Rule or Story, Which is a Better Commonsense Expression for Talking with Large Language Models?

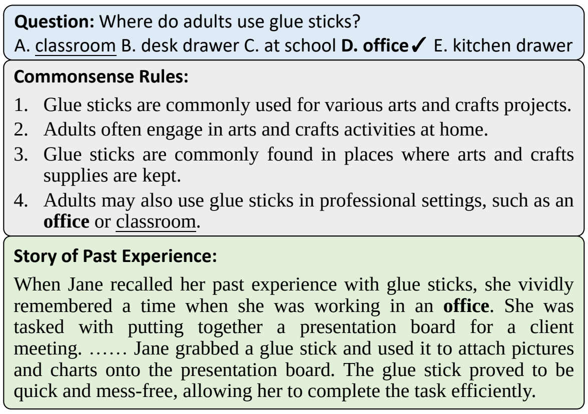
Anonymous ACL submission

Abstract

Building machines with commonsense has been a longstanding challenge in NLP due to the reporting bias of commonsense rules and the exposure bias of rule-based commonsense reasoning. In contrast, humans convey and pass down commonsense implicitly through stories. This paper investigates the inherent commonsense ability of large language models (LLMs) expressed through storytelling. We systematically investigate and compare stories and rules for retrieving and leveraging commonsense in LLMs. Experimental results on 28 commonsense QA datasets show that stories outperform rules as the expression for retrieving commonsense from LLMs, exhibiting higher generation confidence and commonsense accuracy. Moreover, stories are the more effective commonsense expression for answering questions regarding daily events, while rules are more effective for scientific questions. This aligns with the reporting bias of commonsense in text corpora. We further show that the correctness and relevance of commonsense stories can be further improved via iterative self-supervised fine-tuning. These findings emphasize the importance of using appropriate language to express, retrieve, and leverage commonsense for LLMs, highlighting a promising direction for better exploiting their commonsense abilities.

1 Introduction

Building machines with commonsense has been a longstanding goal in AI and NLP (McCarthy, 1959; Brachman and Levesque, 2023). Despite advancements in large language models (LLMs), incorporating commonsense knowledge in these models remains a significant challenge (Ismayilzada et al., 2023; Bian et al., 2023; Li et al., 2022), due to the reporting bias of commonsense knowledge and the exposure bias of commonsense reasoning (Gordon and Van Durme, 2013; Shwartz and Choi, 2020). The reporting bias arises because many aspects of commonsense are rarely stated explicitly in language. For example, "*A person is late*" may appear more frequently than "*A person arrives on time*" in text corpora (Gordon and Van Durme, 2013). Furthermore, commonsense rules are often left implicit and omitted in human language reasoning, leading to exposure bias. For example, the commonsense rule "humans need air to breathe" is usually ignored in cases like "*The room was getting too stuffy, and I opened the windows*" as it is commonly known.



Question: Where do adults use glue sticks?
A. classroom B. desk drawer C. at school **D. office**✓ E. kitchen drawer

Commonsense Rules:

1. Glue sticks are commonly used for various arts and crafts projects.
2. Adults often engage in arts and crafts activities at home.
3. Glue sticks are commonly found in places where arts and crafts supplies are kept.
4. Adults may also use glue sticks in professional settings, such as an **office** or classroom.

Story of Past Experience:

When Jane recalled her past experience with glue sticks, she vividly remembered a time when she was working in an **office**. She was tasked with putting together a presentation board for a client meeting. …… Jane grabbed a glue stick and used it to attach pictures and charts onto the presentation board. The glue stick proved to be quick and mess-free, allowing her to complete the task efficiently.

Figure 1: Comparison between rules and a story written by ChatGPT. The rules only provide useful knowledge until the 4^{th} rule and also include an incorrect answer option, "classroom". The story presents a detailed scenario where an adult uses glue sticks in an office.

To enhance the commonsense ability of NLP models, current studies usually express commonsense as rules. For instance, commonsense rules structured as knowledge graphs of concepts and events (Ilievski et al., 2021; Hwang et al., 2021; Sap et al., 2019a; Speer et al., 2017) are incorporated to support rule-based logical reasoning (Zhang et al., 2023b; Wang et al., 2023c,d,e,f; Liu et al., 2023c). Recently, as studies reveal that LLMs like GPT-3 (Brown et al., 2020) and ChatGPT (OpenAI, 2022) have already learned abundant commonsense (Shwartz et al., 2020), there is a current trend to extract commonsense knowledge from the models' memory, also expressed as rules like in Figure 1, and enhance LLMs by reintegrating this knowledge into the models (Liu et al., 2023a; Yao et al., 2023; Liu et al., 2022b; West et al., 2022).

However, commonsense is more than just rules (Brachman and Levesque, 2022). Humans acquire commonsense by recognizing prototypical patterns, extracting memories of similar past experiences, and contrasting them with the current novel situation to make decisions, as supported by psychological studies (Schacter and Addis, 2007; Klein, 2004; Tulving, 2002; Schank, 1983; Schank and Abelson, 1977). Our commonsense is often conveyed and passed down through stories such as myths and fairy tales (Cassirer et al., 1946), with only a limited portion expressed in rules. Renowned AI theorist and cognitive psychologist Roger Schank argues in his book "*Tell Me a Story: Narrative and Intelligence*" that "*knowledge is stories*" (Schank, 1995). He emphasizes that humans struggle to learn and remember abstract rules derived from past experiences but can more easily remember a good story, because "*stories give life to past experience*".

As a result, human-written text corpora mainly convey commonsense through stories, with limited instances of explicit rules and logical reasoning. In this way, models trained on these corpora acquire commonsense and reasoning abilities implicitly. Studies show that LLMs exhibit a strong storytelling ability, generating narratives that adhere to real-world logic (Bhandari and Brennan, 2023; Eldan and Li, 2023; Wen et al., 2023; Jiayang et al., 2023). However, these models may not effectively learn commonsense rules and explicit reasoning through mimicking human behaviors, as shown by recent studies (Bian et al., 2023; Li et al., 2022).

These observations lead to a critical question: Which is the better commonsense expression for talking with LLMs—rule or story? Specifically, this paper aims to answer the following two questions: (1) Which expression is more effective for retrieving commonsense from the memory of LLMs? (2) Which expression is more suitable for LLMs to leverage commonsense in solving problems?

To answer the questions, we systematically compare stories and rules as commonsense expressions for talking with LLMs. We use a total of 28 commonsense QA datasets for experiments. For the first question, we instruct LLMs to generate stories and rules based on commonsense questions, as shown in Figure 1. We compare the confidence and the accuracy of commonsense generation using stories and rules, showing that LLMs are more confident and more accurate at retrieving commonsense as stories than as rules. For the second question, we compare the confidence of generating the correct answers with stories or rules as contexts, showing that LLMs can more confidently leverage stories than rules for reasoning. The QA accuracy results further demonstrate that the story is a more effective commonsense expression for answering questions regarding daily events, while the rule is more effective for scientific commonsense QA. This phenomenon aligns with the reporting bias of commonsense in the text corpora. Moreover, stories and rules complement each other, i.e., combining them can further enhance the answer accuracy.

In-depth analyses reveal two main issues in generating commonsense stories: commonsense hallucination and semantic drifting. To address these problems, we propose an iterative self-supervised fine-tuning (self-SFT) method. We ask the model to generate stories given the training set of 8 datasets and design a scoring method to rank the stories based on their consistency with commonsense and similarity with the question. We filter the stories based on the scores and use them to fine-tune the model. The tuned model is then used to generate stories in the next iteration. Experimental results show that the self-SFT method leads to further accuracy improvements, highlighting the potential for LLMs to self-improve their commonsense abilities.

The main contributions of this paper are:

1. We systematically investigate and compare the effects of using stories and rules as commonsense expressions for retrieving and leveraging commonsense in LLMs. To our best knowledge, this is the first study to investigate the effects of specific commonsense expressions in LLMs.

2. We show that the story is a more effective expression for retrieving commonsense from LLMs and for leveraging commonsense in answering questions regarding daily events.

3. We identify two main issues that hinder commonsense story generation: commonsense hallucination and semantic drifting, and propose an iterative self-SFT method to improve the accuracy and relevance of stories generated by LLMs.

2 Background

Commonsense QA. Answering commonsense questions has become one of the standard tasks for evaluating LLMs (Srivastava et al., 2022). In this paper, we use 28 commonsense QA datasets for experiments, covering different domains of commonsense. These datasets are CommonsenseQA (Talmor et al., 2019), OpenBookQA (Mihaylov et al., 2018), Winograd Schema Challenge (WSC) (Levesque et al., 2012), PIQA (Bisk et al., 2020), SocialIQA (Sap et al., 2019b), ARC (easy and challenge set) (Clark et al., 2018), QASC (Khot et al., 2020), HellaSWAG (ActivitiNet and WikiHow set) (Zellers et al., 2019), NumerSense (Lin et al., 2020), AI2 Science Questions (AI2Sci, elementary and middle school set) (AllenAI, 2017), CommonsenseQA 2.0 (Talmor et al., 2021), SWAG (Zellers et al., 2018), WinoGrande (Sakaguchi et al., 2020), Com2Sense (Singh et al., 2021), SciQ (Welbl et al., 2017), QuaRel (Tafjord et al., 2019a), QuaRTz (Tafjord et al., 2019b), CycIC, ComVE (Task A) (Wang et al., 2019), COPA (Roemmele et al., 2011), PROST (Aroca-Ouellette et al., 2021), CODAH (Chen et al., 2019), Story Cloze Test (SCT) (Mostafazadeh et al., 2016), αNLI (Bhagavatula et al., 2020), and WinoVenti (Do and Pavlick, 2021). We use their development set for evaluation.

