LEROS: Learning Explicit Reasoning on Synthesized Data for Commonsense Question Answering

Chenhao Wang[1,2], **Pengfei Cao**[1], **Jiachun Li**[1,2], **Yubo Chen**[1,2], **Kang Liu**[1,2,3,*], **Xiaojian Jiang**[4], **Jiexin Xu**[4], **Li Qiuxia**[4], **Jun Zhao**[1,2]
[1]The Laboratory of Cognition and Decision Intelligence for Complex Systems, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Shanghai Artificial Intelligence Laboratory
[4]China Merchants Bank
{chenhao.wang, pengfei.cao, jiachun.li, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn
{jiangxiaojian, jiexinx, annielqx}@cmbchina.com

Abstract

Recent work shows large language models can be prompted to generate useful rationales for commonsense question answering (CQA), which can improve the performance of both themselves and other models. However, the cost of deployment and further tuning is relatively expensive for the large models. Some work explores to distill the the rationale-generation ability to convenient small-sized models, yet it typically requires human-authored QA instances during the distillation. In this paper, we propose a novel framework that leverages both knowledge graphs and large language models to synthesize rationale-augmented CQA data. Based on it, we train LEROS, a model that can generate helpful rationales to assist generic QA models to accomplish unseen CQA tasks. Empirical results demonstrate LEROS can substantially enhance the performance of QA models on five unseen CQA benchmarks, providing better gains than both same-sized counterpart models trained with downstream data and 10x larger language models. Our work reveals a novel way to integrate knowledge from both knowledge graphs and large language models into smaller models. The codes and synthesized resources are publicly available at https://github.com/wchrepo/leros.

Keywords: Commonsense Question Answering, Commonsense Knowledge, Rationale Generation

1. Introduction

Commonsense question answering (CQA) is a challenging natural language processing task. It requires the understanding of questions based on unstated background knowledge. In comparison with directly predicting the answers, previous work shows that adding useful rationales (e.g. relevant knowledge and reasoning details) beforehand can lead to better performance and interpretability (Shwartz et al., 2020; Liu et al., 2022b), which forms a Question-Rationale-Answer paradigm dubbed as *explicit reasoning*. For example, as shown in Figure 1, for the question "`What can owls do`", adding a rationale such as "`Owls are birds. Birds can fly.`" can help the model predict the answer.

However, obtaining high-quality rationales is non-trivial. Previous work (Mitra et al., 2019; Chen et al., 2020; Xu et al., 2022) attempts to extract knowledge from commonsense knowledge graphs (CKG) (Speer et al., 2017; Hwang et al., 2021) and other sources, which is limited by the coverage and retrieval availability of the knowledge sources. Some other work explores to use neural models to generate rationales on-the-fly (Rajani et al., 2019; Shwartz et al., 2020; Bansal et al., 2022). Especially, recent work elicits large language models (LLM) to generate "Chain-of-Thoughts" (Wei et al., 2022), which can not only boost their own QA performance, but also provide transferable rationales for assisting other models (Liu et al., 2022b; Saha et al., 2023). However, such ability only emerges on models with large sizes (typically >10B parameters), which are **expensive to deploy and inconvenient to further tune when needed** (e.g. optimizing for special use). Therefore, some work develops more convenient and controllable small-sized models by distilling the rationale-generating ability from LLMs (Liu et al., 2022a; Wang et al., 2022b; Li et al., 2023), yet such work **relies on expensive human-authored QA instances for distillation**.

To address the limitations of the above work, we propose a novel framework that enables small models to learn explicit reasoning on synthesized data. As our work relies solely on synthesized data, it can **(1)** avoid the use of expensive human-authored QA instances, **(2)** show generalization performance on CQA benchmarks in zero-shot setting, and **(3)** provide a strong start point for further tuning. To achieve that, we take the best of both commonsense knowledge graphs and large language models to synthesize rationale-augmented QA instances, and train a rational-generation model,

[*]Corresponding author.

LREC-COLING 2024, pages 10303–10315

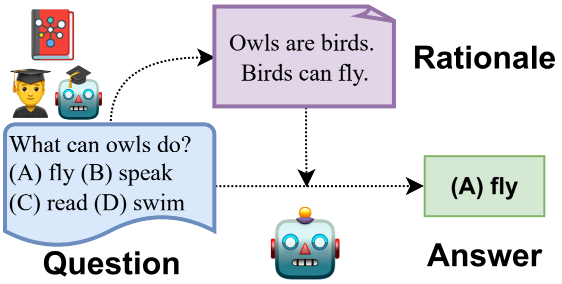

Figure 1: The illustration of explicit reasoning. The rationales are obtained from a source. Then the answers are predicted based on the rationales.

LEROS. Specifically, to ensure the quality of synthesized instances, we propose model-feedback-based prompting and refining strategies to obtain instances with high consistency and helpfulness. To make LEROS ready for providing on-demand rationales for given questions, we propose a two-stage training process, ensuring LEROS learn both adequate knowledge from CKGs and the generalized rationalization ability from LLMs. Finally, the trained LEROS can generate helpful and readable commonsense rationales, assisting a generic QA model to accomplish unseen CQA tasks. When specialized data are available, LEROS can be further tuned to generate better rationales.

We summarize our contribution as follows.

- We propose a novel framework to synthesize rationale-augmented CQA data and train small rationale-generation models. It combines the strengths of both CKGs and LLMs, avoiding the use of human-authored QA data.

