COCONUT: Contextualized Commonsense Unified Transformers for Graph-Based Commonsense Augmentation of Language Models

Jun-Hyung Park[1] **Mingyu Lee**[2] **Junho Kim**[2] **SangKeun Lee**[2,3]
[1]BK21 FOUR R&E Center for Artificial Intelligence, Korea University
[2]Department of Artificial Intelligence, Korea University
[3]Department of Computer Science and Engineering, Korea University
{irish07, decon9201, monocrat, yalphy}@korea.ac.kr

Abstract

In this paper, we introduce COCONUT to effectively guide the contextualization of structured commonsense knowledge based on large language models. COCONUT employs a contextualized knowledge prompting scheme to gather high-quality contextualization examples from a large language model. These examples are subsequently distilled into small language models to enhance their contextualization capability. Extensive evaluations show that COCONUT considerably improves commonsense reasoning performance across diverse benchmarks, models, and settings, exhibiting its flexibility and universality in generating contextualized commonsense knowledge. Notably, COCONUT consistently outperforms the state-of-the-art technique by an average of 5.8%[1].

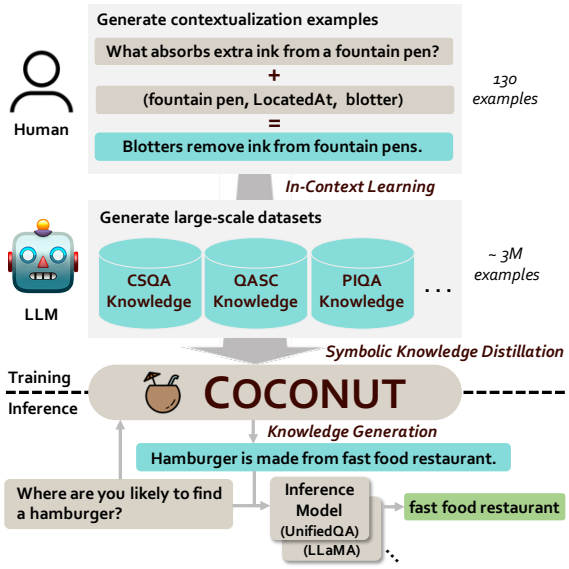

Figure 1: Illustration of COCONUT. COCONUT learns to contextualize structured knowledge within commonsense questions via contextualized knowledge prompting. During inference, COCONUT generates contextualized knowledge, which can be readily integrated by the concatenation with given questions.

1 Introduction

Commonsense reasoning constitutes a significant challenge within natural language processing. While scaling language models using considerably more data and parameters has led to significant improvements in commonsense reasoning tasks (Brown et al., 2020), several studies have demonstrated that pre-trained language models possess a limited understanding of commonsense knowledge (Sakaguchi et al., 2020; Talmor et al., 2022). These have triggered approaches to integrate external knowledge into language models to improve their commonsense reasoning abilities.

To enhance the commonsense capability of language models, typical approaches draw commonsense knowledge from symbolic commonsense knowledge graphs (CSKGs) (Speer et al., 2017; Hwang et al., 2020), which are repositories encapsulating hand-crafted commonsense knowledge about objects, concepts, and events. These approaches augment language model representations with the structural and relational information in CSKGs (Lin et al., 2019; Zhang et al., 2022). Even though there exists an open question regarding whether pre-trained language models already encode the knowledge in CSKGs, substantial research has indicated that these approaches (Zhou et al., 2021; Lourie et al., 2021; Yasunaga et al., 2021; Zhang et al., 2022) facilitate language models to utilize the knowledge, leading to improvements on commonsense reasoning performance.

These approaches essentially involve contextualization, which refers to the interpretation and application of knowledge tuples within the specific context provided by a commonsense reasoning example. Contextualization is a crucial but difficult step due to the diversity and obscurity of the underlying commonsense knowledge that grounds

[1]The code is available at https://github.com/irishev/Coconut

Findings of the Association for Computational Linguistics: ACL 2024, pages 5815–5830
August 11-16, 2024

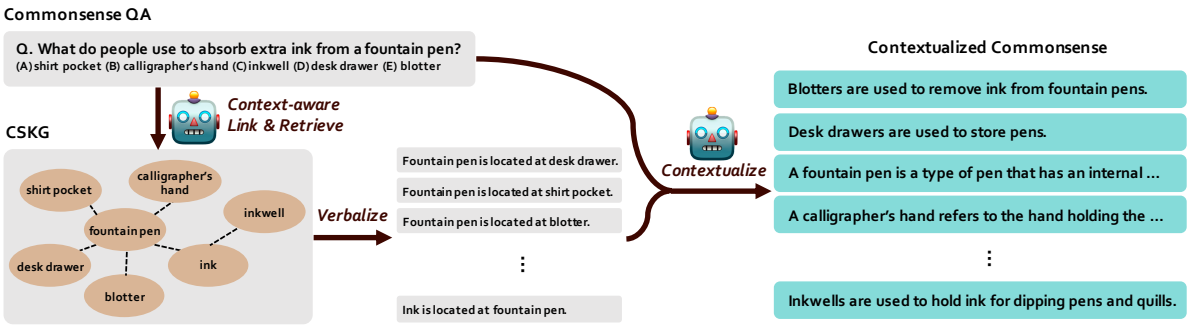

Figure 2: Illustration of contextualized knowledge prompting. Considering the given context, a large language model selectively links and retrieves relevant knowledge from CSKGs (§2.2), and then elaborates on the retrieved knowledge (§2.3). We generate the examples of these two contextualization processes, and subsequently distill the generated examples into COCONUT models.

the reasoning process (Liu et al., 2022a). Since most CSKGs provide simple, abstract descriptions of commonsense knowledge without specifying where and how the knowledge can be applied, existing methods largely lean on language models for contextualization. However, considering the limited commonsense capabilities of language models, this could lead to spurious contextualization, potentially degrading the commonsense reasoning performance. Particularly, small language models, known to have poor knowledge and reasoning ability to fill empty contexts, are expected to be more vulnerable to spurious contextualization.

In this paper, we propose a novel framework, called COCONUT (***CO**ntextualized **CO**mmonse**N**se **U**nified **T**ransformers*), which augments language models with contextualized commonsense knowledge. COCONUT contextualizes the structured knowledge in CSKGs for specific commonsense questions, trained by explicit, direct guidance from large language models. To overcome the lack of data and costly human annotations, we present a contextualized knowledge prompting scheme, where humans construct a few contextualization examples and then large language models extend the human-curated data into a million-scale via prompting. Following the scheme, we generate examples to guide the contextualization, and subsequently train COCONUT models on the generated data via symbolic knowledge distillation (West et al., 2022).

