EVALUATING TEXTUAL DIVERSITY METRICS WITH CONTROLLED MIXTURES OF DIVERSE PERSPECTIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

The prevalence of large language models (LLMs) as tools for information access and creativity necessitates the development of rigorous methods to assess the diversity of LLM-generated content in open-ended contexts. While many diversity metrics have been proposed in the literature, incorporating both varied representations of text and methods for measuring diversity, we identify a significant gap between existing abstract studies of diversity, and the practical realities of measuring diversity in text. Drawing on the interdisciplinary framework proposed by Stirling, which decomposes diversity into three necessary dimensions (variety, balance, and disparity), we add a fourth requirement specific to open-ended natural language, invariance to composition, and operationalize the framework to develop a generalizable procedure for validating proposed diversity measurements of natural language grounded in real datasets of diverse text. Concretely, we: (1) Demonstrate how to generate controlled mixtures of diverse real-world, open-ended text from existing datasets with processes that make the four fundamental diversity criteria directly accessible; and (2) Empirically measure and statistically test which diversity metrics correspond with the underlying conceptual diversity. Our evaluation bridges the gap between theoretical notions of diversity and empirical work in natural language processing, providing grounding for diversity measurement in open-ended text and a methodology for validating future proposed metrics.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as primary tools for information access, idea exploration, and creative tasks (Trippas et al., 2024; Jiang et al., 2025; Moon et al., 2025; Xu et al., 2025; Shaer et al., 2024), and have also been integrated into search engine results for millions of users.[1] When users seek recommendations or explore perspectives on a controversial topic, the diversity of the responses directly shapes what users discover and consider Hao et al. (2026).

In open-ended generation settings such as brainstorming and exploration, it is the diversity of the underlying ideas and perspectives expressed which matters most. However, existing metrics for output diversity, which span lexical, semantic, and entailment-based approaches (Zhu et al., 2018; Shaib et al., 2025; Reimers & Gurevych, 2019; Stasaski & Hearst, 2022), often operate on features of the text, measuring variation in word choice, n-gram overlap, or embedding distance between responses. It is unclear whether these metrics actually capture substantive variation at the conceptual level.

Beyond natural language processing and machine learning, disciplines such as ecology have long studied diversity over discrete, distinct items. Stirling (2007) proposed an interdisciplinary framework for characterizing diversity over abstract items that identifies three necessary, but individually insufficient, dimensions. These include *variety* (how many distinct types are present), *balance* (how evenly they are distributed), and *disparity* (how different the types are from each other). Figure 1 illustrates these attributes.

Our goal is to bridge the gap between the underlying concepts in a set of texts and the natural language expression that contains them. First, we extend Stirling's framework to include a fourth requirement, specific to text: invariance to surface-level structural *composition*; such as whether

[1] https://www.search.google/ways-to-search/ai-overviews/

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

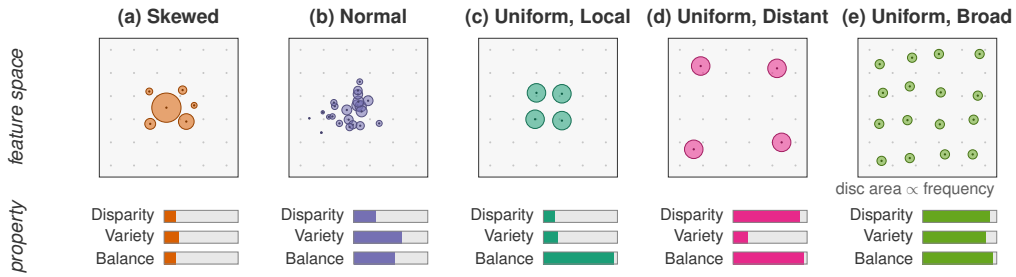


Figure 1: Three fundamental components of diversity. *Disparity* reflects the difference between items, *Variety* reflects the quantity of distinct items, *Balance* reflects the distribution of the items.

diversity exists across or within sentences, passages, or independent text sets. Then, we introduce a meta-evaluation framework to empirically test whether existing diversity metrics are actually sensitive to these four dimensions when the underlying concepts are known *a priori*. To achieve this, we leverage existing datasets of hierarchically grouped diverse text to construct text sets with controlled diversity properties, and we determine whether diversity metrics respond appropriately.

Our results indicate that many widely-used diversity metrics fail to reflect all of the desired criteria simultaneously, though the Vendi Score (Friedman & Dieng, 2023) shows promise likely owing to its ability to essentially count the latent modes in a space, reflecting simultaneously the variety of modes, the balance between them, and the disparity of the distribution.

Additionally, we present this meta-evaluation as a reusable framework with open-source code [2] and modular experiment design for validating future metrics against these four diversity dimensions.

2 Fundamental Properties of Diversity

Measuring diversity has attracted interest in a wide variety of domains, including ecology (Pielou, 1969), economics (Gini, 1912; Rhoades, 1993), and information retrieval (Carbonell & Goldstein, 1998). Much of the historical work on diversity assumes that diversity is measured over a set of discrete and distinguishable items. We recognize a central tension between that body of work and the emerging landscape of diversity measurement for text and LLMs, where measurements often have highly specific ad-hoc goals tailored to texts which are complex, compositional, and lack clear distinguishability. Despite the complexity of the content, strategies often rely on generic measurements such as embedding similarity or lexical overlap.

Generality is certainly a desirable property of diversity, so we aim to bridge this gap by treating natural language as a surface form induced over some set of latent concepts. We identify essential abstract properties of diversity, which are directly measurable over the latent concepts but not the surface form. We then generate sets of text where the underlying properties are both controllable and measurable, and empirically test the *sensitivity* of diversity metrics, computed over the surface forms, to the essential properties as measured over the corresponding latent concepts – ie, whether diversity metrics applied to the surface form can recover or distinguish the latent properties.

2.1 Formal Definitions of Latent Properties

In this section, we go through each of the four diversity properties and formalize them. Table 1 contains a summary of these formalizations.

Let $S = \{s_1, s_2, \ldots, s_N\}$ be a response set consisting of N sentences. Each sentence s_i is a surface form associated with a set of latent concepts $C_i = \{c_{i1}, c_{i2}, \ldots, c_{iN}\}$. Let $\mathcal{C} = \bigcup_{i=1}^{N} C_i$ denote the universe of latent concepts appearing in S. And following the notation of Stirling (2007), we denote a diversity function $\Delta(S)$. For a concept $c \in \mathcal{C}$, define its frequency in the response set as $f(c) = \sum_{i=1}^{N} \mathbf{1}[c \in C_i]$. Further, we define the latent conceptual disparity between two concepts $c_a, c_b \in \mathcal{C}$ by a distance function $d(c_a, c_b) : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_{\geq 0}$, where larger values indicate greater semantic or conceptual divergence.

