ChatGPT is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models

Ning Bian^{1,2}, **Xianpei Han**^{1,2,3,*}, **Le Sun**^{1,2,3}, **Hongyu Lin**^{2}, **Yaojie Lu**^{2}, **Ben He**^{1,2},
Shanshan Jiang^{4}, **Bin Dong**^{4}
^{1}University of Chinese Academy of Sciences, Beijing, China
^{2}Chinese Information Processing Laboratory ^{3}State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
^{4}Ricoh Software Research Center (Beijing) Co., Ltd., Beijing, China
`bianning21@mails.ucas.ac.cn {xianpei,sunle,hongyu,yaojie}@iscas.ac.cn`
`benhe@ucas.ac.cn {shanshan.jiang, bin.dong}@cn.ricoh.com`

Abstract

Large language models (LLMs) have made significant progress in NLP. However, their ability to memorize, represent, and leverage commonsense knowledge has been a well-known pain point. In this paper, we specifically focus on ChatGPT, a widely used and easily accessible LLM, and ask the following questions: (1) Can ChatGPT effectively answer commonsense questions? (2) Is ChatGPT aware of the underlying commonsense knowledge for answering a specific question? (3) Is ChatGPT knowledgeable in commonsense? (4) Can ChatGPT effectively leverage commonsense for answering questions? We conduct a series of experiments on 11 datasets to evaluate ChatGPT's commonsense abilities, including answering commonsense questions, identifying necessary knowledge, generating knowledge descriptions, and using knowledge descriptions to answer questions again. Experimental results show that: (1) ChatGPT can achieve good QA accuracies in commonsense tasks, while still struggling with certain domains of datasets. (2) ChatGPT is knowledgeable, and can accurately generate most of the commonsense knowledge using knowledge prompts. (3) Despite its knowledge, ChatGPT is an inexperienced commonsense problem solver, which cannot precisely identify the needed commonsense for answering a specific question. These findings raise the need to explore improved mechanisms for effectively incorporating commonsense into LLMs like ChatGPT, such as better instruction following and commonsense guidance.

Keywords: Large Language Model, Commonsense Knowledge, Question Answering

1. Introduction

Commonsense knowledge is a foundational aspect of human cognition, encompassing our innate understanding of the world and our capacity to reason within it. It includes knowledge about the spatial, physical, social, temporal, and psychological aspects of the typical everyday life, as well as an awareness of social norms, beliefs, and values (Liu and Singh, 2004; Davis, 2023). The integration of commonsense knowledge is essential for developing NLP systems that can comprehend and generate language similar to humans. However, acquiring and representing commonsense in machines has posed a long-standing challenge (Li et al., 2021; Zhang et al., 2022; Zhou et al., 2023), primarily due to the implicit and context-dependent nature of commonsense (Gordon and Van Durme, 2013; Shwartz and Choi, 2020). In recent years, there has been a growing interest in addressing the commonsense problem within NLP models, with the aim of enabling more human-like language generation and understanding (Bauer et al., 2018; Wang et al., 2020; Jiang et al., 2021; Liu et al., 2021; Sun et al., 2022; Liu et al., 2022; He et al., 2023; Cui and Chen, 2023; Chen et al., 2023).

Recently, large language models (LLMs) such as ChatGPT have made substantial advancements in a wide range of NLP tasks, including inference, contextual understanding (Tang et al., 2023), and chain-of-thought reasoning (Wei et al., 2022). These achievements suggest that LLMs possess a certain degree of commonsense knowledge (West et al., 2022; Gu et al., 2022; Zhao et al., 2023). However, the challenge of commonsense still remains a significant limitation for LLMs (Zhou et al., 2020; Li et al., 2022; Bhargava and Ng, 2022; Bang et al., 2023; Kondo et al., 2023). Despite their increasing abilities, the extent of LLMs' understanding and reasoning capabilities regarding commonsense knowledge remains unclear.

In this paper, we specifically focus on ChatGPT to evaluate the commonsense abilities in LLMs. ChatGPT is a prominent and widely used representative of LLMs, due to its high performance and ease of access. We pose the following key questions on the commonsense abilities of ChatGPT: (1) *Can ChatGPT effectively answer commonsense questions?* (2) *Is ChatGPT aware of the underlying commonsense knowledge for answering a specific question?* (3) *Is ChatGPT knowledgeable in com-*

---

* Corresponding Author

LREC-COLING 2024, pages 3098–3110
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

monsense? (4) *Can ChatGPT effectively leverage commonsense for answering questions?* Answering these questions is crucial for understanding the capabilities and limitations of LLMs and for developing better methods to evaluate and improve their performance on commonsense tasks. To do so, we employ 11 commonsense QA datasets that cover a wide range of 8 commonsense domains, including physical, social, temporal, and numerical reasoning, etc. Our evaluation methodology consists of four key steps. First, we present commonsense questions to the GPT-3, Instruct GPT (text-davinci-003), and ChatGPT, and assess the accuracy of their responses. This step helps us gauge the models' ability to accurately answer commonsense questions. Next, we investigate whether ChatGPT possesses an understanding of the underlying commonsense knowledge necessary for answering these questions. We prompt the models to describe the required knowledge and evaluate the accuracy and appropriateness of their descriptions. Finally, we explore the models' capacity to leverage commonsense knowledge for reasoning. We utilize the knowledge generated in the previous experiments as context and ask the models to answer the questions again, in order to evaluate whether the models can effectively leverage the identified knowledge in their reasoning process. We further compare their performance using "golden" knowledge.

