# Retrieval-augmented Diffusion Language Model for Generative Commonsense Reasoning

Anonymous ACL submission

## Abstract

The ability of *generative commonsense reasoning* (GCR) reflects how well an AI system can produce trustworthy outputs that align with real-world commonsense knowledge. Despite the growing research efforts towards improved GCR, current studies still fall short in robustness and token-by-token generation. In this work, we propose a novel **R**etrieval-**a**ugmented **Di**ffusion Language Model **for G**enerative **C**ommonsense **R**easoning **(RaDi4GCR)**. RaDi4GCR not only allows for gradually refining the output via the denoising process, but also improves generation quality by injecting contextually relevant retrieved information, especially for low-resource scenarios that purely relying on parametric knowledge would suffer from. A comprehensive evaluation on the CommonGen benchmark demonstrates that RaDi4GCR significantly outperforms the state-of-the-art baseline (a 9.5% improvement in terms of SPICE), as well as surpassing multiple cutting-edge LLMs (such as GPT-4o and Llama3).

## 1 Introduction

Benefiting from the revolution of generative AI, modern systems based on large language models (LLMs) can smoothly interact with users and automate numerous tasks, relieving people from the tedious process of searching, extracting, and digesting information on their own. However, developing AI systems that can perform human-like natural communication and reasoning remains an open challenge. A fundamental problem is the so-called commonsense intelligence (Choi, 2022). In this work, we focus on *generative commonsense reasoning* (GCR), which reflects how well an AI system can generate trustworthy outputs that align with real-world commonsense knowledge. Although LLMs (e.g., GPT series (Brown et al., 2020) and Llama (Touvron et al., 2023)) have demonstrated remarkable performance in many tasks, such as machine translation (Zhu et al., 2024) and arithmetic reasoning (Li et al., 2023), they still fall short in robustness on GCR. The primary limitations are that: (1) The generating capability is constrained to the embedded parametric knowledge, which may result in the problem of *hallucination* (Huang et al., 2024). For instance, the generated content sometimes violates real-world commonsense knowledge (Cui et al., 2024). (2) Most LLMs are trained in a generalized autoregressive manner (Brown et al., 2020). The generation process is formulated as a Markov chain, where the next token is predicted with the condition of previously generated tokens. As a result, autoregressive LLMs usually suffer from higher latency in inference (Gu et al., 2018). Moreover, the diversity of generated sequences has been limited since the next-token prediction objective subsequently decoded high-likelihood, non-diverse sequences that only captured a limited input context (Gao et al., 2024a).

Motivated by the aforementioned shortcomings, we propose a novel **Re**trieval-**a**ugmented **Di**ffusion Language Model **for G**enerative **C**ommonsense **R**easoning **(RaDi4GCR)** by effectively fusing retrieval-augmented generation (RAG) and the diffusion language model. Specifically, by integrating RAG, RaDi4GCR enables us to inject factual or contextually relevant retrieved knowledge, potentially reducing the burden and reliance on the internal model parameters. Benefiting from the adoption of the diffusion language model, RaDi4GCR can handle continuous latent representations (Ho et al., 2020) and generate sequences in a non-autoregressive manner, leading to faster decoding and iterative refinement (Ghazvininejad et al., 2019; Lee et al., 2018). In order to effectively evaluate the effectiveness of RaDi4GCR, we conduct a series of experiments based on the benchmark dataset CommonGen (Lin et al., 2020). The experimental results show that: (1) RaDi4GCR not only

significantly outperforms the state-of-the-art baseline (a 9.5% improvement in terms of the key metric SPICE), but also shows superior performance over multiple cutting-edge LLMs (such as GPT-4o and Llama3) under the same RAG setting.

## 2 Related Work

### 2.1 Diffusion Language Models

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) have emerged as a new paradigm of generative models and have achieved significant progress in various generation tasks, such as image synthesis (Dhariwal and Nichol, 2021; Ho et al., 2020), video generation (Ho et al., 2022) and human motion generation (Tevet et al., 2023). Inspired by their success in computer vision, researchers have explored adopting diffusion models for text generation tasks, where the discrete nature is a key challenge (Li et al., 2022; Gong et al., 2023; Yuan et al., 2024; Hoogeboom et al., 2021; Savinov et al.; Lin et al., 2023). Hoogeboom et al. (2021); Savinov et al. transfer the diffusion models to discrete space and adopt the discrete diffusion model for text generation tasks; Li et al. (2022); Gong et al. (2023); Yuan et al. (2024) encode discrete text tokens to continuous word embeddings, thus leveraging the strong power of continuous diffusion models; GENIE (Lin et al., 2023) introduces a novel diffusion language model pretraining framework for text generation tasks that consists of an encoder and a diffusion-based decoder, which can generate text by gradually transforming a random noise sequence into a coherent text sequence. In this work, we adopt GENIE as the base for integrating RAG.