We extensively evaluate and analyze COCONUT with popular QA models on diverse commonsense reasoning benchmarks. Experimental results show that COCONUT delivers significant performance improvements on commonsense reasoning in both zero-shot and fine-tuned settings, demonstrating the efficacy of the proposed framework. Notably, COCONUT consistently outperforms state-of-the-art knowledge augmentation methods by an average of 5.8%. The main contributions of this work are summarized as follows:

- We propose COCONUT, a novel framework that augments models with contextualized commonsense knowledge from structured knowledge in CSKGs.
- We present a novel contextualized knowledge prompting scheme to generate contextualization examples from commonsense questions and CSKGs using a large language model.
- We demonstrate the efficacy of providing contextualized knowledge through extensive experiments.

2 Contextualized Knowledge Prompting

In this section, we present a contextualized knowledge prompting scheme, which generates contextualization examples using commonsense questions and CSKGs. Specifically, we prompt language models to generate examples of context-aware link and knowledge contextualization, as illustrated in Figure 2.

2.1 Notation

We first define ConceptNet, utilized as a CSKG in this work, as $G = (V, E)$, where V is the set of nodes and E is the set of edges. Its knowledge tuple $\{s, r, o\} \in E$ consists of a source concept $s \in V$, a relation type r, and a target concept $o \in V$. For example, given a knowledge tuple {food, LocatedAt, refrigerator}, the source concept

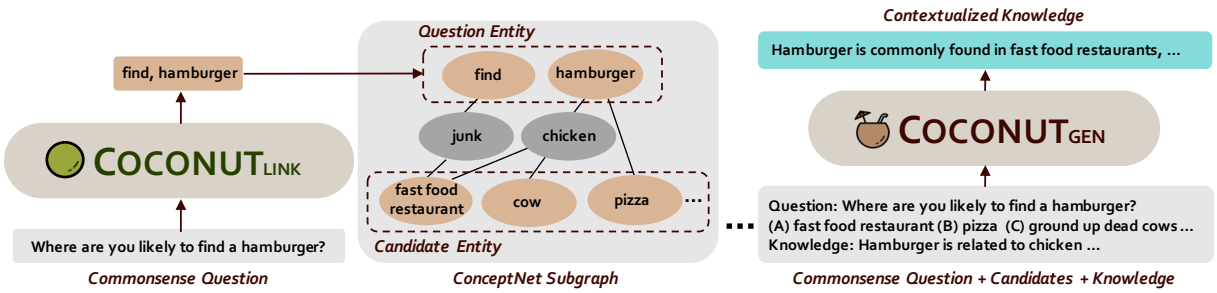

Figure 3: Illustration of generating contextualized knowledge using COCONUT linker and generator.

"food" is related to the target concept "refrigerator" by the relation "LocatedAt", which represents that food is typically located at a refrigerator. In addition, a multiple-choice commonsense question-answering task requires predicting the answer a given a question q_i and a finite set of candidates $C_i = \{c_i^0, c_i^1, ...c_i^N\}$, where N is the number of candidates and $a \in C_i$. For example, a CommonsenseQA example consists of the question "The accountant used a calculator regularly, he kept one at home and one at the what?", the set of candidates {"desk drawer", "desktop", "office", "wristwatch", "city hall"}, and the answer "office".

2.2 Generating Context-aware Link Examples via Prompting

Typical knowledge integration methods (Yasunaga et al., 2021, 2022) retrieve the knowledge by linking entities in a question and its candidates with the matching nodes in CSKGs and then selecting the knowledge tuples that connect an entity in the question with an entity in the candidates. However, numerous irrelevant entities may be linked during the process, resulting in excessive and redundant knowledge tuples to be retrieved.

To address the issue, we present a context-aware knowledge retrieval method to link entities highly relevant to the context. We first identify all entities that appear in a question q_i and each of its candidates c_i^j as follows:

$$Z_{q_i} = \{z | z \in \text{ngram}(q_i), z \in V\}$$
$$Z_{c_i^j} = \{z | z \in \text{ngram}(c_i^j), z \in V\}, \quad (1)$$

where ngram$(\cdot)$ is a function that extracts word n-grams from an input text. Then, we manually select relevant and helpful entities in q_i. Following this process, we craft ten examples per commonsense question-answering task, and write a prompt template as follows:

```
Find words or phrases relevant to the question.
Examples:

Question: How does getting paid feel?
Keywords: paid, feel

...

Question: Where can meat last a long time?
Keywords:
```

where the blue, green, and orange texts denote an instruction, a demonstration, and a new example, respectively. Given this prompt template filled with a new question, a large language model generates relevant entities for every question in a commonsense question-answering dataset as follows:

$$t_l \sim P_{lm}(t|T(q_i)) \quad (2)$$

where P_{lm} is the probability distribution of the language model prediction and $T(\cdot)$ is a function that fills the prompt template using a given input. The generated result t_l is a textual sequence of relevant entities "$x_{q_i}^1, x_{q_i}^2, ...$" separated by comma. By using the set of generated entities X_{q_i}, we can extract relevant knowledge tuples as follows:

$$K_i^j = \bigcup_{x \in X_{q_i}} \bigcup_{y \in Z_{c_i^j}} \text{path}(x, y, E) \quad (3)$$

where path(s, t, E) is a function that finds the set of knowledge tuples connecting s to t in the set of edges E. It is noteworthy that we generate link examples only for commonsense questions involving more than four words.

2.3 Generating Knowledge Contextualization Examples via Prompting

We generate contextualized knowledge statements from retrieved knowledge tuples using a large language model. We first construct few-shot examples and design templates to prompt effective contextualized knowledge. Our primary goal is to provide

more detailed information about how a knowledge statement from CSKGs can be connected to a given situation. To achieve this, we first verbalize the retrieved knowledge tuples using our templates and craft ten demonstration examples per task, each of which provides a description that contextualizes the knowledge. We utilize two prompt formats considering two possible commonsense generation scenarios: (1) commonsense fact generation and (2) contrastive commonsense generation.

