Retrieval Augmentation for Commonsense Reasoning: A Unified Approach

Wenhao Yu[1], **Chenguang Zhu**[2], **Zhihan Zhang**[1], **Shuohang Wang**[2],
Zhuosheng Zhang[3], **Yuwei Fang**[2], **Meng Jiang**[1]
[1]University of Notre Dame, Indiana, USA
[2]Microsoft Cognitive Services Research, Washington, USA
[3]Shanghai Jiaotong University, Shanghai, China
[1]{wyu1, zzhang23, mjiang2}@nd.edu;
[2]{chezhu, shuow, yuwfan}@microsoft.com; [3]zhangzs@sjtu.edu.cn

Abstract

A common thread of retrieval-augmented methods in the existing literature focuses on retrieving encyclopedic knowledge, such as Wikipedia, which facilitates well-defined entity and relation spaces that can be modeled. However, applying such methods to commonsense reasoning tasks faces two unique challenges, i.e., the lack of a general large-scale corpus for retrieval and a corresponding effective commonsense retriever. In this paper, we systematically investigate how to leverage commonsense knowledge retrieval to improve commonsense reasoning tasks. We proposed a unified framework of **R**etrieval-**A**ugmented **Co**mmonsense reasoning (called RACo), including a newly constructed commonsense corpus with over 20 million documents and novel strategies for training a commonsense retriever. We conducted experiments on four different commonsense reasoning tasks. Extensive evaluation results showed that our proposed RACo can significantly outperform other knowledge-enhanced method counterparts, achieving new SoTA performance on the CommonGen[1] and CREAK[2] leaderboards. Our code is available at https://github.com/wyu97/RACo.

1 Introduction

Recent work has shown that scaling language models with considerably more data and parameters, such as GPT3-175B (Brown et al., 2020), could drive significant advances in commonsense reasoning tasks. Nevertheless, such models make predictions by only "looking up information" stored in their parameters, making it difficult to determine what knowledge is stored or has been already forgotten by the neural network (Guu et al., 2020). Besides, storage space is limited by the size of the neural network. In order to memorize more world knowledge, one must train ever-larger networks, which can be prohibitively expensive and slow.

The solution that may seem obvious at first glance is to grant language models free access to open-world sources of commonsense knowledge in a plug-and-play manner, instead of memorizing all world knowledge. To achieve this capability, language models must be able to *retrieve* relevant commonsense knowledge from an unbounded set of situations. Then, the language models can leverage the input text, as well as the retrieved information to produce the desired output.

Compared with the large-scale language model counterparts, e.g., UNICORN (Lourie et al., 2021), retrieval-augmented methods have three remarkable advantages: first, the knowledge is not stored implicitly in the model parameters, but is explicitly acquired in a plug-and-play manner, leading to great scalability; second, the paradigm generates text based on some retrieved references, which alleviates the difficulty of generating from scratch (Li et al., 2022); third, knowledge corpus can be constantly edited and updated by experts, making the model aware of the latest information. Besides, compared with knowledge graph inference model counterparts, e.g., QA-GNN (Yasunaga et al., 2021), retrieval-augmented methods allow more flexibility in accessing and using knowledge from different sources, because of the nature of commonsense knowledge, which cannot all be contained in a single knowledge graph defined by a certain schema (Yu et al., 2022b).

A common thread of retrieval-augmented methods in the existing literature focuses on retrieving encyclopedic knowledge such as Wikipedia, which lends itself to a well-defined space of entities and relations that can be modeled (Karpukhin et al., 2020; Lewis et al., 2020b; Yu et al., 2022a). However, retrieval-augmented methods for commonsense reasoning have been rarely studied in the literature. In this paper, we propose a unified frame-

[1]https://inklab.usc.edu/CommonGen/leaderboard.html
[2]https://www.cs.utexas.edu/~yasumasa/creak/leaderboard.html

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4364 - 4377
December 7-11, 2022

	RACo (this work)	AristoRoBERTa (Mihaylov et al., 2018)	Re-T5 (Wang et al., 2021)	KFCNet (Li et al., 2021)	OpenCSR (Lin et al., 2021)
Number of corpus types	**3**	1	1	1	1
Number of commonsense tasks	**4**	1	1	1	1
Number of docs for retrieval	**20M**	5K	0.8M	0.8M	1M

Table 1: Comparison of RACo to a few recent commonsense retrieval works in the field. Our work provides a more comprehensive and larger-scale multi-source commonsense corpus that can generalize to various tasks.

work of **R**etrieval-**A**ugmented **Co**mmonsense reasoning (RACo) to solve various commonsense tasks. RACo first retrieves relevant commonsense documents from a large-scale corpus, then combines the input text with the retrieved documents to produce the desired output. However, there are two main challenges in training a RACo model.

The first challenge to address is *what* commonsense knowledge to retrieve. Different from encyclopedic knowledge used in open-domain QA tasks, commonsense knowledge is very diverse, containing everyday events and their effects, facts about beliefs and desires, and properties of objects in human's daily life. Since commonsense involves various aspects including human interaction and object properties in everyday life, we collected a over 20 million commonsense documents collection from both open-domain knowledge sources (e.g., OMCS) that cover multiple domains of commonsense, and domain-specific sources (e.g., ATOMIC) that focus on particular commonsense types.