### 2.2 RAG and Its Usage with Diffusion Models

RAG (Lewis et al., 2020; Guu et al., 2020; Ram et al., 2023) has been proposed as a promising approach to cope with LLMs' inherent problem of hallucination, especially on knowledge-intensive tasks (Gao et al., 2024b), such as question answering and fact verification. By retrieving relevant information from external knowledge sources, RAG provides additional supplement contexts to LLMs through techniques like in-context learning (Ram et al., 2023), thereby improving their generation quality and reliability. The existing studies on RAG mainly focus on knowledge-intensive tasks, while the way of incorporating fine-grained commonsense knowledge has not been well explored yet. In recent years, the CV community has also started to explore leveraging RAG to improve image generation. RDM (Blattmann et al., 2022) performs image synthesis conditioned on the CLIP (Radford et al., 2021) embedding of retrieved nearest neighbors images; knn-diffusion (Sheynin et al.) leverages large-scale retrieval methods to train a substantially small and efficient text-to-image diffusion model without any text and generates out-of-distribution images by simply swapping the retrieval database at inference time; Re-Imagen (Chen et al.) uses retrieved information to produce high-fidelity and faithful images even for rare or unseen entities. Motivated by their progress, this work explores the integration of RAG with diffusion language models for generative commonsense reasoning tasks.

### 2.3 Generative Commonsense Reasoning

Lin et al. (2020) proposed the task of generative commonsense reasoning together with the CommonGen benchmark as generating coherent sentences following common sense given a set of commonsense concepts. After that, Zhao et al. (2022) revisited the performance of pre-trained models (PTM) on the generative commonsense reasoning task and proposed a pre-ordering approach to elaborately manipulate the order of the given concepts before generation. Recently, DimonGen (Liu et al., 2023) proposed a new benchmark focusing on the diversity in generative commonsense reasoning based on CommonGen. SituatedGen (Zhang and Wan) introduced a challenging task to incorporate geographical and temporal contexts into generative commonsense reasoning as a complement to CommonGen. As a further endeavor to the aforementioned studies, we focus on how to improve GCR by effectively fusing RAG and the diffusion language model.

## 3 Retrieval-augmented Diffusion Language Model

In this section, we detail how to tailor RaDi4GCR for the generative commonsense reasoning task. We start by briefly introducing the continuous diffusion model and the CommonGen task. Then we describe the overall framework of RaDi4GCR, followed by the details on training and inference.

## 3.1 Preliminary

### 3.1.1 Continuous Diffusion Model

The continuous diffusion model (Ho et al., 2020) is a latent variable model that contains two processes: *forward diffusion process* and *reverse denoising process*, each of which is a Markov chain.

The forward diffusion process gradually corrupts the data sample by adding Gaussian noise according to a variance schedule $\beta_1 ... \beta_t$. Specifically, given a data point $x_0$ sampled from a real-world data distribution $x_0 \sim q(x_0)$, for each time step $t \in \{1, 2, ..., T\}$, the corrupted latent variable $x_t$ is sampled from the distribution:

$$x_t \sim q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

where $\beta_t \in (0, 1)$ is the variance schedule that controls the noise scale added at each time step $t$. As $t$ increases, data sample $x_t$ will gradually and ultimately be corrupted to the standard Gaussian noise $x_T \sim \mathcal{N}(x_T; 0, \mathbf{I})$.

The reverse denoising process starts from the standard Gaussian noise $x_T \sim \mathcal{N}(x_T; 0, \mathbf{I})$ and inverts the diffusion process by gradually reconstructing the data sample through the learned parameterized denosing distribution.

$$x_{t-1} \sim p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2)$$

where $\mu_\theta$ and $\Sigma_\theta$ represent the approximated mean and variance of the denoising Gaussian distribution. Following DDPM (Ho et al., 2020), the variance $\Sigma_\theta$ is ignored and the parameterized mean $\mu_\theta$ is further factorized as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)) \quad (3)$$

where $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\epsilon_\theta$ predicts the noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ determing $x_t$ from $x_0$ at each time step $t$, which is parameterized by a neural network like U-Net (Ronneberger et al., 2015) or Transformer (Vaswani, 2017).

