Evaluating the Evaluation of Diversity in Commonsense Generation

Anonymous ACL submission

Abstract

In commonsense generation, given a set of input concepts, a model must generate a response that is not only commonsense bearing, but also capturing multiple diverse viewpoints. Numerous evaluation metrics based on form- and content-level overlap have been proposed in prior work for evaluating the diversity of a commonsense generation model. However, it remains unclear as to which metrics are best suited for evaluating the diversity in commonsense generation. To address this gap, we conduct a systematic meta-evaluation of diversity metrics for commonsense generation. We find that form-based diversity metrics tend to consistently overestimate the diversity in sentence sets, where even randomly generated sentences are assigned overly high diversity scores. We then use an Large Language Model (LLM) to create a novel dataset annotated for the diversity of sentences generated for a commonsense generation task, and use it to conduct a meta-evaluation of the existing diversity evaluation metrics. Our experimental results show that content-based diversity evaluation metrics consistently outperform the form-based counterparts, showing high correlations with the LLM-based ratings. We recommend that future work on commonsense generation should use content-based metrics for evaluating the diversity of their outputs.

1 Introduction

Commonsense reasoning—the ability to make plausible assumptions about ordinary scenarios—is a core requirement for robust Natural Language Generation (NLG) systems (Lin et al., 2020). In the task of Generative Commonsense Reasoning (GCR), an NLG model is expected to generate sentences that are both *quality-bearing* (i.e. logically coherent and commonsense-aware) and *diverse* (i.e. offering varied perspectives on the same input concepts) (Liu et al., 2023a; Yu et al., 2022; Hwang et al., 2023).

Inputs: {Walk, Dog, Take, Park, Couple}

Set 1:
The **couple takes** their **dog** for a **walk** in the **park**.
The **couple** decided to **take** a **walk** in the **park** without **taking** their **dog**.
Every evening, the **couple takes** a **walk** in the **park** with their **dog**.
The **dog** enjoys when the **couple takes** it for a **walk** in the **park**.

self-BLEU-3: 0.486 **VS-embed-0.5 : 2.689** ✅

Set 2:
A **couple take** their **dog** for a **walk** in the **park** every morning.
Every morning, the **couple** and their **dog take** a **walk** in the **park**.
Every evening, the **couple takes** a **walk** in the **park** with their **dog.**
In the **park**, a **walk** is **taken** every evening by the **couple** with their **dog.**

self-BLEU-3: 0.593 ❌ **VS-embed-0.5:** 1.916

Figure 1: An example from the CommonGen (Lin et al., 2020) dataset comparing two sets of generated sentences. `self-BLEU-3` indicates Set-2 to be more diverse, which simply repeats near-identical paraphrases. In contrast, `Vendi Score (VS)-embed-0.5` aligns well with the notion of meaningful textual diversity.

While recent neural architectures have significantly improved the quality of commonsense generation, reliably evaluating the diversity of generated outputs remains an open challenge. Quality evaluation typically relies on comparing generated outputs against a set of human-written reference sentences using metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or SPICE (Anderson et al., 2016). A GCR method that produces outputs that have a high overlap with human-written reference sentences is considered to be of *high quality*. In contrast, diversity is assessed by comparing the outputs among themselves. A variety of diversity metrics have been proposed (Li et al., 2016; Zhang et al., 2024) and can be broadly categorised into two groups: **form-based** vs. **content-based**. Form-based diversity metrics such as self-BLEU (Zhu et al., 2018) and distinct (Li et al., 2016), measure the token/word overlap between pairs of sentences using n-grams, whereas content-based diversity metrics such as self-CosSim (Cox et al., 2021) and Vendi-Score (Friedman and Dieng, 2023) capture semantic variations using sentence embeddings.

A central question arises: *Which diversity metrics best capture meaningful variations in commonsense generation, and under what conditions?* For instance, as shown in Figure 1, given the five input concepts *walk, dog, take, park* and *couple*, a GCR method must produce sentences that contain all of the input concepts and their diverse commonsense relations. Although both Set-1 and Set-2 contain commonsense-making sentences covering all input concepts, Set-2 contains direct paraphrases or random word-order shuffles. Consequently, Set-2 is less diverse compared to Set-1. However, the form-based diversity metrics (e.g. self-BLEU3) assign high diversity scores to Set-2 than to Set-1, overestimating the diversity in GCR. As we later see in our meta-evaluations (§ 5.3), form-based diversity metrics tend to assign high diversity scores even for randomly generated nonsensical sentences, which is counter-intuitive. On the other hand, content-based diversity metrics (e.g. VS-embed-0.5) seem less susceptible to such issues and correctly predict Set-1 to have a higher diversity than Set-2.

We conduct a comprehensive meta-evaluation of covering 12 diversity metrics for GCR using three standard GCR datasets. For this purpose, we create a large-scale diversity-annotated dataset. Prior work studying diversity (Tevet and Berant, 2021) in NLG has shown difficulty in obtaining reliable diversity ratings via crowdsourcing. However, Zhang et al. (2024) showed that LLMs could be used to evaluate the diversity in GCR with a moderate-level of agreement with linguistically trained human annotators. We follow their work and create a dataset where an LLM provides a pairwise preference rating for two sets of sentences covering the same input concepts. A human evaluation on a subset of our dataset shows that the LLM-based diversity ratings to be well-aligned with the human judgments with an average accuracy of 79.4%.

Next, we measure the pairwise preference agreement between the LLM-based ratings and diversity metrics for high vs. low quality generations. We find that,

1. Form-based diversity metrics produce reliable evaluations for high quality generations, but often fail to distinguish genuine diversity for the lower-quality generations, and

2. Content-based metrics produce consistently reliable evaluations for both high and low quality generations.

Our datasets/code are submitted to ARR, and will be publicly released upon paper acceptance.

2 Related Work

Diversity in NLG: Diverse output generation is a critical requirement for many NLG applications (Tevet and Berant, 2021) such as storytelling (Li et al., 2018), question generation (Pan et al., 2019) and machine translation (Shen et al., 2019). Strategies proposed for improving diversity in NLG include sampling methods that prune the probability distribution over the next-token predictions such as nucleus sampling (Holtzman et al., 2019) and top-k sampling (Fan et al., 2018). Setting high temperature for the decoder (Peeperkorn et al., 2024) can sometimes increase the diversity in the generated output but must be done with care as it can decrease the quality (Zhang et al., 2024).

Diversity in GCR: Diversification in GCR presents an additional layer of complexity because we must generate both diverse as well as commonsense bearing outputs. Datasets such as CommonGen (Lin et al., 2020) and DimonGen (Liu et al., 2023a) provide a set of concepts and a set of sentences that describe various commonsense relations among those concepts, while ComVE (Wang et al., 2020) requires a GCR method to explain why a given counterfactual statement (*e.g.*, "A shark interviews a fish") does not make commonsense. Prior work in diversification for GCR has injected external knowledge from a knowledge graph (Yu et al., 2022; Hwang et al., 2023), retrieved diverse sentences from an external corpora (Liu et al., 2023a)), or use in-context learning to instruct an LLM (Zhang et al., 2024) to elicit diverse outputs. However, our goal in this paper is *not to propose diversification methods* for GCR, but to conduct *a meta-evaluation of existing metrics* proposed in prior work for evaluating the diversity of GCR.