Our experiments provide insights into the commonsense problem of ChatGPT: (1) ChatGPT and Instruct GPT can achieve good QA accuracies on commonsense tasks, while still struggling with certain domains of datasets. (2) ChatGPT is knowledgeable and can accurately generate most of the commonsense knowledge using knowledge prompts. (3) ChatGPT is an inexperienced commonsense problem solver, which cannot precisely identify the needed commonsense knowledge for solving a specific question. Furthermore, ChatGPT cannot effectively leverage knowledge in context for answering questions.

The main contributions of this paper are:

- We investigate the commonsense ability of ChatGPT in detail by conducting experiments to answer four key questions.
- We design a series of experiments to evaluate ChatGPT's ability to memorize, represent and leverage commonsense knowledge, including answering commonsense questions, identifying and generating necessary knowledge, and leveraging commonsense knowledge for reasoning.
- By identifying the strengths and weaknesses of ChatGPT's ability in commonsense knowledge and reasoning, we provide insights into the development of more advanced language models that can effectively leverage and reason about commonsense knowledge.

2. What is Commonsense?

Commonsense knowledge is "a huge portion of human experience, encompassing knowledge about the spatial, physical, social, temporal, and psychological aspects of typical everyday life" (Liu and Singh, 2004; Brachman and Levesque, 2022). This type of knowledge is often taken for granted and is typically acquired through years of experience and socialization within a particular culture.

To establish the necessary background and preliminary concepts for our study, we summarize several main categories of commonsense as following: **General commonsense** refers to knowledge that is widely shared and assumed to be true by most people, such as *the sun rises in the east and sets in the west*. **Physical commonsense** involves intuitive knowledge about the physical world, such as *objects fall to the ground when dropped* and *water flows downhill*. **Social commonsense** involves knowledge about social norms, customs, and practices, such as *it is polite to say "thank you" when making requests*. **Science commonsense** involves knowledge about basic scientific principles, such as *gravity pulls all objects on Earth to Earth's center*. **Event commonsense** involves knowledge about the sequence of events and the causal relationships between them, such as *if a glass is knocked over, the liquid inside will spill*. **Numerical commonsense** involves knowledge about numbers, such as *human has two hands and ten fingers*. **Prototypical commonsense** involves knowledge about typical or prototypical examples of concepts, such as *a swallow is a kind of bird, and a bird has wings*. **Temporal commonsense** involves knowledge about time, such as *traveling abroad requires a longer time than taking a walk*.

3. Can ChatGPT Effectively Answer Commonsense Questions?

In this section, we evaluate the performance of ChatGPT to answer commonsense questions. We use 11 commonsense QA datasets covering 8 commonsense domains, including general, physical, social, science, event, numerical, prototypical, and temporal knowledge. The 11 datasets are CommonsenseQA (Talmor et al., 2019), OpenBookQA (Mihaylov et al., 2018), WSC (Levesque et al., 2012), PIQA (Bisk et al., 2020), Social IQA (Sap et al., 2019), ARC (Easy set) (Clark et al., 2018), QASC (Khot et al., 2020), HellaSWAG (Zellers et al., 2019), NumerSense (Lin et al., 2020), ProtoQA (Boratko et al., 2020), and MC-TACO (Zhou et al.,

Dataset	Domain	Example (Bold texts are the answers)
CommonsenseQA	General	Choose your answer to the question: Where are you likely to find a hamburger? **A. fast food restaurant**, B. pizza, C. ground up dead cows, D. mouth, E. cow circus
OpenBookQA	General	Choose your answer to the question: If a person walks in the opposite direction of a compass arrow they are walking A. west, B. north, C. east, **D. south**
WSC	General	Choose subsentence A or B that completes the sentence: The trophy doesn't fit into the brown suitcase because A. the trophy is too small. **B. the suitcase is too small.**
PIQA	Physical	Choose one that is correct: **A. ice box will turn into a cooler if you add water to it.** B. ice box will turn into a cooler if you add soda to it.
Social IQA	Social	Taylor taught math in the schools after studying to be a teacher. Choose the most suitable answer for the question: What does Taylor need to do before this? **A. get a certificate**, B. teach small children, C. work in a school
ARC	Science	Choose your answer to the question: Which technology was developed most recently? **A. cellular telephone**, B. television, C. refrigerator, D. airplane
QASC	Science	Choose your answer to the question: What is described in terms of temperature and water in the air? A. storms; **B. climate**; C. mass; D. seasonal; E. winter; F. density; G. length
HellaSWAG	Event	Choose your answer to the question: We see a chair with a pillow on it. A. a man holding a cat does curling. B. a man holding a cat starts hitting objects on an item. C. a man holding a cat is wrapping a box. **D. a man holding a cat sits down on the chair.**
NumerSense	Numerical	a square is a shape with <mask> equally lengthed sides. (**four**)
ProtoQA	Prototypical	Use simple words separated by commas to name something in your life that could cause you to lose weight. (**Eating less, exercising more, stress.**)
MC-TACO	Temporal	Select all feasible answers for the question: Carl Laemmle, head of Universal Studios, gave Einstein a tour of his studio and introduced him to Chaplin. At what time did Einstein return home? **A. 8:00 PM**; B. a second later; **C. a hour later**

Table 1: Domains and examples of the commonsense QA datasets used in this paper.

2019). The datasets, domains, and an example for each dataset are shown in Table 1.

