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

Anonymous ACL submission

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

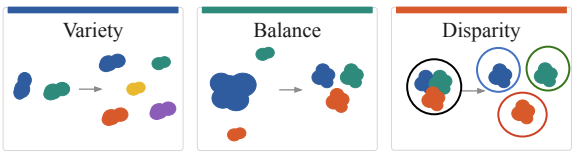


Figure 1: The three components of diversity in EMBDIVBENCH. In each panel, the left configuration has lower diversity and the right has higher diversity along the corresponding component. **Variety** counts the number of groups present; **Balance** measures how evenly points are distributed across groups;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

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)—variety (how many "categories"—cf. §4.1—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).

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: researchers can measure diversity along many dimensions, as long as texts can be mapped to an embedding space capturing the relevant diversity dimension. Although some studies have compared different 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. We address both gaps; specifically, we make the following **contributions**:

- A **taxonomy** (§3) that surveys and classifies **21**[1] embedding-based diversity measures into an organized taxonomy.
- **EMBDIVBENCH** (§4), a benchmark that tests measures on the variety, balance and disparity component 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, with recommendations for each component.
- Data and code are publicly available at `https://anonymous.4open.science/r/EmbDivBench-4963/` and can be used to compare new embedding-based diversity measures.

[1]Our full survey in Appendix A covers 26 measures; five are excluded from the benchmark for reasons discussed there.

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

2 Related Work

Diversity measurement beyond embeddings. Diversity has been studied across many disciplines—in ecology, economics, and the social sciences—and Shannon Entropy is a popular approach in many of these areas (Leinster and Cobbold, 2012; Nehring and Puppe, 2002). Within NLP, the dominant non-embedding measures are 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 these measures are useful, they remain relatively simple and only capture surface variation. In this work, we focus on the complementary—embedding-based diversity measures—an increasingly promising and 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 Cosine Distance has been used the 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).

Embedding-based measures have also been proposed and used in other fields, such as Hamiltonian Diversity in chemistry (Hu et al., 2024; Du et al., 2022), Energy over graph-distance embeddings in graph generation (Velikonivtsev et al., 2024), and

Rényi Kernel Entropy in computer vision (Jalali et al., 2023). Such measures could potentially also be applied to text. We include such measures in our taxonomy and test them on our benchmark. We give an overview and definitions of all included measures in App. §A.

Evaluation of diversity measures. As diversity measurement itself has received limited systematic evaluation, 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 lower or higher diversity, 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).

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 the measurement of diversity in NLP and apply Stirling's variety, balance and disparity (V-B-D) decomposition to classify existing measures. However, no prior work has empirically evaluated whether existing diversity measures correctly reflect shifts in variety, balance, and disparity; our benchmark EMBDIVBENCH fills this gap (§4).

3 Unified Taxonomy of Diversity Measures

We describe the current fragmented landscape of 21 embedding-based diversity measures we survey (§3.1) and introduce a unifying taxonomy (§3.2).

3.1 A Fragmented Landscape

Diversity is a term that is not well-defined in NLP (§1). Moreover, the literature on embedding-based diversity measurement is strikingly inconsistent in how it defines and uses measures. We collect a total of 21 different embedding-based diversity measures (cf. Tab. 1). Further, we 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. (2022), *DistSum* by Yang et al. (2025) and *Remote-Clique* (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 *Isocontour 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. (2021).

Similar objectives, different realisations. Beyond name-computation collisions, several measures aim to capture the same concept but operationalize it differently. For example, both Cox et al. (2021) and Yang et al. (2025) partition the embedding space into discrete bins and apply Shannon entropy to bin counts. However, Cox et al. (2021) 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 or similarities into a single score, without first creating additional structures such as clusters or graphs. An example is the frequently used Average Pairwise Similarity (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 spacial properties of the embeddings. For example, the Convex Hull Volume measures the volume of the smallest convex region enclosing all embeddings (Yu et al., 2022) 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 normalised 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., 2021).

Some measures consist of steps that fall into different families. We resolve such cases based on their dominant step. 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.

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—instantiated 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.

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

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

Concretely, for each component (cf. Fig. 1) we build a sequence of five datasets 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:

- n: the total number of element[2] 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.

[2]Note that generally for the variety, balance and disparity components, "elements" are needed (Estève et al., 2025). For our case, elements just correspond to the vectors or texts in the considered dataset.

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 (van Dam, 2019). In NLP, this concept of diversity has only recently begun to be adopted (Estève et al., 2025; Nguyen and Ploeger, 2025).

Our benchmark design adopts the variety-balance-disparity decomposition directly. 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 (cf. 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). For each component, we build a sequence of five datasets that vary one component while holding the other two fixed.

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 distribution conditions, ranging from uniform to strongly skewed, while holding variety ($k = 50$) and disparity ($m = 10$) fixed. The strongly skewed condition assigns 50% of data points to one category, 30% to another, and the remaining 20% to all other categories. 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 §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, class 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 $\mathbf{s}_1, \ldots, \mathbf{s}_m$ from a Gaussion $\mathcal{N}(\mathbf{0}, \sigma^2_{\text{super}} I)$. 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 $\mathbf{c}_i \sim \mathcal{N}(s_{m(i)}, \sigma^2_{\text{category}} I)$, where $m(i)$ denotes the super centre assigned to category i. Then, each element of category i is drawn from $\mathcal{N}(\mathbf{c}_i, \sigma^2_{\text{element}} I)$. 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 §B.1.

As a result of the synthetic construction, each component of diversity can be manipulated inde-

pendently: 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 because 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 and this makes it the tier most directly relevant to downstream use. App. Tab. 3 illustrates its structure with examples. 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 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.

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-8B[3] 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.

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-v2[4] ($d = 384$) (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.

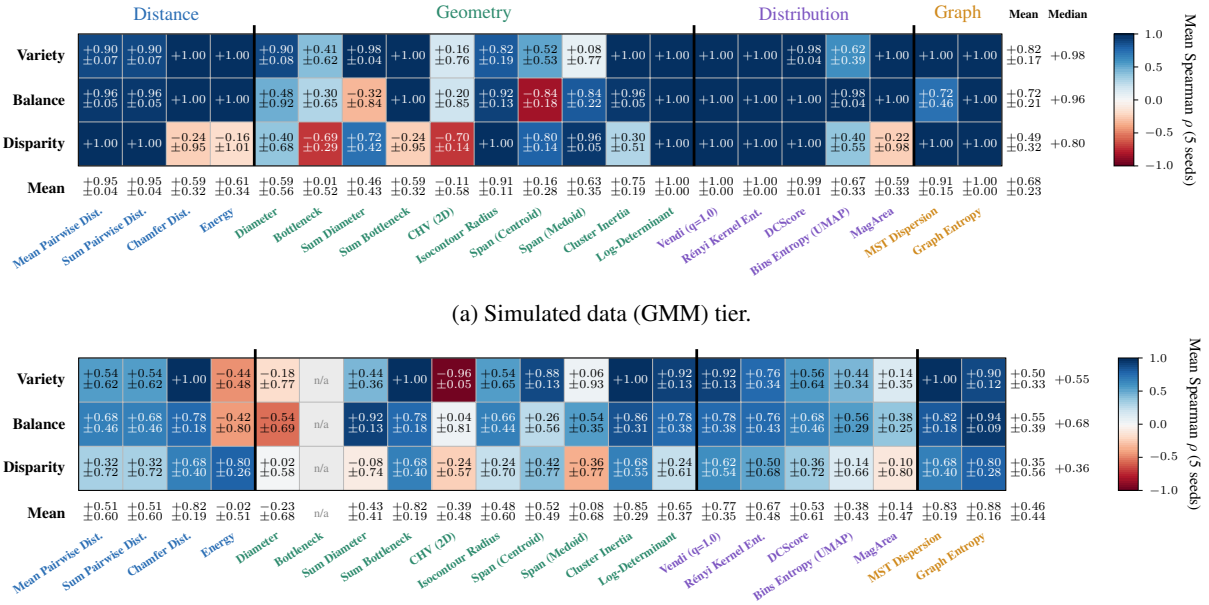
5.1 Analysis and Discussion

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 baseline; however, even on this easy and controlled setting, it does not perform perfectly (mean $\rho < 1$ on both variety and balance). On natural text, the situation is worse: its correlation drops to roughly half its value on the simulated data tier on average. 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, measures

[3] https://huggingface.co/Qwen/Qwen3-Embedding-8B

[4] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2



(a) Simulated data (GMM) tier.