Evaluating Quality in GCR: Quality metrics in GCR primarily assess coherence, logical consistency, and their correlation with human judgments (Sai et al., 2022; Yu et al., 2022). Popular metrics use n-gram overlaps (e.g. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)), which measure the lexical overlap between a generated text and a human-written reference. BLEU (Papineni et al., 2002), for instance, computes the mean n-gram precision of a candidate sentence against human-written references, while semantic metrics (e.g. SPICE (Anderson et al., 2016), BERTScore (Zhang et al., 2020)) capture semantic textual similarity. BERTScore (Zhang et al., 2020) uses contextualised word embeddings to measure the semantic overlap between tokens in paired sentences. Despite their wide use, quality metrics alone are insufficient for evaluating NLG tasks, especially in GCR.

Evaluating Diversity Metrics: Our work builds upon studies such as Tevet and Berant (2021), who used human annotations to assess diversity metrics in NLG. There are several important distinctions between their work and ours:

a) **Task-specific Focus:** They did not consider commonsense relations in the outputs they evaluate, which is an important requirement in GCR.

b) **Generation Variability:** They assumed the availability adjustable decoding parameters (e.g. temperature) to control diversity. However, Zhang et al. (2024) showed that simply increasing temperature can harm the quality of commonsense generation. Instead, we use controlled perturbations (e.g. random shuffling and LLM-based paraphrasing) to generate outputs with varying diversity.

c) **Annotation Methodology:** Whereas Tevet and Berant (2021) relied on crowdsourced human annotators—faced with low agreement and high cost—we leverage LLMs as reference-free annotators (Wang et al., 2023; Liu et al., 2023b; Fu et al., 2024). Recent studies have successfully used LLMs for evaluations in NLG tasks (Kocmi and Federmann, 2023; Liu et al., 2023b) and Zhang et al. (2024) reported a moderate level of agreement between human and LLM-based diversity ratings in GCR. Our own human evaluation confirms that LLM-based diversity ratings achieve 79.4% accuracy with expert human annotators.

In summary, while there has been extensive work on diversifying NLG outputs and evaluating quality in GCR, the evaluation of diversity metrics—especially in the context of commonsense generation—remains underexplored. Our work fills this gap by providing a systematic meta-evaluation of both form-based and content-based diversity metrics in GCR.

3 Diversity Metrics for GCR

In this section, we describe the diversity metrics used in our meta-evaluation.

Form-based Diversity: Self-BLEU (Zhu et al., 2018) measures the average n-gram overlap between all pairs of sentences within a set.[1] We use self-BLEU-3/4 (i.e. $n = 3, 4$) in our experiments. Inspired by ecology and quantum mechanics, VS (Friedman and Dieng, 2023) was proposed as a diversity metric in computer vision. VS is the exponential of the Shannon's entropy over the eigenvalues of the pairwise similarity (kernel) matrix of a set of sentences, computed using either the n-gram overlap or sentence embeddings (see Appendix A for further details.) Pasarkar and Dieng (2024) extended the original VS by introducing an order parameter q, which adjusts its sensitivity to the frequency of the items. A smaller q (e.g. $q = 0.5$) increases the sensitivity to larger variances, capturing diversity more effectively in imbalanced scenarios, while $q = \infty$ is more robust against the intraclass variance, focusing on the most dominant features. For the form-based diversity measurement using VS, the kernel matrix is constructed using a bag-of-n grams representation.

Distinct-k (Li et al., 2016) calculates the ratio of the unique k-grams to the total number of k-grams, and is one of the widely-used metrics for evaluating corpus diversity. It adjusts the bias towards generating longer sequences, ensuring that diversity is not artificially inflated by the sentence length. Similarly, Entropy-k quantifies the uniformity of the k-gram distribution within the text. Higher values for both Distinct-k and Entropy-k reflect greater diversity.

Content-based Diversity: To measure diversity at content level, self-CosSim (Cox et al., 2021) calculates the average pairwise cosine similarity between the generated sentences using their sentence embeddings. On the other hand, Chamfer Distance (Jones et al., 2006) measures diversity by calculating the average of the minimum pairwise distances between embeddings, reflecting proximity to the nearest neighbour (see Appendix B). We also use VS for content-based diversity, where the kernel matrix is built from sentence embeddings. For consistency across metrics, we use embeddings obtained via SimCSE (Gao et al., 2021).

4 Meta-Evaluation of Diversity Metrics

We propose an LLM-based annotation method for creating a diversity rated dataset for our meta-

[1] We subtract self-BLEU scores by 1, such that higher scores indicate greater pairwise diversity.

evaluation in § 4.1, and a method to create sentence sets with different quality levels from the CommonGen dataset in § 4.2.

4.1 LLM-based Diversity Annotation

A reliable diversity metric must align well with the human notion of diversity, independently of the quality of the generation. For example, randomly permuting the word order or including nonsensical words in a sentence are not considered by humans to be improving diversity. Therefore, a reliable diversity metric must also not assign high diversity scores for such cases. However, obtaining reliable human diversity ratings at scale is costly. Moreover, Tevet and Berant (2021) showed that human diversity judgments often conflate text quality and variety. Consequently, to conduct a large-scale meta-evaluation over existing diversity metrics, we elicit diversity ratings from an LLM. LLMs have been used as annotators for multiple NLG tasks (Wang et al., 2023; Liu et al., 2023b; Fu et al., 2024). In particular, Zhang et al. (2024) reported a moderate level of agreement between LLM and human diversity ratings in a GCR task.

We consider two types of diversities (Tevet and Berant, 2021) in our annotation:

Form-based Diversity: A diverse set of sentences must exhibit minimal lexical overlap, avoiding repetitive word usage while preserving clarity and fluency.

Content-based Diversity: A diverse set of sentences must exhibit distinct semantic content *centred on the same input*, ensuring that each sentence offers a different perspective on the topic rather than talking about unrelated topics.

We follow a two-step approach to mitigate any choice-ordering bias (Zheng et al., 2023; Wang et al., 2024a,b) when eliciting a pairwise preference order between two sets of sentences shown to the annotator LLM. First, we require that the annotator LLM assign a numerical score (1–5), where higher ratings indicate stronger diversity. We repeat this process multiple times by randomly ordering the sentences in each set shown to the **annotator LLM** as well as which set is shown first in the instruction. We then aggregate the LLM ratings and predict the set with the higher rating to be the more diverse set in the pair. We include eight few-shot examples rated by three human annotators following the same instructions as given to the annotator LLM to further improve the prompt.[2]

We use `GPT-4o` as the annotator LLM, which has shown superior performance in a broad range of annotation tasks.[3] Moreover, we conducted a human evaluation using a subset of sentence set pairs in our dataset to validate the LLM-based diversity ratings (further details described in Appendix D). From this human validation, we found that `GPT-4o` to have a high level of agreement with human diversity judgements with an average accuracy of 79.4%, which confirms the reliability of LLM-based annotations. An example of an LLM-based diversity judgement by `GPT-4o` is shown in Figure 2.

4.2 Candidate sets

Diversity would be of interest only when the generation quality is high. Therefore, a reliable diversity metric must be able to accurately evaluate the diversity of generations of varying qualities. For this purpose, we propose a method to create sentence sets that have varying levels of generation quality to be used later in our meta-evaluations. Specifically, we use the CommonGen dataset (Lin et al., 2020) where a GCR model must generate a coherent sentence that contains all of the input concepts, reflecting their commonsense relations. We use the official CommonGen test set, which includes 1,497 examples, each containing 3–5 input concepts on average. We create sets of sentences of *high* and *low* generation quality as described respectively in section 4.2.1 and section 4.2.2 by prompting three **generator LLMs**[4]: `GPT-4-turbo` (Achiam et al., 2023), `Llama3.1-8b` (Dubey et al., 2024), and `Qwen 2.5-14b` (Hui et al., 2024). Due to space limitations, we show the detailed instructions provided to the generator LLMs, an empirical quality evaluation, and example generations in Appendix C.

