LINKED: Eliciting, Filtering and Integrating Knowledge in Large Language Model for Commonsense Reasoning

Anonymous ACL submission

Abstract

Large language models (LLMs) sometimes demonstrate poor performance on knowledge-intensive tasks, commonsense reasoning is one of them. Researchers typically address these issues by retrieving related knowledge from knowledge graphs or employing self-enhancement methods to elicit knowledge in LLMs. However, noisy knowledge and invalid reasoning issues hamper their ability to answer questions accurately. To this end, we propose a novel method named ***eL**iciting, f**I**ltering and i**N**tegrating **K**nowledge in large languag**E** mo**D**el* (LINKED). In it, we design a reward model to filter out the noisy knowledge and take the marginal consistent reasoning module to reduce invalid reasoning. With our comprehensive experiments on two complex commonsense reasoning benchmarks, our method outperforms SOTA baselines (up to **9.8%** improvement of accuracy on WinoGrande). Besides, to measure the positive and negative impact of the injected knowledge, we propose a new metric called effectiveness-preservation score for the knowledge enhancement works. Finally, through extensive experiments, we conduct an in-depth analysis and find many meaningful conclusions about LLMs in commonsense reasoning tasks.

1 Introduction

Commonsense reasoning is one of the key abilities for models to reach artificial general intelligence (AGI). To measure it, researchers designed commonsense reasoning tasks (Talmor et al., 2019; Zellers et al., 2019; Sakaguchi et al., 2020), which require models to answer questions based on commonsense knowledge (see Figure 1 for examples). In recent works, large language models (LLMs) (e.g. PaLm (Chowdhery et al., 2022), GPT-4 (OpenAI, 2023), Llama2 (Touvron et al., 2023)) have improved performances in this task compared to small models. Nevertheless, there is still a considerable gap between them and humans. For instance,

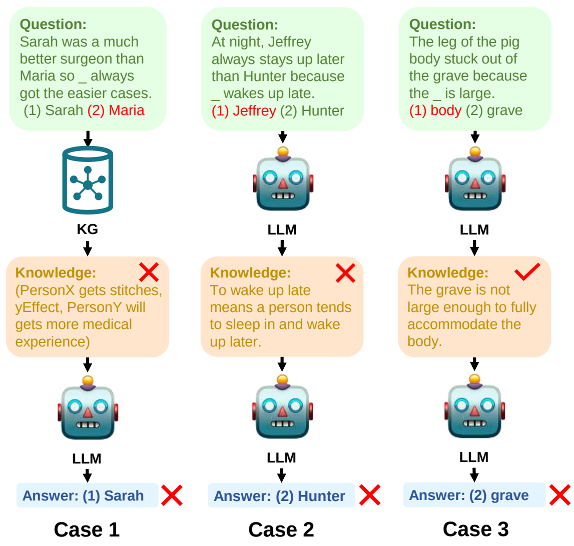

Figure 1: Some failed cases of traditional knowledge enhancement methods on complex commonsense reasoning tasks.

on WinoGrande (Sakaguchi et al., 2020), the accuracy of Llama2-70B is 80.2%, lagging more than ten points behind the 94.1% accuracy of humans (Touvron et al., 2023).

To further improve LLM's commonsense reasoning abilities, a series of works are proposed (Wei et al., 2022; Wang et al., 2023a; Wu et al., 2023), which can be mainly divided into two different lines: **(1) Retrieval augmentation.** This kind of method retrieves knowledge corresponding to the question from knowledge graphs (KGs), then integrates it into the model's input as supplementary information (Chen et al., 2023; Wang et al., 2023a). **(2) Self-enhancement.** This method employs a chain-of-thought (CoT) like prompting technique, empowering LLMs to generate the knowledge required for reasoning in the form of a rationale (Wei et al., 2022; Wang et al., 2023c; Li et al., 2023b).

