Knowledge-enhanced Prompt for Open-domain Commonsense Reasoning

Chen Ling[1], **Xuchao Zhang**[2], **Xujiang Zhao**[3], **Yifeng Wu**[1],
Yanchi Liu[3], **Wei Cheng**[3], **Haifeng Chen**[3], **Liang Zhao**[1]

[1]Emory University, Atlanta, GA {chen.ling, yifeng.wu, liang.zhao}@emory.edu
[2]Microsoft Research, Redmond, WA xuchaozhang@microsoft.com
[3]NEC Laboratories America, Princeton, NJ {xuzhao, yanchi, weicheng, haifeng}@nec-labs.com

Abstract

Neural language models for commonsense reasoning often formulate the problem as a QA task and make predictions based on learned representations of language after finetuning. However, without providing any finetuning data and predefined answer candidates, *can neural language models still answer commonsense reasoning questions only relying on external knowledge?* In this work, we investigate a unique yet challenging problem - open-domain commonsense reasoning that aims to answer questions without providing any answer candidates and finetuning examples. Our proposed method leverages neural language models to iteratively retrieve reasoning chains on the external knowledge base, which does not require task-specific supervision. The reasoning chains can help to identify the most precise answer to the commonsense question and its corresponding knowledge statements to justify the answer choice. We conduct experiments on two commonsense benchmark datasets. Compared to other approaches, our proposed method achieves better performance both quantitatively and qualitatively.

Introduction

Large-scale pretrained language models (PLMs) learn to implicitly encode basic knowledge about the world by training on an extremely large collection of general text corpus and refining on downstream datasets, which have recently taken over as the primary paradigm in NLP. Although PLMs have excelled in many downstream tasks, they still face two major cruxes in reasoning-related tasks: 1) PLMs frequently encounter difficulties when the required knowledge is absent from the training corpus or the test instances are not formulated as question-answering format, and 2) PLMs base their predictions on implicitly encoded knowledge that is incapable of handling structured reasoning and does not offer explanations for the chosen response. As shown in Figure 1, if we are presented with a question that its domain is different from examples seen during the training. For medical-domain questions like *What are both **Family Doctor** and **Surgeon** refer to?*, we aim to generate an abstracted meaning for both entities without providing any answer candidates. However, without providing any finetuning instances, the state-of-the-art PLM T5-3b (Kale and Rastogi 2020) would generate

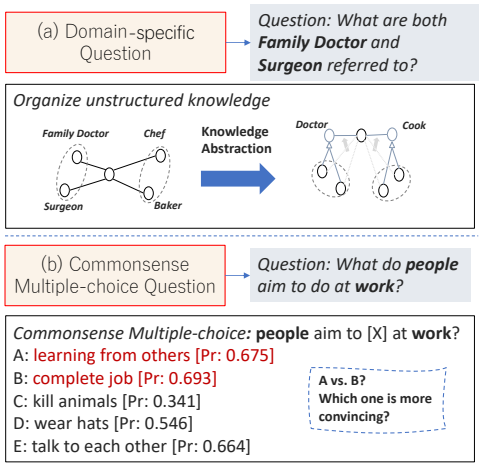


Figure 1: Two cruxes of using PLMs in commonsense reasoning: 1) Without finetuning, PLMs may not handle out-of-distribution or domain-specific reasoning questions. 2) PLMs need pre-existing answer candidates and they generally cannot justify their prediction results.

an irreverent answer: *quizlet*. In addition, for commonsense questions: *People aim to* [MASK] *at work*, the paradigm of prompt learning with PLMs often formulate the problem to multiple-choice QA and calculate the likelihood of the whole sentence by filling in the blank with each answer candidate. However, both answers *learning from others* and *complete job* can fit in the semantics of the question. PLMs cannot provide justification for why a certain answer can be chosen. Both cases reveal that the prediction of commonsense reasoning requires robust and structured reasoning to integrate the explicit information offered by the question context and external knowledge.