We sample 100 questions from the development set of each dataset, except for ProtoQA, which has only 52 questions in its development set. We use GPT-3 (davinci, Brown et al., 2020), Instruct GPT (text-davinci-003), and ChatGPT (we use the "GPT-3.5" web interface[1]) for evaluation. For GPT-3, we use 4-shot in-context learning, as GPT-3 cannot effectively answer questions in a zero-shot setting. For Instruct GPT and ChatGPT, we use zero-shot inference and design prompt templates for different datasets (shown in Table 1).

From results in Table 2, we can see that:

Instruct GPT and ChatGPT demonstrate high accuracy in answering commonsense questions. We evaluate the performances of different LLMs on 11 commonsense QA datasets. The results in Table 2 show that both Instruct GPT and ChatGPT achieve good performance across most datasets. Notably, ChatGPT achieves the highest accuracy of 94% on the ARC dataset and 94.2% on the ProtoQA dataset. This indicates that ChatGPT is capable of accurately answering various types of commonsense questions. However, there are still large accuracy gaps between models and humans, as shown in Table 2.

The ability of models to leverage commonsense is probably improved by instruction tuning and human alignment. A notable observation from the results in Table 2 is the significant improvement achieved by Instruct GPT and ChatGPT compared to GPT-3. This improvement is probably due to the incorporation of instruction tuning and human alignment during training (Ouyang et al., 2022). In addition to the pre-training, these techniques may enable the models to better leverage and reason with commonsense knowledge, demonstrating the importance of instruction and alignment in enhancing the models' performance.

Overall, ChatGPT achieves higher accuracy than Instruct GPT in most domains, demonstrating the effectiveness of the RLHF technique in enhancing knowledge-leveraging ability. However, Instruct GPT slightly outperforms ChatGPT on certain datasets including CommonsenseQA (p = 0.238 by T-test) and Social IQA (p = 0.179). This is because ChatGPT tends to be cautious and refuses to answer questions when information is insufficient, resulting in outputs like "Based on the information

---

[1] chat.openai.com

Dataset	Domain	GPT-3	Instruct GPT	ChatGPT	Human
CommonsenseQA	General	38.0	**81.0**	74.0	88.9
OpenBookQA	General	22.0	65.0	**73.0**	89.3
WSC	General	46.0	**78.0**	**78.0**	92.1
PIQA	Physical	48.0	77.0	**78.0**	94.5
Social IQA	Social	36.0	**71.0**	62.0	86.9
ARC	Science	27.0	88.0	**94.0**	–
QASC	Science	25.0	**75.0**	74.0	93.0
HellaSWAG	Event	19.0	61.0	**67.0**	95.7
NumerSense	Numerical	45.0	63.0	**79.0**	89.7
ProtoQA	Prototypical	67.3	84.6	**94.2**	–
MC-TACO	Temporal	20.0	**53.0**	52.0	75.8

Table 2: Evaluation results (% accuracy) of large language models on commonsense QA datasets. Human accuracies are adopted from the original papers of these datasets except WSC by Bender (2015).

given, it is not possible to determine ...". These outputs are considered incorrect because they indicate an inability to generate a direct and correct answer. In CommonsenseQA, ChatGPT produces 4 such responses out of 100 questions, while in Social IQA, there are 13 such instances out of 100. This highlights the need for further research on how to balance the caution and accuracy in models where there is insufficient information. Achieving this requires the model to understand the necessary knowledge to answer the question and be aware of whether the model possesses that knowledge.

Although ChatGPT performs well on most commonsense knowledge domains, they still struggle with certain domains. The experiments in Table 2 show that ChatGPT lags behind on datasets regarding social, event, and temporal (Social IQA, HellaSWAG, and MC-TACO datasets), with the ChatGPT's performances below 70%. This shows that ChatGPT still has drawbacks on the social, event, and temporal domains of commonsense QA, which is consistent with Chang and Bergen (2023). We believe this is because these domains of commonsense require a deeper understanding of human behavior and social interactions, and they appear infrequently in text corpora. ChatGPT needs to go beyond superficial semantic understanding and learn about human behaviors to better learn these domains of commonsense.

4. Is ChatGPT Aware of the Commonsense needed for QA?

In Section 3, we found that ChatGPT performs well on commonsense QA datasets. This intrigues us to explore whether ChatGPT is experienced experts that are aware of what knowledge is needed and can leverage the knowledge for question answering, or if they are inexperienced problem solvers that rely on memorizing a large amount of information that covers the answers.

Question: How to make outdoor pillow?
A. Blow into *tin can* and tie with rubber band. ✗
B. Blow into *trash bag* and tie with rubber band. ✓

ChatGPT Response:
1. Understanding of what an outdoor pillow is. ✓
2. Knowledge of how to make an outdoor pillow.
3. Familiarity with the materials required to make an outdoor pillow. ✓
4. Understanding of how to inflate a container such as a tin can or trash bag.
5. Knowledge of how to tie a rubber band.

Human Evaluation:
1. Understanding of what an outdoor pillow is.
3. Familiarity with the materials required to make an outdoor pillow.
a. Knowing that tin can is hard and trash bag is soft.
b. Knowing that human prefers soft pillows.

Table 3: An example of necessary knowledge generated by ChatGPT and human evaluation. The question is from the PIQA dataset.