- With solely the synthesized data, we train LEROS, a model capable of generating helpful and readable rationales for unseen commonsense questions, using only 0.7B parameters (approximately 0.5% of GPT-3 175B).

- Experiment results show LEROS can substantially improve the performance of QA models on five unseen CQA benchmarks. **(1)** Trained with synthesized data at an API cost of $\sim$\$100, it can directly bring more average performance gains than 10x larger language models and previous rationale models that are trained with human-authored CQA data. **(2)** When feedback of downstream QA benchmarks and further tuning are available, LEROS shows even better improvement. **(3)** The rationales generated by LEROS are useful for different QA models, including the models that are not used during training LEROS, such as LLaMA2-7B.

2. Related Work

2.1. Exploiting Knowledge for Commonsense Question Answering

Incorporating knowledge is a common practice in CQA tasks. Most previous work (Lin et al., 2019; Feng et al., 2020; Yasunaga et al., 2021; Guan et al., 2022) exploits knowledge from commonsense knowledge graphs, which have limited coverage and require well-designed retrieval heuristics. To directly acquire relevant knowledge given the questions, some work trains rationale generation models using human-annotated QA rationales (Rajani et al., 2019; Jhamtani and Clark, 2020; Aggarwal et al., 2021), commonsense knowledge graphs (Wang et al., 2020) or encyclopedia corpora (Bansal et al., 2022). Recently, large language models become another competitive source to generate rationales (Liu et al., 2022b). Some work further trains flexible smaller rationale generation models (Liu et al., 2022a; Wang et al., 2022b; Li et al., 2023) through distilling LLMs. Our methods pursue a similar target but avoid using the human-authored training data of target benchmarks.

Zero-shot commonsense question answering is a closely related field, which focuses on the inference and pretraining approaches to improve the generalized performance on unseen CQA benchmarks without the supervision of corresponding training data (Shwartz et al., 2020; Bosselut et al., 2021; Dou and Peng, 2022; Li et al., 2022). An effective way is to utilize the commonsense knowledge to synthesize QA instances and train a zero-shot QA model (Ma et al., 2021; Zhang et al., 2022; Kim et al., 2022; Wang et al., 2023a). These methods focus on optimizing a QA model to directly rank QA pairs, while our work improves the zero-shot QA performance via building the rationale generation models, which can provide rationales that are readable and usable for different QA models.

2.2. Synthesizing Data via Eliciting Large Language Models

The knowledge and ability of Large language models can be elicited with appropriate prompts (Petroni et al., 2019; Sung et al., 2021; Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023b). The advances in this line also open new doors to data synthesizing. There have been efforts to synthesize various language resources by prompting LLMs, including both symbolic commonsense knowledge (West et al., 2022; Wang et al., 2022a) and wide-range training data (Sclar et al., 2022; Wang et al., 2022c; Honovich et al., 2022). Such resources can be used for constructing smaller models for special purposes. In this work, we utilize synthesized commonsense questions to distill reasoning rationales from large language models, which are valuable resources for enhancing the generalized reasoning ability of smaller models.

3. Methods

3.1. Overview

For clarity, we let an instance of multiple-choice CQA be a textual pair (q, a^*), where q includes the question stem $\hat{q}$ and a set of answer choices A, and $a^* \in A$ is the correct answer. We assume there is a generic QA model[1] M_{QA} that can predict the probability of each answer $p(a|q), a \in A$. When the rationales k are provided and concatenated with q, the QA model M_{QA} may yield a different probability prediction, denoted as $p(a|q \circ k)$. Leaving M_{QA} unchanged, our methods train a model M_{QK} to generate helpful k for a given q, optimizing the probability of the correct answer $p(a^*|q \circ k)$. In this paper, we assume the zero-shot setup. The models have no access to the training data of target benchmarks. Instead, we synthesize a set of instances $\mathcal{D} = (q_i, a_i^*, k_i)_{i=1}^{|\mathcal{D}|}$ for training model M_{QK}, and directly evaluate it with M_{QA} on the target benchmarks. The overall framework is shown in Figure 2, including three parts: data preparation, model training, and inference.

3.2. Data Preparation

The first step is to prepare synthetic (q, a^*, k) instances. To take the best of CKGs and LLMs, we decompose the procedure into synthesizing CQA instances (q, a^*), augmenting rationales k, and refining high-quality instances (upper left in Figure 2).

Synthesizing QA Instances Inspired by previous zero-shot CQA research based on synthetic data (Ma et al., 2021), we utilize the knowledge from CKGs to synthesize CQA instances. Specifically, we sample knowledge triples of (h, r, t) format, e.g. `(owls, capableOf, fly)`. According to the type of relation r, we use verbalizing templates to convert the (h, r) into a question stem, e.g. `"What can owls do?"`, and take t as the correct answer a^*. We sample several other triples (h_i', r, t_i') that share the same relation r with (h, r, t) but have different heads and tails, and take t_i' as a distractor for the question. We then concatenate the question stems and shuffled choices to obtain q, such as `"What can owls do? (A) fly (B) speak"`.