Finally, the continuous diffusion model is trained by optimizing the variational lower bound (VLB) on the negative log likelihood $-logp_\theta(x_0)$, which can be simplified to the mean squared error loss between $\epsilon$ and $\epsilon_\theta$ (Ho et al., 2020):

$$\mathcal{L}_\theta = \mathbb{E}_{t \sim Unif(1..T), \epsilon \sim \mathcal{N}(0, \mathbf{I})}[||\epsilon - \epsilon_\theta(x_t, t)||^2] \quad (4)$$

### 3.1.2 The Evaluation Task on GCR

Our proposed approach is evaluated on CommonGen (Lin et al., 2020), which is a widely used benchmark dataset for evaluating GCR. Given a set of unordered concepts $C = \{c_1, c_2, ..., c_k\}$, where $c_i \in \mathcal{C}$ is either a common object or an action and $\mathcal{C}$ denotes the concept vocabulary, the goal is to generate a coherent sentence $y \in \mathcal{Y}$ using all given concepts. Meanwhile, the sentence should describe an everyday scenario that should not contradict real-world commonsense knowledge.

## 3.2 The Framework of RaDi4GCR

Inspired by the previous studies on the integration of RAG for text generation (Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Ram et al., 2023), RaDi4GCR consists of three stages for GCR as illustrated in Figure 1, which will be further discussed in the following sections.

### 3.2.1 Retriever

In the first stage, a retriever is employed to retrieve relevant documents from an external knowledge base to provide factual or contextually relevant knowledge (the leftmost module of Figure 1). In this work, we investigate two types of retrievers for retrieval augmentation. (1) *Sparse retriever*. Specifically, we adopt BM25 (Robertson and Zaragoza, 2009) as the sparse retriever, which represents documents into bag-of-words representation. The relevance score of a document w.r.t. a specific query is determined based on the term frequency of the query terms in the document and the number of documents in the corpus that contains the query terms. (2) *Dense retriever*. We fine-tune the DPR model (Karpukhin et al., 2020), which employs a dual-encoder structure to generate query and document embeddings using separate query and document encoders. The similarity between these embeddings is calculated using the inner product.

### 3.2.2 Text Input Encoder

After retrieving the relevant information, it is crucial to bridge the gap between the discrete nature of textual data and the continuous input space of the diffusion model to effectively apply it to the GCR task. To address this challenge, we employ a 6-layer transformer encoder architecture to map the input text sequences into continuous embeddings (the middle module of Figure 1).

Formally, given a concept set $C = \{c_1, c_2, \ldots, c_m\} = \{w_1^c, w_2^c, ..., w_m^c\}$ comprising $m$ concepts $c_m$, each of which is a text token $w_m^c$, and the corresponding top-$k$ relevant

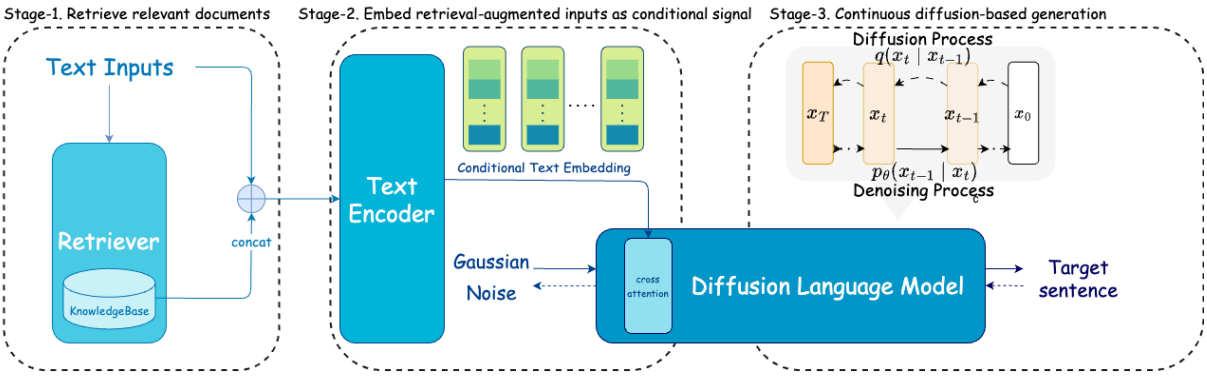


Figure 1: The overall framework of RaDi4GCR. The process contains three stages: 1) in Stage-1, the inputs are served as queries and relevant documents are retrieved through a retriever; 2) in Stage-2, the inputs and retrieved documents are concatenated and embedded to continuous embedding as the conditional signal through a text encoder; 3) in Stage-3, the diffusion language model performs the diffusion and denoising process to generate target sequences conditioned on the embeddings of Stage-2.

