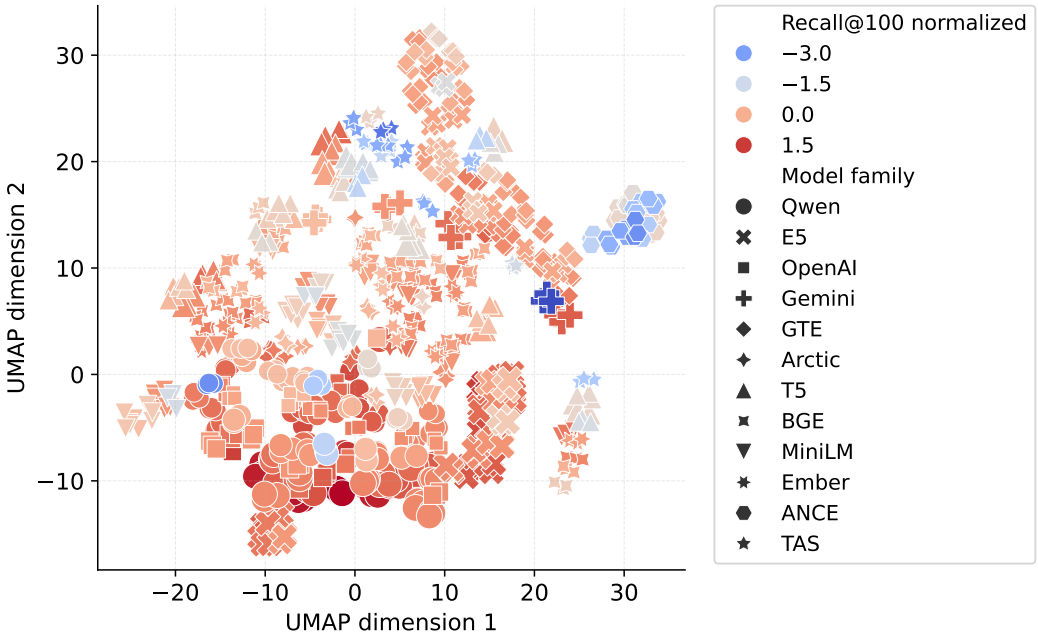


Figure 16: Clustering of Unified Topological Signatures and their normalized Recall@100.

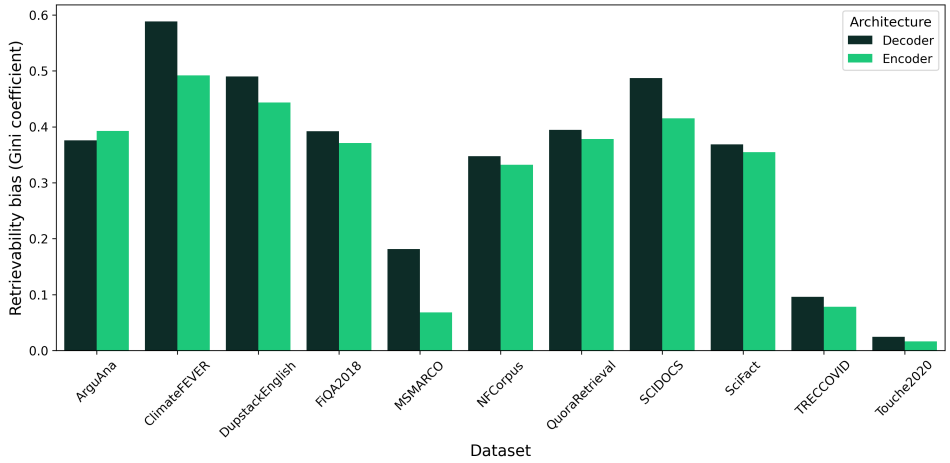


Figure 17: Retrievability bias of encoder and decoder models on different datasets.