(b) Natural text (Wikipedia) tier.

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. Columns are grouped and coloured by the four families (Distance-, Geometry-, Distribution-, Graph-based). The two rightmost columns give, over measures, the mean ρ ($\pm$ mean std) and the median ρ; the bottom row gives the per-measure mean ρ ($\pm$ mean std).

generally perform well, but a few perform "catastrophically" (high negative correlation) on some components which leads to a lower mean than median. Seven 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, 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). The difficulty of the natural text tier might come from the noise of real text 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, variety scores are higher, but for natural text, variety scores are lower than balance scores.

Family rankings and the best/worst measures stay the same. The family ranking holds across tiers: Graph > Distribution > Distance > Geometry (simulated data averages $+0.95/+0.85/+0.78/+0.50$, natural text $+0.86/+0.50/+0.46/+0.36$). On the simulated data tier, the seven well-performing measures span all four families, though unevenly. Graph-based measures degrades least across tiers (0.09, versus 0.36 for distribution); geometry stays weakest. At the measure level, graph entropy is the best-performing measure and convex hull volume is the worst on both tiers.

6 Robustness Checks

Sentence embeddings are known to be sensitive to model choice and representational dimensionality (Ethayarajh, 2019; Tsukagoshi and Sasano, 2025). Because our results are produced on top of one fixed embedding setup, a natural question is whether the observed findings reflect real differences between measures, or just the specific embedder and dimension we chose. We therefore test this 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 $d \in \{256, 384, 1024, 2048, 4096\}$. Re-

sults are shown in Fig. 5 (Appendix).

Most findings are stable across dimensions, with a few exceptions. The seven measures that perform well at $d = 384$ also do so for all other tested dimensions. Similarly, the distribution-based and graph-based measures stay the strongest, and the geometry-based measures stays the weakest. Only Vendi (q=1), Rényi Kernel Entropy, and Graph Entropy perform perfectly across all dimensions. Some measures worsen as d grows (e.g., MST on balance and MagArea on disparity) and a few improve as d grows (e.g., Isocontour Radius).

6.2 Embedding models

We re-encode the natural text tier datasets with two different semantic embedding models (i.e., *Qwen3-Embedding-8B*[5] and *e5-mistral-7b-instruct*[6]) and an embedding model sensitive to style rather than semantics (i.e., *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 §C), 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 a 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.

Recommendations. The best available diversity measure depends on the diversity component of interest, the embedding model chosen, and the available computational budget (see App. Tab. 4 for computational cost of each measure). We make the following recommendations:

- **Do not use Mean Pairwise Distance.** Although being intuitive and popular, mean pairwise distance is performing worse than 5 others on simulated data and worse than 10 other measures on real text.
- **Use Graph Entropy.** Graph Entropy has the highest aggregated performance scores on three components of diversity, compared to any other measure. However, for specific domains and diversity component other measures might be better. For component-specific shortlists:
  - **Variety:** Chamfer Distance, Sum Bottleneck, and MST Dispersion
  - **Balance:** Graph Entropy
  - **Disparity:** Graph Entropy (less stable for Qwen3) and MST dispersion
- **Choose your embedding model carefully.** Choose an embedding model that (primarily) captures the concept of diversity you want to capture (e.g.semantic diversity). Even if different embedding models capture the same diversity dimension (e.g., semantic), differences are expected to be smaller but can still affect measure rankings. We recommend practitioners to confirm that the embedding they use is sensitive to their diversity dimension of interest.

[5] https://huggingface.co/Qwen/Qwen3-Embedding-8B

[6] https://huggingface.co/intfloat/e5-mistral-7b-instruct

[7] https://huggingface.co/AnnaWegmann/Style-Embedding

[8] Note that both Qwen3 and E5 are optimised for tasks beyond pure semantic representation: Qwen3 for retrieval and ranking, and E5-Mistral with InfoNCE at low temperature (τ=0.02), which produces embeddings concentrated in a narrow cone. We therefore base our main conclusions on MiniLM.

Limitations

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—rather than sweeping them systematically. We report Vendi score at $q = 1$, energy order at $\gamma = 1$, histogram entropy on a 5×5 UMAP grid (following Cox et al., 2021), and cluster inertia with $K = 200$ clusters (following Yang et al., 2025); convex hull volume is reported only on 2D UMAP projections, since direct computation is intractable at our embedding dimensionality. A systematic sweep over these parameter values may shift the relative rankings we report.

Our benchmark only captures part of the picture. Diversity measures should not be evaluated solely based on their performance on our benchmark. Our benchmark only tests the measures against the three diversity components (variety, balance and disparity). Furthermore, our evaluation only considers whether datasets are correctly ordered according to the diversity components tested. 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). Moreover, there are other aspects that should be considered when selecting a measure, such as its actionability (Delobelle et al., 2024).

Our benchmark does not test measures against the diversity axioms of Mironov and Prokhorenkova (2025). Our benchmark assesses empirical behaviour across three controlled components. It does not test the formal axiomatic properties (monotonicity, uniqueness, continuity) studied by Mironov and Prokhorenkova (2025). The two approaches are complementary: a metric may satisfy these axioms yet respond weakly on our three-component diversity benchmark.

Sentence-level units 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 longer-unit diversity therefore remains an open question for future work.

English Wikipedia only. Our natural-text tier is built exclusively from English Wikipedia, whose domains/topics 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.

Wikipedia inherits several confounds inherent to natural text evaluation

1. **Topic labels as imperfect proxy.** We use Wikipedia topic labels to create the natural text tier. However, not all sentences within a given article are directly related to the article's topic, introducing noise into the topic clusters.

2. **Disparity via hierarchy.** On our natural text tier, disparity is operationalized as the number of domains spanned. The 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.

3. **Embedding-model artefacts.** Because the measurements are computed over sentence embeddings, the measurements are also influenced by properties of the embedding space itself (anisotropy, hubness) in addition to properties of the underlying data. Understanding how the structure of natural-text embedding spaces influences embedding-based diversity measures is an important direction for future work.

Ethical Considerations

Risk of misuse. Our work is methodological and introduces no new data or models: all text comes from English Wikipedia (an editorially moderated public source), and the released benchmark consists of dataset configurations and evaluation code only. We do not foresee immediate ethical risks beyond standard methodological caveats common to any NLP evaluation work.

PII and offensive content. The natural text tier draws sentences from English Wikipedia, so public-

figure names that may appear in biographical articles are public information by Wikipedia's own biographical standards and are not personally identifying in the sense of contact details, addresses, or government identifiers; we apply no additional name redaction. Wikipedia articles may discuss sensitive topics (e.g., conflicts, controversial historical figures) in encyclopaedic context; we apply no offensive-content filtering.

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*.

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.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 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. 2021. 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 Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, 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, Christophe Servan, Thomas Lavergne, and Agata Savary. 2026. A Diversity Diet for a Healthier Model: A Case Study of French ModernBERT. 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. 2024. 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.

FB Karsdorp, Enrique Manjavacas, and Lauren Fonteyn. 2022. Introducing functional diversity: A novel approach to lexical diversity in (historical) corpora. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

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.

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. 2012. 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.

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. pages 3808–3816.

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.

Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15):707–719.

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.

Fedor Velikonivtsev, 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. *arXiv preprint*.

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, 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.

Yutong Xie, Ziqiao Xu, Jiaqi Ma, and Qiaozhu Mei. 2021. How much space has been explored? measuring the chemical space covered by databases and machine-generated molecules. *arXiv preprint arXiv:2112.12542*.

Vikas Yadav, Hyuk joon Kwon, Vijay Srinivasan, and Hongxia Jin. 2024. Explicit over Implict: 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, Mingqi Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Measuring Data Diversity for Instruction Tuning: A Systematic Analysis and A Reliable Metric. *arXiv preprint*.

Yu Yu, Shahram Khadivi, and Jia Xu. 2022. 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. *arXiv preprint*.