The second challenge is to address *how* to retrieve relevant commonsense knowledge from the corpus. Different from training a dense retriever on Wikipedia (Karpukhin et al., 2020), the heuristic of taking "documents containing correct answers" as positive candidates cannot be used because the output answer in commonsense reasoning tasks is usually not a substring of retrieved documents. For example, in binary question answering, the answer is *True* or *False* but it does not appear in the retrieved documents. Therefore, we propose novel strategies to construct question-document pairs for commonsense dense retriever training.

Overall, our main contributions in this work can be summarized as follows:

1. We collected and publicized a collection of over 20 million documents from three knowledge sources for commonsense knowledge retrieval.

2. We presented a unified framework of **R**etrieval-**A**ugmented **Co**mmonsense reasoning (RACo), and proposed novel strategies for training a strong commonsense knowledge retriever.

3. We evaluated our RACo on four types of commonsense reasoning tasks. Our experiments showed RACo can significantly outperform other knowledge-enhanced counterparts, achieving new SoTA on CommonGen and CREAK leaderboards.

2 Related Work

Though large-scale language models yield state-of-the-art performance on many commonsense reasoning tasks, their pre-training objectives do not explicitly guide the models to reason with commonsense knowledge such as the relation and composition of daily concepts in our lives (Zhou et al., 2021), leading to unsatisfactory performance in many real-world scenarios (Talmor et al., 2021; Zhu et al., 2022). Existing work has mainly explored two directions to improve their commonsense reasoning ability. The first is to pre-train or post-train a language model on commonsense corpora (Bosselut et al., 2019; Lourie et al., 2021; Zhou et al., 2021). When the commonsense materials are appropriately selected, this simple strategy could demonstrate significantly superior performance than vanilla pre-trained language models (Zhou et al., 2021). Notable methods include COMET (Bosselut et al., 2019), CALM (Zhou et al., 2021), UNICORN (Lourie et al., 2021), etc. Nonetheless, these methods still suffer from the same drawbacks as the pre-trained language models introduced in §1. The second is to explicitly introduce external knowledge from commonsense knowledge graphs to augment the limited textual information. (Lin et al., 2019; Ji et al., 2020). A KG often provides comprehensive and rich entity features and relations so models can easily traverse links to discover how entities are interconnected to express certain commonsense knowledge. Notable methods include KagNet (Lin et al., 2019), GRF (Ji et al., 2020), QA-GNN (Yasunaga et al., 2021), GreaseLM (Zhang et al., 2022), etc. However, commonsense knowledge lies at an unbounded set of facts and situations that usually cannot be covered by a single knowledge graph defined by a cer-

tain schema. Reasoning over multiple knowledge graphs is a challenging task.

Retrieval-augmented method is a new learning paradigm that fuses pre-trained language models and traditional information retrieval techniques (Lewis et al., 2020b). A few recent methods have explored retrieving *in-domain* commonsense documents from a task-relevant corpus to improve commonsense reasoning performance (Mihaylov et al., 2018; Wang et al., 2021; Li et al., 2021). We provide a detailed comparison in Table 1. Different from existing methods that focus on retrieving knowledge from *in-domain* corpus, our proposed RACo leverages a much larger and general commonsense corpus collected from multiple sources that provide supportive evidences for various commonsense reasoning tasks. Meanwhile, we proposed several novel strategies for training a commonsense retriever that can be generalized to different commonsense reasoning tasks.

3 Proposed Method

In this section, we elaborate on how to leverage commonsense knowledge retrieval from a large-scale corpus to improve various commonsense reasoning tasks, including commonsense corpus construction (§3.1), commonsense document retriever (§3.2) and commonsense document reader (§3.3). The architecture of RACo is shown in Figure 1.

3.1 Commonsense Corpus Construction

Commonsense knowledge includes the basic facts about situations in everyday life, which is shared by most people and implicitly assumed in communications (Li et al., 2022). Commonsense knowledge has two important properties: *large* and *diverse*.

Regarding the scale of knowledge, many commonsense corpus contains millions of statements. For example, Wiktionary has more than one million word definitions and descriptions in English. Meanwhile, the commonsense knowledge is diverse, involving various aspects including human interaction and object properties. For example, OMCS[3] covers multiple domains of commonsense such as everyday events and their effects (e.g., mop up the floor if we split food over it), facts about beliefs and desires (e.g., study hard to win scholarship), and properties of objects (e.g., goat has four legs). The diversity of knowledge is beneficial for retrieval-augmented methods because it enables

[3] https://en.wikipedia.org/wiki/OMCS

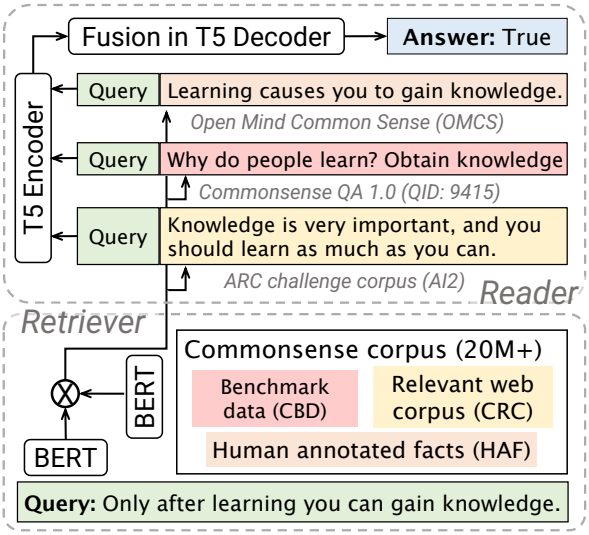

Figure 1: RACo has two major components: (i) a document retriever and (ii) a document reader. Specifically, the document retriever aims to fetch a handful of relevant documents from a large document collections. The document reader takes the input text, as well as the support documents to produce the desired output.