Although these methods have made some progress, they still suffer from two main challenging problems: **(1) Noisy knowledge:** For the retrieval augmentation method, the limited-scale KGs cannot cover the knowledge required for complex commonsense reasoning scenarios. Thus, we can only retrieve noisy knowledge from KGs, which is irrelevant to our questions. As shown in Figure 1-1, for the question in WinoGrande, models need commonsense knowledge that describes the relation between "*be a better surgeon*" and "*get the easier cases*", but the most relevant knowledge "(*PersonX gets stitches*, *yEffect*, *PersonY will gets more medical experience*)" from ATOMIC-2020 (Hwang et al., 2021) is still far from what is required. For the self-enhancement method, some works have pointed out that the rationale generated by the LLM itself may contain severe noise (Zhao et al., 2023; Gao et al., 2023; Trivedi et al., 2023) that is harmful to reasoning. For example, in Figure 1-2, the generated knowledge indicates "*To wake up late means wake up later*", which is a piece of noisy information and leads to LLM's incorrect response "*Answer: Hunter*". **(2) Invalid reasoning:** Sometimes, even if reasonable knowledge is provided to LLM, it may still result in incorrect answers because of the invalid reasoning issue (Kojima et al., 2022; Lyu et al., 2023; Lanham et al., 2023). As illustrated in Figure 1-3, while the rationale "*The grave is not large enough to fully accommodate the body*" is correct for the question, LLMs still fail to draw the correct conclusions based on it. In our pilot experiment, the noisy knowledge issue accounts for 34% in all of the failure cases and the invalid reasoning issue accounts for 28%. Hence, these two issues are not negligible for further improving the LLM's commonsense reasoning abilities. [1]

In this paper, we propose a novel method named 𝕃𝕀ℕ𝕂𝔼𝔻 (***eLiciting, fIltering and iNtegrating Knowledge in large languagE moDel***) to enhance the commonsense reasoning abilities of LLMs with effective knowledge. **Firstly, we design the reward model to filter out the noisy knowledge generated by LLMs.** We define the confidence level of knowledge based on its contribution to question-answering and use it as a supervision signal for training the reward model. **Then, we propose marginal consistent reasoning to reduce invalid reasoning**. Given a rationale, the traditional CoT-like methods only perform the reasoning process once, which may lead to wrong outputs when the probability distribution of candidate answers is relatively random. To avoid it, we use one effective rationale, execute multiple rounds of reasoning based on it and select the answer with the highest marginal probability.

We evaluate our method on two commonsense reasoning benchmarks WinoGrande (Sakaguchi et al., 2020) and Hellaswag (Zellers et al., 2019). Since the traditional metric accuracy can not measure how much noisy knowledge the enhancement method brings, we propose a new metric named **effectiveness-preservation score (EPS)** to mitigate this gap. This metric measures both the positive and negative impact a knowledge augmentation method has on the model's reasoning. Extensive experiments on the two datasets show that our method brings significant improvements over baselines.

We summarize the contribution of this paper as follows: (1) We propose a novel method 𝕃𝕀ℕ𝕂𝔼𝔻 to enhance the performance of LLMs in commonsense reasoning tasks. Additionally, we introduce a novel metric EPS to evaluate both the effectiveness and harmfulness of knowledge augmentation methods. (2) In our method, we not only train a reward model to mitigate noisy knowledge in LLM's generations, but also devise the marginal consistent reasoning module to solve invalid reasoning problems. (3) We conduct extensive experiments on two benchmarks, demonstrating that our method outperforms SOTA methods. Impressively, we observe **9.8%** absolute accuracy improvement and **12.5%** EPS improvement over WinoGrande. Furthermore, we get several meaningful conclusions about LLM's commonsense reasoning based on the experimental results. We will release the source code if this paper is accepted.

2 Related Work

2.1 Chain-of-thoughts Prompting

Chain-of-thoughts prompting is a type of method that elicits the knowledge inside LLMs to enhance its reasoning abilities (Kojima et al., 2022; Wei et al., 2022). It shows excellent performance on tasks such as mathematical reasoning, symbolic reasoning, and commonsense reasoning, Recently, a line of new works focuses on further improving its performance. Some works design new structures for the mid steps of reasoning (Wang et al., 2023c; Yao et al., 2023a; Besta et al., 2023). Another series of work in the field of CoT is trying to generate higher quality rationales by referring to external

[1] In this experiment, we randomly choose 50 examples from failed cases on two commonsense benchmarks and analyze the corresponding reason.