Augmenting Rationales For each synthesized QA instance, a primitive way to provide rationale is to verbalize the source CKG triple (h, r, t) into a textual statement k_{source}, such as `"Owls can fly"`. Since it simply retells the source knowledge for the question, training with it can make the model gulp down knowledge from CKGs, but that is not adequate for generalization. Therefore, we query an LLM (e.g. GPT-3.5) to obtain additional rationales. Specifically, as shown in Figure 3, we start with a small set of seed examples in the format of $q \circ k \circ a$, and prompt the LLM to complete $k' \circ a'$ for a synthesized question q'. It can generate rationales that are relevant but not identical to k_{source}, e.g. `"Owls are birds"`, which could be useful for learning generalized reasoning. Such rationales generated by LLMs are denoted as k_{llm}. Also, we let the generated answer be a_{llm}, and allow the model to answer `"None"` if there is no proper choice.

Consistency and Helpfulness Refining So far, we have synthesized rationale-augmented QA instances. However, there may be errors in the CKGs and the responses of LLMs, which result in flawed instances. Therefore, we introduce two refining strategies based on the feedback of models.

First, in the previous step, we deliberately make the LLM generate both the rationale k_{llm} and the answer a_{llm}. We assume that k_{llm} is usable only when a_{llm} is equal to the correct answer a^*. We will remove the instance if $a_{llm} = None$ or $a_{llm} \neq a^*$, because it indicates either the instance is flawed or the LLM is unable to handle the question. We call this strategy **consistency refining**.

Second, even if the LLM gives the correct answer, the generated rationale k_{llm} could be irrelevant and unhelpful. Hence, we use the generic QA model M_{QA} to predict $p(a|q)$ and $p(a|q \circ k)$, which are used for calculating the feedback score of helpfulness. Specifically, inspired by the knowledge reward of Liu et al. (2022a), we define the helpfulness score $S \in (-1, 1)$ as:

$$S(k|q, a^*, A) = \frac{1}{2} \left[\tanh\left(\log p(a^*|q \circ k) - \max_{\substack{a' \in A \\ a' \neq a^*}} \log p(a'|q \circ k) \right) - \tanh\left(\log p(a^*|q) - \max_{\substack{a' \in A \\ a' \neq a^*}} \log p(a'|q) \right) \right] \tag{1}$$

where larger $S > 0$ indicates that the rationale can help the QA model favour the correct answer over the distractors more. Let S_0 be a threshold, we reserve the instances when $S > S_0$. We call this strategy **helpfulness refining**.

In order to improve the generation quality, the refining is conducted for each LLM-querying round, and the results are randomly added back to the prompt examples in subsequent rounds. The strategy is denoted as **refining prompting**.

[1] In terms of implementation, M_{QA} can be QA models trained on generic QA data or large language models prompted with in-context learning, which are not specialized on target CQA benchmarks.

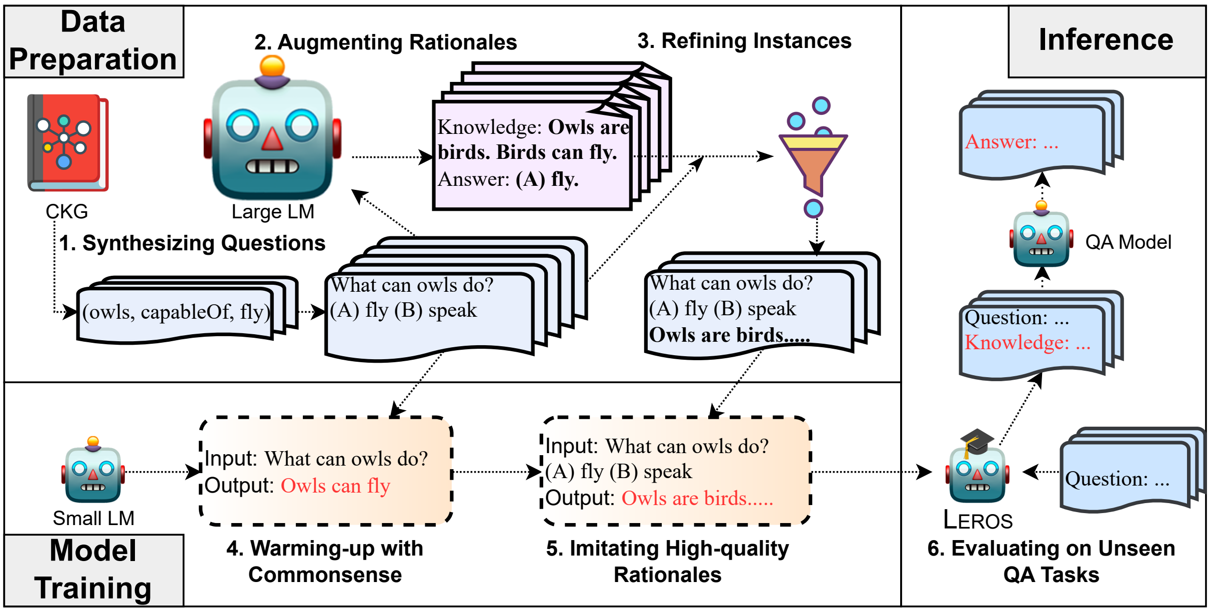

Figure 2: Overview of the proposed framework. **Upper Left:** Synthesizing questions and rationales. **Lower Left:** Training the LEROS model. **Right:** Zero-shot inference on target CQA benchemarks.