[2] https://anonymous.4open.science/r/DiversityEvaluationAnonymous-52E5

Property	Formalization	Example
Quantity	$\Delta(S) \propto \lvert\mathcal{C}\rvert$	$\Delta(A,B,C,D) > \Delta(A,A,A,A)$
Balance	$\Delta(S) \propto E\big(\{f(c)\}_{c\in\mathcal{C}}\big)$, E is Schur-concave	$\Delta(A,B,A,B) > \Delta(A,A,A,B)$
Disparity	$\Delta(S) \propto \sum_{a,b} w_{ab}\, d(c_a, c_b)$, for weight w.	$\Delta(A,Z,A,Z) > \Delta(A,B,A,B)$
Composition	$\phi(S) = \phi(S') \rightarrow \Delta(S) = \Delta(S')$	$\Delta(A,E,A,E) = \Delta(A,E,AE)$

Table 1: Formal properties of concept diversity. A diversity function Δ, computes the diversity score of a set S. We define $\mathcal{C}$ as the set of distinct concepts represented in S, $f(c)$ as the frequency of a concept in S, and $\phi(S) = \{f(c)\}_{c\in\mathcal{C}}$, an oracle function which extracts the set of concepts in S.

Quantity Following Stirling's *variety* principle, an effective diversity metric should be sensitive to the quantity of concepts contained within the set. See Table 1 for a formal expression of the property. A key distinction between this abstract notion and the realities of measuring diversity in text is that, in practice, the notion of a textual unit within a set of texts may not align exactly with the underlying conceptual units present across the set of texts. A text may contain multiple concepts, and a concept may occur in multiple texts. We want diversity metrics which reflect the quantity of underlying concepts, regardless of how many times they appear or how many concepts are expressed in a text.

Assuming knowledge of the underlying concepts, we operationalize Quantity (Table 1) as the count of the set of concepts.

Balance Another important criterion, also identified by Stirling, is the balance or evenness of the concepts. An effective diversity metric for text should be sensitive to not only the absolute number of concepts, but also the relative frequencies by which the concepts are expressed: the skewness or the evenness of the concepts across the set. Even if a set nominally contains multiple concepts, they may be represented in varying proportions that can influence the downstream interpretation. Consider for instance the following responses to the query "Where should I travel this summer?": *Italy. Naples. The Amalfi Coast. Paris. Positano.* On the surface, multiple recommendations are present, but in practice a diversity metric that neglects the balance will provide no indication that the overwhelming majority suggests a trip to the Amalfi coast. For a user-facing system performing open-ended generation, imbalanced output could result in markedly different experiences across user sessions, which – especially in potentially sensitive information-seeking contexts - may or may not be the desired outcome. A diversity metric ought to quantify this.

Assuming knowledge of the underlying concepts, we operationalize Balance (Table 1) as the output of a Schur-concave function over the counts of the concepts. A common choice for this function is Pielou's evenness (Pielou, 1969), which is effectively the normalized entropy:

$$E = \frac{H}{\ln(S)} \quad \text{where } H = -\sum_{i=1}^{S} p_i \ln(p_i) \tag{1}$$

Degree of Difference Though often difficult to define, Stirling's framework also considers *disparity*, or the degree of difference between the items in the set, to be a central aspect of diversity. This is especially important in natural language, where two concepts or sentences that are nominally distinct may actually be very similar under some notion of conceptual distance; other concepts may be genuinely far apart. Continuing the example of travel recommendations, consider two different response sets: (A) *Marseille. Paris. Lyon*; vs (B) *Guangzhou. Paris. Yosemite*. While both contain completely distinct items, the second set spans a much wider conceptual range. Diversity metrics should account for this property, reflecting the breadth of coverage within a conceptual space.

Operationalizing the degree of difference is challenging, as it requires both a defined distance function between all pairs of concepts, and a systematic way to combine the pairwise distances to produce an aggregate disparity measure.

Compositional Invariance In addition to the preceding three positive principles, we introduce a fourth criterion specific to natural language: namely, that the diversity of a set of texts should be invariant to the structural composition of the concepts. The same ideas, whether expressed across multiple sentences or packed into a single sentence, should yield the same diversity measurement. Ideas should not be hidden by their inclusion alongside others, nor should a metric incorrectly measure re-combinations of redundant ideas as something new. This criterion is notably distinct from one that might exist in economics: if two firms addressing different markets merge, it would be appropriate to consider the competitive landscape more highly concentrated, as those firms are unlikely to operate independently.

3 Diversity in NLP

Recently, textual diversity has become an important concern in natural language processing research, with relevance to both the social impacts of AI-generated content, and the utility of higher diversity in reasoning and synthetic data applications. Unlike in fields where the units over which diversity is measured are well-defined and clearly distinct (such as species or firms), diversity in open-ended LLM output text lacks clear boundaries between texts which may express multiple or complex concepts.

3.1 Diverse views of Diversity

There are multiple strands of LLM research addressing different aspects of model response diversity and heterogeneity. Some work focuses on *pluralism*, the extent to which model responses reflect or represent the distribution of human perspectives, and the variety of ground-truth topics a model is capable of discussing, including moral norms (Benkler et al., 2023), cultural alignments (Tao et al., 2024), political opinions (Feng et al., 2023) and subjective stances (Hayati et al., 2024). Other work focuses on *creativity*, assessing whether models recycle similar ideas, structures, and topics in open-ended generation (Moon et al., 2025; Xu et al., 2025; Jiang et al., 2025), or offer novelty compared to other methods like search (Wright et al., 2025).

A relevant distinction here, perhaps a consequence of overloaded terminology, is whether diversity is studied as an intrinsic quality, measured as the variation within a set of texts, or extrinsically, measured as representativeness with respect to reference responses or a reference distribution. In this work, we focus on diversity as an intrinsic property of a set of texts.

3.2 Diversity Quantification

Diversity in generated text has been measured using a variety of metrics, ranging from lexical and semantic similarity measures to methods that group outputs into idea-level categories, as well as approaches that quantify coverage or distributional spread.

Traditional metrics focus on lexical-level variation, capturing differences in surface-level wording rather than the underlying ideas. These include distinct n-gram counts (Li et al., 2016), variations on BLEU (Zhu et al., 2018), compression ratios (Shaib et al., 2025), Levenshtein distance (Levenshtein, 1966), and the Type-Token Ratio (TTR).

Semantic metrics look beyond verbatim words to capture the underlying ideas. These use cosine similarity over contextualized vector representations such as token-level BERTScore (Zhang et al., 2020) and sentence-level SBERT (Reimers & Gurevych, 2019) to measure the distance between generated outputs in a latent semantic space. Several researchers have integrated embedding-based distances into broader diversity frameworks (Du & Black, 2019; Lai et al., 2020). Several recent studies have applied these metrics to evaluate diversity in LLM-generated text (Tevet & Berant, 2021; Reviriego et al., 2024; Kirk et al., 2024; Shypula et al., 2025; Guo et al., 2025; Padmakumar & He, 2024).

Embedding-based similarity can lack nuance, so another emerging approach is to focus on the underlying concepts by grouping text into equivalence classes of similar ideas Xu et al. (2025); Zhang et al. (2025b); Farquhar et al. (2024). Other work focuses on coverage, assessing not just internal diversity but the extent to which a model can surface the full range of relevant ideas (Hayati et al., 2024; Liu et al., 2026). This stream of research most closely approaches the abstract goal

of measuring the latent diversity in a set of items, but can struggle to model nuanced differences between items, as reflected by Stirling's disparity criterion.

3.3 A General Framework of Diversity Metrics

Here, we cover a variety of existing and popular diversity metrics for text. Every metric we consider can be understood along two largely independent axes: (1) the *feature space* in which texts are compared, some of which are similarity functions $f(s_i, s_j)$ between texts, others of which are distance functions $d(s_i, s_j)$; and (2) the *aggregation strategy* by which a matrix F of pairwise similarities (or D of distances) across texts is reduced to a single diversity score for the response set: $g(D \text{ or } F) \in \mathbb{R}$, where $D \in \mathbb{R}^{n \times n}$.