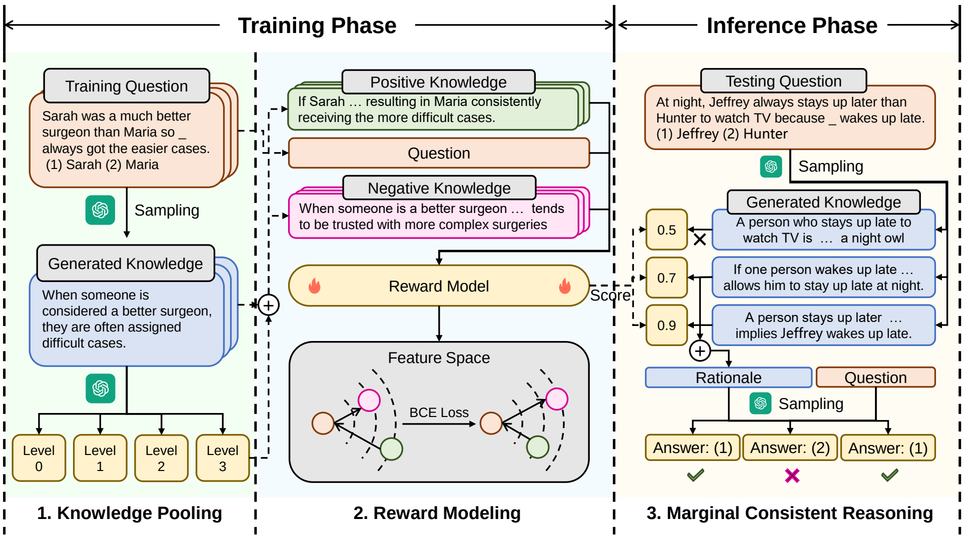

Figure 2: The main architecture of our proposed method $\mathbb{LINKED}$.

knowledge sources or tools (Wang et al., 2023b; Yao et al., 2023b; Zhao et al., 2023). By bringing fact-based knowledge into the model, LLMs can generate responses that have fewer hallucinations.

2.2 Knowledge Enhancement for LLMs

LLM has suffered from serious hallucination issues. To solve the problem, researchers retrieve knowledge from external sources to enhance the models. Several works get knowledge through search engines, they finetune models to imitate human's searching actions (Nakano et al., 2021) or use in-context learning to let the model generate API calls (Gao et al., 2023; Trivedi et al., 2023; Lu et al., 2023). Other work uses KGs (such as ConceptNet (Speer et al., 2017)) as knowledge resources, they train a retriever, use it to get sub-graphs or triples from the KG and embed this extra information into the input prompt of models (Yasunaga et al., 2021; Baek et al., 2023; Chen et al., 2023). Therefore, LLMs can generate knowledge-enhanced outputs.

3 Methodology

3.1 Overview

Figure 2 demonstrates the main architecture of our $\mathbb{LINKED}$ method, which is divided into two phases. In the training phase, we aim to train a reward model to address the issue of noisy knowledge. To this end, we first prepare the training data and define the confidence level of the knowledge to distinguish knowledge of different quality (§ 3.2). Then, we train the reward model using a ranking task based on the annotated data (§ 3.3). As for mitigating the invalid reasoning issue, we propose the marginal consistent reasoning module in the inference phase. We prompt LLMs to conduct multiple reasoning processes on one effective rationale and choose the final answer based on the marginal majority vote (§ 3.4). Below, we will provide a detailed illustration of each module in our approach.

3.2 Knowledge Pooling

Previous studies have demonstrated that LLMs inherently contain a vast amount of commonsense knowledge (Wang et al., 2022; Liu et al., 2022; Yuan et al., 2023). Thus, here we use LLM itself as the knowledge source. When provided with a question q in the training data, we use in-context learning to prompt the model and generate multiple pieces of related knowledge, denoted as $\mathcal{K}_q$. Then we instruct LLMs to predict answers to q, considering two scenarios: with access to k in $\mathcal{K}_q$ and without it:

$$r(q) = \mathcal{M}(q, P_d) \quad r(q,k) = \mathcal{M}(q, P_k, k)$$

Here, P_d is the prompt for LLMs to generate direct answer $r(q)$, while P_e is the prompt for LLMs to generate the answer $r(q,k)$ based on the provided knowledge k. $\mathcal{M}$ represents output of LLMs. Therefore, for each knowledge piece k, we can classify it into four different confidence levels according to the correctness of $r(q)$ and $r(q,k)$, which is defined as follows:

- ***Useful*** (Level-0) $\iff$ $r(q) \neq a^* \wedge r(q,k) = a^*$

Level	Question	Knowledge
0	The house on the hill needed some work on the floors but not the cabinets as the _ were ancient. **(1) floors** (2) cabinets (3) None	The fact that the floors needed work indicates that they were in poor condition and required attention or repairs.
1	Maria looked at Katrina, stretched out a hand and then _ accepted the handshake to introduce. (1) Maria **(2) Katrina** (3) None	When someone stretches out their hand, it is typically a gesture inviting a handshake as a form of introduction.
2	The woman wanted to put her hand inside the glove but the _ was too large. (1) hand **(2) glove** (3) None	The glove being too large implies that the hand of the woman was smaller in comparison.
3	So _ was worried because Randy forgot to study for the upcoming test and Robert studied. (1) Randy (2) Robert **(3) None**	Based on the information given, we cannot definitively determine whether Randy or Robert was worried.

Table 1: Some examples for questions, knowledge, and related knowledge level. We denote the correct option using red marking. The options chosen by the model before and after introducing knowledge are represented by underlining and **bold**, respectively.

- ***Harmless*** (Level-1) $\iff r(q) = a^* \wedge r(q,k) = a^*$
- ***Useless*** (Level-2) $\iff r(q) \neq a^* \wedge r(q,k) \neq a^*$
- ***Harmful*** (Level-3) $\iff r(q) = a^* \wedge r(q,k) \neq a^*$

Here a^* is the correct answer. Table 1 shows examples for each knowledge level. Notably, for a pair <q, k>, the effectiveness of knowledge k in enabling the model to answer the question q correctly decreases from level-0 to level-3. Level-0 knowledge can enhance LLMs to answer questions correctly that they couldn't initially handle. In contrast, level-3 knowledge leads to incorrect responses to commonsense questions that LLMs typically answer correctly. Hence, the level of knowledge can gauge its effectiveness and harmfulness, offering valuable supervised learning signals for us to train a reward model.

3.3 Reward Modeling

In this section, we focus on training a reward model to filter out noisy knowledge.

Training Data: We collect a set of <q,k> pairs and the corresponding knowledge level through the knowledge pooling module. To prepare training data, we need to further classify them into positive and negative examples with the label l. Considering the contribution of knowledge to answering questions, here a piece of knowledge k is defined as positive to the query q when its level is 0 or 1, otherwise, it is negative. We remove questions that related to only positive or negative knowledge during implementation.

Training Objective: According to former works, it is hard to differentiate the quality of commonsense knowledge using hard labels (Li, 2022; Li et al., 2023a). Thus, here we design the training objective as ranking rather than classification. Specifically, we encourage the reward model to give effective knowledge a higher score than the noisy one through the following objective function $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = \max\{f(q,k_t;\theta) - f(q,k_f;\theta)\}$$

Here k_t, k_f represent positive and negative knowledge in $\mathcal{K}$ to the question q, $f(\cdot;\theta)$ is the score predicted by the reward model. We use Deberta (He et al., 2023) model as a CrossEncoder to encode both q and k simultaneously, then produce a confidence score f between 0 and 1:

$$f(q,k;\theta) = CrossEncoder(q,k)$$

3.4 Marginal Consistent Reasoning

According to Wang et al. (2023c)'s work, the randomness in the model's output sampling may cause the invalid reasoning issue. As shown in the CoT case of Figure 3, even with a reasonable rationale, if we only sample the answer once, there remains a significant possibility of generating an incorrect option. From this perspective, to mitigate the problem, we need to adopt a more stable approach when sampling the answer.

In previous CoT-like works (Wang et al., 2023c; Zhao et al., 2023; Yao et al., 2023a), self-consistency is a critical method to make the final output more stable by exploring a large set of rationales. The key idea behind it can be expressed using the following formula:

$$\arg\max_a P(a|q) = \arg\max_a \sum_k P(a,k|q)$$

$$\sum_k P(a,k|q) \approx \frac{frequency(a)}{n} \propto frequency(a)$$

where a is the answer to question q, k is the generated rationale, and n is the sampling count. Based on it, we can choose the answer that receives the majority vote as the final prediction because of its highest frequency. However, when addressing difficult questions, The quality of each generated rationale is relatively random, leading to unstable answer distributions across different samplings based