Commonsense knowledge and knowledge-augmented reasoning. In commonsense research, there is a growing consensus that integrating knowledge can improve the commonsense ability of NLP models (Bian et al., 2021). Typically, commonsense knowledge is expressed by concise and clear rules as either triples of <head, relation, tail> like in ConceptNet (Speer et al., 2017), or simple sentences like in Open Mind Common Sense (Singh et al., 2002). Recently, researchers turn to retrieving commonsense knowledge from pre-trained LLMs like GPT-3, assuming LLMs have already learned abundant commonsense from large-scale human-written text corpora (Wang et al., 2023a; Chen et al., 2023b; Liu et al., 2023a; Li et al., 2023; Yao et al., 2023; Wang et al., 2023b; Zhou et al., 2022; Wang et al., 2022a; Wei et al., 2022; Liu et al., 2022a,b; Gu et al., 2022; Yu et al., 2022; Paranjape et al., 2021; West et al., 2022; Bosselut et al., 2021; Shwartz et al., 2020; Latcinnik and Berant, 2020; Rajani et al., 2019). They instruct LLMs with prompts and examples to generate commonsense rules as concise sentences. In this paper, we follow their assumption and compare stories and rules as the commonsense expression for retrieving commonsense knowledge from LLMs.

There have been numerous works that inject commonsense rules into NLP models to improve commonsense reasoning, either by pre-training on knowledge bases (Ma et al., 2021; Chang et al., 2020; Mitra et al., 2019; Zhong et al., 2019), or incorporating knowledge rules in the input of language models (Shi et al., 2023; Wang et al., 2023c,f; Bian et al., 2021). There is also a chain of works that use graph-based reasoning for inference (Wang et al., 2023e; Yasunaga et al., 2021; Lv et al., 2020; Lin et al., 2019). In this paper, we exploit the commonsense knowledge in the memory of LLMs as stories or rules to support commonsense reasoning.

Story generation by LLMs. Recent studies have found that LLMs can perform well on story generation (Eldan and Li, 2023; Wen et al., 2023; Peng et al., 2022). Bhandari and Brennan (2023) compare stories generated by OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023), and Alpaca (Taori et al., 2023) with human-written stories, showing that these two kinds of stories exhibit a remarkable similarity in terms of readability and topics. The LLM-generated stories are even more accessible than traditional children's stories. Xie et al. (2023) compared GPT-3 with story generation models before LLMs, demonstrating that LLMs generate stories of significantly high quality, even comparable with human authors. In this paper, we analyze and exploit the inherent commonsense embedded in the storytelling ability of LLMs.

Large language models. This study focuses on three LLMs: ChatGPT (OpenAI, 2022), Alpaca (Taori et al., 2023), and Vicuna (Chiang et al., 2023; Zheng et al., 2023). ChatGPT was developed by OpenAI and is one of the state-of-the-art LLMs demonstrating robust abilities for generating human-like text. Our experiments are conducted using the *gpt-3.5-turbo* API. Alpaca is an open-source LLM with 7B parameters that achieves a good balance between performance and efficiency. It was fine-tuned from the LLaMA-7B model (Touvron et al., 2023) using 52K instructions gathered via a "self-instruct" methodology (Wang et al., 2022b). Vicuna is another open-source LLM trained by fine-tuning the LLaMA2 model on about 125,000 user-shared conversations with ChatGPT from ShareGPT. We use Vicuna v1.5 with 7B parameters. During experiments, we set the temperature of these LLMs to 0 when answering questions and to default when generating stories and rules.

3 Commonsense Retrieval from LLMs as Stories and Rules

In this section, we answer the first question: *Which expression, story or rule, is more effective for re-*

trieving commonsense from the memory of LLMs? First, we compare the confidence in generating stories and commonsense rules. Then, we employ an automatic evaluation to assess the accuracy of commonsense within the generated stories and rules.

3.1 Confidence of Commonsense Generation

To assess the confidence in generating stories and rules using LLMs, we ask LLMs to write corresponding stories of past experiences and commonsense rules given questions from commonsense QA datasets as input (as shown in Figure 1). Specifically, we randomly select 100 questions from each dataset and instruct Alpaca and Vicuna models to generate 5 stories and 5 rules using specific prompts (shown in Table 4 in Appendix A). We use perplexity to indicate the generation confidence of LLMs, which has been a longstanding confidence measure for language models (Jiang et al., 2021). However, there is a notable difference in word usage between stories and rules. Stories tend to incorporate less common words such as people's names and specific scenes, while rules typically consist of more general and common words. To account for this variation in word frequencies, we subtract the text perplexity with the perplexity of randomly shuffled word lists, which is a common practice in psychological linguistic studies to account for the word frequency effects (Humphries et al., 2006; Pallier et al., 2011; Zaccarella et al., 2017; Labache et al., 2019; Zhang et al., 2023a). The confidence is measured by the "Perplexity Reduction (PR)":

$$\mathrm{PR}(t) = \mathrm{PPL}(\mathrm{shuffle}(t)) - \mathrm{PPL}(t) \qquad (1)$$

Here, $\mathrm{PPL}(\cdot)$ denotes the perplexity calculation function, and $\mathrm{shuffle}(t)$ refers to the shuffling of the text t by words. A higher PR indicates that the LLM is more confident with the text.

Finding 1. When retrieving commonsense from LLMs, stories can result in more confident commonsense generation than rules. The results in Figure 2 show that stories have significantly higher PR than rules for both Vicuna and Alpaca models ($p \ll 0.001$). This effect is more obvious for Vicuna, and we believe this is because Vicuna has better storytelling and instruction-following abilities than Alpaca. These observations align with the reporting bias, where commonsense rules are less prevalent in the training text of LLMs, resulting in lower confidence in generating them. Conversely, human language is more likely to convey commonsense as stories, so LLMs develop an ability to generate stories with more confidence.

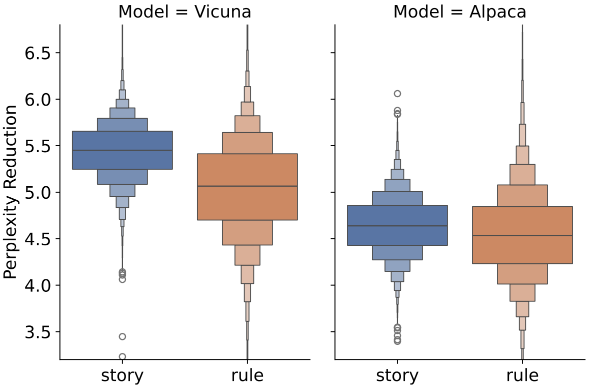


Figure 2: Comparison of perplexity reduction between generating stories and rules. Sample size $N = 14,000$ for each setting.

Setting	ChatGPT	Vicuna	Alpaca
Story	**99.42%**	**98.82%**	**95.39%**
Rule	98.56%	96.21%	93.25%

Table 1: Commonsense accuracy of stories and rules.

3.2 Accuracy of Commonsense Generation

To assess which expression incorporates more accurate commonsense knowledge, we ask ChatGPT to determine if each story or rule aligns with commonsense, responding with either "yes" or "no" (prompt shown in Table 6 in Appendix A). We use the same 100 questions for each dataset in Section 3.1, randomly selecting one story and one rule for evaluation, which results in a total of 2,800 stories and 2,800 rules for each model.

Finding 2. LLMs generate more accurate stories than rules in terms of commonsense. Table 1 shows that the commonsense accuracy of stories is higher than that of rules for all three models ($p < 0.0005$). This verifies that the story is a more accurate commonsense expression than the explicit commonsense rule for retrieving commonsense knowledge from LLMs, highlighting the potential of stories as valuable commonsense sources.

4 Leveraging Commonsense in Stories and Rules for Problem Solving

This section answers the second question: *Which expression, story or rule, is more suitable for LLMs to leverage commonsense in solving problems?* We compare the confidence of reasoning with stories or rules as contexts. Then, we assess the performance of commonsense QA by employing either stories or rules as contexts and perform detailed analyses.

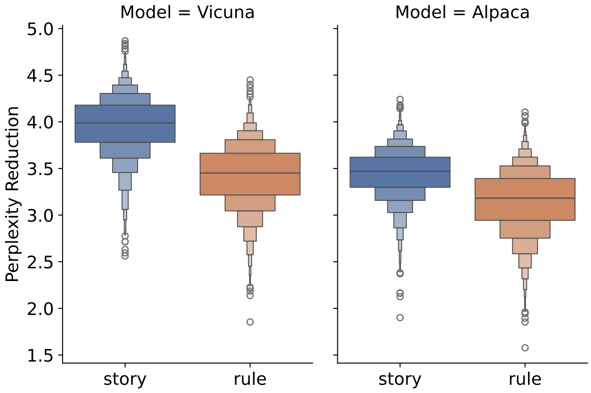


Figure 3: Comparison of perplexity reduction in generating the correct answer with stories or rules as context. Sample size $N = 14,000$ for each setting.