# J Local retrievability analysis

## J.1 Predictions with raw embeddings

As our results have shown, the topological signatures reveal significant differences between retrievable and non-retrievable documents. We investigate whether such separations exist in the raw embedding space, as this would eliminate the need of computing signatures. For this, we use UMAP to visualize the raw embedding space and compare it with the UTS space, color-coding the retrievability of documents. As Figures 19–24 show, the UTS space provides a much clearer separation than the raw embedding space, motivating our approach. These findings are supported by the lower accuracy of the baseline prediction model, as reported in the results section. Finally, even if the raw embedding

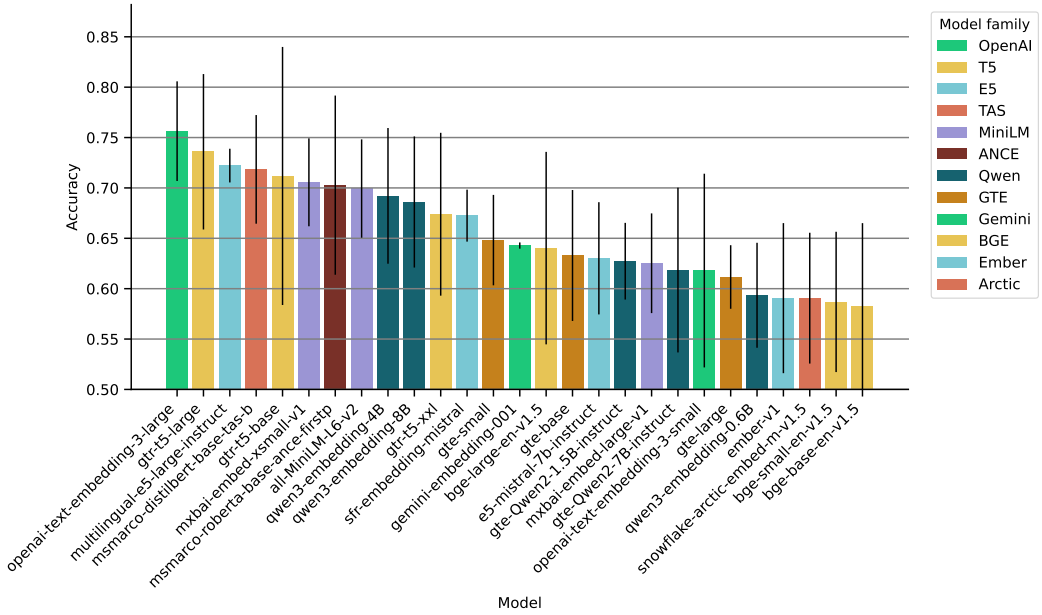


Figure 18: Comparison of accuracies by different models for predicting retrievability.

space shows a partial separation (e.g. Figure 23), such trends do not extrapolate well to embedding spaces of other datasets, unlike our proposed UTS vectors.

### J.2 Characteristics of retrievable documents

To gain deeper insight into the topological features that influence document retrievability, we analyze the differences between retrievable and non-retrievable clusters. For this, we select 10 data points from each of the clusters and compare their topological signatures. As Figure 25 shows, the local topology has distinct values for several descriptors, highlighting the importance of our unified approach. Further, we find that the largest differences lie in properties related to the density of the embedding space, suggesting that less retrievable documents lie in isolated regions. Beyond density, we also find that the neighborhood of highly retrievable documents has lower intrinsic dimensions. These observations are generalized by the feature importance analysis of the prediction model as summarized by Figure 26.

### J.3 Retrievability of embedding models

We investigate whether embeddings differ in their ability to predict document retrievability. For this, we fit prediction models for each embedding model and evaluate the accuracy using group cross-validation. As Figure 18 shows, there are large differences in the accuracy between different models, with OpenAI's text-embedding-3-large achieving the best results. This suggests that the structural characteristics of local topologies vary across models, motivating further exploration of retrievability.

