Improving Diversity of Commonsense Generation by Large Language Models via In-Context Learning

Anonymous ACL submission

Abstract

Generative Commonsense Reasoning (GCR) requires a model to reason about a situation using commonsense knowledge, while generating coherent sentences. Although the quality of the generated sentences is crucial, the diversity of the generation is equally important because it reflects the model's ability to use a range of commonsense knowledge facts. Large Language Models (LLMs) have shown proficiency in enhancing the generation quality across various tasks through in-context learning (ICL) using given examples without the need for any fine-tuning. However, the diversity aspect in LLM outputs has not been systematically studied before. To address this, we propose a simple method that diversifies the LLM generations, while preserving their quality. Experimental results on three benchmark GCR datasets show that our method achieves an ideal balance between the quality and diversity. Moreover, the sentences generated by our proposed method can be used as training data to improve diversity in existing commonsense generators.

1 Introduction

Commonsense reasoning is the ability to make logical deductions about concepts encountered in daily life, and is considered as a critical property of intelligent agents (Davis and Marcus, 2015). Concepts are mental representations of classes and are expressed using words in a language (Liu et al., 2023). Given the inputs, the GCR task requires a model to generate a *high quality* sentence that is grammatical and adheres to commonsense, evaluated by its similarity to a set of human-written reference sentences covering the same set of concepts (Lin et al., 2020).

Often there exists multiple relationships between a given set of concepts, leading to alternative reasoning paths that take *diverse* view points. For example, given the four concepts *dog*, *frisbee*, *throw* and *catch*, different sentences can be generated as shown in Figure 1. Although all sentences shown in Figure 1 are grammatical, the bottom set expresses diverse view points (e.g. from the dog's as well as the man's) compared to the set at the top. Apart from the generation quality, diversity is also an important factor in text generation because the low-diversity texts tend to be dull, repetitive or biased towards a particular view point (Tevet and Berant, 2021). Diversity is an important consideration in many Natural Language Generation (NLG) applications, such as story generation (Li et al., 2018), paraphrase generation (Gupta et al., 2018), and GCR (Yu et al., 2022; Liu et al., 2023). Existing methods promote diversity through special decoding strategies, such as nucleus sampling (Holtzman et al., 2019), or encoding interventions such as random noise injection (Gupta et al., 2018) or Mixture of Experts (MoE) approaches (Shen et al., 2019).

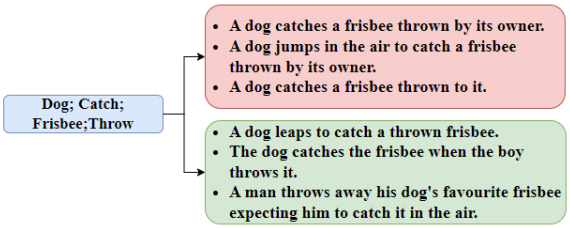


Figure 1: Example sentences generated that cover a given set of concepts from the CommonGen (Lin et al., 2020) dataset. The model is required to generate commonsense bearing and coherent sentences containing all of the input concepts. The responses shown at the bottom (in green) are considered by human annotators to be more diverse than those at the top (in red).

We propose In-Context Diversification (ICD), a computationally-efficient and accurate method to improve the diversity in GCR, where the sentences are generated from a pre-trained LLM, and strikes a fine-balance between the output diversity and quality. ICD uses an ICL approach to increase the diversity of the sentences generated by an LLM, while maintaining the quality of the generation. ICD is a two-step process where it first lets an

LLM to freely generate high-quality sentences that are grammatical, commonsense bearing and cover all the given input concepts. Next, ICD uses a user-specified diversity metric to evaluate the diversity of the generated sentences. If the diversity is low, ICD provides feedback to the LLM, instructing it to generate more diverse sentences considering the already generated sentences.

Given that ICD is using LLMs to generate diverse sentences via ICL and without updating the parameters of the LLMs, an interesting and open question is *whether an LLM can accurately judge the diversity of a given set of sentences, covering a common set of concepts*. To answer this question, we conduct an experiment where we instruct `GPT3.5-turbo` to judge the diversity of the set of input sentences according to a five-scale grading system, and convert the predicted grades into binary judgements (i.e. diverse vs. non-diverse). We compare the LLM-assigned grades against those by a group of human annotators, and find a moderate-level (Cohen's Kappa of 0.409) agreement between human vs. LLM judgements, demonstrating that LLMs can indeed be instructed to obtain diversity judgements for GCR tasks.

We evaluate ICD on three GCR tasks/datasets: CommonGen (Lin et al., 2020), ComVE (Wang et al., 2020), and DimonGen (Liu et al., 2023). We find that our proposed ICD balances diversity and quality appropriately, improving their harmonic mean by at least 6% over that of a default baseline. Moreover, the sentences generated by ICD can be used as training data to improve diversity in a Seq2Seq model (Sutskever et al., 2014; Lewis et al., 2020), producing results that are comparable to the models that are trained on knowledge graphs or human-written text corpora (Liu et al., 2021; Fan et al., 2020; Li et al., 2021). *An anonymised version of the source code is submitted to ARR and will be made public upon paper acceptance.*

2 Related Work

Diverse Text Generation. A variety of methods have been proposed to enhance the diversity of NLG. Sampling-based decoding is an effective method to increase the generation diversity. Holtzman et al. (2019) proposed nucleus sampling to generate diverse content at the generation stage. Truncated sampling (Fan et al., 2018) prunes and then samples the tokens based on the probability distribution. Furthermore, Shen et al. (2019) proposed an MoE approach to diversify translation outputs. Moreover, incorporating external corpora in the MoE further promotes diversity, such as by using a knowledge graph (Yu et al., 2022; Hwang et al., 2023) or by a collection of retrieved sentences (Liu et al., 2023). Although LLMs have reported superior performance in numerous Natural Language Processing (NLP) tasks (Touvron et al., 2023; OpenAI, 2023b,a), to the best of our knowledge, diversifying their generations in commonsense reasoning with ICL has not been explored in prior work on GCR.

In-Context Learning. Recent studies demonstrate that LLMs can exhibit robust few-shot performance on a variety of downstream tasks through ICL (Brown et al., 2020). ICL is a technique for instructing an LLM using one or more examples for a particular text generation task. The generated text is conditioned on both the input as well as the instruction prompt. Wang et al. (2023) show that in ICL, label words in the demonstration examples function as anchors, which aggregate semantic information to their word representations in the shallow (closer to the input) layers, while providing that information to the final predictions performed by the deeper (closer to the output) layers. In contrast to fine-tuning-based methods, ICL is computationally lightweight because it does not update the parameters of the LLM. Therefore, ICL is an attractive method when integrating task-specific knowledge to an LLM by simply changing the prompt and the few-shot examples (Dong et al., 2022).