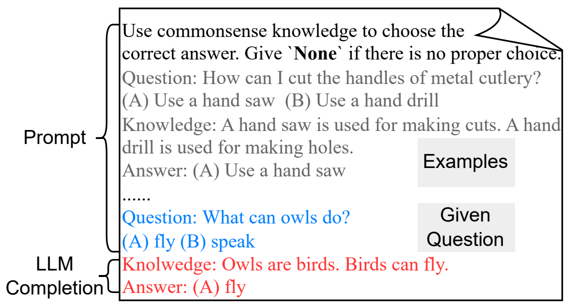

Figure 3: The illustration of prompting a LLM to complete rationales and answers.

3.3. Training LEROS

After data preparation, we start to train LEROS based on a pretrained sequence-to-sequence language model. The loss function is defined as:

$$L(\theta) = \frac{1}{|y|} \sum_{t=1}^{|y|} - \log p_\theta(y_t | x, y_{<t}) \tag{2}$$

where x is the input sequence and y is the target output sequence. During the two training stages (lower left in Figure 2), they are defined differently.

Stage 1: Warming-up with Commonsense To make the model internalize abundant commonsense knowledge from CKGs, we train the model to generate the source knowledge given a synthetic question stem, i.e. $x = \hat{q}, y = k_{source}$. For example, given `"What can owls do?"` as the input, the model is trained to predict `"owls can fly"`. This objective is similar to the generative commonsense knowledge completion task (Bosselut et al., 2019), which predicts t given (h, r).

Stage 2: Imitating High-quality Rationales To make the model generate useful rationales for questions, we further train the model on the refined question-rationale instances, i.e. $x = q, y = k_{llm}$. For example, given a question `"What can owls do \n (A) fly (B) speak"` as the input, the target is to predict `"Owls are birds. Birds can fly"`. In this way, the LEROS model learns to imitate the helpful rationales generated by LLMs.

Optional Further Tuning After the above stages, LEROS can be directly applied to unseen CQA tasks. In addition, it can serve as a good base model for further tuning. Besides fine-tuning with specialized rationale data, we can optimize it through reinforcement learning with the feedback of QA models on downstream tasks, as did in Liu et al. (2022a). We leave it as an optional step and discuss it in the later experiment section.

3.4. Inference

During inference, following Liu et al. (2022b), for each test question q, we first sample a set of rationales from LEROS with an additional blank rationale. The set is denoted as $K(q)$. We then enhance the generic QA model with $K(q)$ to predict the answer. Specifically, we concatenate the q with each k and use the QA model to predict the probability of each answer $p(a|q \circ k)$. The final predicted answer $\hat{a}$ is given as:

$$\hat{a} = \arg\max_{a \in A} \max_{k \in K(q)} p(a|q \circ k) \tag{3}$$

	Initial Data ($\mathcal{D}^{syn}$)	Queried Data	Consistent Data	Refined Data ($\mathcal{D}^{refine}$)
Train	823K	441K	303K	173K
Dev	67K	/	/	1K

Table 1: The statistics of synthetic datasets.

where each choice is linked with a rationale that maximizes its probability, and the final prediction is the choice with the highest probability.

4. Experiments Setup

In this section, we describe the experimental setting for evaluating the proposed framework.

4.1. Data

Knowledge Source For synthesizing data, We use ATOMIC-2020 (Hwang et al., 2021) and CWWV subset of CSKG (Ma et al., 2021) as the source CKGs. ATOMIC2020 contains knowledge triples across 23 relations, involving commonsense about social interaction, physical entities and general events. CWWV contains aligned commonsense knowledge across 14 relations from ConceptNet, WordNet, Wikidata and VisualGenome.

Synthesized Instances For ATOMIC2020, we extract knowledge triples from the official training and development set and synthesize 666K and 59K QA instances respectively. For CWWV, we reuse the synthesized instances from Ma et al. (2021), which contain 157K instances for training and 8K for validation. We pair these instances with their source knowledge. The combined synthetic QA dataset is named as $\mathcal{D}^{syn}$. In addition, we augment rationales for 441K instances from $\mathcal{D}^{syn}$ by querying `gpt-3.5-turbo-0301`[2] with the default generation setting. We use 12 human-authored examples and 10 synthesized instances with the highest helpfulness in previous rounds for prompting. During querying, the prompt contains random 3 examples and random 10 questions to be answered. We only generate one rationale and one answer for each question. After consistency and helpfulness refining ($S_0 = 0.01$), we obtain 174K high-quality QA instances with rationales, from which we sample 173K and 1K instances respectively for training and validation. We name the dataset as D^{refine}. The statistics are summarized in Table 1.

Evaluation Benchmarks For zero-shot evaluation, we evaluate the models on the following five benchmarks: CommonsenseQA (**CSQA**) (Talmor et al., 2019), **QASC** (Khot et al., 2020), PhysicalIQA (**PIQA**) (Bisk et al., 2020), SocialIQA (**SIQA**) (Sap et al., 2019), and WinoGrande (**WG**) (Sakaguchi et al., 2020). As their test sets are hidden and have submission restrictions, we mainly report the accuracy on their development sets.

4.2. Model Implementation

Generic QA Models For the feedback in data synthesizing and most of the evaluation experiments, we use `UnifiedQA-large`[3] (Khashabi et al., 2020) as the QA model. It is a generic QA model based on `T5-large`[4] (770M parameters) (Raffel et al., 2019) and trained on eight QA tasks. These tasks do not include the evaluation benchmarks used in our experiments and thus the model is evaluated in the zero-shot setting.