3.4 Feature Spaces

The feature space encodes the aspect of the text that is used to quantify conceptual overlap. In general, the feature space induces a pairwise function satisfying the principles of non-negativity $f(a, b) \geq 0$ and symmetry $f(a, b) = f(b, a)$. We consider three feature spaces commonly employed in the literature on textual diversity: lexical, semantic, and mutual entailment.

Lexical features. Lexical features encode surface-level textual properties, essentially which words are present in the input. We operationalize lexical similarity for diversity measurement as either BLEU score Papineni et al. (2002) or the Levenshtein distance Levenshtein et al. (1966). Though limited in its ability to reflect conceptual similarities across vocabulary differences, lexical approaches can be fast to compute, interpretable, and effective in restricted response spaces, capable of revealing surprisingly high degrees of homogeneity in LLM responses (Bhatt & Diaz, 2024).

Semantic features. Semantic features compare texts at the semantic level, replacing a lexical overlap metric with the distance between vector embeddings of the text. These embeddings can be computed using a pretrained sentence encoder (e.g., Sentence-BERT, Reimers & Gurevych, 2019). Given two embeddings, we define the similarity function as the cosine similarity, $\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \, \|\mathbf{y}\|}$, and the distance function as the cosine distance $\frac{1}{2}(1 - \cos(\mathbf{x}, \mathbf{y}))$ These embeddings are more likely to reflect the conceptual similarity of texts independent of their lexical relationship.

Entailment features. Semantic equivalence, increasingly used for diversity quantification Stasaski & Hearst (2022); Zhang et al. (2025b) and uncertainty quantification (Farquhar et al., 2024) of texts can also be defined by a natural language inference (NLI) model which produces a label of *entailment*, *contradiction*, or *neutral* for a pair of input texts. Two texts are considered equivalent if an NLI model predicts mutual (bidirectional) entailment between them; this in turn creates a partition of the input set S into equivalence classes, as such it is only compatible with the clustering aggregation strategy below. These equivalence classes are clearly interpretable and map nicely to the abstract notion of distinguishable items, but may not reflect graded similarities and may introduce a heavy computational requirement at scale as every input text may require decomposition, and every pair of texts requires a full forward pass of an NLI model to evaluate.

3.5 Aggregators

Once a distance matrix $D_{ij} = d(s_i, s_j)$ for a set of items has been computed, an aggregator g reduces this matrix of pairwise comparisons to a single diversity score. In this subsection we discuss the four common aggregation strategies investigated. While other aggregation strategies have been proposed, we focus on the primary methods that have been used for analysis of diversity in text.

Mean Distance. Similarity scores can be aggregated by calculating an average over their pairwise distances. A common approach is to measure the *mean pairwise distance* (MPD), the average distance between each pair of items. This measure increases as the dissimilarity between the items increases, and is sensitive to dissimilar outliers but unable to penalize redundancy. Another approach is to measure the *mean neighbor distance* (MND), the mean of each item's distance to its nearest neighbor. This metric in particular penalizes redundant items; but it is fundamentally limited in that it cannot capture the global structure of the concept space.

Aggregator	Range	Formula
Mean Pairwise Distance (MPD)	[0, 1]	$\frac{2}{\lvert S\rvert(\lvert S\rvert-1)} \sum_{i<j} d(s_i, s_j)$
Mean Neighbor Distance (MND)	[0, 1]	$\frac{1}{\lvert S\rvert} \sum_i \min_{j\neq i} d(s_i, s_j)$
Number of Classes (NC)	$\mathbb{Z}_{\geq 1}$	$\lvert S/\sim\rvert, \quad s_i \sim s_j \iff 1 - d(s_i, s_j) \geq \theta$
Vendi Score (VS)	$[1, n]$	$\exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right), \quad \lambda_i$ eigenvalues of D

Table 2: Aggregation methods for a matrix D of distance scores.

Partitions. Another common approach for measuring diversity is to use the feature space to partition the items into distinct equivalence classes using a similarity threshold θ. In practice the number of classes (NC) is then counted to obtain a diversity score for the set.

Vendi Score. The von Neumann diversity (Vendi) score (VS), introduced by Friedman & Dieng (2023), computes the exponential of the entropy over the eigenvalues of the similarity matrix. Conceptually, the Vendi score reflects the number of conceptual modes in the space implied by the similarity kernel.

4 Evaluations

We conduct three experiments, each testing the sensitivity of the diversity metrics to fundamental properties of diversity. Leveraging existing datasets which contain labeled diverse perspectives, we artificially construct text sets as mixture of distinct concepts with different properties. This provides sets of texts where the ground truth diversity properties are known and measurable.

4.1 Datasets

The Perspectrum dataset (Chen et al., 2019), introduced for the task of perspective decomposition, contains over 900 debatable claims and corresponding human written perspectives on the claims. In Perspectrum, each perspective is associated with a **claim** (e.g.: *"A government has the obligation to lessen the economic gap between its rich and poor citizens"*), a list of **topic** labels (e.g.: *['politics', 'economy', 'society']*), a **stance** (*SUPPORT* or *UNDERMINE*), and a perspective **cluster** grouping human-authored sentences by the perspective they convey. Perspective clusters in Perspectrum were formed by sourcing multiple human-authored responses on the same claim, grouping them by the perspective they embody, and were manually verified.

The DiverseSumm dataset (Huang et al., 2024) covers 245 distinct news stories of 10 articles each, with each story associated with 2.5 questions and each question with 3.4 answer clusters. This structure provides our conceptual hierarchy, where each question reveals a latent concept that unifies the answers, but each answer may offer a different perspective on the question. In DiverseSumm, data was generated by GPT-3.5-Turbo, and validated by human annotators.

In Perspectrum, we treat each perspective group as a distinct "concept." Each human-written perspective sentence is thus a distinct instantiation of the concept. In DiverseSumm, each event has multiple questions and each question has multiple answer groups. We treat the answer groups as the distinct "concepts," and each answer as an instantiation of the concept.

4.2 Quantity and Balance

Our first validation experiment measures the sensitivity of each metric to the number (quantity) and proportion (balance) of latent concepts.

Simulation We sample latent concepts from a finite Dirichlet process, then instantiate each sampled concept as a randomly-chosen sentence corresponding to the concept, and measure the diversity of

the resulting sets of text. We operationalize this simulation using the Pólya urn model (Hoppe, 1984) of the Dirichlet process. Let $\mathcal{C} = \{1, \ldots, K\}$ be a fixed set of latent concepts, where each concept k is associated with a pool $\mathcal{S}_k$ of sentences expressing it. We begin with an urn containing, for each concept k, an initial mass of c/K, where c is the concentration parameter. At each step, we draw a concept with probability proportional to its mass:

$$P(z_t = k \mid z_1, \ldots, z_{t-1}) = \frac{c/K + n_k^{(t)}}{c + t - 1}, \tag{2}$$

where $n_k^{(t)} = \sum_{i<t} \mathbb{1}[z_i = k]$ is the number of times concept k has been drawn so far. In this way, as we proceed, we reinforce the probability of concepts that have already been drawn. When c is low, the process primarily samples from a single latent concept; when c is high, many latent concepts are sampled. We run 300 sampled sets, of size up to 10, and varying the concentration parameter $c \in \{0.1, 0.3, 1.0, 3.0, 10.0\}$, which provides a variety of expected concentrations.