documents $D_k = \{d_1, d_2, \ldots, d_k\}$ retrieved from an external knowledge source $\mathcal{D}$, where $d_k$ is a specific relevant document that contains $j$ distinct tokens: $d_k = \{w_1^{d_k}, w_2^{d_k}, \ldots, w_j^{d_k}\}$, the input sequence is constructed by concatenating $C$ and $D_k$:

$$s = [C; D_k] = \{w_1^c, w_2^c, \ldots, w_m^c, w_1^{d_1}, w_2^{d_1}, \ldots, w_j^{d_k}\} \quad (5)$$

which is then encoded into the continuous embedding representation $H_s$ by the input encoder:

$$H_s = \{h_1, h_2, ..., h_n\} = Encoder(s) \quad (6)$$

where $h_i$ is the encoded continuous embedding of the $i$-th token in $s$. By leveraging the input encoder, the discrete input texts are transformed into the continuous space, making them compatible with the diffusion language model.

#### 3.2.3 Diffusion-based Generator

In the third stage, the diffusion-based generator is leveraged for generating high-quality sequences based on the input concepts and retrieved documents. In this work, we construct our diffusion language model generator of RaDi4GCR based on GENIE (Lin et al., 2023). Specifically, we adopt a 6-layer transformer decoder architecture with cross-attention layers as our generator, denoted as $\epsilon_\theta^{'}(t, x_t, H_s)$, which predicts the Gaussian noise $\epsilon$ in Eq-4 at each denoising step conditioned on the current time step $t$, the continuous latent representation of the target text $x_t$, and the source input representation $H_s$.

**Cross-Attention Conditioning** To integrate RAG with the diffusion language model, it is essential to effectively incorporate retrieved relevant documents into the generating process. In RaDi4GCR, we draw inspiration from the retrieval-augmented diffusion framework **RDM** proposed in (Blattmann et al., 2022). Specifically, we incorporate the cross-attention mechanism within the diffusion decoder model to enable conditional generation (the rightmost module of Figure 1). During the reverse denoising process, at each denoising step $t$, the augmented state $x_t^{'}$ is computed by performing cross-attention on the retrieval augmented input source embedding computed in Eq-6:

$$x_t^{'} = CrossAttn(x_t, H_s) \quad (7)$$

The predicted Gaussian noise $\epsilon_\theta$ at the time step $t$ in Eq-4 is then calculated by the diffusion generator as:

$$\epsilon_\theta^{'}(t, x_t, H_s) = \epsilon_\theta(CrossAttn(x_t, H_s), t) = \epsilon_\theta(x_t^{'}, t) \quad (8)$$

### 3.3 Training and Inference

To train RaDi4GCR for CommonGen, in the diffusion process, the target sentence $S$ is converted to continuous embedding representation using the text encoder and incrementally corrupted to Gaussian noise $x_t$ in $t$ steps as in Eq-1. Then we use the concept sets $C$ as queries and retrieve top-$k$ relevant documents $D_r$ and compute the retrieval augmented source input embedding $H_s$ as in Eq-6. Finally, we perform the reverse denoising process with $x_t$ conditioning on $H_s$ as in Eq-8 to predict

the Gaussian noise added at step $t$ and compute the loss $\mathcal{L}_\theta$ as in Eq-4.

In the inferencing stage, we start at the time step $t$ and directly sample $x_t$ from a standard Gaussian distribution. Similar to the training stage, we retrieve top-$k$ relevant documents using source concept sets as queries and compute $H_s$, which is then used as the condition in the denoising process with $x_t$. After arriving at step $t = 0$, the predicted final outputs $x_0^{'}$ are decoded to discrete text tokens based on the clamping trick (Li et al., 2022) that replaces the predicted $x_0^{'}$ with its closest word embedding.

# 4 Experiments

## 4.1 Dataset and External Corpus

The proposed RaDi4GCR model is trained and verified on the CommonGen datasets (Lin et al., 2020), which is constructed by utilizing several existing corpora to sample frequent commonsense concepts, and the golden sentences are collected by employing AWT crowd-sourcing workers. Basic statistics of the datasets can be found in Table 1.

| Statistics | Train | Dev | Test |
|---|---|---|---|
| # Concept-Sets | 32,651 | 993 | 1,497 |
| # Sentences | 67,389 | 4,018 | 6,042 |
| Avg. Sentences per Set | 2.06 | 4.04 | 4.04 |
| Avg. Sentence Length | 10.54 | 11.55 | 13.34 |

Table 1: Statistics of CommonGen datasets.