A Detailed Taxonomy of Embedding-based Diversity Measures

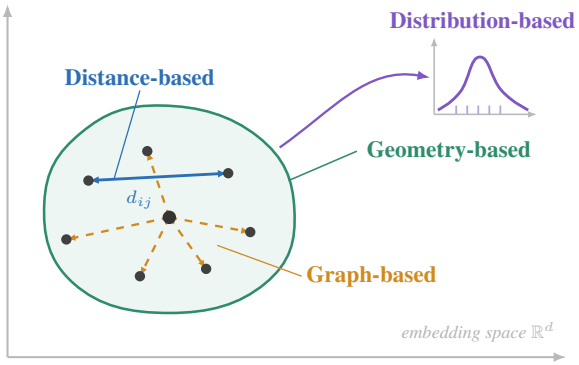


Figure 3: Illustration behind the four measure families on a toy vector space.

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

A.0.1 Notation and Conventions

Throughout this Appendix, we use the following notation:

- $X = \{x_1, \ldots, x_n\}$: 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)
- $s_{ij} = sim(\mathbf{e}_i, \mathbf{e}_j)$: similarity between embeddings (cosine similarity, unless specified)

A.1 Distance-based Measures

They directly measure 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. Average the similarity scores of each sample to all other samples in the embedding space. A higher value indicates that samples are semantically closer (i.e., less diverse), whereas a lower value implies greater semantic diversity.

Computation. Compute the average of all off-diagonal pairwise similarities (i.e. between distinct samples):

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

In principle, this measure generalizes to any function capable of quantifying similarity or distance between two samples. However, within the context of semantic embeddings, **cosine similarity** is the most prevalent choice. When employed, this measure is often referred to as **Self-cosSim** (Tevet and Berant, 2021; Zhang et al., 2024; Miranda et al., 2025). Some works also refer to the distance-based variant (specifically using cosine distance) as **Remote-Clique** (Cox et al., 2021). Since cosine distance is $1 -$ cosine similarity, these measures are mathematically equivalent (linearly related).
Complexity. $\mathcal{O}(n^2 d)$

A.1.2 Pairwise Dist. Sum: Max Dispersion (MD) / DistSum

Intuition. By summing all pairwise distances, it captures the overall spread of samples in the embedding space. Higher values indicate greater diversity.
Computation. Both terms refer to the same formulation, originally introduced as Max Dispersion (MD) by Yu et al. (2022) and later adopted as DistSum by Yang et al. (2025).
Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, the Max Dispersion (or DistSum) is defined as the total sum of pairwise distances:

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

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

A.1.3 Chamfer Distance

Intuition. Chamfer Distance is a concept originally from Computational Geometry and Computer Vision (specifically point cloud processing), where it is used to comparing *two* sets of points by measure how close every point in one cloud is to its nearest neighbour in the other. 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 semantic diversity, prior work adapts this idea to a *single* dataset (Cox et al., 2021; Zhang et al., 2025a). Instead of comparing two clouds, the measure computes, for each sample, the distance to its nearest neighbour. This reflects the *local compactness* of the embedding space: datasets with tight clusters have smaller nearest-neighbour distances, while more dispersed datasets yield larger values.
Computation. the Chamfer Distance is defined as the average of the minimum pairwise distances:

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

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

A.1.4 KNN Distance

[9] **Intuition.** Generalizes Chamfer Distance by measuring the average distance from each sample to its k-th nearest neighbour (Yang et al., 2025). This formulation captures local neighbourhood density: the larger the distance to the k-th neighbour, the more sparsely the dataset is distributed.
Computation. Let $N_k(x_i)$ denote the k-th nearest neighbour of sample x_i. Then the KNN Distance is defined as:

$$\mathrm{KNN}(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} d\big(x_i, N_k(x_i)\big). \quad (4)$$

When $k = 1$, it is identical to the Chamfer Distance.
Complexity. $\mathcal{O}(n^2 d + n \log n)$ for pairwise distance computation and neighbour search.

A.1.5 Energy(γ)

Intuition. Measures dataset diversity as the negative potential energy of a system of equally charged particles distributed in the embedding space (Velikonivtsev et al., 2024; Mironov and Prokhorenkova, 2025). Each pair of samples repels each other inversely proportional to their distance raised to the power γ. Thus, when samples are far apart, the potential energy (and hence the diversity score) increases. The measure encourages uniform spacing and penalises close or duplicate points.

[9] Not implemented because at $k = 1$ this reduces to Chamfer Distance, which we already evaluate, and we do not explore additional values of k in this work.

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$ and pairwise distances $d_{ij} = d(\mathbf{e}_i, \mathbf{e}_j)$, the Energy measure of order $\gamma > 0$ is defined as:

$$\text{Energy}_\gamma(\mathbf{X}) = -\frac{2}{n(n-1)} \sum_{i<j} \frac{1}{d_{ij}^\gamma}. \quad (5)$$

The negative sign ensures that configurations with smaller pairwise distances (i.e., higher "repulsive energy") yield lower diversity values, and increasing all pairwise distances increases the overall diversity score. In the special case $\gamma = 1$, the measure corresponds to the physical energy of a system of point charges under Coulomb repulsion.
Complexity. $\mathcal{O}(n^2 d)$ (pairwise distance computation for all sample pairs)

A.2 Geometry-based Measures

A.2.1 Diameter

Intuition. This measure finds the largest distance between any two embeddings in the embedding space of a dataset (Mironov and Prokhorenkova, 2025; Xie et al., 2021). It shows how far apart the most different embeddings are in the embedding space, figuratively the "width" of the data cloud.
Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, define pairwise distances $d_{ij} = d(\mathbf{e}_i, \mathbf{e}_j)$. The Diameter is the largest of these pairwise distances:

$$\text{Diameter}(\mathbf{X}) = \max_{i<j} d_{ij} = \max_{i<j} d(\mathbf{e}_i, \mathbf{e}_j). \quad (6)$$

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

A.2.2 Bottleneck

Intuition. Measures dataset diversity as the smallest distance between any two data points (Mironov and Prokhorenkova, 2025; Xie et al., 2021).
Computation. For a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with pairwise distances d_{ij}, the Bottleneck is defined as:

$$\text{Bottleneck}(\mathbf{X}) = \min_{i<j} d_{ij}. \quad (7)$$

Caveats. As a standalone 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. $\mathcal{O}(n^2 d)$

A.2.3 Sum Diameter

Intuition. Sum 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** sample (Mironov and Prokhorenkova, 2025; Xie et al., 2021).
Computation. Let $d_{ij} = d(\mathbf{e}_i, \mathbf{e}_j)$ denote pairwise distances between embeddings. For each sample x_i, find its maximum distance to all other samples and sum these maxima:

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

Optionally, the result can be normalized by dividing by n.
Complexity. $\mathcal{O}(n^2 d)$

A.2.4 Sum Bottleneck

Intuition. While the Bottleneck measure is determined only by the closest pair of samples (making it sensitive to (near) duplicates), SumBottleneck aggregates the minimum distance from each data point to its nearest neighbour for **every** sample.
Computation. Let $d_{ij} = d(\mathbf{e}_i, \mathbf{e}_j)$ denote pairwise distances between embeddings. For each sample x_i, find its minimum distance to all other samples and sum these minima:

$$\text{SumBottleneck}(\mathbf{X}) = \sum_{i=1}^{n} \min_{j \neq i} d_{ij}. \quad (9)$$

Optionally, the result can be normalized by dividing by n, yielding the mean nearest-neighbour distance.
Caveats. Unlike Diameter or Sum Diameter, this measure primarily reflects *local* density rather than the overall spread of the dataset: low values indicate clustering or redundancy among neighbours, while high values indicate that even the closest pairs are well separated.
Complexity. $\mathcal{O}(n^2 d)$

A.2.5 Convex Hull Volume

[10] **Intuition.** This measure treats each embedding as a data point, and the convex hull represents the

[10] In 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.

smallest convex region containing all points in the embedding space (Yu et al., 2022). Larger hull volume indicates that samples are more widely distributed, thus having a greater diversity.

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, the convex hull $\mathcal{H}(\mathbf{X})$ is defined as the smallest convex set containing all embeddings:

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

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

$$\mathrm{CHV}(\mathbf{X}) = \mathrm{Vol}_d\big(\mathcal{H}(\mathbf{X})\big). \tag{10}$$

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; worst-case $\mathcal{O}(n^2)$ in degenerate configurations.