Sensitivity Analysis. The concentration parameter modulates both quantity and balance of concepts: the higher the parameter, the more latent concepts are sampled, and the more uniform the sample is over those concepts. But these effects are not independent; the quantity of concepts K depends on the number of samples and the concentration parameter, and the balance of concepts E depends on the number of samples, the concentration parameter, and the quantity of concepts. To isolate the effect of each property on a diversity metric, we regress the metric's score on K, operationalized as the raw count of concept groups sampled, and E, measured as Pielou's evenness (see Equation 1). The coefficient for K is thus the metrics sensitivity to quantity at fixed balance; and the coefficient for E is the sensitivity to balance at fixed quantity.

The table in Figure 2 shows the estimated effect size of each of the features on the final diversity score, evaluated using the Perspectrum perspective groups as the concepts. Appendix B contains the table for DiverseSumm. We summarize key findings below.

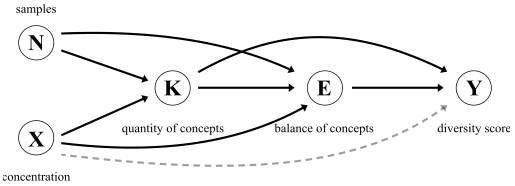


Figure: Our proposed model of the effect of the Dirichlet concentration parameter on Δ.

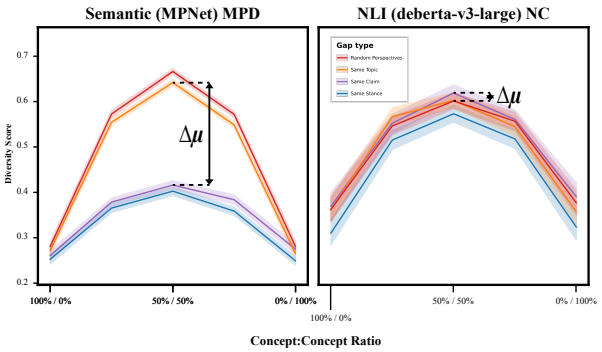
Similarity Function	Aggregation	X	K	E	N	R^2
BLEU	MPD	+0.018	+0.337**	+0.215**	-0.189**	0.17
	MND	+0.017	+0.308**	+0.133**	-0.630**	0.34
	NC	-0.000	+0.005	+0.008*	+0.992**	0.99
NLI (DeBERTa-v3-Large)	NC	+0.022*	+0.247**	+0.001	+0.720**	0.78
NLI (RoBERTa-MNLI)	NC	+0.020*	+0.260**	+0.000	+0.709**	0.78
Semantic (MPNet)	MPD	+0.023	+0.416**	+0.291**	-0.256**	0.27
	MND	+0.001	+0.402**	+0.139**	-0.803**	0.52
	NC	+0.004	+0.499**	+0.019	+0.393**	0.62
	VS	+0.009	+0.404**	+0.052**	+0.601**	0.79
Semantic (Qwen3-8B)	MPD	+0.025	+0.430**	+0.291**	-0.268**	0.29
	MND	+0.004	+0.416**	+0.145**	-0.817**	0.54
	NC	+0.001	+0.657**	+0.020	+0.165**	0.58
	VS	+0.010	+0.417**	+0.063**	+0.591**	0.80

Table (right): Estimated coefficients for the direct (X) and indirect (K, E) effects. Significance markers denote thresholds (* $p < 0.05$, ** $p < 0.01$).

Figure 2: Causal mediation model (left) and regression results (right), for the Perspectrum dataset. Causal mediation results for DiverseSumm are included in Appendix B

Finding: Many metrics are sensitive to the variety of concepts Perhaps surprisingly, NLI-based metrics are not more sensitive to the variety of concepts than semantic / embedding based approaches. We find that embedding-based metrics, and most strongly an embeddinb-based clustering metric, are sensitive to the number of underlying concepts. This suggests that using clustered embeddings to approximate the latent concepts is a relatively promising approach for diversity quantification, and perhaps superior to simply applying NLI models to the data. Model quality also appears to matter, with the largest and most advanced embedding model outperforming the smaller embedding model on the clustering-based diversity metric.

Finding: Few metrics are sensitive to the balance of concepts Measuring a fair balance across concepts appears to be much harder, with few metrics having large coefficients. Even the best performing approach for measuring the balance of concepts, the mean pairwise distance aggregator applied to embeddings, has a relatively low coefficient.



Feature Space	Aggregation	rand - topic $\Delta\mu$	topic - claim $\Delta\mu$	claim - stance $\Delta\mu$
lexical (BLEU)	MPD	*0.00838*	***0.015***	*0.0115*
	MND	*0.0216*	-0.0187	***0.0281***
	NC	0.01	-0.0179	*0.0415*
lexical (Levenshtein)	MPD	*0.00561*	***0.0147***	*0.00703*
	MND	0.0125	-0.00712	*0.0212*
	NC	0.04	-0.063	0.0298
semantic (mpnet)	MPD	***0.0244***	***0.226***	*0.0134*
	MND	0.00809	*0.0177*	*0.0169*
	NC	0.014	***0.227***	*0.116*
	VS	*0.0432*	***0.388***	*0.0601*
semantic (qwen3-8b)	MPD	*0.00876*	***0.134***	***0.0247***
	MND	0.00625	0.00658	*0.0157*
	NC	0.006	***0.243***	***0.207***
	VS	*0.0292*	***0.289***	***0.0872***
nli (roberta-mnli)	NC	0.056	-0.0801	*0.11*
nli (deberta-v3-large)	NC	-0.004	-0.0521	*0.147*

Figure 3: Example plots for Perspectrum showing the effect of the concept ratio on diversity score (left), and a table summarizing the hierarchical difference in means between perspective gap types (right). Full set of plots in Appendix C and Appendix D. Holm-adjusted one-sided p-values < 0.05 in italics, < 0.01 in bold italics.

4.3 Disparity

The next validation experiment measures whether diversity metrics are sensitive to the disparity of concepts. Under the third criterion, a set of texts containing distant concepts ought to be measured as more conceptually diverse than a set of texts composed of related concepts.

Simulation We create synthetic concept sets by interpolating between two distinct perspective groups. Given two groups of n perspectives each, we create a synthetic group with $k \in \{0, \ldots, n\}$ perspectives from the first group, and $n - k$ perspectives from the second group. We expect that the diversity of groups where k is close to $\frac{n}{2}$ to exhibit the most diversity. To test whether the metrics are sensitive to the distance between groups, we exploit the structure of the dataset to create pairs of concept groups that exhibit different levels of similarity. Paired concept groups can be chosen from different distances from one another: In Perspectrum, the same stance on the same claim (**stance**), different stances on the same claim (**claim**), the same topic category (**topic**), or completely at random across the entire dataset (**rand**). In DiverseSumm, different answers in the same answer group, different answers for the same question, and different answers for the same event all provide distinct disparities. In our experiments, we set $n = 4$ and at each gap-disparity sample 500 pairs of concept groups.

Finding: Diversity scores form an arc-shaped curve Across all settings, we see that diversity at the extrema (when $k = 0$ or $k = n$) converges; this is expected because we are only sampling from a single perspective cluster. We see that across all metrics, the score increases as the perspective clusters are more mixed when there is a 50:50 balance of perspective groups. See Figure 3 for a selected pair of curves, and Appendix C and Appendix D for the full set.

