Contextual Diversity Measure (CDM) for Controllable Story Generation in Large Language Models

Richard R. W. Susilo[1], Hanna Suominen[1,2,3], Patrik Haslum[1]
[1]School of Computing, The Australian National University, ACT, Australia,
[2]School of Medicine and Psychology, The Australian National University, ACT, Australia,
[3]Department of Computing, University of Turku, Finland
Correspondence: richard.susilo@anu.edu.au

Abstract

Scenario-based text generation has broad applications across education and creative writing, but remains underexplored in controllable text generation. We introduce the Contextual Diversity Measure (CDM), a metric that quantifies semantic diversity for scenario generation under fixed abstract semantic constraints, and validate it through controlled experiments. Statistical analysis across four embedding models demonstrates that CDM successfully distinguishes between high-diversity and low-diversity text pairs, with all tests achieving statistical significance at $p < 0.05$ on both the manually curated and LLM-generated subsets of the dataset. Effect sizes range from small-to-medium (Cohen's d: 0.292–0.508) on the former and medium-to-large (Cohen's d: 0.677–1.195) on the latter. Baseline comparisons indicate that CDM achieves excellent discrimination accuracy (100% and 91.9%, respectively), with discriminative power up to $5.5\times$ greater than the best baseline.

1 Introduction

Scenario-based text generation has broad applications across the educational domain (e.g. teaching case studies (Zhang et al., 2019; Guo et al., 2020; Cai et al., 2026), medical simulations (Zheng et al., 2024), language learning (Almazova et al., 2021)) and creative writing (Golden, 2018; Bai et al., 2024). While Large Language Models (LLMs) have demonstrated remarkable text generation capabilities, they face critical challenges: they frequently hallucinate details (Ji et al., 2023; Wang et al., 2025), violate user-specified constraints (Zhang et al., 2023; Liang et al., 2024), or struggle to maintain output diversity across generated scenarios (Chang et al., 2024).

Although Controllable Text Generation (CTG) has been extensively studied in Natural Language Processing, existing work primarily focuses on controlling attributes such as sentiment (Chen et al., 2019; Dathathri et al., 2020; Zhang and Song, 2022), writing style (Prabhumoye et al., 2018; He et al., 2020; Reif et al., 2022), and writing structure (Fan et al., 2018; Goldfarb-Tarrant et al., 2020; Fang et al., 2021). However, the problem of structure-preserving scenario generation, where the goal is to generate contextually diverse texts while maintaining a fixed underlying semantic constraint, remains understudied.

This variant of the CTG problem occurs naturally in tasks such as data augmentation and scenario-based training. Consider ethics training, where instructors need to generate multiple case studies that present the same ethical dilemma — preserving the roles, actions, and relationships that define it — but set in different professional contexts such as medicine, law, or engineering. Here, contextual diversity is essential, as a single scenario fails to expose learners to the range of situations they will encounter, while scenarios that vary too freely may no longer test the same dilemma. Yet no existing metric quantifies contextual diversity under fixed semantic constraints for evaluating generation systems or serving as an optimisation target.

In this paper, we introduce and validate the Contextual Diversity Measure (CDM) for scenario generation from structured semantic constraints. Unlike existing metrics, CDM quantifies diversity at the level of contextual framing, producing numeric scores that can both evaluate generation quality and be integrated into model training objectives.

The main outcomes of this study are as follows:

- We introduce CDM to measure contextual diversity in structure-preserving scenario generation under fixed semantic constraints.
- We validate CDM through controlled experiments on both the manually curated and LLM-generated subsets, with all statistical tests achieving significance at $p < 0.05$ and small-to-medium effect sizes (Cohen's d:

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1125–1138

0.292–0.508) on the manually curated subset, and medium-to-large effect sizes (Cohen's d: 0.677–1.195) on the LLM-generated subset.

- We demonstrate that CDM outperforms existing baselines, achieving perfect discrimination accuracy (100%) on the manually curated subset and the highest accuracy (91.9%) on the LLM-generated subset, with discriminative power up to $5.5\times$ greater than the best baseline.

2 Related Work

Existing text similarity and diversity metrics, such as BERTScore (Zhang et al., 2020), lexical diversity measures (Distinct-N) (Li et al., 2016), and Self-BLEU (Zhu et al., 2018; Shu et al., 2019), operate at the sentence or document level, with BERT and BLEU referring to Bidirectional Encoder Representations from Transformers and Bilingual Evaluation Understudy, respectively. However, these metrics primarily measure surface-level similarity (Deutsch and Roth, 2021) rather than contextual diversity or semantic content (Mathur et al., 2020; Fabbri et al., 2021).

Topic modelling (Blei et al., 2003; Bianchi et al., 2021) is an approach for identifying broad thematic categories in text by analysing word co-occurrence patterns across documents. However, while topic modelling can classify text into different topics, it cannot incorporate structured semantic constraints, nor can it generate new text across different topics while maintaining such constraints.

Alternatively, semantic frame analysis (Baker et al., 1998; Das et al., 2014) captures event structures by identifying event types and their participants. However, frame semantics approaches such as FrameNet (Baker et al., 1998) abstract over lexical variation, limiting their ability to quantify contextual differences across lexical realisations of the same frame structure (Belcavello et al., 2020).

Finally, masked language modelling (Devlin et al., 2019), a superficially related problem, predicts missing tokens conditioned on a single sentence's context. This constrains predictions to that sentence's domain. Our task requires cross-domain instantiations where the same semantic position is filled differently across contexts, and crucially, a mechanism to quantify the diversity between them.

In summary, existing approaches rely on surface-level similarity, in-domain prediction, or thematic classification, and none can quantify contextual diversity under fixed semantic constraints. CDM addresses this gap directly.

3 Problem

3.1 Task Definition

A **structured semantic constraint** is a constraint that defines the semantic content and relationships in a text without dictating the specific words used to express them. In our case, it consists of abstract entities and events (predicate-argument structures specifying what actions occur and which entities fill which semantic roles), based on Semantic Role Labeling (SRL).