A.2.6 Isocontour Radius

Intuition. Computes the geometric mean of the standard deviations across all embedding dimensions (Lai et al., 2020; Yang et al., 2025). It treats each component as an independent direction of variation and models the distribution of embeddings as a multivariate Gaussian. The resulting value corresponds to the radius of the ellipsoidal isocontour that encloses most data points, capturing how widely the dataset spreads in all dimensions. A larger radius indicates greater diversity.

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i = [e_{i,1}, \ldots, e_{i,d}]^\top \in \mathbb{R}^d$, compute the standard deviation σ_j of the j-th embedding dimension across all samples. The Isocontour Radius is defined as the geometric mean of these per-dimension standard deviations:

$$\mathrm{IsocontourRadius}(\mathbf{X}) = \left(\prod_{j=1}^{d} \sigma_j \right)^{\frac{1}{d}}. \tag{11}$$

Equivalently, if the embeddings are approximately Gaussian-distributed with diagonal covariance $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, then this measures the mean radius of the ellipsoidal isocontour:

$$(x - \boldsymbol{\mu})^\top \Sigma^{-1} (x - \boldsymbol{\mu}) = c^2, \tag{12}$$

where $\boldsymbol{\mu}$ is the mean vector of embeddings and c is a constant controlling contour level.

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

A.2.7 Span (Centroid)

Intuition. It measures the average distance from each embedding to the centroid of all embeddings in the dataset (Cox et al., 2021). 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:

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{e}_i. \tag{13}$$

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

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

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

$$\mathrm{Span}_{\mathrm{centroid}}(\mathbf{X}) = \mathrm{Percentile}_{90}(D). \tag{14}$$

Typical distance functions include cosine or Euclidean distance.

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

A.2.8 Span (Medoid)

Intuition. It measures 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., 2021). Unlike the Span (Centroid), Span (Medoid) is less sensitive to outliers.

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ and pairwise distances $d_{ij} = d(\mathbf{e}_i, \mathbf{e}_j)$, identify the medoid:

$$x_m = \arg\min_{x_i \in \mathbf{X}} \sum_{j=1}^{n} d_{ij}. \tag{15}$$

Compute the mean distance of all samples to this medoid:

$$\mathrm{Span}_{\mathrm{medoid}}(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} d(\mathbf{e}_i, \mathbf{e}_m). \qquad (16)$$

Higher values indicate a higher diversity.
Complexity. $\mathcal{O}(n^2 d)$ (pairwise distance matrix required to identify medoid)

A.2.9 Cluster Inertia

Intuition. In physics, the *moment of inertia* describes how difficult it is to change an object's rotational state. This resistance happens because the mass is distributed far from the axis, making the object "hard to control". By analogy, this measure quantifies the dispersion of data points (mass) around their cluster centroids. A high inertia value implies that samples are scattered far from the centre (high resistance to collapsing into a single cluster), indicating a diverse dataset; whereas low inertia suggests samples are tightly packed and redundant.
Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, partition $\mathbf{X}$ into K clusters $\{C_1, \ldots, C_K\}$ using K-means (Yang et al., 2025; Du and Black, 2019). Let $\boldsymbol{\mu}_j$ denote the centroid of cluster C_j. The Cluster Inertia is defined as the total within-cluster sum of squared distances (Euclidean):

$$\mathrm{Inertia}(\mathbf{X}) = \sum_{j=1}^{K}\sum_{x_i \in C_j} \|\mathbf{e}_i - \boldsymbol{\mu}_j\|^2. \qquad (17)$$

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

A.2.10 Log-Determinant Diversity (LDD)

Intuition. Measures dataset diversity as the volume spanned by all sample embeddings in a feature space (Wang et al., 2025a; Yang et al., 2025). It quantifies how linearly independent the representations are: if samples are redundant, their embeddings occupy a smaller subspace, resulting in a lower determinant. Thus, higher LDD values indicate greater diversity.
Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, construct the similarity matrix $\mathbf{K}$ with elements $k_{ij} = \cos(\mathbf{e}_i, \mathbf{e}_j)$ or another positive semidefinite kernel. The Log-Determinant Diversity is defined as:

$$\mathrm{LDD}(\mathbf{X}) = \log\det(\mathbf{K} + \epsilon\mathbf{I}), \qquad (18)$$

where ϵ is a small constant for numerical stability. The determinant measures the hyper-volume spanned by the embeddings. The logarithm is taken to ensure numerical robustness.
Complexity. $\mathcal{O}(n^3)$ (matrix determinant via Cholesky or eigenvalue decomposition)

A.3 Distribution-based Measures

A.3.1 Vendi Score

Intuition. Inspired by ecology, this one measures diversity as the effective number of unique elements in a dataset, derived from the entropy of the eigenvalue spectrum of a similarity matrix (Friedman and Dieng, 2023; Mironov and Prokhorenkova, 2025; Limbeck et al., 2024). It captures both the richness and balance of samples' semantic representations. It achieves its maximum value (n) when all samples are orthogonal (fully dissimilar) and its minimum value (1) when all samples are identical.
Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, construct the similarity matrix $\mathbf{S}$ with elements $s_{ij} = \exp(-d(\mathbf{e}_i, \mathbf{e}_j))$ or cosine similarity. Let $\tilde{\mathbf{S}} = \mathbf{S}/\mathrm{tr}(\mathbf{S})$ be its normalized form, and $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of $\tilde{\mathbf{S}}$. The Vendi Score is then defined as the exponential of the Shannon entropy of the eigenvalue distribution:

$$\mathrm{Vendi}(\mathbf{X}) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right). \qquad (19)$$

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

A.3.2 Vendi-q

[11] **Intuition.** Extends the original Vendi Score by introducing an order parameter q that controls sensitivity to rare or common samples (Pasarkar and Dieng, 2024; Yang et al., 2025). Smaller q values emphasize rare, unique samples, whereas larger q

[11] Not implemented because at $q = 1$ this reduces to Vendi Score, which we already evaluate, and we do not explore additional values of q in this work.

values give greater weight to frequent or highly similar ones. When $q = 1$, the measure reduces to the standard Vendi Score based on Shannon entropy.

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, construct a similarity matrix $\mathbf{S}$ with entries $s_{ij} = \exp(-d(\mathbf{e}_i, \mathbf{e}_j))$ or cosine similarity. Normalize it as $\tilde{\mathbf{S}} = \mathbf{S}/\mathrm{tr}(\mathbf{S})$, and let $\lambda_1, \ldots, \lambda_n$ denote its eigenvalues. The Vendi-q diversity score is defined as the Hill-number generalization of the Vendi Score:

$$\mathrm{Vendi}_q(\mathbf{X}) = \left(\sum_{i=1}^{n} \lambda_i^q \right)^{\frac{1}{1-q}}, \qquad q > 0. \quad (20)$$

For $q \to 1$, the expression converges to the Shannon-based Vendi Score:

$$\lim_{q \to 1} \mathrm{Vendi}_q(\mathbf{X}) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i \right).$$

Complexity. $\mathcal{O}(n^3)$ (eigenvalue decomposition of the $n \times n$ similarity matrix)

A.3.3 Rényi Kernel Entropy (RKE)

Intuition. Measures dataset diversity through the information-theoretic entropy of the similarity matrix in kernel space. It estimates the effective number of modes or clusters in the data distribution by treating normalized pairwise similarities as a probability distribution (Mironov and Prokhorenkova, 2025; Jalali et al., 2023). The RKE is based on the matrix-based Rényi entropy from quantum information theory and captures how spread out the data are in the embedding space: more uniformly distributed samples yield higher entropy (greater diversity).

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, construct a normalized kernel (similarity) matrix $\mathbf{S}$ with entries:

$$S_{ij} = \frac{1}{n} k(\mathbf{e}_i, \mathbf{e}_j),$$

where $k(\mathbf{e}_i, \mathbf{e}_j)$ is a positive semi-definite kernel (e.g., Gaussian or cosine). Let $\lambda_1, \ldots, \lambda_n$ denote its eigenvalues such that $\sum_i \lambda_i = 1$. The Rényi Kernel Entropy of order $\alpha > 0$ is defined as:

$$\mathrm{RKE}_\alpha(\mathbf{X}) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} \lambda_i^\alpha \right). \quad (21)$$

The most common case is $\alpha = 2$, for which:

$$\mathrm{RKE}_2(\mathbf{X}) = -\log \left(\frac{1}{n^2} \sum_{i,j} S_{ij}^2 \right) = -\log\left(\|\mathbf{S}\|_F^2 \right), \quad (22)$$

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

Complexity. $\mathcal{O}(n^2 d)$ 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.3.4 DCScore

Intuition. DCScore measures 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. Thus, DCScore reflects the dataset's effective number of distinguishable samples while being computationally more efficient.