To answer this question, we sample 20 questions from each commonsense QA dataset and ask ChatGPT "What knowledge is necessary for answering this question? {question} {answer choices (if applicable)}". We sample 10 correctly and 10 incorrectly answered questions for each dataset to ensure a fair comparison. In cases where there are insufficient incorrectly answered questions (specifically, there are 6 for ARC and 3 for ProtoQA), we take all incorrectly answered questions and sample more correctly answered questions to fill up the 20 questions. In total, ChatGPT identified 855 pieces of knowledge as necessary for answering these questions, with an average of 3.9 pieces of knowledge per question.

Human annotators with a solid understanding of commonsense reasoning are employed to manually evaluate the precision and recall of each generated piece of knowledge. Precision refers to the proportion of relevant knowledge correctly included in the response, while recall refers to the extent to which the necessary knowledge is appropriately

Dataset	Domain	Correct	Wrong	Overall
CommonsenseQA	General	65.83 / 94.17 / 75.86	50.00 / 72.50 / 57.79	57.92 / 83.33 / 66.82
OpenBookQA	General	80.50 / 100.00 / 87.94	35.83 / 55.83 / 42.81	58.17 / 77.92 / 65.37
WSC	General	80.00 / 87.50 / 83.21	57.50 / 83.33 / 65.90	68.75 / 85.12 / 74.56
PIQA	Physical	60.00 / 80.00 / 64.90	53.36 / 88.33 / 63.25	56.78 / 84.17 / 64.08
Social IQA	Social	53.00 / 90.00 / 63.43	28.17 / 40.00 / 32.05	40.58 / 65.00 / 47.74
ARC	Science	73.57 / 100.00 / 82.80	45.00 / 83.33 / 55.36	65.00 / 95.00 / 74.57
QASC	Science	67.17 / 100.00 / 78.79	68.33 / 88.33 / 73.48	67.75 / 94.17 / 76.13
HellaSWAG	Event	64.00 / 95.00 / 74.10	47.55 / 73.00 / 57.31	55.77 / 84.00 / 65.70
NumerSense	Numerical	44.00 / 95.00 / 58.29	44.00 / 89.17 / 58.21	44.00 / 92.08 / 58.25
ProtoQA	Prototypical	65.88 / 98.04 / 76.96	48.33 / 88.89 / 58.73	63.25 / 96.67 / 74.23
MC-TACO	Temporal	47.50 / 80.00 / 58.00	26.17 / 61.67 / 35.57	36.83 / 70.83 / 46.79

Table 4: Precision / Recall / F1 scores of ChatGPT generated necessary knowledge for correct- and wrong-answered questions.

covered. To synthesize the precision and recall scores into a single metric, we calculate the F1 score for each response. The annotators are provided with the question, the model's response, and the corresponding answer choices (if applicable). They are guided by predefined criteria for evaluating the responses. Specifically, they first assess whether each piece of knowledge is necessary for answering the question. Then, they judge whether the question is answerable based on reasoning upon the labeled necessary knowledge. If the question is still unanswerable, the annotators need to fill in the missing knowledge to answer the question.

For example, Table 3 shows a response of ChatGPT that outlines the necessary knowledge for answering a specific question. During the manual evaluation, expert annotators assess the usefulness of each piece of knowledge. In this example, knowledge pieces 1 and 3 are labeled as relevant and useful for answering the question accurately. Knowledge 2 is identified as overgeneralized, as it merely repeats information already presented in the question without providing additional insights and details. Knowledge 4 and 5 are labeled as unnecessary for answering the question, as they cannot contribute to distinguishing between answer options. After this assessment, the annotators determine whether the question can be answered based on the two identified pieces of useful knowledge (1 & 3). However, it is found that additional knowledge is still needed to provide a correct answer. Consequently, two additional pieces of knowledge, a and b, are added to supplement the reasoning process. Thus, the precision and recall for this response are 2/5 and 2/4, so the F1 score is 44.44%. The results are shown in Table 4. We can see that:

ChatGPT is an inexperienced problem solver, which struggles to precisely identify the necessary commonsense knowledge to answer a specific question. As shown in Table 4, the overall F1 scores range from 45% to about 75% on most of the commonsense QA datasets. These results demonstrate that ChatGPT is an inexperienced problem solver and cannot accurately identify the necessary knowledge to answer a specific commonsense question.

Specifically, the model performs relatively well in the science domain, achieving 74.57% F1 on ARC and 76.13% on QASC. However, the model exhibits the lowest performances on social and temporal domains, i.e., Social IQA and MC-TACO, which is consistent with the results in Section 3. This discrepancy in F1 scores is likely because scientific commonsense knowledge is more prevalent in the text corpus than social and temporal knowledge. For instance, textbooks frequently discuss scientific concepts such as "*climate is described by temperature and humidity*", but rarely mention social commonsense like "*students don't like taking big exams*". This shows a deficiency of LLMs like ChatGPT in social and temporal knowledge, highlighting the importance of developing more effective training strategies to inject these domains of commonsense knowledge.