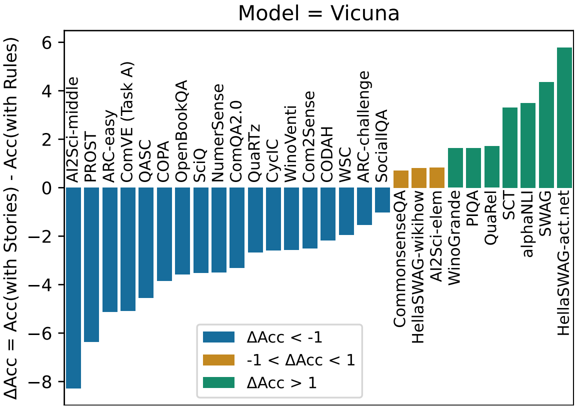


Figure 4: Comparison between the accuracy (%) with stories and with rules for Vicuna.

4.1 Confidence in Commonsense Reasoning

To evaluate the confidence of reasoning and generating the correct answer in commonsense QA given stories and rules as contexts, we compare the perplexity reduction of sequence "context, question (and options if applicable), correct answer" with context as either stories or rules. Specifically, we employ the same set of 100 questions, the stories, and the rules of each dataset in Section 3.1. The perplexity reduction is calculated similarly:

$$\begin{aligned}&\mathrm{PR}([c, q, a]) \\ &= \mathrm{PPL}([\mathrm{shuffle}(c), q, a]) - \mathrm{PPL}([c, q, a])\end{aligned} \tag{2}$$

where c, q, and a are the context, question, and correct answer, respectively.

Finding 3. LLMs are more confident in commonsense reasoning based on stories than on rules. Figure 3 shows a significantly higher perplexity reduction when generating the correct answers for commonsense questions using stories than using rules as contexts ($p < 0.0003$). This discrepancy reflects the exposure bias in commonsense reasoning: explicit rules are seldom used by people to reason and solve commonsense problems, resulting in more sparse examples in the text corpora for training LLMs. This finding affirms that LLMs can more naturally leverage stories of past experiences for reasoning in commonsense QA.

4.2 Effectiveness in Commonsense QA

Next, we assess the accuracy of zero-shot commonsense QA using stories or rules, comparing them with a baseline setting without contextual commonsense. For each question, LLMs first generate five stories or rules, which are then concatenated as context for answering questions using the same model (with prompts shown in Table 5 in Appendix A). The results are presented in Table 1, and the accuracy differences between using stories and rules are shown in Figure 4 for Vicuna and Figure 7 in Appendix C.1 for the other two models.

Finding 4. Story is a more effective commonsense expression for answering questions regarding daily events, while the rule is more effective for scientific commonsense QA, which aligns with the reporting bias of commonsense. As shown in Table 2 and Figure 4, for datasets like HellaSWAG, SWAG, SCT, and αNLI, which involve daily events and stories, leveraging stories as context leads to higher accuracies across all models (only except ChatGPT on the SCT dataset). These datasets involve tasks like selecting correct follow-up behaviors for sequences of events (HellaSWAG and SWAG), choosing appropriate story endings (SCT), or determining events between the beginning and end of a script (αNLI). We attribute this to the influence of reporting bias of different types of commonsense in text corpora, which shapes the commonsense ability of LLMs. Commonsense in our daily life events, such as "*how to add toothpaste onto a toothbrush*", is subject to more pronounced reporting bias than other forms of commonsense (Shwartz and Choi, 2020). This type of commonsense is inherently more implicit and, consequently, more suitable to be expressed as stories.

In contrast, datasets focusing on scientific commonsense at the elementary or middle school level, including OpenBookQA, ARC, QASC, AI2Sci, and SciQ, show better performance when provided with commonsense rules. This is because scientific knowledge is typically structured as rules in textbooks and encyclopedias. The questions and options often include scientific terms, and scientific

Datasets	ChatGPT (gpt-3.5-turbo)				**Vicuna**				Alpaca			
	Base	Story	Rule	Both	Base	Story	Rule	Both	Base	Story	Rule	Both
†HellaSWAG-act.net	**67.57**	60.96	60.86	62.83	45.75	**48.07**	42.29	**48.40**	29.29	**32.23**	29.54	**33.60**
†SWAG	**69.50**	61.58	61.22	62.30	48.00	**48.28**	43.92	46.19	30.44	**35.54**	32.42	**36.47**
†αNLI	76.21	**77.64**	75.94	**79.28**	59.57	**64.51**	61.02	**64.67**	59.67	**61.44**	58.89	**62.37**
†SCT	**96.48**	95.94	96.26	**97.35**	79.82	**82.24**	78.93	**83.60**	**86.50**	84.15	83.31	83.57
QuaRel	70.04	77.82	**80.36**	80.29	55.60	**60.89**	59.18	60.15	51.82	**57.55**	54.32	**58.63**
PIQA	84.20	84.07	**84.40**	**84.90**	65.61	**67.03**	65.39	**67.59**	54.25	**54.91**	54.58	**57.82**
WinoGrande	63.38	**70.51**	68.33	69.78	57.74	**60.11**	58.48	**60.28**	50.08	**52.53**	50.24	51.42
‡AI2Sci-elem	**92.68**	87.80	**92.68**	90.24	51.22	**62.81**	61.98	**63.03**	42.62	46.72	**52.85**	47.97
†HellaSWAG-wikihow	**73.62**	69.45	62.75	68.29	28.89	**31.44**	30.63	28.69	26.43	**25.78**	25.64	25.42
CommonsenseQA	74.98	**76.00**	75.84	**77.89**	35.60	**47.05**	46.34	**49.01**	29.01	**33.94**	33.52	**34.70**
SocialIQA	70.21	70.02	**70.44**	**71.47**	**46.24**	42.64	43.66	43.73	42.16	**43.46**	42.69	**44.10**
‡ARC-challenge	82.27	79.93	**84.95**	82.94	42.62	50.00	**51.54**	50.87	36.79	**40.13**	36.36	**40.74**
WSC	73.33	**82.11**	80.70	**83.03**	62.46	63.00	**64.94**	**65.81**	55.48	**56.74**	54.61	55.63
CODAH	**85.77**	81.47	82.55	83.09	58.38	56.45	**58.63**	**59.88**	**44.32**	41.20	38.69	41.12
††Com2Sense	70.46	66.28	**75.45**	70.72	52.30	53.20	**55.70**	54.74	49.68	**50.45**	50.38	50.38
WinoVenti	75.41	77.66	**79.09**	**79.61**	58.46	58.79	**61.35**	60.11	52.49	53.60	**54.57**	**56.99**
CycIC	64.26	68.59	**74.49**	70.62	43.03	43.44	**46.02**	45.40	33.91	35.94	**38.68**	**39.61**
QuaRTz	72.40	77.60	**82.81**	78.33	52.74	58.52	**61.20**	59.23	57.72	58.36	**65.62**	64.83
††CommonsenseQA2.0	64.90	63.15	**70.09**	65.76	50.68	50.12	**53.43**	52.90	48.44	48.96	**49.55**	48.70
NumerSense	73.50	72.50	**76.00**	**76.50**	47.00	44.00	**47.50**	47.50	28.00	40.00	**54.00**	51.50
‡SciQ	**93.30**	92.08	91.98	92.69	65.35	68.19	**71.71**	**71.87**	47.90	**57.26**	54.45	56.40
‡OpenBookQA	78.00	**80.00**	78.92	**82.93**	34.80	41.28	**44.86**	**46.62**	33.80	35.27	**36.49**	**37.15**
COPA	**96.40**	94.30	95.11	95.26	69.74	75.86	**79.71**	**82.89**	**81.20**	75.80	75.80	79.00
‡QASC	75.70	**77.43**	77.14	**79.14**	27.09	40.53	**45.08**	41.64	22.59	**28.57**	26.14	27.68
††ComVE (Task A)	92.26	87.18	**94.71**	90.94	49.85	48.59	**53.67**	48.54	52.77	**56.94**	54.09	**58.33**
‡ARC-easy	92.46	92.11	**93.86**	92.81	59.30	63.15	**68.28**	65.29	51.67	53.68	**54.74**	**56.59**
PROST	53.00	49.90	**62.90**	53.15	31.29	32.40	**38.77**	38.06	30.10	31.40	**33.50**	32.20
‡AI2Sci-middle	88.80	**92.00**	90.40	92.00	60.80	59.20	**67.48**	66.13	52.00	**55.28**	46.77	49.59

Table 2: Accuracy (%) in zero-shot commonsense QA under different settings: Base - without context, Story - with stories as context, Rule - with rules as context, Both - with both stories and rules as context. Datasets are sorted and grouped by accuracy differences between using stories and using rules in Vicuna, as depicted in Figure 4. † Datasets related to daily events. ‡ Scientific commonsense datasets. †† Datasets related to negation.

concepts and descriptions are not commonly presented in the form of stories in the training corpora.