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i = \Phi(x_i) \in \mathbb{R}^d$, construct a kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with entries

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

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

$$P[i, j] = \frac{\exp(K[i, j]/\tau)}{\sum_{k=1}^{n} \exp(K[i, k]/\tau)}, \quad (23)$$

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:

$$\mathrm{DCScore}(\mathbf{X}) = \mathrm{tr}(\mathbf{P}) = \sum_{i=1}^{n} P[i, i]. \quad (24)$$

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

A.3.5 Bins Entropy

Intuition. Measures how evenly sample embeddings are distributed across a discretized low-dimensional projection of the embedding space (Yang et al., 2025; Cox et al., 2021). It operationalizes diversity as the spatial evenness of points after after projecting the high-dimensional embeddings (e.g., with UMAP or PCA) into a 2D grid. Datasets whose samples occupy many bins with similar frequencies exhibit higher entropy, indicating a higher diversity.

Computation. Given embeddings $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ projected into a 2D space $\mathbf{Y} = \mathrm{UMAP}(\mathbf{E})$, divide the projected space into a $B_x \times B_y$ grid of bins $\{b_1, \ldots, b_B\}$, 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:

$$\begin{aligned} \mathrm{Entropy}_{\mathrm{bins}}(\mathbf{X}) &= -\sum_{b=1}^{B} f_b \log f_b, \\ \text{where } f_b &= \frac{n_b}{n}, \ \sum_b f_b = 1. \end{aligned} \tag{25}$$

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

$$\mathrm{Entropy}_{\mathrm{norm}} = \frac{\mathrm{Entropy}_{\mathrm{bins}}}{\log(B)}. \tag{26}$$

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

A.3.6 MagArea (Metric Space Magnitude Area)

Intuition. Measures dataset diversity by quantifying the *size* or *magnitude* of the embedding space as a metric object (Limbeck et al., 2024). The concept originates from category-theoretic metric geometry, where the *magnitude* of a finite metric space reflects the effective number of distinct points, taking into account pairwise distances. MagArea extends this idea by computing the *area under the magnitude curve* as the distance-scaling parameter τ varies, thereby capturing both local and global diversity across multiple distance scales.

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, define the similarity kernel $\mathbf{Z}_\tau$ parameterized by scale $\tau > 0$:

$$Z_{ij}(\tau) = \exp(-\tau\, d(\mathbf{e}_i, \mathbf{e}_j)), \tag{27}$$

where $d(\mathbf{e}_i, \mathbf{e}_j)$ denotes the pairwise distance (e.g., cosine or Euclidean). The *magnitude* of the metric space at scale τ is:

$$\mathrm{Mag}(\tau) = \mathbf{1}^\top \mathbf{Z}_\tau^{-1} \mathbf{1}, \tag{28}$$

where $\mathbf{1}$ is the all-ones vector. The **MagArea** is defined as the integral of this magnitude function over a range of scales:

$$\mathrm{MAGAREA}(\mathbf{X}) = \int_0^\infty \mathrm{Mag}(\tau)\, d\tau. \tag{29}$$

In practice, this integral is approximated numerically by evaluating $\mathrm{Mag}(\tau)$ at a discrete set of logarithmically spaced scales τ_k and summing with trapezoidal integration.

Complexity. $\mathcal{O}(Tn^3)$, where T is the number of sampled τ values (each step requires inversion of the $n \times n$ similarity matrix $\mathbf{Z}_\tau$).

A.3.7 Species(q)

[12] **Intuition.** A similarity-aware generalization of Shannon and Hill-number diversity that accounts not only for the number of items (species or samples) but also for how similar they are (Leinster and Cobbold, 2012; Mironov and Prokhorenkova, 2025). Each item contributes to diversity proportionally to how *unique* it is relative to all others in the dataset. When items are very similar (e.g., embeddings clustered together), their contribution is down-weighted; when dissimilar, they contribute more strongly. Species(q) thus measures the *effective number of distinct items*, extending classical entropy-based diversity to continuous similarity spaces.

Computation. Given embeddings $\mathbf{X} = \{x_1, \ldots, x_n\}$ with similarities $S_{ij} = \cos(\mathbf{e}_i, \mathbf{e}_j)$ or other normalized kernel values ($0 \le S_{ij} \le 1$), define the *similarity-adjusted abundance* of each item as

$$(\mathbf{S}p)_i = \sum_{j=1}^{n} S_{ij} p_j, \tag{30}$$

[12] We do not implement Species(q) in our benchmark: it belongs to the same similarity-sensitive Hill-number family as the Vendi Score and dcscore, differing only in aggregation (row-sum vs spectral)

where p_i denotes the normalized weight of each sample ($p_i = 1/n$ for uniform datasets). The Species diversity of order q is defined as

$$^{q}D_{\mathbf{S}}(p) = \left(\sum_{i=1}^{n} p_i \, (\mathbf{S}p)_i^{q-1} \right)^{\frac{1}{1-q}}, \quad q \geq 0,\, q \neq 1. \tag{31}$$

As $q \to 1$, the measure converges to a Shannon-like form:

$$^{1}D_{\mathbf{S}}(p) = \exp\left(-\sum_{i=1}^{n} p_i \ln(\mathbf{S}p)_i \right), \tag{32}$$

and for $S = I$, it reduces to the standard Hill numbers: $^{q}D(p) = (\sum_i p_i^q)^{1/(1-q)}$. In the embedding-based adaptation (Mironov & Prokhorenkova, 2025), S_{ij} is derived from cosine or RBF similarity between embeddings, and $^{q}D_{\mathbf{S}}$ quantifies the effective number of semantically distinct samples in the dataset.

Complexity. $\mathcal{O}(n^2 d)$ (pairwise similarity computation and matrix–vector product per iteration); can be reduced to $\mathcal{O}(nkd)$ with sparse or k-NN similarity graphs.

A.3.8 Functional Diversity

[13] **Intuition.** Measures the effective number of functionally distinct items by integrating semantic distances into the Hill-number diversity framework (Karsdorp et al., 2022). Originally developed for ecological and later for lexical diversity, it quantifies not only how many unique elements exist (variety), but also how distinct they are (disparity) and how evenly they occur (balance). In a semantic context, functional diversity generalizes to embeddings by replacing word-type distances with embedding-based semantic distances. Higher values indicate that samples are semantically distinct and uniformly distributed across the embedding space.

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with relative frequencies p_i and pairwise distances d_{ij} (e.g., cosine distance between embeddings), the functional diversity of order q is defined as:

$$^{q}D_{\text{func}} = \left(\sum_{i=1}^{n} p_i \left(\sum_{j=1}^{n} p_j \, \exp(-\beta \, d_{ij}) \right)^{q-1} \right)^{\frac{1}{1-q}}, \tag{33}$$

where β controls distance sensitivity: higher β emphasizes near distances (local distinctiveness), while smaller β emphasizes global structure. At $q = 0$, diversity depends only on the number of distinct items (richness); at $q = 1$, it converges to the exponential of Shannon entropy; and at $q = 2$, it corresponds to the inverse of the Gini–Simpson index.

Complexity. $\mathcal{O}(n^2 d)$ (pairwise distance computation and summation across all types)

A.4 Graph-theory-based Measures

A.4.1 MST Dispersion

Intuition. Originally adapted from phylogenetics and ecology (e.g., Faith's PD), MST Dispersion measures diversity by constructing the Minimum Spanning Tree (MST) of the embedding graph. It sums the lengths of the edges required to connect all samples with the minimum total cost (Cox et al., 2021; Atwal et al., 2025).

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, 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 $\mathrm{MST}(G)$ that connects all n nodes with the smallest possible total edge weight. The MST Dispersion is defined as:

$$\mathrm{MST}(\mathbf{X}) = \frac{1}{n-1} \sum_{(i,j) \in \mathrm{MST}(G)} d_{ij}. \tag{34}$$

Optionally, distances can be computed using cosine or Euclidean metrics; normalization by $(n-1)$ ensures comparability across datasets of different sizes.