**Commonsense Knowledge Corpus**. We adopt the RACo (Yu et al., 2022) corpus as our external knowledge base for retrieval. RACo is a collection of over 20 million commonsense documents from three knowledge sources, including 1) human annotated facts (HAF); 2) commonsense benchmark datasets (CBD); and 3) commonsense relevant corpus (CRC) for commonsense knowledge retrieval, as illustrated in table 2.

| Corpus | # Instance | Avg. Word |
|---|---|---|
| HAF-corpus | 3,561,762 | $11.06 \pm 5.86$ |
| CBD-corpus | 2,881,609 | $12.78 \pm 9.31$ |
| CRC-corpus | 14,587,486 | $17.76 \pm 10.4$ |

Table 2: Statistics of the RACo corpus.

## 4.2 Evaluation Metrics

We evaluate the performance from two perspectives: *token-level similarity* and *semantic-level similarity*. Token-level similarity is assessed by comparing the generated sentences with reference human-written sentences, where BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) are used. Semantic-level similarity is measured through CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). We note that **SPICE** is viewed as the primary metric since it demonstrates the highest alignment with human evaluation (Lin et al., 2020).

## 4.3 Baseline Methods

In this work, we compare RaDi4GCR with two types of representative methods.

**Retrieval-augmented Baselines:** 1) **DKMR**$^2$ (He et al., 2022) is a retrieval-enhanced method for generative commonsense reasoning that uses metric-guided distillation to improve the ranker and a progressive distillation strategy to improve the retriever. 2) **KFCNet** (Li et al., 2021) is a novel knowledge filtering and contrastive learning framework for retrieval-augmented commonsense sequence generation. 3) **RACo** (Yu et al., 2022) proposed a unified framework of retrieval-augmented commonsense reasoning. 4) **MORE** (Cui et al., 2024) proposed a novel multi-modal retrieval augmentation framework that leverages both text and images to enhance the commonsense ability of language models.

**Cutting-edge LLMs** To evaluate the effectiveness of utilizing the diffusion language model for text generation, we compare the performance of our proposed RaDi4GCR framework against several LLM baselines under the same RAG setting. Specifically, we exploit **Llama3-8B-Instruct**, **Llama3-70B-Instruct** (Dubey et al., 2024), and **GPT-4o** (Hurst et al., 2024) as representative baselines.

## 4.4 Implementation Details

For the diffusion generator, we fine-tune RaDi4GCR based on the released GENIE checkpoint that is pretrained on 160GB text data from news, books, stories, and web text (Lin et al., 2023). The model consists of a 6-layer encoder and a 6-layer decoder, with totally around 140 million parameters. Following GENIE, we use Adam opimizer (Kingma and Ba, 2015) with the learning rate of 5e-5 and batch size 128. All models were trained on a total of 120000 steps with 7200 warm-up steps. The diffusion step is set to 2000 with the *sqrt* noise schedule. For the dense

| Method | BLEU-4 | ROUGE-L | METEOR | CIDEr | SPICE* |
|---|---|---|---|---|---|
| KFCNet† | 51.46 | 47.52 | 38.92 | 20.98 | 39.15 |
| RACo† | 42.76 | 48.19 | 35.80 | 18.89 | 33.89 |
| RACo‡ | 45.91 | 62.96 | 40.90 | 17.50 | 39.23 |
| MORE_OPT2.7b† | 32.78 | 57.07 | 32.15 | 17.04 | 32.94 |
| DKMR$^2$† | 64.19 | 49.22 | 46.01 | **24.85** | 43.37 |
| **GENIE w/o retrieval†** | 19.6 | 36.0 | - | 10.3 | 23.4 |
| **RaDi4GCR w/ BM25-top10‡** | **66.56** | **75.47** | **53.65** | 23.64 | **47.49** |

Table 3: The overall experiment results of RaDi4GCR on CommonGen (v1.0) compared with baseline methods. * denotes the primary metric that should be focused on. † denote results reported in their paper. ‡ denotes results reproduced in our experiments. The best performances are in bold, and the second-best ones are underlined.

retriever, we fine-tuned the DPR model following the same training setting as RACo (Yu et al., 2022). For the sparse retriever, we built the inverted index of RACo Commonsense corpus using Pyserini (Lin et al., 2021). The RAG pipeline is constructed based on the FlashRAG toolkit (Jin et al., 2024). All the metrics are computed using NLGEval (Sharma et al., 2017). All models are trained using 2x A100 80GB GPUs.

# 5 Results and Analysis

## 5.1 Results Overview

The experimental results are shown in Table 3 , where we present the performance of our best-performing model, **RaDi4GCR w/ BM25-top10** that adopts BM25 sparse retrieval and utilizes the top-10 relevant documents, compared with results from baselines. As demonstrated in the table, our proposed approach outperforms all baseline methods across most evaluation metrics. RaDi4GCR achieved the highest scores in BLEU-4 (66.56), ROUGE-L (75.47), METEOR (53.65), and SPICE (47.49), and competitive scores in CIDEr (23.64), reflecting its superior ability to generate coherent and contextually relevant commonsense sentences.