ChatGPT cannot effectively distinguish between relevant and irrelevant commonsense knowledge for a specific question, which usually generates knowledge with noise. As shown in Table 4, the precisions are significantly higher than the recalls of the generated necessary knowledge ($p < 0.05$ on all datasets except OpenBookQA with $p = 0.09$). The average recall of all datasets is 84.42% and the average precision is 55.88%. This shows that while the model can identify most of the commonsense knowledge for questions, it struggles with accurately identifying which pieces of knowledge are essential and usually generates irrelevant or overgeneralized knowledge. For example, knowledge 2 in Table 3 is overgeneralized because the question itself is "how to make outdoor pillow", and knowledge 4 and 5 are irrelevant given the difference between the two choices. We believe this is because the model relies on keyword and topic matching, rather than a full understanding of the logical relationships within the question. This further demonstrates that ChatGPT is still in-

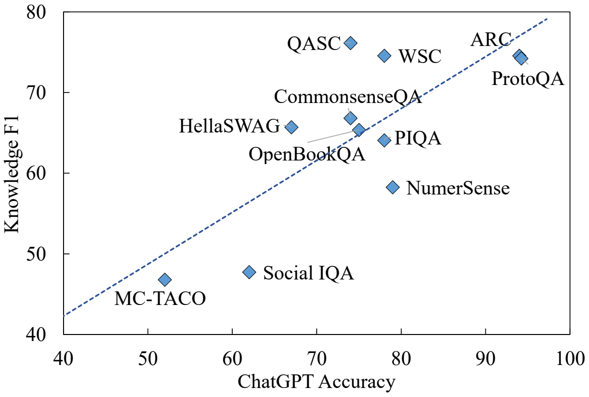

Figure 1: Correlation between necessary knowledge F1 and the answer accuracy of ChatGPT.

experienced problem solvers. Therefore, LLMs need to increase their self-awareness of the task at hand and distinguish crucial information from irrelevant background information. They should determine which parts of a question are necessary for decision-making, such as "blow into tin can/trash bag" in Table 3, and which parts are extraneous.

The commonsense ability of ChatGPT can be improved by enhancing their knowledge-awareness. Figure 1 demonstrates a strong correlation between the overall F1 scores of generated necessary knowledge and the answer accuracies of ChatGPT, with a Pearson coefficient of 0.77 ($p = 0.006$). Furthermore, Table 4 shows that the knowledge F1 scores for correctly answered questions are significantly higher than those for incorrectly answered questions ($p < 0.05$ on OpenBookQA, WSC, Social IQA, ARC, and MC-TACO datasets). These findings suggest that accurately identifying necessary knowledge is crucial for correctly answering commonsense questions. Consequently, enhancing the model's self-awareness of necessary knowledge may improve its performance on downstream tasks including commonsense QA.

5. Is ChatGPT Knowledgeable in Commonsense?

This section answers the question: *To what extent do ChatGPT possess commonsense knowledge?* To answer this question, similar to Shwartz et al. (2020), we manually construct knowledge-querying prompts based on the generated necessary knowledge in Section 4. For example, as shown in Table 5, based on knowledge 1 in Table 3, we ask ChatGPT knowledge-querying questions like "Do you have an understanding of what an outdoor pillow is?" and manually label each generated knowledge description as correct or incorrect. We collect a total of 775 knowledge descriptions for the questions used in the experiments, with an average of 3.5 descriptions per question. The average length of the knowledge descriptions is 136.1 words.

Question 1: Do you have an understanding of what an outdoor pillow is?

ChatGPT Response:

As an AI language model, I have an understanding that an outdoor pillow is a type of pillow that is designed to be used outside, such as on a patio or deck, and is typically made of weather-resistant materials to withstand exposure to the elements.

Regarding the two options you provided, neither option appears to be a conventional method for making an outdoor pillow. In option A, blowing into a tin can and tying it with a rubber band does not create a cushion-like surface suitable for use as a pillow. In option B, blowing into a trash bag and tying it with a rubber band may create a cushion-like surface, but it is unlikely to be durable or comfortable for prolonged use as an outdoor pillow.

Therefore, I would suggest exploring other methods or materials that are specifically designed for making outdoor pillows, such as using weather-resistant fabrics and stuffing materials.

......

Table 5: An example of generated knowledge by ChatGPT. The question is in Table 3.

Table 5 shows an example of a knowledge-querying question and the generated knowledge description. The description says "*blowing into a trash bag and tying it with a rubber band may create a cushion-like surface, but it is unlikely to be durable or comfortable for prolonged use as an outdoor pillow*", but it contradicts with the correct answer. So, this description is labeled as incorrect.

From the results in Table 6 we can see that:

ChatGPT is knowledgeable and contains most of the commonsense knowledge for accurately answering questions. The results in Table 6 show that the generated knowledge descriptions of ChatGPT can achieve over 70% accuracy on most commonsense QA datasets, achieving an average accuracy of 82.66%. This means ChatGPT can generate accurate commonsense knowledge descriptions given knowledge-querying questions. Thus, ChatGPT can serve as commonsense knowledge bases and provide support for downstream tasks. However, the accuracy is low on Social IQA (54.92%). We believe this is because social commonsense, such as "*The person who receives help, rather than gives it, should say thank you*", is not commonly described in texts. This highlights the importance of developing specific approaches to inject social commonsense knowledge into LLMs.

ChatGPT contains misleading and overgeneralized commonsense knowledge. We further conduct a manual evaluation of the relevance and informativeness of the knowledge descriptions. We

Dataset	Correct	Wrong	Overall
CommonsenseQA	100.00	83.83	91.92
OpenBookQA	84.83	100.00	92.42
WSC	90.00	74.17	82.08
PIQA	85.00	62.14	73.57
Social IQA	58.33	51.50	54.92
ARC	91.67	97.62	95.83
QASC	88.33	89.17	88.75
HellaSWAG	80.00	70.83	75.42
NumerSense	85.17	84.50	84.83
ProtoQA	80.29	100.00	83.25
MC-TACO	95.00	77.50	86.25

Table 6: Accuracies (%) of ChatGPT generated knowledge descriptions for correct- and wrong-answered questions.