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

[13] Not implemented because it is mathematically identical to Leinster's Species(q) with $S_{ij} = \exp(-\beta \, d_{ij})$ (an RBF kernel). Both belong to the similarity-sensitive Hill-number framework, of which the Vendi Score is the spectral counterpart in our benchmark.

A.4.2 Graph Entropy

Intuition. Measures dataset diversity as the uncertainty of pairwise distances between samples represented as a fully connected graph (Yu et al., 2022). Each node corresponds to an embedding, and each edge weight reflects distance between two nodes. By computing the entropy of normalized edge weights, this measure 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. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, we 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, we normalize its pairwise distances to obtain a local distance distribution:

$$f_i(d_{ij}) = \frac{d_{ij}}{\sum_{k=1}^{n} d_{ik}}. \tag{35}$$

The local distance entropy of node x_i is then:

$$I(x_i) = -\sum_{j=1}^{n} f_i(d_{ij}) \log f_i(d_{ij}),^{14} \tag{36}$$

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

$$\text{GraphEntropy}(\mathbf{X}) = \sum_{i=1}^{n} I(x_i). \tag{37}$$

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

A.4.3 Hamiltonian Diversity (HamDiv)

[15] **Intuition.** Measures dataset diversity as the total length of the shortest Hamiltonian circuit over all sample embeddings (Hu et al., 2024; Mironov and Prokhorenkova, 2025). Each sample is treated as a node in a complete weighted graph, where edge weights are pairwise distances between embeddings. The measure captures both *quantity* (number of samples) and *dissimilarity* (pairwise distance) by finding the minimal "travel cost" required to visit all nodes exactly once. Datasets whose embeddings are widely separated yield longer Hamiltonian circuits, indicating higher diversity.

Computation. Given a dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ with embeddings $\mathbf{e}_i \in \mathbb{R}^d$, construct a complete undirected graph $G = (V, E)$ with $V = \{1, \ldots, n\}$ and edge weights $d_{ij} = d(\mathbf{e}_i, \mathbf{e}_j)$. Let $\mathbf{Y} = (y_{ij})$be the adjacency indicator matrix where $y_{ij} = 1$ if the Hamiltonian circuit traverses the edge from node i to j, and 0 otherwise. The Hamiltonian Diversity is defined as the length of the shortest Hamiltonian circuit:

$$\text{HamDiv}(\mathbf{X}) = \min_{\mathbf{Y}} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} d_{ij}\, y_{ij}, \tag{38}$$

subject to:

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} y_{ij} = 1, \qquad \forall j = 1, \ldots, n, \tag{39}$$

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} y_{ij} = 1, \qquad \forall i = 1, \ldots, n, \tag{40}$$

$$\sum_{\substack{i,j \in S \\ i \neq j}} y_{ij} \leq |S| - 1, \quad \forall\, S \subsetneq \{1, \ldots, n\},\ |S| \geq 2. \tag{41}$$

The last constraint is the subtour elimination constraint, preventing the solution from decomposing into multiple disjoint cycles and ensuring a single circuit visits all nodes exactly once. When $n = 2$, $\text{HamDiv}(\{x_1, x_2\}) = 2\, d_{12}$; for $n \leq 1$, the value is 0.

Complexity. Exact computation is NP-hard (equivalent to solving the Travelling Salesman Problem). Approximation algorithms such as: Dynamic programming ($\mathcal{O}(2^n n^2)$) for small n; 2-opt local search ($\mathcal{O}(n^2)$) for medium sets ($n \leq 1000$); Greedy or Christofides ($\mathcal{O}(n^3 \log n)$) for large n can be used for practical estimation.

B Benchmark Construction Details

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

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

[15] Not implemented in our benchmark: computing the shortest Hamiltonian circuit is NP-hard and becomes prohibitive at our dataset scale. A reference implementation is included in our code release for use on smaller datasets.

Table 2: Distribution conditions, ordered from most balanced (top) to least balanced (bottom).

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

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 characterise dimension-dependent measure behaviour; results in the main body use d=384.

Hierarchical Gaussian mixture. Super-category centres are drawn from $\mathcal{N}(\mathbf{0}, \sigma^2_{\text{super}} I_d)$ with $\sigma_{\text{super}} = 5\sqrt{d}$; category centres from $\mathcal{N}(\mathbf{s}_{m(i)}, \sigma^2_{\text{category}} I_d)$ with $\sigma_{\text{category}} = \sqrt{d}$; and each element of category i from $\mathcal{N}(\mathbf{c}_i, \sigma^2_{\text{element}} I_d)$ 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 package[16]) 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 bit-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 FPS Selection of Domains and Topics

Two independent selection settings. Variety and balance share one selection pool (10 domains $\times$ 5 topics = 50 topics); disparity uses a second pool (50 domains, one or more topics each) seeded with an offset of $+1000$ so the two pools are effectively independent. Both pools use the same FPS procedure described below.

FPS over Qwen3-Embedding-8B label embeddings. Domain and topic labels are encoded once with Qwen3-Embedding-8B and L2-normalised; embeddings are cached on disk and reused across seeds. Because the embeddings are unit-norm, Euclidean FPS is monotonically equivalent to cosine FPS. For each FPS run, the starting index is drawn from a seeded permutation of the candidate pool, which empirically gives well-separated starts