Following this, **scenario generation** is the task of producing multiple coherent text realisations that faithfully adhere to a given structured semantic constraint while varying their contextual framing. Each realisation must preserve the specified semantic constraint, maintaining the same entities, events, and role assignments, but may instantiate the abstract elements with different concrete lexical choices, allowing the same underlying meaning to be expressed across diverse domains and contexts. We refer to this variation in domain and context, while preserving the semantic constraint, as **contextual diversity**.

This formulation naturally arises when generation systems produce multiple realisations from a shared template or constraint, such as in data augmentation, paraphrase generation, or scenario-based training.

3.2 Example

Consider the following example of a structured semantic constraint where abstract entities and events are represented as predicate-argument structures:

Predicate	Role	Filler
conduct	ARG0	ENT_1
	ARG1	OBJ_1
	ARGM-LOC/in	LOC_1
become	ARG1	LOC_1
	ARG2	LOC_2
	ARGM-TMP	ATT_1

To interpret this example, we can read the predicate-argument structure as follows:

```
ENT_1 conducts OBJ_1 in LOC_1.
LOC_1 becomes LOC_2 ATT_1.
```

This structure allows for syntactic flexibility while maintaining the same semantic roles. For example, the constraint can also be expressed as:

```
ENT_1 is conducting OBJ_1 in
LOC_1 that has ATT_1 become
LOC_2.
```

These abstract identifiers can be filled with different concrete words or phrases to generate contextually diverse scenarios, as illustrated:

Identifiers	Instantiation 1	Instantiation 2	Instantiation 3
ENT_1	Armin	Jessica	Marcus
OBJ_1	fieldwork	market research	archaeological surveys
LOC_1	country	region	territory
LOC_2	conflict zone	economic hotspot	war zone
ATT_1	recently	recently	recently

From these three possible instantiations, here are some examples as complete texts:

> “Armin is conducting fieldwork in a country that has recently become a conflict zone.”

> “Jessica is conducting market research in a region that has recently become an economic hotspot.”

The examples above follow the same syntactic structure, but the constraint also permits different structural realisations, such as:

> “In a territory that has recently become a war zone, Marcus conducts archaeological surveys.”

This task differs from several similar problems. Unlike paraphrase generation, which preserves meaning while varying surface form, our task preserves the semantic constraint while deliberately shifting the domain and context. Unlike lexical substitution, which replaces words with synonyms within the same domain, our task requires coherent cross-domain shifts across all positions. Finally, unlike masked language modelling, which predicts words independently at each position, our task requires that all filled positions are contextually distinct across generations.

4 Formal Problem Definition

4.1 Abstract Representations

Let $\mathcal{R} = \{R^{(1)}, R^{(2)}, \ldots, R^{(m)}\}$ denote the set of m abstract representations in the dataset. For a given abstract representation $R \in \mathcal{R}$, we formally define it as a tuple:

$$R = \left(V, S, \{\psi_j\}_{j=1}^{n_s}\right)$$

where V and S are representation-specific spaces and $\{\psi_j\}_{j=1}^{n_s}$ are role-assignment functions defined over a universal semantic role space Ξ, as shown:

Entity Space:

$$V = \{v_1, v_2, \ldots, v_{n_v}\} \subset \mathcal{V}$$

represents a finite set of n_v entities specific to the representation R, where $n_v = |V|$ denotes the cardinality of V, and $\mathcal{V}$ denotes the universal entity space.

Predicate Space:

$$S = \{s_1, s_2, \ldots, s_{n_s}\} \subset \mathcal{S}$$

represents a finite set of n_s predicates specific to the representation R, where $n_s = |S|$ denotes the cardinality of S and $\mathcal{S}$ denotes the universal predicate space.

Role Space: Let $\Xi = \{\xi_1, \xi_2, \ldots, \xi_{n_z}\}$ denote the universal semantic role space, where $n_z = |\Xi|$ denotes the cardinality of Ξ and each ξ_i represents a semantic role such as ARG0, ARG1, or ARGM-LOC.

Role Assignment: For each predicate $s_j \in S$, we define $\psi_j : \Xi \to V \cup \{\emptyset\}$ as a function mapping semantic roles to entities, where

$$\psi_j(\xi) = \begin{cases} v_i & \text{if entity } v_i \text{ fills role } \xi \text{ for predicate } s_j, \\ \emptyset & \text{if role } \xi \text{ is not assigned to predicate } s_j. \end{cases}$$

4.2 Instantiation

For a given abstract representation R, we generate k instantiations $\{T_1, T_2, \ldots, T_k\}$. Each generated instantiation T_i is represented as a sequence of phrases, where each phrase represents a concrete lexical instantiation of an abstract semantic element:

$$T_i = \langle w_{i,1}, w_{i,2}, \ldots, w_{i,n} \rangle$$

where n denotes the number of phrases[1] in T_i, and each $w_{i,j}$ represents the phrase at position j in instantiation T_i. For instance, from the example in Section 3.2, Instantiation 1 can be read as $T_1 = \langle$"Armin", "fieldwork", "country", $\ldots\rangle$.

5 Definition of the CDM Metric

We introduce a geometric decomposition that analyses semantic changes at each filled position relative to the overall instantiation-level shift. For each semantic position index $j \in [n]$, we measure the semantic diversity among the k phrases corresponding to the position j across the k generated texts.

5.1 Centroid Direction

We first establish a reference direction by computing the centroid of each instantiation's phrase-level embeddings and taking the direction between them.