Training Leros We initialize Leros based on `T5-large`. For warming-up with commonsense, we train the model on $\mathcal{D}^{syn}$ for $50,000$ steps and set batch size to 128, learning rate to 1×10^{-5}. The learning rate is warmed up in the first 100 steps and linearly decayed to 0 in the remaining steps. For imitating high-quality rationales, we train the model on $\mathcal{D}^{refine}$ with the same hyperparameters. During each of the training stages, we save checkpoints with the lowest loss on the validation data.

Inference Setting During inference, we use nucleus sampling (Holtzman et al., 2019) ($p = 0.7$) to sample 10 rationales from Leros for each question. For the QA model, the concatenated question-rationale input format is `"{q} \n {k}"`, which is in line with the context format of UnifiedQA. We feed the concatenated input to the QA model and normalize the average log-likelihood for each choice to obtain the probability $p(a|q \circ k)$. The final prediction is given with Equation 3.

4.3. Baselines and Model Variants

We include the following zero-shot baselines in the experiments.

- **UQA** represents the `UnifiedQA-large` model without the rationale input, which provides the base performance.
- **UQA**$_{syn}$ represents a `UnifiedQA-large` variant which is fine-tuned on $\mathcal{D}^{syn}$ to predict answer without the rationale input. It is similar to previous zero-shot CQA methods based on synthetic data (Ma et al., 2021).
- **Few-shot GPT-3.5-turbo** represents the rationales generated by `GPT-3.5-turbo-0301` with few-shot prompting.

[2]https://platform.openai.com/docs/api-reference/chat/

[3]https://huggingface.co/allenai/unifiedqa-t5-large

[4]https://huggingface.co/t5-large

- **Few-shot GPT-3** represents rationales generated by GPT-3 (13B) with few-shot prompting.
- **Self-talk GPT-3** represents the rationales generated by GPT-3 (13B) with self-talk prompting (Shwartz et al., 2020).

As a comparison, we also include the following rationale models that were trained with the feedback from the training data of target benchmarks.

- **RAINIER** is a rationale-generation model proposed in Liu et al. (2022a). It is first trained with the rationales generated by GPT-3 and then tuned through reinforcement learning with QA model feedback. It requires the training data of target benchmarks. Both RAINIER and LEROS are based on `T5-large` (770M).
- **LEROS**$_{RL}$ is a variant model of LEROS, which is initialized with LEROS and further applied the reinforcement learning of RAINIER.

Besides, **Gold Rationale** represents the human-authored rationales for some benchmarks, which provides upper bound performance. Specifically, for CSQA, we use the explanations from ECQA (Aggarwal et al., 2021); for QASC, we use the composed facts provided in the official data.

4.4. Other QA Models

To assess the transferability of generated rationales, we also evaluate LEROS with other generic QA models besides `UnifiedQA-large` and its siblings. These QA models are listed as follows.

- **RoBERTa-Large-CSKG** (Ma et al., 2021) is a representative zero-shot CQA model trained on synthesized QA instances. Its usage is to concatenate the question with each choice and scoring the entire sequence for ranking. As the model is not trained with rationales, to make it work with LEROS, we simply add the generated rationale before the input sequence. Besides, **DeBERTa-v3-Large-CAR** (Wang et al., 2023a) is a similar state-of-the-art zero-shot CQA model that improves the data synthesizing. We apply it with LEROS in the same way.
- **Llama 2** (Touvron et al., 2023) is a famous family of open large language models. We use a vanilla version **Llama-2-7B** and an RLHF fine-tuned version **Llama-2-chat-7B** in the experiments. To make them serve as QA models and work with LEROS, we add a 5-shot prompt before the input question. For comparison, we also implement Self-Consistency with Chain-of-Thought (Wang et al., 2023b) (**CoT-SC**) as a baseline to elicit the model's own knowledge.

5. Results and Analyses

5.1. Main Results

Overall Performance Table 2 shows the performance of LEROS and baselines on the development sets. From the results, we find **(1)** the rationales generated by LEROS increase the zero-shot performance of UnifiedQA on the five benchmarks by 6.3% on average, which indicates that our methods can provide helpful knowledge for the QA model and improve the performance on unseen CQA tasks. **(2)** Among the benchmarks, QASC (+13.5%), CommonsenesQA (+6.46%) and SocialIQA (+5.93%) have larger improvement, while WinoGrande (+1.66%) only has slight improvement. We think it is because the latter one less overlaps the domain of source CKGs and has greater reasoning difficulty. The performance on test sets (Table 3) is in line with the above observations.

Few/Zero-shot Baselines **(1)** All few-shot or zero-shot methods in Table 2 bring improvement to the performance on the basis of UQA, and GPT-3.5-turbo provides the best performance as it is optimized on human feedback and has possibly the largest model size. **(2)** In comparison with GPT-3-based prompting baselines, LEROS brings better average performance gains with much fewer parameters (770 million versus 13 billion), which shows the effectiveness of our methods to exploit knowledge from both CKGs and LLMs. **(3)** UQA$_{syn}$ is fine-tuned on the same synthesized QA instances for training LEROS but yields less improvement, which indicates that enhancing QA models with explicit rationales is a strong way to improve zero-shot performance. **(4)** On CommonsenseQA and SocialIQA, LEROS has the closest performance with few-shot GPT-3.5-turbo, because the two benchmarks have overlapped domain with the source CKGs of synthesized data. It indicates that LEROS can help small models make better use of in-domain knowledge and narrow the gap with much larger models.