[16] https://pypi.org/project/Wikipedia-API/

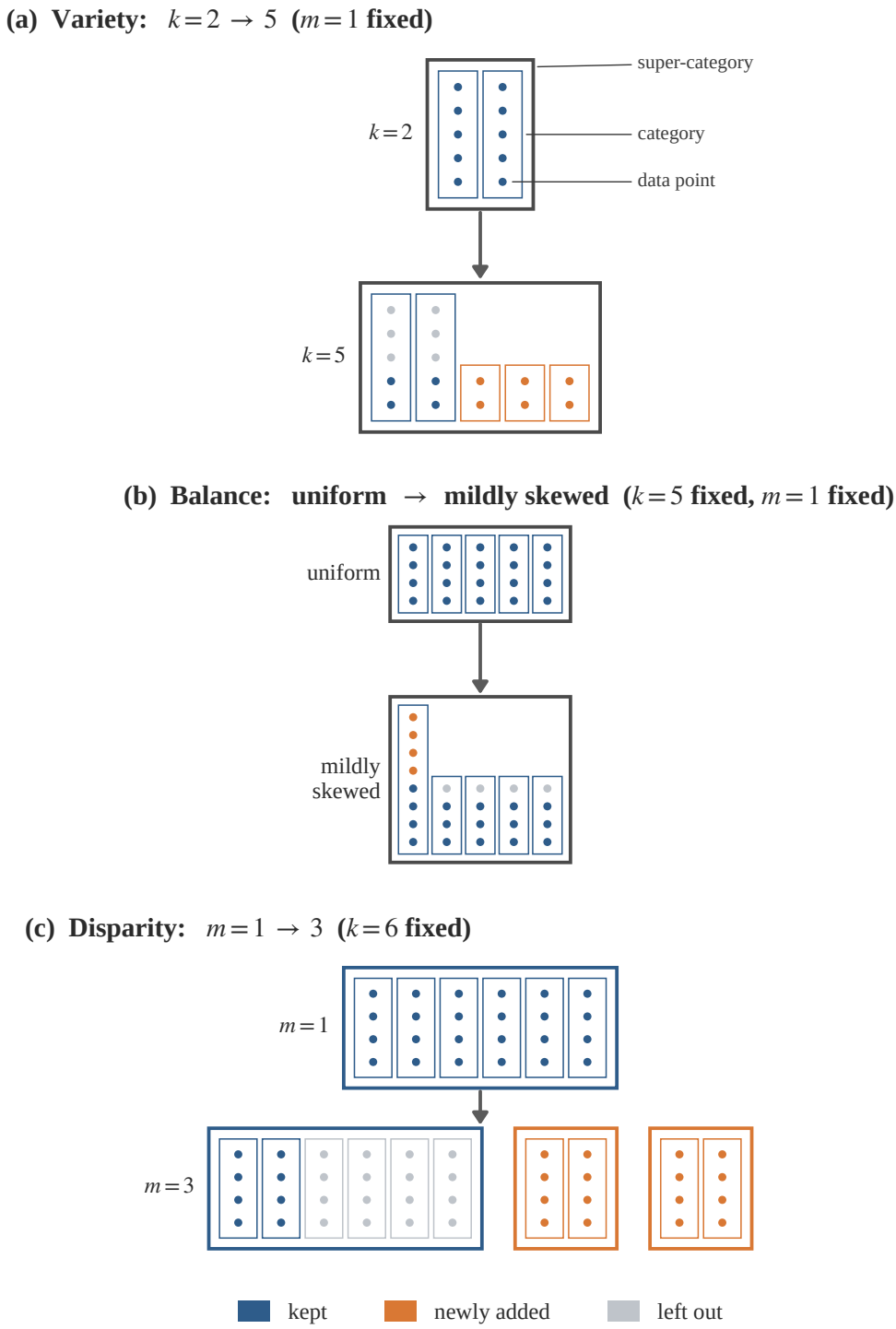


Figure 4: Incremental construction across the three components, illustrated with small example values. Each dataset is a hierarchy of *super-categories* (thick outer box) containing *categories* (thin inner column) containing *data points* (dot). The total dataset size n is conserved across levels in every panel. **Variety** (k=2→5, with m=1 fixed): each kept category's data points are truncated to a prefix, and 3 new categories are added. **Balance** (uniform → 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→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 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.

across seeds; the algorithm then proceeds greedily, at each step picking the candidate that is farthest from the currently selected set. The domain FPS is seeded by the run seed for the variety/balance pool and by seed $+ 1000$ for the disparity pool. Within each chosen domain, the topic FPS uses a stable per-(seed, domain) sub-seed derived as the first 8 hex digits of $\mathrm{SHA256}(\text{seed} \mid \text{domain})$, so topic selection is deterministic given the chosen domain.

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 $1 \to 3$, articles inspected per round $4{,}000 \to 20{,}000$, and per-article sentence cap $5 \to 40$. 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 (which avoids both semantic drift from deep subcategory descent and lead-paragraph dominance from any single article), 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 stylised opening sentences characteristic of Wikipedia articles.

C Results for Robustness Check

We elaborate on results for the robustness check. We display results on the natural text tier for *Qwen3-Embedding-8B*[17] compared to MiniLM in Fig. 6 and results on Wikipedia for *Style Embedding*[18] compared to MiniLM in Fig. 7.

Qwen3 improves correlations and creates the same family ranking. Several findings with MiniLM hold with Qwen3, a high performing embedding model on MTEB[19] (Enevoldsen et al., 2025): no measure performs perfectly any component, the family ranking (Graph $>$ Distribution $>$ Distance $>$ Geometry) is preserved, and disparity remains the hardest component. Further, most measures reach higher correlation with Qwen3 embeddings, with a few decreasing correlation on disparity.

D Statistics For Data

Each dataset contains n=10,000 elements. The simulated tier sweeps 5 embedding dimensionalities, with 5 seeds and 15 conditions (5 variety levels, 5 balance conditions, 5 disparity levels) per dimensionality, giving $5 \times 5 \times 15 = 375$ datasets.

[17] https://huggingface.co/Qwen/Qwen3-Embedding-8B

[18] https://huggingface.co/AnnaWegmann/Style-Embedding

[19] Note 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.

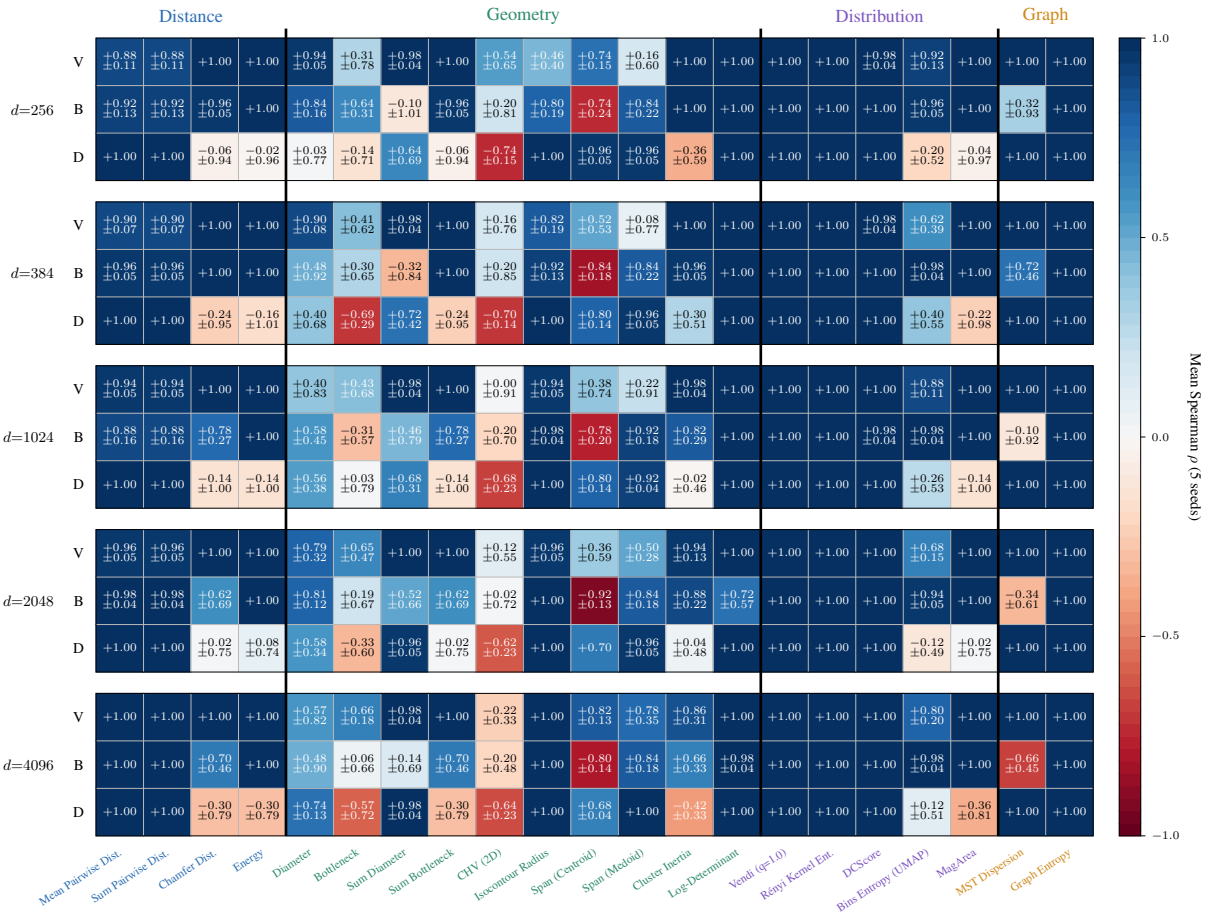


Figure 5: Per-componen t mean Spearman ρ (5 seeds) on the **simulated data (GMM)** tier across embedding dimensions $d \in \{256, 384, 1024, 2048, 4096\}$ (top to bottom) for the Variety (V), Balance (B) and Disparity (D) components. Columns are grouped and coloured by the four measure families (Distance-, Geometry-, Distribution-, Graph-based); the standard deviation is shown beneath the mean when it exceeds 0.01.