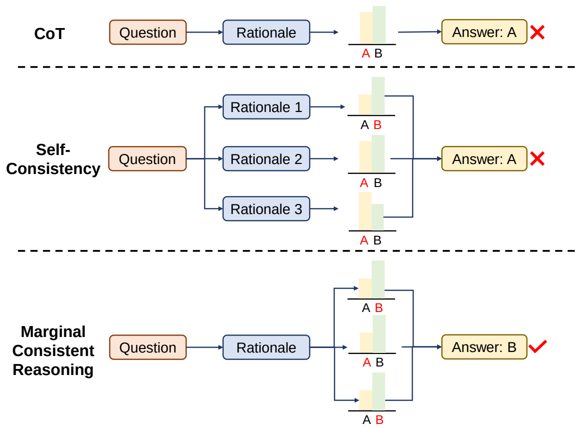

Figure 3: Comparison of different reasoning processes. The bars represent the probability distribution of options and the option marked in red indicates the final prediction in this sampling round.

on them. Therefore, we cannot guarantee the '≈' in the above equation to hold within a limited number of samplings. Like the Self-Consistency case in Figure 3, since it samples the answer only once for each rationale, it is easy to generate the wrong option when the probability distribution of different answers is relatively uniform (see Rationale 2 in the case).

To mitigate the above problem, we implement the marginal consistent reasoning module. The principle behind it is as below:

$$\arg\max_a P(a|q) \approx \arg\max_a P(a|k_0, q)$$

$$P(a|k_0, q) \approx \frac{frequency(a)}{n} \propto frequency(a)$$

Since it is unstable to continue to generate answers based on k in an auto-regressive manner, we use an effective rationale k_0 as the condition to shift the calculation goal from joint probability $P(a, k|q)$ to marginal probability $P(a|k_0, q)$. Hence, the search space for generating answers becomes smaller, which makes the sampling more stable. Besides, we also perform multi-round samplings for the answers. Through it, we can alleviate the impact of random rationale on the answer distribution, decreasing uncertainty during the sampling process. To make our method effective, we require a piece of k_0 that supports the correct answer's generation, holding the first '≈' in the equation. This is precisely the problem that is addressed in §3.3.

Specifically, the process of this module is illustrated in Figure 3. For each question, we utilize the reward model to rate the generated knowledge, select the top-k pieces of it and concatenate them to create an effective rationale. Then we integrate it into the input and prompt the LLM to conduct multi-round reasoning. The final output is determined by taking the majority vote on the answers. Through this module, we can mitigate the invalid reasoning issue by enhancing the stability of the LLM's reasoning process.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

We conduct experiments on two representative commonsense reasoning datasets: **WinoGrande** (Sakaguchi et al., 2020) and **HellaSwag** (Zellers et al., 2019). WinoGrande contains 44k binary choice questions, which are improved in scale and difficulty based on the WSC task (Levesque, 2011). HellaSwag contains 70k multiple-choice questions with four options each, focusing on completing WikiHow articles. Both datasets employ adversarial filtering to generate hard questions that require commonsense knowledge for solutions, posing a significant challenge for models. For each dataset, we use 5,000 questions from the original training set for training the reward model, and 500 questions from the development set as our testing set.

4.1.2 Baselines

We compare our method with these baselines:

- **Fine-tune**: We directly fine-tune `roberta-large` model (Liu et al., 2019) on the training data to predict the answer.
- **Retrieval augmentation**: We train a dense passage retrieval (DPR) model (Karpukhin et al., 2020) to fetch the most relevant documents from the constructed commonsense knowledge resource for the question. Then, the retrieved documents are used as supplementary knowledge input to the LLM for answering the question.
- **Self-enhancement**: We implement several self-augmentation methods, including: **Few-shot**, **CoT** (Wei et al., 2022), **CoT-SC** (Wang et al., 2023c), **Self-Refine** (Madaan et al., 2023), **Least-to-Most** (Zhou et al., 2023). Similar to our approach, these methods leverage the internal knowledge of the LLM to improve reasoning abilities.

We illustrate the details and prompts when implementing these baselines in the appendix A.

