

Measuring Data Diversity with Embeddings: A Taxonomy and Benchmark for Diversity Measures

Cantao Su, Anna Wegmann, Esther Ploeger, Dong Nguyen

Department of Information & Computing Sciences, Utrecht University

Utrecht, The Netherlands

Correspondence: c.su@uu.nl

Abstract

Diversity is a critical property of text corpora, with direct implications for training data curation, test set design, and model generalization. However, its measurement remains fragmented and poorly understood. Existing embedding-based diversity measures vary widely in how they are computed and can yield substantially different diversity scores on the same data, yet no systematic comparison of their empirical behaviour exists. We make two contributions toward addressing this gap. First we present a taxonomy of 21 embedding-based diversity measures, organizing them into four families based on their underlying computations. Second, we introduce **EmbDivBench**, a benchmark that evaluates these measures along three conceptually distinct components—variety, balance and disparity—on both simulated and natural text data. Although Mean Pairwise Distance is the most commonly used measure, our experiments show that it is outperformed by several alternatives on the three diversity components, with Graph Entropy being the best choice overall. Lastly, we distil these findings into practical recommendations for diversity measure selection.

1 Introduction

Data diversity has demonstrated practical value across multiple areas of NLP. First, diversity in training data is a driver of model capability: corpora that are more diverse have shown performance improvement across a range of tasks (Liu and Zeldes, 2023; Yadav et al., 2024; Chen et al., 2024), and can compensate for smaller data sizes (Schiller et al., 2024; Estève et al., 2026). Second, it helps to improve fairness and inclusiveness (Blodgett et al., 2018), counteract the dominance of English- and Western-centric data (Pouran Ben Veyseh et al., 2022), and reduce toxic output and bias (Chu et al., 2024; Agarwal et al., 2026; Gehman et al., 2020). Third, diversity is such an intrinsic property of hu-

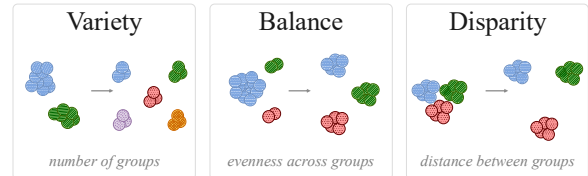


Figure 1: The three components of diversity in EMB-DIVBENCH. In each panel, the left configuration has lower diversity and the right has higher diversity along the corresponding component. **Variety** counts the number of groups present; **Balance** measures how evenly points are distributed across groups; **Disparity** captures how far apart the groups lie in the embedding space.

man language that generating sufficiently diverse outputs is itself a user expectation (Estève et al., 2025; Lee et al., 2022).

Despite its importance, the measurement of diversity has not kept pace. First, “diversity” remains a vague and overloaded term in NLP (Nguyen and Ploeger, 2025; Estève et al., 2025): Datasets are described as “diverse” without specifying what *dimension* (e.g., stylistic vs. semantic) and *component* (e.g., variety vs. balance in Fig. 1) of diversity is meant (Zhao et al., 2024; Shaib et al., 2025; Guo et al., 2024; Nguyen and Ploeger, 2025). Outside NLP, diversity has long been decomposed into three distinct components (Fig. 1, see also §4.1)—variety (how many categories the data spans), balance (how evenly data are distributed across them), and disparity (how different the categories are from one another) (Stirling, 2007). These components have more recently been used in NLP, for example, to theoretically categorize measures into what components they are likely to be sensitive to (Estève et al., 2025). However, different measures have been named and computed inconsistently (§3.1) and used in isolation without an overview organizing the differences in how they are computed, or how they relate. We introduce a unifying taxonomy of existing embedding-based diversity measures

based on how they are computed (§3.2).

Second, despite the increasing use of diversity measures, the measurements themselves are rarely evaluated, with a few exceptions (e.g., [Tevet and Berant 2021](#); [Yang et al. 2025](#); [Zhang et al. 2025a](#)). However, systematic evaluation is crucial: Otherwise, it remains unclear which measure researchers should use and whether newly proposed measures truly improve over existing ones. We introduce EMBDIVBENCH, an empirical benchmark that, contrary to previous approaches (§2), tests sensitivity to variety, balance, and disparity through controlled manipulations of each component.

Among approaches to diversity measurement, embedding-based methods have become increasingly popular ([Shaib et al., 2025](#); [Guo et al., 2025](#); [Zhu et al., 2025](#); [Yang et al., 2025](#); [Fan et al., 2025](#); [Friedman and Dieng, 2023](#)). The success of representation learning in NLP makes it possible to represent different dimensions (e.g., semantic or stylistic) of text as points in a continuous vector space ([Reimers and Gurevych, 2019](#); [Wegmann et al., 2022](#)). Embedding-based diversity measures are thus widely applicable: diversity can be measured along many dimensions, as long as texts can be mapped to an embedding space capturing the relevant diversity dimension. Although some studies have compared diversity measures ([Estève et al., 2025](#); [Yang et al., 2025](#)), none have provided an exhaustive categorization or an empirical analysis of embedding-based diversity measures. To address these gaps, we make the following **contributions**:

- **A taxonomy** (§3) that surveys and classifies 21 embedding-based diversity measures into four organized families.
- **EMBDIVBENCH** (§4), a benchmark that tests measures on the variety, balance and disparity components through two controlled complementary tiers: a simulated tier that affords exact control over each component, and a natural-text tier for the semantic dimension.
- **Practical recommendations** (§7) for diversity measure selection, on each component.
- **Data and code** are publicly available on Github¹ and can be used to compare new diversity measures.

We find that disparity is the hardest component for diversity measures and that natural text is more difficult for measures than the simulated data tier.

Mean Pairwise Distance—a popular embedding-based diversity measure—is outperformed by several alternatives on all three components, with Graph Entropy performing the best overall.

2 Related Work

Non-embedding-based diversity measures. Diversity has been studied across many fields—e.g., in ecology, economics, and sociology. In many of these fields, Shannon Entropy is a popular approach ([Leinster and Cobbold, 2012a](#); [Nehring and Puppe, 2002](#)). Within NLP, the dominant non-embedding measures are usually surface-form statistics over tokens or n -grams, such as type-token ratio and its variants ([Covington and McFall, 2010](#)), Distinct- n ([Li et al., 2016](#)), Self-BLEU ([Zhu et al., 2018](#)), and n -gram entropy ([Han et al., 2022](#)). While useful, they only capture surface variation ([Stasaski and Hearst, 2022](#); [Tevet and Berant, 2021](#)). Such non-embedding-based measures have also been included in toolkits, to apply them over dependency-parsed corpora ([Estève and Dobrovoljc, 2026](#)) or between language pairs ([van der Goot et al., 2025](#)). In contrast, we focus on embedding-based diversity measurement—a promising and increasingly popular way to measure data diversity.

Embedding-based diversity measures: many but scattered. Advances in representation learning have made it possible to encode discrete objects—sentences, images, molecules—as points in a continuous space whose structure reflects similarity, opening up a class of diversity measures that operate directly on these representations rather than on surface form or predefined categories. Within NLP, a growing number of works exploit this: Mean Pairwise Distance has been used most frequently ([Tevet and Berant, 2021](#); [Shaib et al., 2025](#); [Miranda et al., 2025](#); [Ploeger et al., 2024](#)). Other used measures in NLP include Cluster Inertia ([Du and Black, 2019](#); [Yang et al., 2025](#); [Zhu et al., 2025](#)), the geometric mean of radii across all embeddings ([Lai et al., 2020](#); [Wang et al., 2025b](#); [Yang et al., 2025](#)), the Vendi Score ([Friedman and Dieng, 2023](#); [Yang et al., 2025](#); [Zhang et al., 2024](#); [Ba et al., 2025](#)), and DCScore ([Zhu et al., 2025](#); [Ba et al., 2026](#)). Across this literature, these measures are largely introduced and used as general-purpose diversity measures, without systematically distinguishing the different components of diversity they may capture.

Embedding-based measures have also been pro-

¹<https://github.com/nlpsoc/EmbDivBench>

posed and used in other fields, such as Hamiltonian Diversity in chemistry (Hu et al., 2024a; Du et al., 2022), Energy over graph-distance embeddings for graph generation (Velikonitsev et al., 2024), and Rényi Kernel Entropy in computer vision (Jalali et al., 2023). Because such measures could also be applied to text, we also include them in our taxonomy and test them on our benchmark. An overview of all included measures is in App. §A.

Evaluation of diversity measures. In most of the work above, diversity measures serve as a tool rather than as the object of study. Diversity measurement itself has received limited systematic evaluation, and researchers have recently called for more attention to this issue (Yang et al., 2025; Zhao et al., 2024; Nguyen and Ploeger, 2025). Prior work has evaluated measures by assessing their scores on text sets of known diversity level, e.g., by varying the temperature parameter (Tevet and Berant, 2021; Zhu et al., 2025), the number of paraphrases (Zhang et al., 2025a) and by asking crowd workers to produce text sets with higher or lower diversity (Tevet and Berant, 2021). Crucially, none of these evaluations distinguishes between the variety, balance and disparity components of diversity.

Diversity measures have also been correlated against human judgements of diversity (Han et al., 2022; Zhu et al., 2025). However, eliciting human judgements of text diversity is challenging and costly, can be influenced by other factors like text quality, and is only feasible for small sets of texts (Zhang et al., 2025a; Tevet and Berant, 2021). A recent study by Zhang et al. (2025a) compares diversity measures against LLM-annotated judgements of diversity. While more scalable, such evaluations are dependent on the quality of the LLM annotations and factors like prompt design.

Diversity measures have also been evaluated by correlating them with performance on a downstream task (Yang et al., 2025); however, such correlations do not necessarily demonstrate that the measure captures diversity, as performance gains may result from other factors captured by the measure. Studies have also examined diversity measures by correlating them against each other (Zhang et al., 2025a; Shaib et al., 2025).

Most directly related to our work is Estève et al. (2025), who survey diversity measurement in NLP and apply Stirling’s variety, balance and disparity (V-B-D) decomposition to assign existing measures to the component their formula should target, with-

out empirical testing. Instead, our taxonomy classifies embedding-based measures by how the score is computed (§3.2), and EMBDIVBENCH tests their component sensitivity empirically (§4). To the best of our knowledge, this is the first empirical component-based evaluation of diversity measures in NLP.

3 Unified Taxonomy of Diversity Measures

We describe the current fragmented landscape of embedding-based diversity measurement (§3.1) and unify 21 measures into a taxonomy (§3.2).

3.1 A Fragmented Landscape

The literature on embedding-based diversity measurement is strikingly inconsistent in how it defines and uses measures. We collect 21 different embedding-based diversity measures (Tab. 1) and highlight three patterns of fragmentation that motivate the need for a unified taxonomy.

Different names, identical computation. The same computation is sometimes introduced under multiple names by different researchers. For example, the sum of pairwise distances over a dataset is called *Max Dispersion* by Yu et al. (2022a), *Dist-Sum* by Yang et al. (2025) and *Remote-Clique* by Cevallos et al. (2018). Another example is that the geometric mean of per-dimension standard deviations is called *Diversity* by Lai et al. (2020) but *Radius* by Yang et al. (2025).

Same name, different computations. Conversely, the same name is sometimes used for distinct computations. The term *Remote-Clique* denotes the sum of pairwise distances in Cevallos et al. (2018), but the mean of pairwise distances in Cox et al. (2021a).

Similar objectives, different realizations. Beyond name-computation collisions, several measures aim to capture the same concept but operationalize it differently. For example, both Cox et al. (2021a) and Yang et al. (2025) partition the embedding space into discrete bins and apply Shannon entropy to bin counts. However, Cox et al. (2021a) realize this via UMAP projection and 2D grid binning, and Yang et al. (2025) via K -means clustering (“Partition Entropy”).

3.2 Four Families of Diversity Measures

We survey existing measures (App. §A) and organize them into four families (Tab. 1), based on the type of mathematical operation that transforms a set of embeddings into a scalar diversity score. We later use these families to make more general conclusions about the quality of different measurement approaches. The families are:

Distance-based measures aggregate pairwise distances into a single score, without first creating additional structures such as clusters or graphs. An example is the frequently used Mean Pairwise Distance (e.g., in Tevet and Berant 2021; Wang et al. 2025a; Guo et al. 2024), which computes the mean cosine similarity across all embedding pairs.

Geometry-based measures quantify the spatial properties of the embeddings. For example, the Convex Hull Volume measures the volume of the smallest convex region enclosing all embeddings (Yu et al., 2022a) and Cluster Inertia is another popular one (Du and Black, 2019; Yang et al., 2025; Zhu et al., 2025).

Distribution-based measures quantify distributional properties in the embedding space, for example using entropy. An example measure is the Vendi Score, which is the exponential of the Shannon Entropy of the eigenvalues of the normalized similarity matrix (Friedman and Dieng, 2023).

Graph-based measures first build a graph with embedding vectors and then quantify its structural properties. For example, MST Dispersion is the total edge length of the minimum spanning tree of the fully connected graph over the embeddings, with pairwise distances as edge weights (Cox et al., 2021a).

Some measures consist of steps that fall into different families. We resolve such cases based on their dominant step, or, where the dominant step is itself ambiguous, on the framing of the original work that introduced the measure. Cluster Inertia, for instance, partitions embeddings with K -means and then sums the squared distances from each embedding to its assigned centroid. We classify it as geometry-based (rather than distance-based) because it captures the spatial distribution of embeddings around cluster centroids. See App. §A for a complete catalogue of measures with formulas, complexity, and references.

Distance-based <i>Aggregates pairwise distances; no constructed structures.</i>	Mean Pairwise Dist.
	Sum Pairwise Dist.
	Energy
	Diameter ²
	Bottleneck
	Sum Diameter
Geometry-based <i>Quantifies spatial properties of embeddings.</i>	Sum Bottleneck
	Chamfer Dist.
	Convex Hull Vol. (2D)
	Span (Centroid)
	Span (Medoid)
	Cluster Inertia
Distribution-based <i>Quantifies distributional properties of embeddings.</i>	Geo. Mean of Std.
	Log-Determinant
	Vendi Score
	Rényi Kernel Ent.
	DCScore
	Bins Entropy (UMAP)
Graph-based <i>Summarizes a graph built over the embeddings.</i>	MagArea
	MST Dispersion
	Graph Entropy

Table 1: Taxonomy of embedding-based diversity measures into four families.

4 Benchmark Design

Our benchmark evaluates diversity measures along **three components of controlled variation** following Stirling (2007)’s decomposition of diversity (§4.1)—variety, balance and disparity—initiated across **two complementary tiers**: a simulated data tier and a natural text tier (§4.2).

Relational and controlled evaluation. We evaluate measures relationally rather than absolutely, because in practice diversity measures are usually used comparatively (e.g., to choose among candidate corpora), which requires only that datasets be ordered correctly by diversity. Concretely, for each component (cf. Fig. 1) we build a sequence of five datasets—which we refer to as five diversity *levels*—in which only that one component varies, while the others are fixed. We then check whether each measure responds in the correct direction across the sequence. Every (tier, component) configuration is replicated under 5 random seeds to separate stable measure behaviour from sampling variability.