Feedback Tuning In Table 2, even without the training data of target benchmarks, LEROS has already achieved better performance than RAINIER. Moreover, these methods can be complementary. Initialized with LEROS and further tuned with the reinforcement learning process of RAINIER, the LEROS$_{RL}$ variant provides even better performance. We conjecture that LEROS can learn both knowledge from CKGs and the rationale generation ability of advanced LLMs via extensive synthesized instances, although the synthesized instances are worse than real benchmark-specific training data in question quality and complexity. Therefore, LEROS provides a strong foundation for further tuning.

Changing QA Models We apply LEROS to different UnfiedQA variants and other generic QA model

Method	Rationale Source	QA Model	Dataset CSQA	QASC	PIQA	SIQA	WG	Average	Avg. Gain
	Gold Rationale	UQA	89.92	83.37	-	-	-	-	-
Few/Zero-shot	-	UQA	61.43	43.09	63.66	53.84	53.35	55.07	+0.00
	-	UQA$_{syn}$	62.24	52.27	66.05	55.42	55.25	58.25	+3.17
	Few-shot GPT-3.5-turbo	UQA	70.02	66.52	71.82	61.00	58.64	65.60	+10.53
	Self-talk GPT-3 (13B)	UQA	63.31	49.89	65.23	51.89	52.96	56.66	+1.58
	Few-shot GPT-3 (13B)	UQA	66.34	53.24	65.25	58.29	**55.56**	59.74	+4.66
	(Ours) LEROS (770M)	UQA	**67.89**	**56.59**	**67.57**	**59.77**	55.01	**61.37**	**+6.29**
Feedback Tuning	RAINIER (770M)	UQA	67.24	54.97	65.67	57.01	56.91	60.36	+5.09
	(Ours) LEROS$_{RL}$ (770M)	UQA	**70.35**	**60.15**	**69.53**	**64.32**	**59.27**	**64.72**	**+9.65**

Table 2: Few/Zero-shot and feedback-tuned results on the benchmarks (development sets).

	QASC	PIQA	SIQA	WG	Avg.
UQA	45.65	65.50	57.21	54.67	55.76
UQA+RAINIER	54.13	67.09	59.01	**57.39**	59.41
UQA+LEROS	**55.33**	**67.67**	**60.90**	56.14	**59.81**

Table 3: Results on the benchmarks (test sets).

QA Model→ Rationale Model↓	UQA (small)	UQA (base)	UQA (large)	UQA (3b)
-	39.07	45.51	55.07	66.51
RANIER	48.60	54.77	60.36	67.85
LEROS	**49.05**	**56.12**	**61.37**	**67.91**

Table 4: Average performance of applying different UnifiedQA variants with LEROS.

implements. The average performance is shown in Table 4 and Table 5. From Table 4, we find LEROS can consistently bring gains for different sizes of QA models, which is in line with Liu et al. (2022a). From Table 5, we find LEROS improve the performance of both previous zero-shot CQA models and the latest open large language models (i.e. `Llama2-7B`), even though these models are implemented in a completely different way from UnifiedQA. The results demonstrate that the rationales generated by LEROS contain transferable knowledge and are useful for different models.

Method	Average
RoBERTa-Large-CSKG (Ma et al., 2021)	64.0
LEROS + RoBERTa-Large-CSKG	**65.2**
DeBERTa-v3-Large-CAR (Wang et al., 2023a)	70.2
LEROS + DeBERTa-v3-Large-CAR	**71.1**
Llama2-7B (Few-shot)	53.4
Llama2-7B (CoT-SC)	55.8
LEROS + Llama2-7B (Few-shot)	**57.2**
Llama2-chat-7B (Few-shot)	58.6
Llama2-chat-7B (CoT-SC)	61.9
LEROS + Llama2-chat-7B (Few-shot)	**63.0**

Table 5: Average performance of applying QA models other than UnfiedQA in few/zero-shot setting.

	CSQA	QASC	PIQA	SIQA	WG	Avg.
None	61.43	43.09	63.66	53.84	53.35	55.07
LEROS	**67.89**	**56.59**	**67.57**	**59.77**	**55.01**	**61.37**
-WM	66.42	54.75	67.03	58.96	53.83	60.20
-IM	56.76	39.96	61.75	53.89	53.20	53.11
-CS	65.27	55.37	66.16	58.47	54.78	60.01
-HP	66.42	56.26	66.92	58.96	54.75	60.66
Source K	58.89	47.30	63.55	54.20	51.85	55.16
CKG Path	65.11	51.30	66.16	57.16	54.93	58.93
LLM SynQ.	61.34	49.46	66.16	58.29	53.51	57.75

Table 6: Performance of different variants of LEROS. (**-WM**): Removing warming-up with commonsense. (**-IM**): Removing imitating high-quality rationales. (**-CS**): Removing consistency refining. (**-HP**) Removing helpfulness refining. (**Source K**): Training on the source knowledge k_{source} rather than k_{llm}. (**CKG Path**): Training on sampled knowledge paths rather than k_{llm}. (**LLM SynQ**): Training on LLM generatd question instances.

Knowledge	(Cake, UsedFor, feed to guests)
Rule-based Question	What can cake be used for? (A) record achievement **(B) feed to guests {correct}** (C) lose the weight
LLM-based Question	Which of the following foods would be a good option for serving guests? (A) Pizza (B) Salad **(C) Cake {correct}** (D) Tacos

Table 7: Synthesized question instances with a rule-based method and a LLM-based method. The LLM generates more fluent questions but it also provides inappropriate distractors.