3 In-context Diversification

We consider the problem of generating a set of diverse sentences that express commonsense reasoning, either by covering a set of given concepts (in CommonGen and DimonGen) or by providing an explanation for a given counterfactual statement (in ComVE). Formally, given a sequence (a set of concepts or a statement) $\mathcal{X} = \{x_1, \ldots, x_m\}$, the goal of GCR is to generate a set of grammatically correct and commonsense bearing sentences $\mathcal{Y} = \{y_1, \ldots, y_n\}$, where y_i is the i-th output generated by the model with probability $p(y_i|\mathcal{X})$. Moreover, we require that the generated sentences $\{y_1, \ldots, y_n\}$ to be lexically as well as semantically diverse.

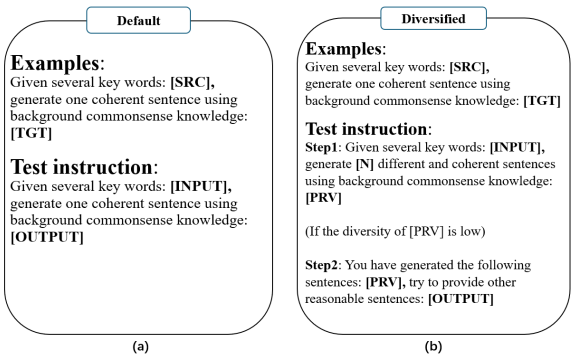


Figure 2: An example of default and diversified prompts is shown for an instance selected from the CommonGen dataset. Here, the default prompt shown in Figure 2a is taken from Li et al. (2023). Few-shot examples are included in each prompt where **[SRC]** denotes the set of input concepts and **[TGT]** the corresponding sentences in CommonGen. For a given set of **[INPUT]** concepts, the LLM is then required to generate sentences at the slot **[OUTPUT]**. As shown in Figure 2b, ICD uses the diversified prompt, which operates in two steps. Step 1 generates a set of **[N]** sentences, **[PRV]**. We check for the diversity among the sentences in **[PRV]**, and if it is low, we use the prompt in Step 2 to generate the final set of sentences.

3.1 Sentence Generation

To explain our proposed ICD, let us consider GCR on CommonGen, where we must generate a set of sentences $\mathcal{Y}$, such that each sentence contains *all* of the input concepts $\mathcal{X}$ as shown in Figure 2a. Given an LLM, we can design a prompt that contains a task-specific instruction and one or more examples containing the input concepts (denoted by **[SRC]** in Figure 2) and the corresponding human-written sentences containing **all** given input concepts (denoted by **[TGT]**) to instruct the LLM to generate output sentences $\mathcal{Y}$ (denoted by **[OUTPUT]**) for a given set of input concepts $\mathcal{X}$ (denoted by **[INPUT]**). We refer to a prompt of this nature as a default prompt, and the corresponding set of generated sentences by $\mathcal{S}_{\text{def}}$.

Note that the default prompt does not necessarily guarantee that the generated set of sentences will be diverse and an LLM could return sentences that are highly similar to each other. To address this issue, we propose a diversified prompt as shown in Figure 2b. Specifically, the diversified prompt operates in two steps. In Step 1, we require that the LLM generate N number of sentences that are *different*, in addition to being coherent and commonsense bearing. Next, we use a suitable diversity metric to evaluate the level of diversity among the generated set of sentences. If the diversity of the generated sentences is low, in Step 2, we show those sentences to the LLM and instruct it to generate sentences that are *different* to those. As the criteria for triggering Step 2, we check whether the exact same sentence has been generated multiple times by the LLM during Step 1. The final set of generated sentences is denoted by $\mathcal{S}_{\text{div}}$.

Algorithm 1 In-Context Diversification (ICD)

Input: Generated sets of sentences $\mathcal{S}_{\text{def}}$ and $\mathcal{S}_{\text{div}}$, respectively from default and diversified prompts, the number of desired output sentences N, and a diversity metric f.
Output: Output set of sentences $\mathcal{S}^*$

$\mathcal{S}^* \leftarrow \emptyset$
$\alpha \leftarrow 0$
for $\mathcal{S} \in (\mathcal{S}^-_{\text{def}} \cup \mathcal{S}_{\text{div}})$ **do**
  if $(|\mathcal{S}| == N) \wedge (f(\mathcal{S}) \geq \alpha)$ **then**
    $\alpha \leftarrow f(\mathcal{S})$
    $\mathcal{S}^* \leftarrow \mathcal{S}$
  end if
end for
return $\mathcal{S}^*$

3.2 Diversity-based Sampling

Because of the limited availability of human-written reference sentences for evaluating GCR models, there exists a trade-off between quality vs. diversity when generating sentences for GCR tasks.[1] Simply maximising for diversity often leads to generations that do not cover the input concepts in a natural way. For example, a randomly selected set of sentences would be highly diverse, yet unlikely to capture the input concept sets. On the other hand, if we force an LLM to generate sentences that contain all of the input concepts, it might find difficult to generate semantically diverse sentences and resort to trivial lexical or syntactic diversity tricks such as morphological inflections or word-order permutations.

To address this issue, we propose a diversity-based sampling method shown in Algorithm 1. Consider that the default prompt provides a set $\mathcal{S}_{\text{def}}$ of sentences that have not been optimised for diversity (likely to have a higher quality), while on the other hand the diversified prompt provides a set $\mathcal{S}_{\text{div}}$ of sentences that are further refined for diversity (likely to have a higher diversity). We wish to find a set of sentences that simultaneously satisfies the following criteria: (a) must contain exactly N number of sentences, as specified by the user, and (b) must have a high diversity score, measured using a user-specified diversity metric $f(\in \mathbb{R}_{\geq 0})$. We formalise this as a subset search

[1] This trade-off is further empirically verified in § 5.1.

problem, where we compute the union $\mathcal{S}_{\text{def}} \cup \mathcal{S}_{\text{div}}$ and search for the subset $\mathcal{S}^*$ that jointly satisfies those criteria following the procedure detailed in Algorithm 1. Although the total number of subsets of size N is $\binom{|\mathcal{S}_{\text{def}} \cup \mathcal{S}_{\text{div}}|}{N}$, it is sufficiently small for the values of $N (N \leq 6)$ in our GCR tasks, which makes this subset search fast in practice.