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

Table 2: Statistics for the commonsense corpus. The total size of these corpora exceeds 20M documents.

relevance comparison across different sources, and offers textual knowledge to easily augment the input of generation models by concatenation. To build a large-scale commonsense corpus covering diverse sources, we collected commonsense documents from the following three aspects: (i) human annotated facts; (ii) commonsense benchmark datasets; (iii) commonsense relevant web corpus. The statistics can be found in Table 2.

Human annotated facts (HAF). It contains factual commonsense either annotated by human annotators or written by domain experts, including OMCS (Havasi et al., 2010), ATOMIC (Sap et al., 2019a), Wiktionary (Meyer and Gurevych, 2012).

Commonsense benchmark datasets (CBD). It includes training data from 19 commonsense benchmark datasets, such as α-NLI (Bhagavatula et al., 2020). See Appendix A.1 for more details.

Commonsense relevant corpus (CRC). It consists of raw statements about commonsense from the web, usually after some simple filtering. We

obtained a filtered version from AI2 commonsense corpus, which is a merged corpus collected from ARC (Clark et al., 2018), QASC (Khot et al., 2020) and GenericsKB (Bhakthavatsalam et al., 2020).

3.2 Commonsense Document Retrieval

Given a collection of M commonsense documents, the goal of our retriever is to map all the documents in a low-dimensional vector, such that it can efficiently retrieve the top-k documents relevant to the input text. Note that M can be very large (e.g., over 20 million in our experiments) and k is usually small (e.g., 10 or 20 in our experiments).

In this work, we follow the neural document retriever DPR (Karpukhin et al., 2020) to employ two independent BERT (Devlin et al., 2019) models to encode the query and the document separately, and estimate their relevance by computing a single similarity score between their [CLS] token representations. Specifically, the document encoder $E_D(\cdot)$ which maps any text document to a low-dimensional real-valued vectors and builds an index for all the M documents used for retrieval. At runtime, it applies a different query encoder $E_Q(\cdot)$ that maps the input question to a vector of the same dimension as the document vector, and retrieves top-k documents of which vectors are the closest to the question vector. The similarity between the question and the document is calculated by the dot product of their vectors.

$$\text{sim}(q, d) = E_Q(q)^\intercal E_D(d). \quad (1)$$

Recent efforts have shown that DPR transfer poorly to other domains (Li and Lin, 2021; Kulshreshtha et al., 2021). Thus, the primary challenge of training a strong commonsense retriever is to appropriately construct positive pairs and hard negative pairs (Karpukhin et al., 2020; Xiong et al., 2021). To do this, we propose novel strategies to construct question-document pairs that can be used for training a strong commonsense retriever.

3.2.1 Positive Training Pairs

In open-domain document retrieval, it is often the case that positive training pairs are available explicitly. For example, DPR treated Wikipedia documents that contain the correct answer as positive documents (Karpukhin et al., 2020). However, such training pairs might not be applicable on commonsense reasoning tasks because the output (e.g., *True / False* in a binary question answering) is not supposed to be a sub-string of retrieved documents.

In order to train a strong commonsense dense retriever, we propose two novel strategies to construct positive training pairs, as described below.

Explanation as positive document. The first method for constructing positive training pairs is to take human annotated explanations as positive documents. For examples, taking the question *"Where do people go to pray? (A) church"* from CommonsenseQA1.0 as input, the explanation annotated in Aggarwal et al. (2021) is *"People go to a church to pray"*; similarly, a positive document for the question *"When food is reduced in the stomach, nutrients are being deconstructed"* in OpenBookQA (Mihaylov et al., 2018) could be *"Digestion is when stomach acid breaks down food"*. The explanations have two important properties. First, they contain commonsense knowledge, such as *people praying in church*, in the form of natural language. Second, they can be used to help select the correct choice in commonsense reasoning tasks. So, we take advantage of the high correlation of natural language explanations with the input query, defining the input query as q and the corresponding generated explanation as d to train the retriever.

Ground truth output as positive document. The second method for constructing positive training pairs is to directly use ground truth outputs in generation tasks as positive documents. The ground truth output can be seen as a natural positive document that the retriever should retrieve. For example, in the CommonGen (Lin et al., 2020) dataset, the ground truth output for an input concept set *{dance, kid, room}* is *"a group of kids are dancing around a living room"*. We define the input sequence in a generation task as q and its corresponding ground truth output as d to train the retriever. During training, the vector distance between them are minimized. During inference, though the ground truth documents are no longer in the commonsense corpus, the retriever can still fetch relevant documents similar to the ground truth output such as *"a couple of kids are dancing on the floor (ARC corpus)"*, which provides relevant contexts describing the potential reaction between the input concepts *"kid"* and *"dance"*, hence helps generate desired outputs.

3.2.2 Negative Training Pairs

For negative pairs, we adopt the trick of in-batch negatives, which has been shown as an effective strategy for learning a dual-encoder model and used in the many recent dense retrieval models (Lee

et al., 2019; Karpukhin et al., 2020).