Commonsense fact generation addresses reasoning over common facts such as "Birds have two wings." or "Winning lottery tickets give a lot of money.". Humans write contextualization examples from a question, its answer, and the retrieved knowledge. We convert the Winogrande (Sakaguchi et al., 2020), CommonsenseQA 2.0 (Talmor et al., 2022), Com2sense (Singh et al., 2021), ComVE (Wang et al., 2020), and GenericsKB (Bhakthavatsalam et al., 2020) datasets into the true or false examples and then utilize them as the seeds. An example prompt template used in commonsense fact generation is as follows:

```
Write new knowledge based on the given
knowledge to explain the correctness of a
question. Examples:

Question: The letter that Joel has written …
Answer: True
Given Knowledge: writer is related to write …
New Knowledge: Joel being the writer of the
letter is supported by the statement that …

…

Question: Sarah Jane's watch smashed when …
Answer: False
Given Knowledge: smash is related to activity …
New Knowledge:
```

Since the knowledge tuple extraction using the candidates of converted examples (true or false) may not provide meaningful relational knowledge, we consider the generated question entities X_{q_i} as the candidate entities and then extract knowledge tuples K_i. Given the prompt template filled with a new question q_i, its answer a, and the set of relevant knowledge tuples K_i, a large language model generates contextualized knowledge as follows:

$$t_c \sim P_{lm}(t|T(q_i, a, K_i)) \tag{4}$$

The generated result t_c is a contextualized knowledge statement.

Contrastive commonsense generation addresses reasoning by comparing the plausibilities of multiple candidates, such as "A rose garden provides a vast and continuous source of nectar to bees, while a bouquet of flowers is not a natural environment for bees". Humans write a reason why the answer candidate is more plausible than the other candidate based on the context and retrieved knowledge. We utilize the CommonsenseQA (Talmor et al., 2019), PhysicalIQA (Bisk et al., 2020), SocialIQA (Sap et al., 2019), OpenBookQA (Mihaylov et al., 2018), QASC (Khot et al., 2020), ARC (Bhakthavatsalam et al., 2021), and SyntheticQA (Wang et al., 2023) datasets. An example prompt template used in contrastive commonsense generation is as follows:

```
Write new knowledge based on the given
knowledge to explain the correct and wrong
options for a question. Examples:

Question: How does getting paid feel?
Correct Option: satisfaction
Wrong Option: bill collectors to happy
Given Knowledge: pride is related to
satisfaction. …
New Knowledge: Getting paid is intrinsically
linked to the feeling of satisfaction, …

…

Question: Where can meat last a long time?
Correct Option: freezer
Wrong Option: butcher shop
Given Knowledge: meat is located at freezer. …
New Knowledge:
```

Given this prompt template filled with a new question q_i, its answer c_i^m, one of its wrong candidate c_i^n, and the set of relevant knowledge tuples $J = K_i^m \cup K_i^n$, a large language model generates contextualized knowledge as follows:

$$t_c \sim P_{lm}(t|T(q_i, c_i^m, c_i^n, J)) \tag{5}$$

Note that the answer information is only used in the prompting stage to generate more accurate contextualization examples, and completely excluded in the training and inference with COCONUT models.

3 COCONUT

In this section, we introduce COCONUT, a framework designed to augment language models with contextualized commonsense knowledge. We present two COCONUT models: (1) COCONUT linker and (2) COCONUT generator. These two models generate contextualized knowledge from commonsense reasoning examples, as described in Figure 3. Then, we describe the integration of the contextualized knowledge generated by the COCONUT models.

3.1 COCONUT Linker

We train a COCONUT linker on the generated context-aware link examples in §2.2. Given a question q_i, the objective function for training can be formulated as:

$$\mathcal{L}_{link} = \mathbb{E}_{t_l \sim P_{lm}(t|T(\cdot))}[-\sum_{j=1}^{|t_l|} \log P_l(t_l^j | q_i, t_l^{<j})], \tag{6}$$

where P_l is the probability distribution of the COCONUT linker prediction. Using a trained COCONUT linker, we can generate relevant entities from a new question q_h without task-specific prompt templates:

$$\hat{t}_l \sim P_l(t|q_h) \tag{7}$$

From the set of relevant entities $\hat{X}_{q_h}$ in $\hat{t}_l$, we can extract relevant knowledge tuples as follows:

$$\hat{K}_h^j = \bigcup_{x \in \hat{X}_{q_h}} \bigcup_{y \in Z_{c_h^j}} \text{path}(x, y, E) \tag{8}$$

3.2 COCONUT Generator

We train a COCONUT generator on the generated knowledge contextualization examples in §2.3. Given a question q_i, its candidates C_i, and its relevant knowledge tuples K_i, the objective function for training can be formulated as:

$$\begin{aligned}\mathcal{L}_{gen} =& \mathbb{E}_{t_c \sim P_{lm}(t|T(\cdot))} \\ & [-\sum_{j=1}^{|t_c|} \log P_g(t_c^j | q_i, C_i, K_i, t_c^{<j})],\end{aligned} \tag{9}$$

where P_g is the probability distribution of the COCONUT generator prediction. Using a trained COCONUT generator, we can generate contextualized knowledge from a new question q_h, its candidates C_h, its set of relevant knowledge tuples K_h:

$$\hat{t}_c \sim P_g(t|q_h, C_h, K_h) \tag{10}$$

3.3 Contextualized Knowledge Integration

We prompt an inference model, i.e., a language model or a question-answering model, by concatenating each generated knowledge statement k_i^m to the question q_i and candidates C_i. We concatenate a question and its generated knowledge statement by following the default question-answering prompt format of the inference model. For example, UnifiedQA (Khashabi et al., 2020) uses a format that involves context, question, and candidate fields with symbols in order, while using "\n" as a delimiter. Therefore, the concatenation process $q_i \circ C_i \circ k_i^m =$ "{$\mathtt{q_i}$} \n {$\mathtt{k_i^m}$} \n (A) {$\mathtt{c_i^0}$} ...".

Following the pre-defined format, we calculate the probability of each candidate for each concatenated knowledge statement and average the probabilities to aggregate the scores. Given the generated knowledge set $\hat{K}_i = \{\hat{k}_i^0, \hat{k}_i^1, ..., \hat{k}_i^M\}$, the score of a candidate c_i^n is calculated as follows:

$$\text{score}(q_i, c_i^n, K_i) = \frac{\sum_j^M p_{inf}(c_i^n | q_i, k_i^j)}{M}. \tag{11}$$

where p_{inf} denotes an inference model. The final prediction $\hat{c}_i$ is the candidate that maximizes the score as follows:

$$\hat{c}_i = \arg\max_{x \in C_i} \text{score}(q_i, x, K_i). \tag{12}$$

4 Experiments

COCONUT establishes new state-of-the-art results on our evaluation benchmarks, significantly improving the commonsense reasoning performance of diverse inference models.