Notably, RaDi4GCR significantly outperforms the multi-modal RAG baseline, MORE, which utilizes both relevant texts and images. In contrast, RaDi4GCR relies solely on retrieved knowledge in a single modality of text. These results emphasize RaDi4GCR's strong ability to effectively utilize retrieved information to enhance generation. The overall results underscore the model's robust capability to leverage retrieved knowledge for improved text generation, positioning it as a highly effective solution for the task of generative commonsense reasoning.

## 5.2 Analysis

In order to further assess the effectiveness of our proposed methods, we examine the performance of different variants of RaDi4GCR to demonstrate the impacts of various aspects of RaDi4GCR.

### 5.2.1 Impact of Retrieval Augmentation

In Table 3, the last two bold rows demonstrate results of the pre-trained diffusion language model GENIE fine-tuned on CommonGen and our proposed RaDi4GCR with the top 10 relevant documents from bm25 sparse retrieval augmented generation. While the pre-trained diffusion language model shows moderate performance, after augmenting the generation with retrieval from external knowledge corpus, our proposed RaDi4GCR outperforms the vanilla diffusion language model by a large margin. These results demonstrate the effectiveness of leveraging retrieval augmentation with the diffusion language model for generative commonsense reasoning tasks.

### 5.2.2 Impact of Different Retrieval Methods

As illustrated in Figure 2, the sparse retrieval method that adopts BM25 generally outperforms the dense retrieval method that adopts DPR in metrics of both token-level and semantic level. Although dense retrievers provide substantial performance improvements when compared to traditional sparse retrievers (Arabzadeh et al., 2021), the observed counterfactual result may stem from the distinct task design of CommonGen, which requires the model to generate coherent commonsense sentences that contain the provided concepts. While dense retrieval methods retrieve relevant information based on the semantic similarity between the query and the candidate documents computed by the inner products of their dense embeddings, sparse retrieval methods like BM25 retrieve relevant information based on the token matching

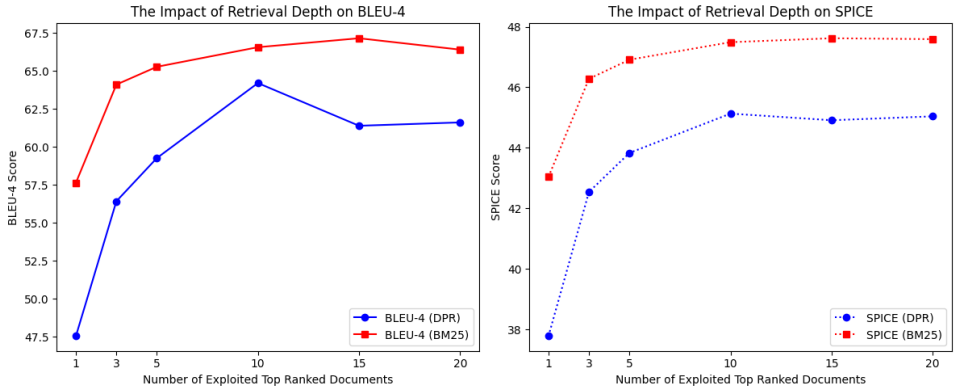


Figure 2: The experiment results of RaDi4GCR with different retrieval methods and different retrieval depths on BLEU-4 and SPICE.

| Method | BLEU-4 | ROUGE-L | METEOR | CIDEr | SPICE |
|---|---|---|---|---|---|
| Diffusion Language Model (140M) | | | | | |
| w/o RAG | 19.6 | 36.0 | - | 10.3 | 23.4 |
| RaDi4GCR w/ BM25 top3 | 64.10 | 73.17 | 51.72 | 22.77 | 46.28 |
| RaDi4GCR w/ BM25 top5 | 65.27 | 74.43 | 52.63 | 23.25 | 46.91 |
| RaDi4GCR w/ BM25 top10 | **66.56** | **75.47** | **53.65** | **23.64** | **47.49** |
| Llama3-8B-instruct | | | | | |
| w/o RAG | 18.76 | 42.84 | 29.49 | 9.10 | 27.01 |
| w/BM25 top3 | 29.42 | 51.63 | 35.08 | 12.97 | 35.44 |
| w/ BM25 top5 | 29.53 | 51.50 | 34.92 | 13.13 | 35.21 |
| w/ BM25 top10 | 29.47 | 51.79 | 35.07 | 13.05 | 35.33 |
| Llama3-70B-instruct | | | | | |
| w/o RAG | 17.85 | 42.55 | 29.90 | 8.05 | 27.19 |
| w/ BM25 top3 | 26.62 | 49.52 | 33.48 | 12.04 | 32.57 |
| w/ BM25 top5 | 26.21 | 49.52 | 33.42 | 12.01 | 32.74 |
| w/ BM25 top10 | 26.69 | 49.09 | 33.54 | 12.11 | 32.60 |
| GPT-4o | | | | | |
| w/o RAG | 38.45 | 57.87 | 37.54 | 15.82 | 37.69 |
| w/ BM25 top3 | 38.16 | 57.87 | 37.56 | 15.81 | 37.59 |
| w/ BM25 top5 | 38.29 | 58.08 | 37.62 | 15.86 | 37.33 |
| w/ BM25 top10 | 38.77 | 57.62 | 37.43 | 15.85 | 36.83 |