3.3 Commonsense Document Reader

After retrieving commonsense documents, the reader takes the input text along with the retrieved documents to produce the desired output. Sequence classification tasks are considered as a target sequence of length one. In our work, we use the fusion-in-decoder (FiD) (Izacard and Grave, 2021) model as the reader. Specifically, each retrieved document is concatenated with the input text, then independently encoded by the T5 (Raffel et al., 2020) encoder. Then, the T5 decoder performs cross-attention over the concatenation of the resulting representations of all the retrieved documents.

4 Experiments

4.1 Tasks and Datasets

Multi-choice question answering. Give a question, an intelligent system is asked to select one correct answer from the choices offered as a list. We conducted experiments on CommonsenseQA1.0 (Talmor et al., 2019) and OpenBookQA (Clark et al., 2018). CommonsenseQA1.0 (CSQA1.0) contains 12,102 questions with one correct answer and four distractor answers. OpenBookQA (OBQA) consists of 5,957 elementary-level questions with one correct answer and three distractor answers. For evaluation, the primary metric on these two tasks is accuracy (ACC.).

Commonsense fact verification. Given a commonsense claim, an intelligent system is expected to verify the statement in natural text against facts. For example, the statement *"A pound of cotton has the same weight as a pound of steel"* in the CommonsenseQA2.0 (Talmor et al., 2021) should be identified as *true*. We conducted experiments on two commonsense fact verification datasets, including CommonsenseQA2.0 (Talmor et al., 2021) and CREAK (Onoe et al., 2021). CommonsenseQA2.0 was collected via gamification, which includes 14,343 assertions about everyday commonsense knowledge. CREAK is designed for commonsense reasoning about entity knowledge, which consists of 13,000 assertions about entities. For evaluation, the primary metric is accuracy (ACC.).

Constrained commonsense generation. Given a set of concepts such as *"dog, frisbee, catch, throw"*, the task is to generate a coherent sentence describing an everyday scenario such as *"a man throws a frisbee and his dog catches it"*. Our experiments were conducted on the benchmark dataset Commongen (Lin et al., 2020). It consists of 79,000 commonsense descriptions over 35,000 unique concept-sets. The average input / output length is 3.4 / 10.5 words. All examples in the dataset have 4-6 references. The task is evaluated by SPICE (Anderson et al., 2016), BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), in which SPICE is the primary metric for leaderboard ranking.

Counterfactual explanation generation. Given a counterfactual statement, the task is to generate reasons why the statement does not make sense. Our experiments were conducted on the benchmark dataset ComVE from SemEval-2020 Task 4 (Wang et al., 2020). It contains 11,997 examples. The average input/output length is 7.7 / 9.0 words. All ground truth have 3 references. The task is evaluated by SPICE (Anderson et al., 2016), BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), in which BLEU-4 is the primary metric for leaderboard ranking.

4.2 Baseline Methods

We compared our RACo with various kinds of baseline methods. In addition of comparing with pre-trained language models, such as BART (Lewis et al., 2020a) and T5 (Raffel et al., 2020), we also compared with three kinds of commonsense knowledge augmented methods as introduced below.

Commonsense-aware language models (CLM). They are trained with external commonsense corpus or datasets, in order to embed commonsense knowledge into their parameters. During fine-tuning, the language models make predictions without accessing to any external corpus. In the experiments, we compared our model with CALM (Zhou et al., 2021) and UNICORN (Lourie et al., 2021).

Knowledge graph reasoning models (KGM). KGs are incorporated into models for augmenting the limited information in the input texts. We compared our model with KagNet (Lin et al., 2019), GRF (Ji et al., 2020), KG-BART (Liu et al., 2021), QA-GNN (Yasunaga et al., 2021), MoKGE (Yu et al., 2022c) and GreaseLM (Zhang et al., 2022).

Retrieval augmented models (RAM). We compared with a recent retrieval-augmented method named KFCNet (Li et al., 2021) for constraint commonsense generation. In addition, we also compared with using sparse retriever such as BM25

Methods ↓	K-type	CommonGen					ComVE				
		BL-4	RG-L	MET	CIDEr	SPICE*	BL-4*	RG-L	MET	CIDEr	SPICE
BART-large	-	26.30	41.98	30.90	13.92	30.60	19.22	44.86	27.10	11.04	35.14
T5-large	-	28.60	42.97	30.10	14.96	31.60	22.77	51.83	26.66	11.42	34.62
CALM	CLM	29.50	-	31.90	15.61	33.20	23.50	52.56	27.41	11.87	35.23
UNICORN	CLM	39.86	44.56	34.52	17.26	30.20	24.46	52.75	27.88	12.40	35.79
KG-BART	KGR	30.90	44.54	32.40	16.83	32.70	-	-	-	-	-
GraphRF	KGR	-	-	-	-	-	22.07	44.32	25.96	11.62	33.09
MoKGE	KGR	-	-	-	-	-	22.87	52.03	27.01	11.75	34.88
KFCNet	RAM	41.97	46.13	36.22	17.39	33.11	-	-	-	-	-
BM25+FiD	RAM	42.17	47.95	35.57	18.74	33.16	24.39	52.76	28.26	12.67	35.28
RACO	RAM	**42.76**	**48.19**	**35.80**	**18.89**	**33.89**	**25.30**	**53.29**	**28.62**	**12.76**	**36.37**

Table 3: Compared with commonsense-aware language models (CLM) and knowledge graph reasoning models (KGR) counterparts, our retrieval-augmented commonsense reasoning (RACO) can outperform the baseline methods and achieved state-of-the-art performance on the CommonGen and ComVE benchmarks. *: primary metric.