In this work, we focus on the Open-domain Commonsense Reasoning task, which requires machines to make human-like presumptions about the type and essence of ordinary situations ***without*** presenting any answer candidates and finetuning examples. Addressing open-domain commonsense reasoning problems would be more ordinary than formulating the commonsense reasoning problem into QA task since we might not have any pre-existing answer candidates for the question or the training resources and data necessary to do fine-tuning. The intricacy of this task requires us to incor-

porate explicit information provided by external knowledge in the reasoning step. However, open-domain commonsense reasoning is still under-explored because there are two obstacles that existing external knowledge-enhanced commonsense reasoning methods cannot handle. *1) the difficulty of retrieving relevant information for structured reasoning*. Retrieving relevant from external knowledge under the open-domain setting can be quite challenging since the overall searching is neither learning-based nor guided by certain targets. Existing methods (Ma et al. 2021; Bian et al. 2021) have been leveraging learning-based ranking algorithms to retrieve knowledge, but it requires substantial pretraining or finetuning on the testing domain, which is not applicable in our setting. The other line of works (Lin et al. 2019; Yasunaga et al. 2021) builds local knowledge graphs between the question and provided answer candidates, but their algorithm cannot be executed if there are no pre-existing candidates. *2) the difficulty of enhancing explainability in the retrieved knowledge.* The current PLMs more perform like a black-box model that lacks adequate explanations of the answer selection. Other than knowing the correct answer, it would be more essential to see what rationale can be taken from external knowledge to support PLMs making human-like presumptions. Existing methods either directly sample sentences as knowledge statements based on the question (Liu et al. 2021) or require learning steps to generate explanations (Paranjape et al. 2021). None of both approaches can generate explanations of the answer choice under the zero-shot and open-domain settings as required in our case.

In this paper, we present the external Knowledge-Enhanced Prompting method (KEP) to solve the open-domain commonsense reasoning task. We utilize the implicitly stored knowledge in PLMs to iteratively recover reasoning chains from the organized external knowledge base, as opposed to alternative methods that need direct supervision of the reasoning processes. Additionally, each retrieved reasoning path acts as the explicit justification for the answer selection. Following is a summary of the work's main contributions. 1) We formulate the novel open-domain commonsense reasoning problem and identify its unique challenges. 2) We iteratively collect reasoning chains from the external structured knowledge base using the implicit information stored in PLMs as guidance. 3) The proposed approach is capable of identifying the most appropriate answer and automatically producing the corresponding explanations.

Related works

External Knowledge for Commonsense Reasoning. Many studies have demonstrated that PLMs empirically cannot perform well on reasoning-related tasks only relying on their implicitly stored knowledge during training (Jiang et al. 2020). Combining PLMs and external knowledge for reasoning has recently gained lots of attention, methods have been invented to inject commonsense knowledge into language models, either by pretraining on knowledge bases (Ma et al. 2021), finetuning the model on the test domain (Bian et al. 2021), or leveraging structured knowledge base (e.g., ConceptNet) (Yasunaga et al. 2021) so that it can reason with additional retrieved knowledge. However, none of these works can be trivially adapted to solve the open-domain commonsense reasoning problem since they either require substantial pretraining/finetuning or pre-existing answer candidates for the question. Lastly, our work is conceptually related to open-domain Question-answering but our reasoning step is conducted on a knowledge graph, which is a more robust and organized external knowledge source.

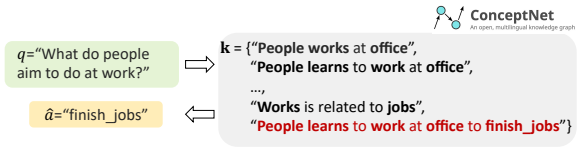


Figure 2: Example of the open-domain commonsense reasoning: the model takes the question as input and returns supporting knowledge statements with the predicted answer.

Explanations for Commonsense Reasoning. Other than predicting the correct answer, it is also important to explore what are explicit reasoning steps behind the answer selection. Other than works that require direct supervision to predict explanation (Paranjape et al. 2021), (Bosselut, Le Bras, and Choi 2021) proposed to leverage knowledge graph to acquire reasoning paths as the explanation in an unsupervised way. However, this approach requires predefined answers to guide the reasoning, which is not applicable in open-domain commonsense reasoning. The other line of works (Shwartz et al. 2020; Liu et al. 2021) have also been utilizing model-generated text as the clarification of the commonsense question and empirically demonstrated the performance can be boosted by augmenting the query with knowledge statements. However, purely relying on the language model still lacks the model transparency, and the generated knowledge statement cannot be empirically served as the answer explanation.

Proposed Method

In this section, we first introduce the problem formulation, and then discuss the detailed framework of the proposed method, which can be divided into three components: 1) entity extraction and linking, 2) local knowledge graph expansion, and 3) explanation generation and answer prediction.