4.1 Experimental Setup

Datasets. Consistent with Liu et al. (2022a), we first evaluate the commonsense reasoning performance on eight seen datasets: OpenBookQA (Mihaylov et al., 2018), ARC easy/hard (Bhakthavatsalam et al., 2021), CommonsenseQA (Talmor et al., 2019), QASC (Khot et al., 2020), PhysicalIQA (Bisk et al., 2020), SocialIQA (Sap et al., 2019), and Winogrande (Sakaguchi et al., 2020). In addition, we evaluate the performance on four unseen commonsense reasoning datasets: NumerSense (Lin et al., 2020), RiddleSense (Lin et al., 2021), QuaRTz (Tafjord et al., 2019), and HellaSwag (Zellers et al., 2019). The official train, dev, and test splits of these benchmarks are employed for training and evaluation purposes.

Models. For prompting, we use LLaMA-65B (Touvron et al., 2023). We evaluate two combinations of the COCONUT linker and generator: Using T5-large and T5-3B as the generators, which are denoted as COCONUT-large and COCONUT-3B, respectively, while fixing the T5-small-based linker. For inference models, we mainly use UnifiedQA (Khashabi et al., 2020) and UnifiedQAv2 (Khashabi et al., 2022)). Note that we do not fine-tune inference models on downstream tasks unless mentioned otherwise.

Method	#Params	OBQA	ARC_e	ARC_h	CSQA	QASC	PIQA	SIQA	WNGR	Avg.
UnifiedQA-large	0.77B	69.8	68.1	55.2	61.4	43.1	63.4	52.9	53.3	58.7
+ Self-talk GPT-3 Curie	+ 13B	-	-	-	63.3	49.8	65.2	51.8	52.9	-
+ DREAM	+ 11B	-	-	-	64.5	49.4	64.7	51.5	56.1	-
+ GKP GPT-3 Curie	+ 13B	68.8	71.1	56.5	66.3	53.2	64.2	58.2	55.5	61.7
+ Rainier-large	+ 0.77B	69.6	67.7	55.2	67.2	54.9	65.6	57.0	56.9	61.8
+ Rainier-large + Vera	+ 6B	73.4	71.1	57.2	68.3	55.5	67.5	57.0	57.7	63.5
+ COCONUT-large (ours)	+ 0.83B	**75.2**	**75.8**	**61.5**	**74.8**	**67.0**	**74.6**	**67.3**	**57.9**	**69.3**
+ GKP GPT-3 Davinci	+ 175B	74.6	75.4	64.6	70.2	63.8	67.7	58.7	56.6	66.5
+ GKP GPT-3 Davinci + Vera	+ 180B	77.6	80.0	67.6	71.9	66.2	70.4	59.4	57.2	68.8
+ LLaMA-65B + ConceptNet	+ 65B	75.4	81.6	65.6	69.2	62.7	75.6	59.0	56.5	68.2
+ COCONUT-3B (ours)	+ 3B	**80.8**	**80.9**	**68.9**	**80.9**	**75.3**	**79.6**	**64.0**	**58.8**	**73.7**

Table 1: Comparison with knowledge prompting methods using UnifiedQA-large on **seen** datasets. "#Params" denotes the total number of parameters of used models and '+' denotes adding knowledge models and their number of parameters. We report the accuracy on the development set.

Method	#Params	NumerSense	RiddleSense	QuaRTz	HellaSwag	Avg.
UnifiedQA-large	0.77B	32.5	28.3	69.3	36.2	41.6
+ GKP GPT-3 Curie	+ 13B	38.0	35.7	69.0	37.3	45.0
+ Rainier-large	+ 0.77B	30.0	30.1	70.3	35.7	41.5
+ COCONUT-large (ours)	+ 0.83B	**41.5**	**36.1**	**72.9**	**39.6**	**47.5**
+ COCONUT-3B (ours)	+ 3B	**42.0**	**40.9**	**74.2**	**42.0**	**49.8**

Table 2: Comparison with knowledge prompting methods using UnifiedQA-large on **unseen** datasets. We report the accuracy on the development set.

Method	OBQA	ARC_h	CSQA	PIQA	SIQA
KagNet	-	-	69.0	-	-
CALM	60.9	-	71.3	75.7	69.2
Unicorn	-	-	72.6	82.2	75.5
FiD	67.8	-	74.1	-	-
RACo	71.3	-	75.8	-	-
QA-GNN	67.8	44.4	73.4	79.6	75.7
GreaseLM	66.9	44.7	74.2	79.6	75.5
Dragon	72.0	48.6	74.0	81.1	76.8
COCONUT (ours)	**76.3**	**61.3**	**76.7**	**82.3**	**76.9**

Table 3: Comparison with graph reasoning, commonsense aware, and retrieval augmented models. We use COCONUT-large as the knowledge model and fine-tuned UnifiedQAv2-large as the inference model. We report the accuracy on the development set.

Baselines. We compare COCONUT with diverse knowledge augmentation methods, categorized as:

- **Knowledge prompting methods** involve generated knowledge prompting (GKP) with GPT-3 (Liu et al., 2022b), Self-talk (Shwartz et al., 2020), DREAM (Gu et al., 2022), Rainier (Liu et al., 2022a), and Rainier with Vera (Liu et al., 2023) where knowledge descriptions are elicited from other language models. We further use LLaMA-65B prompted with knowledge tuples in ConceptNet as the baseline of augmented prompting of knowledge.[2]
- **Knowledge graph reasoning models** incorporate external CSKGs to enhance the limited information present in the input texts, such as KagNet (Lin et al., 2019), QA-GNN (Yasunaga et al., 2021), GreaseLM (Zhang et al., 2022), and Dragon (Yasunaga et al., 2022).
- **Commonsense aware language models** are trained using an external commonsense corpus or datasets to embed knowledge into their parameters, such as CALM (Zhou et al., 2021) and Unicorn (Lourie et al., 2021).
- **Retrieval augmented models** focus on retrieving relevant knowledge from commonsense corpora, such as RACo (Yu et al., 2022) and FiD (Izacard and Grave, 2021).

4.2 Main Results

We first compare COCONUT with state-of-the-art knowledge prompting methods using UnifiedQA-large (Khashabi et al., 2020). As shown in Table 1, COCONUT-large surpasses the best baseline,

[2]We use the input format consistent with that of $\text{COCONUT}_{\text{GEN}}$ (i.e., Question + Options + Verbalized ConceptNet Paths). We use contextualization demonstrations that are used in contextualized knowledge prompting while excluding answer information. We use the exact-match-based knowledge retrieval method.