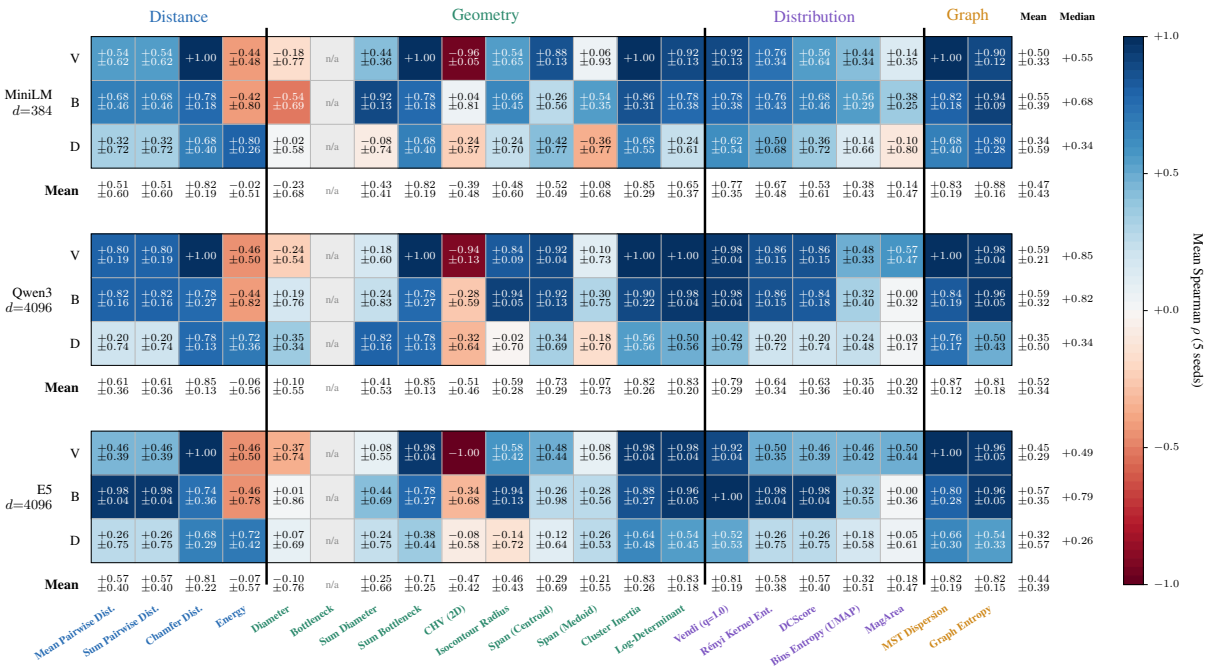


Figure 6: Per-component mean Spearman ρ (5 seeds) on the **natural text (Wikipedia)** tier for three sentence encoders: *all-MiniLM-L6-v2* (d=384, top), *Qwen3-Embedding-8B* (d=4096, middle), and *E5* (d=4096, bottom), along the Variety (V), Balance (B) and Disparity (D) components. Columns are grouped and coloured by the four measure families (Distance-, Geometry-, Distribution-, Graph-based). The two rightmost columns per panel give the mean ρ ($\pm$ mean std) and median ρ over measures; the row beneath each panel gives the per-measure mean ρ ($\pm$ mean std) over components. Note that `bottleneck` collapses to zero across all five levels for nearly every seed, making Spearman ρ undefined (*n/a*).

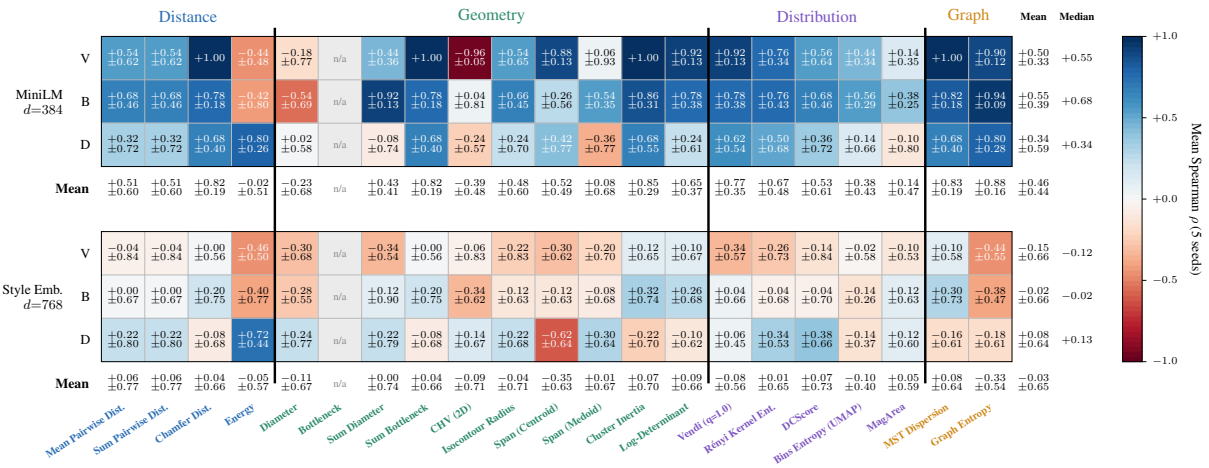


Figure 7: Per-component mean Spearman ρ (5 seeds) on the **Wikipedia** tier for two sentence encoders: *all-MiniLM-L6-v2* (d=384, top) and the *Style Embedding* model of Wegmann et al. (2022) (d=768, bottom), along the Variety (V), Balance (B) and Disparity (D) components. Columns are grouped and coloured by the four measure families (Distance-, Geometry-, Distribution-, Graph-based). The two rightmost columns per panel give the mean ρ ($\pm$ mean std) and median ρ over measures; the row beneath each panel gives the per-measure mean ρ ($\pm$ mean std) over components. `bottleneck` collapses to zero across all five levels for nearly every seed, making Spearman ρ undefined (*n/a*).

The natural-text tier uses 5 seeds and the same 15 conditions, giving $5 \times 15 = 75$ datasets. In total, the benchmark comprises 450 datasets.

E 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. 4 shows, per-measure cost spans more than four orders of magnitude, with MagArea alone accounting for ∼87% of the total (∼21.8 h) and the remaining 20 measures together completing in ∼3.4 h.

F 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.[20] 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.[21] `e5-mistral-7b-instruct` (7B parameters) (Wang et al., 2024), is released under the MIT licence.[22] `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.[23] All four models are used strictly for sentence embedding extraction, consistent with their intended use.

Crawling library. The Wikipedia crawler (§**??**) uses the `Wikipedia-API` Python package[24], 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 organised 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).[25] 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 meta-evaluation research. The curated Wikipedia data is released under CC BY-SA 4.0, inheriting the licence of the underlying Wikipedia text; The simulated datasets and source code are released under CC BY 4.0. The intended use of EmbDivBench is the meta-evaluation of embedding-based diversity metrics within NLP/ML research; we do not anticipate use beyond this scope.

G Use of AI Assistants

We used Claude Code and GitHub Copilot for coding tasks—primarily command lookup and occasional function generation, with all generated

[20] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[21] https://huggingface.co/Qwen/Qwen3-Embedding-8B

[22] https://huggingface.co/intfloat/e5-mistral-7b-instruct

[23] https://huggingface.co/AnnaWegmann/Style-Embedding

[24] https://pypi.org/project/Wikipedia-API/

[25] https://en.wikipedia.org/wiki/Wikipedia:Copyrights

Table 4: 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×	$O(n^2 d)$
		Sum Pairwise Dist.	1.00×	$O(n^2 d)$
		Energy	1.02×	$O(n^2 d)$
	🔵	Chamfer Dist.	1.05×	$O(n^2 d)$
Geometry		Convex Hull (2D)[‡]	$<$0.01×	$O(n^{1.14} d + n \log n)$
		Span (Centroid)	0.01×	$O(nd)$
		Isocontour Radius	0.01×	$O(nd)$
		Diameter	0.99×	$O(n^2 d)$
		Bottleneck	0.99×	$O(n^2 d)$
		Span (Medoid)	1.03×	$O(n^2 d)$
		Sum Diameter	1.03×	$O(n^2 d)$
	🔵	Sum Bottleneck	1.03×	$O(n^2 d)$
		Cluster Inertia	1.17×	$O(nkd \cdot \text{iters})$
		Log-Determinant	0.55×	$O(n^3 + n^2 d)$
Distribution		Vendi Score	0.03×	$O(nd^2 + d^3)$
		DCScore	0.16×	$O(n^2 d)$
		Rényi Kernel Ent.	0.20×	$O(n^2 d)$
		Bins Entropy (UMAP)[†]	1.93×	$O(n^{1.14} d)$
		MagArea	108.8×	$O(n_{ts} \cdot n^3)$
Graph	🔴🟢	Graph Entropy	1.13×	$O(n^2 d)$
	🔴🔵	MST Dispersion	3.49×	$O(n^2 d)$

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 (`iters` $\leq$ 300). Coloured dots mark the recommendations for different components: blue for variety, green for balance and red for disparity.

code reviewed and tested before use. We also used Claude for rephrasing and grammatical correction during writing.