4.2.1 High-Quality Sentence Sets

We propose the following strategies to create sentence sets with high generation quality.

Default: Note that CommonGen was developed as a dataset for evaluating the quality and not diversity of GCR methods. Therefore, it contains only a small number of human-written sentences covering the input concepts in a test case. Moreover,

[2]Prompt shown in Appendix D.

[3]LLM Leaderboard

[4]To prevent any confusion with the `GPT-4o` that we used as the annotator LLM in § 4.1, we collectively call those models as the **generator LLMs**.

Query:

Set 1: Step by step, she sits down to watch her father catch fish on the shore
Step by step, she sits down and watches as her father catches fish.
On a sunny day, we sit by the river, watch the currents, and hope to catch some fish.
By the river on a sunny day, we sit and watch the currents, hoping to catch some fish.

Set 2: He sits on the dock to watch and catch fish as they swim up the steps.
Step by step, she sits down to watch her father catch fish on the shore.
On a sunny day, we sit by the river, watch the currents, and hope to catch some fish.
I sit on the stairs to watch the fish-catching competition as participants step forward one by one.

LLM Score "Diversity_Set1": **3**, "Diversity_Set2": **4**

LLM Reason

Set 1 has a diversity score of 3 due to noticeable redundancy. The first two sentences are nearly identical, both describing a step-by-step process of sitting to watch a father catch fish. The last two sentences also mirror each other, focusing on sitting by the river and observing currents with the hope of catching fish, offering limited exploration of the theme. In contrast, Set 2 scores a 4 for diversity as it introduces more varied scenarios. It includes unique elements like sitting on a dock to watch fish swim up steps and a fish-catching competition, enriching the theme with different perspectives.

Figure 2: An example of annotating for diversity using `GPT-4o` for two sets of sentences generated for the same input concepts. `GPT-4o` assigns a higher diversity rating for Set-2, indicating it to be more diverse than Set-1. LLM Reason is provided here for explanation purposes only and is not used in the meta-evaluation.

these human-written sentences do not adequately cover all possible commonsense bearing sentences that can be generated from the input concepts. To address this issue, we prompt the generator LLMs with the same instructions as given to the human annotators in CommonGen to generate four sentences for each test case. We call it the **Default** set of sentences for a test case.

Paraphrasing: We randomly select one or more sentences from the **Default** set and instruct the generator LLMs to create their paraphrases. We then replace the non-selected sentences in each Default set with the generated paraphrase sentences. We expect the diversity of a set of sentences to decrease when we include more paraphrasing sentences. Specifically, we consider three variants of this method. Let the Default set contain four sentences $\{A, B, C, D\}$, and a A^* be the paraphrase of A, selected randomly from the set. We then define: **Para-1** = $\{A, A^*, B, C\}$, **Para-2** = $\{A, A^*, B, B^*\}$, and **Para-3** = $\{A, A^*, A^{**}, B\}$.

4.2.2 Low-Quality Sentence Sets

To evaluate the ability of a diversity metric to accurately distinguish genuine diversity from nonsensical or random corruptions made to a sentence, we create a set of low generation quality sentences for each input concept set in CommonGen test dataset as follows.

Nonsensical: We prompt[5] the generator LLMs to produce sentences that are syntactically valid and include all of the input concepts, but do not make any commonsense or illogical.

NounShuff: We run a part-of-speech tagger and randomly shuffle nouns and pronouns within each sentence, while leaving other words unchanged. This process disrupts semantic consistency while retaining some semblance of syntactic framing, serving as an intermediate case of corruption.

RndShuff: We take each sentence from the **Default** set and randomly shuffle *all of the words* in it to produce sequences that are devoid of coherent sentence structure or meaning.

5 Experiments

5.1 Settings and Evaluation Metrics

To obtain statistically stable diversity ratings, we run the annotator LLM (i.e. `GPT-4o`) with the *temperature* set to 0.6, and average the results over five independent runs. All experiments are conducted on two (Nvidia A6000 and 4090) GPUs for `Qwen2.5-14B` and `Llama3.1-8B` models. For `GPT-4-turbo`, we use the OpenAI API, with the temperature set to 0 to increase determinism in the generations. We use 1024-dimensional[6] Sim-

[5] Prompt shown in Appendix C.
[6] `huggingface.co/princeton-nlp/sup-simcse-roberta-large`

	Diversity Metric	GPT-4-turbo	Qwen2.5-14B	Llama3.1-8B
Form	self-BLEU-3	48.4	50.7	52.7
	self-BLEU-4	49.0	51.9	53.0
	VS-ngram-0.5	49.2	57.7	56.1
	VS-ngram-1	49.0	57.8	56.2
	VS-ngram-inf	47.5	58.9	56.5
	Distinct-4	64.0	69.0	61.7
	Entropy-2	62.9	74.0	62.5
Content	Chamfer	80.6	78.9	71.9
	self-cosSim	76.9	80.0	71.9
	VS-Embed-0.5	**80.7**	80.8	**73.2**
	VS-Embed-1	79.3	**81.1**	73.1
	VS-Embed-inf	76.0	79.9	71.9

Table 1: Meta-evaluation of the accuracy of the diversity metrics on the CommonGen test dataset with each of the generator LLMs.

	Diversity Metric	ComVE	DimonGen
Form	self-BLEU-3	77.3	59.7
	self-BLEU-4	76.9	59.4
	VS-ngram-0.5	76.7	60.0
	VS-ngram-1	77.0	59.8
	VS-ngram-inf	77.2	58.8
	Distinct-4	73.8	62.2
	Entropy-2	74.2	62.2
Content	Chamfer	77.0	**67.8**
	self-cosSim	76.4	66.6
	VS-Embed-0.5	**77.4**	67.2
	VS-Embed-1	76.8	67.6
	VS-Embed-inf	76.4	66.6

Table 2: Accuracy of diversity metrics on ComVE and DimonGen datasets.

CSE (Gao et al., 2021) sentence embeddings for all content-based diversity metrics.

We define the *accuracy* of a target diversity metric as the percentage of pairwise decisions that agree with those of the annotator LLM. For example, given a pair of sentence sets $(\mathcal{S}_1, \mathcal{S}_2)$, if both the annotator LLM and the target diversity metric consider $\mathcal{S}_1$ to be more diverse than $\mathcal{S}_2$, it is counted as a correct prediction. To prevent diversity evaluations from being influenced by the quality of the sentence sets, we ensure that both sentence sets in a pair to have the same generation quality (i.e. both sets must be either high quality or low quality). Moreover, to ensure meaningful comparisons, we filter out any sentence set pairs where the annotator LLM's average diversity ratings differ by less than 0.5. After this filtering step, the resulting sentence pair sets generated with `GPT-4-turbo`, `Llama3.1-8B`, and `Qwen-2.5-14B` and used for evaluations contain, respectively, 1414, 1916, and 1864 instances.