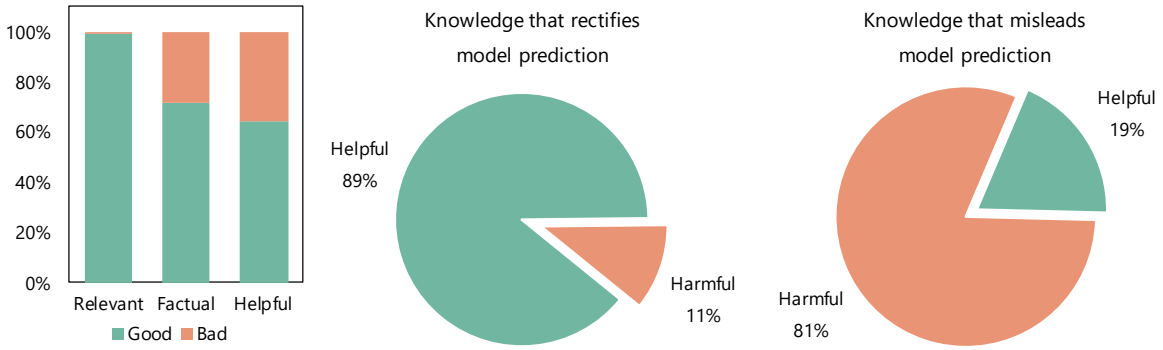

Figure 4: Human evaluation of generated knowledge. **Left:** Percentage of good knowledge statements along each axis. **Right:** Agreement between human and machine on the helpfulness of selected knowledge.

Example	A revolving door is convenient for two-direction travel, but it also serves as a security measure at a what? **(A) bank** (B) library (C) department store (D) mall (E) new york
COCONUT-large (ours)	Revolving doors provide security measures within banks, reducing the likelihood of theft or unauthorized entry. They don't serve as convenient for robbery prevention in libraries.
GPT-3-Davinci	A revolving door serves as a barrier.
Example	What do people aim to do at work? **(A) complete job** (B) learn from each other (C) kill animals (D) wear hats (E) talk to each other
COCONUT-large (ours)	People aim to complete jobs at work to fulfill their responsibilities and fulfill the demands of their jobs. Killing animals would be considered criminal behavior.
GPT-3-Davinci	People aim to do their job well.

Table 4: Comparison of knowledge generated by COCONUT-large and GPT-3 Davinci from Liu et al. (2022b).

Rainier-large + Vera, on seen datasets by an average of 5.8%. In addition, COCONUT-3B outperforms GPT-3 Davinci + Vera using 60 times more parameters by an average of 4.9%. We observe that directly prompting LLaMA-65B with ConceptNet knowledge tuples during inference is not ideal, since it shows performance worse than that of COCONUT-large and COCONUT-3B, while using significantly more parameters and computations. These results show COCONUT's capability to effectively and efficiently augment language models with commonsense knowledge by learning how to contextualize structured knowledge.

Moreover, as shown in Table 2, COCONUT consistently stands out on unseen datasets. Particularly, COCONUT-large excels over GPT-3 Curie using by an average of 2.5%, using significantly fewer parameters. The superior performance on unseen datasets shows the generalization capability of COCONUT, rooted in CSKGs with the general, widely applicable commonsense knowledge.

Our evaluation of COCONUT extends to UnifiedQAv2 (Khashabi et al., 2022). As shown in Table 3, COCONUT effectively augments UnifiedQAv2 fine-tuned without knowledge, outperforming all the knowledge graph reasoning models, commonsense aware language models, and retrieval augmented models. These results underscore COCONUT's advantage in providing contextualized knowledge that is readily integrated by a wide range of inference models.

4.3 Human Evaluation

We conduct a human evaluation on CommonsenseQA to study the quality of generated knowledge and the interpretability of its impact on task performance. We sample 1,200 knowledge statements generated by COCONUT-large (100 knowledge statements per each evaluation dataset) and evaluate in terms of relevance, factuality, and helpfulness. As shown in Figure 4, we can observe that most generated knowledge is related, factually correct, and helpful for the model's reasoning. Specifically, COCONUT achieves 99.2% relevance, 71.7% factuality, and 64.2% helpfulness. Among the rectifying knowledge, 89% are deemed helpful by humans, and among the misleading knowledge, 81% are deemed harmful. These results show that

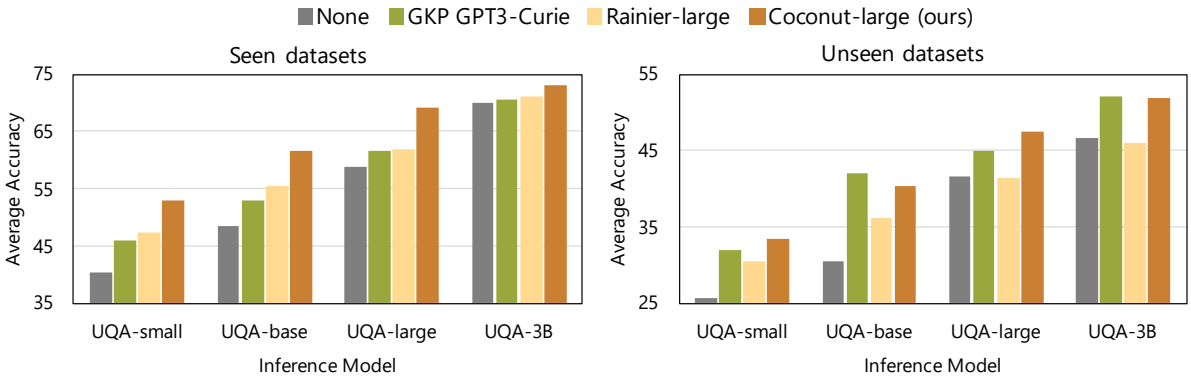

Figure 5: Results of COCONUT-large varying sizes of inference models.

Method	OBQA	ARC_e	ARC_h	CSQA	QASC	PIQA	SIQA	WNGR	Avg.
COCONUT-large	**75.2**	**75.8**	**61.5**	**74.8**	**67.0**	**74.6**	**67.3**	**57.9**	**69.3**
w/o $\text{COCONUT}_{\text{LINK}}$[3]	72.0	72.3	55.8	72.8	62.5	68.4	57.3	57.0	64.9
w/o $\text{COCONUT}_{\text{LINK}}$ & $\text{COCONUT}_{\text{GEN}}$	69.8	68.1	55.2	61.4	43.1	63.4	52.9	53.3	58.7

Table 5: Ablation study on **seen** datasets.

Method	NumerSense	RiddleSense	QuaRTz	HellaSwag	Avg.
COCONUT-large	**41.5**	**36.1**	**72.9**	**39.6**	**47.5**
w/o $\text{COCONUT}_{\text{LINK}}$[3]	36.3	33.2	70.3	36.7	44.1
w/o $\text{COCONUT}_{\text{LINK}}$ & $\text{COCONUT}_{\text{GEN}}$	32.5	28.3	69.3	36.2	41.6

Table 6: Ablation study on **unseen** datasets.