4 Experimental Settings

4.1 Tasks and Datasets

We evaluate ICD on three GCR tasks as follows.

Constrained Commonsense Reasoning: In CommonGen (Lin et al., 2020) benchmark, a model is required to generate a sentence covering a given set of concepts such that background commonsense knowledge associated with the input concepts is reflected. This dataset contains 3.5K distinct concept sets (train = 32651, dev = 993, and test = 1497) with corresponding human written sentences (train = 67389, dev = 4018, and test = 6042). Each instance contains on average 3-5 input concepts.

Commonsense Explanation Reasoning: ComVE (Wang et al., 2020) is part of the SemEval 2020 commonsense validation task, where for a given counterfactual statement, a model is required to generate an explanation providing a reason describing why the statement is nonsensical. This dataset contains 10K (train = 8532, dev = 476, and test = 992) examples, where each example contains three reference outputs.

Diversified GCR: DimonGen (Liu et al., 2023) involves generating diverse sentences that describe the relationships between two given concepts. It is a challenging task because it requires generating reasonable scenarios for a given pair of concepts without any context. This dataset contains 17109 instances (train = 15263, dev = 665, test = 1181), where each instance has 3-5 references.

4.2 Evaluation Metrics

We measure both the quality and diversity of the sentences generated by models using the metrics described next.

4.2.1 Quality Metrics

We compare a generated sentence by a model against a set of human-written references to evaluate the quality of the generation using several metrics: BLEU (Papineni et al., 2002) measures n-gram precision against human reference texts, SPICE (Anderson et al., 2016) measures the semantic propositional overlap between two sentences, and BERTScore (Zhang et al., 2020) uses contextualised word embeddings to measure the semantic similarity between tokens in two sentences. In alignment with prior works (Yu et al., 2022; Liu et al., 2023; Hwang et al., 2023), when multiple candidate sentences are generated for a test case, we select the highest-scoring candidate for evaluating quality.

4.2.2 Diversity Metrics

Pairwise Diversity: We use self-BLEU (Zhu et al., 2018) to measure n-gram overlap among sentences within each generated set. The metric computes the average sentence-level similarity between all pairwise combinations of the generations in the generation set. Note that unlike BLEU, self-BLEU does *not* require human generated references for measuring diversity. We use self-BLEU3/4 (corresponding to $n = 3$ and 4) in our experiment. Lower self-BLEU scores indicate higher lexical diversity.

Corpus Diversity: To measure the variety within our generated text corpus, we employ Distinct-k (Li et al., 2016), which calculates the ratio of unique k-grams to the total number of k-grams. This metric is particularly useful for adjusting the bias of LLMs toward generating longer sequences, ensuring that diversity is not artificially inflated by the sentence length. Additionally, we use Entropy-k to evaluate the distributional uniformity of k-gram occurrences, considering word frequencies for a more nuanced view of diversity. Higher Distinct-k and Entropy-k scores indicate higher diversity.

Semantic Diversity: All previously described diversity metrics are limited to evaluating lexical diversity. To measure diversity at a semantic level, we propose self-cosSim, which is the average pairwise cosine similarity between generated sentences, computed using sentence embeddings obtained from SimCSE (Gao et al., 2021). Likewise, we define the self-BERTScore as a diversity metric that averages the BERTScores for all generated sentence pairs. Lower self-cosSim and self-BERTScore values indicate higher semantic diversity.

4.2.3 Combined Metrics

We would prefer GCR models that have both high quality and high diversity. To incoporate both as-

pects into a single metric, we compute the **Harmonic Mean** between (a) the self-BLEU-4 as the diversity metric, and (b) BERTScore as the quality metric. As discussed in § 3.2, there exists a trade-off between quality and diversity in GCR. Therefore, the harmonic mean is suitable when averaging quality and diversity scores.[2]

Alihosseini et al. (2019) proposed Fréchet BERT Distance (FBD) as a joint metric for simultaneously measuring both the quality and diversity of NLG. FBD is inspired by the Fréchet Inception Distance (FID), proposed by Heusel et al. (2017), for measuring the quality of image generation. Specifically, FBD computes the pooler output[3] of a sentence as its embedding (Devlin et al., 2019) and represents a set of sentences using the mean vector and the covariance matrix computed from their sentence embeddings. Next, Wasserstein-2 distance is computed between the set of reference sentences and the set of generated sentences, which captures both the distance between the means as well as variance in the distributions. Lower FBD scores indicate high combined performance.

4.3 Implementation Details

We use `GPT3.5-turbo` and `Vicuna-13b-v1.5`[4] as LLMs with temperature set to 1.1 in our experiments. By using two LLMs with significantly differing number of parameters and by including, Vicuna, an open source LLM, we plan to improve the reliability and reproducibility of our results. Max response length is set to 25 tokens. The inference times for CommonGen, ComVE and DimonGen datasets are respectively 5-6, 2-3 and 1-2 hours. The cost of running ICD with `GPT3.5-turbo` are ca. $6, $4 and $4 respectively for CommonGen, ComVE and DimonGen datasets. On the other hand, the costs of fine-tuning on `GPT3.5-turbo` are much higher at $58.8 for CommonGen, $24.7 for ComVE and $32.0 for DimonGen. Moreover, fine-tuning with LoRA (Hu et al., 2022) with rank of 8 and alpha of 16 on Vicuna takes ca. 34 hours. We use `BART-large`[5] for MoE-based models. We use the `GPT3.5-turbo` to generate sentences for the CommonGen train/dev/test sets using the default, diversified and for ICD. For model training, we use the Adam optimiser (Kingma and Ba, 2015) with a batch size of 64, a learning rate of 3e-5 and a beam size of 5. All of the MoE-based models are trained for 20 epochs and required to generate $k = 3$ sentences. All experiments, except with GPT3.5-turbo, are conducted on a single RTX A6000 GPU.

[2]Any diversity and quality metric could be combined for computing the harmonic mean. We use self-BLEU-4 and BERTScore due to their reliability shown in preliminary evaluations.

[3]The last layer's hidden-state of the first token of the sequence is further processed by a Linear layer and a Tanh activation function.