Finding 5. Stories and rules can complement each other, further enhancing the QA accuracy of LLMs. As shown in Table 2, when employing both stories and rules as contextual inputs, LLMs can achieve higher QA accuracy on 10, 12, and 13 datasets compared to using either stories or rules alone. This highlights that the combination of both stories and rules enables LLMs to leverage the unique strengths of each, resulting in a more comprehensive and precise understanding of the presented questions and underlying commonsense. For example, although rules are better for expressing scientific commonsense, stories still play a crucial role in scientific commonsense QA by providing essential contextual information. As shown in Table 2, using both stories and rules on the OpenBookQA dataset further improves answer accuracy for all three models.

4.3 Analyses

4.3.1 Error Analysis

To gain deeper insights into the influence of generated stories, we conduct an error analysis. We focus on questions that are initially answered correctly by the Vicuna model, but the answers change to incorrect when considering the stories as context. We randomly select 10 such questions from each dataset, except for the AI2Sci-elem dataset which has only 8 such questions. Each question comprises five stories, resulting in a total of 1390 stories for error analysis. We manually classify error cases into 5 primary types: **Semantic Drifting** (34.0% – the story drifts away from the question), **Uncommon or Incorrect Scenarios** (26.6% – the story does not represent common real-world situations or contains errors), **Incorrect Answering** (18.6% – the story is accurate, but the predicted answer is wrong), **Inconsideration of Options** (16.2% – the story does not align with any answer options), and **Inclusion of Wrong Options** (4.6% – the story emphasizes a wrong answer). The pie chart is shown in Figure 8 in Appendix C.2.

This analysis highlights two main issues of the stories: commonsense hallucination and semantic drifting. A commonsense hallucination occurs when LLMs are misled by incorrect answer options, generating stories that are uncommon or against commonsense, leading to the uncommon or incorrect scenarios error. Semantic drifting refers to LLMs generating stories whose topics deviate from

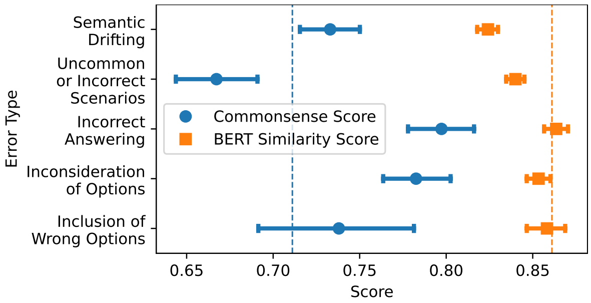


Figure 5: The average scores of stories generated by Vicuna of different error types. The dashed lines are the overall average scores among all questions. Error bars indicate 95% confidence intervals.

the question, making them unhelpful in answering the question. These two error types are the primary reasons for incorrect model answers, accounting for over 60% of total errors. Besides, 18.6% of the stories are correct and relevant, yet the model fails to effectively use them to answer questions correctly, suggesting room for further improvement in LLMs' ability to leverage contextual information.

To quantitatively analyze the generated stories, we employed two scoring methods: commonsense scores and BERT similarity scores. For the commonsense score, we use the Vera model (Liu et al., 2023b), a T5-based model trained on extensive commonsense statements from knowledge bases. This model outputs a score of the correctness for a given text according to commonsense. For the BERT similarity score, we calculate the cosine similarity between semantic representations of stories and questions using the BERT-large model (Devlin et al., 2019). The two scores range between 0 and 1.

There is a correlation between error types and the corresponding scores. Figure 5 shows that stories with semantic drifting have significantly lower semantic similarities with the questions than the overall average and other error types. Moreover, stories describing uncommon or incorrect scenarios show both lower similarity and commonsense scores in contrast to the overall average and other error types except for semantic drifting. These findings further support the commonsense hallucination and semantic drifting issues.

4.3.2 Influence of Story on Answer Accuracy

Further analysis shows notable correlations between answer accuracy and the two scores, the commonsense score and BERT similarity score, at the dataset level (shown in Figure 9 and 10 in Appendix C.3). Specifically, answer accuracy demonstrates a robust correlation with the commonsense score (Pearson coefficient 0.612, $p < 0.001$). In comparison, the correlation with the BERT similarity score is weaker but still positive (Pearson coefficient 0.226, $p = 0.003$). This is because a story involving commonsense hallucination may mislead the model by providing incorrect information, leading to wrong answers, while a story deviating from the question merely offers no relevant information, resulting in a weaker influence. These observations underscore the substantial influence of semantic drifting and commonsense hallucination issues on commonsense QA based on stories.

4.3.3 Analysis of Datasets Related to Negation

An interesting phenomenon of leveraging commonsense in stories is that LLMs are still not good at handling negations in commonsense. On datasets that involve negation, including CommonsenseQA 2.0 and Com2Sense which require models to assess statement correctness, and ComVE (Task A) which requires identifying statements contradicting commonsense, using rules demonstrates higher accuracy compared to using stories (Table 2). Specifically, the Vicuna model tends to more frequently give incorrect "yes" responses to questions with a correct answer of "no" (69.1% of error cases in CommonsenseQA 2.0 and 90.2% in Com2Sense) than the opposite cases. Furthermore, for error cases where the correct answer is "no", the model generates stories with more commonsense errors (the average commonsense scores of stories for incorrectly answered "no" questions are significantly lower than the opposite with $p < 0.003$). This disparity indicates a challenge for LLMs in handling negations within commonsense (Chen et al., 2023a). Training corpora include few negative commonsense examples like "*a stapler is not used for sewing*", resulting in LLMs generating more hallucinations misled by the given incorrect statements.

5 Iterative Self-Supervised Fine-Tuning

After identifying the two issues in generating stories of past experiences, this section presents our iterative self-SFT approach to address these issues.

5.1 Method

Our self-SFT method contains three steps in each iteration: generating, filtering, and training.

In the generating step, we use LLMs to generate stories for questions in the training set. Specifically, we generate five stories for each question. In the filtering step, we first select the generated

Datasets	Without SFT	Iter-1	Iter-2	Iter-3
Seen	51.96	52.32	**53.26**	52.12
Unseen	55.30	**55.57**	55.16	55.08

Table 3: Average accuracy (%) of commonsense QA by Vicuna with and without iterative self-SFT. Accuracy for each dataset is shown in Table 9 in Appendix E.

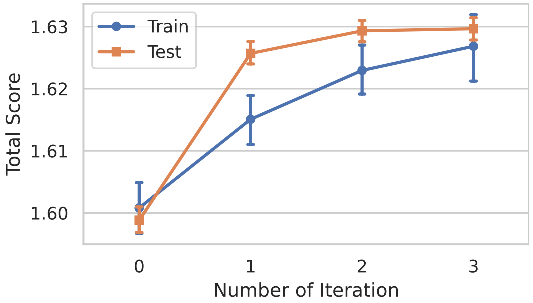


Figure 6: Changes of total score alongside the iteration of self-SFT. Error bars are 95% confidence intervals.

stories based on changes in the responses to the questions, i.e., a story is considered helpful if it can rectify an initially incorrect response as correct given it in the context. These helpful stories are then scored using a scoring method. The design of the scoring method is crucial for mitigating the commonsense hallucination and semantic drifting issues. We employ the two scores in Section 4.3: the commonsense score by the Vera model for commonsense correctness and the BERT-based semantic similarity for relevance. The total score is the sum of the two scores. We retain the $K\%$ top-scored stories for the subsequent training step. In the training step, we fine-tune an LLM using the filtered stories as output, with inputs following the prompt in Section 3.1. The fine-tuned model is used for the generating step in the next iteration.

5.2 Experiments

Following Liu et al. (2023a) and our analyses, we train our model on 8 datasets: OpenBookQA, AI2Sci (elementary and middle school set), WinoGrande, HellaSWAG (ActivitiNet and WikiHow set), SWAG, and αNLI. The other datasets are used for unseen evaluation. We fine-tune the Vicuna model using LoRA tuning (Hu et al., 2022), with hyper-parameters in Appendix D.

The commonsense ability of LLMs can be further self-improved via iterative self-supervised fine-tuning. Table 3 shows that on the datasets used for fine-tuning (seen), the average accuracy at all three iterations outperforms the original Vicuna model without SFT. The most significant improvement is at iteration 2, rising from 51.96% to 53.26%. This verifies the effectiveness of self-SFT in enhancing the quality of generated stories. Furthermore, self-SFT demonstrates improvements on unseen datasets at iteration 1, suggesting the method's ability to generalize to other commonsense QA datasets not seen during training. Performance decreases at later iterations can be attributed to over-fitting on seen datasets. Ablation and hyper-parameter studies are shown in Appendix E.

Our approach is effective in addressing the issues and improving the quality of generated stories. To further assess the effect of our method in mitigating commonsense hallucination and semantic drifting issues during story generation, we analyze the score variations across training iterations. Figure 6 shows that, along with the training iterations, the total score (sum of commonsense score and BERT similarity) consistently improves for both the training and testing phases.