COCONUT effectively helps the inference models by generating relevant and accurate knowledge.

4.4 Analysis

Analysis on generated knowledge. Table 4 illustrates the knowledge descriptions generated by COCONUT and GPT-3-Davinci from Liu et al. (2022b) on examples from the CommonsenseQA validation set. From the generation results, we can observe that COCONUT generates more detailed knowledge descriptions about objects and their interactions. In contrast, GPT-3 leans towards providing descriptions of broad and general knowledge. Since inference models are possibly deficient in specific knowledge or reasoning processes required to infer answers from the provided knowledge, the lack of detailed descriptions can result in spurious contextualization, thereby degrading performance. Indeed, the UnifiedQA-large model finds the correct answer with knowledge generated by COCONUT, while failing with knowledge generated by GPT-3.

[3]We use the exact-match-based knowledge retrieval method.

Scaling trends of the inference model. We compare the commonsense reasoning performance of knowledge prompting methods with varying sizes of inference models. The results are presented in Figure 5. We observe two dominant trends from the results. Firstly, on both the seen and unseen datasets, the performance improvements from COCONUT-large are consistently better than the baseline knowledge prompting methods with a similar size. Secondly, when the inference models have a significantly larger size than that of COCONUT, we still observe performance improvements, showing that COCONUT may handle the knowledge absent in language models by integrating CSKGs.

Ablation Study. To better understand the contributions to performance improvements, we execute a series of ablation studies on COCONUT. As shown in Tables 5 and 6, we observe that both $\text{COCONUT}_{\text{GEN}}$ and $\text{COCONUT}_{\text{LINK}}$ contribute to the performance improvement. Particularly, $\text{COCONUT}_{\text{LINK}}$ achieves more significant improvements on the commonsense benchmarks with relatively long contexts, such as PIQA and SIQA, possibly due to its selection of important concepts.

5 Related Work

While large language models yield state-of-the-art performance on many commonsense reasoning tasks, their pre-training objectives do not explicitly guide them to reason using commonsense knowledge (Zhou et al., 2021), resulting in unsatisfactory performance in many real-world scenarios (Talmor et al., 2022; Zhu et al., 2023). To address the limitation, existing work has explored augmented language models to improve their commonsense reasoning ability. A typical approach is incorporating external knowledge from CSKGs, thereby supplementing the limited textual information (Lin et al., 2019; Yasunaga et al., 2021; Zhang et al., 2022; Yasunaga et al., 2022). Another approach involves training a language model on commonsense corpora (Lourie et al., 2021; Zhou et al., 2021). Recently, a line of research (Shwartz et al., 2020; Paranjape et al., 2021; Wei et al., 2022; Liu et al., 2022b) has proposed to generate knowledge by prompting language models due to the lack of scalability in utilizing CSKGs. Some recent methods have explored retrieving in-domain commonsense documents from a task-relevant corpus to improve commonsense reasoning capabilities (Wang et al., 2021; Li et al., 2021; Yu et al., 2022).

COCONUT provides two distinct advantages over existing commonsense augmentation methods. First, COCONUT alleviates the inherent limitation of CKSGs, the lack of coverage, by introducing language models in contextualization. Since language models have wide coverage and strong expressive power, they can effectively complement the coverage of CSKGs. Second, COCONUT utilizes CSKGs as pivots in knowledge generation, which facilitates to generate a wide range of accurate knowledge and achieves generalization on small language models.

6 Conclusion

In this paper, we have proposed COCONUT that contextualizes structured knowledge from CSKGs, guided by large language models. Our experimental results have verified that COCONUT outperforms state-of-the-art knowledge augmentation methods on diverse commonsense benchmarks. These show that large language models can explicitly guide the contextualization, leading to significant improvements in commonsense reasoning. Furthermore, our analyses suggest that prompting with structured knowledge may be a promising approach to address hallucination in knowledge prompting.

Limitations

While we have demonstrated that COCONUT effectively improves the commonsense reasoning performance by integrating contextualized commonsense knowledge from CSKGs, there are some limitations that present promising avenues for future research. First, COCONUT can generate unsafe or nonfactual knowledge, inheriting the limitation from language models. To address this, we plan to investigate the alignment of COCONUT with social, culture-specific, and ethical values. In addition, ConceptNet utilized by COCONUT involves limited types of knowledge, posing limitations in addressing domain-specific, eventual, or factual knowledge. Thus, we plan to extend the source of structured knowledge by introducing more diverse CSKGs, and to integrate retrieval augmentation from large knowledge corpora.

Acknowledgements

This work was supported by the Basic Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2021R1A2C3010430) and Institute of Information Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)).

References

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

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have solved direct-answer question answering? try arc-da, the direct-answer AI2 reasoning challenge. *CoRR*, abs/2102.03315.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yuling Gu, Bhavana Dalvi, and Peter Clark. 2022. DREAM: improving situational QA by first elaborating the situation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1115–1127. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. *CoRR*, abs/2010.05953.

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, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training. *CoRR*, abs/2202.12359.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1896–1907. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

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, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2918–2928. Association for Computational Linguistics.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 2829–2839. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6862–6868. Association for Computational Linguistics.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1504–1515. Association for Computational Linguistics.

Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. 2022a. Rainier: Reinforced knowledge introspector for commonsense question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8938–8958. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022b. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3154–3169. Association for Computational Linguistics.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1264–1287. Association for Computational Linguistics.

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 *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13480–13488. AAAI Press.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4179–4192. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *CoRR*, abs/1904.09728.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4615–4629. Association for Computational Linguistics.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-Lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 883–898. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. Quartz: An open-domain dataset of qualitative relationship questions. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5940–5945. Association for Computational Linguistics.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of AI through gamification. *CoRR*, abs/2201.05320.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 307–321. International Committee for Computational Linguistics.

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, Online Event,*

August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3056–3062. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13520–13545. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, November 16-20, 2020*, pages 38–45.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. In *NeurIPS*.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 535–546. Association for Computational Linguistics.

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 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4364–4377. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. 2021. Pre-training text-to-text transformers for concept-centric common sense. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chenguang Zhu, Yichong Xu, Xiang Ren, Bill Yuchen Lin, Meng Jiang, and Wenhao Yu. 2023. Knowledge-augmented methods for natural language processing. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM 2023, Singapore, 27 February 2023 - 3 March 2023*, pages 1228–1231. ACM.

A Technical Appendix

A.1 Additional Experimental Details

Implementation details. We implement COCONUT on PyTorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020). All our experiments are conducted on four NVIDIA RTX A6000 GPUs. We train the models in bfloat16 mixed-precision for efficiency. The training of COCONUT-large took 117 hours, while that of COCONUT-3B took 359 hours. The hyperparemeter settings are described in Table 7. The verbalization templates are shown in Table 8.