[4]https://huggingface.co/lmsys/vicuna-13b-v1.5

[5]https://huggingface.co/facebook/bart-large

5 Results and Discussion

5.1 Commonsense Generation

We compare the commonsense generations made by ICD against those using the default and diversified prompts. For this purpose, we use `GPT3.5-turbo` as the LLM and use the same 10 few-shot examples in all prompts for ICL. Further templates of the default and diversified prompts used for each task are given in Appendix B. To assess the impact of ICL, we compare against **fine-tune** method, wherein `GPT3.5-turbo` is fine-tuned on the entire training set in each dataset. Specifically, we use multiple human-written sentences, available in the training data for the three datasets to separately fine-tune the models for each task. It is noteworthy that the **fine-tune** method uses a substantially larger dataset for training (e.g., 67,389 sentences from CommonGen) compared to the 10 examples used by the ICL-based approaches. We use self-BLEU-3 as the diversity metric f in Algorithm 1 for ICD in this evaluation. The outcomes, presented in Table 1, highlight the diversity and quality metrics of these methods across the CommonGen, ConVE, and DimonGen datasets. Additionally, a **human** baseline is introduced to evaluate the diversity of sentences written by humans, where we pair-wise compare the human-written sentences for each input in the instances in the benchmark datasets using diversity metrics. Note that however, the **human** baseline must not be considered as an upper-bound for diversity because there are only a smaller number of human-written sentences per instance in the benchmark datasets.

From Table 1, we see that **fine-tune** generates sentences that have high semantic and corpus diversity, and outperforms the **human** baseline. However, recall that **fine-tune** requires a much larger training set and is computationally costly compared to all ICL-based methods. Moreover, we see that ICD can strike a good balance between quality and diversity in the sentences generated. Among

	Semantic Diversity ⇓		Corpus Diversity ⇑		Pairwise Diversity ⇓		Quality ⇑				Combined	
	self-cosSim	self-BERTScore	Entropy-4	Distinct-4	self-BLEU-3	self-BLEU-4	BLEU-3	BLEU-4	SPICE	BERTScore	Harmonic ⇑	FBD ⇓
CommonGen												
Human	67.3	60.6	10.9	91.0	25.4	17.6	-	-	-	-	-	-
Fine-tune	*64.7*	*55.9*	***11.4***	*91.1*	26.9	17.9	41.2	32.1	30.3	64.2	72.1	51.9
default	88.6	85.1	10.6	62.8	70.0	63.9	***50.4***	***40.8***	***31.0***	***70.1***	47.7	55.6
diversified	72.9	62.3	***11.4***	86.7	34.9	26.4	37.6	28.8	27.6	61.8	67.2	54.9
ICD	**72.4**	**60.1**	***11.4***	**90.3**	***24.4***	***16.0***	41.8	32.7	28.6	64.9	***73.2***	***50.8***
ComVE												
Human	62.7	47.0	9.6	96.1	12.4	8.1	-	-	-	-	-	-
Fine-tune	*55.0*	*34.0*	9.9	*97.4*	*7.7*	4.8	*23.4*	*16.0*	31.1	49.3	*65.0*	*48.8*
default	82.2	66.0	10.0	78.5	42.1	35.9	**22.8**	**15.9**	***34.3***	***50.4***	56.4	55.2
diversified	71.2	48.5	***10.2***	91.0	18.1	12.8	20.6	13.8	31.2	48.4	62.2	53.5
ICD	**67.4**	**41.9**	***10.2***	**91.6**	**8.3**	***4.7***	21.1	14.2	32.2	49.0	**64.7**	**54.1**
DimonGen												
Human	56.8	47.0	10.1	85.6	14.7	8.7	-	-	-	-	-	-
Fine-tune	*43.4*	*33*	10.4	*98.7*	6.8	*3.4*	*17.7*	*10.7*	15.5	42	*58.5*	*51.6*
default	75.7	71.3	10	83.2	43.4	37.3	**15.9**	**9.5**	***16.4***	***44.5***	52.1	68.2
diversified	57.1	46.9	***10.5***	95.9	11.2	6.5	11.4	6.4	15.2	39.9	55.9	69.0
ICD	**56.7**	**45.7**	10.4	**96.3**	***6.5***	**3.5**	13.2	7.6	15.4	41.7	**58.2**	**68.0**

Table 1: Diversity and quality scores on CommonGen, ComVE and DimonGen with `GPT3.5-turbo` LLM. Best results on each task for each metric are shown in *italics*, while the best performing ICL results are shown in **bold**.

Method	SCS ⇓	SBS ⇓	SB-3⇓	BLEU-3⇑	SPICE⇑	HM ⇑	FBD ⇓
Fine-tune	59.6	49.9	22.8	35.8	27.6	69.9	52.4
Default	82.2	73.8	52.9	***44.6***	***29.1***	60.2	56.2
Diversified	***59.1***	53.3	23.6	32.6	24.3	68.6	53.2
ICD	59.3	***49.8***	***11.3***	34.2	25.5	***73.4***	***51.0***

Table 2: GCR on CommonGen using Vicuna-13b. ICD uses self-BLEU-3. Here, SCS: self-CosSim, SBS: self-BERTScore, SB-3: self-BLEU3, HM: Harmonic Mean. Best results for each metric are shown in *italics*, while the best performing ICL results are shown in **bold**.

Method	SCS ⇓	SBS ⇓	SB-3⇓	BLEU-3⇑	SPICE⇑	HM ⇑	FBD ⇓
self-BLEU-3	72.4	60.1	**24.4**	41.8	28.6	**73.2**	**50.8**
self-CosSim	**68.8**	62.5	34.5	**42.0**	**28.8**	68.9	**50.8**
self-BERTScore	71.3	**56.3**	29.2	40.0	28.0	70.3	51.1

Table 3: Comparing the effect of using different diversity metrics, f, in Algorithm 1 for ICD. We use `GPT3.5-turbo` as the LLM and the best results on CommonGen dataset are in **bold**. Here, SCS: self-CosSim, SBS: self-BERTScore, SB-3: self-BLEU3, HM: Harmonic Mean.

the ICL-based methods, ICD achieves the best diversity scores on all diversity metrics in all three datasets. It also exhibits higher diversity compared against the human-written references. Moreover, ICD outperforms default and diversified according to the Combined metrics. ICD also achieves a Harmonic Mean comparable to that of the **fine-tune** baseline. Although default reports the best quality scores, it has low diversity, and is consistently outperformed by diversified and ICD on diversity metrics. On the other hand, diversified generally scores lower on the quality metrics. Compared to default and diversified, ICD enhances generation diversity while maintaining a satisfactory level of quality. Note that **fine-tune** is not an ICL setting (the focus of this paper) and is included only as a baseline to demonstrate the level of performance that can be achieved by fine-tuning on a much larger dataset. Despite this ICD outperforms **fine-tune** on the Pairwise Diversity in three datasets, and Combined metrics in the CommonGen dataset.