5.2. Ablation Study

To further analyze the effectiveness of different parts of the proposed framework, we show the ablation results of several LEROS variants. All of them are evaluated with `UnifiedQA-Large`.

Refining and Training As shown in Table 6, we first remove different parts of the proposed frame-

Task	Question/**Rationale**	Category
CSQA	If there is a place that is hot and arid, what could it be? (A) bland (B) lifeless (C) sandy (D) neutral (E) freezing **Hot and arid can mean a place that is dry and inhospitable.**	attribute
QASC	What can measure pounds? (A) animals (B) lamphreys (C) a mouse (D) a ruler (E) humans (F) surveyor (G) a scale (H) a microscope **Measuring pounds is done using a scale.**	use
PIQA	how do you blame someone? (A) say they did it. (B) say you did it for them. **Blaming someone involves saying they did something wrong.**	subevent
SIQA	Ash always performed better at his workplace after a warm cup of coffee. What will Ash want to do next? (A) start a new task (B) take some nyquil (C) go home **After having a warm cup of coffee, people usually feel refreshed and want to continue their work.**	behavior
WG	Angela did a bunch of crunches and sit-ups but Cynthia didn't, consequentially _ had six- pack abs. (A) Angela (B) Cynthia **Doing crunches and sit-ups is a common exercise to get six-pack abs.**	taxonomy

Table 8: Examples of helpful rationales generated by LEROS.

work and evaluate the resulting models. Generally, these models all yield worse performance than the fully trained LEROS. Without imitating high-quality rationales, the performance is greatly damaged, which shows the importance of training on rationale-augmented data. Removing the refining process marginally decreases the performance gains, which shows that models can learn with noisy instances but high-quality instances are more useful.

Alternative Synthesizing Strategies As CKGs and LLMs are independent sources, we also examine several alternative strategies for synthesizing questions and rationales. Specifically, we try to use the source knowledge of questions **(Source K)** or sample multi-hop connection paths from CKGs based on concepts mentioned in the question (**CKG Path**) to create question-specific rationales. We also try to use LLMs instead of rules to generate questions based on a knowledge triple (**LLM SynQ**). These variant methods all yield worse performance. Interestingly, the results show the LLM is worse than rule-based methods at synthesizing questions for given knowledge. It is partly because the LLM is not good at generating distractors, as shown in Table 7. Also, the CKG Path variant provides strong performance, which shows the importance of relevant knowledge. Note that we do not include the results of directly generating questions without providing knowledge from CKGs, because the LLM-generated questions are highly repetitive, even if we add previously generated instances for prompting. We leave better strategies for synthesizing QA instances with LLMs as future work.

Training Data Size To investigate the impact of data size, we further train models using 10%, 30%, 50%, 70% and 90% of training data respectively. As a comparison, we try to remove refining prompting (i.e. add refined high-quality instances into the prompts) in synthesizing rationales and train corresponding variant models. The performance curve is shown in Figure 4. We find that as the data size increases, the performance improvement tends to converge. Meanwhile, applying refining prompting can improve the data efficiency, achieving better performance with the same synthesizing budget.

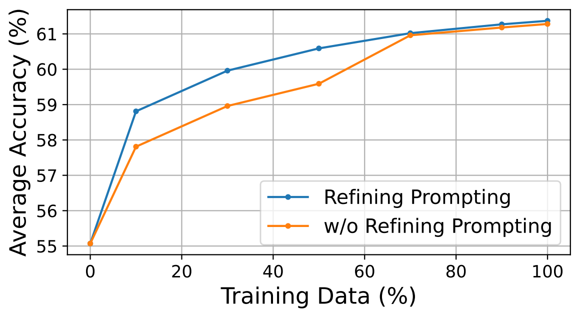

Figure 4: The performance curve of altering the size of training data.

5.3. Manual Analyses

For further analyses, we randomly select 100 instances from the evaluated benchmarks and manually annotate whether the rationale generated by LEROS is relevant, factual, and helpful for the question. Generally, 86% of the rationales are annotated as relevant, 69% are factual and 55% are helpful. For instances where the rationale rectifies the answer, 89% of the rationales are helpful. We show some of the helpful examples of rationales in Table 8 and mark the knowledge category that they express. These instances show that although LEROS is trained with synthesized question instances, it can generalize on unseen commonsense question answering tasks, providing helpful and readable evidence. On the other hand, we also find 70% of generated rationales are single facts between two concepts, indicating the multi-hop reasoning ability requires further improvement as the model only learns from synthesized questions with low complexity.

6. Conclusion

In this paper, we propose a novel framework for explicit reasoning on commonsense question answering, which takes the best of commonsense knowledge graphs and large language models to synthesize rationale-augmented QA data. Based on solely synthesized data, we train a rationale generation model that can provide textual rationales for unseen questions. Empirical results show the model improves the performance of QA models on five unseen CQA benchmarks, surpassing previous methods that require training data of target benchmark and 10x larger language models. It can directly work with different generic QA models or serve as a good start for further tuning. This work shows a novel and effective way to transfer commonsense knowledge from both symbolic sources (CKGs) and neural sources (LLMs) to smaller special-purpose models. It also reveals enlightening phenomena for LLM-based synthesized resources.