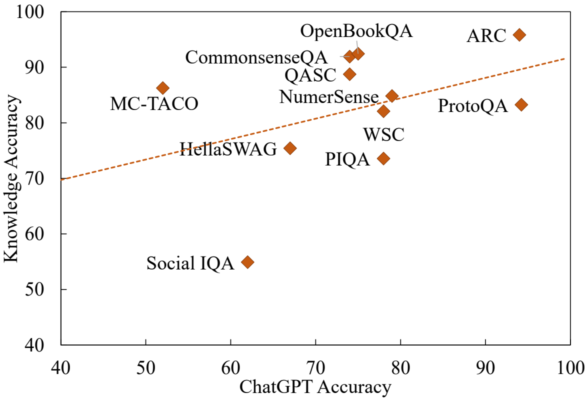

Figure 2: Correlation between generated knowledge accuracy and answer accuracy of ChatGPT.

find that 26.25% of the descriptions include irrelevant and misleading information, and 15.00% of the descriptions are overgeneralized and fail to provide the specific knowledge necessary to answer the question. Overgeneralized knowledge means correct but unhelpful or irrelevant general knowledge for the given questions. For example, the description in Table 5 mentions “exploring other methods or materials that are specifically designed for making outdoor pillows”, which is unhelpful and misleading for answering the question. We believe this is because of noisy and redundant information in the training data of LLMs, which impairs the ability to accurately judge the relevance of information. These findings emphasize the need for instructing LLMs to generate relevant and informative knowledge descriptions that are helpful for QA.

There is a gap between knowing and leveraging commonsense knowledge in ChatGPT. We conduct a correlation analysis between the accuracy of generated knowledge descriptions and answer accuracy, shown in Figure 2. Our result reveals a weak correlation between the two accuracies, with a Pearson coefficient of 0.40 ($p = 0.227$). Notably, while the answer accuracies are low on both the Social IQA and the MC-TACO datasets, there was a significant difference in the accuracy of knowledge descriptions between them: it was low for Social IQA (54.92%) but high for MC-TACO (86.25%). Table 6 further shows that the difference in description accuracy between correctly and incorrectly answered questions is relatively small compared to the results in Table 4 ($p > 0.05$ for all datasets except CommonsenseQA with $p = 0.01$). This shows that a good knowledge description does not necessarily translate to a correct answer. We believe this is because answering commonsense questions not only requires knowledge but also other abilities like reasoning and making inferences under insufficient information conditions.

6. Can ChatGPT Effectively Leverage Commonsense for Reasoning?

This section answers the question: *Can ChatGPT leverage commonsense knowledge in context for reasoning and answering questions?* After answering the knowledge-querying questions in Section 5, we ask the model to answer the commonsense questions again given the generated knowledge descriptions as context, and evaluate whether the answers will change. Specifically, we added these knowledge descriptions before the prompts used in Section 3. The prompts remained the same, and the knowledge descriptions served as additional context to evaluate the impact of knowledge on answer changes. This minimal interference with the prompts aims to facilitate a fair comparison between responses with and without contextual knowledge.

Results in Table 8 show that:

ChatGPT cannot effectively leverage the generated commonsense descriptions if we only add them to the question context. Our analysis of answer changes before and after using knowledge descriptions shows that in most datasets there are no obvious and consistent accuracy improvements given the commonsense descriptions in contexts. Table 7 shows an example that a previously incorrect answer remains unchanged after generating knowledge descriptions. There are both wrong-to-correct changes and correct-to-wrong changes, as well as a significant proportion of unchanged answers, for the example shown in Table 7. In the case of the Social IQA dataset, the accuracy of knowledge generation is low, leading to more correct answers being modified to become wrong. This shows that ChatGPT cannot effectively exploit its own generated knowledge descriptions to answer questions, and the accuracy of the knowledge generation has a huge impact on the answering results. We believe this is because the model already possesses the generated knowledge, thus

Question: How to make outdoor pillow?
A. Blow into *tin can* and tie with rubber band. ✗
B. Blow into *trash bag* and tie with rubber band. ✓

ChatGPT's Answer before Knowledge Description:
Neither option is a good way to make an outdoor pillow. Blowing into a tin can or a trash bag and tying it with a rubber band will not provide adequate support or comfort for an outdoor pillow. Over time, the air will leak out and the pillow will become flat and uncomfortable. A better option for making an outdoor pillow would be to use durable and weather-resistant materials such as outdoor fabric and stuffing or foam designed for outdoor use. Sewing or using fabric glue to attach the fabric together and stuffing it with the appropriate material will result in a more functional and comfortable outdoor pillow.

ChatGPT's Answer after Knowledge Description:
Neither of these options is a suitable way to make an outdoor pillow. Inflating a tin can or a trash bag and tying it with a rubber band would not provide the necessary level of firmness or support for a pillow, and the materials used may not be durable enough to withstand exposure to outdoor elements.
A better option for making an outdoor pillow would be to use weather-resistant fabric and outdoor-rated stuffing or foam inserts, as I mentioned earlier. This will ensure that the pillow is comfortable, durable, and able to withstand exposure to the elements.

Table 7: An example of the answer before and after knowledge descriptions by ChatGPT. The question is from the PIQA dataset.

adding redundant knowledge is not useful.