As an open source alternative LLM to `GPT3.5-turbo`, we repeat this evaluation with `Vicuna-13b` (Zheng et al., 2023) in Table 2. The same 10 few-shot examples as used with `GPT3.5-turbo` are used in this experiment for the ICL-based methods. Table 2 reconfirms ICD's ability to balance both quality and diversity according to the Combined metrics (i.e. Harmonic Mean and FBD) on this dataset. Interestingly, we see that methods that use Vicuna-13b to be more diverse compared to those that use `GPT3.5-turbo`, while the latter showing better generation quality.

In Table 3, we use different diversity metrics as f in Algorithm 1 to study the effect on text generation of ICD. We see that self-BLUE-3 and self-CosSim perform similarly across both diversity and quality metrics. Therefore, any of those diversity metrics can be used with ICD to obtain comparable performance. We see that self-BERTScore shows a slightly lower performance, except when measured against itself (i.e. SBS), which indicates some level of over-fitting to the diversity metric being used.

	Semantic Diversity ⇓		Corpus Diversity ⇑		Pairwise Diversity ⇓		Quality ⇑				Combined	
	self-cosSim	self-BERTScore	Entropy-4	Distinct-4	self-BLEU-3	self-BLEU-4	BLEU-3	BLEU-4	SPICE	BERTScore	Harmonic Mean ⇑	FBD ⇓
KG-BART	-	-	-	-	-	-	42.1	30.9	32.7	-	-	-
EKI-BART	-	-	-	-	-	-	46.0	36.1	33.4	-	-	-
KFCNet-w/o FC	-	-	-	-	-	-	50.2	42.0	35.9	-	-	-
KFCNet	-	-	-	-	-	-	**57.3**	**51.5**	**39.1**	-	-	-
MoE	89.3	81.9	9.7	61.6	63.1	56.6	49.0	38.5	33.5	70.6	53.8	61.7
MoKGE	88.7	80.6	**9.9**	**65.2**	60.4	53.6	48.8	38.4	33.1	70.3	55.9	60.8
default+MoE	90.8	84.2	9.7	61.2	65.6	58.8	51.8	41.3	34.7	73.1	52.7	61.9
diversified+MoE	**85.3**	**79.9**	9.8	63.2	**58.3**	52.6	51.4	41.4	34.6	71.6	57.0	**54.5**
ICD+MoE	90.4	82.3	9.8	64.9	58.4	**50.5**	53.2	43.1	35.4	**73.8**	**59.3**	62.5

Table 4: Downstream evaluation of the LLM-generated sentences. Top block methods use human-generated resources for training, while the ones in the bottom block are trained on LLM-generated sentences. MoE approaches are shown in the middle block and bottom blocks. `BART-large` is used as the generator for MoE-based methods. Best results for each metric is shown in **bold**, while the best performing MoE for quality is shown in **underline**.

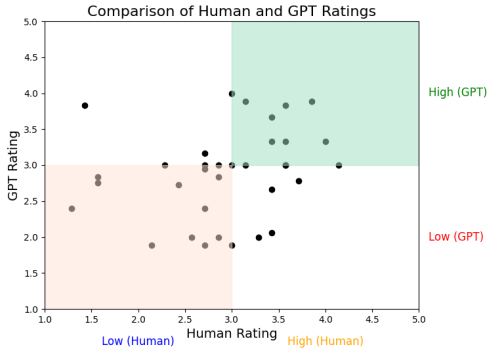


Figure 3: Human vs. GPT3.5 diversity ratings for randomly sampled sets of sentences generated by ICD. Cohen's $\kappa = 0.409$ indicates a moderate agreement.

5.2 Downstream Evaluation

The experiments presented in § 5.1 show the ability of our proposed ICD to generate diverse and commonsense bearing sentences. Therefore, an important question with practical implications is whether we can use the sentences generated by ICD as additional training data to improve both diversity and quality of previously proposed models on the GCR task, which could be seen as a downstream (extrinsic) evaluation.

For this purpose we select the MoE (Shen et al., 2019), which diversifies the generation by selecting outputs from a mixture of experts. Each expert is assigned a randomly generated sequence of tokens, which is used as a prefix for all inputs sent to that expert. For each input, an expert is selected according to the value of a latent variable, which is trained using the hard-EM algorithm. We follow Liu et al. (2023) and train three experts that retrieve sentences from the collection of sentences generated by ICD for concept sets in the CommonGen train split (210846 sentences in total). We use `BART-large` (Lewis et al., 2020) as the base model, which has shown to produce high quality commonsense generations (Zhang et al., 2023) as the generator for all experts (see Appendix A for further details). We denote this method by ICD+MoE.

As baselines for comparisons, we repeat the above process using the sentences generated by default and diversified, which we denote respectively as default+MoE and diversified+MoE in Table 4. Moreover, we compare the performance against two previously proposed MoE models: MoE (Shen et al., 2019) and MoKGE (Yu et al., 2022). MoE relies solely on the base model, whereas MoKGE requires each expert to use different sets of concepts from the ConceptNet (Speer et al., 2017) knowledge graph (KG). Because Yu et al. (2022) do not evaluate their MoKGE method on CommonGen, we ran their original implementation[6] on CommonGen and report results in Table 4.