Notation. Three variables parameterize every dataset in the benchmark:

²Although the names of Diameter and Bottleneck suggest geometric constructs, they are simply the maximal and minimal pairwise distance, respectively; we therefore place them, together with their variants (Sum Diameter, Sum Bottleneck), in the distance-based family.

- n : the total number of elements in a dataset, fixed at $n = 10,000$ across all configurations.
- k : the number of **categories** in a dataset. A category is the basic diversity unit that an element belongs to.
- m : the number of **super-categories** into which these k categories are grouped, with $m \leq k$.

Each tier instantiates “category” and “super-category” differently: for the simulated data tier, categories are Gaussian “centres” and super-categories are “super-centres” that Gaussian centres are grouped under; for the natural text tier, categories are “topics” and super-categories are broader “domains” in Wikipedia’s hierarchy.

4.1 Three Components of Controlled Variation

The variety–balance–disparity decomposition originates in [Stirling \(2007\)](#), who unified definitions of diversity from ecology, economics, and science-and-technology studies into a single conceptual framework. Stirling argues that any substantive claim about diversity must distinguish three orthogonal questions: *how many* categories are present (**variety**), *how evenly* items are spread across those categories (**balance**), and *how different* the categories themselves are from one another (**disparity**). Subsequent work has shown that each component can be operationalized independently, provided the data comes with a categorical structure ([van Dam, 2019](#)). In NLP, this concept of diversity was adopted as early as [Lion-Bouton et al. \(2022\)](#), but has been taken up more broadly only in the last few years ([Estève et al., 2025](#); [Nguyen and Ploeger, 2025](#)).

Mapping elements to categories Applying the variety–balance–disparity decomposition requires deciding how to map *elements* to *categories*. This mapping can be performed in different ways. For example, it can be based on metadata in the dataset (e.g., explicit topic or language labels), inferred through clustering, or every element can be treated as its own category (e.g., a case discussed in [Estève et al., 2025](#)). We note that many diversity dimensions of interest in NLP do not map cleanly onto categories. For example, although language varieties (e.g., dialects) are often treated as discrete entities, in practice the boundaries are less clear cut ([Chambers and Trudgill, 1998](#)). See Limitations for a further discussion on this.

Operationalization Our benchmark design adopts the variety–balance–disparity decomposition directly. Here, elements correspond to the vectors or texts in the considered dataset. In the simulated data tier, the categories are defined by the data-generating process. In the natural text tier, they are external metadata (Wikipedia topics) used as a proxy for semantic content. Note that the mapping is part of the benchmark’s ground truth only; the measures themselves receive no category labels and only operate on the raw embeddings.

Each dataset in our benchmark contains $n = 10,000$ elements that fall into a number of categories, which are themselves organized into a smaller number of super-categories. Then, *variety* is the number of categories; *balance* is how evenly the points are distributed across them; and *disparity* is the separation between categories (see Fig. 1). Instead of manipulating the separation of categories directly, we use the number of super-categories as a proxy: fewer super-categories pack categories into related groups (low disparity), more super-categories spread them across unrelated ones (high disparity). The five diversity levels of each component are as follows.

Variety component. A sequence of datasets with increasing number of categories $k \in \{10, 20, 30, 40, 50\}$, holding balance (i.e., uniform distribution) and disparity ($m = 10$) fixed.

Balance component. A sequence of datasets with five balance levels, ranging from uniform to strongly skewed, while holding variety ($k = 50$) and disparity ($m = 10$) fixed. The full set of conditions is summarized in App. Tab. 2.

Disparity component. A sequence of datasets varying the number of super-categories $m \in \{10, 20, 30, 40, 50\}$, while holding the variety $k = 50$ and balance (uniform distribution) fixed. For instance, at $m = 10$, the 50 categories cluster within 10 super-categories (low disparity); at $m = 50$, each category belongs to a distinct super-category (high disparity).

4.2 Two Complementary Tiers

The simulated data tier (§4.2.1) contains no text: data samples are drawn from Gaussian mixtures whose structure exactly realizes the three components, isolating measure behaviour from any embedding model effect. In contrast, the natural text tier (§4.2.2) provides a more natural setting with

factors that are more difficult to control. More details can be found in Appendix §B.

4.2.1 Simulated Data Tier

Embedding-based diversity measures operate on vector space in $\mathbf{R}^d$, so a clean test case is a simulated vector dataset where the three components of variation can be exactly controlled by construction parameters. The simulated data tier provides this: data is drawn from a hierarchical Gaussian mixture model. This tier functions as a test case with a clear ground truth: every factor is controlled by construction, category labels are not noisy proxies but are part of the data-generating structure itself, and there is no embedding model in the loop. The flip side of this realization is that simulated samples of vector spaces are isotropic and Gaussian, while real text embeddings occupy a more structured and non-Gaussian space (Cai et al., 2021; Timkey and van Schijndel, 2021); the following tier introduces this complexity from real text.

Construction conditioned on n, k and m . We sample m super-category centres $s_1, \dots, s_m$ from a Gaussian $\mathcal{N}(\mathbf{0}, \sigma_{\text{super}}^2 I)$. Next, we sample k centres. For that, we first distribute each category as evenly as possible over the m super-centres. For each of the k categories, we then draw a category centre $c_i \sim \mathcal{N}(s_{m(i)}, \sigma_{\text{category}}^2 I)$, where $m(i)$ denotes the super centre assigned to category i . Then, each element of category i is drawn from $\mathcal{N}(c_i, \sigma_{\text{element}}^2 I)$. The three scale parameters satisfy $\sigma_{\text{super}} > \sigma_{\text{category}} > \sigma_{\text{element}}$, so super-centres are well separated, category centres are grouped within their super-centre, and data points are grouped within their class. Details can be found in Appendix §B.1.

As a result of the synthetic construction, each component of diversity can be manipulated independently: the number of category centres k controls variety; the distribution of the fixed n data points across the k categories controls balance; and the number of super-centres m controls disparity—because super-centres are far apart while category centres cluster tightly around their assigned super-centre, a small m bunches the k categories into a few tight groups (low disparity), while a large m spreads them apart (high disparity).

4.2.2 Natural Text Tier

Next, we look at a natural text setting. We use semantic diversity as an example, because this is

the diversity dimension that embedding-based measures have primarily been applied to and it is one of the main dimensions of diversity that researchers have been interested in (Tevet and Berant, 2021; Han et al., 2022). This tier uses real natural language text drawn from Wikipedia, making it more directly relevant to downstream use.

Annotating semantics directly at the sentence level is difficult and often ambiguous (Pavlick and Kwiatkowski, 2019), so we operationalize it through *topic-level categorical structure*: a sentence’s category is the Wikipedia topic of the article it is drawn from, and the topic serves as a proxy for semantic content. Concretely here, domains correspond to the broader super-centres and topics to smaller class centres in the simulated data tier; App. Tab. 3 illustrates its structure with examples.

Construction. The construction has two stages: selecting domains (super-categories) and topics (categories), and collecting sentences from them.

(1) To obtain domains that are semantically well-spread rather than clustered in a single region of the Wikipedia hierarchy, we embed every domain label with embedding model Qwen3-Embedding-8B³ and select via **farthest-point sampling** (FPS), a standard spread-maximizing algorithm that iteratively picks the candidate maximally distant from those already chosen. Within each selected domain, we apply FPS again over its topics.

(2) For each selected topic, we then collect its sentences from the category tree. First, we draw from articles directly under the topic, descending into subcategories only when there are not enough sentences; this avoids the semantic drift that can occur when digging too far down the category tree. Second, we cap each article’s contribution at five sentences (which prevents any single article from dominating a topic) and sample them randomly (which avoids overrepresenting the stylized leading sentences characteristic of Wikipedia articles). Further details are in App. B.2.

Label noise. Because sentences in our dataset are sampled from articles under a topic’s category, not every sentence is necessarily about its Wikipedia topic, and we treat such cases as noise. For example, the sentence “Gerhard Fisher studied electronics at the University of Dresden before emigrating to the United States.” is a background sentence about the inventor of the hand-held metal detector,

³Qwen/Qwen3-Embedding-8B

a device used in prisons, but the sentence itself is not semantically related to its Wikipedia topic *Prisons*. Noise like this comes with using real-world text data, so instead of removing it, we quantify it: one of the authors samples 750 sentences from all 75 datasets (i.e., 10 sentences per dataset) and checks whether each sentence is semantically about its Wikipedia topic. About 17.9% of the sentences do not support their Wikipedia topics. Within each component, the noise rate does not vary substantially across the five levels, so it degrades all five datasets alike, and also the levels are separated by wide intervals (k and m both move in steps of 10), so the noise should not affect the relative ordering of datasets by component. We therefore do not expect label noise to invalidate the benchmark. More details can be found in App. C.

5 Results

We display results for both the simulated data and natural text tier in Fig. 2. For the diversity components, we denote V for variety, B for balance and D for disparity.

Evaluation metric. For each diversity component (§4.1), each tier (simulated data and natural text; §4.2), and each seed (5 seeds), we obtain a sequence of 5 datasets that systematically vary along one single component. For each diversity measure M , we compute M on every dataset in the sequence and check whether its values induce the correct dataset ordering with respect to the component being varied. We report Spearman’s rank correlation ρ between the ground truth ordering and values of M , averaged over five seeds.

Embedding configurations. For the natural text tier, we use embedding model MiniLM-L6-v2⁴ ($d = 384$) (Reimers and Gurevych, 2019), since the model is optimized to align semantically similar sentence pairs. The simulated data tier uses no embedding model, but instead is sampled from a $d = 384$ -dimensional vector space.

5.1 Analysis and Discussion

No measure responds to a single component only. We regard a measure as *component-specific* if it responds to a single component ($\rho \geq 0.9$) while remaining insensitive to the other two ($|\rho| \leq 0.2$). By this criterion, no measure is component-specific on either tier: those responding strongly to one

component always respond to at least one other (Fig. 2a, 2b).

Mean Pairwise Distance (a frequently used measure) underperforms. Mean Pairwise Distance is a common choice in prior work (Tevet and Berant, 2021; Shaib et al., 2025; Miranda et al., 2025). It performs well on our simulated data; however, even on this easy and controlled setting, it does not perform perfectly (mean $\rho < 1$ on both variety and balance; Fig. 2a). On natural text, the situation is worse: on average, its correlation drops to roughly half its value on the simulated data tier (Fig. 2b). On both tiers, several alternatives—Log-Determinant, Vendi $q=1.0$, Rényi Kernel Entropy, and Graph Entropy—consistently outperform Mean Pairwise Distance across the three components.

Real text is more challenging than the simulated data setting. On the simulated data tier (Fig. 2a), measures generally perform well, but a few perform “catastrophically” (high negative correlation) on some components which leads to a lower mean than median. Six distinct measures reach a mean $\rho \geq 0.9$ and across-seed standard deviation ≤ 0.1 for simulated data on all three components. In contrast, on natural text (Fig. 2b), no measure reaches this bar and the gap between mean and median closes. It is no longer driven by a few catastrophic measures but is broad, with nearly all measures weakening. Some of the top measures on the simulated tier drop severely (e.g., Log-Determinant drops from $+1.00$ to $+0.24$ on disparity in Fig. 2). The difficulty of the natural text tier might come from the noise of real text (App. §C) or from properties of the embedding space produced by the used embedding models, or from both. The drop demonstrates the necessity to benchmark measures on real text as well.

Disparity is the hardest component on both tiers. For both tiers, disparity reaches the lowest mean and median scores. For the simulated data tier (Fig. 2a), variety scores are higher, but for natural text (Fig. 2b), variety scores are lower than balance scores.

On both tiers, every family contains comparatively good measures and the best/worst ones stay the same. Averaged over the three components, graph- and distribution-based measures are generally stronger than geometry- and distance-based ones on both tiers, and graph-based mea-

⁴[sentence-transformers/all-MiniLM-L6-v2](#)

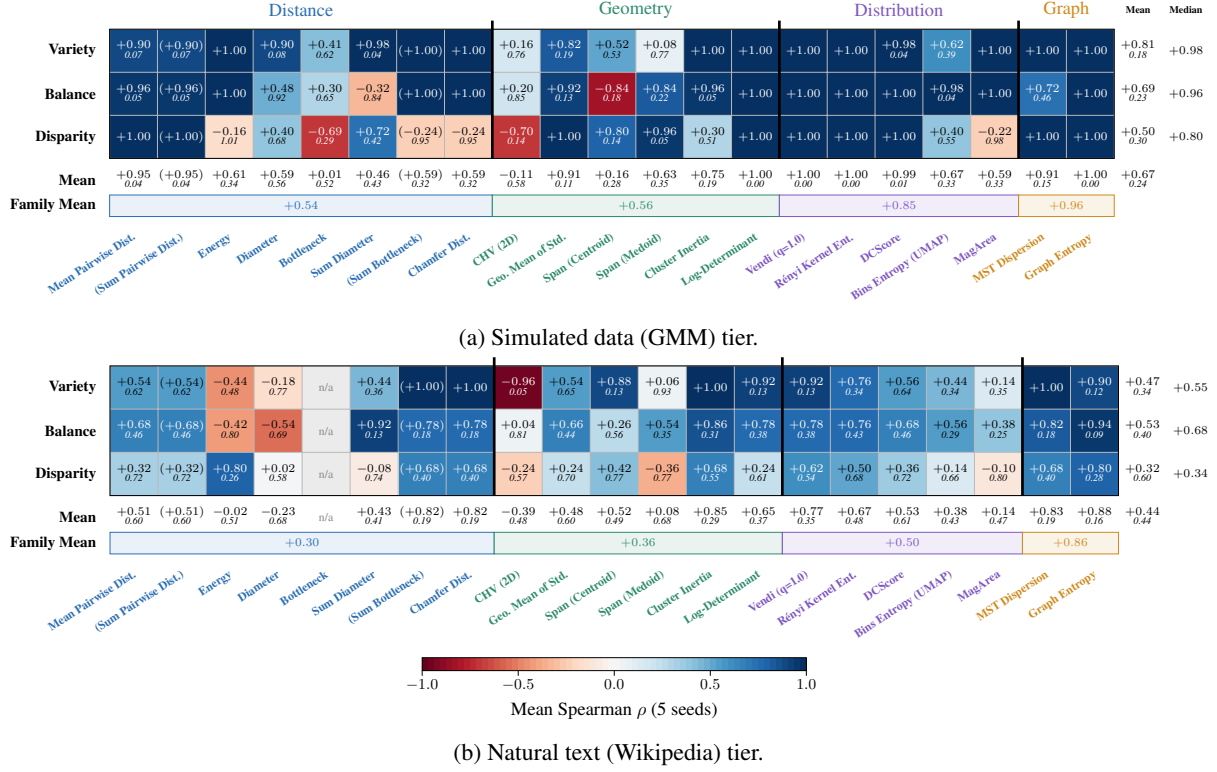


Figure 2: EMBDIVBENCH results for the (a) simulated data and the (b) natural text tier. We display per-measure mean Spearman’s rank correlation ρ between the ground-truth ordering and the measure values (averaged over 5 seeds) for the Variety, Balance and Disparity components; the smaller italic number below each ρ is the across-seed standard deviation. Columns are grouped and coloured by the four families. The two rightmost columns give the mean ρ and the median ρ ; the bottom row gives the per-measure mean ρ . Note that two pairs of measures are the mean and the sum of the same quantity (Mean Pairwise Dist. / Sum Pairwise Dist., and Chamfer Dist. / Sum Bottleneck) and therefore at our fixed dataset size n , they yield identical ρ ; the parenthesized member of each pair is excluded from the family and mean aggregates so that it is not double-counted.

asures degrade least across tiers (0.10, versus 0.35 for distribution). Even with the noticeable gap between family averages, each family contains a strong member: the six measures meeting our bar (mean $\rho \geq 0.9$ with $SD \leq 0.1$) on the simulated tier span all four families, and on natural text the best member of each family lies within a gap of 0.11 (Chamfer Dist. +0.82, Cluster Inertia +0.85, Vendi +0.77, Graph Entropy +0.88). Therefore, we do not regard any family as inherently inferior, as a poorly designed measure can be designed for any family. So a low average is evidence about the members we survey rather than about the approach itself. Across tiers, Graph Entropy performs the best and Convex Hull Volume (CHV) the worst.