Datasets. Detailed information of datasets are shown in Table 9.

Human Evaluation. We asked three NLP experts to annotate the quality of generated knowledge. We randomly selected 1,200 examples, 100 examples per dataset. Each annotator evaluated the quality of knowledge statements along three axes: (1) Relevance: whether it is relevant to the situation in a question; (2) Factuality: whether it contains only correct statements; and (3) Helpfulness: whether it helps solve a question correctly. For each example, the annotator chose "good" or "bad". We do not reveal whether the knowledge rectifies or misleads the model prediction for objectivity.

A.2 Additional Results and Analyses

Extended experimental results. Full results on seen and unseen datasets are reported in Table 10 and 11, respectively.

Qualitative analysis. Tables 12 and 13 present examples showcasing both the strengths and weaknesses of contextualized knowledge prompting and COCONUT. In the first two examples of Table 12, the contextualized knowledge prompting effectively generates well-contextualized knowledge descriptions that suit the implicit intentions of the questions, while knowledge tuples barely provide the information to distinguish the answer and wrong options. Yet, some examples are not as favorable, as shown in the last two examples. In Table 13, we observe several undesirable cases from the generation results of COCONUT including a direct guidance to the answer, a knowledge description seemingly plausible but not relevant to the question, and nonsense descriptions.

Hyperparameter	Value
Contextualized Knowledge Prompting	
Maximum number of extracted knowledge tuples	5
Number of sampled texts	1
p in nucleus sampling	0.8
Maximum length	128
COCONUT *Training*	
Maximum input length	512
Maximum output length	128
Batch size	128
Training steps	200,000
Optimizer	Adam
Learning rate	2e-5
β_1	0.9
β_2	0.999
ϵ	1e-8
Warmup steps	1,000
Learning rate scheduling	Linear decay
Inference	
Maximum number of extracted knowledge tuples	128
Number of sampled texts	10
p in nucleus sampling	0.8
Maximum input length of COCONUT	512
Maximum output length of COCONUT	128
Maximum input length of inference models	512
Maximum output length of inference models	128

Table 7: Hyperparamter settings.

Relation	Verbalized
RelatedTo	_ is related to _
FormOf	_ is an inflected form of _
IsA	_ is a specific instance of _
PartOf	_ is a part of _
UsedFor	_ is used for _
NotUsedFor	_ is not used for _
CapableOf	_ can do _
NotCapableOf	_ cannot do _
AtLocation	_ is located at _
Causes	_ causes _
HasFirstSubevent	_ begins with _
HasLastSubevent	_ concludes with _
HasProperty	_ can be described as _
NotHasProperty	_ cannot be described as _
MotivatedByGoal	_ is a step toward accomplishing _
ObstructedBy	_ can be prevented by _
Desires	_ wants _
NotDesires	_ does not want _
Synonym	_ has a very similar meaning to _
Antonym	_ is opposite to _
DistinctFrom	_ is not _
DerivedFrom	_ appears within _
SymbolOf	_ symbolically represents _
MannerOf	_ a specific way to do _
LocatedNear	_ is found near _
HasContext	_ is used in the context of _
SimilarTo	_ is similar to _
EtymologicallyRelatedTo	_ has a common origin with _
EtymologicallyDerivedFrom	_ is derived from _
CausesDesire	_ makes someone want _
MadeOf	_ is made of _
Entails	_ happens with _
InstanceOf	_ is an example of _
HasA	_ belongs to _
HasSubevent	_ happens as a subevent of _
HasPrerequisite	_ is a dependency of _
CreatedBy	_ creates _
DefinedAs	_ is _
ReceivesAction	_ can be done to _

Table 8: ConceptNet verbalization templates.

Name	Train Ex.	Dev Ex.	Train Statements
Seen			
OpenBookQA	4957	500	14871
ARC easy	2251	570	6753
ARC hard	1119	299	3357
CommonsenseQA	9741	1221	38964
QASC	8134	926	56938
PIQA	16113	1838	16113
SocialIQA	33410	1954	66820
SynQA	486778	-	973556
Winogrande	40398	1267	80796
CommonsenseQA 2.0	9264	2541	9264
Com2sense	1608	782	1608
ComVE	20000	1994	20000
GenericsKB	1904144	-	1904144
Total	2537917	13892	3193184
Unseen			
NumerSense	0	200	0
RiddleSense	0	1021	0
QuaRTz	0	384	0
HellaSwag	0	10042	0
Total	0	11647	0

Table 9: Statistics of Datasets.

Method	#Params	OBQA	ARC_e	ARC_h	CSQA	QASC	PIQA	SIQA	WNGR	Avg.
UnifiedQA-small	0.06B	48.6	43.5	35.8	32.0	19.0	53.2	41.9	49.4	40.4
+ COCONUT-large	+ 0.83B	60.2	56.8	46.8	59.0	51.0	58.5	44.2	48.5	53.1
UnifiedQA-base	0.22B	60.2	53.9	44.8	45.3	25.3	58.5	47.8	52.1	48.5
+ COCONUT-large	+ 0.83B	70.6	66.8	51.2	70.8	58.6	67.5	54.4	53.3	61.7
UnifiedQA-large	0.77B	69.8	68.1	55.2	61.4	43.1	63.4	52.9	53.3	58.7
+ COCONUT-large	+ 0.83B	75.2	75.8	61.5	74.8	67.0	74.6	67.3	57.9	69.3
+ COCONUT-3B	+ 3B	80.8	80.9	68.9	80.9	75.3	79.6	64.0	58.8	73.7
UnifiedQA-3B	0.77B	79.0	77.9	70.2	71.7	62.1	75.7	60.7	63.3	70.1
+ COCONUT-large	+ 0.83B	78.8	78.1	64.5	77.1	70.1	77.9	72.1	65.7	73.0
+ COCONUT-3B	+ 3B	83.6	82.1	69.2	81.1	77.0	81.3	69.5	66.4	76.3
UnifiedQAv2-small	0.06B	46.4	44.6	38.1	39.6	27.2	61.2	55.2	57.5	46.2
+ COCONUT-large	+ 0.83B	55.6	53.5	42.8	48.2	41.3	61.3	59.2	57.4	52.4
UnifiedQAv2-base	0.22B	60.4	56.3	48.2	58.5	46.5	67.7	63.2	60.2	57.6
+ COCONUT-large	+ 0.83B	67.4	66.8	52.5	69.5	61.0	72.1	67.5	61.4	64.8
UnifiedQAv2-large	0.77B	69.8	69.1	61.5	71.7	59.6	75.6	71.0	74.9	69.2
+ COCONUT-large	+ 0.83B	76.2	74.9	61.1	76.2	68.9	77.6	75.4	75.0	73.1
+ COCONUT-3B	+ 3B	79.8	80.5	69.6	81.0	76.5	81.6	75.0	75.8	77.5
UnifiedQAv2-3B	3B	81.8	77.4	72.6	80.8	73.9	83.4	76.3	82.2	78.6
+ COCONUT-large	+ 0.83B	81.0	80.2	66.2	79.4	72.5	80.5	77.6	77.5	76.9
+ COCONUT-3B	+ 3B	84.8	82.1	73.2	82.6	78.2	83.6	77.5	83.6	80.7
LLaMA-7B	7B	27.8	73.0	33.9	58.2	51.5	78.9	46.9	51.9	57.5
+ COCONUT-3B	+ 3B	45.8	79.3	53.2	70.7	59.4	79.1	57.3	56.5	64.6
LLaMA-13B	13B	29.8	78.2	43.1	60.0	56.6	79.3	47.5	51.9	59.1
+ COCONUT-3B	+ 3B	62.9	80.2	65.3	71.9	61.2	79.3	57.7	57.0	65.4