For each phrase $w_{i,j}$ in an instantiation $T_i = \langle w_{i,1}, w_{i,2}, \ldots, w_{i,n}\rangle$, where each phrase may consist of one or more words, we apply a word embedding function $\phi : \mathcal{W} \to \mathbb{R}^d$ that maps words to d-dimensional embeddings. The embedding for each phrase is computed as the normalised mean of its constituent word embeddings:

$$\bar{\phi}(w_{i,j}) = \frac{1}{|w_{i,j}|} \sum_{q=1}^{|w_{i,j}|} \phi(w_{i,j}^{(q)}), \text{ and}$$

$$\hat{e}_{i,j} = \frac{\bar{\phi}(w_{i,j})}{\|\bar{\phi}(w_{i,j})\|_2}.$$

where $|w_{i,j}|$ is the number of words in the phrase and $w_{i,j}^{(q)}$ is the q-th word in the phrase.

We define the **centroid** of the instantiation, C_i, as the mean of its normalised word embeddings:

$$C_i = \frac{1}{n} \sum_{j=1}^{n} \hat{e}_{i,j} \in \mathbb{R}^d.$$

For each pair of instantiations T_i and T_l, we can compute the vector from C_i to C_l and normalise it to obtain the primary direction of semantic change:

$$\eta(i,l) = \frac{C_l - C_i}{\|C_l - C_i\|_2}.$$

[1]The number of phrases n is consistent across all instantiations T_i from the same abstract representation R, as n corresponds to the number of fillable elements in the semantic constraint.

5.2 Geometric Decomposition

We decompose each word-level change[2] into two orthogonal components relative to the centroid direction $\eta(i,l)$: one aligned with the overall semantic shift, and one independent of it, illustrated in Figure 1.

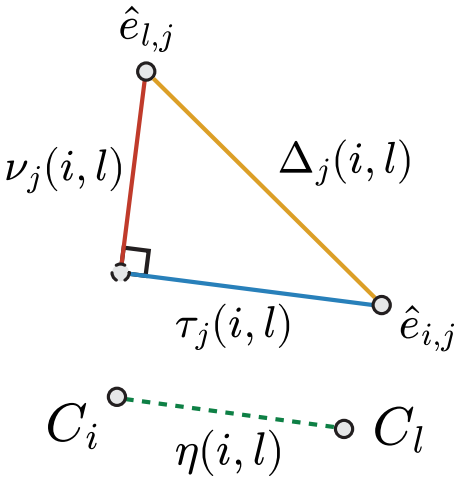

Figure 1: Visualisation of the geometric decomposition of the word change vector $\Delta_j(i,l)$ into a directional component $\tau_j(i,l)$ and an orthogonal component $\nu_j(i,l)$ relative to the centroid direction $\eta(i,l)$. Shown in 2D for illustration; the actual decomposition operates in $\mathbb{R}^d$ where d is the embedding dimension. See Appendix B for an extended visualisation with three instantiations.

We define the **Word Change Vector** as the change in word embedding from instantiation T_i to instantiation T_l at position j as

$$\Delta_j(i,l) = \hat{e}_{l,j} - \hat{e}_{i,j}.$$

We then project $\Delta_j(i,l)$ onto the centroid direction $\eta(i,l)$ to obtain the **Projection Vector**:

$$p_j(i,l) = (\Delta_j(i,l) \cdot \eta(i,l))\, \eta(i,l).$$

The **Directional Component** represents the magnitude of the word change along the centroid direction, defined as

$$\tau_j(i,l) = \|p_j(i,l)\|_2 .$$

The **Orthogonal Component** captures contextual variation beyond the main thematic shift, defined as

$$\nu_j(i,l) = \|\Delta_j(i,l) - p_j(i,l)\|_2 .$$

[2]Each filler $w_{i,j}$ is a phrase that may consist of one or more words, and we use word-level change as shorthand for a change at an individual filled semantic position. Formally, each filler is represented by the normalised mean embedding of its constituent word embeddings.

Intuitively, the directional component τ_j measures how much of a word's change aligns with the overall semantic shift between the two instantiations, while the orthogonal component ν_j captures the remaining variation that is independent of this shift. These two components are semantically distinct: τ_j tracks domain-level shift (e.g., moving from an academic to a business context), while ν_j captures residual word-level variation that is not explained by the global context change. A principled diversity metric should be sensitive to both, as two instantiations may share a broad domain shift yet differ substantially in their specific lexical choices, or vice versa.

We combine these two components into a single score using a weighted sum, allowing the balance between directional and orthogonal contributions to be configured:

$$g_j(i,l) = \zeta \times \Big(\lambda \times \tau_j(i,l) + (1-\lambda) \times \nu_j(i,l)\Big)$$

where ζ is the amplification factor and λ is the balance parameter weighting the directional component; both are tunable parameters. The weighted combination naturally parameterises the trade-off between the two orthogonal sources of diversity, with $\lambda = 0$ and $\lambda = 1$ recovering pure orthogonal and directional sensitivity, respectively; the default $\lambda = 0.5$ reflects no prior preference between the two. The amplification factor ζ rescales the raw geometric quantities, which are typically small due to embedding normalisation, into a range where the subsequent transformation operates with meaningful sensitivity.

To bound the score to $[0, 1]$ and introduce sensitivity around a natural midpoint, we apply a scaled $\tanh$ transformation:

$$G_j(i,l) = \frac{1}{2}\left(1 + \tanh\left(\gamma\sqrt{2}\left(g_j(i,l) - \frac{\sqrt{2}}{2}\right)\right)\right)$$

where γ is the steepness parameter controlling the non-linearity of the transformation. The midpoint $\frac{\sqrt{2}}{2}$ is geometrically motivated, corresponding to a word change vector at $45°$ relative to the centroid direction at equal directional and orthogonal contributions. Hyperparameter values used across all experiments are reported in Appendix C.