6 Robustness Checks

Sentence embeddings are known to be sensitive to model choice and their dimensionality (Ethayarajh, 2019; Tsukagoshi and Sasano, 2025). Because our results are produced on top of one fixed embedding

setup, we test the robustness of our findings along two aspects: On the simulated data tier we vary the dimension d of the GMM (§6.1); on the natural text tier we vary the embedding model used to encode the sentences (§6.2).

6.1 Dimensionality

We rerun the simulated data (GMM) tier at five dimensions $d \in \{256, 384, 1024, 2048, 4096\}$. Results are shown in Fig. 5 (Appendix).

Most findings are stable across dimensions, with a few exceptions. Four of the six measures that perform well (i.e., $\rho \geq 0.9$ with $SD \leq 0.1$ on all three components) at $d = 384$ —Vendi ($q=1$), Rényi Kernel Entropy, DCScore and Graph Entropy—also do so for all other tested dimensions; Mean Pairwise Distance falls below the bar at $d = 256$ and $d = 1024$, and Log-Determinant at $d = 2048$. At the family level, the distribution- and graph-based measures stay the strongest, and the distance- and geometry-based measures the weak-

est. The distance family, however, overtakes geometry at every dimensionality other than $d = 384$, though their mean ρ differs by at most 0.13 at any dimensionality; we therefore do not read their relative order as a stable finding.

6.2 Embedding models

We re-encode the natural text tier datasets with two different semantic embedding models (*Qwen3-Embedding-8B*⁵ and *e5-mistral-7b-instruct*⁶) and an embedding model sensitive to style rather than semantics (*Style Embedding*⁷ from Wegmann et al., 2022). Results are shown in App. Fig. 6 and Fig. 7.

Embeddings need to capture the intended diversity dimension. As expected, diversity measures using Style Embedding cannot capture semantic diversity.

Findings can differ for a different semantic embedding. Most findings with MiniLM hold for Qwen3 and E5 (see §D), but individual measures can diverge. For example, Mean Pairwise Distance is weak on both balance and variety with MiniLM, reasonable on both with Qwen3, and strong on balance but weak on variety with the similarly-sized E5.⁸ We leave further investigation to future work.

7 Conclusion

In this work, we collect and categorize 21 embedding-based diversity measures into four families (Tab. 1). Further, we introduce the EMBDIVBENCH benchmark to evaluate diversity measures along three components: variety, balance and disparity. We evaluate the measures in a perfectly controlled simulated data setting— independent from a specific definition of diversity (e.g., stylistic vs. semantic). Surprisingly, we find that a majority of measures already behave unexpectedly on this easy and controlled setting. Additionally, EMBDIVBENCH includes a semantic tier to exemplify a specific diversity dimension (i.e., semantic) with real text data. The real data setting shows that the requirements for diversity measures might be quite different on an embedding space as more measures fall short of our expectations. No

measure performs perfectly. We invite the community to expand our benchmark with other real data settings for different diversity dimensions. All measures benchmarked here and more are implemented in the *emb-diversity*⁹ package, which we release as follow-up work (Su et al., 2026). The package makes it straightforward to apply measures to new data, or to add newly proposed measures and compare them against the ones surveyed here.

Recommendations. The most suitable diversity measure depends on various factors, including the diversity component of interest, the embedding model chosen, and the available computational budget (see App. Tab. 7 for computational cost of each measure). We make the following recommendations based on our empirical benchmark results:

- **Do not use Mean Pairwise Distance.** Although intuitive and popular, Mean Pairwise Distance performs worse than several other measures on simulated and text data.
- **Select measures based on the component(s) of interest.** No measure performs best on all three components. We therefore recommend identifying the component of interest first and consulting our component-specific shortlists. Our recommendations should be read as the currently best available option rather than a future proof quality judgement.
 - **Variety:** Chamfer Distance¹⁰ and MST Dispersion
 - **Balance:** Graph Entropy
 - **Disparity:** MST Dispersion

If no single component is of primary interest, Graph Entropy is currently the best-supported single choice, followed by Vendi ($q=1$). See App. E for the detailed rationale.

- **Choose your embedding model carefully.** Choose an embedding model that (primarily) captures the dimension of diversity you want to capture (e.g., semantic diversity). Even among models targeting the same diversity dimension, differences between embedding models can affect measure behaviour and rankings (§6.2). We recommend practitioners to confirm that the embedding they use is sensitive and faithful to their diversity dimension of interest.

⁵[Qwen/Qwen3-Embedding-8B](#)

⁶[intfloat/e5-mistral-7b-instruct](#)

⁷[AnnaWegmann/Style-Embedding](#)

⁸Because Qwen3 and E5 are optimized for retrieval and ranking rather than for embedding distances that reflect semantic similarity, we base our main conclusions on MiniLM.

⁹<https://github.com/nlpsoc/emb-diversity/>

¹⁰Sum Bottleneck is rank-equivalent to Chamfer Distance at a fixed dataset size and therefore interchangeable here (App. §A).

Limitations

Categorical structure required. The V-B-D decomposition needs a partition of the data into categories: variety counts them, balance is the distribution over them, and disparity is the separation between them. Both our tiers provide one by construction, but some diversity dimensions are graded rather than categorical (e.g., formality is relative, see Wegmann et al., 2026). This raises a broader question for future work: to what extent the V-B-D decomposition, which was developed outside NLP, is suitable for measuring diversity in natural language data. Future work could also investigate alternative mappings from texts to categories: some measures already carry an implicit mapping (e.g., DCScore treats each sample as its own class; App. §A), and a measure whose implicit mapping aligns with the categories used to construct a dataset may be favoured by that alignment.

Confounds inherited from Wikipedia. First, topic labels are an imperfect proxy: not all sentences within a given article are directly related to the article’s topic, introducing noise into the topic clusters (see our analysis in App. C). Second, disparity on our natural text tier is operationalized as the number of domains spanned. This proxy assumes that cross-domain topics lie further apart in embedding space than within-domain topics, but Wikipedia’s editorial taxonomy does not enforce this. For example, the topic *History of mathematics* (under domain History) and topic *Mathematician* (under domain Mathematics) belong to different domains yet are semantically close.

Our benchmark only captures part of the picture. Our benchmark tests the measures against the three diversity components and only considers the ordering of datasets; it does not consider the raw diversity scores or change in those scores. Future work could investigate to what extent score differences are perceptible to humans or have meaningful downstream effects on trained models (Kocmi et al., 2024). Our work also does not test the formal axiomatic properties (monotonicity, uniqueness, continuity) studied by Mironov and Prokhorenkova (2025); these two approaches are complementary, as a measure may satisfy these axioms yet respond weakly on our three-component benchmark. Other aspects, such as actionability (Delobelle et al., 2024), should also inform measure selection.

Hyperparameter settings untested. Several measures involve hyperparameters that we fix to a single value—typically a published default or a setting inherited from prior work (see App. Tab. 4). A systematic sweep over these parameter values may shift the relative rankings we report.

Sentence-level units and English Wikipedia only. Our benchmark does not evaluate the measures on paragraph- or document-level units, which would require different pipelines (mean-pooling, sliding-window aggregation, or long-context encoders). Whether the relative behaviour of measures we report transfers to diversity measurement with longer units remains therefore an open question for future work. Also, our natural-text tier is built exclusively from English Wikipedia, whose category hierarchy, article-length distribution, and topical coverage reflect the editorial choices of one language community. Still, the taxonomy and pipeline we introduce are not language-specific, and could be extended to other Wikipedia languages in future work.

Embedding model artefacts. Because the measurements in the natural text tier are computed over sentence embeddings, they are also influenced by properties of the embedding space itself (anisotropy, hubness) in addition to properties of the underlying data. One concrete step for future work would be calibrating the simulated tier to real embedding statistics: e.g., drawing simulated data from distributions whose covariance is estimated from real sentence embeddings instead of isotropic Gaussians, to better simulate the anisotropy of natural embedding spaces, which would help clarify the extent to which the performance drop from the simulated to the natural text tier (§5) can be explained by embedding-space artefacts.

Ethical Considerations

Risk of misuse, PII and offensive content. Our work is methodological and introduces no new data or models: all text is from English Wikipedia, and the released benchmark consists of dataset configurations and evaluation code only. We apply no name redaction, as public-figure names in biographical articles are public information and carry no contact or identification details; we also apply no offensive-content filtering, as sensitive topics (e.g., controversial figures) appear in encyclopaedic context. We do not foresee ethical risks beyond standard methodological caveats.

Acknowledgments

This work is supported by the ERC Starting Grant DataDivers 101162980.

References

- Aanchal Agarwal, Anup Bera, and Tanusree De. 2026. [SafeText: A Unified Approach for Detecting and Mitigating Toxicity and Bias in Textual Data](#). In *Safe, Secure, Ethical, Responsible Technologies and Emerging Applications*, pages 71–92, Cham. Springer Nature Switzerland.
- Tevin Atwal, Chan Nam Tieu, Yefeng Yuan, Zhan Shi, Yuhong Liu, and Liang Cheng. 2025. [Privacy-Preserving Synthetic Review Generation with Diverse Writing Styles Using LLMs](#). *arXiv preprint*.
- Yang Ba, Mohammad Sadeq Abolhasani, Michelle V. Mancenido, and Rong Pan. 2026. [Measuring Dataset Diversity from a Geometric Perspective](#). *arXiv preprint*.
- Yang Ba, Mohammad Sadeq Abolhasani, and Rong Pan. 2025. [Predict training data quality via its geometry in metric space](#). In *New Perspectives in Graph Machine Learning*.
- C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. 1996. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483.
- Su Lin Blodgett, Johnny Wei, and Brendan O’Connor. 2018. [Twitter Universal Dependency Parsing for African-American and Mainstream American English](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.
- Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2021. [Isotropy in the contextual embedding space: Clusters and manifolds](#). In *International Conference on Learning Representations*.
- Alfonso Cevallos, Friedrich Eisenbrand, and Sarah Morell. 2018. [Diversity Maximization in Doubling Metrics](#). In *29th International Symposium on Algorithms and Computation (ISAAC 2018)*, volume 123 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 33:1–33:12, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.
- Jack K Chambers and Peter Trudgill. 1998. *Dialectology*. Cambridge University Press.
- Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abidin. 2024. [On the Diversity of Synthetic Data and its Impact on Training Large Language Models](#). *arXiv preprint*.
- Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. [Fairness in Large Language Models: A Taxonomic Survey](#). *SIGKDD Explor. Newsl.*, 26(1):34–48.
- Michael A. Covington and Joe D. McFall. 2010. [Cutting the Gordian Knot: The Moving-Average Type-Token Ratio \(MATTR\)](#). *Journal of Quantitative Linguistics*, 17(2):94–100.
- Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian Von Der Weth, and Brian Y. Lim. 2021a. [Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–35.
- Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian von der Weth, and Brian Y. Lim. 2021b. [Directed Diversity: Leveraging Language Embedding Distances for Collective Creativity in Crowd Ideation](#). In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–35.
- Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. [Metrics for what, metrics for whom: Assessing actionability of bias evaluation metrics in NLP](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21669–21691, Miami, Florida, USA. Association for Computational Linguistics.
- Wenchao Du and Alan W Black. 2019. [Boosting Dialog Response Generation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 38–43, Florence, Italy. Association for Computational Linguistics.
- Yuanqi Du, Tianfan Fu, Jimeng Sun, and Shengchao Liu. 2022. [Molgensurvey: A systematic survey in machine learning models for molecule design](#). *arXiv preprint arXiv:2203.14500*.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Sibli, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. [MMTEB: Massive multilingual text embedding benchmark](#). In *The Thirteenth International Conference on Learning Representations*.
- Louis Estève and Kaja Dobrovoljc. 2026. [DELTA: A toolkit for measuring linguistic diversity in dependency-parsed corpora](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 75–85, Rabat, Morocco. Association for Computational Linguistics.
- Louis Estève, Christophe Servan, Thomas Lavergne, and Agata Savary. 2026. [A Diversity Diet for a Healthier Model: A Case Study of French Modern-BERT](#). In *64th Annual Meeting of the Association for*

- Computational Linguistics (ACL 2026)*, San Diego, United States.
- Louis Estève, Marie-Catherine de Marneffe, Nurit Melnik, Agata Savary, and Olha Kanishcheva. 2025. A survey of diversity quantification in natural language processing: The why, what, where and how. *arXiv preprint*.
- Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.
- Ziqing Fan, Siyuan Du, Shengchao Hu, Pingjie Wang, Li Shen, Ya Zhang, Dacheng Tao, and Yanfeng Wang. 2025. Combatting dimensional collapse in LLM pre-training data via submodular file selection. In *The Thirteenth International Conference on Learning Representations*.
- Dan Friedman and Adji Bousso Dieng. 2023. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.
- Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.
- Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.
- Seungju Han, Beomsu Kim, and Buru Chang. 2022. Measuring and improving semantic diversity of dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 934–950, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Xiuyuan Hu, Guoqing Liu, Quanming Yao, Yang Zhao, and Hao Zhang. 2024a. Hamiltonian diversity: effectively measuring molecular diversity by shortest hamiltonian circuits. *Journal of Cheminformatics*, 16(1):94.
- Xiuyuan Hu, Guoqing Liu, Quanming Yao, Yang Zhao, and Hao Zhang. 2024b. Hamiltonian diversity: effectively measuring molecular diversity by shortest Hamiltonian circuits. *Journal of Cheminformatics*, 16(1):94.
- Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. 2023. An information-theoretic evaluation of generative models in learning multi-modal distributions. *Advances in Neural Information Processing Systems*, 36:9931–9943.
- Folger Karsdorp, Enrique Manjavacas, and Lauren Fonteyn. 2022. Introducing functional diversity: A novel approach to lexical diversity in (historical) corpora. In *Proceedings of the Computational Humanities Research Conference 2022, CHR 2022, Antwerp, Belgium, December 12-14, 2022*, volume 3290 of *CEUR Workshop Proceedings*, pages 114–126. CEUR-WS.org.
- Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.
- Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286.
- Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, Density, and Homogeneity: Quantitative Characteristic Metrics for Text Collections. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1739–1746, Marseille, France. European Language Resources Association.
- Jing Yang Lee, Kong Aik Lee, and Woon Seng Gan. 2022. A Randomized Link Transformer for Diverse Open-Domain Dialogue Generation. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.
- Tom Leinster and Christina A Cobbold. 2012a. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489.
- Tom Leinster and Christina A. Cobbold. 2012b. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Katharina Limbeck, Rayna Andreeva, Rik Sarkar, and Bastian Rieck. 2024. Metric space magnitude for