6 Conclusion

This paper systematically compares stories and rules as commonsense expressions to retrieve and leverage commonsense knowledge in LLMs. Experimental results show that the story is a better expression for retrieving commonsense from LLMs. LLMs generate stories with more confidence and higher commonsense accuracy than rules. Moreover, the story is a more effective commonsense expression for answering questions regarding daily events, while the rule is more effective for scientific commonsense QA. This phenomenon aligns with the reporting bias of commonsense. Stories and rules can complement each other to further enhance answer accuracy. We provide further insights through in-depth analyses, highlighting two challenges in generated stories: commonsense hallucination and semantic drifting. We show that the correctness and relevance of commonsense stories can be improved via iterative self-supervised fine-tuning, underscoring the potential for self-improvement in the commonsense ability of LLMs.

This paper suggests a new perspective, going beyond the common practice of expressing commonsense as rules. Our results and findings emphasize the importance of using the appropriate language to express, retrieve, and leverage commonsense for LLMs to further exploit their potential. The full extent of LLMs in handling commonsense is not yet fully realized, calling for future research to refine and improve the commonsense abilities of LLMs.

Limitations

This study specifically investigates several popular LLMs, including ChatGPT, Vicuna, and Alpaca, while excluding the exploration of other LLMs such as GPT-4, Mistral (Jiang et al., 2023), and Google's Bard (Thoppilan et al., 2022). The selection of models is based on considerations of popularity, availability, and cost. Future research could provide valuable insights by examining whether similar findings hold for these models and conducting performance comparisons with the models included in this study.

The assessment of commonsense accuracy in Section 3.2 relies on automatic labeling by ChatGPT. Therefore, the accuracy presented in Table 1 serves solely for comparing stories and rules, and should not be regarded as absolute accuracy. To ensure the quality of this automatic evaluation, we conduct a manual review of the stories and rules labeled as incorrect by ChatGPT and confirm that they are indeed incorrect in terms of commonsense. However, due to its labor-intensive nature, manually labeling all generated stories and rules is impractical for us. Future studies need to incorporate human evaluation to provide a more comprehensive and nuanced understanding of the generated commonsense knowledge by LLMs.

This paper employs two scores to assess the quality of generated stories and to filter training data. However, these two scores—the commonsense score and BERT similarity score—may inherently exhibit biases as they rely on model-based scoring methods. For instance, the Vera model for commonsense score is fine-tuned from the T5 model using commonsense statements synthesized from commonsense knowledge bases, potentially leading the score to favor statements aligned with these knowledge bases. We anticipate detailed analyses of potential biases in these models and scores.

Lastly, in the manual error analysis process, we discover that some commonsense QA datasets are not actually asking about commonsense, despite being recognized as commonsense questions. Some datasets are automatically constructed based on knowledge graphs, probably leading to unreasonable questions or insufficient information to answer the question. Other manually constructed datasets may face the challenge that different annotators may have entirely different understandings of what commonsense is. We follow the common practices in commonsense studies and use these datasets in our commonsense QA experiments. Further investigation into the existing commonsense QA datasets is a task for future studies.

References

AllenAI. 2017. Ai2 science questions v2.1 (october 2017). http://data.allenai.org/ai2-science-questions/.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. PROST: Physical reasoning about objects through space and time. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4597–4608, Online. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Prabin Bhandari and Hannah Marie Brennan. 2023. Trustworthiness of children stories generated by large language models. *ArXiv preprint*, abs/2308.00073.

Ning Bian, Xianpei Han, Bo Chen, and Le Sun. 2021. Benchmarking knowledge-enhanced commonsense question answering via knowledge-to-text transformation. 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 12574–12582. AAAI Press.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *ArXiv preprint*, abs/2303.16421.

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.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. 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 4923–4931. AAAI Press.

R.J. Brachman and H.J. Levesque. 2023. *Machines like Us: Toward AI with Common Sense*. MIT Press.

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*.

E. Cassirer, E.A. Cassirer, S.K.K. Langer, O. Hansen-Love, and S. Marić. 1946. *Language and Myth*. Dover Books on Literature, Philosophy, History, Religion. Dover Publications.

Ting-Yun Chang, Yang Liu, Karthik Gopalakrishnan, Behnam Hedayatnia, Pei Zhou, and Dilek Hakkani-Tur. 2020. Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 74–79, Online. Association for Computational Linguistics.

Jiangjie Chen, Wei Shi, Ziquan Fu, Sijie Cheng, Lei Li, and Yanghua Xiao. 2023a. Say what you mean! large language models speak too positively about negative commonsense knowledge. *ArXiv preprint*, abs/2305.05976.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Qianglong Chen, Guohai Xu, Ming Yan, Ji Zhang, Fei Huang, Luo Si, and Yin Zhang. 2023b. Distinguish before answer: Generating contrastive explanation as knowledge for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13207–13224, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nam Do and Ellie Pavlick. 2021. Are rotten apples edible? challenging commonsense inference ability with exceptions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2061–2073, Online. Association for Computational Linguistics.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *ArXiv preprint*, abs/2305.07759.

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, and Peter Clark. 2022. DREAM: Improving situational QA by first elaborating the situation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1127, Seattle, United States. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Colin Humphries, Jeffrey R Binder, David A Medler, and Einat Liebenthal. 2006. Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of cognitive neuroscience*, 18(4):665–679.

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.

Filip Ilievski, Pedro Szekely, and Bin Zhang. 2021. Cskg: The commonsense knowledge graph. In *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pages 680–696. Springer.

Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. 2023. Crow: Benchmarking commonsense reasoning in real-world tasks. *ArXiv preprint*, abs/2310.15239.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *ArXiv preprint*, abs/2310.06825.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11518–11537, Singapore. 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.

Gary Klein. 2004. *The Power of Intuition: How to Use Your Gut Feelings to Make Better Decisions at Work*. Currency/Doubleday.

Loic Labache, Marc Joliot, Jérôme Saracco, Gaël Jobard, Isabelle Hesling, Laure Zago, Emmanuel Mellet, Laurent Petit, Fabrice Crivello, Bernard Mazoyer, et al. 2019. A sentence supramodal areas atlas (sensaas) based on multiple task-induced activation mapping and graph analysis of intrinsic connectivity in 144 healthy right-handers. *Brain Structure and Function*, 224(2):859–882.

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

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

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.

Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023. Guiding large language models via directional stimulus prompting. *ArXiv preprint*, abs/2302.11520.

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.

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.

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.

Jiacheng Liu, Ramakanth Pasunuru, Hannaneh Hajishirzi, Yejin Choi, and Asli Celikyilmaz. 2023a.

Crystal: Introspective reasoners reinforced with self-feedback. *ArXiv preprint*, abs/2310.04921.

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

Weize Liu, Guocong Li, Kai Zhang, Bang Du, Qiyuan Chen, Xuming Hu, Hongxia Xu, Jintai Chen, and Jian Wu. 2023c. Mind's mirror: Distilling self-evaluation capability and comprehensive thinking from large language models. *ArXiv preprint*, abs/2311.09214.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. 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 8449–8456. AAAI Press.

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. 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 13507–13515. AAAI Press.

John McCarthy. 1959. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pages 75–91, London. Her Majesty's Stationary Office.

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.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. How additional knowledge can improve natural language commonsense question answering? *ArXiv preprint*, abs/1909.08855.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt. 2022, Nov 30.

Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. 2011. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527.

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.

Xiangyu Peng, Siyan Li, Sarah Wiegreffe, and Mark Riedl. 2022. Inferring the reader: Guiding automated story generation with commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7008–7029, Abu Dhabi, United Arab Emirates. 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.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

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, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial*

Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 3027–3035. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Daniel L Schacter and Donna Rose Addis. 2007. The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362:773–786.

R. C. Schank and R. P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum, Hillsdale, NJ.

R.C. Schank. 1995. *Tell Me a Story: Narrative and Intelligence*. Rethinking theory. Northwestern University Press.

Roger C Schank. 1983. *Dynamic memory: A theory of reminding and learning in computers and people*. cambridge university press.

Haochen Shi, Weiqi Wang, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu, and Yangqiu Song. 2023. Qadynamics: Training dynamics-driven synthetic qa diagnostic for zero-shot commonsense question answering. *ArXiv preprint*, abs/2310.11303.

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.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, page 1223–1237, Berlin, Heidelberg. Springer-Verlag.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, abs/2206.04615.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019a. QUAREL: A dataset and models for answering questions about qualitative relationships. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7063–7071. AAAI Press.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019b. QuaRTz: An open-domain dataset of qualitative relationship questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5941–5946, Hong Kong, China. 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.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of ai through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Endel Tulving. 2002. Episodic memory: From mind to brain. *Annual review of psychology*, 53:1–25.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. Association for Computational Linguistics.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023a. Boosting language models reasoning with chain-of-knowledge prompting. *ArXiv preprint*, abs/2306.06427.

PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022a. Pinto: Faithful language reasoning using prompt-generated rationales. In *The Eleventh International Conference on Learning Representations*.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023b. Scott: Self-consistent chain-of-thought distillation. *ArXiv preprint*, abs/2305.01879.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023c. Car: Conceptualization-augmented reasoner for zero-shot commonsense question answering. *ArXiv preprint*, abs/2305.14869.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023d. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13111–13140, Toronto, Canada. Association for Computational Linguistics.