Finding: Most metrics are insensitive to disparity Across all metrics (See Appendix C), we perform single-tailed t-tests (Holm-adjusted for multiple testing) according to the *a priori* hierarchy of expected disparities: that rand $>$ topic $>$ claim $>$ stance. For each adjacent pair in the hierarchy, we consider the hypothesis that the mean of diversity scores for the more disparate pairing, at a 50/50 blend, is higher than the mean of the less disparate pairing. We find that across metrics, the null hypothesis is rarely rejected. The table in Figure 3 shows a summary of the differences in means.

4.4 Composition

In natural language, concepts are not expressed in isolation. A passage, or even a sentence, may contain multiple concepts composed together. This decomposition poses a practical challenge for the application of diversity metrics at scale, potentially requiring fine-grained decomposition of any input text.

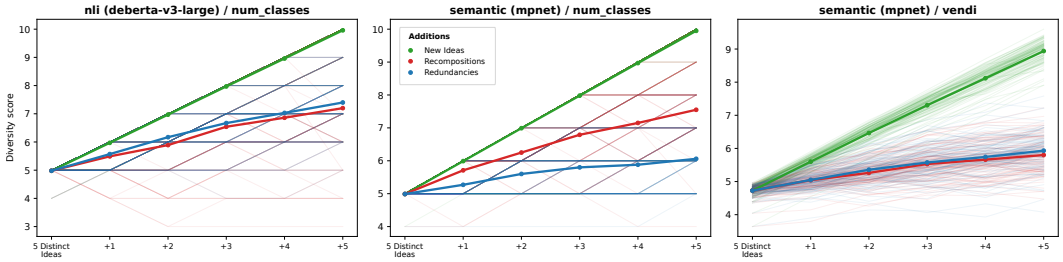


Figure 4: A selection of composition-sensitivity plots for Perspectrum. At the origin, y represents the diversity score of a set consisting of five distinct perspectives. New sentences are added to the set (either containing a brand new perspective, a paraphrase of an existing perspective, or a concatenation of such paraphrases). Hairlines indicate each of 100 trials; the bold lines indicate the median result.

Simulation We test the robustness of diversity metrics to the structural composition of the input by measuring the sensitivity of the metric to the recombination of redundant concepts. For each trial, we begin with a set of 5 conceptually sentences and measure the diversity score of that set. Then, sequentially, we take two of the existing concepts within the set, concatenate concept-equivalent sentences (sourced from the dataset) into a single text, and add this text to the set. We also compare to two baselines: adding a genuinely conceptually new sentence at each step, and adding a single redundant sentence, to isolate whether the metric is sensitive to the inclusion of any new texts, redundant perspectives, or simply re-composed redundant perspectives. If a metric is sensitive to the composition of perspectives, we expect to see the diversity increase as we add newly composed (yet redundant) concepts.

Finding: Sensitivity to redundancy and recomposition varies significantly across methods. The NLI-based equivalence metric (left) overestimates diversity by treating both redundant and recomposed ideas as distinct, resulting in a relatively high growth rate under both the recomposition and the redundant settings. In contrast, the embedding-based clustering metric (center) partially distinguishes recompositions from redundancies by demonstrating low growth rate as redundant ideas are added, while the Vendi Score (right) shows low sensitivity to both due to its reliance on global embedding variance rather than fine-grained structure. Appendix E and Appendix F contain the full set of plots across all diversity functions and datasets.

5 Conclusion

In this work, we introduced a novel technique to validate whether metrics of textual diversity capture properties of the underlying conceptual diversity. Our methodology: (1) Identifies and operationalizes a set of central diversity properties (Stirling, 2007) that act as criteria for a diversity metric; (2) Generated controlled mixtures of human-authored perspectives using Perspectrum (Chen et al., 2019); and (3) Measures the sensitivity of diversity metrics to the latent properties of the generated mixtures.

First, this procedure surfaces gaps in the ability of some commonly used diversity metrics to reflect underlying conceptual diversity in sets of text. We find that popular diversity metrics are often ineffective at capturing multiple fundamental properties of diversity in the text input, particularly in settings where the number of texts or concepts is variable or indeterminate. Some widely used approaches, such as counting clusters or equivalence classes of the input text, fail to consistently reflect the conceptual disparity between the resulting groups themselves, or distinguish between new ideas and redundant texts, or recompositions of the same ideas.

Second, our work establishes a reusable framework for evaluating future diversity metrics. The proposed evaluation procedure provides a way to systematically test whether a metric is sensitive to quantity, balance, disparity, and composition under controlled latent concept manipulations. We aim for this method to serve as a practical guide when developing diversity metrics, and to help future work identify which properties a newly proposed diversity metric is sensitive to, and where it may fail to align with conceptual diversity.

LIMITATIONS

While our work introduces a novel and broadly generalizable framework for principled evaluation of diversity metrics, and evaluates against an exhaustive cartesian product of feature spaces and diversity aggregation functions, there are myriad possible extensions to this work which cover specific targeted domains. These extensions are not addressed, but are an exciting opportunity to use this new methodology should validated diverse text data be available in other areas. Our findings intentionally use structured and validated perspective datasets, and while this allows for reliable experimentation, it also limits generalization to all possible settings of text. Nonetheless, we view this approach as a significant step toward more principled and grounded evaluation of diversity metrics.

REPRODUCIBILITY

We strive to make these results as reproducible as possible. All parameters used for experiments are explained in the main text of the paper – these are primarily choices for the number of samples taken as part of our controlled data mixtures. We have also produced a code repository that makes it easy to run the full pipeline of the project, including all the simulations and experiments. An anonymized version of the repository is available at: https://anonymous.4open.science/r/DiversityEvaluationAnonymous-52E5

AI USAGE

We used LLM-based AI systems in three ways during the research and writing process. First, "Deep Research" LLM chatbots were used as part of the literature review, to help identify relevant work across fields. Second, LLM-based coding agents were used to refactor human-written software and package it neatly to assist with the reproducibility of the experimental results. Third, LLMs were used to assist with LaTeX formatting, specifically with writing formulas, creating tikz diagrams, and laying out figures and tables.

REFERENCES

Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. Assessing llms for moral value pluralism, 2023. URL https://arxiv.org/abs/2312.10075.

Shaily Bhatt and Fernando Diaz. Extrinsic evaluation of cultural competence in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16055–16074, 2024.

Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 335–336, 1998.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. Seeing things from a different angle:discovering diverse perspectives about claims. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 542–557, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1053. URL https://aclanthology.org/N19-1053/.

Wenchao Du and Alan W Black. Boosting dialog response generation. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 38–43, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1005. URL https://aclanthology.org/P19-1005/.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11737–11762, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.656. URL https://aclanthology.org/2023.acl-long.656/.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=g97OHbQyk1.

Corrado W Gini. Variability and mutability, contribution to the study of statistical distributions and relations. *Studi Economico-Giuridici della R. Universita de Cagliari*, 1912.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models, 2025. URL https://arxiv.org/abs/2412.10271.

Qianyue Hao, Fengli Xu, Yong Li, and James Evans. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, pp. 1–7, 2026.

Shirley Anugrah Hayati, Minhwa Lee, Dheeraj Rajagopal, and Dongyeop Kang. How far can we extract diverse perspectives from large language models? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5336–5366, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.306. URL https://aclanthology.org/2024.emnlp-main.306/.