- evaluating the diversity of latent representations. *Advances in Neural Information Processing Systems*, 37:123911–123953.
- Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. 2022. [Evaluating diversity of multiword expressions in annotated text](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Yang Janet Liu and Amir Zeldes. 2023. [Why Can’t Discourse Parsing Generalize? A Thorough Investigation of the Impact of Data Diversity](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.
- Brando Miranda, Alycia Lee, Sudharsan Sundar, Allison Casasola, Rylan Schaeffer, Elyas Obbad, and Sanmi Koyejo. 2025. [Beyond Scale: The Diversity Coefficient as a Data Quality Metric for Variability in Natural Language Data](#). *arXiv preprint*.
- Mikhail Mironov and Liudmila Prokhorenkova. 2025. [Measuring diversity: Axioms and challenges](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 44396–44411. PMLR.
- Klaus Nehring and Clemens Puppe. 2002. [A Theory of Diversity](#). *Econometrica*, 70(3):1155–1198.
- Dong Nguyen and Esther Ploeger. 2025. [We need to measure data diversity in NLP — better and broader](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8812–8821, Suzhou, China. Association for Computational Linguistics.
- Amey P. Pasarkar and Adji Bousso Dieng. 2024. [Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning](#). In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3808–3816. PMLR.
- Ellie Pavlick and Tom Kwiatkowski. 2019. [Inherent Disagreements in Human Textual Inferences](#). *Transactions of the Association for Computational Linguistics*, 7:677–694.
- Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. [What is “Typological Diversity” in NLP?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5681–5700, Miami, Florida, USA. Association for Computational Linguistics.
- Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Nguyen. 2022. [MINION: a Large-Scale and Diverse Dataset for Multilingual Event Detection](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2286–2299, Seattle, United States. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Benjamin Schiller, Johannes Daxenberger, Andreas Waldis, and Iryna Gurevych. 2024. [Diversity Over Size: On the Effect of Sample and Topic Sizes for Topic-Dependent Argument Mining Datasets](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10870–10887, Miami, Florida, USA. Association for Computational Linguistics.
- Chantal Shaib, Venkata S Govindarajan, Joe Barrow, Jiuding Sun, Alexa Siu, Byron C Wallace, and Ani Nenkova. 2025. [Standardizing the measurement of text diversity: A tool and comparative analysis](#). In *Proceedings of The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 36–46, Mumbai, India. Association for Computational Linguistics.
- Katherine Stasaski and Marti Hearst. 2022. [Semantic diversity in dialogue with natural language inference](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, United States. Association for Computational Linguistics.
- Andy Stirling. 2007. [A general framework for analysing diversity in science, technology and society](#). *Journal of The Royal Society Interface*, 4(15):707–719.
- Cantao Su, Menan Velayuthan, Esther Ploeger, Dong Nguyen, and Anna Wegmann. 2026. [emb-diversity: A tool for embedding-based measurement of data diversity](#). *Preprint*, arXiv:2607.19848.
- Guy Tevet and Jonathan Berant. 2021. [Evaluating the Evaluation of Diversity in Natural Language Generation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.
- William Timkey and Marten van Schijndel. 2021. [All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

- 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hayato Tsukagoshi and Ryohei Sasano. 2025. [Redundancy, Isotropy, and Intrinsic Dimensionality of Prompt-based Text Embeddings](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25915–25930, Vienna, Austria. Association for Computational Linguistics.
- Alje van Dam. 2019. [Diversity and its decomposition into variety, balance and disparity](#). *Royal Society Open Science*, 6(7):190452.
- Rob van der Goot, Esther Ploeger, Verena Blaschke, and Tanja Samardzic. 2025. [DistalS: a comprehensive collection of language distance measures](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 307–318, Suzhou, China. Association for Computational Linguistics.
- Fedor Velikonitvsev, Mikhail Mironov, and Liudmila Prokhorenkova. 2024. [Challenges of generating structurally diverse graphs](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 57993–58022. Curran Associates, Inc.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Improving Text Embeddings with Large Language Models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.
- Peiqi Wang, Yikang Shen, Zhen Guo, Matthew Stallone, Yoon Kim, Polina Golland, and Rameswar Panda. 2025a. [Diversity measurement and subset selection for instruction tuning datasets](#). In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.
- Zaitian Wang, Jinghan Zhang, Xinhao Zhang, Kunpeng Liu, Pengfei Wang, and Yuanchun Zhou. 2025b. [Diversity-oriented data augmentation with large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22265–22283, Vienna, Austria. Association for Computational Linguistics.
- Anna Wegmann, Cristina Aggazzotti, Rafael Rivera Soto, and Dong Nguyen. 2026. [A survey on representing linguistic style: Challenges and opportunities](#). To appear at EMNLP 2026.
- Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. [Same author or just same topic? towards content-independent style representations](#). In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.
- Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. 2018. [Practical Diversified Recommendations on YouTube with Determinantal Point Processes](#). In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2165–2173, Torino, Italy. ACM.
- Yutong Xie, Ziqiao Xu, Jiaqi Ma, and Qiaozhu Mei. 2023. [How much space has been explored? measuring the chemical space covered by databases and machine-generated molecules](#). In *The Eleventh International Conference on Learning Representations*.
- Vikas Yadav, Hyuk joon Kwon, Vijay Srinivasan, and Hongxia Jin. 2024. [Explicit over Implicit: Explicit Diversity Conditions for Effective Question Answer Generation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6876–6882, Torino, Italia. ELRA and ICCL.
- Yuming Yang, Yang Nan, Junjie Ye, Shihan Dou, Xiao Wang, Shuo Li, Huijie Lv, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. [Measuring data diversity for instruction tuning: A systematic analysis and a reliable metric](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18530–18549, Vienna, Austria. Association for Computational Linguistics.
- Yu Yu, Shahram Khadivi, and Jia Xu. 2022a. [Can data diversity enhance learning generalization?](#) In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Yu Yu, Shahram Khadivi, and Jia Xu. 2022b. [Can Data Diversity Enhance Learning Generalization?](#) In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4933–4945, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Tianhui Zhang, Bei Peng, and Danushka Bollegala. 2024. [Improving diversity of commonsense generation by large language models via in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9226–9242, Miami, Florida, USA. Association for Computational Linguistics.
- Tianhui Zhang, Bei Peng, and Danushka Bollegala. 2025a. [Evaluating the evaluation of diversity in commonsense generation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 24258–24275, Vienna, Austria. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models](#). *arXiv preprint*.

Dora Zhao, Jerone Andrews, Orestis Papakyriakopoulos, and Alice Xiang. 2024. [Position: Measure dataset diversity, don’t just claim it](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 60644–60673. PMLR.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Texygen: A benchmarking platform for text generation models](#). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR ’18*, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

Yuchang Zhu, Huizhe Zhang, Bingzhe Wu, Jintang Li, Zibin Zheng, Peilin Zhao, Liang Chen, and Yatao Bian. 2025. [Measuring diversity in synthetic datasets](#). In *Forty-second International Conference on Machine Learning*.

A Detailed Taxonomy of Embedding-based Diversity Measures

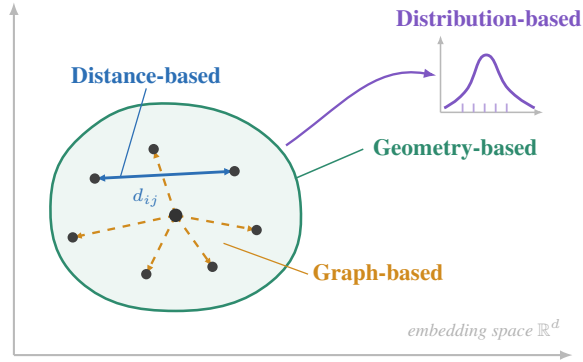


Figure 3: Illustration behind the four measure families on a toy vector space.

This Appendix presents the full taxonomy of embedding-based diversity measures surveyed in this work, organized by their underlying mathematical formulation. Fig. 3 provides a visual intuition for the four families.

Notation and Conventions Throughout this Appendix, we use the following notation:

- $X = \{x_1, \dots, x_n\}$: a dataset of n samples

- $\mathbf{e}_i \in \mathbb{R}^d$: the d -dimensional embedding of sample x_i

- $d_{ij} = d(\mathbf{e}_i, \mathbf{e}_j)$: distance between embeddings (cosine distance, unless specified)

- $s_{ij} = \text{sim}(\mathbf{e}_i, \mathbf{e}_j)$: similarity between embeddings (cosine similarity, unless specified)

Comparability across dataset sizes. Several measures are defined as sums over elements or pairs (e.g. Sum Pairwise Distance, Sum Diameter, Sum Bottleneck, Graph Entropy, MST Dispersion) and therefore grow with n , so their values are not comparable across datasets of different sizes. Since n is fixed at 10,000 across every dataset in our benchmark, this does not affect any result we report.

A.1 Distance-based Measures

This family directly measures diversity using raw pairwise distances or similarities, without building any additional structures such as graphs, cluster assignments, or geometric constructs (e.g., convex hulls) that require further processing beyond distance computation.

A.1.1 Mean Pairwise Distance

Intuition. This measure averages the distance of each sample to all other samples in the embedding space. A higher value indicates that samples are further apart (i.e., more diverse), whereas a lower value implies less diversity.

Computation. Compute the average of all off-diagonal pairwise distances:

$$\text{AvgDist}(\mathbf{X}) = \frac{2}{n(n-1)} \sum_{i < j} d_{ij} \quad (1)$$

In principle, this measure generalizes to any function capable of quantifying similarity or distance between two samples. However, within the context of embeddings, **cosine similarity** is the most prevalent choice. When employed, this measure is often referred to as **Remote-Click** (Cox et al., 2021b). Some works also refer to the similarity-based variant **Self-cosSim** (Tevet and Berant, 2021; Zhang et al., 2024; Miranda et al., 2025). Since cosine similarity is $1 - \text{cosine distance}$, these measures are mathematically equivalent (linearly related).

Complexity. $\mathcal{O}(n^2d)$

A.1.2 Sum Pairwise Distance

Intuition. By summing all pairwise distances, this measure captures the overall spread of samples in the embedding space. Different terms refer to the same formulation, originally introduced as Max Dispersion (MD) by Yu et al. (2022b) and later adopted as DistSum by Yang et al. (2025).

Higher values indicate greater diversity.

Computation. Compute the sum of all pairwise distances:

$$\text{MD}(\mathbf{X}) = \text{DistSum}(\mathbf{X}) = \sum_{i < j} d_{ij}, \quad (2)$$

Complexity. $\mathcal{O}(n^2d)$

A.1.3 Chamfer Distance

Intuition. This measure originates from computational geometry and computer vision (specifically point cloud processing), where it is used to compare *two* sets of points by measuring how close every point in one set is to its nearest neighbour in the other set. For point sets A and B , the Chamfer Distance averages: (i) the distance from each point in A to its nearest neighbour in B , and (ii) the distance from each point in B to its nearest neighbour in A .

In the context of diversity research in NLP, prior work adapts this idea to a *single* dataset (Cox et al., 2021b; Zhang et al., 2025a). Instead of comparing two sets, the measure averages, for each sample, the distance to its nearest neighbour. This reflects the *local compactness* of the embedding space: datasets with tighter clusters have smaller nearest-neighbour distances, while more dispersed datasets yield larger values.

Computation. Compute the average of the minimum pairwise distances:

$$\text{Chamfer}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n \min_{j \neq i} d_{ij} \quad (3)$$

Complexity. $\mathcal{O}(n^2d)$

A.1.4 Energy(γ)

Intuition. This measure originates from classical mechanics. It takes a dataset as a system of equally charged particles distributed in the embedding space and measures its diversity as the negative potential energy (Velikonitsev et al., 2024; Mironov and Prokhorenkova, 2025). Each pair of

samples repels each other, inversely proportional to their distance raised to the power γ . The measure encourages uniform spacing and penalizes close pairs; for exact duplicates ($d_{ij} = 0$), the measure diverges to $-\infty$. Thus, when samples are further apart, the negative potential energy (and hence the diversity score) increases.

Computation. The Energy measure of parameter $\gamma > 0$ is defined as:

$$\text{Energy}_\gamma(\mathbf{X}) = -\frac{2}{n(n-1)} \sum_{i < j} \frac{1}{d_{ij}^\gamma}. \quad (4)$$

The negative sign ensures that configurations with smaller pairwise distances (i.e. higher repulsive energy) receive lower diversity values, so that increasing all pairwise distances increases the overall diversity score. The value is always negative, and higher values (i.e. closer to zero) indicate greater diversity. In the special case of $\gamma = 1$, the measure corresponds to the potential energy of a system of point charges under Coulomb repulsion.

Complexity. $\mathcal{O}(n^2d)$ (pairwise distance computation for all sample pairs)

A.1.5 Diameter (maximal pairwise distance)

Intuition. This measure finds the largest distance between any two embeddings of a dataset (Mironov and Prokhorenkova, 2025; Xie et al., 2023). It shows how far apart the most different embeddings are in the space, figuratively the “width” of the point cloud.

Computation. The Diameter is the largest of these pairwise distances, defined as:

$$\text{Diameter}(\mathbf{X}) = \max_{i < j} d_{ij} = \max_{i < j} d(\mathbf{e}_i, \mathbf{e}_j). \quad (5)$$

Complexity. $\mathcal{O}(n^2d)$ (pairwise distance computation)

A.1.6 Bottleneck (minimal pairwise distance)

Intuition. This measure computes the smallest distance between any two data points (Mironov and Prokhorenkova, 2025; Xie et al., 2023).

Computation. The Bottleneck is defined as:

$$\text{Bottleneck}(\mathbf{X}) = \min_{i < j} d_{ij}. \quad (6)$$

Caveats. As a stand-alone diversity measure, this measure can be misleading—a single pair of near-duplicates can collapse its value regardless of the rest of the distribution.

