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

Anonymous ACL submission

Abstract

Scenario-based text generation has broad applications across education and creative writing, but remains underexplored in the controllable text generation problem domain. We introduce the Contextual Diversity Measure (CDM), a metric that quantifies semantic diversity for scenario generation given abstract semantic role labeling 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 significance at $p < 0.05$ and small-to-medium effect sizes (Cohen's d from 0.292 to 0.508). Baseline comparisons show that CDM achieves perfect discrimination accuracy (100%) with the best-performing variant producing a discriminative power $1.5\times$ 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., 2025), 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 in text generation: 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 diversity (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 (He et al., 2020; Prabhumoye et al., 2018; Reif et al., 2022), and writing structure (Fan et al., 2018; Goldfarb-Tarrant et al., 2020; Fang et al., 2021). To the best of our knowledge, the problem of structure-preserving scenario generation, where the goal is to generate contextually diverse texts while maintaining a fixed underlying semantic structure, remains understudied.

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 meaning (Mathur et al., 2020; Fabbri et al., 2021). Topic modelling (Blei et al., 2003; Bianchi et al., 2021) is an approach to identify 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).

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 with all statistical tests achieving significance at $p < 0.05$ and small-to-medium effect sizes (Cohen's d: 0.292–0.508).

- We demonstrate that CDM outperforms existing baselines in both accuracy, achieving perfect classification (100%), and discriminative power by producing $1.5\times$ greater effect.

2 Problem

2.1 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 in contextual framing. Each realisation must preserve the specified semantic structure, maintaining the same entities, events, and role assignments, but can instantiate the abstract elements with different concrete lexical choices, allowing the same underlying meaning to be expressed across diverse domains and contexts. We call this variation in domain and context while preserving semantic structure as **contextual diversity**.

2.2 Example

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

Table 1: **Events:** Event predicates and their semantic role assignments within an abstracted SRL graph R.

Consider the following example of a structured semantic constraint consisting of abstract entities and events shown in Table 1.