All previously proposed GCR methods are exclusively trained using human-created data (e.g. sentences written by human and/or manually compiled KGs such as ConceptNet), whereas the methods described thus far in this section are trained on sentences generated by an LLM (`GPT3.5-turbo`). Therefore, to evaluate the feasibility of using LLM-generated sentences for training GCR models, we include the following previously proposed GCR models that are trained using a combination of corpora and KGs: KG-BART (Liu et al., 2021),EKI-BART (Fan et al., 2020) and KFCNet (Li et al., 2021). For KFCNet, we present its two results – KFCNet w/o FC, which relies only on sentences including the input concepts, without further processing, and KFCNet, which additionally ranks candidates and adds contrastive modules for the encoder and the decoder (Li et al., 2021). However, note that those methods do *not* consider diversifica-

[6] https://github.com/DM2-ND/MoKGE

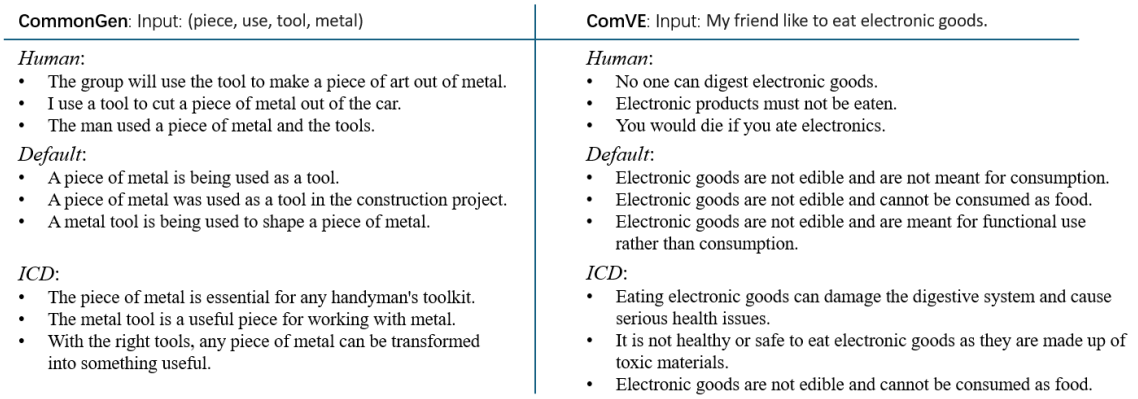


Figure 4: Sentences generated by default prompt and ICD against those by humans on CommonGen and ComVE test instances. ICD generates more diverse and high quality sentences than default.

tion, and do not report performance using diversity metrics. Therefore, we report only their published results for generation quality in Table 4.

From Table 4 we see that diversified+MoE always outperforms the original MoE in all diversity metrics, which shows that sentences generated from LLMs can be used to diversify MoE-based GCR. ICD+MoE closely matches the performance of diversified+MoE on diversity metrics, while consistently outperforming both diversified+MoE and default+MoE on quality metrics. In particular, the quality metrics reported by ICD+MoE (underlined in Table 4) are competitive against those obtained by the models that are trained on human-compiled resources (in the top block), except against KFCNet. This finding hints at potential improvement gains for GCR by using hybrid training resources that combine both human-compiled and LLM-generated data, which we highlight as an interesting future research direction.

5.3 Diversity-Awareness of LLMs

Given that we use LLMs to produce diverse generations via ICL, it remains an open and interesting question whether an LLM would agree with humans on the diversity of a given set of sentences. To answer this question, we use randomly selected 210 sentences (35 sets, each containing 6 sentences) generated by ICD (using self-BLEU-3 as the diversity metric) for the input concept sets in the CommonGen dataset. We instruct `GPT3.5-turbo` to rate the diversity of a given set of sentences according to five diversity ratings 1-5 with 1 being highly similar, while 5 being highly diverse.[7] We provide the same instruction as the annotation guidelines for eight human-annotators, who are graduate students in NLP. To reduce the subjective variability in human judgements, we average and then normalise the ratings following the Likert scale.

In Figure 3, we plot the GPT-assigned ratings against those by humans. We further split the ratings into *high* vs. *low* diversity ratings depending on whether the rating is greater or lesser than 3. The majority of the data points are distributed along the diagonal quadrants and a Cohen's Kappa of 0.409 indicating a moderate level of agreement between GPT and human ratings for diversity.

The generated sentences using the default prompt, ICD and the human references in CommonGen and ComVE datasets for a single test instance are shown in Figure 4. From Figure 4 we see that the sentences generated using the default prompt often results in significant token overlap, thereby lowering the diversity. On the other hand, ICD generates both lexically and semantically diverse sentences, covering the the diverse viewpoints in the human references.

6 Conclusion

We proposed, ICD, an ICL-based method for achieving the optimal balance between diversity and quality in text generation via LLMs. Our experiments, conducted on three GCR tasks, demonstrate that ICD significantly improves the diversity without substantially compromising the quality. Furthermore, we found that by training on the sentences generated by ICD, we can improve diversity in previously proposed GCR methods.

7 Limitations

This study primarily focuses on the generation of English sentences using pre-trained LLMs, a limi-

[7] Detailed prompt templates are shown in Appendix B.

tation shaped by the datasets we employed. Specifically, we used the ComVE (Wang et al., 2020), CommonGen (Lin et al., 2020) and DimonGen (Liu et al., 2023) datasets, which are well-regarded for evaluating diversified commonsense reasoning in English. Therefore, our evaluation of the generation quality was limited to English, which is a morphologically limited language. Future research could expand this scope to include multilingual pretrained models, thereby encompassing a broader linguistic spectrum.

Our approach is primarily geared towards optimizing the trade-off between diversity and quality in text generation. Consequently, we maintained consistent default instructions across all experiments, adopting the standard commonsense generation prompts used in Li et al. (2023) as our default instructions.

We conducted our experiments using both a closed model (i.e. `GPT3.5-turbo`) as well as an open-source one (i.e. `Vicuna-13b-v1.5`) to promote the reproducibility of our results, which are reported using multiple public available benchmarks. However, there exist many other LLMs with varying numbers of parameters and trained on different corpora. Therefore, we consider it is important to evaluate our proposed method on a broad range of LLMs to verify the generalisability of our proposed method. However, conducting such a broad analysis can be computationally costly and expensive. For example, although GPT-4 is known to have superior text generation capabilities, it incurs substantially greater costs (being 30 times more expensive than `GPT3.5-turbo` at the current pricing). Nevertheless, ICD is adaptable and could be extended to other LLMs.

8 Ethical Considerations

In this work, we did not create or release any manually annotated data. Our work is based on the publicly available datasets, CommonGen, ComVE, and DimonGen. To the best of our knowledge, no ethical issues have been reported for those datasets. Therefore, we do not foresee any data-related ethical issues arising from our work.

However, LLMs are known to generate responses that may reflect societal biases and potentially harmful content. We have not verified whether the `GPT3.5-turbo` and `Vicuna-13b` LLMs that we use in our experiments have similar problems. Therefore, it is important to test on existing benchmarks for social biases and harmful generations before the proposed method is deployed to diversify existing GCR methods used by human users.

To elicit human judgements of diversity for the sentences generated by ICD, we use annotators who are familiar working with LLMs. It is possible that their subjective (and possibly biased) viewpoints might have influenced the ratings provided. Therefore, will be important to conduct the evaluation involving a group of annotators with different backgrounds to validate the findings reported in this analysis.