5.3 Contextual Diversity Metric

A specific score per semantic position j is given by

$$\text{CDM}_j = \binom{k}{2}^{-1} \sum_{i=1}^{k} \sum_{l=i+1}^{k} G_j(i,l)\,.$$

Therefore, we have the overall CDM score across all semantic positions, $\forall j \in [n]$, which can be defined as

$$\begin{aligned}\text{CDM} &= \frac{1}{n}\sum_{j=1}^{n} \text{CDM}_j \\ &= \frac{2}{nk(k-1)} \sum_{j=1}^{n}\sum_{i=1}^{k}\sum_{l=i+1}^{k} G_j(i,l)\,.\end{aligned}$$

6 Methods for Validating CDM

To validate CDM, we conduct a controlled experiment to demonstrate that our diversity metric meaningfully distinguishes between texts that express the same semantic content in different contexts.

6.1 Dataset

The dataset consists of 218 instances, each associated with an abstract representation R. For each instance, we construct 3 text realisations that faithfully express the constraints specified in R. The three texts are:

- **Reference Text (RF):** The original labelled text derived from the abstract SRL representation (academic domain).
- **High Diversity (HD):** Maximises contextual distance from the reference text by instantiating the abstract representation in a distinct semantic domain (e.g., business, archaeology, journalism).
- **Low Diversity (LD):** Minimises contextual distance from the reference text by instantiating the abstract representation in a closely related semantic domain (e.g., a different academic scenario).

In total, the dataset contains 654 realisations. Of these, a core subset of 24 realisations was manually curated with fully specified abstract representations comprising 135 lexical instantiations, while the remaining 630 realisations were generated using two LLMs, Claude (Opus 4.6) and GPT (GPT-5.4), with each model producing 300 and 330 realisations, respectively, using the curated instances as few-shot examples.

The manually curated subset serves as a controlled benchmark, as these were constructed by domain experts and validated through human review to ensure faithful adherence to the semantic constraints. The LLM-generated subset provides a larger-scale evaluation to test whether CDM generalises beyond the curated instances. Further details on the data curation and generation process are provided in Appendix A.

From these three text realisations, we create two groups for comparison: $P_1 = \{\text{RF, HD}\}$ representing high-diversity pairs and $P_2 = \{\text{RF, LD}\}$ representing low-diversity pairs. By design, we expect $\text{CDM}(P_1) > \text{CDM}(P_2)$, where P_1 pairs should exhibit significantly higher diversity scores than P_2.

6.2 Word Embeddings

In our experiments, we instantiate ϕ using four different pre-trained embedding models to evaluate the robustness of our diversity metric, as illustrated:

Type	**Model** (ϕ)	**Dimension** (d)	**Parameters**
Static[3]	GloVe	300	120M
	Word2Vec	300	900M
	FastText	300	300M
Contextual[4]	MiniLM	384	22.7M

6.3 Statistical Analysis

To validate our hypothesis that $\text{CDM}(P_1) > \text{CDM}(P_2)$, we employ the following three complementary statistical tests:

6.3.1 Two-Sided Wilcoxon Signed-Rank Test

We first test whether the paired difference in CDM scores differs significantly from zero using a two-sided Wilcoxon signed-rank test. The null hypothesis is:

$$H_0 : \text{median}\left(\text{CDM}\left(P_1\right) - \text{CDM}\left(P_2\right)\right) = 0.$$

A significant result ($p < 0.05$) indicates that P_1 and P_2 produce detectably different diversity scores across different semantic positions.

[3]Static embeddings assign a single fixed vector to each word regardless of context.

[4]Contextual embeddings generate representations depending on surrounding words.

6.3.2 One-Sided Wilcoxon Signed-Rank Test

Since we expect P_1 to produce a larger diversity score than P_2, we test this directional hypothesis using a one-sided Wilcoxon signed-rank test:

$$H_1 : \text{median}\left(\text{CDM}\left(P_1\right) - \text{CDM}\left(P_2\right)\right) > 0.$$

A significant result ($p < 0.05$) indicates that P_1 consistently scores higher than P_2 across different semantic positions.

6.3.3 Cohen's d Effect Size

To quantify the magnitude of the difference between the high- and low-diversity pairs, we compute a paired-samples Cohen's d over the matched differences. For each abstract representation p and semantic position j, the paired difference is

$$\delta_{p,j} = \text{CDM}_j^{(p)}(P_1) - \text{CDM}_j^{(p)}(P_2),$$

yielding $N = \sum_{p=1}^{m} n_p$ comparisons, where n_p is the number of semantic positions in representation p. The effect size is then

$$d = \frac{\bar{\delta}}{s_\delta}, \quad \bar{\delta} = \frac{1}{N}\sum_{p,j}\delta_{p,j},$$
$$s_\delta = \sqrt{\frac{1}{N-1}\sum_{p,j}\left(\delta_{p,j} - \bar{\delta}\right)^2}.$$

A positive value of d indicates that CDM assigns higher diversity scores to the high-diversity pairs P_1 than to the low-diversity pairs P_2. Following standard conventions, we interpret $|d| \geq 0.2$ as a small effect, $|d| \geq 0.5$ as a medium effect, and $|d| \geq 0.8$ as a large effect. A large effect size indicates that CDM produces substantially different scores for P_1 and P_2.

6.4 Baseline

Since each group consists of only two texts, existing sentence-level similarity and diversity metrics can be directly applied as baselines. We compare CDM against the following baseline methods:

- **BERTScore (Zhang et al., 2020):** A learned metric that computes token-level similarity using contextualised embeddings from pre-trained BERT models.
- **Distinct-1 and Distinct-2 (Li et al., 2016):** Lexical diversity metrics that measure the ratio of unique unigrams and bigrams to total tokens.

Model	Manually Curated Subset			LLM Generated Subset		
	W2	W1	CD	W2	W1	CD
FastText	***	***	0.508‡	***	***	1.195§
GloVe	**	**	0.292†	***	***	0.677‡
MiniLM	**	***	0.402†	***	***	1.076§
Word2Vec	*	**	0.339†	***	***	1.114§

Table 1: **Statistical Test Results:** CDM score test results across embedding models. **W2**: Wilcoxon two-sided, **W1**: Wilcoxon one-sided, **CD**: Cohen's d. All p-values are from Wilcoxon signed-rank tests. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Effect size: † small ($d \geq 0.2$), ‡ medium ($d \geq 0.5$), § large ($d \geq 0.8$).