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

```
ENT_1 conducts OBJ_1 in LOC_1.
LOC_1 become 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 to generate contextually diverse scenarios. Table 2 shows three valid instantiations:

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

Table 2: **Abstract Mapping:** Three valid lexical instantiations of the abstract constraint from Table 1.

From these three valid instantiations, here are some examples in complete text format:

> “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 above examples follow the same syntactic structure, but the constraint also permits different structural realizations, such as:

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

3 Formal Problem Definition

3.1 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 $R^{(p)} \in \mathcal{R}$, we formally defined it as a tuple

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

where we have the following three spaces and a role assignment:

Entity Space:

$$V^{(p)} = \left\{v_1^{(p)}, v_2^{(p)}, \ldots, v_{n_v}^{(p)}\right\} \subset \mathcal{V}$$

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

Predicate Space:

$$S^{(p)} = \left\{s_1^{(p)}, s_2^{(p)}, \ldots, s_{n_s}^{(p)}\right\} \subset \mathcal{S}$$

represents a finite set of n_s predicates specific to the representation $R^{(p)}$, where $n_s = \left|S^{(p)}\right|$ denotes the cardinality of $S^{(p)}$ 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 $s_j^{(p)} \in S^{(p)}$, we define $\psi_j^{(p)} : \Xi \to V^{(p)} \cup \{\emptyset\}$ as a partial function mapping semantic roles to entities, where

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

3.2 Instantiation

For a given abstract representation $R^{(p)}$, we generate k instantiation $\{T_1^{(p)}, T_2^{(p)}, \ldots, T_k^{(p)}\}$. Each generated instantiation $T_i^{(p)}$ is represented as a sequence of position-aligned words, where each word represents a concrete lexical instantiation of an abstract semantic element:

$$T_i^{(p)} = \{w_{i,1}^{(p)}, w_{i,2}^{(p)}, \ldots, w_{i,n}^{(p)}\}$$

where n denotes the number of position-aligned words[1] in $T_i^{(p)}$, and each $w_{i,j}^{(p)}$ represents the i-th instantiated word at position j. For instance, based on Table 2, Instantiation 1 can be read as $T_1^{(p)} = \{\text{“Armin”}, \text{“fieldwork”}, \text{“country”}, \ldots\}$.

4 Definition of the CDM Metric

In the remainder of this paper, we fix an arbitrary abstract representation $R \in \mathcal{R}$ and omit the superscript (p) for notational clarity, as all definitions are understood to be with respect to this fixed representation unless stated otherwise.

We introduce a geometric decomposition that analyzes word-level semantic changes relative to the overall sentence-level shift. For each abstract entity index $j \in [n]$, we measure the semantic diversity among the k words corresponding to the position j across the k generated texts.

4.1 Centroid Direction

For each word $w_{i,j}$ in a sentence $T_i = \{w_{i,1}, w_{i,2}, \ldots, w_{i,n}\}$, we apply a word embedding function $\phi : \mathcal{W} \to \mathbb{R}^d$ that maps words from word space $\mathcal{W}$ to d-dimensional normalized embeddings:

$$\hat{e}_{i,j} = \frac{\phi(w_{i,j})}{\|\phi(w_{i,j})\|_2}, \quad \forall i \in [k], \forall j \in [n].$$

We define the centroid of the sentence as the mean of its normalized word embeddings:

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

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

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

4.2 Geometric Decomposition

For each aligned word pair $(w_{i,j}, w_{l,j})$ at position j, we decompose the semantic change into two orthogonal components relative to the centroid direction $\eta(i,l)$.

[1] Note that the number of position-aligned words n is consistent across all instantiation T_i generated from the same abstract representation R, as n corresponds to the number of fillable elements in the semantic constraint.

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

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

The **Directional Component** represent the word change in the direction of the centroid direction defined as

$$\tau_j(i,l) = \left|(\Delta_j(i,l) \cdot \eta(i,l))\, \eta(i,l)\right|.$$

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

$$\nu_j(i,l) = \left\|\Delta_j(i,l) - \tau_j(i,l)\right\|_2.$$

The total changes of the geometric decomposition are given by

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

where ζ is the amplification factor and λ is the balance parameter weighting the directional component; both are trainable parameters.

The final score for each aligned word pair $(w_{i,j}, w_{l,j})$ at position j are given by

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

where γ is the steepness parameter controlling the nonlinearity of the transformation.

4.3 Contextual Diversity Metric

A specific distance per entity index 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 position distance across all the word embeddings $\hat{e}_{i,j}, \forall j \in [n]$, which can be defined as

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

5 Methods for Validating CDM

To validate our CDM, we conduct a controlled experiment using synthetic scenario generation dataset. The objective is to demonstrate that our diversity metric meaningfully distinguishes between texts that express the same semantic content in different contexts.

5.1 Dataset

Table 3 presents the dataset statistics used in the experimentation.

Statistic	Value
Number of abstract representations	8.0
Average entities per representation	9.4
Average predicates per representation	7.0
Average words per generated T_i	43.4

Table 3: Dataset statistics

For each abstract representation R, we construct 3 text realisations that faithfully express the constraints specified in R. The three texts are:

- **Reference Text (RF):** The original labeled text from which the abstract SRL representation was derived. All reference texts belong to the academic domain, as they were sampled from academic scenario descriptions.
- **High Diversity (HD):** Maximizes contextual distance from the reference text by instantiating the abstract representation in a distinct semantic domain (e.g., business, archaeology, journalism instead of academia).
- **Low Diversity (LD):** Minimizes contextual distance from the reference text by instantiating the abstract representation in a closely related semantic domain (e.g., a different academic scenario).

From these three text realizations, 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.

5.2 Word Embeddings

In our experiments, we instantiate ϕ using four different pre-trained embedding models to evaluate

the robustness of our diversity metric across varied embedding approaches, as shown in Table 4.

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

Table 4: **Word Embedding Models**: Word embedding models used for diversity evaluation. Static models use fixed word vectors. Contextual models generate context-dependent representations.

5.3 Statistical Analysis

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

5.3.1 Two-Sided Wilcoxon Signed-Rank Test

We first test whether the position-level differential distances differ 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.

5.3.2 One-Sided Wilcoxon Signed-Rank Test

Since we expect P_1 to produce 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 score higher than P_2 across different semantic positions.

5.3.3 Cohen's d Effect Size

We compute Cohen's d to observe the Effect size, which quantifies the magnitude of that difference:

$$d = \frac{\bar{\Delta}}{s_\Delta}, \quad \bar{\Delta} = \frac{1}{m \times n} \sum_{p=1}^{m} \sum_{j=1}^{n} \Delta_j\left(\delta\right)$$

where $\bar{\Delta}$ is the mean differential distance across all positions and s_Δ is the standard deviation. Following standard conventions, we interpret $|d| \geq 0.2$ as small, $|d| \geq 0.5$ as medium, and $|d| \geq 0.8$ as large effects. A large effect size would indicate that our metric produces substantially different scores for the P_1 and P_2.

5.4 Baseline

Since each group consists of only two texts, we can apply existing sentence-level embedding and diversity metrics to evaluate CDM's ability to distinguish the two groups. We compare CDM against the following baseline methods:

- **BERTScore (Zhang et al., 2020):** A learned metric that computes token-level similarity using contextualized 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.
- **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 models to distinguish between the two groups by measuring: (1) **Accuracy:** the proportion of scenarios where $P_1 > P_2$, and (2) **Sum Difference:** the total value of $P_1 - P_2$ across all scenarios, quantifying each method's power to distinguish between high-diversity and low-diversity pairs.

6 Statistical Test Results

As shown in Table 5, both two-sided and one-sided Wilcoxon signed-rank tests yield p-values well below the significance threshold ($\alpha = 0.05$) across all embedding models, 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.

Model	Statistical Tests		
	Wilcoxon (Two-sided)	Wilcoxon (One-sided)	Cohen's d
FastText	8.02×10^{-5}	4.01×10^{-5}	0.508
GloVe	8.02×10^{-3}	4.01×10^{-3}	0.292
MiniLM	1.42×10^{-3}	7.11×10^{-4}	0.402
Word2Vec	1.33×10^{-2}	6.67×10^{-3}	0.339

Table 5: **Statistical Test Results:** Position-level differential distance test results across embedding models and distance metrics. All p-values are from Wilcoxon signed-rank tests with $m \times n \approx 130$ observations. Cohen's d values indicate effect sizes.

Of the four experimented models (Table 5), FastText demonstrates the strongest statistical evidence, achieving the lowest p-values and largest effect size (Cohen's $d = 0.508$), while Word2Vec shows the weakest, though still highly significant results. All embedding models achieve small-to-medium effect sizes, ranging from 0.292 to 0.508. FastText consistently produces the largest effect size (0.508), followed by Minimal Language Model (MiniLM) (0.402), Word2Vec (0.339), and Global Vectors for Word Representation (GloVe) (0.292). While effect sizes vary across models, all demonstrate meaningful discrimination between the two conditions.

7 Performance Results

Among baseline methods, BERTScore achieves perfect discrimination accuracy (100%), correctly identifying $P_1 > P_2$ in all eight scenarios, followed by Sentence Similarity at 87.5%, Distinct-1 and Distinct-2 both at 75.0%, and Self-BLEU at 62.5%, as shown in Table 6. However, BERTScore produces the smallest sum difference (+0.132) among all methods, indicating weak discriminative power despite perfect accuracy. On the other hand, CDM variants demonstrate strong performance, where three CDM variants (FastText, MiniLM, and Word2Vec) achieve perfect accuracy (100%), while (GloVe) achieves 87.5% accuracy.

As shown in Table 6, CDM produces larger sum difference than the best baseline by sum difference, Sentence Similarity, and outperforms the difference in 6 out of 8 scenarios. Critically, CDM (FastText) produces the largest sum difference (+0.998) across all methods, representing a 7.6× improvement over BERTScore (best accuracy baseline) and a 1.5× improvement over Sentence Similarity (best sum difference baseline).

Different embedding models yield varying discriminative power in CDM. FastText achieves the highest sum difference with perfect accuracy (+0.998, 100%), followed by GloVe (+0.881, 87.5%), MiniLM (+0.845, 100%), and Word2Vec (+0.624, 100%). Despite GloVe's slightly lower accuracy, it outperforms both MiniLM and Word2Vec in sum difference, suggesting that accuracy alone does not fully capture discriminative strength. FastText's superior performance aligns with its statistical test results (Table 5), where it demonstrated the strongest effect size (Cohen's $d = 0.508$).