## K Ablation studies

We perform a series of ablation studies to evaluate the robustness of our method.

**Descriptor sensitivity** First, we evaluate the effect of removing descriptors from the signature, to evaluate how the prediction results change. As shown in Table 7, the full signature consistently achieves the best overall performance, while density and isotropy descriptors contribute most strongly to predictive performance. The last row corresponds to the baseline performance when using only effective rank as feature.

**Normalization strategy** Furthermore, in Table 8, we investigate the robustness when using different normalization strategies and find that all methods maintain similar performance. To complement these results, we also provide the minimum, mean, and maximum values in Table 10 for each descriptor used in the UTS signature. The observed ranges indicate that the descriptors do not exhibit extreme outliers, supporting the robustness of the normalization procedure.

**Model choice** Lastly, we evaluate different prediction models in Table 9 to assess the influence of classifier choice on downstream performance. Overall, model choice has a measurable but limited impact on the results. Across all prediction tasks, the Random Forest consistently achieves the highest balanced accuracy and is therefore used throughout our experiments. While the MLP performs competitively on some tasks, it generally underperforms the Random Forest and exhibits no clear advantage. The Nearest Neighbor classifier yields the weakest overall results, suggesting that the relationship between topological descriptors and model properties cannot be captured solely through local similarity.

Table 7: Balanced accuracy (mean $\pm$ standard deviation across cross-validation folds) for predicting model-specific properties after successively removing descriptor groups.

| | **Model** | **Model Family** | **Architecture** | **Activation** |
|---|---|---|---|---|
| **All** | $\mathbf{0.66 \pm 0.02}$ | $\mathbf{0.68 \pm 0.07}$ | $\mathbf{0.98 \pm 0.01}$ | $\mathbf{0.84 \pm 0.10}$ |
| Remove PH descriptors | $0.64 \pm 0.02$ | $0.67 \pm 0.07$ | $0.98 \pm 0.01$ | $0.82 \pm 0.07$ |
| Remove ID descriptors | $0.62 \pm 0.02$ | $0.67 \pm 0.07$ | $0.98 \pm 0.01$ | $0.84 \pm 0.09$ |
| Remove Density descriptors | $0.41 \pm 0.01$ | $0.40 \pm 0.07$ | $0.98 \pm 0.01$ | $0.70 \pm 0.20$ |
| Remove Isotropy descriptors | $0.39 \pm 0.01$ | $0.47 \pm 0.07$ | $0.98 \pm 0.01$ | $0.78 \pm 0.20$ |

Table 8: Balanced accuracy (mean $\pm$ standard deviation across cross-validation folds) obtained using different normalization procedures.

| | **Model** | **Model Family** | **Architecture** | **Activation** |
|---|---|---|---|---|
| **Max normalization** | $\mathbf{0.66 \pm 0.02}$ | $\mathbf{0.68 \pm 0.07}$ | $\mathbf{0.98 \pm 0.01}$ | $\mathbf{0.84 \pm 0.10}$ |
| Mean normalization | $0.66 \pm 0.04$ | $0.68 \pm 0.09$ | $0.98 \pm 0.01$ | $0.83 \pm 0.05$ |
| Robust scaler | $0.65 \pm 0.04$ | $0.68 \pm 0.09$ | $0.98 \pm 0.01$ | $0.84 \pm 0.05$ |
| Standard scaler | $0.65 \pm 0.04$ | $0.68 \pm 0.09$ | $0.98 \pm 0.01$ | $0.84 \pm 0.06$ |

Table 9: Impact of classifier choice on downstream prediction tasks. Balanced accuracy (mean $\pm$ standard deviation across cross-validation folds) obtained using different prediction models.