Fred M Hoppe. Pólya-like urns and the ewens' sampling formula. *Journal of mathematical biology*, 20(1):91–94, 1984.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 570–593, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.32. URL https://aclanthology.org/2024.naacl-long.32/.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond), 2025. URL https://arxiv.org/abs/2510.22954.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL https://arxiv.org/abs/2310.06452.

Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1739–1746, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.215/.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100, 2024.

V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, February 1966.

Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pp. 707–710. Soviet Union, 1966.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yuhan Liu, Fangyuan Xu, Vishakh Padmakumar, Daphne Ippolito, and Eunsol Choi. No single best model for diversity: Learning a router for sample diversity, 2026. URL https://arxiv.org/abs/2604.02319.

Kibum Moon, Adam E. Green, and Kostadin Kushlev. Homogenizing effect of large language models (llms) on creative diversity: An empirical comparison of human and chatgpt writing. *Computers in Human Behavior: Artificial Humans*, 6:100207, 2025. ISSN 2949-8821. doi: https://doi.org/10.1016/j.chbah.2025.100207. URL https://www.sciencedirect.com/science/article/pii/S294988212500091X.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity?, 2024. URL https://arxiv.org/abs/2309.05196.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Evelyn C Pielou. An introduction to mathematical ecology. 1969.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Pedro Reviriego, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández. Playing with words: Comparing the vocabulary and lexical diversity of chatgpt and humans. *Machine Learning with Applications*, 18:100602, 2024. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2024.100602. URL https://www.sciencedirect.com/science/article/pii/S2666827024000781.

Stephen A Rhoades. The herfindahl-hirschman index. *Fed. Res. Bull.*, 79:188, 1993.

Orit Shaer, Angelora Cooper, Osnat Mokryn, Andrew L Kun, and Hagit Ben Shoshan. Ai-augmented brainwriting: Investigating the use of llms in group ideation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642414. URL https://doi.org/10.1145/3613904.3642414.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores, 2025. URL https://arxiv.org/abs/2403.00553.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the diversity and quality of llm generated content, 2025. URL https://arxiv.org/abs/2504.12522.

Katherine Stasaski and Marti A. Hearst. Semantic diversity in dialogue with natural language inference, 2022. URL https://arxiv.org/abs/2205.01497.

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

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346, 09 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae346. URL https://doi.org/10.1093/pnasnexus/pgae346.

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.25. URL https://aclanthology.org/2021.eacl-main.25/.

Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. What do users really ask large language models? an initial log analysis of google bard interactions in the wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 2703–2707, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657914. URL https://doi.org/10.1145/3626772.3657914.

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. Epistemic diversity and knowledge collapse in large language models. *arXiv preprint arXiv:2510.04226*, 2025.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *Proceedings of the National Academy of Sciences*, 122(35): e2504966122, 2025. doi: 10.1073/pnas.2504966122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2504966122.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025a.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*, 2025b.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pp. 1097–1100, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210080. URL https://doi.org/10.1145/3209978.3210080.

A Licensing Information

Artifact	Citation	URL	License
Perspectrum	Chen et al. (2019)	https://github.com/CogComp/perspectrum	CC
DiverseSumm	Huang et al. (2024)	https://github.com/salesforce/DiverseSumm	Apache-2.0
MPNet	Reimers & Gurevych (2019)	https://huggingface.co/sentence-transformers/all-mpnet-base-v2	Apache 2.0
Qwen3-embedding-8B	Zhang et al. (2025a)	https://huggingface.co/Qwen/Qwen3-Embedding-8B	Apache 2.0
RoBERTa-Large-MNLI	Liu et al. (2019)	https://huggingface.co/FacebookAI/roberta-large-mnli	MIT
DeBERTa-v3-Large-MNLI	Laurer et al. (2024)	https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	MIT

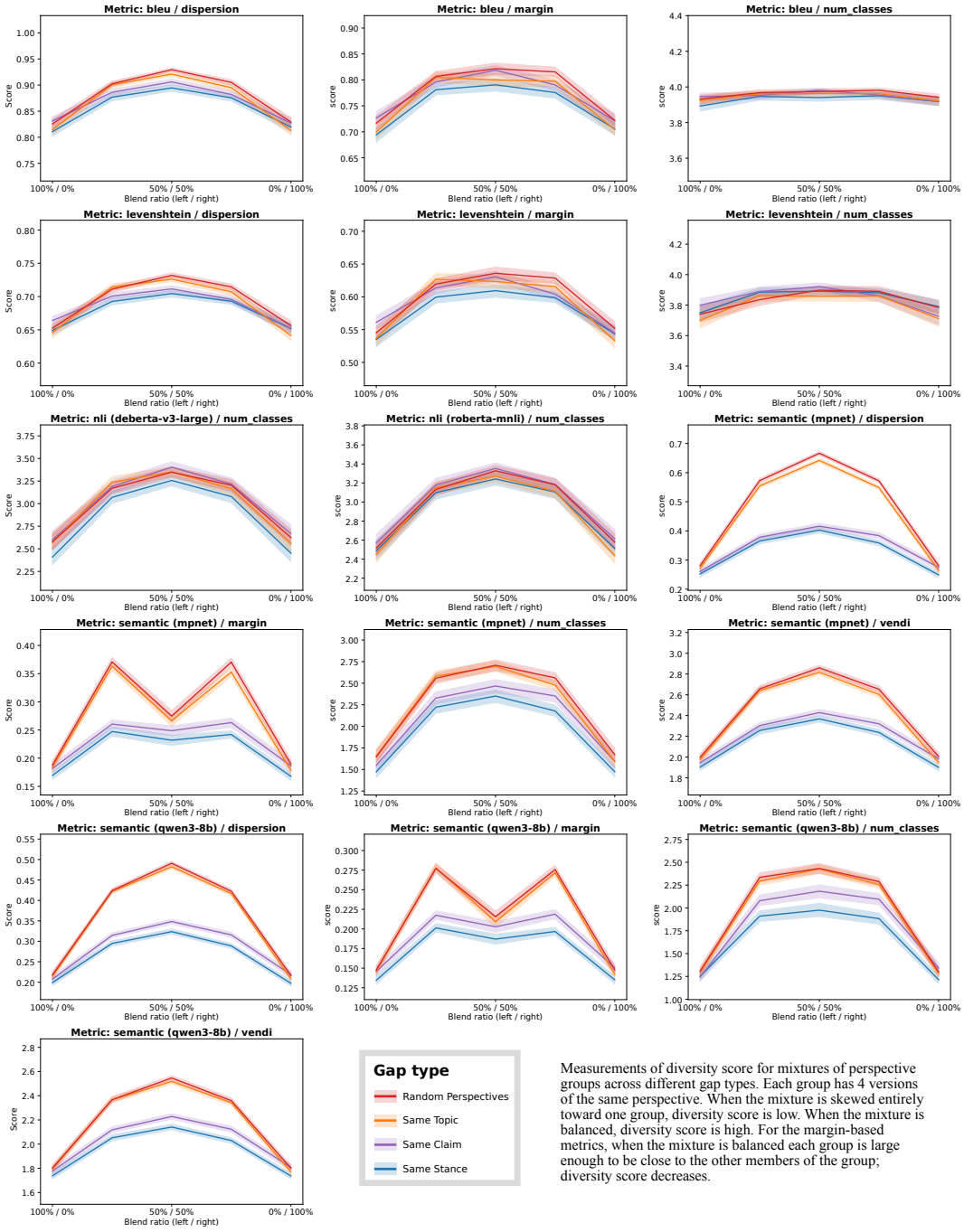
Table 3: Artifacts and Licensing Information