Table 4: Comparison between LLM with RAG and RaDi4GCR. Both LLM and RaDi4GCR adopt the same retrieval setting of BM25 sparse retrieval.

between the candidate documents and the query, e.g., the given concept sets, which could retrieve and provide more useful relevant information that contains the concept tokens, thus better integrated with the diffusion language models for generating target sequences.

#### 5.2.3 Impact of Different Retrieval Depth

As illustrated in Figure 2, when the number of adopted relevant documents is below top-10, the performance of RaDi4GCR, whether using a dense (DPR) or sparse (BM25) retriever, improves as the number of retrieved relevant documents increases. This result aligns with our expectations, as incorporating more relevant documents provides additional complementary information, thereby enhancing the model's reasoning and generation capabilities.

Specifically, for RaDi4GCR with BM25, increasing the number of adopted relevant documents from top-3 to top-10 yields a 3.7% improvement in BLEU-4 and a 2.5% improvement in SPICE. In contrast, for RaDi4GCR with DPR, the same increase in the number of relevant documents leads to significantly larger gains: 12.1% in BLEU-4 and 5.7% in SPICE. These results suggest that RaDi4GCR benefits more from additional context when using a dense retriever (DPR) compared to a sparse retriever (BM25).

However, for both retrieval methods, when the number of adopted relevant documents exceeds 10, the performance on BLEU-4 and SPICE remains unchanged or even decreases. This finding is consistent with prior work (Yu et al., 2022) as incorporating more relevant documents may also introduce more noisy information, thus degrading the

model's performance. Compared to BM25, the performance degradation of DPR is more prominent, which aligns with with the findings in (Cuconasu et al., 2024) as introducing semantically aligned yet non-relevant documents increased the complexity of inputs thus potentially misguides models away from the correct response.

#### 5.2.4 Comparison of LLM and DLM

To evaluate the performance of the diffusion language model (DLM) as the generator compared with LLMs, in this work we also conduct extensive experiments of leveraging LLMs as the generator within the same RAG pipeline. The comparison of LLM with RAG and RaDi4GCR is illustrated in Table 4. We report the results of our fine-tuned RaDi4GCR as well as the results of different LLMs in a few-shot prompt setting. The prompt used in the experiments can be found in A.

As demonstrated in the first block of the table, after fine-tuning with RAG, the DLM beats all other LLMs with only around 140 million parameters across all metrics, which demonstrates the efficiency and effectivity of DLM on the generative commonsense reasoning tasks and highlights the potential of DLM as an alternative of generalized LLMs for domain-specific text generation tasks.

From the second and third blocks of the table, we found that the Llama3 model with 8 billion parameters consistently outperforms its 70 billion-parameter counterpart across all retrieval-augmentation settings. For both versions of the Llama3 model, utilizing RAG in the generation enhances the performances prominently. In the last block, although the performance of GPT-4o outperforms Llama3 of both 8B and 70B even without RAG, the overall performance still falls behind that of RaDi4GCR. Different from Llama3, adding RAG to the generation barely improves the performance.

These findings underscore that the generator model is crucial in a RAG pipeline for the generative commonsense reasoning task. While LLMs demonstrate impressive general-purpose capabilities, they struggle with commonsense reasoning due to their lack of commonsense knowledge without task-specific fine-tuning, which is costly in computational resources. In contrast, DLM with RAG could effectively capture and integrate commonsense knowledge after fine-tuning with significantly smaller model sizes. When compared with other fine-tuned LMs of competitive model size, DLM still strongly outperforms all the baselines, as demonstrated in Table 3, which demonstrates its effectiveness. Its ability to gradually refine output sequences in reasoning steps allows for more coherent and contextually grounded responses, making them well-suited for generative commonsense reasoning.