Problem Formulation

We aim to solve open-domain commonsense reasoning questions using knowledge from a PLM and a structured knowledge graph G. The knowledge graph $G = (V, E)$ (e.g., ConceptNet) is a multi-relational graph, where V is the set of entity nodes, $E \subseteq V \times R \times V$ is the set of edges that connect nodes in V, where R represents a set of relation types. Specifically, for open-domain commonsense reasoning questions q (i.e., given a question q without providing answer candidates), the target of this work is to determine 1) a local knowledge graph $G_q \in G$ contains relevant information of q; 2) a set of knowledge statements $\mathbf{k} = \{k_1, k_2, ..., k_m\}$; and 3) an entity $\hat{a}$ extracted from $\mathbf{k}$ that is precise to answer the question q. For example in Figure 2, to answer the open-domain commonsense question "what do people aim to do at work?", we aim at first extracting all plausible knowledge statements from the external knowledge base that can provide us logical information to answer the question. Among all the

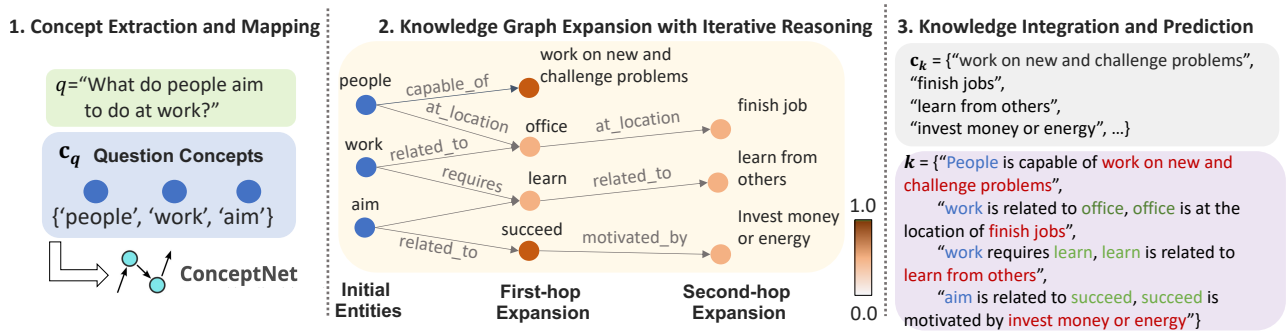


Figure 3: The framework of the proposed method, which consists of 1) concept extraction and entity linking; 2) local knowledge graph expansion with iterative reasoning steps; and 3) knowledge integration and final answer prediction.

statements, we select the most precise one (i.e., *people learn to work at the office to finish jobs*) and extract the answer $\hat{a}$ = *finish_jobs* such that the following joint likelihood can be maximized.

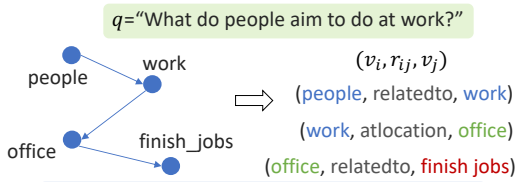
$$P(\hat{a}, \mathbf{k}|q, G_q) = P(\mathbf{k}|q, G_q) \cdot P(\hat{a}|\mathbf{k})$$

However, the expressiveness and generalization capability of existing frameworks are still limited to solving the open-domain commonsense reasoning problem due to two critical challenges. Firstly, retrieving knowledge statements that can explicitly reveal the reasoning steps is challenging, and most existing works (Shwartz et al. 2020; Liu et al. 2021) solely utilize PLMs $p_\theta(k|q)$ to sample a set of question-related statements $\mathbf{k} = \{k \sim p_\theta(k|q)\}$. This approach makes the sampling process time-consuming and lacks knowledge awareness, let alone the interpretability of the generated knowledge. Secondly, another line of work (Yasunaga et al. 2021; Lin et al. 2019) has been trying to build a local knowledge graph G between q and all the provided answer candidates to enhance the interpretability of the answer selection. However, as we are dealing with open-domain or domain-specific questions without any given answer candidates, building the local graph also becomes infeasible.

Next, we discuss how to initiate the local knowledge graph and iteratively reason over it to find all plausible knowledge statements and the most convincing answer. We demonstrate the overall framework in Figure 3.