Methods ↓	K-type	CSQA1.0 ACC.	OBQA ACC.
T5-large	-	70.14	66.02
UNICORN	CLM	71.60	70.02
KagNet	KGR	69.00	-
QA-GNN	KGR	73.40	67.80
GreaseLM	KGR	74.20	66.90
BM25+FiD	RAM	74.12	67.75
RACO	RAM	**75.76**	**71.25**

Table 4: RACo achieves better performance than other knowledge-enhanced method counterparts.

Methods ↓	K-type	CSQA2.0 ACC.	CREAK ACC.
T5-large	-	54.60	77.32
UNICORN	CLM	54.90	79.51
GreaseLM	KGR	-	77.51
BM25+FiD	RAM	58.75	83.03
RACO	RAM	**61.75**	**84.17**

Table 5: RACO outperforms the baseline methods and achieved state-of-the-art performance on the CREAK.

Retriever in RACO ↓	CSQA1.0		OBQA	
	Hit@5	Hit@10	Hit@5	Hit@10
BM25	46.33	50.93	50.21	60.24
DPR_{Wiki}	4.40	5.78	28.41	37.58
DPR_{RACO}	**61.71**	**65.68**	**75.23**	**85.41**

Table 6: Retrieval accuracy on dev sets, measured as the percentage of retrieved documents that contain the ground truth document, which is annotated in Mihaylov et al. (2018); Aggarwal et al. (2021). In addition, DPR_{Wiki} directly uses the DPR trained on Wikipedia for commonsense retrieval without any fine-tuning process. DPR_{RACO} trains the commonsense dense retrieval using our proposed training pairs construction strategy.

to retrieve knowledge from our constructed commonsense corpus and use FiD (Izacard and Grave, 2021) as generator to produce outputs.

4.3 Automatic Evaluation

4.3.1 RACO v.s. Baseline Methods

Comparison with non-retrieval methods. To observe the effectiveness of retrieval on commonsense reasoning tasks, we first compared model performance with and without commonsense retrieval. As shown in Table 3-5, compared with BART and T5 that directly encode the input query and produce output without using external knowledge, our proposed RACO can improve the commonsense reasoning performance by a large margin. Specifically, RACO improved BLEU-4 by +8.44% on the commonsense generation tasks, improved accuracy by +5.43% on the multiple choice question answering tasks, and improved accuracy by +6.15% on the commonsense verification tasks. Therefore, we concluded that RACO can leverage the retrieval of relevant references from commonsense corpora to help language models produce better outputs in various commonsense reasoning tasks.

Comparison with other knowledge-enhanced methods. As mentioned in the §4.2, the commonsense reasoning ability of a language model can be enhanced by fine-tuning on commonsense corpora or reasoning over multi-hop relations on knowledge graphs. As indicated by Table 3-5, compared with commonsense-aware language models (CLM), retrieval augmented model explicitly leverage relevant commonsense knowledge, demonstrating superior performance on all datasets. Compared with knowledge graph reasoning methods (KGR), it can achieve better performance on all six datasets.

Retriever training set ↓	CSQA1.0 ACC.	OBQA ACC.	CSQA2.0 ACC.	CREAK ACC.	CommonGen BL-4	CommonGen SPICE*	ComVE BL-4*	ComVE SPICE	Avg.
OBQA	71.09	69.55	57.57	81.47	33.70	33.83	27.55	38.29	58.67
CommonGen	73.36	66.34	57.49	83.15	**38.34**	**36.71**	27.71	37.79	59.46
CSQA1.0	74.84	**71.56**	59.76	82.85	35.29	35.05	27.28	35.05	60.50
All datasets	**75.08**	70.40	**60.21**	**84.01**	37.26	36.03	**28.05**	**38.17**	**60.80**

Table 7: Model performance (on dev sets) of using commonsense retrievals trained on different datasets. Training with question-document pairs from all datasets yield the best average performance on six tasks. *: primary metric.

Corpus size ↓	Corpus name HAF	CBD	CRC	Dataset name CSQA2.0	CREAK
0	(no retrieval)			53.80	77.32
3.56M	✓			56.57	78.91
2.88M		✓		56.65	78.93
13.6M			✓	56.86	82.82
6.44M	✓	✓		57.46	79.09
17.1M	✓		✓	58.16	82.52
16.5M		✓	✓	58.39	82.98
20.1M	✓	✓	✓	**59.66**	**83.85**

Table 8: Performance on dev sets of retrieving commonsense knowledge from different size of corpus.

4.3.2 Effects on Commonsense Retriever

To evaluate the effectiveness of commonsense retrieval, we compare the performance of different retriever training settings, including BM25, DPR_{Wiki}, and DPR_{RACo}. Specifically, DPR_{Wiki} directly uses the DPR trained on Wikipedia for commonsense retrieval without any fine-tuning process. DPR_{RACo} trains the commonsense dense retrieval using our our proposed training pairs construction strategy. As shown in Table 6, we can observe DPR_{Wiki} performs the worst among all retrievers. Our proposed DPR_{RACo} can significantly improve the retrieval performance, compared to BM25. It is important to note that the performance of retrieval is not necessarily linearly related to the performance of final output. However, in general, the more relevant the retrieved content, the more helpful it is to produce better outputs during the reading step. The observation can also be drawn from the comparison of BM25+FiD and RACo in Tables 3-5.