Several scenarios prove challenging for the baseline methods. In S03, CDM variants consistently produce strong differences (ranging from +0.123 to +0.256), with CDM (GloVe) achieving +0.256, approximately $10.7\times$ larger than the best-performing baseline in this scenario, BERTScore (+0.024). In S05, all CDM variants maintain positive differences (ranging from $+0.051$ to $+0.101$), while all baseline models produce near-zero or negative values (ranging from $+0.002$ to -0.199). Even though one CDM variant (GloVe, -0.022) produces a negative value in S08, our best-performing variant in this scenario (Word2Vec, $+0.134$) outperforms the best baseline, Sentence Similarity ($+0.120$).

8 Discussion

8.1 Principal Results

Statistical Test: Our controlled experiments 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 metric successfully captured these intended diversity patterns, with small-to-medium effect sizes ($|d|$ ranging from 0.292 to 0.508) demonstrating

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 6: **Baseline Comparison:** Diversity scores across all scenarios for each metric. P_1 represents high-diversity groups, P_2 represents low-diversity groups. Positive differences indicate $P_1 > P_2$. **Acc** shows the percentage of scenarios where P_1 outperforms P_2. For CDM, parentheses indicate the embedding model used. All metrics are scaled to $[0, 1]$ representing dissimilarity (1 = completely dissimilar, 0 = identical).

that the differences were practically meaningful. The convergence of results across four diverse embedding, demonstrates that CDM captures genuine semantic diversity patterns independent of the underlying embedding architecture.

The range of effect sizes (0.292 to 0.508) also provided insight into practical utility: while all fell within the small-to-medium range by Cohen's conventions, the $1.74\times$ difference between the strongest (FastText) and weakest (GloVe) models demonstrated that embedding choice meaningfully influenced discriminative power. This suggested that while CDM's core principle was sound across embeddings, optimization could be done by selecting embeddings aligned with their specific diversity assessment needs.

Performance: Baseline comparisons validated that three CDM variants achieved perfect discrimination accuracy (100%), with CDM (FastText) producing the largest sum difference (+0.998) among all methods, representing a $7.6\times$ improvement over the best-performing baseline accuracy metric, BERTScore (+0.132), and a $1.5\times$ improvement over the best-performing baseline by sum difference, Sentence Similarity (+0.661). The perfect accuracy indicated that CDM successfully distinguished between high-diversity and low-diversity pairs across all scenarios, while the large sum difference demonstrated that CDM provided strong discriminative signals suitable for both evaluation and optimization in scenario generation tasks.

8.2 Research Significance Statement

BERTScore's performance with perfect accuracy (100%) yet minimal discrimination (+0.132) reveals a critical limitation of prior work: achieving the correct ranking is insufficient for diversity metrics. The near-zero differences (+0.002 to +0.034) suggest that BERTScore captures diversity so weakly that it would be ineffective as an optimization target. This highlights why CDM's larger margins are practically important as they provide clear gradient signals for generation systems.

The divergence between accuracy and sum difference across CDM variants reveals an important insight: FastText achieves both perfect accuracy (100%) and maximum discrimination (+0.998), while GloVe trades some accuracy (87.5%) for strong discrimination (+0.881, second-highest). This suggests that discriminative power and classification accuracy, while related, capture different aspects of metric quality, the former being more important for gradient-based optimization.

9 Conclusion

We introduce CDM as a position-aligned metric that quantifies contextual diversity by decomposing word-level semantic changes into directional and orthogonal components relative to sentence-level shifts across scenario realizations from abstract semantic role labeling representations.

Our results demonstrate that CDM successfully quantifies contextual diversity with statistical significance ($p < 0.05$) and small-to-medium effect sizes (Cohen's d: 0.292–0.508), achieves perfect classification accuracy (100%) and produces discriminative power $1.5\times$ stronger than the best baseline metric.

Our CDM addresses an understudied problem in controllable text generation which is to measuring contextual diversity under fixed semantic constraints, and providing quantifiable measures that existing metrics cannot capture.

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) optimizing 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, as a contextual diversity measure, quantitative metric for contextual diversity but has not yet integrated it into automated generation systems. Future work could explore incorporating CDM into LLM generation pipelines building on recent controllable generation frameworks. This could take place either as a decoding constraint (e.g., diversity-promoting beam search) or as a training objective. Our validation experiments use only 8 abstract representations, and larger-scale validation across more diverse domains would strengthen confidence in the metric's generalizability.

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(null):993–1022.

Xuan Cai, Xuesong Bai, Zhiyong Cui, Danmu Xie, Daocheng Fu, Haiyang Yu, and Yilong Ren. 2025. Text2scenario: Text-driven scenario generation for autonomous driving test. *Preprint*, arXiv:2503.02911.

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.

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, Randy Bennett, and 1 others. 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.