ChatGPT's performance improvement in commonsense QA is not significant even using golden knowledge. We use two human-annotated commonsense explanation datasets for the CommonsenseQA dataset, CoS-E (Rajani et al., 2019) and ECQA (Aggarwal et al., 2021), as the golden knowledge in context and ask the ChatGPT to generate the answers. We discover that there are only 4/10 wrong → correct answers given CoS-E explanations, and 8/10 wrong → correct answers given ECQA explanations while with 1/10 correct → wrong answer. This shows that ChatGPT cannot answer all questions correctly even given the golden knowledge explanations. We believe this is because ChatGPT lacks the ability to use knowledge for complex commonsense reasoning, such as negation. For example, here is a question that requires reasoning of negation: "*What would not be true about a basketball if it had a hole in it but it did not lose its general shape? A. punctured, B. popular in America, **C. full of air**, D. gone, E. round*". The CoS-E explanation for this question is "*Air cannot stay in any object that has a hole in it.*", but ChatGPT still predicts the wrong answer A and explains "*If a basketball has a hole in it, it is punctured and air can escape from it.*". These results suggest that LLMs require further guidance and improvement to better leverage and reason about commonsense knowledge in context.

Dataset	C → W	W → C
CommonsenseQA	1/10	2/10
OpenBookQA	0/10	5/10
WSC	2/10	1/10
PIQA	1/10	1/10
Social IQA	5/10	1/10
ARC	0/14	2/6
QASC	1/10	4/10
HellaSWAG	3/10	4/10
NumerSense	4/10	4/10
ProtoQA	3/17	2/3
MC-TACO	4/10	2/10

Table 8: Numbers of changed/total answers after the generation of knowledge descriptions. C → W means a correct answer changes to a wrong answer, and W → C means a wrong answer changes to a correct answer.

7. Related Work

Recent studies have shown that LLMs such as GPT-3, ChatGPT, and GPT-4 have made significant progress in various NLP tasks, including QA, text generation, and translation (Brown et al., 2020). However, there is a growing concern about their ability to understand and reason about commonsense knowledge (Tamborrino et al., 2020; Cui et al., 2021; Bhargava and Ng, 2022). Recent studies have focused on evaluating the ability of LLMs to understand commonsense knowledge (Davison et al., 2019; Liu et al., 2020; Niu et al., 2021; Ma et al., 2021; Klein and Nabi, 2021; Porada et al., 2022; Laskar et al., 2023). For example, Zhou et al. (2020) evaluates several LLMs on a set of commonsense reasoning tasks and found that they have a certain degree of commonsense knowledge, but there is still a gap between models and humans. Wang et al. (2021) studies the generalizability of models for commonsense inference and found that the ability relies heavily on whether the objects to predict are seen during training. Cohn and Hernandez-Orallo (2023) conduct qualitative investigations of the spatial commonsense reasoning ability of ChatGPT and Bard. In this paper, we evaluate the commonsense abilities of ChatGPT including answering commonsense questions, identifying and generating necessary knowledge, and leveraging knowledge for reasoning.

8. Conclusions and Discussions

In this paper, we investigate the commonsense abilities of ChatGPT and found that ChatGPT is a knowledgeable but inexperienced problem solver: (1) While ChatGPT can achieve good accuracies in commonsense QA, it still struggles with certain domains of QA, including social and temporal commonsense. (2) ChatGPT is knowledgeable in commonsense, which can accurately generate most of the commonsense knowledge using knowledge prompts. (3) ChatGPT is an inexperienced commonsense problem solver. It struggles to precisely identify the underlying commonsense knowledge for a given question and often generates knowledge with a high noise rate. Furthermore, ChatGPT cannot effectively leverage commonsense knowledge in contexts to answer commonsense questions.

The above findings raise several promising directions for the future of LLMs:

(1) Although current ChatGPT is knowledgeable, they are still not experienced problem solvers. Therefore, it is critical to investigate better mechanisms for utilizing commonsense knowledge in LLMs, such as instruction tuning, better commonsense-guided reasoning, etc.

(2) There are still several types of commonsense knowledge missing in LLMs, such as social and temporal commonsense. Therefore it is critical to design knowledge injection approaches for these knowledge types. Furthermore, it is important to design lightweight commonsense updating methods to keep the knowledge up-to-date.

(3) Because ChatGPT does not release its full details, such as training data, hyper-parameters, and checkpoints, and evaluating an “artificial general intelligence” model is very difficult, it is crucial to construct benchmarks with wider coverage, and design evaluation methods that provide a more comprehensive and unbiased assessment of LLMs.

9. Limitations

This study provides valuable insights into the commonsense abilities of ChatGPT, but there are several limitations that could be acknowledged and addressed in future research.

Human evaluations of LLMs’ commonsense performance and abilities, such as answer accuracy and the F1 score and accuracy of generated necessary knowledge, are labor-intensive and time-consuming. The manual analysis in this paper required approximately 80 to 100 human-hours in total. Additionally, it can be difficult even for humans to clarify which pieces of commonsense knowledge are necessary for answering a specific question, as commonsense knowledge is often implicit and automatic for humans (Ellis, 2008). Future studies should explore automated methods for evaluating LLMs’ performance and assessing their commonsense abilities. For example, researchers could develop methods that leverage knowledge retrieval or knowledge graphs to evaluate the generated knowledge of LLMs.