5.2 Meta-Evaluation of Diversity Metrics

Table 1 shows the accuracy of form-based (top group) vs. content-based (bottom group) GCR diversity metrics on the CommonGen dataset. We observe that content-based diversity metrics—– specifically self-cosSim, Chamfer, and VS-Embed variants—consistently achieve higher accuracy than form-based diversity metrics such as the corpus-level diversity metrics (e.g. Entropy, Distinct) or the n-gram-based diversity metrics (e.g. self-BLEU, VS-n-gram variants) across all generator LLM outputs. In particular, VS-Embed-0.5 and VS-Embed-1 consistently report the best accuracy, suggesting that content is more important than form when evaluating diversity in GCR. Form-based metrics primarily focus on lexical overlap, overlooking the deeper semantic nuances that characterise the diversity. Although Entropy and Distinct reflect some aspects of overall lexical variety and frequency distributions, they fail to capture semantic richness. Even when these metrics sometimes outperform self-BLEU, they still fall short of content-based metrics.

To ensure our findings generalise beyond CommonGen, we extend the meta-evaluation to two additional commonsense generation datasets: ComVE (Wang et al., 2020) and DimonGen (Liu et al., 2023a). ComVE requires a GCR method to explain why a counterfactual statement is nonsensical, while DimonGen focuses on generating diverse sentences describing relationships between two given concepts. Both tasks require outputs that are diverse and commonsense-bearing. Zhang et al. (2024) provide three sets of generated sentences for each dataset, along with a pre-evaluation of output quality. We compare each pair of sentence sets generated for the same input using the diversity ratings returned by our annotator LLM (i.e. `GPT-4o`), and contrast these with the diversity scores produced by each target metric, as shown in Table 2.

Consistent with the trends observed on CommonGen, **content-based metrics** (e.g. VS-Embed-0.5, Chamfer) consistently achieve the highest agreement with `GPT-4o` on both ComVE and DimonGen. For example, VS-Embed-0.5 performs best on ComVE, whereas Chamfer excels on DimonGen. Although form-based metrics show competitive accuracies on the ComVE dataset, their performance drops on DimonGen. These findings confirm that content-based metrics offer a more reliable and consistent approach for evaluating

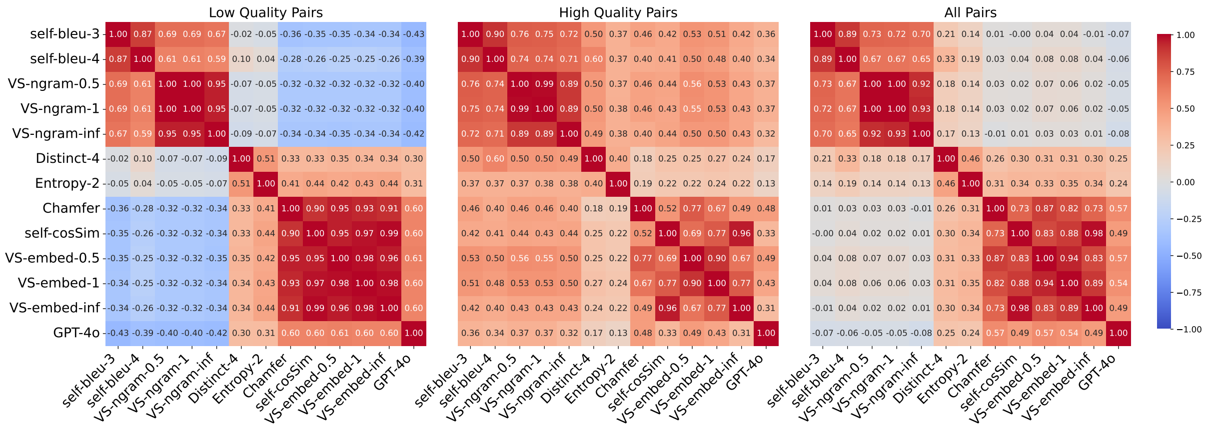


Figure 3: Inter-annotator agreement (measured using Cohen's Kappa) between two diversity metrics when used to make pairwise preference orderings for sentence sets generated for the same input concepts in CommonGen test cases. Agreement with the annotator LLM (i.e. `GPT-4o`) is also shown.

		GPT-4-turbo		Qwen2.5-14b		Llama3.1-8b	
	Diversity Metric	**High**	**Low**	**High**	**Low**	**High**	**Low**
Form	self-BLEU-3	73.5	27.6	68.4	35.3	66.6	39.8
	self-BLEU-4	72.0	30.0	67.1	38.7	64.3	42.5
	VS-ngram-0.5	73.7	28.8	69.7	47.2	66.5	46.5
	VS-ngram-1	73.4	28.8	69.5	47.6	66.6	46.8
	VS-ngram-inf	71.0	27.8	69.7	48.0	67.0	46.8
	Distinct-4	61.7	65.9	58.6	79.4	56.3	66.6
	Entropy-2	59.2	65.9	57.0	88.5	49.6	**74.4**
Content	Chamfer	**80.2**	80.8	67.5	**88.9**	73.6	70.4
	self-cosSim	72.3	80.7	71.7	87.2	74.4	69.6
	VS-Embed-0.5	**80.2**	**81.1**	72.3	88.2	**76.9**	69.8
	VS-Embed-1	77.7	80.6	**73.0**	88.1	**76.9**	69.5
	VS-Embed-inf	71.2	80.7	71.5	87.3	74.4	69.6

Table 3: Accuracy of diversity metrics across different levels of quality in sentence sets, generated by three generator LLMs. The content-based diversity metrics consistently perform better than the form-based metrics.

text diversity, especially in diverse commonsense generation tasks. While form-based metrics have close alignment with content-based metrics on the ComVE dataset, their performance is not always consistent (see Appendix F).

5.3 Diversity Metrics and Generation Quality

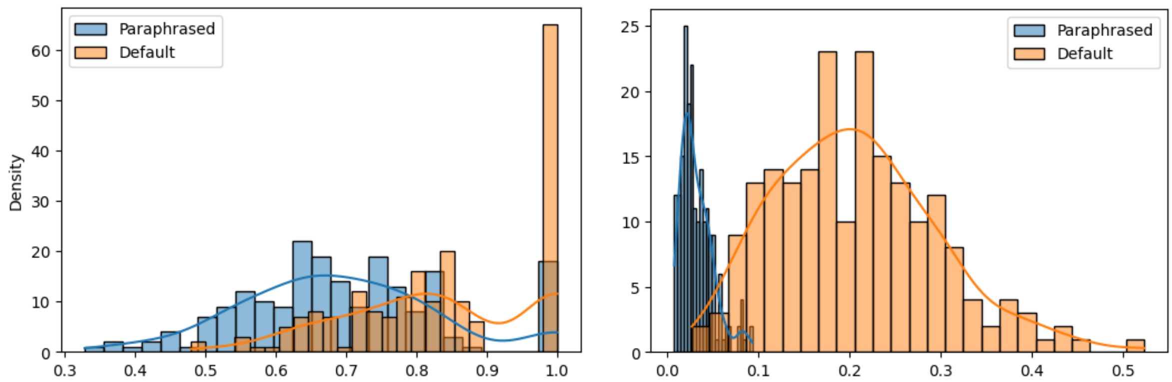
In Table 3, we conduct a meta-evaluation of diversity metrics for their ability to reliably estimate diversity in both high and low quality generations. We see that form-based metrics perform particularly well when the generation quality is high, however, their accuracy drops drastically (even below 40%) for low quality sets, demonstrating their sensitivity to inherent noise in the n-gram overlaps. In contrast, content-based metrics maintain consistently high accuracy, regardless of generation quality. In particular, VS-Embed-0.5 and VS-Embed-1 approach or exceed 70% accuracy in all comparisons, even for shuffled or nonsensical scenarios, demonstrating statistically significant improvements (Appendix E) over form-based metrics.