Complexity. $\mathcal{O}(n^2d)$

A.1.7 Sum Diameter (sum of maximal pairwise distances)

Intuition. This measure sums up the maximum distance from each data point to its farthest data point. While the Diameter measure relies on a single pair of extreme points (making it sensitive to outliers), Sum Diameter aggregates the maximum separation for every data point (Mironov and Prokhorenkova, 2025; Xie et al., 2023).

Computation. For each sample x_i , find its maximum distance to all other samples and sum these maxima:

$$\text{SumDiameter}(\mathbf{X}) = \sum_{i=1}^n \max_{j \neq i} d_{ij}. \quad (7)$$

Complexity. $\mathcal{O}(n^2d)$

A.1.8 Sum Bottleneck (sum of minimal pairwise distances)

Intuition. While the Bottleneck measure is determined only by the closest pair of samples (making it sensitive to (near) duplicates), this measure aggregates the minimum distance from each data point to its nearest neighbour for every data point (Mironov and Prokhorenkova, 2025; Xie et al., 2023).

Computation. For each sample x_i , find its minimum distance to all other samples and sum these minima:

$$\text{SumBottleneck}(\mathbf{X}) = \sum_{i=1}^n \min_{j \neq i} d_{ij}. \quad (8)$$

Caveats. Unlike Diameter or Sum Diameter, this measure primarily reflects *local* density rather than the overall spread of the dataset.

Complexity. $\mathcal{O}(n^2d)$

A.2 Geometry-based Measures

This family treats the embeddings of a dataset as a point cloud in $\mathbb{R}^d$ and quantifies diversity through the geometric properties of that configuration—its extent, volume, or spread around a representative point, etc.

A.2.1 Convex Hull Volume

¹¹ **Intuition.** This measure treats each embedding as a data point. The convex hull then represents the smallest convex region containing all points in the embedding space (Yu et al., 2022b). A higher value indicates that the samples are more widely distributed, indicating greater diversity.

Computation. The convex hull $\mathcal{H}(\mathbf{X})$ is defined as the smallest convex set containing all embeddings:

$$\mathcal{H}(\mathbf{X}) = \left\{ \sum_{i=1}^n \alpha_i \mathbf{e}_i \mid \alpha_i \geq 0, \sum_{i=1}^n \alpha_i = 1 \right\}.$$

The Convex Hull Volume is then the d -dimensional volume of $\mathcal{H}(\mathbf{X})$:

$$\text{CHV}(\mathbf{X}) = \text{Vol}_d(\mathcal{H}(\mathbf{X})). \quad (9)$$

In practice, the volume can be efficiently approximated using the Quickhull algorithm (Barber et al., 1996), which recursively partitions the space to compute the convex envelope of the points. When embeddings are high-dimensional, the algorithm operates on the subset of boundary points that define the hull.

Complexity. Expected $\mathcal{O}(n \log n)$ using Quickhull

A.2.2 Geometric Mean of Standard Deviations

Intuition. This measure was introduced by Lai et al. (2020) under the name Diversity and adopted by Yang et al. (2025) under the name Radius. It is the geometric mean of the per-dimension standard deviations of the embeddings. Treating each embedding dimension as an independent axis of variation, the per-dimension standard deviations define an axis-aligned ellipsoid whose semi-axis lengths are $\sigma_1, \dots, \sigma_d$. The product $\prod_j \sigma_j$ is proportional to the volume of this ellipsoid. Taking its d -th root converts this product into an average per-axis spread. A larger value indicates greater dispersion, and hence greater diversity.

Computation. Let σ_j denote the standard deviation of the j -th embedding dimension across all samples. The measure is defined as the geometric mean of these per-dimension standard deviations:

$$\text{GeoMeanStd}(\mathbf{X}) = \left(\prod_{j=1}^d \sigma_j \right)^{\frac{1}{d}}. \quad (10)$$

¹¹In our benchmark we compute CHV after reducing embeddings to two dimensions with UMAP, as direct computation in the original embedding space ($d = 384$) is intractable.

Note that the measure collapses to zero whenever any standard deviation is zero.

Complexity. $\mathcal{O}(nd)$ (computing the standard deviation along each embedding dimension).

A.2.3 Span (Centroid)

Intuition. This measure computes the average distance from each embedding to the centroid of all embeddings in the dataset (Cox et al., 2021b). The centroid represents the centre of the embedded dataset, and Span quantifies how far, on average, the embeddings deviate from this central point. Datasets with larger Span values are more widely distributed around the mean embedding, indicating higher diversity.

Computation. First, compute the centroid of all embeddings:

$$\mu = \frac{1}{n} \sum_{i=1}^n \mathbf{e}_i. \quad (11)$$

Then, compute the set of distances from all embeddings to this centroid:

$$D = \{d(\mathbf{e}_i, \mu) \mid i = 1, \dots, n\}.$$

To ensure robustness against outliers, Cox et al. (2021b) define the measure not as the maximum, but as the **90th percentile** of these distances:

$$\text{Span}_{\text{centroid}}(\mathbf{X}) = \text{Percentile}_{90}(D). \quad (12)$$

Complexity. $\mathcal{O}(nd)$

A.2.4 Span (Medoid)

Intuition. This measure takes the mean distance from all samples to the medoid, i.e., the sample whose average distance to all other samples is minimal (Cox et al., 2021b). Higher values indicate a higher diversity. Cox et al. (2021b) introduce it under the name Sparseness; we label it Span (Medoid) to pair it with Span (Centroid) (§A.2.3). Note that the two differ in both the reference point and the aggregation: Span (Centroid) takes a percentile of distances to a mean-based reference, whereas this measure takes a mean of distances to a median-like reference.

Computation. First, the medoid is identified as:

$$m = \arg \min_{x_i \in \mathbf{X}} \sum_{j=1}^n d_{ij}. \quad (13)$$

Compute the mean distance of all samples to this medoid:

$$\text{Span}_{\text{medoid}}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n d(\mathbf{e}_i, \mathbf{e}_m). \quad (14)$$

Complexity. $\mathcal{O}(n^2d)$ (pairwise distance matrix required to identify medoid)

A.2.5 Cluster Inertia

Intuition. In physics, the *moment of inertia* describes how difficult it is to change an object’s rotational state. Objects with more mass distributed farther from the axis of rotation have a larger moment of inertia. By analogy, embeddings lying farther from their cluster centroid contribute more to cluster inertia. A higher inertia value indicates that samples are scattered farther from their cluster centroids, indicating a more diverse dataset.

Computation. Given a dataset $\mathbf{X}$, partition the embeddings into K clusters $\{C_1, \dots, C_K\}$ using K -means (Yang et al., 2025; Du and Black, 2019). Let μ_j denote the centroid of cluster C_j . The Cluster Inertia is defined as the total within-cluster sum of squared distances:

$$\text{Inertia}(\mathbf{X}) = \sum_{j=1}^K \sum_{\mathbf{e}_i \in C_j} \|\mathbf{e}_i - \mu_j\|^2. \quad (15)$$

Note that unlike most other measures in our taxonomy, Cluster Inertia is not deterministic because K -means is sensitive to its random initialization.

Complexity. $\mathcal{O}(nKdt)$, where t is the number of K -means iterations until the cluster algorithm converges (i.e., the cluster assignments stabilize).

A.2.6 Log-Determinant Diversity (LDD)

Intuition. This measure quantifies dataset diversity as the volume spanned by the sample embeddings in feature space: the determinant of a kernel matrix equals the squared volume of the parallelepiped spanned by the underlying feature vectors (Kulesza and Taskar, 2012), and this volume was proposed as a dataset diversity measure by Wang et al. (2025a). The more similar the embeddings are to one another, the smaller this volume. Higher values thus indicate greater diversity.

Computation. First, construct the kernel matrix $\mathbf{K}$ with elements $k_{ij} = \cos(\mathbf{e}_i, \mathbf{e}_j)$. Since cosine similarity is equivalent to the inner product between L2-normalized embeddings, this kernel matrix is

positive semidefinite. More generally, other positive semidefinite kernels can also be used. The measure is defined as:

$$\text{LDD}(\mathbf{X}) = \log \det(\mathbf{K} + \epsilon \mathbf{I}), \quad (16)$$

Since the cosine kernel matrix has rank at most $\min(d, n)$, $\mathbf{K}$ is singular whenever $n > d$, and its log-determinant is then $-\infty$; we therefore add a term $\epsilon \mathbf{I}$ (with $\epsilon = 10^{-6}$) to shift the eigenvalues by ϵ , following Yang et al. (2025).

Complexity. $\mathcal{O}(n^2d + n^3)$ (kernel construction, then Cholesky factorization)

A.3 Distribution-based Measures

This family reads global properties of how the embedded dataset is distributed—over the eigenvalue spectrum of a similarity matrix, over the bins of a discretized projection, or over the elements themselves. Most measures in this family are interpretable as an effective number of distinct elements or modes, several of them obtained through an entropy.

A.3.1 Vendi Score

Intuition. Inspired by ecology, Friedman and Deng (2023) propose this measure, which quantifies diversity as the effective number of unique elements in a dataset, derived from the entropy of the eigenvalue spectrum of a similarity matrix (Mironov and Prokhorenkova, 2025; Limbeck et al., 2024). It achieves its maximum value (n) when all samples are orthogonal (fully dissimilar) and its minimum value (1) when all samples are identical.

Computation. First, construct the similarity matrix $\mathbf{S} = [s_{ij}]$. Let $\tilde{\mathbf{S}} = \mathbf{S}/\text{tr}(\mathbf{S})$ be its normalized form. The eigenvalues $\lambda_1, \dots, \lambda_n$ of $\tilde{\mathbf{S}}$ are non-negative and sum to 1, so that they can be treated as a probability distribution and their entropy can be calculated. The Vendi Score is then defined as the exponential of the Shannon entropy of the eigenvalues:

$$\text{Vendi}(\mathbf{X}) = \exp\left(-\sum_{i=1}^n \lambda_i \log \lambda_i\right). \quad (17)$$

Complexity. Generally $\mathcal{O}(n^2d + n^3)$ for the eigenvalue decomposition of the $n \times n$ similarity matrix. However, when the embedding dimension d is much smaller than the sample size ($d \ll n$), it reduces the complexity to $\mathcal{O}(nd^2)$.

A.3.2 Rényi Kernel Entropy (RKE)

Intuition. This measure quantifies dataset diversity through the information-theoretic entropy of the similarity matrix in kernel space. Proposed by Jalali et al. (2023) and further theoretically analysed by Mironov and Prokhorenkova (2025), it estimates the effective number of modes or clusters in the data distribution by computing entropy of the eigenvalues of the normalized similarity matrix. More uniformly distributed samples yield higher entropy (greater diversity).

Computation. First, construct a normalized kernel matrix $\mathbf{S}$ with entries $S_{ij} = \frac{1}{n} s_{ij}$, where $[s_{ij}]$ is a positive semi-definite kernel with unit diagonal ($s_{ii} = 1$). Since the kernel is positive semi-definite with unit diagonal, $\text{tr}(\mathbf{S}) = 1$, so its eigenvalues $\lambda_1, \dots, \lambda_n$ are non-negative and sum to one, and can be treated as a probability distribution over spectral modes. The Rényi Kernel Entropy of order $\alpha > 0$ is defined as:

$$\text{RKE}_\alpha(\mathbf{X}) = \frac{1}{1 - \alpha} \log\left(\sum_{i=1}^n \lambda_i^\alpha\right). \quad (18)$$

The most common case is $\alpha = 2$, for which the score admits a closed form that avoids eigendecomposition:

$$\text{RKE}_2(\mathbf{X}) = -\log(\|\mathbf{S}\|_F^2) = -\log\left(\frac{1}{n^2} \sum_{i,j} s_{ij}^2\right). \quad (19)$$

where $\|\cdot\|_F$ is the Frobenius norm. Intuitively, low pairwise similarity (small $\|\mathbf{S}\|_F^2$) corresponds to a high RKE score, indicating a dataset with many distinct regions or modes.

Complexity. $\mathcal{O}(n^2d)$ for kernel construction and $\mathcal{O}(n^2)$ for Frobenius norm computation (for $\alpha = 2$). More generally, computing the full eigenvalue spectrum has complexity $\mathcal{O}(n^3)$.

A.3.3 DCScore

Intuition. This measure computes dataset diversity from a classification perspective, interpreting diversity as the degree to which each sample can be distinguished from others (Zhu et al., 2025). It views diversity evaluation as an n -class classification task, where each sample is treated as its own class. If samples are highly distinct, each is easily classified into its own category; if samples are similar, they are likely to be confused. A higher value indicates greater diversity.

Computation. First, construct a kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with entries

$$K[i, j] = \text{Kernel}(\mathbf{e}_i, \mathbf{e}_j),$$

using an inner-product or RBF kernel. Then apply a softmax-based classification function to obtain the class probability matrix $\mathbf{P}$:

$$P[i, j] = \frac{\exp(K[i, j]/\tau)}{\sum_{k=1}^n \exp(K[i, k]/\tau)}, \quad (20)$$

where τ is a temperature hyperparameter controlling classification resolution. Finally, the diversity score is computed as the trace of $\mathbf{P}$, i.e., the sum of probabilities that each sample is correctly classified into its own class:

$$\text{DCScore}(\mathbf{X}) = \text{tr}(\mathbf{P}) = \sum_{i=1}^n P[i, i]. \quad (21)$$

The score ranges from 1 (all samples identical) to n (all samples distinct). Smaller τ increases sensitivity to smaller differences, while larger τ smooths classification probabilities.

Complexity. $\mathcal{O}(n^2d + n^2)$ for general kernels (pairwise similarity + softmax summarization).

A.3.4 Bins Entropy

Intuition. This measure quantifies how evenly sample embeddings are distributed across a discretized low-dimensional projection of the embedding space (Cox et al., 2021b), by projecting the high-dimensional embeddings (e.g., with UMAP) into a 2D grid. Datasets whose samples occupy many bins with similar frequencies exhibit higher entropy, indicating greater diversity. Yang et al. (2025) apply the same idea of taking the Shannon entropy over bin counts, but discretize via K-means clustering rather than a grid over a projection. We follow Cox et al. (2021b) and use UMAP throughout.

Computation. Given embeddings $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ projected into a 2D space $\mathbf{Y} = \text{UMAP}(\mathbf{E})$, divide the projected space into a $B_x \times B_y$ grid of bins $\{b_1, \dots, b_B\}$, where $B = B_x B_y$. Let f_b denote the normalized frequency (proportion) of points in bin b . The Shannon entropy is then computed as:

$$\begin{aligned} \text{Entropy}_{\text{bins}}(\mathbf{X}) &= - \sum_{b=1}^B f_b \log f_b, \\ \text{where } f_b &= \frac{n_b}{n}, \sum_b f_b = 1. \end{aligned} \quad (22)$$

Optionally, normalize by $\log(B)$ to bound the measure between $[0, 1]$:

$$\text{Entropy}_{\text{norm}} = \frac{\text{Entropy}_{\text{bins}}}{\log(B)}. \quad (23)$$

Note that UMAP is a stochastic algorithm, so different runs can produce different results.