Table 10: Full experimental results on seen datasets.

Method	#Params	RiddleSense	NumerSense	QuaRTz	HellaSwag	Avg.
UnifiedQA-small	0.06B	16.8	4.0	54.2	27.4	25.6
+ COCONUT-large	+ 0.83B	26.8	21.5	57.0	28.8	33.5
UnifiedQA-base	0.22B	20.0	15.5	57.0	29.6	30.5
+ COCONUT-large	+ 0.83B	30.2	37.0	61.2	32.9	40.3
UnifiedQA-large	0.77B	32.5	28.3	69.3	36.2	41.6
+ COCONUT-large	+ 0.83B	36.1	41.5	72.9	39.6	47.5
+ COCONUT-3B	+ 3B	40.9	42.0	74.2	42.0	49.8
UnifiedQA-3B	3B	30.5	38.3	69.5	48.2	46.6
+ COCONUT-large	+ 0.83B	38.4	46.0	73.7	49.5	51.9
+ COCONUT-3B	+ 3B	45.4	48.0	71.9	51.8	54.3
UnifiedQAv2-small	0.06B	26.2	7.0	56.2	25.7	28.8
+ COCONUT-large	+ 0.77B	26.7	8.0	58.6	26.3	29.9
UnifiedQAv2-base	0.22B	32.3	27.5	66.4	28.7	38.7
+ COCONUT-large	+ 0.77B	36.3	32.0	65.9	30.6	41.2
UnifiedQAv2-large	0.77B	32.0	37.6	72.7	40.0	45.6
+ COCONUT-large	+ 0.77B	38.7	46.0	74.5	41.7	50.2
+ COCONUT-3B	+ 3B	43.0	45.0	73.4	43.8	51.3
UnifiedQAv2-3B	3B	40.5	51.2	74.0	56.7	55.6
+ COCONUT-large	+ 0.77B	42.0	46.0	75.0	56.1	54.8
+ COCONUT-3B	+ 3B	48.3	49.5	72.9	59.5	57.6

Table 11: Full experimental results on unseen datasets.

Question / Knowledge	Related Knowledge Tuples	Plausible
Sammy wanted to go to where the people were. Where might he go? (A) race track **(B) populated areas** (C) the desert (D) apartment (E) roadblock → Populated areas suggest many people are present, suggesting Sammy would likely seek these locations to interact and socialize with others. An apartment generally houses fewer people compared to larger, public spaces, so this location would likely not satisfy Sammy's intent.	people - AtLocation - populated areas people - AtLocation - apartment	✓
The fox walked from the city into the forest, what was it looking for? (A) pretty flowers (B) hen house **(C) natural habitat** (D)storybook (E) dense forest → Foxes are native forest animals and require natural habitats to survive and thrive. A hen house represents a man-made artificial structure that isn't conducive to supporting the wild nature of foxes.	fox - AtLocation - natural habitat fox - AtLocation - hen house	✓
Where do you put your grapes just before checking out? (A) mouth **(B) grocery cart** (C) super market (D) fruit basket (E) fruit market → Grocery carts serve as temporary holding places for items during shopping trips, providing convenience and mobility. Grapes placed in a grocery cart are ready for purchase and transportation, as opposed to those in the mouth, which are part of a snack or meal.	grape - AtLocation - mouth grape - AtLocation - grocery cart	✗
Seeds exhibit little or no dormancy, losing their ability to germinate after about five years. **(A) True** (B) False → Seeds generally have varying degrees of dormancy, often exhibiting longer lifespans exceeding five years without compromising their germination capabilities.	seed - RelatedTo - germinate. seed - RelatedTo - grow.	✗

Table 12: Qualitative analysis on generated examples by contextualized knowledge prompting.

Task	Question / Knowledge	Plausible
CSQA	A revolving door is convenient for two direction travel, but it also serves as a security measure at a what? **(A) bank** (B) library (C) department store (D) mall (E) new york → A revolving door is a door that can be turned to open and close. It is also a security measure at banks and other financial institutions.	✓
QASC	Climate is generally described in terms of what? (A) sand (B) occurs over a wide range (C) forests (D) Global warming ... **(G) local weather conditions** → Climate is defined in terms of local weather conditions observed over a prolonged period, considering factors like temperature, precipitation, humidity, atmospheric pressure, and seasonal variability. Desert is primarily a geological term linked to arid landscapes, rather than a meteorological descriptor used to define or quantify climate patterns.	✓
PIQA	Make outdoor pillow. (A) Blow into tin can and tie with rubber band. **(B) Blow into trash bag and tie with rubber band.** → You have to blow a large amount of air into the pillow. You can then tie it with a rubber band. You can't blow into a tin can.	✓
SIQA	Jordan was in charge of taking the food on the camping trip and left all the food at home. ... **(A) horrible that he let his friends down on the camping trip** (B) happy that he doesn't ... → They are feeling horrible. This is because they have ruined the camping trip. When someone leaves home and does not cook, they are generally feeling happy or satisfied.	✗
WG	Sarah was a much better surgeon than Maria so _ always got the easier cases. (A) Sarah **(B) Maria** → Maria is the sister of Sarah. Sarah is also the sister of Maria. She was the sister of Maria's mother. Maria was a doctor who worked with Maria. Maria is a doctor who worked with Sarah.	✗

Table 13: Qualitative analysis on generated examples by COCONUT.