| | **Model** | **Model Family** | **Architecture** | **Activation** |
|---|---|---|---|---|
| **Random Forest** | $\mathbf{0.66 \pm 0.02}$ | $\mathbf{0.68 \pm 0.05}$ | $\mathbf{0.98 \pm 0.01}$ | $\mathbf{0.84 \pm 0.10}$ |
| MLP | $0.57 \pm 0.04$ | $0.66 \pm 0.05$ | $0.96 \pm 0.01$ | $0.77 \pm 0.09$ |
| Nearest Neighbor | $0.49 \pm 0.07$ | $0.63 \pm 0.07$ | $0.93 \pm 0.04$ | $0.78 \pm 0.10$ |

Table 10: Distribution of descriptor values across all embedding spaces.

| **Descriptor** | **Min** | **Mean** | **Max** |
|---|---|---|---|
| Magnitude Area | 43.8468 | 1314.3146 | 1735.8951 |
| Magnitude Dimension | 2.2974 | 4.9628 | 5.6931 |
| PCA Dimension | 34.0000 | 118.6526 | 310.0000 |
| TwoNN Dimension | 4.7281 | 20.0096 | 37.0267 |
| Silhouette Score | 0.0059 | 0.0376 | 0.2335 |
| Isolation Score | 0.0000 | 0.1255 | 0.4551 |
| Uniformity | -3.9189 | -2.0565 | -0.0423 |
| Euler Characteristic | -20990.0 | -5368.3743 | 2324.0 |
| Effective Rank | 59.9828 | 814.8092 | 2572.4165 |
| Average Pairwise Cosine Similarity | 0.0077 | 0.4758 | 0.9894 |
| Mean Persistence Lifetime (Cosine) | 0.0037 | 0.2636 | 0.6032 |
| Mean Persistence Midlife (Cosine) | 0.0019 | 0.1318 | 0.3016 |
| Vendi Score (Cosine) | 1.0078 | 1.3272 | 1.4449 |
| Persistent Homology Dimension (Cosine) | 0.0000 | 10.7721 | 21.7708 |
| Persistent Homology Entropy (Cosine) | 6.4838 | 13.4932 | 14.2780 |
| Spread (Cosine) | 1.0106 | 1.7332 | 2.6893 |

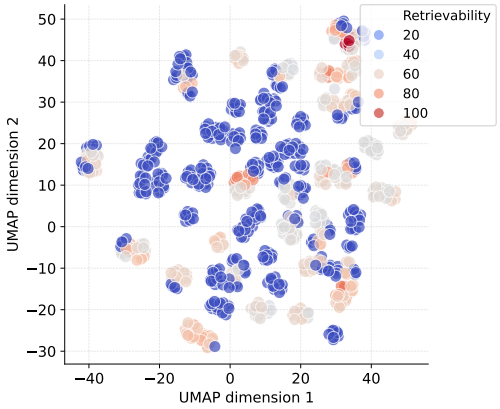


Figure 19: ArguAna raw embeddings for text-embedding-3-large.

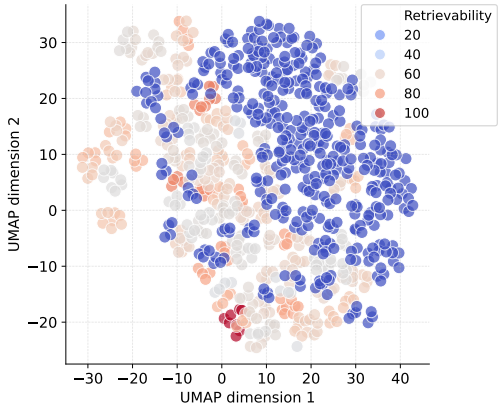


Figure 20: ArguAna local UTS for text-embedding-3-large.