4.3.3 Effects on Multi-dataset Training

As shown in Table 7, we compare the model performance of retrievers trained by different set of question-document pairs. For example, the first line represents the retriever is trained with only question-document pairs (5,000 in total) from the OBQA dataset. The last line represents using question-document pairs from all six datasets. From the table, we can observe when the retriever is trained on only one dataset, it might not work well on other datasets because of differences in data distribution. Instead, training with multiple datasets demonstrates better generalizability.

4.3.4 Effects on Commonsense Corpus

To validate the effect of the number and content of corpora on our proposed method, we test the corresponding model performance under different corpora, including choosing a corpus, or any combination of multiple corpora. In Table 8, we show the performance of CSQA2.0 and CREAK on different commonsense corpora. It is worth noting that compared with other data, CSQA2.0 and CREAK can more realistically reflect the impact of different corpora on model performance, mainly because these two datasets are *not* based on any commonsense knowledge source during the collection process, so the coverage of the problem is much wider than other four datasets that are collects from a certain knowledge source. For example, CSQA1.0 and CommonGen are collected based on ConceptNet.

4.3.5 Effects on Number of Documents

We also compared model performance with different numbers of retrieved documents. As shown in Figure 2, as the number of retrieved documents increases, the model performance of RACO on the CommonGen dataset first increases and then remains unchanged on BLEU-4 or even decreases on SPICE (the primary metric on the CommonGen leaderboard), but the GPU memory consumption increases significantly. This is mainly because when the number of retrieved documents increases, more noisy information might be included in the model, which could hurt the performance of the model. Thus, with reasonable computational overhead, we only use 10 documents in our experiments.

4.4 Human Evaluation

We randomly sample 50 generated outputs from the CommonGen dev set (as the test set is not pub-

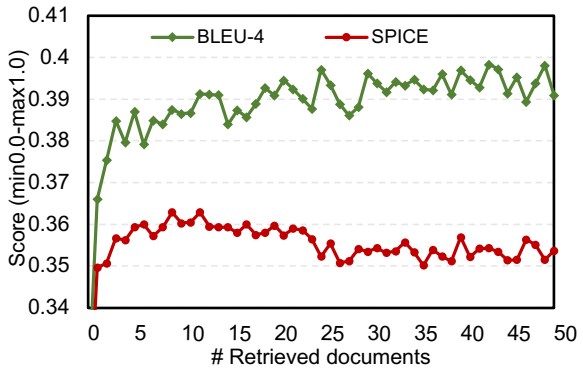

Figure 2: As the number of retrieved documents increases, the model performance of RACo on the CommonGen dataset first increases and then remains unchanged on BLEU-4 or even decreases on SPICE (the primary metric on the CommonGen leaderboard).

	CommonGen		ComVE	
	Fluency	Accuracy	Fluency	Accuracy
T5	2.86	2.78	2.94	2.74
CALM	2.85	2.81	2.93	2.78
UNICORN	2.88	2.84	**2.95**	2.78
RACo	**2.90**	**2.96***	2.94	**2.86***

Table 9: Human evaluations by independent scoring based on accuracy and fluency. * indicates p-value < 0.05 under paired t-test between RACo and baselines.

lic) and 50 generated outputs from the ComVE test set. All evaluations were conducted on Amazon Mechanical Turk (AMT), and each evaluation form was answered by three AMT workers. The generated outputs are evaluated by *fluency* and *accuracy*. Fluency is assessed on the grammatical correctness and readability of the generated outputs disregarding the input text. Besides, accuracy evaluates whether the output generated is correct and reasonable given the input text of each task.

As shown in Table 9, our model significantly outperforms baseline methods in terms of accuracy and fluency on both datasets. In particular, the accuracy of the generated output is greatly improved due to the incorporation of the retrieved commonsense knowledge. Furthermore, since all baseline models are pre-trained on large-scale corpora, they all produce outputs with great fluency. However, compared with baseline methods, the outputs generated by our model on the CommonGen dataset still have better fluency. This is mainly because the retrieved references are semantically complete sentences with good fluency, which might mitigate grammatical errors during the generation process.

1. CSQA2.0 Statement – A private college is usually smaller than a public university in attendance. (**True**) **Retrieved evidence – #1** Private schools are usually small and are worth the cost. **#2** University's are larger then most colleges. **#3** Colleges considered "small" have fewer than 5,000 students
Predictions – T5 and UNICORN: False **RACo:** True
2. ComVE Statement – The sun made my t-shirt wet. **Retrieved evidence – #1** The sun can dry wet clothes. **#2** The sun can dry something that is wet.
Generated outputs – T5: The sun is hot. **UNICORN:** The sun does not make clothes wet. **RACo:** The sun would dry a t-shirt but not make a t-shirt wet.
3. Commongen Input Words – eye look move **Retrieved evidence – #1** She moves her eyes around. **#2** The eye looks towards the peaks. **#3** A woman looks at the camera as she moves each eye individually. **#4** His eyes move across the paper.
Generated outputs – T5: Someone looks at someone, then moves his eyes. **UNICORN:** Someone looks at her and moves her eyes. **RACo:** A man moves his eyes to look at the camera.

Table 10: Case study. RACo corrects the erroneous predictions of baseline models (e.g., T5 and UNICORN) using the retrieved commonsense knowledge.