Local Graph Construction and Expansion

Knowledge Graph Entity Linking. ConceptNet enables a variety of useful context-oriented reasoning tasks over real-world texts, which provides us with the most suitable structured knowledge in the open-domain commonsense reasoning task. To reason over a given commonsense context using knowledge from both the PLM and the knowledge graph G, the first step of the framework is to extract the set of critical entities $\mathbf{c}_q = \{c_q^{(1)}, ..., c_q^{(i)}, ...\}$ from the question q that have the surjective mapping to nodes $V_q \in V$ in the knowledge graph. Since q is often presented in the form of non-canonicalized text and contains fixed phrases, we follow the prior work (Becker, Korfhage, and Frank 2021) to map informative entities $\mathbf{c}_q$ from q to conjunct concept entities in ConceptNet by leveraging the latent representation of the query context and relational information stored in G.



Relation Groups	Merged Relations	Relation Text
antonym/distinctfrom	antonym	is the antonym of
atlocation/locatednear	atlocation	is at location of
relatedto/similarto/synonym	relatedto	is related to

Figure 4: Knowledge statement transformation and cloze-based prompt construction.

Reasoning Over Local Knowledge Graph. To imitate the human reasoning process, we aim to retrieve reasoning paths within L hops from G to form the local knowledge subgraph G_q that has the highest coverage to the question concepts $\mathbf{c}_q$. Ideally, each path in G_q can be regarded as a reasoning chain that helps to locate the most precise answer and its explanation to the question q. However, expanding L-hop subgraph G_q from $\mathbf{c}_q$ is computationally prohibited. Unlike other works (Yasunaga et al. 2021; Lin et al. 2019) that build G_q between the question q and all answer candidates, the open-domain commonsense reasoning problem does not provide any directions (i.e., answer candidates). The typical node size of a 3-hop local knowledge graph with $|\mathbf{c}_q| = 3$ could easily reach $1,000$ on ConceptNet, and many nodes are irrelevant under the current question context.

Reasoning Path Pruning. In order to make the process of reasoning path expansion scalable, we incorporate the implicit knowledge in PLMs to prune irreverent paths. Specifically, we pair the question q with the text of node v along with the reasoning-path-transformed knowledge statement to form a cloze-based prompt $W = [q; v_j; (v_i, r_{ij}, v_j)]$ in order to turn the local graph expansion problem into an explicit reasoning procedure by directly answering the question with explanation. For example in Figure 4, the prompt is formatted as *What do people aim to do at work? <**node**>, because <**reasoning path**>*. Note that we leverage predefined template to transform the triplet (v_i, r_{ij}, v_j) into natural language. Specifically, ConceptNet contains lots of relations ($|R| = 34$) and some of them share similar meanings (e.g., both *antonym* and *distinct_from* have the same mean-

ing *antonym*). Moreover, we predefine templates to transform the reasoning path triplets into natural language. For example, (*work*, *antonym*, *unemployment*) can be translated to *work is the antonym of unemployment*. We also illustrate a few examples of the merged types and templates in Figure 4. To evaluate whether we keep the reasoning path, we propose to compute the commonsense score of the reasoning path, where we use the PLM to score the relevance of each reasoning path given the context of the question. Formally, suppose the logical sentence W consists of N words $W = \{\omega_1, ..., \omega_{n-1}, \omega_n, \omega_{n+1}, ..., \omega_N\}$, the commonsense score $\phi_l(W)$ of the logical sentence W composed at l-th hop expansion is defined as:

$$\phi_l(W) \coloneqq \sum_{n=1}^{N} \log(p_\theta(\omega_n|W_{\setminus n}))/N, \tag{1}$$

where the $W_{\setminus n}$ indicates the masked knowledge statement by replacing the token ω_n to the mask, and the denominator N reduces the influence of the sentence length on the score prediction. Intuitively, $\log(p_\theta(\omega_n|W_{\setminus n}))$ can be interpreted as how probable a word ω_n given the context. For example, by filling *blue* and *red* into the masked logical statement $W_{\setminus n}$ = *The sky is* [MASK], *blue* should have a higher score.

As we iteratively expand G_q, each $\phi_l(W)$ scores a unique reasoning path at a particular $l \in [1, L]$ depth in the graph. As marked in Figure 3, a higher score $\phi_l(W)$ indicates the node v_j should be kept for the next $(l + 1)$ hop expansion.