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

Yujie Wang, Hu Zhang, Jiye Liang, and Ru Li. 2023e. Dynamic heterogeneous-graph reasoning with language models and knowledge representation learning for commonsense question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14048–14063, Toronto, Canada. Association for Computational Linguistics.

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. 2023f. COLA: Contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5253–5271, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Zhihua Wen, Zhiliang Tian, Wei Wu, Yuxin Yang, Yanqi Shi, Zhen Huang, and Dongsheng Li. 2023. GROVE: A retrieval-augmented complex story generation framework with a forest of evidence. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3980–3998, Singapore. Association for Computational Linguistics.

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.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351.

Yunzhi Yao, Peng Wang, Shengyu Mao, Chuanqi Tan, Fei Huang, Huajun Chen, and Ningyu Zhang. 2023. Knowledge rumination for pre-trained language models. *ArXiv preprint*, abs/2305.08732.

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.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Emiliano Zaccarella, Marianne Schell, and Angela D Friederici. 2017. Reviewing the functional basis of the syntactic merge mechanism for language: A coordinate-based activation likelihood estimation meta-analysis. *Neuroscience & Biobehavioral Reviews*, 80:646–656.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. 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.

Guangyao Zhang, Yangwen Xu, Xiuyi Wang, Jixing Li, W. Shi, Yanchao Bi, and Nan Lin. 2023a. A social-semantic working-memory account for two canonical language areas. *Nature Human Behaviour*, 7:1980–1997.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068.

Zheyuan Zhang, Shane Storks, Fengyuan Hu, Sungryull Sohn, Moontae Lee, Honglak Lee, and Joyce Chai. 2023b. From heuristic to analytic: Cognitively motivated strategies for coherent physical commonsense reasoning. *ArXiv preprint*, abs/2310.18364.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv preprint*, abs/2306.05685.

Wanjun Zhong, Duyu Tang, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Improving question answering by commonsense-based pre-training. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8*, pages 16–28. Springer.

Pei Zhou, Karthik Gopalakrishnan, Behnam Hedayatnia, Seokhwan Kim, Jay Pujara, Xiang Ren, Yang Liu, and Dilek Hakkani-Tur. 2022. Think before you speak: Explicitly generating implicit commonsense knowledge for response generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1237–1252, Dublin, Ireland. Association for Computational Linguistics.

Common Prefix: *Jane is answering this question:* *Question: Where do adults use glue sticks?* *Options: A. classroom, B. desk drawer, C. at school, D. office, E. kitchen drawer.*
Prompt for Generating Story: *Jane is reminded of a specific past experience analogous to the situation and the most important information in this question. However, Jane refrains from forming conclusions or making guesses about the answer at this time.* *Write a possible experience as detailed and focused story in a paragraph that Jane may recall and conforms to the common practise. Do not use names in the question or mention the options in the story. Do not output extra sentences.*
Prompt for Generating Rule: *Jane is reminded of specific commonsense rules relevant to the situation and the most important information in this question (without considering the options). However, Jane refrains from forming conclusions or making guesses about the answer at this time.* *List possible commonsense rules as simple knowledge sentences that Jane may recall in a paragraph. Do not output extra sentences.*

Table 4: Prompts for generating stories and rules, with an example question from the CommonsenseQA dataset. The common prefix comes before each prompt.

A Prompts

The prompts we use in this paper for instructing LLMs to generate stories and rules are shown in Table 4. The common prefix which contains the question and the answer options (if applicable) is added before the two prompts below. To avoid responses such as “As an AI language model, I do not have past experiences”, the specific name “Jane” is used instead of “you” in our prompts. It is worth noting that the choice of the name is arbitrary, and any name can be used. The potential influence of name biases in LLMs is a topic for future study.

The prompts for answering questions are shown in Table 5. The prompt for automatically evaluating the stories and rules is shown in Table 6.

These prompts are constructed through a prompt engineering process. This involves testing and comparing different prompt variations to select the most effective ones. Through this process, we can find the prompts to ensure optimal performance in guiding LLMs to generate high-quality stories and rules and to answer questions effectively.

B Details of Commonsense QA Datasets

Table 7 provides information on the number of questions, question types, and accuracy when randomly selecting an answer option for each commonsense QA dataset. All these datasets are in En-

Prompt for Answering Question without Context: *Choose the most suitable answer for the question by selecting the answer letter and do not say anything else: {question} {answer_options}*
Prompt for Answering Question with Story: *Read these experiences:* *{story}* *Analogy to the above text as reference, choose the most suitable answer for the question by selecting the answer letter and do not include anything else: {question} {answer_options}*
Prompt for Answering Question with Rule: *Read these commonsense rules:* *{rule}* *Based on the above text as reference, choose the most suitable answer for the question by selecting the answer letter and do not include anything else: {question} {answer_options}*
Prompt for Answering Question with Both Rule and Story: *Read these experiences:* *{story}* *Read these commonsense rules:* *{rule}* *Based on the above experiences and commonsense rules as reference, choose the most suitable answer for the question by selecting the answer letter and do not include the option content or anything else: {question} {answer_options}*

Table 5: Prompts for answering questions with stories or rules.

Prompt for Evaluating Commonsense Accuracy: *Please evaluate the following sentences for common sense based on your commonsense knowledge:* *{text}* *Does the sentences align with your common sense? Respond with "yes" or "no" only.*

Table 6: Prompt for evaluating the stories and rules with ChatGPT.

glish. It is important to note that, for HellaSWAG-act.net, HellaSWAG-wikihow, SWAG, and PROST datasets, we randomly sample 1,000 questions from each of their development sets. This decision is made due to the excessively large number of questions in their development sets, which would have required significant time and computational resources for evaluation (3,243 for HellaSWAG-act.net, 6,799 for HellaSWAG-wikihow, 20,006 for SWAG, and 18,736 for PROST). An example from each dataset is presented in Table 8.

C Further Analyses of Generated Stories

C.1 Accuracy Differences between Using Stories and Rules for ChatGPT and Alpaca

We show the accuracy differences between using stories and rules for the ChatGPT and Alpaca models in Figure 7, as supplementary to Figure 4.

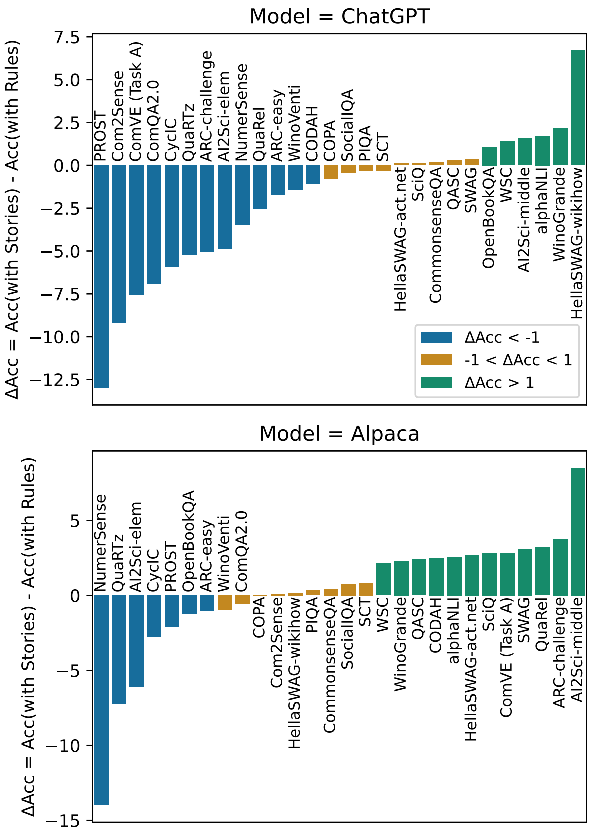


Figure 7: Comparison between the accuracy (%) with stories and with rules for ChatGPT and Alpaca.

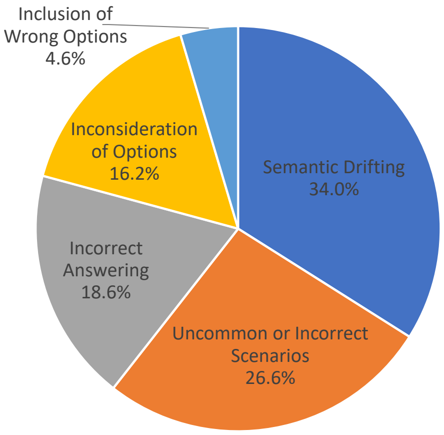


Figure 8: Error analysis of stories generated by Vicuna.

C.2 Error Analysis

We show the pie chart of our error analysis in Figure 8. From the figure we can see that uncommon or incorrect scenarios and semantic drifting are the primary reasons for incorrect model answers, accounting for over 60% of total errors.