4.5 Case Study

Table 10 shows two examples with predictions from different models. We demonstrate a "comparison" statement from CSQA2.0 as the first example. As shown in the table, both T5 and UNICORN make wrong predictions, demonstrating a lack of commonsense knowledge. By leveraging the retrieved evidence from commonsense corpus, our proposed RACo can tell the statement "private college is usually smaller than a public university in attendance" is true. In addition, we show an example from counterfactual explanation generation task as the second example. Among the three outputs shown, the explanation generated by T5 is less associated with the input statement. Compared with the generated outputs from UNICORN, our model can generate a semantically richer and more reasonable explanation. This is mainly because the references retrieved provide strong evidence from the perspective of the sun dries things out.

5 Epilogue

Conclusions. Retrieval-augmented methods have been widely used in many NLP tasks such as open-

domain question answering. However, applying this approach to commonsense reasoning has been neglected in the existing literature. In this paper, we systematically investigate how to leverage commonsense knowledge retrieval to improve commonsense reasoning tasks. We collected a commonsense corpus containing over 20 million documents, and proposed novel strategies for training a commonsense retriever. Extensive experiments demonstrate our method can effectively improve the performance of various commonsense reasoning tasks, achieving new state-of-the-art performance on the CommonGen and CREAK leaderboards.

Future work. A natural extension of this work is to leverage heterogeneous knowledge to improve commonsense reasoning tasks, such as combining structured (i.e., knowledge graph) and unstructured (i.e., retrieved text) knowledge. Such a model will require building a graph reasoning module and a textual reasoning module, and merging the knowledge learned from both, which is a challenging task. The second future direction is to learn a commonsense dense retriever without question-document pairs. For example, in binary question answering, the labels are *True* / *False* that cannot be used to train a commonsense retriever.

Limitations. There are two main limitations. First, RACo retrieves documents from a large-scale corpus, then leverage the retrieved documents to make predictions. So, compared with baseline methods such as T5 and UNICORN, RACo consumes more time and computing resources. Second, due to the diversity and multi-source nature of commonsense knowledge, the retrieved evidence might contain noisy information that can even hurt model performance. A fine-grained filtering or re-ranking module could be a future work.

Acknowledgement

This work was supported in part by NSF IIS-1849816, IIS-2119531, IIS-2137396, IIS-2142827, CCF-1901059, and ONR N00014-22-1-2507.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference for Learning Representation (ICLR)*.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. Codah: An adversarially-authored question answering dataset for common sense. In *Third Workshop on Evaluating Vector Space Representations for NLP*.

Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. Adaprompt: Adaptive model training for prompt-based nlp. *arXiv preprint arXiv:2202.04824*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of North American Chapter of the Association for Computational Linguistics (NAACL)*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *arXiv preprint arXiv:2002.08909*.

Catherine Havasi, Robert Speer, Kenneth Arnold, Henry Lieberman, Jason Alonso, and Jesse Moeller. 2010. Open mind common sense: Crowd-sourcing for common sense. In *Workshops at the Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. 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)*.

Gautier Izacard and Édouard 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*.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736.

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.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference for Learning Representation (ICLR)*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Devang Kulshreshtha, Robert Belfer, Iulian Vlad Serban, and Siva Reddy. 2021. Back-training excels self-training at unsupervised domain adaptation of question generation and passage retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7064–7078.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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*.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

Minghan Li and Jimmy Lin. 2021. Encoder adaptation of dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2110.01599*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839.

Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William Cohen. 2021. Differentiable open-ended commonsense reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4611–4625.

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

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

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S Yu. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Conference on Artificial Intelligence (AAAI)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference for Learning Representation (ICLR)*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Christian M Meyer and Iryna Gurevych. 2012. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2021. Creak: A dataset for commonsense reasoning over entity knowledge. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research (JMLR)*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. 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)*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of ai through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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.

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 (SemEval-14)*.

Han Wang, Yang Liu, Chenguang Zhu, Linjun Shou, Ming Gong, Yichong Xu, and Michael Zeng. 2021. Retrieval enhanced model for commonsense generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3056–3062.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *International Conference for Learning Representation (ICLR)*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546.

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022a. Kg-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022b. A survey of knowledge-enhanced text generation. In *ACM Computing Survey (CSUR)*.

Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao, and Meng Jiang. 2022c. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models. In *International Conference on Learning Representations*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, Bill Yuchen Lin, and Xiang Ren. 2021. Pre-training text-to-text transformers for concept-centric common sense. In *International Conference for Learning Representation (ICLR)*.

Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Lin, Meng Jiang, and Wenhao Yu. 2022. Knowledge-augmented methods for natural language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 12–20.

A Appendix

A.1 Commonsense Retrieval Corpus

We use a combination of 19 commonsense datasets for our largest scale training data retrieval. The datasets include α-NLI (Bhagavatula et al., 2020), SWAG (Zellers et al., 2018), RACE (Lai et al., 2017), CODAH (Chen et al., 2019), CommonsenseQA1.0 (Talmor et al., 2019), CommonsenseQA2.0 (Talmor et al., 2021), WinoGrade (Sakaguchi et al., 2021), ARC (Clark et al., 2018), CREAK (Onoe et al., 2021), OBQA (Mihaylov et al., 2018), PhysicalIQA (Bisk et al., 2020), QASC (Khot et al., 2020), SocialIQA (Sap et al., 2019b), CosmosQA (Huang et al., 2019), MNLI (Williams et al., 2018), VATEX (Wang et al., 2019), Activity (Krishna et al., 2017), SNLI (Bowman et al., 2015) STSB (Cer et al., 2017).