## 6 Conclusion

In this work, we propose a novel method, RaDi4GCR, for GCR by effectively integrating RAG and the diffusion language model. We conduct a series of experiments over the CommonGen dataset. The experimental results demonstrate that RaDi4GCR significantly outperforms the state-of-the-art baseline method, as well as surpassing multiple cutting-edge LLMs. This study highlights the potential of a diffusion language model with RAG as an alternative to autoregressive LLMs for GCR. In particular, the factors, such as different retrieval strategies and the number of relevant documents, significantly affect the performance of RaDi4GCR. Careful examinations of these factors are highly recommended in the development of GCR methods when RAG is used.

## 7 Limitations

To the best of our knowledge, our work has two main limitations that should be further explored in the future. First, our method employs a relatively easy RAG setting (Gao et al., 2024b), where we use a pre-fixed number of relevant documents without further assessing the relevance of retrieved documents w.r.t. the final generation target. As a result, the augmentation could be suboptimal. A possible future work may incorporate more advanced RAG techniques like adaptive retrieval for better integration of the retrieved documents with the diffusion language model. Second, we only verify our approach on the CommonGen dataset. We leave the evaluation of our proposed approach on more advanced scenarios, including other generation tasks and multimodal tasks, to future work.

## References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing.

Negar Arabzadeh, Xinyi Yan, and Charles L. A. Clarke. 2021. Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 2862–2866, New York, NY, USA. Association for Computing Machinery.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. 2022. Retrieval-augmented diffusion models. In *Advances in Neural Information Processing Systems*.

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.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *The Eleventh International Conference on Learning Representations*.

Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus*, 151(2):139–155.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 719–729, New York, NY, USA. Association for Computing Machinery.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. MORE: Multi-mOdal REtrieval augmented generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1178–1192, Bangkok, Thailand. Association for Computational Linguistics.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc.

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*.

Silin Gao, Mete Ismayilzada, Mengjie Zhao, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2024a. DiffuCOMET: Contextual commonsense knowledge diffusion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4809–4831, Bangkok, Thailand. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024b. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. DiffuSeq: Sequence to sequence text generation with diffusion models. In *International Conference on Learning Representations, ICLR*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Xingwei He, Yeyun Gong, A-Long Jin, Weizhen Qi, Hang Zhang, Jian Jiao, Bartuer Zhou, Biao Cheng, Sm Yiu, and Nan Duan. 2022. Metric-guided distillation: Distilling knowledge from the metric to ranker and retriever for generative commonsense reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 839–852, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*, volume 34, pages 12454–12465. Curran Associates, Inc.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Jiajie Jin, Yutao Zhu, Xinyu Yang, Chenghao Zhang, and Zhicheng Dou. 2024. Flashrag: A modular toolkit for efficient retrieval-augmented generation research. *CoRR*, abs/2405.13576.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 1173–1182. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 4328–4343. Curran Associates, Inc.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

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. *Findings of EMNLP*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. 2023. Text generation with diffusion language models: a pre-training approach with continuous paragraph denoise. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Chenzhengyi Liu, Jie Huang, Kerui Zhu, and Kevin Chen-Chuan Chang. 2023. 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.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Nikolay Savinov, Junyoung Chung, Mikolaj Binkowski, Erich Elsen, and Aaron van den Oord. Step-unrolled denoising autoencoders for text generation. In *International Conference on Learning Representations*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. knn-diffusion: Image generation via large-scale retrieval. In *The Eleventh International Conference on Learning Representations*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. 2024. Text diffusion model with encoder-decoder transformers for sequence-to-sequence generation. 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 22–39, Mexico City, Mexico. Association for Computational Linguistics.

Yunxiang Zhang and Xiaojun Wan. Situatedgen: Incorporating geographical and temporal contexts into generative commonsense reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, NeurIPS Datasets and Benchmarks 2023, New Orleans, LA, United States, December 10-16, 2023*.

Chao Zhao, Faeze Brahman, Tenghao Huang, and Snigdha Chaturvedi. 2022. Revisiting generative commonsense reasoning: A pre-ordering approach. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1709–1718, Seattle, United States. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

# A Appendix

The few-shot prompt we used in our experiments is as below:

SYSTEM Prompt

Given several concepts together with several reference documents, write a short and simple sentence that contains *all* the required words. Only give me the sentence and do not output any other words. The sentence should describe a common scene in daily life, and the concepts should be used in a natural way.

**Examples:**

- **Example 1:**
  **Concepts:** dog, frisbee, catch, throw
  **Sentence:** The dog catches the frisbee when the boy throws it into the air.
- **Example 2:**
  **Concepts:** apple, place, tree, pick
  **Sentence:** A girl picks some apples from a tree and places them into her basket.

USER Prompt

**Concepts:** {concepts}; **References:** {references}