Methods	WinoGrande				HellaSwag			
	ACC	ES	PS	EPS	ACC	ES	PS	EPS
Fine-tune	64.0	-	-	-	68.6	-	-	-
Few-shot	70.6	-	-	-	67.8	-	-	-
DPR + LLM	<u>72.6</u>	41.5	<u>85.6</u>	55.9	<u>69.0</u>	23.8	**88.5**	37.6
CoT	69.2	45.6	79.0	57.8	64.4	28.5	79.9	<u>42.0</u>
CoT-SC	71.8	49.7	81.0	61.6	65.8	26.5	82.8	40.1
Self-Refine	61.4	46.3	67.7	55.0	49.0	25.2	59.3	35.3
Least-to-Most	70.2	<u>53.7</u>	77.1	<u>63.3</u>	47.2	<u>28.5</u>	55.3	37.6
Ours	**81.6 (+9.8)**	**66.7 (+13.0)**	**87.8 (+2.2)**	**75.8 (+12.5)**	**71.0 (+2.0)**	**33.1 (+4.6)**	<u>87.4 (-1.1)</u>	**48.0 (+6.0)**

Table 2: Comparison of $\mathbb{LINKED}$ performance with some strong baselines on GPT-3.5. The best results are highlighted in **bold**, while the second-best results are <u>underlined</u>.

4.1.3 Metrics

In traditional commonsense reasoning tasks, accuracy is almost the only metric. Nevertheless, it can not measure how much benefit or harm the knowledge-enhancement method brings. For example, suppose a method produces three pieces of **level-1** knowledge and two pieces of **level-3** knowledge, it performs as well as another method producing three pieces of **level-0** knowledge and two **level-2** knowledge in accuracy. But in practice, the latter performs better since it does not harm the model's original reasoning performance. Therefore, a more detailed metric is needed to evaluate these effects. To make up for the issue, we design a novel metric called **effectiveness-preservation score (EPS)**, which is calculated as follows:

$$ES = \frac{|\{q|r(q,k) = a^* \land q \in \mathcal{Q}_{false}\}|}{|\mathcal{Q}_{false}|}$$

$$PS = 1 - \frac{|\{q|r(q,k) \neq a^* \land q \in \mathcal{Q}_{true}\}|}{|\mathcal{Q}_{true}|}$$

$$EPS = \frac{2 * ES * PS}{ES + PS}$$

where $\mathcal{Q}_{true}$ and $\mathcal{Q}_{false}$ represent sets of correct and incorrect cases of the model directly answering questions under few-shot settings. ES quantifies the method's effectiveness in elevating the model's performance on questions it couldn't answer previously, while PS measures the extent to which the method has caused damage to questions the model could initially respond to correctly. When EPS approaches 1, the method can generate effective knowledge that enhances the model's performance. Conversely, when it approaches 0, it implies the method impedes the model's accurate output.

4.1.4 Implementation Details

We use `gpt-3.5-turbo-0613` provided by OpenAI as the LLM in our work. In our method, we set the temperature to 1.3 and the sample count to 5 when generating knowledge. As for the reasoning step, we set the temperature to 0.7 and the sampling count to 3. `Deberta-v3-large` is selected as our reward model. When fine-tuning it, we set batch size to 4, learning rate to 3×10^{-4}, warm-up steps to 500, epoches to 3. All experiments are conducted using 4 NVIDIA GeForce RTX 3090 GPUs.

4.2 Main Results

The main result of our experiments is presented in Table 2, from which we can obtain three key conclusions: **(1) Our method effectively enhances the LLM's commonsense reasoning performance.** For different datasets and metrics, our work significantly surpasses existing SOTA methods. Impressively, on WinoGrande, our method exhibits a significant improvement in accuracy, boasting an increase of **9.8%**, and enhances EPS by **12.5%**. These results demonstrate its ability to enhance LLM's commonsense reasoning abilities while maintaining a good balance between effectiveness and harmfulness. **(2) Retrieval augmentation method lacks effective introduction of commonsense knowledge.** From the results, we can see that the DPR + LLM method can reduce the damage to original reasoning (high PS of **85.6%** and **88.5%**). However, compared to other baselines, it performs worse in introducing effective knowledge (low ES of **41.5%** and **23.8%**). **(3) Self-enhancement method can cause significant harm to the model's commonsense reasoning.** As we have mentioned in § 1, LLM may generate harmful rationale when applying self-enhancement methods, our results align with this opinion. For example, the PS of the Least-to-Most method on HellaSwag is only **52%**, indicating that almost half of the correct reasoning on questions will turn wrong because of this method.

Method	Wino.	Hella.
LINKED	**81.6**	**71.0**
-w/o Reward model	78.0	68.6
-w/o Marginal consistent reasoning	80.0	69.2
-w/o both	78.4	69.4

Table 3: Ablation experimental results for our approach, here we only use accuracy for evaluation.