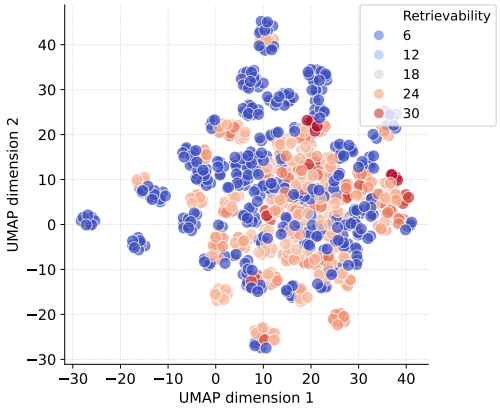


Figure 21: NFCorpus raw embeddings for e5-mistral-7b-instruct.

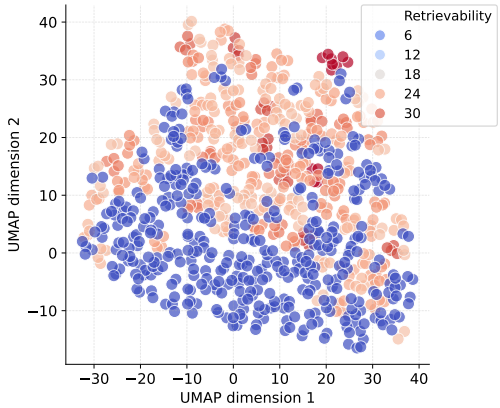


Figure 22: NFCorpus local UTS for e5-mistral-7b-instruct.

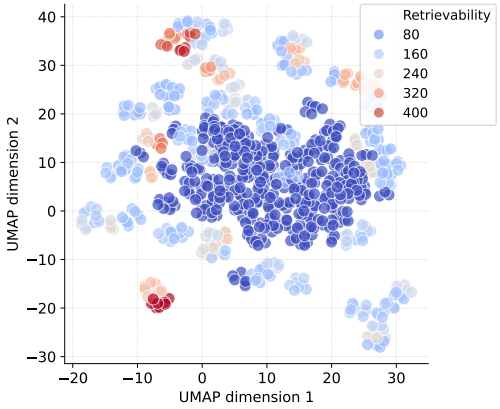


Figure 23: Quora raw embeddings for multilingual-e5-large.

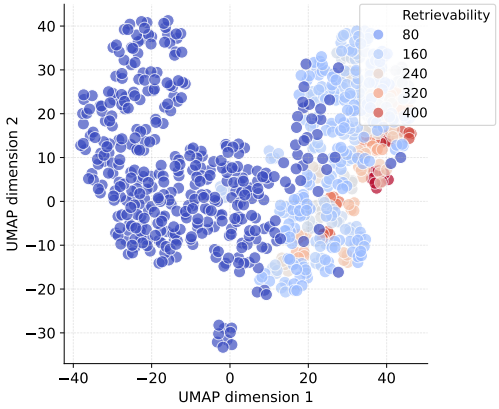


Figure 24: Quora local UTS for multilingual-e5-large.