Knowledge Integration and Prediction

After we obtained the subgraph G_q consisting of all reasoning paths within L-hop with a high commonsense score, all the reasoning paths can be regarded as the supporting knowledge explanation. The final step is to make the answer prediction. We utilize beam search to only keep high-confidence reasoning paths and transform them into natural language by the designed template in the set of knowledge statements $\mathbf{k}$ during the retrieval phase. Starting from each entity in $\mathbf{c}_q$, each reasoning path within L-hop neighbor can then be seen as scoring a path to a particular answer node.

$$\log p_\theta(a|[q;k]) \propto \phi_L = \sum_{l=1}^{L} \phi_l; \quad (\hat{a}, \hat{k}) = \arg\max_{a \in \mathbf{a}, k \in \mathbf{k}} \phi_L,$$

where $\mathbf{a}$ is the set of all explored answer candidates, and the ϕ_L denotes the final score for each answer and can be interpreted as approximating the likelihood of answer a given a singular reasoning path $\{c \rightarrow v_1 \rightarrow v_2 \rightarrow \cdots \rightarrow a\}$. We can thus pick the answer $\hat{a}$ and its explanation $\hat{k}$ with the highest score as the final answer and supporting knowledge.

Experiment

We empirically verify the performance of the proposed method against other methods on commonsense reasoning benchmark datasets under the open-domain setting. In this work, the inference language model p_θ can be any existing masked language model either with the zero-shot setting or finetuned on the external knowledge base, and we leverage the masked language model RoBERTa-large (Liu et al. 2019) since it has larger representative power in commonsense ability with a less model size (Zhou et al. 2020).

Method	CSQA			QASC		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
GPT-3	0.552	– – –	– – –	0.424	– – –	– – –
DeBERTa-large	0.245	0.396	0.598	0.269	0.554	0.618
RoBERTa-large	0.275	0.477	0.682	0.294	0.523	0.578
KEP (Ours)	0.385	0.615	0.778	0.467	0.742	0.821

Table 1: Top-1, 3, and 5 prediction accuracy made by human annotators for each model. Note that GPT-3 generates text autoregressively instead of filling mask in the prompt, it can only generate one answer so we only label its Top-1 accuracy.

Experiment Setting

Dataset. We evaluate our method on two commonsense reasoning benchmarks. *1) CommonsenseQA (CSQA)* (Talmor et al. 2019) is a multiple-choice QA dataset about commonworld scenarios. The dataset is constructed on ConceptNet. *2) QASC* (Khot et al. 2020) is a multiple-choice QA dataset about grade school science. We randomly sample 400 questions from the testing set of both datasets. We also discard the multiple-choice answers in both datasets in order to simulate the open-domain scenario.

Comparison Methods and Evaluation Metrics. Since we are the first to investigate the open-domain commonsense reasoning problem, we have no direct opponents to compare the performance. In this work, we compare our model against state-of-the-art language models: GPT-3 (Brown et al. 2020), Deberta-large (He et al. 2021), and RoBERTa-large (Liu et al. 2019). Since GPT-3 is a generative model, we generate the answer in an autoregressive way. RoBERTa and DeBERTa are masked language models, we then transform the question into a discrete prompt (a few examples are listed in Table 2) and generate the most likely answer. In terms of evaluating the generated answers, since we do not have ground truth to evaluate the prediction correctness, we instead generate answer candidates for each commonsense question and rank them based on their commonsense score (Equation (1)). A human annotator indicates whether there exists a precise answer that fits the semantics in the top-N predicted answers.

Results

Quantitative Analysis. Table 1 summarizes the Top-N accuracy results. For each approach, the test results are obtained by evaluating if there is a precise answer in the top-5 generated answers. As can be clearly seen from the table, our proposed method excels both masked language models by an evident margin. Without accessing external knowledge, standard PLMs still lack knowledge awareness and generally cannot perform well on structured reasoning tasks. In addition, GPT-3 is pretrained on an extremely large corpus and contains 175 billion parameters, and our method KEP (based on RoBERTa-large with only 335 million parameters) demonstrates competitive performance in both datasets. We empirically testify most commonsense knowledge can be derived from the external knowledge base, and it is natural to elicit related knowledge from it.