Our evaluations of ChatGPT use a small number of sampled commonsense questions on each dataset. While this approach allows for a comprehensive analysis of the commonsense abilities of ChatGPT on different domains of commonsense questions without overwhelming the annotators, it is important to consider that the accuracies may be slightly influenced by the specific question sets sampled randomly. Future studies could expand the number of questions to provide a more comprehensive evaluation.

This study specifically focuses on ChatGPT and does not explore other LLMs like GPT-4, LLaMA (Touvron et al., 2023) and Google’s Bard (Thoppilan et al., 2022). We choose ChatGPT in this study to achieve a good balance between popularity, availability, and cost. It would be interesting for future research to explore whether similar findings hold true for these models and to compare their performance against ChatGPT.

10. Acknowledgements

This work is supported by the Natural Science Foundation of China (No. 62122077 and 62106251).

11. 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.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv preprint*, abs/2302.04023.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *MAICS*, pages 39–45.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. 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 12317–12325. AAAI Press.

Yonatan Bisk, Rowan Zellers, Ronan LeBras, 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.

Michael Boratko, Xiang Li, Tim O'Gorman, Rajarshi Das, Dan Le, and Andrew McCallum. 2020. ProtoQA: A question answering dataset for prototypical common-sense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1122–1136, Online. Association for Computational Linguistics.

Ronald J. Brachman and Hector J. Levesque. 2022. Toward a new science of common sense. 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 12245–12249. 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*.

Tyler A Chang and Benjamin K Bergen. 2023. Language model behavior: A comprehensive survey. *ArXiv preprint*, abs/2303.11504.

Qianglong Chen, Guohai Xu, Ming Yan, Ji Zhang, Fei Huang, Luo Si, and Yin Zhang. 2023. Distinguish before answer: Generating contrastive explanation as knowledge for commonsense question answering. *ArXiv preprint*, abs/2305.08135.

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

Anthony G Cohn and Jose Hernandez-Orallo. 2023. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of llms. *ArXiv preprint*, abs/2304.11164.

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2021. On commonsense cues in BERT for solving commonsense tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 683–693, Online. Association for Computational Linguistics.

Wanyun Cui and Xingran Chen. 2023. Free lunch for efficient textual commonsense integration in language models. *ArXiv preprint*, abs/2305.15516.

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

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Nick C Ellis. 2008. Implicit and explicit knowledge about language. *Encyclopedia of language and education*, 6:1–13.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Yuling Gu, Bhavana Dalvi Mishra, and Peter Clark. 2022. Do language models have coherent mental models of everyday things? *ArXiv preprint*, abs/2212.10029.

Jie He, Víctor Gutiérrez-Basulto, Jeff Z Pan, et al. 2023. Buca: A binary classification approach to unsupervised commonsense question answering. *ArXiv preprint*, abs/2305.15932.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online. 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.

Tassilo Klein and Moin Nabi. 2021. Towards zero-shot commonsense reasoning with self-supervised refinement of language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8737–8743, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kazushi Kondo, Saku Sugawara, and Akiko Aizawa. 2023. Commonsense knowledge transfer for pre-trained language models. *ArXiv preprint*, abs/2306.02258.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. *ArXiv preprint*, abs/2305.18486.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, pages 552–561.

Xiang Lorraine Li, Adhiguna Kuncoro, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2021. Do language models learn commonsense knowledge? *ArXiv preprint*, abs/2111.00607.

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, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22:211–226.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. 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.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and Philip S. Yu. 2021. KG-BART: knowledge graph-augmented BART for generative commonsense reasoning. 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 6418–6425. AAAI Press.

Ye Liu, Tao Yang, Zeyu You, Wei Fan, and Philip S. Yu. 2020. Commonsense evidence generation and injection in reading comprehension. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 61–73, 1st virtual meeting. Association for Computational Linguistics.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Satoru Ozaki, Eric Nyberg, and Alessandro Oltramari. 2021. Exploring strategies for generalizable commonsense reasoning with pre-trained models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5474–5483, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Yilin Niu, Fei Huang, Jiaming Liang, Wenkai Chen, Xiaoyan Zhu, and Minlie Huang. 2021. A semantic-based method for unsupervised commonsense question answering. 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 3037–3049, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155.

Ian Porada, Alessandro Sordoni, and Jackie Cheung. 2022. Does pre-training induce systematic inference? how masked language models acquire commonsense knowledge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4550–4557, Seattle, United States. Association for Computational Linguistics.

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.

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.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on 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.

Yueqing Sun, Yu Zhang, Le Qi, and Qi Shi. 2022. TSGP: Two-stage generative prompting for unsupervised commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 968–980, Abu Dhabi, United Arab Emirates. 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, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pre-training is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Online. Association for Computational Linguistics.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners. *ArXiv preprint*, abs/2305.14825.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239.

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

Peifeng Wang, Filip Ilievski, Muhao Chen, and Xiang Ren. 2021. Do language models perform generalizable commonsense inference? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3681–3688, Online. 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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv preprint*, abs/2201.11903.

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.

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 Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yi Zhang, Lei Li, Yunfang Wu, Qi Su, and Xu Sun. 2022. Alleviating the knowledge-language inconsistency: A study for deep commonsense knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:594–604.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2023. Large language models as commonsense knowledge for large-scale task planning. *ArXiv preprint*, abs/2305.14078.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “going on a vacation” takes longer than “going for a walk”: A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Wangchunshu Zhou, Ronan Le Bras, and Yejin Choi. 2023. Commonsense knowledge transfer for pre-trained language models. *ArXiv preprint*, abs/2306.02388.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. 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 9733–9740. AAAI Press.