Complexity. $\mathcal{O}(n^{1.14}d + B)$ (Projection + histogram binning; d = embedding dimension, B = number of bins.)

A.3.5 MagArea (Metric Space Magnitude Area)

Intuition. This measure reads the *magnitude* of the metric space—a scalar that increases as samples become more mutually dissimilar—and reports it as an effective number of distinct samples (Limbeck et al., 2024). Unlike other measures in this family, magnitude is not computed at a single resolution: the similarity matrix is parameterized by a distance-scaling parameter τ , and MagArea aggregates the magnitude across a range of τ , so that diversity is captured jointly at fine and coarse resolutions. A higher value indicates greater diversity.

Computation. First, define the similarity kernel $\mathbf{Z}_\tau$ parameterized by scale $\tau > 0$:

$$Z_{ij}(\tau) = \exp(-\tau d(\mathbf{e}_i, \mathbf{e}_j)), \quad (24)$$

The *magnitude* of the metric space at scale τ is:

$$\text{Mag}(\tau) = \mathbf{1}^\top \mathbf{Z}_\tau^{-1} \mathbf{1}, \quad (25)$$

where $\mathbf{1}$ is the all-ones vector. The **MagArea** is defined as the integral of this magnitude function over a range of scales $T = [0, t_{\text{cut}}]$ (Limbeck et al., 2024, Def. 3.3):

$$\text{MagArea}(\mathbf{X}) = \int_0^{t_{\text{cut}}} \text{Mag}(\tau) d\tau. \quad (26)$$

The upper limit is necessary: as $\tau \rightarrow \infty$, $\mathbf{Z}_\tau \rightarrow \mathbf{I}$ and $\text{Mag}(\tau) \rightarrow n$, so the integral over $[0, \infty)$ diverges. Following Limbeck et al. (2024), we set t_{cut} to the median of the per-dataset convergence scales, where the convergence scale is the value of τ at which the magnitude function reaches 95% of n ; it is determined jointly across the five datasets of a sweep so that their magnitude functions are integrated over a common interval. In practice, the integral is approximated by evaluating $\text{Mag}(\tau)$ at $n_{ts} = 30$ evenly spaced scales in $[0, t_{\text{cut}}]$ and summing with trapezoidal integration; we report the

resulting area divided by t_{cut} , i.e. the mean of the magnitude function over the interval, which is the default behaviour of the reference implementation and does not affect the ordering within a sweep.

Complexity. $\mathcal{O}(n_{ts} n^3)$, where n_{ts} is the number of sampled τ values : at each scale, evaluating $\text{Mag}(\tau)$ requires a Cholesky factorization of the $n \times n$ similarity matrix $\mathbf{Z}_\tau$ followed by a triangular solve, rather than an explicit matrix inversion (Limbeck et al., 2024). The pairwise distance matrix is computed once and reused across scales.

A.4 Graph-based Measures

This family represents the dataset as a fully connected, weighted graph whose nodes are samples and whose edge weights are pairwise distances. Diversity is then measured from the structure of that graph.

A.4.1 MST Dispersion

Intuition. Originally adapted from phylogenetics and ecology, MST Dispersion measures diversity by constructing the Minimum Spanning Tree (MST) of the embedding graph (Cox et al., 2021b; Atwal et al., 2025).

Computation. First, construct a fully connected, weighted graph where each node corresponds to a sample and each edge weight is the pairwise distance:

$$d_{ij} = d(\mathbf{e}_i, \mathbf{e}_j).$$

Compute the Minimum Spanning Tree $\text{MST}(G)$ that connects all n nodes with the smallest possible total edge weight. The MST Dispersion is defined as the total weight of the MST’s edges:

$$\text{MSTDispersion}(\mathbf{X}) = \sum_{(i,j) \in \text{MST}(G)} d_{ij}. \quad (27)$$

Complexity. $\mathcal{O}(n^2 d)$ for pairwise distance computation and $\mathcal{O}(n^2 \log n)$ for MST construction (e.g., Kruskal’s algorithm).

A.4.2 Graph Entropy

Intuition. This measure takes dataset diversity as the uncertainty of pairwise distances between samples represented as a fully connected graph (Yu et al., 2022b). Each node corresponds to an embedding, and each edge weight reflects distance between two nodes. By computing the entropy of normalized edge weights, it captures how evenly pairwise distances are distributed across the

dataset. Higher entropy indicates a more uniformly distributed set of embeddings, suggesting higher dataset diversity.

Computation. First form a complete weighted graph where vertices correspond to samples and edge weights are cosine distances $d_{ij} = 1 - \cos(\mathbf{e}_i, \mathbf{e}_j)$. For each node x_i , normalize its pairwise distances to obtain a local distance distribution:

$$f_i(d_{ij}) = \frac{d_{ij}}{\sum_{k=1}^n d_{ik}}. \quad (28)$$

The local distance entropy of node x_i is then:

$$I(x_i) = - \sum_{j=1}^n f_i(d_{ij}) \log f_i(d_{ij}),^{12} \quad (29)$$

and the overall graph entropy of the dataset is obtained by aggregating across all nodes:

$$\text{GraphEntropy}(\mathbf{X}) = \sum_{i=1}^n I(x_i). \quad (30)$$

Note that since the measure sums $I(x_i)$ over all n nodes, Graph Entropy grows with n and is not comparable across datasets of different sizes.

We categorize it as a graph-based measure, following Yu et al. (2022b)’s description and naming. However, note that a complete graph on all embedding pairs is equivalent to a distance matrix: for each sample, the Shannon entropy is computed over its normalized distances to all others. These entropies are then summed across samples. This measure could therefore arguably also be categorized as a distribution-based measure.

Complexity. $\mathcal{O}(n^2 d)$ (pairwise distance computation and per-node normalization)

A.5 Surveyed but Not Benchmarked

We surveyed ten further embedding-based measures that we do not include in EMBDIVBENCH, either because they reduce to measures we already evaluate, because they are computationally intractable at our dataset size, or because they require an external reference that our controlled datasets do not provide: **KNN Distance** (Yang et al., 2025), **Vendi- q** (Pasarkar and Dieng, 2024), **Species(q)** (Leinster and Cobbold, 2012b), **Functional Diversity** (Karsdorp et al., 2022), **Density**

¹²The $j = i$ term vanishes since $d_{ii} = 0$ implies $f_i(d_{ii}) = 0$; we adopt the standard convention $0 \log 0 := 0$.

(Yang et al., 2025), the dataset-level **Determinantal Point Process** score (Wilhelm et al., 2018), **Hamiltonian Diversity** (Hu et al., 2024b), **#Circles** (Xie et al., 2023), **NovelSum** (Yang et al., 2025) and **Sem-Ent** (Han et al., 2022).

B Benchmark Construction Details

Distribution	Description
Uniform	Equal counts per category
Slightly skewed	Top 20% categories take 40% of points
Mildly skewed	Top 20% categories take 60% of points
Zipfian	Zipfian decay over categories
Strongly skewed	Top 1 category takes 50%; next 4 take 30%

Table 2: The five balance levels, ordered from most balanced (top) to least balanced (bottom).

Incremental construction. Along every component, moving from one level to the next *extends* the previous configuration rather than replacing it, so observed measure responses cannot be attributed to swapping categories between levels. Fig. 4 illustrates the three components with one example transition per component, using small values for visual clarity. Take **variety** as an example: the $k=2$ category set is a strict subset of the $k=5$ set, and the data points retained from each shared category are exactly the first 2 of those used at $k=2$ (since n is fixed and the per-category budget shrinks from 5 to 2). On the **balance** component, the category set is held constant across all conditions and only the allocation of data points changes. On the **disparity** component, the $m=1$ super-category set is a strict subset of the $m=3$ set, and within each retained super-category, the categories are selected by a fixed-order prefix—so changes in the $k=6$ category set between adjacent m levels are driven by super-category extension and the fixed k budget, not by random replacement.

B.1 Simulated Tier Configuration

We sweep five embedding dimensionalities $d \in \{256, 384, 1024, 2048, 4096\}$ to characterize dimension-dependent measure behaviour; results in the main body use $d=384$.

Hierarchical Gaussian mixture. Super-category centres are drawn from $\mathcal{N}(\mathbf{0}, \sigma_{\text{super}}^2 I_d)$ with $\sigma_{\text{super}} = 5\sqrt{d}$; category centres from $\mathcal{N}(s_{m(i)}, \sigma_{\text{category}}^2 I_d)$ with $\sigma_{\text{category}} = \sqrt{d}$; and each element of category i from $\mathcal{N}(c_i, \sigma_{\text{element}}^2 I_d)$

with $\sigma_{\text{element}} = 0.5$ (constant in d). The two larger scales grow with $\sqrt{d}$ because pairwise distances between isotropic Gaussian draws scale with $\sqrt{d}$; this keeps the relative separation of super-category centres, category centres, and intra-category clouds comparable across dimensionalities. The assignment of category i to a super-category is deterministic, $m(i) = i \bmod m$, so categories are distributed as evenly as possible across the m super-categories.

B.2 Natural Text Tier Configuration

Construction has two stages: choosing the domain/topic structure by farthest-point sampling (FPS) over label embeddings (§B.2.1), and collecting sentences for each chosen topic by adaptively crawling its Wikipedia category (§B.2.2). Tab. 3 shows examples of the resulting domain/topic structure.

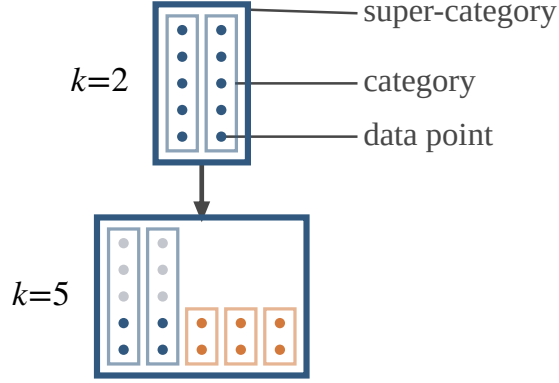
Topic selection. We build the candidate pool from a one-time crawl of Wikipedia’s category hierarchy, accessed through the live MediaWiki API (via the wikipedia-api Python package¹³) on 24 April 2026. The top level (“Life”, “Concepts”, “People”, etc.) is excluded as too broad to be a meaningful categorical unit. Domains (e.g. *Humanities*, *Engineering disciplines*) correspond to super-categories in the benchmark’s notation, and topics within those domains (e.g. *Anthropology*, *Materials science*) correspond to categories. A topic is retained in the pool only if its Wikipedia category has at least 30 direct article pages, which targets topics with enough material to support the largest per-topic sentence count we need. After deduplicating topic labels (a small fraction of topics appear under multiple domains; we keep the one with the most direct pages), the candidate pool contains 191 domains, each with at least 5 suitable topics. Because the live API returns article state at request time, re-running the construction pipeline at a later date will yield a similar but not identical pool as Wikipedia content evolves; the released benchmark archive contains the exact sentences used in our experiments, so reproducing the results in this paper does not require re-crawling.

B.2.1 Farthest-Point Sampling (FPS) of Domains and Topics

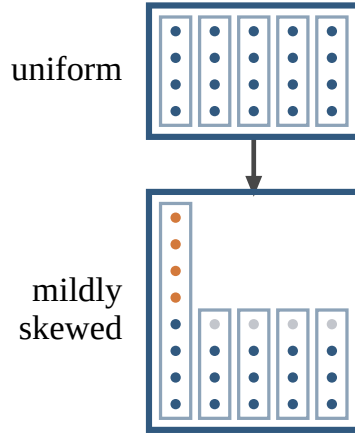
Some domains in the candidate pool can be semantically related (e.g., several concern history or

¹³<https://pypi.org/project/Wikipedia-API/>

(a) **Variety:** $k=2 \rightarrow 5$ ($m=1$ fixed)



(b) **Balance:** uniform $\rightarrow$ mildly skewed ($k=5$ fixed, $m=1$ fixed)



(c) **Disparity:** $m=1 \rightarrow 3$ ($k=6$ fixed)

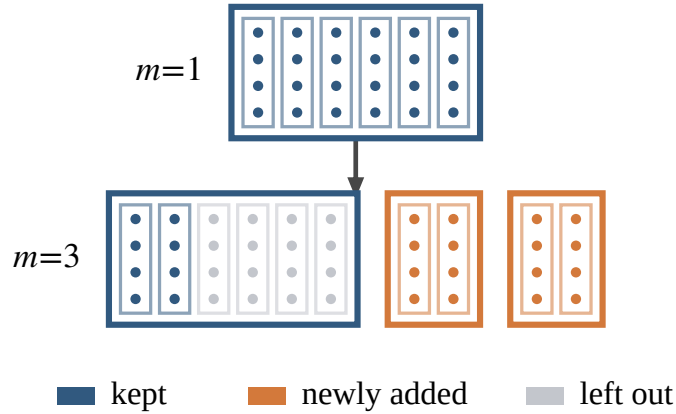


Figure 4: Incremental construction across the three components, illustrated with small example values. Each dataset is a hierarchy of *super-categories* (thick outer box) containing *categories* (thin inner column) containing *data points* (dot). The total dataset size n is conserved across levels in every panel. **Variety** ($k=2 \rightarrow 5$, with $m=1$ fixed): each kept category's data points are truncated to a prefix, and 3 new categories are added. **Balance** (uniform $\rightarrow$ mildly skewed; $k=5$ and $m=1$ both fixed): the category set is held constant; data points are redistributed from low-frequency to high-frequency categories. **Disparity** ($m=1 \rightarrow 3$, with $k=6$ fixed): the kept super-category retains a fixed-order prefix of 2 of its 6 categories (the other 4 are left out), and 2 new super-categories are added. Blue marks elements kept from the lower diversity level, orange marks elements newly added, and gray marks elements that were in the lower level but are left out at the higher level; this colour coding applies at all three levels of the hierarchy.

Domain	Topic	Example sentence
Government by country	National legislatures	A general election is an electoral process to choose most or all members of a governing body at the same time. After being elected the Prime Minister appoints the cabinet ministers and announces them to the Riksdag.
	Parliaments by country	Article 16.4.2° requires that the newly elected Dáil Éireann must convene no later than 30 days after the election.
Humanities	Anthropology	A horse culture is a tribal group or community whose day-to-day life revolves around the herding and breeding of horses.
Engineering disciplines	Materials science	Air and vacuum have high reluctance, while easily magnetized materials such as soft iron have low reluctance.