Dataset	Commonsense Question	Answer Prediction			
		GPT-3	DeBERTa	RoBERTa	KEP (Ours)
CSQA	What do people aim to do at work? → (People aim to [MASK] at work.)	achieve success	burst	succeed	**work on new and challenging problems** (**Work** is done by **People**, **People** desires to **work on new and challenging problems**)
	Where would you find magazines alongside many other printed works? → (You find magazines alongside many other printed works at [MASK].)	magazine publishers	board	home	**bookstore** (**Magazine** is a type of **book**, **book** is at the location of **bookstore**.)
QASC	What is usually important for a doctor to do? → (A doctor is usually important to [MASK].)	one way to treat an infection	help	beforehand	**have checkup** (**doctor** is related to **illness**, **illness** makes you want to **have a checkup**.)
	what is saturated fat at room temperature? → (The saturated fat at room temperature is [MASK].)	solid at room temperature	unchanged	negligible	**solid object** (**fat** is related to **butter**, **butter** is a type of **solid object**.)

Table 2: More examples to illustrate retrieving the reasoning paths on ConceptNet can enhance the language model's reasoning ability. where prompting with generated knowledge reduces the reasoning type and rectifies the prediction. We also present the cloze-based prompt example for each question to let masked language models fill the mask.

Case Study. Next, we demonstrate a few examples from both datasets to see how the retrieved reasoning path can help the PLM to make the correct prediction without any few-shot finetuning steps. As shown in Table 2, masked language models RoBERTa and DeBERTa generally cannot predict the answers that fit the semantic meaning in both questions. In addition, the state-of-the-art PLM GPT-3 can generate suitable answers, but the computational overhead of executing GPT-3 is huge, and the autoregressive way to generate answers is also not grounded. As opposed to existing approaches, by reasoning over the external knowledge graph, KEP can generate precise answers and provide a reasoning chain to support the answer choice without any learning steps.

Conclusion and Future Works

We present an off-the-shelf platform KEP to predict answers for open-domain commonsense reasoning. By leveraging the implicit knowledge stored in PLMs and the external knowledge base, the proposed model is able to retrieve relevant reasoning paths of the question. With the zero-shot and open-domain setting, the work poses a new direction to automated commonsense reasoning. In future works, we plan to substitute the RoBERTa used in this work with other PLMs that are finetuned on commonsense reasoning-related tasks as the underlying model. Since the current framework does not incorporate any learning strategy, the purely PLM-driven reasoning path retrieval may not well handle more complex reasoning problems (e.g., commonsense questions with negation). We further plan to leverage learning steps to enhance the model's reasoning capability. Given the uniqueness of the open-domain commonsense reasoning task, we will also perform a variety of experiments on other datasets.

References

Becker, M.; Korfhage, K.; and Frank, A. 2021. COCO-EX: A tool for linking concepts from texts to ConceptNet. In *EACL*.

Bian, N.; Han, X.; Chen, B.; and Sun, L. 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation. In *AAAI*.

Bosselut, A.; Le Bras, R.; and Choi, Y. 2021. Dynamic Neuro-Symbolic Knowledge Graph Construction for Zero-shot Commonsense Question Answering. In *AAAI*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *ICLR*.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How can we know what language models know? *TACL*, 8: 423–438.

Kale, M.; and Rastogi, A. 2020. Text-to-text pre-training for data-to-text tasks. *arXiv preprint*.

Khot, T.; Clark, P.; Guerquin, M.; Jansen, P.; and Sabharwal, A. 2020. Qasc: A dataset for question answering via sentence composition. In *AAAI*.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint*.

Liu, J.; Liu, A.; Lu, X.; Welleck, S.; West, P.; Bras, R. L.; Choi, Y.; and Hajishirzi, H. 2021. Generated knowledge prompting for commonsense reasoning. *arXiv preprint*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*.

Ma, K.; Ilievski, F.; Francis, J.; Bisk, Y.; Nyberg, E.; and Oltramari, A. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *AAAI*.

Paranjape, B.; Michael, J.; Ghazvininejad, M.; Zettlemoyer, L.; and Hajishirzi, H. 2021. Prompting contrastive explanations for commonsense reasoning tasks. *arXiv preprint*.

Shwartz, V.; West, P.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2020. Unsupervised commonsense question answering with self-talk. *arXiv preprint*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *NAACL*.

Yasunaga, M.; Ren, H.; Bosselut, A.; Liang, P.; and Leskovec, J. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint*.

Zhou, X.; Zhang, Y.; Cui, L.; and Huang, D. 2020. Evaluating commonsense in pre-trained language models. In *AAAI*.