- **Self-BLEU (Zhu et al., 2018; Shu et al., 2019):** An inverse measure of diversity that computes BLEU scores between texts.
- **Sentence Similarity:** Cosine similarity between sentence-level embeddings, measuring semantic similarity between text pairs.

For consistency, all baseline metrics are converted to represent text dissimilarity, scaled to $[0, 1]$, where 1 indicates completely dissimilar texts and 0 indicates identical texts. We compare the ability of these methods to distinguish between the two groups by measuring: (1) **Accuracy:** the proportion of scenarios where $\mathrm{M}(P_1) > \mathrm{M}(P_2)$, and (2) **Sum Difference:** the total value of $\mathrm{M}(P_1) - \mathrm{M}(P_2)$ across all scenarios. Here, M denotes the score of the metric being evaluated.

7 Statistical Test Results

As shown in Table 1, both two-sided and one-sided Wilcoxon signed-rank tests yield p-values well below the significance threshold ($\alpha = 0.05$) across all embeddings, rejecting the null hypothesis H_0 and confirming that CDM produces significantly different scores between high-diversity pairs (P_1) and low-diversity pairs (P_2). The one-sided tests support the alternative hypothesis H_1 that CDM assigns consistently higher diversity scores to P_1 than P_2, demonstrating robust statistical support for our experimental hypothesis. Notably, all embeddings on the LLM-generated subset achieve the highest significance level ($p < 0.001$) across both tests, while the manually curated subset shows varying significance levels ranging from $p < 0.05$ to $p < 0.001$.

On the manually curated subset, all embedding models produce small-to-medium effect sizes (Cohen's d: 0.292–0.508), indicating that the difference between P_1 and P_2 scores is detectable and practically meaningful, though moderate in magnitude. On the LLM-generated subset, effect sizes increase substantially: Word2Vec ($d = 1.114$), MiniLM ($d = 1.076$), and FastText ($d = 1.195$) all achieve large effect sizes ($d \geq 0.8$), while GloVe ($d = 0.677$) achieves a medium effect. The stronger effects on the LLM-generated subset suggest that as the number of instances increases, CDM's discriminative signal strengthens, and the separation between high-diversity and low-diversity pairs becomes more pronounced across all embedding architectures.

8 Performance Results

On the manually curated subset, BERTScore achieves perfect discrimination accuracy (100%) among baseline methods, as shown in Table 2. However, BERTScore produces the smallest sum difference (+0.132), indicating weak discriminative power despite perfect accuracy. In contrast, Sentence Similarity achieves the strongest sum difference among baselines (+0.661), but falls short in accuracy at 87.5%. On the other hand, CDM variants demonstrate strong performance, where three CDM variants (FastText, MiniLM, and Word2Vec) achieve perfect accuracy. Critically, CDM (FastText) produces the largest sum difference (+0.998) across all methods, representing a 7.6× larger sum difference than BERTScore (best accuracy baseline) and a 1.5× improvement over Sentence Similarity (best sum difference baseline). A per-scenario breakdown of these results is provided in Appendix D.

On the LLM-generated subset, CDM (FastText) achieves the highest accuracy among all methods (91.9%), while Sentence Similarity achieves the

Method	Manually Curated Subset			LLM Generated Subset		
	Acc (%)	Mean	Sum	Acc (%)	Mean	Sum
Baseline						
BERTScore	**100.0**	0.017	0.132	74.9	0.008	1.767
Distinct-1	75.0	0.024	0.192	82.9	0.030	6.391
Distinct-2	75.0	0.023	0.185	85.8	0.054	11.345
Self-BLEU	62.5	0.053	0.421	86.3	0.122	25.764
Sentence Similarity	87.5	0.083	0.661	91.0	0.139	29.339
Ours						
CDM (GloVe)	87.5	0.110	0.881	77.6	0.760	159.697
CDM (Word2Vec)	**100.0**	0.078	0.624	88.6	**0.767**	**161.047**
CDM (MiniLM)	**100.0**	0.106	0.845	89.6	0.578	121.951
CDM (FastText)	**100.0**	**0.125**	**0.998**	**91.9**	0.659	138.393

Table 2: **Performance Results:** Diversity scores for each metric on the manually curated and LLM-generated subsets. **Acc** shows the percentage of scenarios where $M(P_1) > M(P_2)$. **Mean** is the average $M(P_1) - M(P_2)$ difference per scenario. **Sum** is the total $M(P_1) - M(P_2)$ difference across all scenarios, where M denotes the metric being evaluated. For CDM, parentheses indicate the embedding model used. All metrics are scaled to $[0, 1]$ representing dissimilarity (1 = completely dissimilar, 0 = identical).

highest accuracy among baseline methods (91.0%). However, the key distinction lies in discriminative power, where CDM variants produce substantially larger sum differences than all baselines, with CDM (Word2Vec) achieving the highest sum difference (161.047), a $5.5\times$ improvement over the best baseline by sum difference, Sentence Similarity (29.339). All four CDM variants produce sum differences exceeding 121, while no baseline surpasses 30, demonstrating that CDM's advantage in discriminative power generalises from the curated subset to the larger LLM-generated subset. Notably, the optimal CDM embedding variant differs across the two subsets, suggesting that the choice of embedding may vary with data characteristics.

9 Discussion

Principal Results: Our controlled experiments have provided robust evidence that CDM behaves as intended: texts instantiated in semantically distant domains consistently achieved higher diversity scores than texts in semantically proximate domains. The convergence of statistical test results across four diverse embeddings demonstrates that CDM captures genuine semantic diversity patterns independent of the underlying embedding architecture.