Table 3: Examples from the Wikipedia Domains/Topics hierarchy. Sentences within the same topic share subject matter; sentences within the same domain but different topics are related but distinguishable; sentences from different domains are semantically distant.

politics). Selecting domains uniformly at random would risk drawing a set concentrated in one region of the label-embedding space, which would undermine the structural assumption behind the natural text tier: that different domains are semantically distant while topics within a domain are related but distinguishable. We therefore use farthest-point sampling (FPS), a greedy procedure that starts from one domain and repeatedly adds the candidate farthest from the already-selected set, thereby spreading the selection across the label-embedding space.

Procedure. Domain and topic labels are encoded once with Qwen3-Embedding-8B and L2-normalized; because the embeddings are unit-norm, Euclidean FPS is monotonically equivalent to cosine FPS. We apply FPS first over domain labels and then, within each chosen domain, apply the same procedure over its topic labels. Given its starting label, FPS is deterministic; the only source of randomness per run is therefore the choice of the starting label, which is drawn under the run seed. The full domain/topic selection is thus reproducible from the run seed alone.

Two selection settings. Variety and balance both hold the domain structure fixed at $m = 10$ and use at most $k = 50$ topics, so they share a single selection of 10 domains with 5 topics each. Disparity instead varies m from 10 to 50 at fixed $k = 50$ and thus requires up to 50 domains; we run a separate domain FPS for it with different seeds and, at each level, select the required number of topics per domain (one or more). Both selections use the same FPS procedure described below.

B.2.2 Sentence Pool Construction

Adaptive expansion schedule. For each chosen topic, we build a sentence pool by crawling its Wikipedia category in up to 5 rounds with linearly increasing parameters: descent depth $1 \rightarrow 3$, articles inspected per round $4,000 \rightarrow 20,000$, and per-article sentence cap $5 \rightarrow 40$. A round only fires when the pool has not yet reached the required size; in the typical case the first round (depth 1, cap 5) suffices, and later rounds are reserved for sparser topics. This keeps the dominant sampling regime close to the “shallow, low cap” setting (i.e., the shallow descent avoids semantic drift from deep subcategory descent and the low per-article cap prevents any single article dominating a topic’s pool), while still allowing thin topics to reach their target pool size.

Crawl mechanics. Within a round, articles are collected breadth-first from the root category: members of the current category are inspected in shuffled order, and we descend into a subcategory only when the current depth has not been exhausted and the article budget has not yet been hit. For each article we extract its full body text, split it into sentences with a simple end-punctuation regex, shuffle the sentences, and add them to the pool subject to the length filter (100–300 characters) and the per-article cap. Sampling sentences in shuffled order rather than as a leading prefix avoids overrepresenting the stylized opening sentences characteristic of Wikipedia articles.

B.3 Hyperparameter Settings

We list the hyperparameter settings for all 21 measures in Tab. 4.

C Topic Label Noise Audit for the Natural Text Tier

The natural text tier uses Wikipedia topic labels as a proxy for sentence-level semantics (§4.2.2), so not every sentence taken from an article is necessarily a good representation of the article’s topic. We therefore manually audit the Wikipedia topic labels.

Setup. We draw a stratified sample of 10 sentences from every dataset of the three components: for variety, the datasets with $k \in \{10, 20, 30, 40, 50\}$ topics ($m = 10$ fixed); for balance, the five distribution conditions ($k = 50$, $m = 10$ both fixed); and for disparity, the datasets with $m \in \{10, 20, 30, 40, 50\}$ domains ($k = 50$ fixed), each across all 5 seeds, yielding $3 \times 5 \times 5 \times 10 = 750$ sentences. One of the authors checks each sentence against the full topic pool of its source dataset and annotates it with a binary label: MATCH if the sentence supports its assigned Wikipedia topic, MISMATCH otherwise. The noise rate is the proportion of MISMATCH over all 750 annotated sentences.

Judgement criteria. A sentence is a MISMATCH when it provides no usable evidence for its assigned topic: it is about a different topic in the pool, it is too generic to point anywhere (e.g., a transition or background sentence), or the sentence splitter returned an incomplete snippet. For the MATCH, apart from exact matches, two other cases are also counted as MATCH. First, a sentence that is about its assigned topic but would also fit a neighbouring topic in the pool is still a match, because it sits in the right region of the embedding space, and the levels of each component are separated by wide intervals (k and m both move in steps of 10), so an ambiguous placement does not damage the structure of our designed benchmark. Second, a sentence that is about its topic only if one knows some external fact is also a MATCH, because embedding models are pre-trained on broad corpora and have access to exactly the world knowledge human annotators may lack. This rule is, however, easier to state than to apply (see below *Annotation difficulty*). Tab. 5 illustrates both labels with examples.

Overall noise level. Of the 750 sentences, 616 match their assigned topic and 134 are mismatched, giving a noise rate of 17.9% (134/750; Wilson 95% CI [15.3%, 20.8%]). The three components give

Measure	Hyperparameters
<i>Distance-based</i>	
Mean Pairwise Dist.	—
Sum Pairwise Dist.	—
Energy	$\gamma = 1.0$, $\epsilon = 10^{-12}$
Diameter	—
Bottleneck	—
Sum Diameter	no division by n
Sum Bottleneck	no division by n
Chamfer Dist.	—
<i>Geometry-based</i>	
Convex Hull Vol. (2D)	2D UMAP projection, random_state = 42
Span (Centroid)	90th percentile
Span (Medoid)	—
Cluster Inertia	$K = 200$
Geo. Mean of Std.	no distance metric
Log-Determinant	cosine kernel, $\tau = 1.0$, normalized, $\epsilon = 10^{-6}$, Cholesky
<i>Distribution-based</i>	
Vendi Score	$q = 1.0$, normalized, dual form
Rényi Kernel Ent.	$\alpha = 2.0$, cosine kernel, $\tau = 1.0$, normalized, $\epsilon = 10^{-12}$
DCScore	cosine kernel, $\tau = 1.0$, normalized
Bins Entropy (UMAP)	5×5 grid, UMAP projection, effective normalization
MagArea	$n_{ts} = 30$ scales
<i>Graph-based</i>	
MST Dispersion	—
Graph Entropy	—

Table 4: Hyperparameter settings for all 21 measures. “—” means the measure takes no further parameters. Measures differ in whether they operate on cosine distances or on a cosine-similarity kernel. More details can be found in our repository at <https://github.com/nlpso/EmbDivBench>

Sentence	Assigned topic	Comment
MATCH — exact matches		
He also contributed a foreword to white supremacist commentator Stefan Molyneux’s book <i>Against the Gods</i> .	New Atheism	The topic is explicit in the sentence itself.
MATCH — neighbouring topics		
In German, for example, the pentagraph <i>tzsch</i> represents the [tʃ] sound of the English digraph <i>ch</i> , and indeed is found in the English word <i>Nietzschean</i> .	Orthography	<i>Orthography</i> and <i>Linguistic terminology</i> both make sense, but they lie close enough in the embedding space.
MATCH — requires background knowledge		
Quickly was founded in California in 1996 and started its trademark licensing program at the same time.	Tea culture	Resolvable only by recognizing Quickly as a Taiwanese bubble tea chain.
George married Edith Jane Stewart Bayley (died in England of a painful internal cancer 8 June 1876) on 18 December 1854.	Government of Australia	Requires knowing the named subject is an Australian politician; reads as a generic biographical note otherwise.
MISMATCH — different topic		
In 1938 he broke the Turkish national record, with a time of 49.0 that was not surpassed for seven years, while winning the 1938 Balkan Games 400 m title in Belgrade.	Jewish emigration from Nazi Germany	The article covers the subject’s emigration, but this sentence reads as another topic in the pool <i>Sports records and statistics</i> .
Meats cooked at high temperatures, especially above 300 °F (149 °C) (as in grilling or pan frying), or that are cooked for a long time tend to form more HCAs.	Animal products	Describes preparation rather than the product itself; closer to another topic <i>Cooking techniques</i> .
MISMATCH — no topical cue		
Gerhard Fisher studied electronics at the University of Dresden before emigrating to the United States.	Prisons	Biographical background; carries no cue to any topic in the pool.
He holds a bachelor’s degree from California State University, Long Beach, and a Master of Fine Arts (MFA) from California State University, Fullerton.	Medical journal editors	The referent of <i>he</i> is only recoverable from the surrounding context.
MISMATCH — incomplete snippet		
Cabinet Ministers Didier Budimbu - 2022-present Aime Ngoy Mukena - 2015–2022 Crispin Atama Tabe - 2012–2015 [...] == References ==	Government of Iran	Not a sentence: a list of names and a section heading picked up by the sentence splitter.

Table 5: Illustrative examples from the manual audit, grouped by the reason for each label. Examples are selected for illustration and are not proportional to the frequency of each case.

comparable rates (variety 20.4%, balance 15.2%, disparity 18.0%), and within each component the rate does not vary substantially across levels (Spearman $\rho = 0.46$, $p = 0.43$ for variety; $\rho = 0.20$, $p = 0.75$ for balance; $\rho = -0.11$, $p = 0.87$ for disparity). Label noise therefore affects the five level datasets of a component to a comparable degree rather than in a growing way, so it degrades all five datasets alike and does not act as a confound for the benchmark design. Table 6 breaks the counts down by component and level.

Annotation difficulty. Assigning a sentence against up to 50 candidate topics is a non-trivial task, and annotator-dependent: even presented with the same sentence and the same topic pool, different annotator may perceive different topics as the closest fit, particularly where several candidates are semantically adjacent. The judgement also depends on the annotator’s background knowledge rather than on the sentence alone. For instance, “*Quickly was founded in California in 1996 and started its trademark licensing program at the same time.*” points clearly to tea culture for a reader who recognizes Quickly as the Taiwanese bubble tea chain brand, and nowhere at all otherwise; “*GMR-2 is not an evolution of GMR-1 but rather a concurrent standard that has been developed by another group of companies.*” is clearly about satellite communication to an engineer. Where the required knowledge is missing, an annotator is more likely to judge a sentence MISMATCH than to recognize its link to the assigned topic (see the knowledge-dependent rows in Tab. 5).

D Results for Robustness Check

We elaborate on results for the robustness check. We display results on the natural text tier for *Qwen3-Embedding-8B*¹⁴ and *e5-mistral-7b-instruct*¹⁵ compared to *MiniLM* in Fig. 6 and results for *Style Embedding*¹⁶ compared to *MiniLM* in Fig. 7.

Qwen3 improves correlations and creates the same family ranking. Several findings with *MiniLM* hold with *Qwen3* and *E5*, a high perform-

Level	MATCH	MISMATCH	Noise
<i>Variety</i>			
$k = 10$	44	6	0.12
$k = 20$	40	10	0.20
$k = 30$	36	14	0.28
$k = 40$	39	11	0.22
$k = 50$	40	10	0.20
<i>Balance</i>			
Uniform	41	9	0.18
Slightly skewed	46	4	0.08
Mildly skewed	43	7	0.14
Zipfian	40	10	0.20
Strongly skewed	42	8	0.16
<i>Disparity</i>			
$m = 10$	39	11	0.22
$m = 20$	42	8	0.16
$m = 30$	42	8	0.16
$m = 40$	41	9	0.18
$m = 50$	41	9	0.18

Table 6: Per-level annotation counts (50 sentences per level) and noise rate (MISMATCH over assignable sentences) for the three components of the natural text tier.

ing embedding model on MTEB¹⁷ (Enevoldsen et al., 2025): no measure performs perfectly on all components, the family ranking (Graph > Distribution > Geometry > Distance) is consistent on all three embedding models, and disparity remains the hardest component. Further, most measures reach higher correlation with Qwen3 embeddings, with a few decreasing correlation on disparity.

E Detailed rationale for recommendations

Graph Entropy has the highest aggregated performance scores across the three components, so if no single component is of primary interest, it is currently the best-supported single choice. As a complement, we also suggest Vendi ($q=1$): of the three measures responding perfectly to all three components at every simulated dimension we test (Fig. 5), Vendi ranks second on natural text averaged over the three semantic embedding models, behind Graph Entropy. It is also considerably cheaper to compute than Graph Entropy (Tab. 7).

Disparity is the hardest component for all measures (§5), and even the recommended MST Dispersion is relatively weak on disparity compared to our variety and balance recommendations. Our disparity recommendation should therefore be read

¹⁴<https://huggingface.co/Qwen/Qwen3-Embedding-8B>

¹⁵<https://huggingface.co/intfloat/e5-mistral-7b-instruct>

¹⁶<https://huggingface.co/AnnaWegmann/Style-Embedding>

¹⁷Note that MTEB does not only include semantic tasks, but also application-oriented tasks like clustering or RAG, which, while they do profit from semantic information—might also have additional encoding requirements.

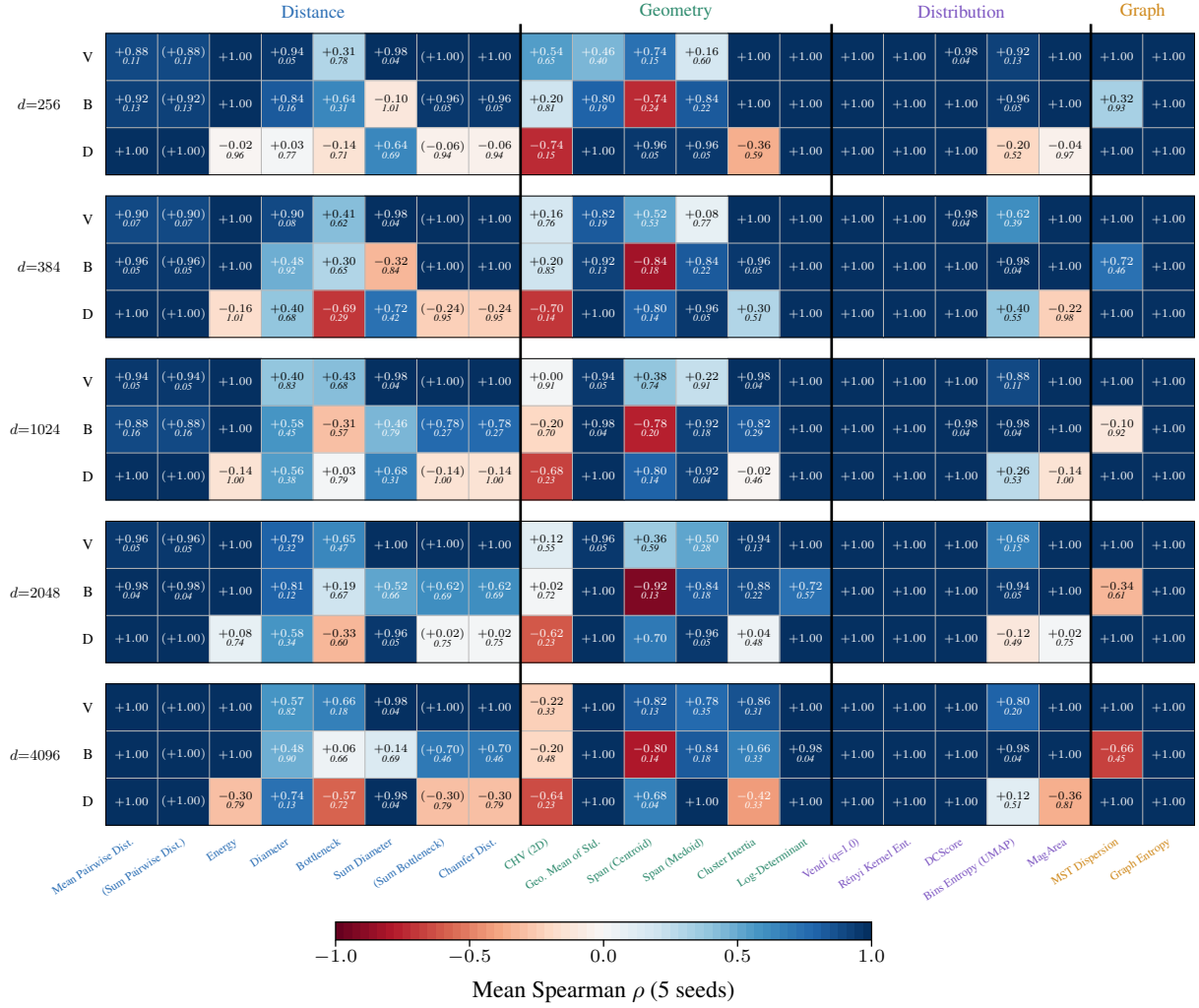


Figure 5: Per-component mean Spearman ρ (5 seeds) on the **simulated data (GMM)** tier across embedding dimensions $d \in \{256, 384, 1024, 2048, 4096\}$ for the Variety (V), Balance (B) and Disparity (D) components; the smaller italic number below each ρ is the across-seed standard deviation, shown when it exceeds 0.01. Columns are grouped and coloured by the four measure families (Distance-, Geometry-, Distribution-, Graph-based). As in Fig. 2, the two rank-equivalent measures are given in parentheses (Sum Pairwise Dist., Sum Bottleneck).