B SENSITIVITY EXPERIMENT - DIVERSESUMM

Similarity Function	Aggregation	X	K	E	N	R^2
BLEU	MPD	+0.059*	+0.130**	+0.090**	-0.097**	0.04
	MND	+0.045*	+0.106**	+0.055*	-0.476**	0.21
	NC	+0.003	+0.015*	+0.002	+0.978**	0.97
NLI (DeBERTa-v3-Large)	NC	+0.008*	+0.018*	+0.003	+0.970**	0.96
NLI (RoBERTa-MNLI)	NC	+0.009*	+0.016*	+0.004	+0.971**	0.96
Semantic (MPNet)	MPD	+0.015	+0.144**	+0.067**	-0.083**	0.03
	MND	+0.025	+0.139**	+0.050*	-0.531**	0.25
	NC	-0.008	+0.187**	+0.009	+0.520**	0.41
	VS	-0.001	+0.150**	+0.003	+0.748**	0.70
Semantic (Qwen3-8B)	MPD	+0.019	+0.154**	+0.114**	-0.092**	0.05
	MND	+0.025	+0.138**	+0.077**	-0.549**	0.28
	NC	+0.001	+0.229**	+0.019	+0.365**	0.28
	VS	+0.003	+0.148**	+0.025	+0.774**	0.74

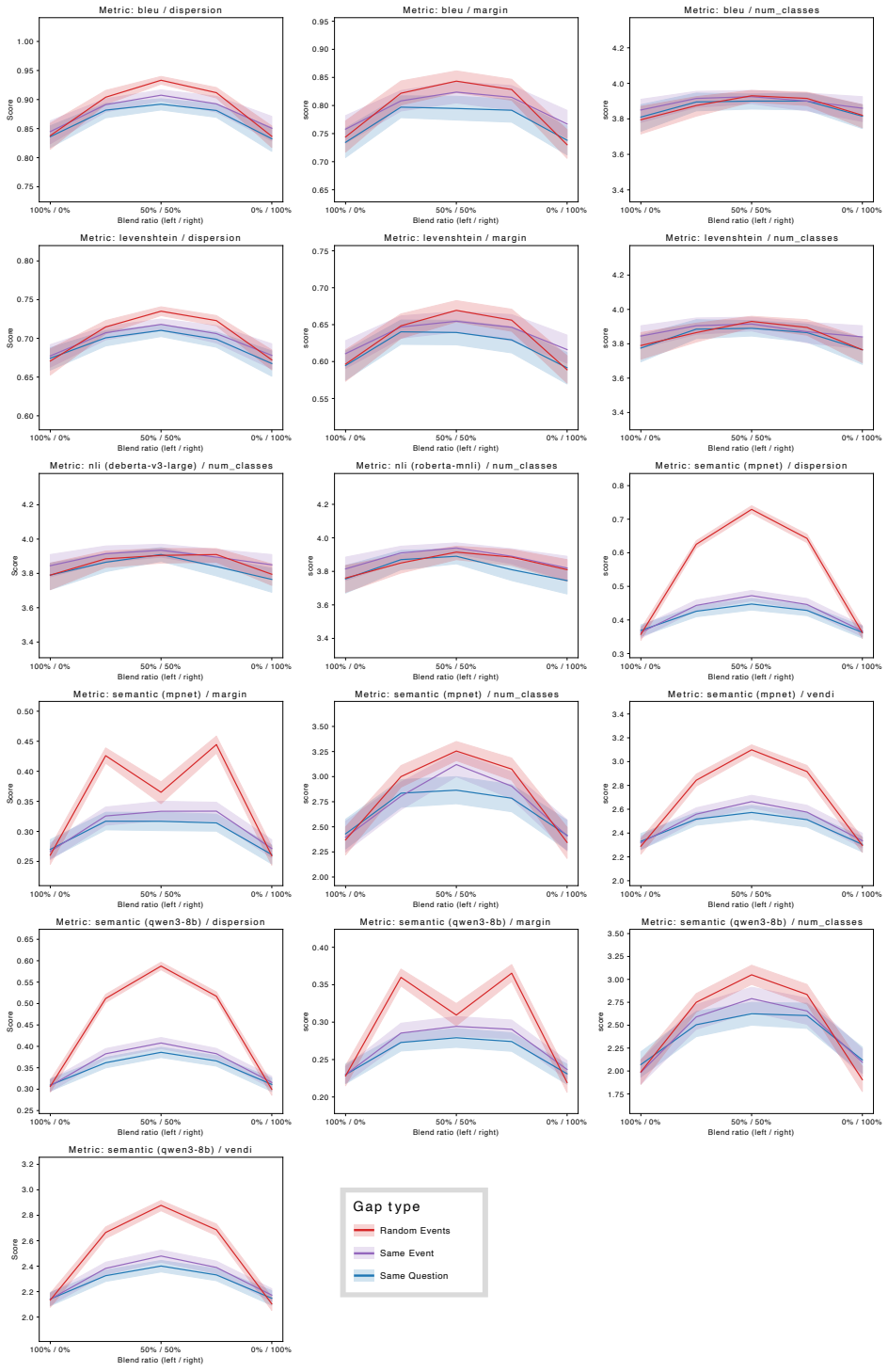
Table 4: Sensitivity to Quantity (K) and Balance (E) for the DiverseSumm dataset. Semantic metrics are much more sensitive to the quantity of the underlying concepts, but no metrics accurately recover the balance of the underlying concepts. Vendi Scores and NLI scores are both sensitive to the number of samples, and well-explained by the model.