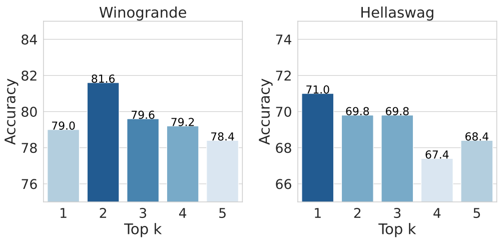

Figure 4: Different top-k value comparison.

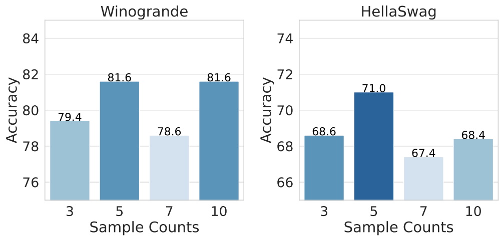

Figure 5: Different sample counts comparison.

Model	Filtered	Normal	None
Llama2-7B	**72.4**	67.8	57.2

Table 4: Accuracy comparison of different injected knowledge types on WinoGrande. 'Filtered' means we inject the filtered knowledge, 'Normal' means we directly inject the generated knowledge, 'None' means we do not inject any knowledge.

4.3 Ablation Study

To further verify the effectiveness of the different components in our method, we conduct ablation experiments (see Table 3). The following conclusions can be drawn from the experimental results: **(1) Effectiveness of both modules:** After we remove any of the two modules, the accuracy decreases, which indicates both the reward model and Marginal Consistent Reasoning module can successfully improve the performance of LLMs. **(2) Importance of the reward model:** Compared to -1.6% and -1.8% on accuracy after removing the marginal consistent reasoning module, the lack of the reward model drops by 3.6% and 2.2%. This demonstrates that our reward model assumes a more prominent role in the model's knowledge enhancement.

4.4 Effectiveness Analysis

4.4.1 Factors influence the performance

In our experiments, various factors can impact the performance. In this section, we aim to draw general conclusions by observing the effects of some.

Top-k knowledge: The top-k knowledge is selected to construct the final rational in the inference time, we change this value and compare their difference, whose results are shown in Figure 4. We find that the optimal value for top-k is **no more than 2**. Increasing it beyond 3 leads to a decline in performance, diminishing the effectiveness of the rationale. Compared to the introduction of a large volume of relevant knowledge, the filtration of knowledge is more crucial for LLMs, given their high sensitivity to input noise.

Sampling counts: In our default settings, we prompt the LLM to generate 5 pieces of knowledge for each question. Here, we change it to figure out whether more sampling counts make it more likely to bring effective knowledge. As illustrated in Figure 5, the number of effective knowledge produced by a model does not directly correlate with the sampling count. Without external sources of knowledge, LLMs exhibit significant quality fluctuations between multiple rounds of generation.

4.4.2 Generalization of filtered knowledge

In essence, we assess the effectiveness of knowledge using signals provided by LLM itself. This leads to a new question: Does this signal possess generality? In other words, can the more effective knowledge selected by our reward model also better enhance other small models' commonsense reasoning abilities? In this section, we aim to figure out this question through experiments.

Here we choose `Llama2-7b-chat` as the small model. Since it can not directly utilize the knowledge from the prompts due to its small scale (the accuracy of it on WinoGrande is around 52%), we first fine-tune it with labeled question-knowledge pairs. After that, we inject different kinds of knowledge into the model, comparing their performance on WinoGrande (see Table 4). We can get that the accuracy increases by **15.2%** after integrating filtered knowledge, which is **4.6%** higher than the injection of normal knowledge. This indicates that the filtered knowledge has generalization across different models in knowledge enhancement sce-

Dataset	Question	Knowledge	Ranking	Human Preference	Reason
WinoGrande	At night, Jeffrey always stays up later than Hunter to watch TV because _ wakes up late. **(1) Jeffrey** (2) Hunter	A person stays up later than another person to watch TV because he does not need to wake up early in the morning ...	1	✓	Contain the reasoning to the correct answer
		If a person ... suggests that Hunter, in this case, wakes up late and consequently stays up later than Jeffrey to watch TV	5	✗	Contain wrong reasoning
HellaSwag	The boy lifts his body above the height of a pole. The boy lands on his back on to a red mat. the boy _ (1) turns his body around on the mat. **(2) gets up from the mat**. (3) ...	When someone falls on their back, it is common for them to turn their body around or get up from the ground afterwards.	1	✓	Contain the reasoning to the correct answer
		When someone lands on their back, they are generally positioned lying down.	5	✗	Too general, no help for answering the question.