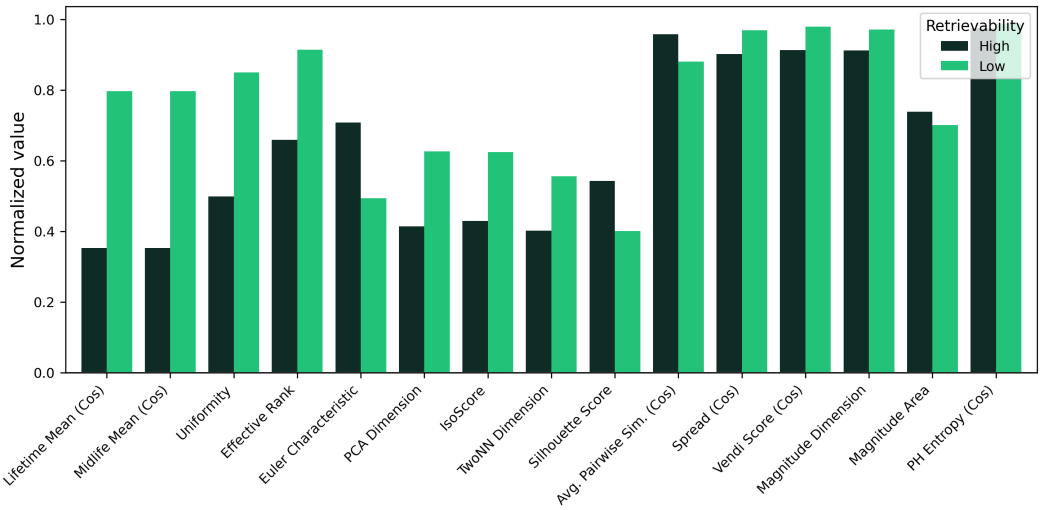


Figure 25: Differences between clusters of highly retrievable and non-retrievable documents for QuoraRetrieval dataset.

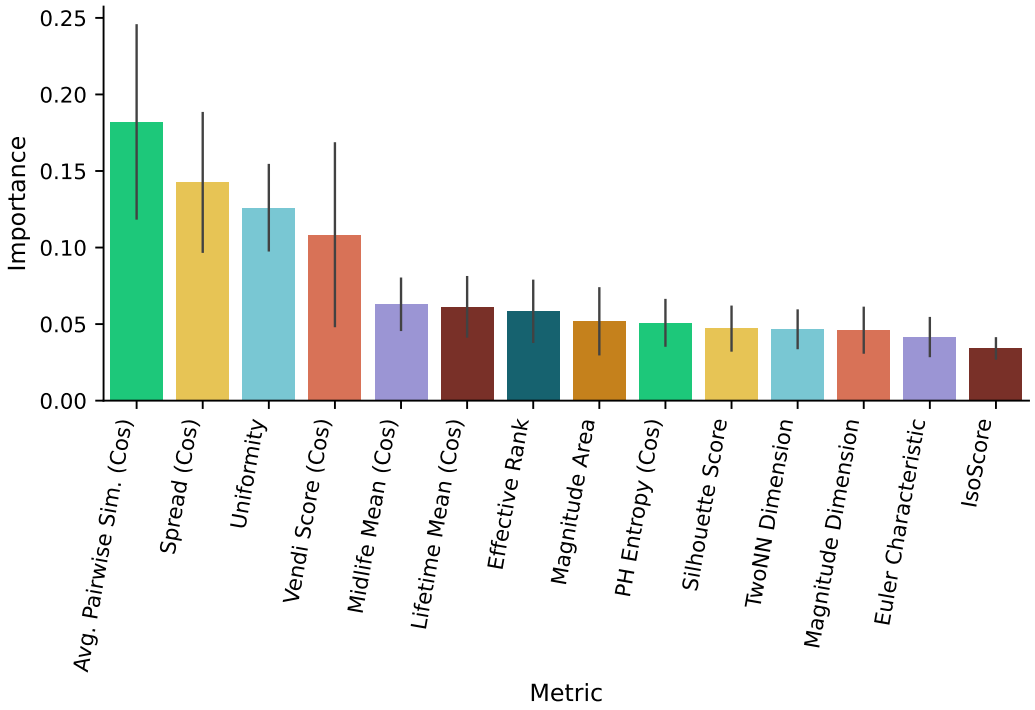


Figure 26: Feature importance for predicting retrievability.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions and claims are explicitly listed in the introduction and briefly summarized in the abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a separate section discussing Limitations and Future Work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: No theoretical results are provided in the work – all experiments are empirical.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We share a GitHub repository with the full code to reproduce all results and data.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Full access to all data and code is provided.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We listed the cross-validation strategy and model training details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All results are reported with standard deviation/error bars across cross-validation folds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We list the number of GPU nodes used for our computations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We read and agree to the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential impact on deriving details about proprietary model architectures.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: There is no data or model with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited all uses of third party libraries, benchmark suites and embedding models used in this work. All licenses are properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide our code and embedding space signature data on GitHub with a detailed documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This research does not involve LLMs as an important component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.