References

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Association for Computational Linguistics, Minneapolis, Minnesota, pages 90–98.

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*. Springer, pages 382–398.

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.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723* .

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM* 58(9):92–103.

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

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* .

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.

Zhihao Fan, Yeyun Gong, Zhongyu Wei, Siyuan Wang, Yameng Huang, Jian Jiao, Xuan-Jing Huang, Nan Duan, and Ruofei Zhang. 2020. An enhanced knowledge injection model for commonsense generation. In *Proceedings of the 28th International Conference on Computational Linguistics*. pages 2014–2025.

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.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *Proceedings of the aaai conference on artificial intelligence*. volume 32 (1).

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., volume 30.

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

EunJeong Hwang, Veronika Thost, Vered Shwartz, and Tengfei Ma. 2023. Knowledge graph compression enhances diverse commonsense generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 558–572. https://aclanthology.org/2023.emnlp-main.37.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. pages 7871–7880.

Bei Li, Rui Wang, Junliang Guo, Kaitao Song, Xu Tan, Hany Hassan, Arul Menezes, Tong Xiao, Jiang Bian, and JingBo Zhu. 2023. Deliberate then generate: Enhanced prompting framework for text generation.

Haonan Li, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. 2021. Kfcnet: Knowledge filtering and contrastive learning for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. pages 2918–2928.

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*. Association for Computational Linguistics, Online, pages 1823–1840.

Chenzhengyi Liu, Jie Huang, Kerui Zhu, and Kevin Chen-Chuan Chang. 2023. DimonGen: Diversified generative commonsense reasoning for explaining concept relationships. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, pages 4719–4731. https://doi.org/10.18653/v1/2023.acl-long.260.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S Yu. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. volume 35, pages 6418–6425.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed: 2023-11-23.

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.

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*. PMLR, pages 5719–5728.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*. volume 31 (1).

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems* 27.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* .

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.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pages 9840–9855.

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, Danushka Bollegala, and Bei Peng. 2023. Learning to predict concept ordering for common sense generation. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Nusa Dua, Bali, pages 10–19.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685* .

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 Appendix

A Mixture of Experts

Given an input x, its corresponding LLM-generated sentences are divided into three random subsets. For each subset $G_i = \{g_1^i, ..., g_k^i\}$, alongside the input x, we concatenate their token sequences with a separate latent variable z_i, resulting in the final input x_i^f. The z_i is a randomly initialised sequence of tokens.

$$x_i^f = z_i[CLS]x[SEP]g_1^i[SEP]...g_k^i \quad (1)$$

We train the model using the hard Expectation-Maximization (EM) approach, where during the E-step, for each x_i^f and its corresponding target y_i^{target}, we identify the input that yields the highest probability as the best training example, with θ representing the generator model parameters:

The model is trained using hard-EM by assigning full responsibility to the expert with the largest joint probability. In the E-step, for each input x_i^{fin} and target y_i^{tgt}, choose the best input with the highest probability, to construct the training examples, where θ is the model's parameters.

$$y_i^{tgt} = \underset{y_j}{\operatorname{argmax}}\, p(y_j|x_i^f;\theta) \quad (2)$$

Subsequently, in the M-step, we use these selected training examples to fine-tune the generator models. During inference, we input all diversified, context-aware inputs into the generator model to yield a range of diverse outputs.

B LLM Prompt Templates

Figure 5 shows the templates that are used for the two GCR tasks: CommonGen and ConVE. The default prompt is adapted from Li et al. (2023) and are task-specific. On the other hand, the diversified prompt modifies the default prompt by appending a task-independent instruction that first

checks whether the diversity of the sentences generated in Step 1 is low, and if presents the generated sentences to the LLM and re-prompts it to generate more diverse set of sentences.

We use `GPT3.5-turbo` to predict the diversity of a given set of sentences using the prompt shown in Figure 6. This prompt uses five diversity categories (i.e. *very similar*, *somewhat similar*, *neutral*, *somewhat diverse*, and *highly diverse*) with increasing diversity with their definitions. Next, the set of sentences to be evaluated for their diversity is presented. Finally, the expected output format of the predictions is described at the end of the prompt. As recommended by Chen et al. (2023), we do not require the LLM to provide reasons for its predictions because it sometimes forces the model to focus on the reason generation than the prediction.

After the LLM's evaluation, the classification are mapped to values from 1 to 5 where 1 being highly similar to 5 being highly diverse. For each sentences set, we request the LLM to perform the evaluation three times and take the average score.

C Human Evaluation

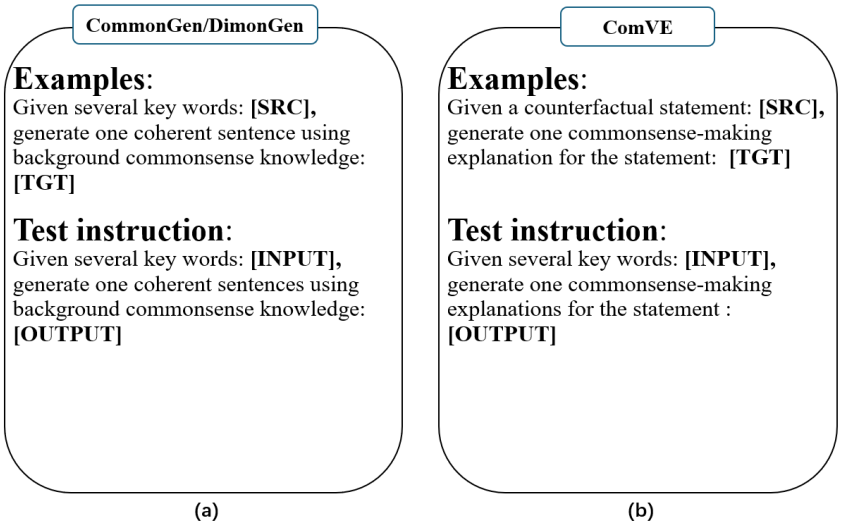
As human-annotators, we recruited eight graduate students from the department of computer science who specialise in NLP and fluent speakers in English. We provided the human annotators with the same set of instructions as we provided to the LLMs. The annotators were allowed deviate from the provided instructions when handling edge cases. Apart from the instruction in Figure 6, we also instruct the human annotators to focus on the diversity, and not so much on their commonsense quality, which we evaluate separately using semi-automatic metrics by comparing against human-written reference sentences in the evaluation benchmarks. Moreover, we informed the annotators that their evaluations would be used in a comparative analysis with the scores generated by an LLM.