We treat each diversity metric as an *annotator* that provides a preference ordering for diversity between two sentence sets, and we measure their pairwise agreements. We use Cohen's Kappa (shown in Figure 3) for this purpose, which is known to be less sensitive to class imbalance, and more reflective of true, non-random agreement. For high quality sets, most metrics achieve fair to substantial levels of agreement, reflecting strong consistency. However, agreements vary considerably in low quality sets. Content-based metrics such as Chamfer, self-cosSim, and VS-Embed variants exhibit near-perfect agreement with each other and maintain Kappa values exceeding 0.6 with `GPT-4o`. Conversely, form-based metrics (e.g. self-BLEU) show poor agreement with `GPT-4o` in low quality sets with negative Kappa values indicating that the observed agreement between these form-based metrics is lower than would be expected by chance. Moreover, the agreements between form- and. content-based metrics remain low, underscoring fundamental differences in how these metrics measure diversity. Notably, Distinct-4 and Entropy-2—although also use n-grams—are less likely to overemphasise repeated phrases or minor word swaps and show a moderate level of agreement with content-based metrics even for low quality sets.

Table 4 shows the average diversity score reported by each metric over the sets of sentences generated from `GPT-4-turbo` according to the high and low quality preserving methods described in

Method	Form-based Diversity⇑							Content-based Diversity⇑				
	self-BLEU-3	self-BLEU-4	VS-ngram-0.5	VS-ngram-1	VS-ngram-inf	Distinct-4	Entropy-2	self-CosSim	Chamfer	VS-Embed-0.5	VS-Embed-1	VS-Embed-inf
High quality candidate sets												
Default	79.53	87.63	3.90	3.79	2.60	93.04	9.52	26.81	20.09	2.67	2.01	1.26
Para-1	73.37	81.86	3.86	3.72	2.48	90.60	9.42	22.04	12.44	2.35	1.76	1.20
Para-2	64.24	73.98	3.80	3.62	2.38	90.93	9.60	20.03	3.09	2.08	1.60	1.18
Para-3	63.32	71.88	3.77	3.57	2.26	89.90	9.66	17.50	9.54	2.08	1.55	1.15
Low quality candidate sets												
Nonsensical	80.96	89.79	3.90	3.81	2.60	90.26	9.12	42.02	35.18	2.59	2.63	2.52
NounShuff	89.06	95.19	3.93	3.87	2.77	98.06	9.82	28.83	22.53	2.77	2.09	1.28
RndShuff	96.75	99.05	3.95	3.89	2.84	99.94	10.23	27.57	21.34	2.72	2.04	1.27

Table 4: Average diversity score of each metric on sentence sets generated using the methods described in § 4.2.



(a) Distribution of self-BLEU-3 scores

(b) Distribution of Chamfer scores

Figure 4: Distribution of diversity scores for self-BLEU-3 (form-based) and Chamfer Distance (content-based) for **Default** and **Paraphrased** high-quality sentence sets. In self-BLEU-3, the two distributions have a high overlap, whereas in Chamfer they are well-separated. This indicates that the Chamfer metric can better distinguish more diverse **Default** sentence sets from the less diverse **Paraphrased** sentence sets than self-BLEU-3.

§ 4.2. For the high quality candidates, as expected, we see that the diversity decreases from the **Default** set as we paraphrase more sentences, as measured by all metrics. We also find that, on average, all metrics assign higher diversity scores to low quality generations than to high quality generations. This is because a random set of sentences could appear to be diverse, covering distinct topics, at both the form and content. This observation highlights an important limitation of existing GCR diversity evaluation metrics: diversity should *not* be evaluated without considering quality. A promising future research direction would be to develop an evaluation metric for GCR that simultaneously incorporates both quality and diversity aspects.

Figure 4 compares the distribution of diversity scores assigned by self-BLEU-3 (form-based) versus Chamfer (content-based) for 200 randomly sampled sentence sets from **Default** and **Paraphrased** (using **Para-2**) high-quality candidate sets. Sentence sets in **Paraphrased** are constructed to be less diverse compared to those in **Default**. We use Kernel Density Estimation (Rosenblatt, 1956) to interpolate the distributions from the frequency histograms. We see that the two distributions for self-BLUE-3 in Figure 4a to have a high overlap, demonstrating its inability to correctly separate high diversity generations in **Default** from the less diverse generations in **Paraphrased**. On the other hand, the two distributions for Chamfer in Figure 4b exhibit a relatively smaller overlap, indicating that Chamfer assigns relatively higher diversity scores to the sentence sets in **Default** than those in **Paraphrased**.

6 Conclusion

We presented a comprehensive meta-evaluation of diversity metrics for commonsense generation, revealing that content-based metrics consistently align with human judgments while form-based metrics tend to overestimate diversity, especially in low-quality generations. Our experiments across multiple datasets demonstrate that metrics such as VS-Embed and Chamfer provide a more robust and reliable assessment of semantic diversity. These findings underscore the importance of incorporating content-level analysis in evaluating commonsense generation. Future research should build on these insights to further enhance the robustness and interpretability of GCR.

7 Limitations

The experiments conducted in this paper were limited to English, a morphologically limited language. Although we would like to extend our meta-evaluation to other languages, we were limited by the lack of availability of commonsense reasoning datasets for languages other than English. In particular, CommonGen (Lin et al., 2020), ComVE (Wang et al., 2020), and DimonGen (Liu et al., 2023a) datasets are specifically designed for evaluating diversified commonsense reasoning only in English. We note however that both form- and content-based diversity metrics considered in our work are not limited to English, and can be easily extended to other languages with suitable tokenisers or multilingual sentence embedding models. For example, a single Kanji character in languages such as Japanese or Chinese can carry meaning on its own, and even n-gram overlap measures defined over character sequences can capture some level of meaning retention between a generated and a reference set of sentences. Therefore, we believe it would be important to conduct similar meta-evaluation for the diversity metrics in commonsense generation for other languages before selecting an appropriate evaluation metric. We hope that the methodology we propose in this paper will be exemplary in such future work.

Our work evaluates diversity metrics primarily within GCR tasks. The candidate sets used in this study were pre-evaluated for quality using official scripts (for CommonGen) or prior work (for ComVE and DimonGen). We use three LLMs as our generative models, a closed model (`GPT4-turbo`) and two open-source models (`Llama3.1-8B` and `Qwen2.5-14B`) to promote the reproducibility of our results, which are reported using multiple publicly available benchmarks. Of course, there is a large number of LLMs being developed, trained on different pre-train data compositions, architectures, parameter sizes and fine-tuned for a plethora of tasks. It is practically impossible to conduct all available LLMs in a conference paper due to the sheer number and the computational costs.

We used `GPT-4o` as the sole LLM-based diversity annotator. Although the prompts and instructions are adaptable to other models, we chose `GPT-4o` due to its superior performance in a range of NLG tasks. Moreover, in our human evaluation, conducted over a subset of the `GPT-4o` rated sentence sets, human judges found those annotations to be of high accuracy (i.e. 79.4% accuracy as shown in § D.1). Therefore, we consider `GPT-4o` to offer a scalable and robust alternative for annotating diversity in sentence sets. However, using LLMs that are comparable or superior to `GPT-4o` could further validate our findings.