Dataset	#Questions	Type	Random Accuracy
CommonsenseQA	1,221	General	20.00%
OpenBookQA	500	Science	25.00%
PIQA	1,838	General	50.00%
SocialIQA	1,954	Social	33.33%
ARC-easy	570	Science	25.00%
ARC-challenge	299	Science	25.00%
QASC	926	Science	12.50%
AI2Sci-elem	123	Science	25.23%
AI2Sci-middle	125	Science	24.92%
WinoGrande	1,267	General	50.00%
WSC	285	General	50.00%
NumerSense	200	Number	0.00%
HellaSWAG-act.net	1,000	Daily event	25.00%
HellaSWAG-wikihow	1,000	Daily event	25.00%
CommonsenseQA2.0	2,541	Yes/No	50.00%
SWAG	1,000	Daily event	25.00%
Com2Sense	782	Yes/No	50.00%
SciQ	1,000	Science	25.00%
QuaRel	278	Science & Comparing	50.00%
QuaRTz	384	Science & Comparing	50.00%
CycIC	1,525	Logical Reasoning	32.16%
ComVE (Task A)	997	Yes/No	50.00%
COPA	500	General	50.00%
PROST	1,000	Physical	25.00%
CODAH	556	General	25.00%
SCT	1,571	Daily event	50.00%
αNLI	1,532	Daily event	50.00%
WinoVenti	4,352	General	50.00%

Table 7: Commonsense QA datasets used in this paper. Random accuracy means the accuracy of randomly choosing an answer option.

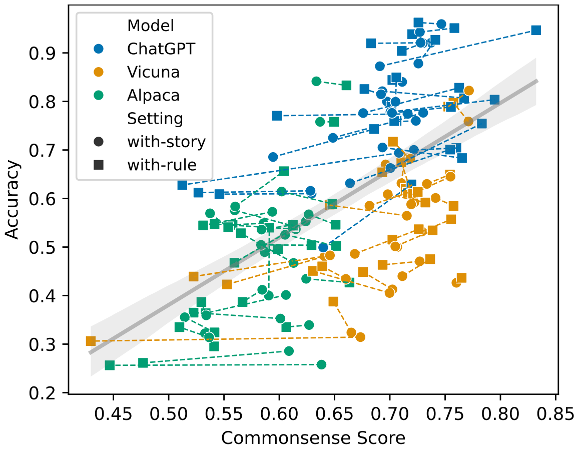


Figure 9: Correlation between answer accuracy of each dataset and commonsense scores of stories and rules.

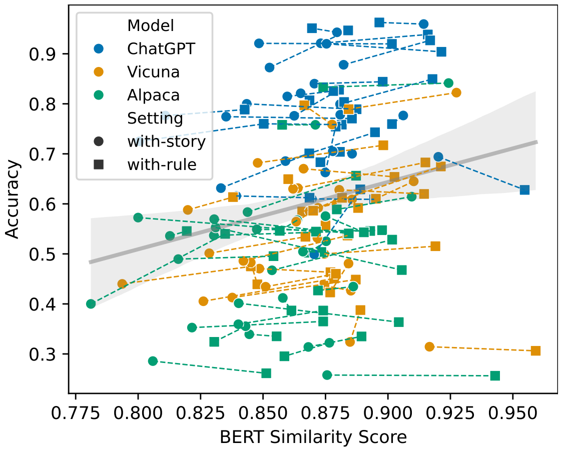


Figure 10: Correlation between answer accuracy of each dataset and BERT similarity scores of stories and rules.

C.3 Correlation Between Answer Accuracy and the Two Scores of Stories

We plot the correlation between answer accuracy and the commonsense score in Figure 9, and the BERT similarity score in Figure 10. The dashed lines are connected between the results of using stories and rules of the same datasets, which further examination reveals positive correlations of the two scores between using stories and using rules on most datasets

D Hyper-parameters for Self-SFT

For the filtering step, we randomly select 200 questions initially answered incorrectly by the Vicuna and can be corrected with at least one generated story for each dataset. The filter ratio K is set to 50%, resulting in 2,269, 2,300, and 2,447 training stories across the 8 datasets for three iterations.

For the training step, we use LoRA tuning (Hu et al., 2022) for training the Vicuna model. LoRA

Dataset	Example
CommonsenseQA	What is another name for a disk for storing information? A. computer store B. computer to store data **C. computer hard drive** D. cd player E. usb mouse
OpenBookQA	Owls spend their nights A. tending to their homes B. sleeping in hollow logs **C. scanning their territory for field mice** D. hanging out with other owls
PIQA	Where can I buy a tennis ball **A. You can purchase a tennis ball at any sports store** B. You can purchase a tennis racket at any sports store
SocialIQA	Aubrey the officer pulled a driver over for speeding on the road. Why did Aubrey do this? A. find a safe place to pull the person over **B. so people don't drive to fast** C. look up the person's license plate number
ARC-easy	Scientists at a local university have been studying the impact that people have on Earth. One of the areas being studied is how the burning of fossil fuels affects the environment. Which effect of fossil fuel burning have the scientists most likely evaluated? A. the production of nitrogen-fixing bacteria B. the mechanical weathering of roads **C. the formation of acid rain** D. the increase in runoff
ARC-challenge	How should a line graph be used to display distance and time data for a moving object? A. The y-axis should be labeled as time, which is the dependent variable. B. The y-axis should be labeled as distance, which is the independent variable. C. The x-axis should be labeled as distance, which is the dependent variable. **D. The x-axis should be labeled as time, which is the independent variable.**
QASC	What may renal failure be treated with? A. Laboratory B. Lymphocytes C. saves lives **D. dialysis** E. Lymph fluid F. dandelions G. ibuprofen H. Protein
AI2Sci-elem	To make an electromagnet, a conductor should be coiled around - A. a glass tube **B. an iron nail** C. a roll of paper D. a wooden stick
AI2Sci-middle	Which best describes the characteristics of a river basin? **A. the land drained by a river and its tributaries** B. the land formed when rivers create estuaries and marshes C. the land at the mouth of a river where water flows into the ocean D. the land formed as a result of a river flooding
WinoGrande	She chose the black car over the green car, because the A. black car has more brighter color. **B. green car has more brighter color.**
WSC	The user changed his password from "GrWQWu8JyC" to "willow-towered Canopy Huntertropic wrestles" as A. grwqwu8jyc was easy to remember. **B. willow-towered canopy huntertropic wrestles was easy to remember.**
NumerSense	a french horn has <how many> keys. (**three**)
HellaSWAG-act.net	Another man practices hurling himself backward over a pole onto a gym mat inside of the gym. several more men A. practice hurling street hurling outside and on a gym floor. B. practice hurling while a coach's hand watches. **C. practice long jumps and backward jumps inside of the gym using the sandbox and gym mats as landing tools.** D. practice pitches outside of the gym interior.
HellaSWAG-wikihow	[header] How to do tiger eye hair [title] Purchase your hair dye. [step] Take a trip to your local drug store or beauty supply store. Depending on the look you're going for, you may want to keep it simple and just choose one color, or you may want to buy four. A. Your hair will stick out more if you use a thicker dye, such as a mousse or gel. [title] Pour 3 ounces of red wine into your bowl. B. [substeps] In the tattoo artist's shop you should find several different colors and strips, and apply those to your hair. Make sure the colors match what you want to do. **C. It's totally up to you! You can buy a blonde highlighting kit that will lighten pieces of your brown hair, auburn dye, gorgeous golden hues, soft brown dyes-whichever dye you think will look good in your hair. The darker your hair is, the less of an effect you'll notice.** D. [substeps] If you want to dye your hair yourself, make sure to use the formula before you apply the dye. Gels are sometimes recommended but are highly expensive, and can be difficult to find at grocery stores.
CommonsenseQA2.0	In the US a senator is a person elected to a six year term? **A. yes** B. no
SWAG	Someone finds people playing chess at one of the long polished tables. She **A. walks down the brightly decorated hall to join them.** B. pats him in the gut with the box. C. faces the building, in wonder, someone and the other recruits stand around watching, uneasy. D. approaches two fat men wearing an earpiece into an office.
Com2Sense	Because the drive was 20 miles long, Beth was able to make it to her destination in under 5 minutes. A. True **B. False**
SciQ	What parts of a human possess the highest concentration of thermoreceptors? A. face and hair B. hand and ears **C. face and ears** D. hands and feet
QuaRel	Lebron James a strong player for the Cavs battles Kevin Durant a thin player for a rebound. Who is likely to get the rebound? A. Durant **B. Lebron**
QuaRTz	Long ago the surface of Venus warmed enough that greenhouse gases escaped into the atmosphere. As a result, the greenhouse effect on that planet **A. increased** B. decreased
CycIC	Rob lauded Will. Charity chastised Will. Who made Will feel happy? A. Daisy B. Cliff **C. Rob** D. Charity E. Joy
ComVE (Task A)	Which statement of the two is against common sense? **A. The cleaner is in charge of the money at the store** B. The cashier is in charge of the money at the store
COPA	The woman was in a bad mood. What was the effect of this? A. She engaged in small talk with her friend. **B. She told her friend to leave her alone.**
PROST	A person drops a bottle, a mirror, an egg, and a shirt from a balcony. Which object is the least likely to break? A. bottle B. mirror C. egg **D. shirt**
CODAH	Kieran is a whale. Kieran **A. is a mammal** B. is a dog C. has six human kids D. is a orange
SCT	I wanted to buy a video game console. I asked my parents, and they came up with an idea. They said if I did my chores, I would be given money to save. I did my chores without being asked every week for a whole summer. What is the end of this story? **A. My parents gave me enough money to buy the console.** B. At the end of the summer I gave the money back to my parents.
αNLI	The beginning of the story: Jim got ready for his first date.The ending of the story: Since then, she has ignored all of Jim's text messages. What happened between the begining and the end of the story? **A. Jim's date wasn't attracted to him.** B. Jim went on the date and said he didn't like the girl.
WinoVenti	The walnut was painted. The walnut is A. edible **B. toxic**

Table 8: Example question for each commonsense QA dataset. Answers are shown in bold.