The performance results demonstrate that CDM matches or exceeds all baselines in discrimination accuracy and consistently exceeds them in discriminative power. Existing metrics such as BERTScore and Sentence Similarity achieve accuracy comparable to CDM's best variants, but their sum differences remain small, meaning they can identify which pair is more diverse without capturing how much more diverse it is. In contrast, CDM produces discriminative power up to $5.5\times$ greater than the best baseline, revealing that existing approaches lack the sensitivity to meaningfully quantify the degree of contextual diversity between text pairs. This highlights that accuracy alone is a misleading indicator of metric quality, as a metric that merely classifies but cannot produce substantial separation provides weak gradient signals for downstream generation systems.

Significance of Applications: CDM's larger margins make it suitable not only for evaluation but also for integration into training objectives, where clear quantitative differences are necessary to guide models towards producing contextually diverse outputs. This property enables three key applications: evaluating and comparing generation systems' ability to produce contextually diverse scenarios, optimising models by incorporating CDM into training objectives to encourage diverse instantiations, and assessing generated output quality by quantifying the degree of contextual variation.

As noted in Section 1, scenario-based text generation has broad applications across education and creative writing. A good example where structure-preserving scenario generation is particularly relevant is ethics training, where instructors may need to generate multiple case studies that test the same ethical dilemma (preserving the semantic constraint) but set in different professional contexts (varying the contextual framing). CDM can evaluate whether the generated scenarios are sufficiently diverse, or serve as an optimisation target to encourage greater contextual variation during generation.

10 Conclusion

Our CDM addresses an understudied problem in controllable text generation: measuring contextual diversity under fixed semantic constraints and providing quantifiable measures that existing metrics cannot capture. Our results have demonstrated that CDM successfully quantifies contextual diversity with statistical significance ($p < 0.05$) and meaningful effect sizes (Cohen's d up to 1.195). Furthermore, CDM matches or exceeds the baselines in accuracy and outperforms them in discriminative power, with margins up to $5.5\times$ greater than the best baseline.

With these properties, CDM provides a principled metric for scenario generation with three key applications: (1) evaluating and comparing existing generation systems' ability to produce contextually diverse scenarios, (2) optimising models by integrating CDM into training objectives to encourage diverse instantiations, and (3) assessing generated output quality by quantifying the degree of contextual variation.

Limitations

This work introduces CDM but does not yet integrate it into automated generation systems or evaluate it as part of specific downstream applications such as data augmentation, scenario-based assessment, or ethics training. Future work could explore incorporating CDM into LLM generation pipelines, building on recent controllable generation frameworks, either as a decoding constraint (e.g., diversity-promoting beam search) or as a training objective. Additionally, evaluating CDM on datasets from new domains would further test its generalisability. To foster such research, our code and dataset are available on GitHub under the MIT License.

Acknowledgements

We gratefully acknowledge Professor Inger Mewburn from The Australian National University for annotating and providing the scenario-based ethics dataset used in our controlled experiments. The original data, used to support research ethics training, are from the US National Institutes of Health.

References

Nadezhda Almazova, Anna Rubtsova, Nora Kats, Yuri Eremin, and Natalia Smolskaia. 2021. Scenario-based instruction: The case of foreign language training at multidisciplinary university. *Education Sciences*, 11(5).

Shurui Bai, Donn Emmanuel Gonda, and Khe Foon Hew. 2024. Write-curate-verify: A case study of leveraging generative ai for scenario writing in scenario-based learning. *IEEE Transactions on Learning Technologies*, 17:1301–1312.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Frederico Belcavello, Marcelo Viridiano, Alexandre Diniz da Costa, Ely Edison da Silva Matos, and Tiago Timponi Torrent. 2020. Frame-based annotation of multimodal corpora: Tracking (a)synchronies in meaning construction. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 23–30, Marseille, France. European Language Resources Association.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(Jan):993–1022.

Xuan Cai, Xuesong Bai, Zhiyong Cui, Danmu Xie, Daocheng Fu, Haiyang Yu, and Yilong Ren. 2026. Text2scenario: Text-driven scenario generation for autonomous driving test. *Automotive Innovation*, 9(1):102–127.

Cheng Chang, Siqi Wang, Jiawei Zhang, Jingwei Ge, and Li Li. 2024. Llmscenario: Large language model driven scenario generation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(11):6581–6594.

Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4925–4931. International Joint Conferences on Artificial Intelligence Organization.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. *Preprint*, arXiv:1912.02164.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Preprint*, arXiv:2007.12626.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Outline to story: Fine-grained controllable story generation from cascaded events. *Preprint*, arXiv:2101.00822.

Paullett Golden. 2018. Contextualized writing: Promoting audience-centered writing through scenario based learning. *International Journal for the Scholarship of Teaching and Learning*, 12(1):6.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Hongwen Guo, Mo Zhang, Paul Deane, and Randy Bennett. 2020. Effects of scenario-based assessment on students' writing processes. *Journal of Educational Data Mining*, 12(1):19–45.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *Preprint*, arXiv:2002.03912.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and Zhiyu Li. 2024. Controllable text generation for large language models: A survey. *Preprint*, arXiv:2408.12599.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1827, Florence, Italy. Association for Computational Linguistics.

Junli Wang, Chenyang Zhang, Dongyu Zhang, Haibo Tong, Chungang Yan, and Changjun Jiang. 2025. A recent survey on controllable text generation: A causal perspective. *Fundamental Research*, 5(3):1194–1203.

Hanqing Zhang and Dawei Song. 2022. DisCup: Discriminator cooperative unlikelihood prompt-tuning for controllable text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3392–3406, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3).

Mo Zhang, Peter W. van Rijn, Paul Deane, and Randy E. Bennett. 2019. Scenario-based assessments in writing: An experimental study. *Educational Assessment*, 24(2):73–90.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Koulong Zheng, Zhiyu Shen, Zanhao Chen, Chang Che, and Huixia Zhu. 2024. Application of ai-empowered scenario-based simulation teaching mode in cardiovascular disease education. *BMC Medical Education*, 24(1):1003.