C DISPARITY EXPERIMENT - PERSPECTRUM

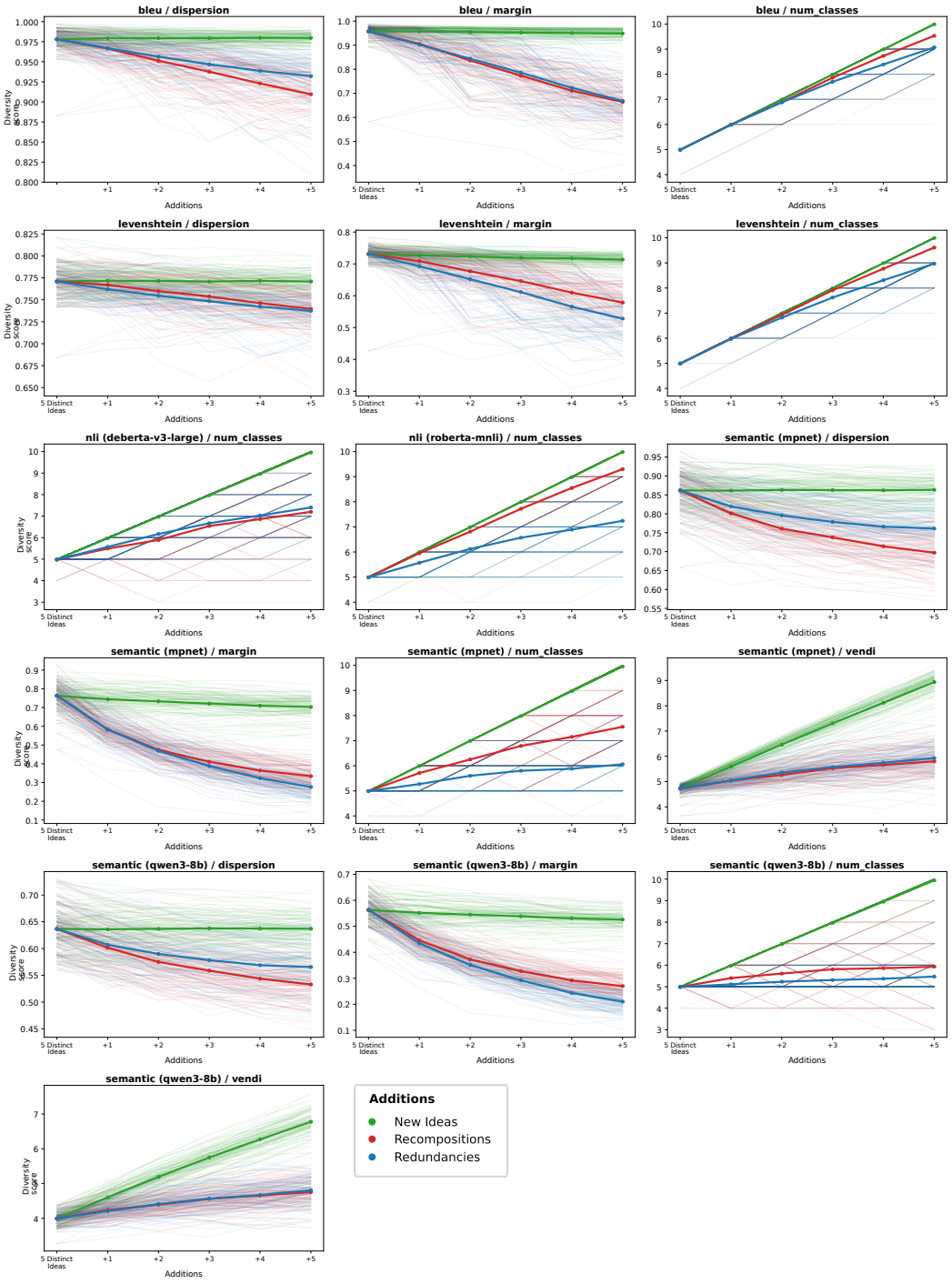


Measurements of diversity score for mixtures of perspective groups across different gap types. Each group has 4 versions of the same perspective. When the mixture is skewed entirely toward one group, diversity score is low. When the mixture is balanced, diversity score is high. For the margin-based metrics, when the mixture is balanced each group is large enough to be close to the other members of the group; diversity score decreases.

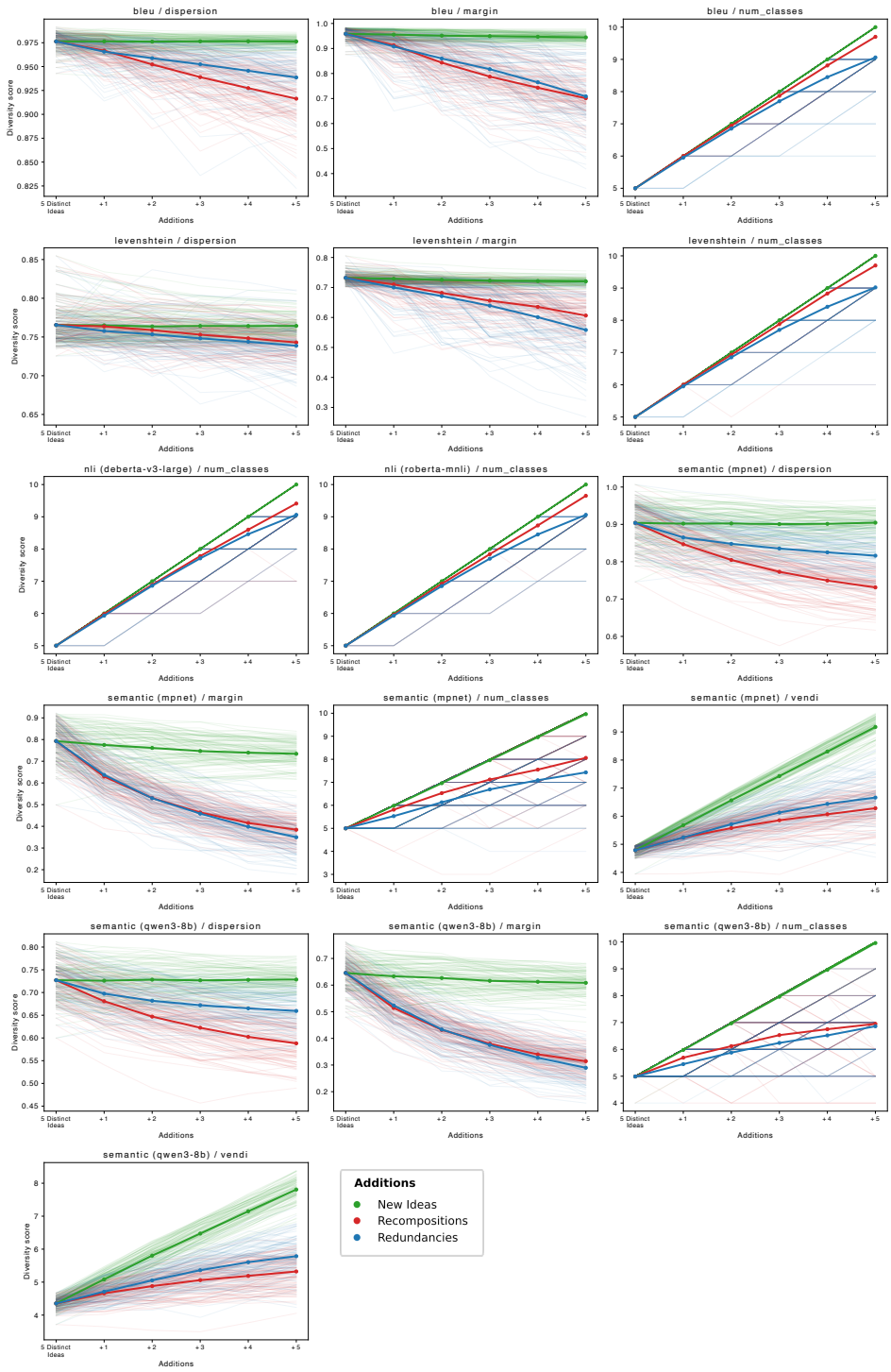
D DISPARITY EXPERIMENT - DIVERSESUMM



E Composition Experiment - Perspectrum



F Composition Experiment - DiverseSumm


bleu / dispersion
bleu / margin
bleu / num_classes
levenshtein / dispersion
levenshtein / margin
levenshtein / num_classes
nli (deberta-v3-large) / num_classes
nli (roberta-mnli) / num_classes
semantic (mpnet) / dispersion
semantic (mpnet) / margin
semantic (mpnet) / num_classes
semantic (mpnet) / vendi
semantic (qwen3-8b) / dispersion
semantic (qwen3-8b) / margin
semantic (qwen3-8b) / num_classes
semantic (qwen3-8b) / vendi
Diversity score
5 Distinct Ideas
+1
+2
+3
+4
+5
Additions
Additions
New Ideas
Recompositions
Redundancies