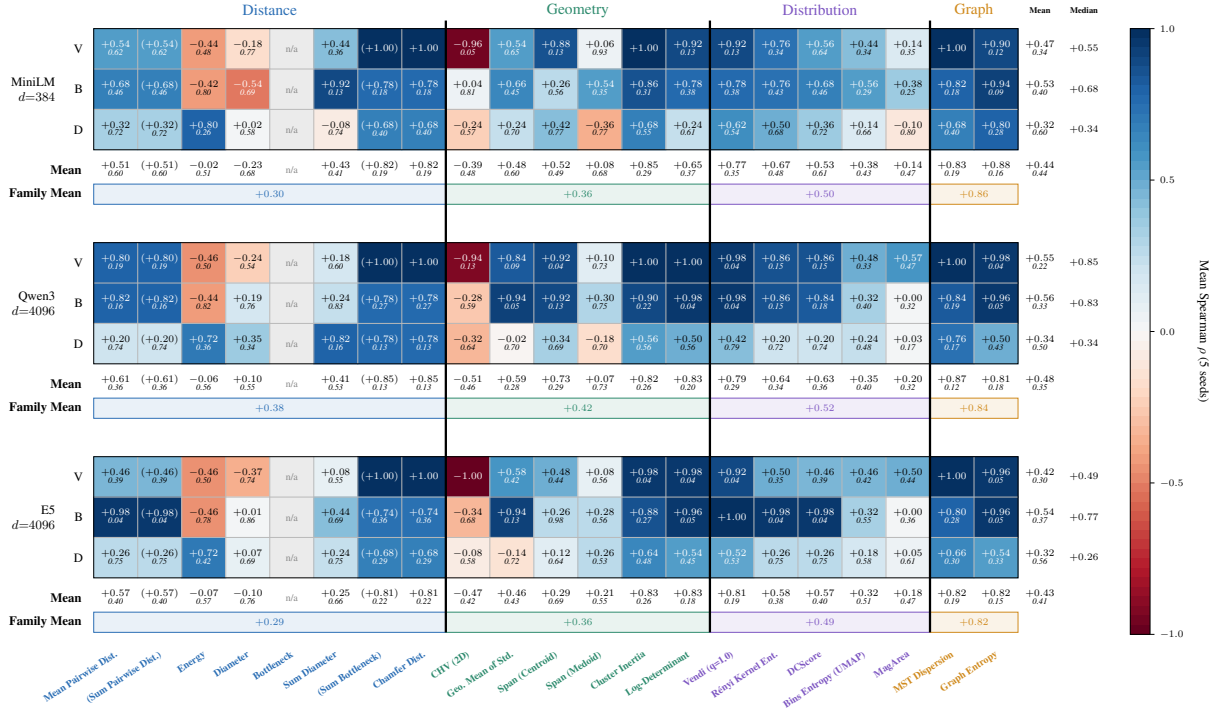


Figure 6: Per-component mean Spearman ρ (5 seeds) on the **natural text (Wikipedia)** tier for three sentence encoders: *all-MiniLM-L6-v2* ($d=384$, top), *Qwen3-Embedding-8B* ($d=4096$, middle), and *E5* ($d=4096$, bottom), along the Variety (V), Balance (B) and Disparity (D) components; the smaller italic number below each ρ is the across-seed standard deviation, shown when it exceeds 0.01. Columns are grouped and coloured by the four measure families (Distance-, Geometry-, Distribution-, Graph-based). The two rightmost columns per panel give the mean ρ and the median ρ over the distinct measures; the row beneath each panel gives the per-measure mean ρ over components, and the coloured strip below it gives the per-family mean ρ , averaged over the distinct measures of that family and the three components. Note that Bottleneck collapses to zero across all five levels for nearly every seed, making Spearman ρ undefined (*n/a*); it is excluded from the Distance average. As in Fig. 2, the two rank-equivalent measures are given in parentheses (Sum Pairwise Dist., Sum Bottleneck) and are counted once in the aggregates.

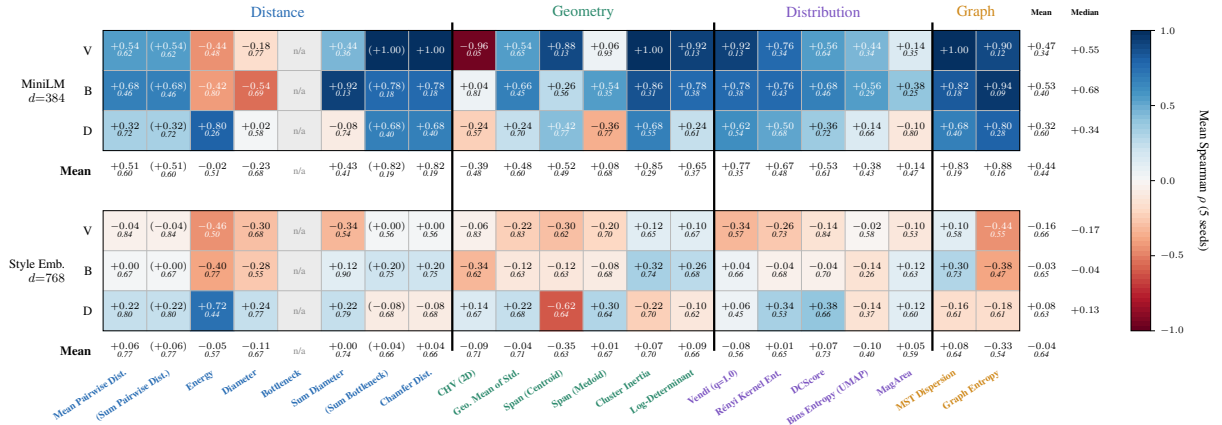


Figure 7: Per-component mean Spearman ρ (5 seeds) on the **natural text (Wikipedia)** tier for two sentence encoders: *all-MiniLM-L6-v2* ($d=384$, top) and the *Style Embedding* model of Wegmann et al. (2022) ($d=768$, bottom), along the Variety (V), Balance (B) and Disparity (D) components; the smaller italic number below each ρ is the across-seed standard deviation. Columns are grouped and coloured by the four measure families (Distance-, Geometry-, Distribution-, Graph-based). The two rightmost columns per panel give the mean ρ and the median ρ over the distinct measures; the row beneath each panel gives the per-measure mean ρ over components. Bottleneck collapses to zero across all five levels for nearly every seed, making Spearman ρ undefined (*n/a*). As in Fig. 2, the two rank-equivalent measures are given in parentheses (Sum Pairwise Dist., Sum Bottleneck) and are counted once in the aggregates.

as the best currently available option rather than a reliable one, and we see measuring disparity as a pressing open problem for future work.

F Statistics For Data

Each dataset contains $n=10,000$ elements. The simulated tier sweeps 5 embedding dimensionalities, with 5 seeds and 15 configurations (5 levels for each of the three components) per dimensionality, giving $5 \times 5 \times 15 = 375$ datasets. The natural-text tier uses 5 seeds and the same 15 configurations, giving $5 \times 15 = 75$ datasets. In total, the benchmark comprises 450 datasets.

G Computational Cost and Budget

Embedding work—Qwen3-Embedding-8B for natural text data creation (§B.2.1) and all-MiniLM-L6-v2 for sentence encoding for measure evaluation—was run on a single A100 GPU (79 GB MIG slice). All other stages, namely simulated dataset generation, the Wikipedia category crawl, and diversity-measure evaluation, ran on CPU with 4–6 cores and up to 64 GB of RAM. A full pass of measure evaluation over all 75 natural-text datasets ($n=10,000$, $d=384$) takes approximately 25 CPU-hours; as Tab. 7 shows, per-measure cost spans more than four orders of magnitude, with MagArea alone accounting for $\sim 87\%$ of the total (~ 21.8 h) and the remaining 20 measures together completing in ~ 3.4 h.

H Intended Use and Licenses

We document the licences and intended use of the models and datasets re-used and created in this work. All third-party resources are used in a manner consistent with their licences.

Sentence encoders. We use all-MiniLM-L6-v2 (22M parameters) (Wang et al., 2020; Reimers and Gurevych, 2019) as the primary sentence encoder for the two tiers; it is released under the Apache 2.0 licence.¹⁸ To assess robustness across encoders of different scale, model family, and training objective, we additionally use three encoders. Qwen3-Embedding-8B (8B parameters) (Zhang et al., 2025b), an LLM-based encoder, is released by Alibaba Cloud under the Apache 2.0 licence.¹⁹ e5-mistral-7b-instruct (7B param-

eters) (Wang et al., 2024), is released under the MIT licence.²⁰ AnnaWegmann/Style-Embedding (125M parameters) (Wegmann et al., 2022), a RoBERTa-based encoder trained for content-independent style representation rather than semantic similarity, is released under the MIT licence.²¹ All four models are used strictly for sentence embedding extraction, consistent with their intended use.

Crawling library. The Wikipedia crawler (§B.2) uses the Wikipedia-API Python package²², released under the MIT licence, for category traversal and article retrieval via the live MediaWiki API. We use it strictly for read-only access to publicly available Wikipedia content, consistent with its intended use.

Wikipedia. The natural text tier (§4.2.2) is constructed by sampling sentences from English Wikipedia articles organized under the categorical hierarchy. Wikipedia text is dual-licensed under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) and the GNU Free Documentation License (GFDL).²³ Re-use for academic research and benchmark construction is permitted; under the share-alike provision, any redistributed Wikipedia-derived corpus must be released under the same licence.

Simulated Data. The synthetic datasets used in the simulated tier (§4.2.1) are generated by our own sampling code from hierarchical Gaussian mixtures. They contain no text, no third-party content, and no personally identifiable information. They are released together with the generation scripts under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence.

EMBDIVBENCH Release. The benchmark—including dataset construction scripts, pre-computed embeddings, the code to call the functions for different measures—is publicly released to support reproducibility and further research on embedding-based diversity measurement. The curated Wikipedia data is released under CC BY-SA 4.0, inheriting the licence of the underlying Wikipedia text; The simulated datasets

¹⁸<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

¹⁹<https://huggingface.co/Qwen/Qwen3-Embedding-8B>

²⁰<https://huggingface.co/intfloat/e5-mistral-7b-instruct>

²¹<https://huggingface.co/AnnaWegmann/Style-Embedding>

²²<https://pypi.org/project/Wikipedia-API/>

²³<https://en.wikipedia.org/wiki/Wikipedia:Copyrights>

Table 7: Compute cost and asymptotic complexity of measures. Relative cost is reported as a multiple of mean pairwise distance on identical inputs ($n=10000$, $d=384$); within each family, rows are ordered by ascending cost.

Family	Measure	Cost	Complexity
Distance	Mean Pairwise Dist.	$1.00\times$	$O(n^2 d)$
	Sum Pairwise Dist.	$1.00\times$	$O(n^2 d)$
	Energy	$1.02\times$	$O(n^2 d)$
	● Chamfer Dist.	$1.05\times$	$O(n^2 d)$
	Diameter	$0.99\times$	$O(n^2 d)$
	Bottleneck	$0.99\times$	$O(n^2 d)$
	Sum Diameter	$1.03\times$	$O(n^2 d)$
	Sum Bottleneck	$1.03\times$	$O(n^2 d)$
Geometry	Convex Hull Vol. (2D) [†]	$< 0.01\times$	$O(n^{1.14} d + n \log n)$
	Span (Centroid)	$0.01\times$	$O(nd)$
	Geo. Mean of Std.	$0.01\times$	$O(nd)$
	Log-Determinant	$0.55\times$	$O(n^3 + n^2 d)$
	Span (Medoid)	$1.03\times$	$O(n^2 d)$
	Cluster Inertia [‡]	$1.17\times$	$O(nKd \cdot \text{iters})$
Distribution	Vendi Score	$0.03\times$	$O(nd^2 + d^3)$
	DCScore	$0.16\times$	$O(n^2 d)$
	Rényi Kernel Ent.	$0.20\times$	$O(n^2 d)$
	Bins Entropy (UMAP) [†]	$1.93\times$	$O(n^{1.14} d)$
	MagArea	$108.8\times$	$O(n_{ts} \cdot n^3)$
Graph	● Graph Entropy	$1.13\times$	$O(n^2 d)$
	●● MST Dispersion	$3.49\times$	$O(n^2 d)$

Cost relative to MPD (≈ 9.6 s wall-time). [†] relies on a 2D UMAP projection, not directly comparable. [‡] Cluster inertia uses KMeans with $K=200$ clusters and default sklearn convergence ($\text{iters} \leq 300$). Coloured dots mark the recommendations for different components: blue for variety, green for balance and red for disparity.

and source code are released under CC BY 4.0. The intended use of EMBDIVBENCH is the evaluation of embedding-based diversity measures within NLP/ML research; we do not anticipate use beyond this scope.

I Use of AI Assistants

We used Claude Code and GitHub Copilot for coding tasks—primarily command lookup and occasional function generation, with all generated code reviewed and tested before use. We also used Claude for rephrasing and grammatical correction during writing.