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.

A Dataset Generation

A.1 Manually Curated Subset

The manually curated subset consists of 8 abstract representations, each with 3 text realisations (Reference, High Diversity, and Low Diversity), totalling 24 realisations comprising 135 lexical instantiations. These were derived from a scenario-ethics dataset originally developed for research ethics training by the US National Institutes of Health. Each instance was individually annotated by a domain expert who carefully marked up the predicate-argument structures following the Semantic Role Labeling framework described in Section 3. The three text realisations per abstract representation were then constructed and validated by the authors to ensure faithful adherence to the semantic constraints, with particular attention to maintaining clear contextual separation between the High Diversity and Low Diversity conditions. This curation process ensures that the manually curated subset serves as a reliable controlled benchmark against which both CDM and the baseline metrics are evaluated.

A.2 LLM-Generated Subset

To evaluate whether CDM generalises beyond the manually curated instances, we generated a larger dataset using two LLMs: Claude and GPT.

Prompt Design. The generation followed a few-shot prompting strategy. Each prompt consisted of: (1) a system-level task description defining the scenario generation constraints, including the requirement to preserve the semantic constraint while varying contextual framing across three diversity conditions (Reference, High Diversity, and Low Diversity); (2) all 8 manually curated instances as few-shot examples, demonstrating the expected input-output format including the reference sentence, the high- and low-diversity alternatives, and the position-aligned list of semantic role fillers; and (3) an output specification requiring structured JSON format with the generated sentences and their corresponding role filler lists. The full prompt template is shown in the box below.

Generation Process. Each model was prompted to generate new instances following the same three-condition format as the curated subset. We used Claude Opus 4.6 with extended thinking and GPT-5.4 with thinking mode enabled. Claude produced 300 realisations and GPT produced 330 realisations, yielding 630 LLM-generated realisations in total.

Note that while the general task formulation in Section 3 permits syntactic flexibility across realisations of the same semantic constraint, the LLM-generated evaluation subset deliberately imposes stricter sentence-structure alignment across the three diversity conditions. This controls for syntactic variation as a confound, ensuring that observed differences in CDM scores between high- and low-diversity pairs reflect contextual rather than structural variation.

Quality Control. All generated instances were manually reviewed to verify structural consistency, ensuring that the sentence structure and semantic roles were preserved across the three diversity conditions, and that the position-aligned role fillers were correctly extracted. Instances that violated structural constraints or contained misaligned role fillers were discarded.

Prompt Template for LLM-Generated Subset

System: You are a scenario generation assistant. Your task is to generate contextually diverse variations of reference sentences.

Given a reference sentence, you must:

1. Identify the key semantic role fillers in the sentence (e.g., the agent, action, location, object, attributes, etc.)
2. Generate two alternative sentences:
   - **HIGH DIVERSITY:** Rewrite the sentence in a completely different domain while preserving the exact same sentence structure and semantic roles.
   - **LOW DIVERSITY:** Rewrite the sentence in a closely related domain while preserving the exact same sentence structure and semantic roles.
3. Extract the list of specific words that fill the semantic roles for each version, in the same order.

[8 few-shot examples omitted for brevity; see supplementary materials (GitHub) for the complete prompt.]

Now generate 100 new examples following the exact same format. Each example should:

- Be a unique reference sentence with a realistic scenario
- Have a high diversity alternative in a distinctly different domain
- Have a low diversity alternative in a closely related domain
- Have specific_words lists that are aligned position-by-position
- Maintain identical sentence structure across all three versions

B Extended CDM Visualisation

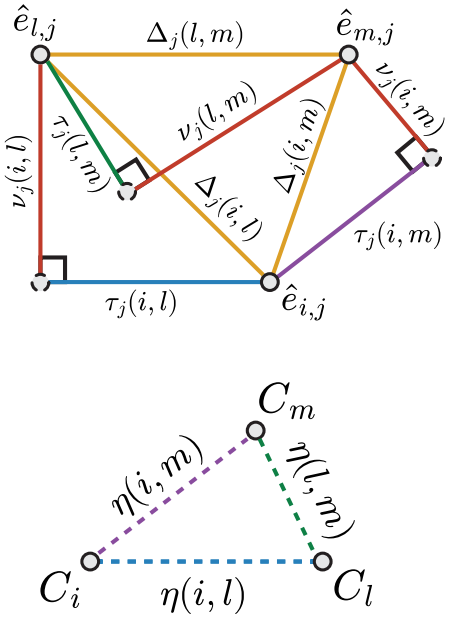

Figure 2: Visualisation of the CDM geometric decomposition for three instantiations (T_i, T_l, T_m). Each pair of instantiations produces its own centroid direction η and corresponding word change vector Δ_j, which is decomposed into directional (τ_j) and orthogonal (ν_j) components. The pairwise scores $G_j(i,l)$, $G_j(i,m)$, and $G_j(l,m)$ are then averaged to produce the final CDM_j for position j, as defined in Section 5.3. Shown in 2D for illustration; the actual decomposition operates in $\mathbb{R}^d$ where d is the embedding dimension.

Figure 2 provides a visual intuition for how CDM differs from conventional distance metrics. For each pair of instantiations, CDM decomposes the word change vector (yellow) into a directional component, τ_j, along the centroid direction, η, (blue, green, and purple for each respective pair) and an orthogonal component, ν_j, perpendicular to it (red). The final CDM score is the pairwise average of these decomposed component magnitudes across all positions. In contrast, conventional metrics such as Euclidean distance would simply measure the magnitude of the yellow vectors directly, while cosine distance would measure the angle between embedding vectors at a given point. By decomposing the word-level change relative to the overall instantiation-level shift, CDM captures not only how much a word changed, but how that change relates to the broader contextual shift between instantiations, a distinction that single-value distance metrics do not make.