8 Ethical Concerns

All experiments conducted in this study use publicly available datasets, CommonGen, ComVE, and DimonGen. To the best of our knowledge no personally identifiable information is included in those datasets and no ethical issues have been reported. The human annotators who participated in our evaluation were over 18 years old adults and have given informed consent to use their diversity annotations for academic research purposes.

It is noteworthy that LLMs have been reported to encode social biases such as gender or racial biases (Kaneko and Bollegala, 2021; Nangia et al., 2020; Kaneko et al., 2022). Although we evaluated quality and diversity of the generations made by LLMs in this work, we have not evaluated how social biases are reflected in their generations. Therefore, it is important to also evaluate the social biases in the diverse LLM generations before a diversification method for GCR is deployed in an NLG application.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

C J Clopper and E S Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian Von Der Weth, and Brian Y. Lim. 2021. Directed diversity: Leveraging language embedding distances for collective creativity in crowd ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–35.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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)*. Association for Computational Linguistics.

Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

EunJeong Hwang, Veronika Thost, Vered Shwartz, and Tengfei Ma. 2023. Knowledge graph compression enhances diverse commonsense generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 558–572, Singapore. Association for Computational Linguistics.

Mark W Jones, J Andreas Baerentzen, and Milos Sramek. 2006. 3d distance fields: A survey of techniques and applications. *IEEE Transactions on visualization and Computer Graphics*, 12(4):581–599.

Masahiro Kaneko and D Bollegala. 2021. Unmasking the mask - evaluating social biases in masked language models. *National Conference on Artificial Intelligence*, pages 11954–11962.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *24th Annual Conference of the European Association for Machine Translation*, page 193.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B 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.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Generating reasonable and diversified story ending using sequence to sequence model with adversarial training. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1033–1043.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chenzhengyi Liu, Jie Huang, Kerui Zhu, and Kevin Chen-Chuan Chang. 2023a. DimonGen: Diversified generative commonsense reasoning for explaining concept relationships. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4719–4731, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Stroudsburg, PA, USA. Association for Computational Linguistics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

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

Amey P Pasarkar and Adji Bousso Dieng. 2024. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3808–3816. PMLR.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Murray Rosenblatt. 1956. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.*, 27(3):832–837.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International conference on machine learning*, pages 5719–5728. PMLR.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. Semeval-2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 307–321.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. 2024a. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *arXiv preprint arXiv:2402.01349*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of EMNLP Workshop*, page 1.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao, and Meng Jiang. 2022. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 1–11.

Tianhui Zhang, Bei Peng, and Danushka Bollegala. 2024. Improving diversity of commonsense generation by large language models via in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9226–9242, Miami, Florida, USA. Association for Computational Linguistics.

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

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

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*, pages 1097–1100.

Supplementary Materials

A Vendi Score (VS)

The VS is a similarity-based diversity metric, inspired by the ecological diversity, which is defined as the exponential of the entropy of the distribution of the species under study. Specifically, VS calculates the exponential of the Shannon entropy of the eigenvalues of a similarity matrix (Friedman and Dieng, 2023). Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be the kernel matrix with entries $K_{i,j} = k(x_i, x_j)$. In our experiments, $k(x_i, x_j)$ is computed as the dot product of the n-gram (for form-based diversity evaluations) or pre-trained embedding (for content-based diversity evaluations) of each sentence pair in a set of sentences. Let us denote the eigenvalues of $\mathbf{K}$ by

$\lambda_1, \lambda_2, \ldots, \lambda_n$. Then, VS is given by (1).

$$VS = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right) \tag{1}$$

The VS could be interpreted as the effective number of dissimilar elements in a sample. This formulation corresponds to a special case where the order $q = 1$. However, it has the limitation that it could not handle imbalanced datasets where rare elements might be under-represented. To address these challenges, the VS has been generalised to include different orders q (Pasarkar and Dieng, 2024) as given by (2).

$$VS_q = \exp\left(\frac{1}{1-q} \log \sum_{i=1}^{n} \lambda_i^q\right) \tag{2}$$

Here, q allows users to control the sensitivity to rare (or common) elements, where $q < 1$ corresponds to high sensitivity to rare elements. The special case of $q = \inf$ forces VS to capture the most dominant elements, making it highly sensitive to redundant elements.

B Chamfer Distance

Chamfer Distance (CD) is a geometric metric commonly used to compute the dissimilarity between two sets of points with embeddings. Given two sets of sentence embeddings $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ and $\mathcal{B} = \{b_1, b_2, \ldots, b_n\}$, CD is defined in (3).

$$\begin{aligned} \mathrm{CD}(A, B) = &\frac{1}{|A|} \sum_{a \in A} \min_{b \in B} \|a - b\|_2^2 \\ &+ \frac{1}{|B|} \sum_{b \in B} \min_{a \in A} \|b - a\|_2^2, \end{aligned} \tag{3}$$

This metric captures how well each sentence embedding in one set is approximated by the closest embedding in the other set.

C Generating Candidate Sets

In this section, we describe further details regarding the high and low quality candidate set generation process. We use three generator LLMs for this purpose: `GPT-4-turbo` (Achiam et al., 2023), `Llama3.1-8b` (Dubey et al., 2024), and `Qwen 2.5-14b` (Hui et al., 2024).

To generate the **Default** set of sentences for each set of input concepts in the CommonGen test cases, we instruct each generator LLM separately with the prompt shown in Figure 5. To generate the paraphrase of a given sentence for the **Para-1**, **Para-2** and **Para-3** sets, we instruct the generator LLMs with the prompt shown in Figure 6. The instruction to generate nonsensical sentences is shown in Figure 7.

An example of the candidate sets of high and low quality sentences generated by the different high and low quality operations are shown in Table 5.

Instruction
Given a set of specific words, write four short and simple sentences that contains all the required words. The sentence should describe a common scene in daily life, and the concepts should be used in a natural way.

Example:
"Concepts":{concept_set},
"Sentences": {sentence_set}

Input:
"Concepts":{concept_set},

Figure 5: The prompt used to instruct generator LLMs to produce the **Default** set of sentences.

Instruction
For each provided sentence, paraphrase it, ensuring that the original meaning is preserved and that all required keywords are included in the paraphrase. You could apply following methods to paraphrase.

- **Passive Voice:** Convert sentences from active to passive voice, focusing on the recipient of the action.
- **Change of Tense:** Adjust the verb tense within the sentence. This could involve changing from present to past, past to future, or any other tense modifications appropriate to the context.
- **Synonym Replacement:** Replace words in the sentence with their synonyms except the provided keywords. Care must be taken to ensure that the synonyms fit naturally within the context of the sentence and maintain the original meaning.

Examples
"Concepts": {concepts set}
"Original_sentences": {original sentences}
"Paraphrases"{paraphrased_sentences}

Input:
"Concepts": {concepts set}
"Original_sentences": {original sentences}

Figure 6: The prompt used to instruct generator LLMs to produce the **Paraphrased** set of sentences.

Instruction
Given a set of specific concepts, write four sentences that are nonsensical and conflict with commonsense in daily life. Each sentence mush contain all the required words.

Example:
"Concepts": {concept_set}
"Sentences": {nonsensical_sentences}

Input:
"Concepts": {concept_set}

Figure 7: The prompt used to instruct generator LLMs to produce the **Nonsensical** set of sentences.