D Addional Generation Examples

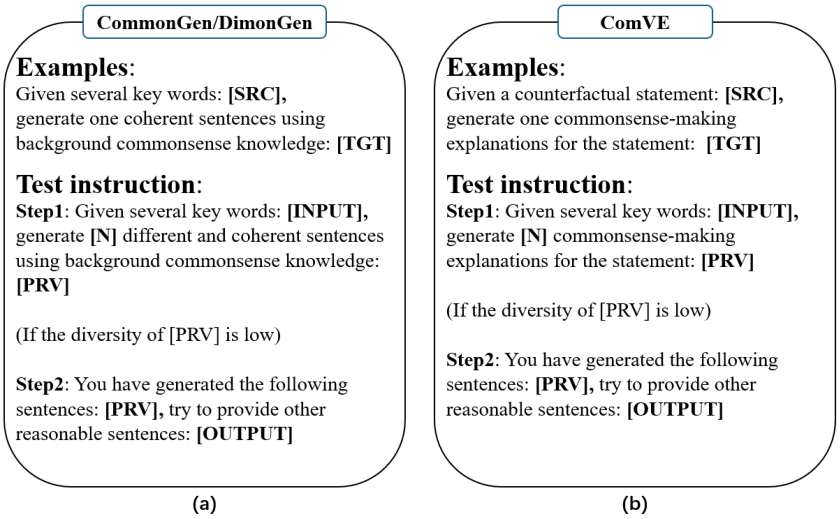
We show additional sentences generated by our proposed methods for the CommonGen, ComVE and DimonGen datasets in the Table 5.

CommonGen	
Keyword	branch ground climb tree jump
Sentences	The squirrel leaped from one branch to another, skillfully maneuvering its way up the tree without touching the ground. A squirrel climbs up a tree branch, jumps onto the ground, and starts to climb another tree. The cat tried to climb the tree, but it fell off the branch and landed on the ground. The monkey enjoyed swinging from branch to branch and would occasionally jump down to the ground. A squirrel jumps from a tree branch to the ground and climbs up another tree. A monkey climbs up a tree branch and jumps down to the ground.
Keyword	sit front table food laugh
Sentences	A couple sits at a table in front, enjoying their meal and sharing laughter. A group of friends sit in front of a table filled with delicious food, laughing and enjoying each other's company. Friends gather around a table, sitting in the front, while they eat and share laughter. People sit in front of a table, laughing while enjoying their food. Colleagues sit together at a table, eating and laughing as they enjoy their meal. Friends sit in front of a table, enjoying good food and laughing together.
ComVE	
Statement	She parachuted into the grocery store from a plane.
Explanation	Parachuting is not a typical mode of transportation for grocery shopping. Grocery stores are not designed to accommodate parachute landings. It is highly unlikely for someone to parachute into a grocery store from a plane.
Statement	You can drive after drinking.
Explanation	Driving under the influence of alcohol is a serious offence and can result in accidents or harm to oneself or others. You should never drink and drive as it impairs your judgment and reaction time. It is illegal and extremely dangerous to drive after consuming alcohol.
DimonGen	
Keyword	cloud mountain
Sentences	As we climbed the mountain, the clouds grew thicker, obscuring our view. The mountain stood tall and majestic, with clouds swirling around its peaks. The mountain peak pierced through the clouds, reaching towards the sky.
Keyword	race victory
Sentences	The underdog pulled off an unexpected victory in the race, leaving the favorite trailing behind. With a burst of speed and determination, the runner sprinted towards the finish line, securing a triumphant victory. After a fierce race, the champion celebrated their victory with a crowd cheering and fireworks lighting up the sky.

Table 5: More examples produced by our proposed ICD method on the CommonGen, ComVE and DimonGen datasets



(a) Default instructions



(b) Diversified instructions

Figure 5: The templates used by the default and the diversified prompt instructions for the CommonGen/DimonGen (shown on the left, (a)) and ComVE (shown on the right, (b)) tasks. Few-shot examples are included in each prompt where **[SRC]** denotes the set of input concepts and **[TGT]** the corresponding sentences in CommonGen. For a given set of **[INPUT]** concepts, the LLM is then required to generate sentences at the slot **[OUTPUT]**.

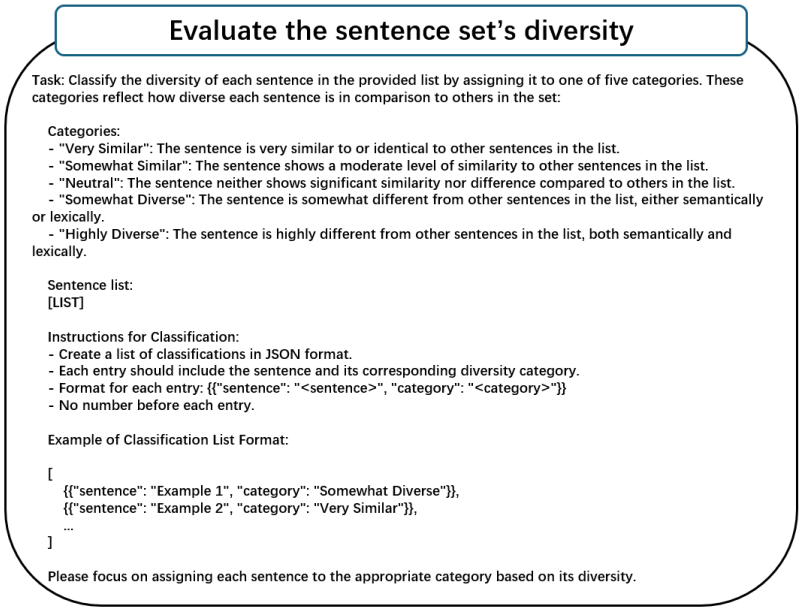


Figure 6: The instructions provided to `GPT3.5-turbo` for predicting the diversity of a given set of sentences. Diversity is predicted according to five categories: *very similar*, *somewhat similar*, *neutral*, *somewhat diverse*, and *highly diverse*. Definitions of the categories are included within the instructions. Next, the set of sentences to be evaluated for their diversity is presented. Finally, the expected output format of the predictions is described at the end of the prompt. As recommended by Chen et al. (2023), we do not require the LLM to provide reasons for its predictions because it sometimes forces the model to focus on the reason generation than the prediction.