C Hyperparameter Configuration

Table 3 reports the hyperparameter values used across all experiments. Each embedding's parameters were held constant across all tests and datasets, and were optimised using grid search.

	Parameter		
Embedding	λ	ζ	γ
MiniLM	0.5	1.0	1.2
FastText	0.5	1.6	1.7
GloVe	0.5	1.9	1.7
Word2Vec	0.5	1.0	1.2

Table 3: Hyperparameter configuration for CDM across embedding models. λ: balance parameter, ζ: amplification factor, γ: steepness parameter.

D Scenario-Level Performance Breakdown

This section provides a scenario-level breakdown of the results summarised in Table 2, based on the manually curated subset.

As shown in Table 4, CDM produces a larger sum difference than the best baseline by sum difference (Sentence Similarity) and outperforms it in 6 out of 8 scenarios. Baseline methods exhibit considerable volatility across scenarios; for instance, Self-BLEU ranges from $+0.253$ (S04) to -0.199 (S05), and Distinct-2 ranges from $+0.130$ (S01) to -0.102 (S05). In contrast, CDM variants demonstrate greater stability, with CDM (FastText) maintaining positive differences across all eight scenarios.

Several scenarios prove particularly challenging for the baselines. In S03, CDM variants consistently produce strong differences (ranging from +0.123 to +0.256), with CDM (GloVe) achieving +0.256, approximately $6.9\times$ larger than the best-performing baseline in this scenario, Distinct-1 (+0.037). In S05, all CDM variants maintain positive differences (ranging from $+0.051$ to $+0.101$), while all baseline methods produce near-zero or negative values (ranging from $+0.002$ to -0.199), indicating that baselines fail entirely to detect the diversity difference in this scenario. Even in S08, where one CDM variant (GloVe, -0.022) produces a negative value, the best-performing CDM variant (Word2Vec, $+0.134$) still outperforms the best baseline, Sentence Similarity ($+0.120$).

Method	Scenarios								Overall	
	S01	S02	S03	S04	S05	S06	S07	S08	Sum	Acc (%)
Baseline										
BERTScore										
P_1	0.066	0.070	0.065	0.069	0.074	0.060	0.042	0.038	0.484	
P_2	0.047	0.036	0.041	0.049	0.072	0.041	0.034	0.032	0.352	**100.0**
Diff	+0.020	+0.034	+0.024	+0.020	+0.002	+0.019	+0.008	+0.006	+0.132	
Distinct-1										
P_1	0.630	0.571	0.611	0.640	0.504	0.553	0.521	0.635	4.666	
P_2	0.548	0.556	0.574	0.588	0.541	0.500	0.510	0.656	4.474	75.0
Diff	+0.082	+0.016	+0.037	+0.052	-0.037	+0.053	+0.011	-0.021	+0.192	
Distinct-2										
P_1	0.833	0.798	0.769	0.833	0.748	0.744	0.670	0.738	6.133	
P_2	0.703	0.784	0.750	0.755	0.850	0.675	0.656	0.774	5.948	75.0
Diff	+0.130	+0.014	+0.019	+0.078	-0.102	+0.069	+0.014	-0.036	+0.185	
Self-BLEU										
P_1	0.765	0.712	0.585	0.897	0.631	0.732	0.452	0.565	5.339	
P_2	0.517	0.662	0.604	0.645	0.830	0.560	0.431	0.669	4.918	62.5
Diff	**+0.248**	+0.050	-0.019	**+0.253**	-0.199	+0.171	+0.021	-0.104	+0.421	
Sentence Sim.										
P_1	0.402	0.318	0.325	0.370	0.230	0.326	0.365	0.308	2.644	
P_2	0.310	0.224	0.306	0.323	0.237	0.174	0.222	0.187	1.983	87.5
Diff	+0.093	**+0.094**	+0.019	+0.047	-0.007	+0.153	+0.143	+0.120	+0.661	
Ours										
CDM (GloVe)										
P_1	0.632	0.577	0.634	0.672	0.509	0.679	0.519	0.591	4.814	
P_2	0.505	0.566	0.379	0.608	0.408	0.506	0.349	0.613	3.933	87.5
Diff	+0.127	+0.011	**+0.256**	+0.065	**+0.101**	+0.173	+0.170	-0.022	+0.881	
CDM (Word2Vec)										
P_1	0.276	0.297	0.329	0.472	0.235	0.307	0.280	0.388	2.584	
P_2	0.219	0.245	0.207	0.430	0.184	0.222	0.199	0.254	1.960	**100.0**
Diff	+0.057	+0.052	+0.123	+0.042	+0.051	+0.086	+0.081	**+0.134**	+0.624	
CDM (MiniLM)										
P_1	0.738	0.592	0.619	0.748	0.628	0.765	0.524	0.640	5.253	
P_2	0.591	0.575	0.397	0.685	0.540	0.559	0.449	0.612	4.408	**100.0**
Diff	+0.147	+0.017	+0.221	+0.063	+0.088	**+0.206**	+0.075	+0.028	+0.845	
CDM (FastText)										
P_1	0.627	0.523	0.544	0.644	0.511	0.674	0.492	0.607	4.621	
P_2	0.451	0.504	0.366	0.551	0.456	0.498	0.299	0.498	3.623	**100.0**
Diff	+0.175	+0.020	+0.178	+0.093	+0.055	+0.175	**+0.193**	+0.109	**+0.998**	

Table 4: **Scenario-Level Comparison:** Per-scenario breakdown of the manually curated subset from Table 2. The **P1** and **P2** rows report $M(P_1)$ and $M(P_2)$; **Diff** reports $M(P_1) - M(P_2)$, where positive values favour the high-diversity pair. **Sum** totals these across all eight scenarios; **Acc** is the percentage with positive differences.