C.1 Quality Evaluation

Before evaluating the diversity of each candidate set, we first assess their quality using the official metrics proposed by Lin et al. (2020) based on `GPT-4o` for evaluating LLM generations. Based on these metrics, we classify the candidate sets into

Input concepts	{dog, throw, frisbee, catch}
	High-quality candidate sets
Default	"The dog catches the frisbee after I throw it." "I throw the frisbee and the dog tries to catch it." "Every time I throw the frisbee, the dog catches it eagerly." "Watching the dog catch the frisbee after I throw it brings joy."
Para-1	"After the frisbee is thrown by me, it is caught by the dog." "The dog catches the frisbee after I throw it." "I throw the frisbee and the dog tries to catch it." "Every time I throw the frisbee, the dog catches it eagerly."
Para-2	"I throw the frisbee and the dog tries to catch it." "The frisbee is thrown by me, and the dog attempts to catch it." "Every time I throw the frisbee, the dog catches it eagerly." "The frisbee is eagerly caught by the dog each time it is thrown by me."
Para-3	"Every time I throw the frisbee, the dog catches it eagerly." "The frisbee is eagerly caught by the dog each time it is thrown by me." "Eagerly, the frisbee is caught by the dog every time it is thrown by me." "Watching the dog catch the frisbee after I throw it brings joy."
	Low-quality candidate sets
Nonsensical	"The frisbee decided to throw the dog sky-high so it could catch the stars." "At midnight, every dog gathers to catch the moon while humans throw frisbees into dreams." "The dog threw a party where frisbees were invited to catch sounds and dances." "During breakfast, the dog learned to throw sunlight so that the frisbee could catch and store it."
NounShuff	"The frisbee catches the it after I throw dog." "dog throw the it and the frisbee tries to catch I." "Every I frisbee throw the time, the it catches dog eagerly." "Watching the dog catch the I after frisbee throw it brings joy."
RndShuff	"catches I dog frisbee it throw the the after." "I to catch throw dog tries and the it frisbee the." "the dog I the it frisbee, throw eagerly every time catches." "the frisbee the catch throw joy I after Watching brings dog it."

Table 5: An example of candidate sets generated by the different high and low quality operations for an input concept set selected from the CommonGen test dataset.

high-quality and low-quality groups using the overall quality score. The quality metrics are defined as follows:

Length: the number of words on average in the generated sentences.

Coverage: the percentage of examples where all given concepts are covered by LLM outputs.

Win_Tie: the percentage of examples where GPT-4o prefers the model outputs over the human-written references (or there is a tie).

Overall Score: the product of scores on Coverage, and Win_Tie Rate.

From Table 6, Table 7 and Table 8, we see that the **Default** generation achieves the best quality among the candidate sets and the outputs generated by `GPT-4-turbo` has the best quality among the three models. Therefore, we use `GPT-4-turbo` to show the result in the main paper. `GPT-4-turbo` also has higher win_tie rate compared with human preference. However, as the number of paraphrases increases (e.g. in Para-2 and Para-3), the Win_Tie decreases. This suggests that the CommonGen evaluator implicitly considers diversity as part of its quality evaluation, even though diversity is not explicitly mentioned in the evaluation instructions. Additionally, the coverage rate declines as the number of paraphrases increases. This highlights that generating diverse outputs while maintaining high coverage remains a challenge for LLMs, even for state-of-the-art models like `GPT-4-turbo`.

D LLM-based Diversity Evaluation

We use GPT-4o as the **Diversity Annotator LLM** for evaluating the diversity in a given set of sentences. Prior work using LLMs for rating NLG tasks have shown that GPT-4o to demonstrate stronger correlations with human ratings (Liu et al., 2023b; Bai et al., 2024). Moreover, as described in § D.1, in a human evaluation over a subset of our dataset showed that GPT-4o to have a high agreement with human diversity preferences.

The prompt that we use to obtain diversity ratings from GPT-4o is shown in Figure 8. This prompt instructs GPT-4o to adhere to commonsense constraints (i.e. nonsensical outputs should not be interpreted to be genuinely "diverse"). We instruct GPT-4o to score each set's diversity according to a five-point scale, from *highly redundant* (1) to *low redundancy across a wide range of aspects* (5). We also require that GPT-4o consider thematic coherence among the sentences in a given set, when evaluating for their diversity such that they all cover the same set of input concepts.

Few-shot Prompting: In-context Learning (ICL) has proven to be an effective strategy for improving text generation and evaluation in many NLG tasks (Brown et al., 2020; Dong et al., 2022). Consequently, to guide GPT-4o towards human-like diversity judgments, we create a set of human-labelled examples illustrating how diverse (or non-diverse) outputs should be rated. Specifically, we asked three linguistically trained annotators to independently evaluate the diversity of 70 sentence sets. Each set comprises of four sentences generated by the same model from the same input concepts. The human annotators followed the same diversity criteria, already described in § 4.1. Specifically, each annotator is instructed to:

1. Assign a 1–5 diversity rating to each sentence set.
2. Rank the sets (if they shared the same input) with their diversity preference. This ranking resolves ties when two sets receive the same numerical score.

Finally, we select the top 8 sentence set pairs with the highest agreement among the human annotators as the few-shot examples to be included in our prompt to GPT-4o.

For each pair of candidate sentence sets, we query GPT-4o five times, each time randomly shuffling the ordering of the sentences presented to GPT-4o as well as the pairwise ordering of the two sets to mitigate any biases resulting from position of the candidates within the prompt. We average the five predicted diversity ratings per sentence set and determine the set with the higher mean rating as the more diverse one.

Method	Length	Coverage	Win_Tie	Overall
Default	12.9	86.5	58.7	50.8
Para-1	12.9	83.1	56.5	47.0
Para-2	13.8	75.6	43.6	32.9
Para-3	14.4	75.1	38.5	28.9
Nonsensical	15.1	95.4	1.3	1.3
NounShuff	12.9	85.2	4.9	4.2
RndShuff	12.9	79.6	0.3	0.2

Table 6: Comparison of length, coverage, win-tie percentage, and overall performance across different methods for the GPT-4-turbo's candidate sets generation.

D.1 Human Verification

To assess the reliability of our LLM-based diversity annotation, we randomly selected 70 pairs of high-quality sentence sets and asked five linguistically trained annotators (graduate students and academics trained in linguistic annotation tasks) to choose their preferred set based on the same diversity criteria used by the LLM. We then compared the human annotations with the LLM's preferences.

To measure the agreement, we calculated the pairwise accuracy between each human annotator's judgments and the LLM annotator's decisions for all pairs. The average pairwise accuracy across all annotators was then computed to represent the overall agreement. The resulting agreement of 79.4% demonstrates that our LLM-based annotations provide an accurate and reliable alternative to human diversity judgments.

Tevet and Berant (2021) highlighted that evaluating text diversity is challenging for crowdsourced human annotators, as judgments can be influenced by individual biases or lack of linguistic training. Consistent with this observation, we calculated Fleiss' Kappa to measure agreement among the five human annotators. The resulting Kappa value of 0.45 indicates a moderate level of agreement among the human annotators, suggesting the difficulty of the task.

Task Description:
You are presented with two sets of sentences, **Set 1** and **Set 2**. Each set contains sentences around a common theme. Your task is to evaluate each set based on their adherence to commonsense (quality) and their diversity, focusing particularly on redundancy within the sets. Subtle differences in reasoning or approach should also be recognized. The sentence sets should be cohesive around the same theme, and diversity should be considered in terms of exploring different aspects of that theme.

Important Notes:
It is crucial to pay close attention to which sentences are in Set 1 and which are in Set 2 when making your evaluations. Do not assume any set is superior by default in quality or diversity.