A.2 Implementation Details

Retriever. We employed two independent pre-trained BERT-base models with 110M parameters (Devlin et al., 2019) as query and document encoders. BERT-base consists of 12 Transformer layers. For each layer, the hidden size is set to 768 and the number of attention head is set to 12. All dense retrievers were trained for 40 epochs with a learning rate of 1e-5. We used Adam (Kingma and Ba, 2015) as the optimizer, and set its hyperparameter ϵ to $1e$-8 and (β_1, β_2) to $(0.9, 0.999)$. The batch size is set as 32 on 8x32GB Tesla V100 GPUs.

Reader. We employed the FiD (Izacard and Grave, 2021) model that is built up on T5-large (Raffel et al., 2020). For model training, we used AdamW (Loshchilov and Hutter, 2019) with batch size 32 on 8x32GB Tesla V100 or A100 GPUs. We experimented with learning rates of 1e-5/3e-5/6e-5/1e-4 and we found that in general the model performed best when set to 3e-5. All reader models were trained with 20,000 steps in total where the learning rate was warmed up over the first 2,000 steps, and linear decay of learning rate.

A.3 Additional Related Work

Pre-training a language model on commonsense corpora is the most straightforward way to learn commonsense knowledge. Meanwhile, it also helps avoid overfitting when fine-tuned on downstream tasks. When the commonsense materials are appropriately selected, this simple strategy could demonstrate significantly superior performance than vanilla pre-trained language models (Zhou et al., 2021). Notable methods include COMET (Bosselut et al., 2019), CALM (Zhou et al., 2021), Unicorn (Lourie et al., 2021) and etc. For example, COMET initialized its parameters from GPT-2 and post-trained on ATOMIC to adapt its learned representations to knowledge generation, and produces novel knowledge tuples that are high quality (Bosselut et al., 2019). Unicorn initialized its parameters from T5 and performed a multi-task training on six commonsense question answering datasets (Lourie et al., 2021). While this development is exhilarating, such commonsense-aware language models still suffer from the following drawbacks: first, they are usually trained offline, rendering the model agnostic to the latest information, e.g., Covid-19 is a disease caused by a coronavirus discovered in 2019. Second, they make predictions by only "looking up information" stored in its parameters, leading to inferior interpretability (Shuster et al., 2021).

Incorporating knowledge graph (KG) is essential for many commonsense reasoning tasks to augment the limited textual information. A KG often provides comprehensive and rich entity features and relations so models can easily traverse links to discover how entities are interconnected to express certain commonsense knowledge. Some recent work explored using graph neural networks (GNN) to reason over multi-hop relational KG paths, yielding remarkable performance on many commonsense reasoning tasks, such as commonsense question answering (Lin et al., 2019; Yasunaga et al., 2021; Zhang et al., 2022), abductive reasoning (Ji et al., 2020; Yu et al., 2022c), and chit-chat dialogue systems (Zhou et al., 2018; Zhang et al., 2020). The most frequently used KG is ConceptNet. For example, Ji et al. (2020) enriched concept representations in the input text with neighbouring concepts on ConceptNet and performed dynamic multi-hop reasoning on multi-relational paths so the knowledge can be embedded into the hidden representations. Nevertheless, the type of commonsense knowledge is restricted by the relations defined in a knowledge graph schema. Meanwhile, commonsense knowledge lies at an unbounded set of facts and situations that usually cannot be covered by a single knowledge graph. Reasoning over multiple knowledge graph is a challenging task.

A.4 Case Study on CSQA2.0

Figure 3 demonstrates the accuracy of T5 and our RACo for different statement types on the

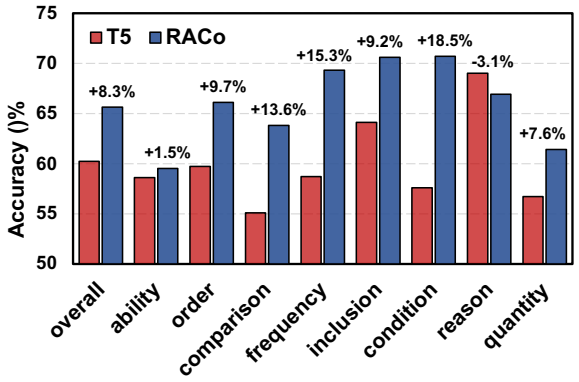

Figure 3: Performance of T5 and our RACo on different commonsense statement types in CSQA2.0.

CSQA2.0 dataset. First, compared to T5, our model can improve by 8.3% accuracy on all dev data (shown in the first column). However, on different statement types, the model performance is different. For example, from the predicted results of T5, the performance on "comparison" statements and "condition" statements is below-average. By introducing the retrieved commonsense knowledge, RACo demonstrated significantly better performance on these two sub-categories, achieving 15.3% and 18.5% improvement, which is significantly higher than the average 8.3% improvement. Nevertheless, we also observe the retrieved evidence might provide noisy information, resulting in performance degradation, such as "reason" related statements. We show an example in Table 10. Statements under these categories are often descriptions or comparisons of factual commonsense, the retrieved documents can thus well complement the necessary knowledge of a given statement. However, some statements require the model to reason in a given scenario, so making correct predictions requires the model to use commonsense knowledge to understand the local contexts. In these statements, retrieved knowledge might even contradict the assumptions, hurting the model performance.