7. Limitations and Ethics Considerations

This work has limitations in some aspects. First, the scope of synthesized questions is still affected by the coverage of source CKGs. Recent CKGs built from large language models can be an alternative source for synthesizing CQA instances for broader domains, yet we have not explored its feasibility. Second, due to the simple structure of synthesized questions, the model cannot learn much about complex reasoning structures and hence brings less improvement on hard CQA tasks (e.g. WinoGrande), which remains a problem to be solved in future work. Third, our framework contains English-specific prompting designs. We only evaluate its effectiveness on English benchmarks. It requires additional adaptation for applying the framework to other languages.

In addition, we mainly focus on the helpfulness of generated rationales for assisting QA models in this work. The LEROS model can also generate human-readable rationales, yet it is not adequate to serve as a reliable source to provide trustworthy knowledge. We have not fully examined the synthesized QA instances used for training the model. The synthesized data are based on publicly available knowledge graphs and pretrained large language models, which could contain unconfirmed bias or toxic information and indirectly affect the trained LEROS model.

8. Acknowledgements

This work is supported by the National Key Research and Development Program of China (No.2022ZD0160503), the National Natural Science Foundation of China (No. 62176257). This work is also supported by the Youth Innovation Promotion Association CAS.

9. Bibliographical 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, Online. Association for Computational Linguistics.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Pratyay Banerjee and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Rachit Bansal, Milan Aggarwal, Sumit Bhatia, Jivat Kaur, and Balaji Krishnamurthy. 2022. CoSe-co: Text conditioned generative CommonSense contextualizer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1128–1143, Seattle, United States. Association for Computational Linguistics.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):4923–4931.

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*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qianglong Chen, Feng Ji, Haiqing Chen, and Yin Zhang. 2020. Improving commonsense question answering by graph-based iterative retrieval over multiple knowledge sources. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2583–2594, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wanyun Cui and Xingran Chen. 2022. Enhancing natural language representation with large-scale out-of-domain commonsense. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *CoRR*, abs/2302.04752.

Zi-Yi Dou and Nanyun Peng. 2022. Zero-shot commonsense question answering with cloze translation and consistency optimization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10572–10580. AAAI Press.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Xin Guan, Biwei Cao, Qingqing Gao, Zheng Yin, Bo Liu, and Jiuxin Cao. 2022. CORN: Co-reasoning network for commonsense question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1677–1686, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *ArXiv*, abs/2212.09689.

Zixian Huang, Ao Wu, Jiaying Zhou, Yu Gu, Yue Zhao, and Gong Cheng. 2022. Clues before answers: Generation-enhanced multiple-choice QA. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3272–3287, Seattle, United States. Association for Computational Linguistics.

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, Online. Association for Computational Linguistics.

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*, pages 1896–1907, Online. Association for Computational Linguistics.

Yu Jin Kim, Beong-woo Kwak, Youngwook Kim, Reinald Kim Amplayo, Seung-won Hwang, and Jinyoung Yeo. 2022. Modularized transfer learning with multiple knowledge graphs for zero-shot commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2244–2257, Seattle, United States. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *CoRR*, abs/2004.05569.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A systematic investigation of commonsense knowledge in large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, Abu Dhabi, United Arab Emirates. 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)*, pages 2829–2839, Hong Kong, China. 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*, pages 8938–8958, Abu Dhabi, United Arab Emirates. 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)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13507–13515.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *NeurIPS*.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. Exploring ways to incorporate additional knowledge to improve natural language commonsense question answering. *CoRR*, abs/1909.08855.

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*, pages 4179–4192, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach weaker agents? teacher explanations improve students via theory of mind. *CoRR*, abs/2306.09299.

Melanie Sclar, Peter West, Sachin Kumar, Yulia Tsvetkov, and Yejin Choi. 2022. Referee: Reference-free sentence summarization with sharper controllability through symbolic knowledge distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9649–9668, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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)*, pages 4615–4629, Online. Association for Computational Linguistics.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4723–4734, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Chenhao Wang, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2022a. CN-AutoMIC: Distilling Chinese commonsense knowledge from pretrained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9253–9265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4129–4140, Online. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. *CoRR*, abs/2305.14869.

Wenya Wang, Vivek Srikumar, Hanna Hajishirzi, and Noah A. Smith. 2022b. Elaboration-generating commonsense question answering at scale. *CoRR*, abs/2209.01232.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022c. Self-instruct: Aligning language model with self generated instructions. *ArXiv*, abs/2212.10560.

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 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*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqi Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. 2022. Human parity on commonsenseqa: Augmenting self-attention with external attention. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 2762–2768. ijcai.org.

Jun Yan, Mrigank Raman, Aaron Chan, Tianyu Zhang, Ryan Rossi, Handong Zhao, Sungchul Kim, Nedim Lipka, and Xiang Ren. 2021. Learning contextualized knowledge structures for commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4038–4051, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Hongming Zhang, Yintong Huo, Yanai Elazar, Yangqiu Song, Yoav Goldberg, and Dan Roth. 2023. CIKQA: Learning commonsense inference with a unified knowledge-in-the-loop QA paradigm. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 114–124, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiarui Zhang, Filip Ilievski, Kaixin Ma, Jonathan Francis, and Alessandro Oltramari. 2022. A study of zero-shot adaptation with commonsense knowledge. In *Automated Knowledge Base Construction (AKBC)*.

10. Language Resource References

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.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. 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 6384–6392. AAAI Press.

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.

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. 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)*, pages 4463–4473, Hong Kong, China. 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.

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, Minneapolis, Minnesota. Association for Computational Linguistics.