Evaluate each set independently based on its own content.

Diversity Evaluation Criteria:
1. Low Redundancy: Sentences should exhibit low lexical and semantic similarity.
2. Degree of Redundancy: Sets with more paraphrased sentences or repetitive themes have lower diversity.
3. Comprehensive Diversity: The sentences in the sets should enrich the theme without compromising realism and common sense.

Diversity Scoring Guidelines (for each set):
5 Points: Sentences explore a wide range of aspects of the theme with low redundancy.
4 Points: Sentences cover different aspects of the theme with minimal redundancy.
3 Points: Sentences have some diversity but noticeable redundancy.
2 Points: Sentences are mostly repetitive with limited exploration of the theme.
1 Point: Sentences are highly redundant with almost no diversity.

Output:
Based on the above criteria, assign a separate score for quality and diversity to each set, ranging from 1 to 5 points.

Examples:
"Set 1": {Sentence Set 1} "Set 2": {Sentence Set 2}
"Diversity_Score_Set1": {score}, "Diversity_Score_Set2": {score}

Figure 8: Instructions provided to GPT-4o for scoring and comparing two sentence sets. The instructions specify a five-point diversity scale ranging from *highly redundant* (score = 1) to *wide range of aspects with minimal redundancy* (score = 5). We also emphasise commonsense consistency and thematic relevance in the instruction. The prompt concludes with a request for a concise output format containing the final scores.

Method	Length	Coverage	Win_Tie	Overall
Default	13.8	62.0	44.0	27.3
Para-1	14.3	55.9	38.5	21.5
Para-2	15.0	46.2	30.4	14.0
Para-3	15.6	46.9	27.1	12.7
Nonsensical	18.4	73.8	1.4	1.0
NounShuff	13.8	60.9	3.7	2.2
RndShuff	13.8	55.7	0.1	0.1

Table 7: Comparison of length, coverage, win_tie percentage, and overall performance across different methods for Qwen2.5-14B's candidate sets generation.

Method	Length	Coverage	Win_Tie	Overall
Default	15.3	60.7	30.1	18.3
Para-1	15.4	57.8	26.4	15.2
Para-2	15.9	57.8	23.9	13.8
Para-3	17.5	55.7	21.4	11.9
Nonsensical	17.1	78.6	2.7	18.1
NounShuff	15.3	59.5	2.9	1.8
RndShuff	15.3	55.6	0.1	0.1

Table 8: Comparison of length, coverage, win_tie percentage, and overall performance across different methods for the Llama3.1-8B model's candidate sets generation.

E Confidence Intervals

To measure statistical significance for the accuracy scores reported by the different diversity evaluation metrics on the CommonGen dataset, we compute the 95% binomial confidence intervals using the Clopper-Pearson test (Clopper and Pearson, 1934) as shown in Figure 9 for all test cases. Additionally, Figure 11 and Figure 10 present confidence intervals for the high-quality and low-quality candidate subsets, respectively. The bars in blue represent form-based metrics, while the green bars correspond to content-based metrics. Across all figures, content-based metrics such as VS-Embed-0.5 and Chamfer consistently exhibit higher accuracies with narrower confidence intervals, highlighting their robustness. In contrast, form-based met-

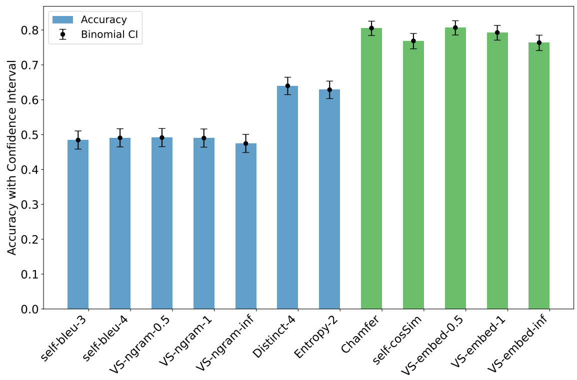


Figure 9: Binomial confidence intervals are superimposed for the accuracies reported by the diversity metrics on the all candidate sentence sets on the CommonGen test dataset

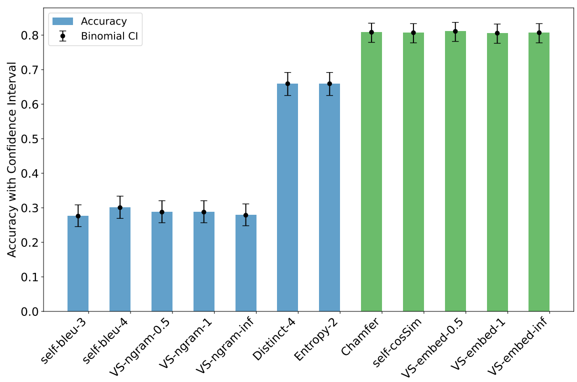


Figure 10: Binomial confidence intervals are superimposed for the accuracies reported by the diversity metrics on the low generation quality candidate sentence sets on the CommonGen test dataset.

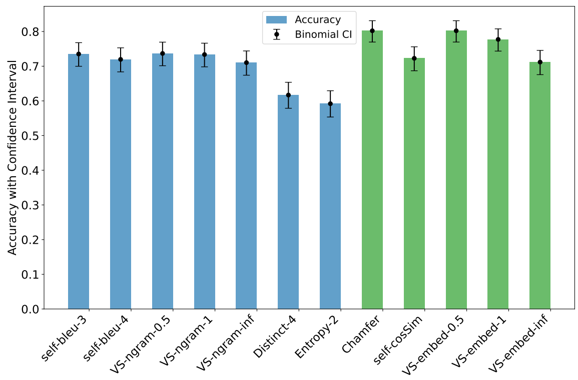


Figure 11: Binomial confidence intervals are superimposed for the accuracies reported by the diversity metrics on the high generation quality candidate sentence sets on the CommonGen test dataset

	Diversity Metric	**Accuracy**
Form	self-BLEU-3	21.7
	self-BLEU-4	20.5
	VS-ngram-0.5	34.7
	VS-ngram-1	34.7
	VS-ngram-inf	35.2
	Distinct-4	29.3
	Entropy-2	24.0
Content	Chamfer	**38.8**
	self-cosine	38.4
	VS-Embed-0.5	38.5
	VS-Embed-1	38.5
	VS-Embed-inf	38.4

Table 9: Accuracy of diversity metrics using low-quality sentence sets generated from the ComVE dataset. We see that form-based metrics perform worse compared to the content-based metrics.

rics such as self-BLEU show lower accuracies and wider intervals, especially in low-quality scenarios. These results emphasise the reliability of content-based metrics for evaluating meaningful diversity in GCR tasks.

F Further Experiments on ComVE

To explore the performance of diversity metrics for low quality sentences, we generated low-quality sentence sets on the ComVE dataset, including Nonsensical, NounShuff and RndShuff sentence sets based on the highest-quality generated set by `Qwen2.5-14B` generator LLM. We also use `GPT-4o` as the annotator LLM, and prompt it to provide pairwise diversity judgements to a given pair of sentence sets, resulting in 1,936 test cases. The accuracy of each diversity metric is shown in Table 9. We see a clear performance gap between form-based and content-based metrics in this setting as well. While content-based metrics achieve the highest accuracy, form-based metrics, such as self-BLEU, consistently underperform. This experiment further shows the limitations of form-based diversity metrics in capturing meaningful diversity.