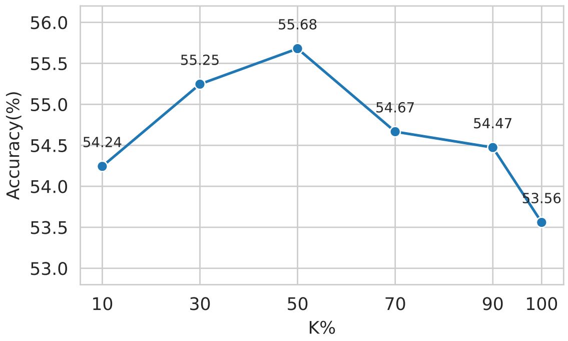


Figure 11: Relationship between average answer accuracy and filter ratio $K\%$.

is a parameter-efficient fine-tuning method that has become a common practice in the LLM era to reduce the overhead of expensive adaptations. The hyper-parameters for LoRA are rank $r = 16$ and $\alpha = 16$. Models are fine-tuned for 3 epochs in each iteration with a batch size of 64 and a learning rate of 3e-4.

E More Results and the Ablation Study of Self-SFT

E.1 Full Accuracy on All Datasets

We show the accuracy before and after self-SFT of each dataset in Table 9. Across 22 datasets, our self-SFT method consistently outperforms the original Vicuna model without SFT (7 seen and 15 unseen) and shows the most accuracy improvements at iterations 1 and 2.

E.2 Effect of Iterative Self-SFT

We compare our self-SFT method with a naive SFT method without iteration and scoring. In this ablated method, we only filter the stories generated by Vicuna on the training sets that can rectify an initially incorrect answer to correct, and use these stories for fine-tuning. The results are shown in Table 9.

From the results, we can see that the ablated method performs worse than our self-SFT method on all datasets, and even worse than the original Vicuna model without fine-tuning on some datasets. This further verifies that our iterative scoring and filtering mechanism is crucial for alleviating the semantic drifting and commonsense hallucination issues of story generation.

Dataset	No SFT	Iter-1	Iter-2	Iter-3	Ablation
HellaSWAG-act.net	48.07	46.63	**48.49**	46.19	47.40
SWAG	48.28	**48.38**	47.88	47.03	47.70
αNLI	64.51	65.47	65.99	**66.51**	64.34
SCT	82.24	82.78	**82.80**	82.40	82.50
QuaRel	**60.89**	59.93	59.42	60.73	59.23
PIQA	67.03	**69.07**	68.04	68.24	67.90
WinoGrande	60.11	61.42	**61.83**	60.10	58.80
AI2Sci-elem	62.81	62.60	**65.57**	64.23	60.16
HellaSWAG-wikihow	**31.44**	30.65	30.97	29.34	30.90
CommonsenseQA	**47.05**	44.99	45.07	43.74	43.57
SocialIQA	42.64	41.59	**43.45**	41.64	42.58
ARC-challenge	50.00	**50.84**	47.32	48.66	46.49
WSC	63.00	**65.60**	64.79	63.96	63.16
CODAH	56.45	58.04	57.43	**58.76**	56.83
Com2Sense	**53.20**	51.15	52.17	52.69	51.48
WinoVenti	58.79	**59.80**	59.39	58.90	58.32
CycIC	43.44	**44.39**	39.42	40.32	42.95
QuaRTz	58.52	59.53	**60.63**	56.69	59.11
CommonsenseQA2.0	50.12	**51.15**	50.22	50.79	48.92
NumerSense	**44.00**	40.50	43.50	43.00	43.00
SciQ	68.19	68.21	68.47	**70.87**	67.80
OpenBookQA	41.28	43.43	42.11	**44.33**	43.20
COPA	75.86	**77.08**	77.06	76.75	74.80
QASC	40.53	39.72	39.33	**40.70**	36.61
ComVE (Task A)	48.59	**49.15**	48.35	48.14	48.55
ARC-easy	**63.15**	62.46	62.21	61.97	61.75
PROST	32.40	**35.46**	34.04	32.62	31.00
AI2Sci-middle	59.20	60.00	**63.20**	59.20	58.40

Table 9: Accuracy (%) of commonsense QA by Vicuna with and without iterative Self-SFT and with ablation study. Bold dataset names are seen datasets during self-SFT.

E.3 Effect of Hyper-parameter: Filter Ratio

We investigate the influence of the filter ratio $K\%$ by experimenting with various values of K. We assess the average answer accuracy with K taking values from the set $\{10, 30, 50, 70, 90, 100\}$ using a single iteration, testing on randomly selected 100 questions from each dataset (2,800 questions in total). We maintain a consistent number of training steps across different K values. As illustrated in Figure 11, our method achieves optimal results when $K\%$ is set to 50%.

E.4 An Example of Generated Stories Alongside the Self-SFT Iterations

We show an example of the generated stories by the Vicuna model before (iteration 0) and alongside (iteration 1-3) the self-SFT iterations. The commonsense score and BERT similarity both increase along with the iteration, and the answer becomes correct after iteration 2. This further demonstrates the effectiveness of our self-SFT method in generating better stories.

F Computational Costs

For the QA evaluations across the 28 datasets, we use an RTX 3090 GPU. Generating the five stories or rules for a single commonsense question typically takes about 20-30 seconds. Once the stories or rules are generated, answering a question with

Iteration	Story	Commonsense Score	BERT Similarity	Predicted Answer
0	Jane recalls a time when she was in her friend's kitchen and saw her using a glue stick to attach a decoration to a cake. Jane watched as her friend carefully applied the glue to the back of the decoration, then pressed it onto the cake. Jane was impressed by how smoothly the glue stuck to the cake and how easily the decoration stayed in place. She had never thought about adults using glue sticks before, but now she could see how useful they could be for a variety of tasks.	0.657	0.827	E. kitchen drawer ✗
1	As a child, Jane used to spend hours crafting with glue sticks. She would often use them to attach paper, cardboard, and other materials to create various projects. She would usually keep a glue stick and a small container of glue near her workspace, which was usually a table in the living room or her bedroom. She would also sometimes use glue sticks to fix small broken objects around the house, such as toys or household items.	0.692	0.830	C. at school ✗
2	Jane was working as an intern at a local law firm. One day, her supervisor asked her to help organize the **office** supplies. As she opened the drawer of the desk she was assigned, Jane noticed a package of glue sticks neatly arranged next to the stapler and scissors. Curiosity piqued, she inquired about the glue sticks. Her supervisor explained that the **office** frequently used them for various tasks such as attaching documents to presentation boards or assembling reports. Jane was surprised but intrigued by this unconventional use of glue sticks in an **office** setting.	0.809	0.848	D. office ✓
3	I remember one specific experience where I saw adults using glue sticks. It was during my first day of an **office** internship. I was assigned to help organize some files in the storage room when I noticed a group of adults huddled around a table. Curiosity got the better of me, so I approached them to see what they were doing. To my surprise, they were using glue sticks to attach important documents onto presentation boards. It was fascinating to watch them carefully apply the glue in a controlled manner, ensuring the papers were securely fixed. I realized that adults indeed use glue sticks in an **office** setting for various tasks.	0.810	0.874	D. office ✓

Table 10: An example of the generated stories of past experiences alongside the iteration of self-SFT. The question is shown in Figure 1.

them as context takes less than a second. Therefore, a complete evaluation of all datasets consumes approximately 170 hours, which spans slightly over a week.

In the iterative self-SFT method employed for fine-tuning the Vicuna model, we use an A100 GPU (80G). The training step in each iteration of the self-SFT method takes around 12 minutes. Additionally, the generation step in each iteration requires about 70 hours to process the 8 datasets.

G Licenses of Datasets and Models

The CommonsenseQA, NumerSense, HellaSWAG, SWAG, and WinoVenti datasets are licensed under the MIT License. The OpenBookQA, QASC, WinoGrande, and PROST datasets are licensed under the Apache-2.0 license. The PIQA dataset is licensed under the Academic Free License v. 3.0 license. The SocialIQA, CommonsenseQA2.0, QuaRel, and QuaRTz datasets are licensed under the CC-BY license. The SciQ dataset is licensed under the CC-BY-NC-3.0 license. The ARC and ComVE (Task A) datasets are licensed under the CC-BY-SA license. The αNLI dataset is licensed under the CC-BY-NC-4.0 license. The COPA dataset is licensed under the BSD 2-Clause license. The CODAH dataset is licensed under the Open Data Commons Attribution license. The AI2Sci, Com2Sense, CycIC, and SCT datasets have unknown licenses.

The terms of use for ChatGPT API are in https://openai.com/policies/terms-of-use. The Alpaca and Vera models are licensed under the MIT License. The Vicuna model is licensed under the Llama 2 Community License Agreement. The BERT model is licensed under the Apache-2.0 license.