Table 5: Examples in case study. The correct answer to the question is **bolded**, some noisy reasoning is marked in red, and some correct reasoning is marked in blue.

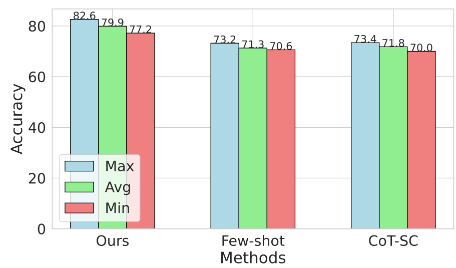

Figure 6: The robustness experiment on Winogrande.

narios, highlighting the critical value of our work in downstream applications.

4.4.3 The robustness of our method

In real application, a question may be asked repeatedly, and the model also needs to generate responses in multiple rounds. In that case, we aim to investigate whether our method can maintain consistent performance in multi-turn generation scenarios. As depicted in Figure 6, we conduct five repetitions of our method (only the inference phase) and two baselines, recording the maximum, average, and minimum accuracy values for comparison. It shows that throughout multiple rounds of generation, our work maintains a consistent edge over the performance of baselines ($> 7\%$ on accuracy).

4.5 Case Study

In this section, we want to further validate the effectiveness of our reward model by humans. We randomly choose a question for each benchmark and compare knowledge with different ranks provided by our reward model (see Table 5). For the first question, the knowledge ranked 1st contains the key evidence that leads to the correct answer (marked in blue), while the knowledge ranked 5th contains the wrong statement (marked in red) without any evidence to support it. As for the second question, the knowledge ranked 1st also contains the reasonable reasoning path to the correct answer, but the knowledge ranked 5th just describes the information in the question without any useful evidence to answer it. In conclusion, we demonstrate that knowledge with higher scores in our work is also more reasonable from a human perspective, indicating that the reward model can be aligned with humans to a certain extent.

5 Conclusion

In this paper, we propose a novel method named LINKED to enhance the LLM's performance on commonsense reasoning tasks. Specifically, we train a reward model to filter out noisy knowledge in LLM's generation and take the marginal consistent reasoning module to reduce invalid reasoning. Besides, we design a new metric named EPS to evaluate both the effectiveness and harmfulness of different knowledge enhancement methods, which the former metric can not. We conduct comprehensive experiments on Wingrande and HellaSwag benchmarks, and experimental results demonstrate that our method significantly outperforms previous baselines.

Limitations

While our method significantly improves LLM's performance in commonsense reasoning tasks, it has two primary limitations: (1) The black-box nature of the LLM we study hinders our ability to delve deeper into the model and explain why the filtered knowledge is effective. (2) Due to time and resource constraints, we were unable to conduct extensive prompt design work, which could have further improved our method's performance.

References

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. CoRR, abs/2306.04136.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of thoughts: Solving elaborate problems with large language models. CoRR, abs/2308.09687.

Zichen Chen, Ambuj K. Singh, and Misha Sra. 2023. Lmexplainer: a knowledge-enhanced explainer for language models. CoRR, abs/2303.16537.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. CoRR, abs/2204.02311.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: researching and revising what language models say, using language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 16477–16508. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

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.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 6769–6781. Association for Computational Linguistics.

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

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning. CoRR, abs/2307.13702.

Hector J. Levesque. 2011. The winograd schema challenge. In Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011. AAAI.

Jiachun Li, Chenhao Wang, Yubo Chen, Kang Liu, and Jun Zhao. 2023a. What events do pre-trained language models learn from text? probing event-based commonsense knowledge by confidence sorting. In Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part I, volume 14302 of Lecture Notes in Computer Science, pages 669–681. Springer.

Xiang Li. 2022. Probabilistic commonsense knowledge.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making language models better reasoners with step-aware verifier. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5315–5333.

Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin

Choi. 2022. Rainier: Reinforced knowledge introspector for commonsense question answering. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 8938–8958. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. CoRR, abs/2304.09842.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. CoRR, abs/2301.13379.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. CoRR, abs/2112.09332.

OpenAI. 2023. GPT-4 technical report. CoRR, abs/2303.08774.

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.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4149–4158. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 10014–10037. Association for Computational Linguistics.

Chenhao Wang, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2022. Cn-automic: Distilling chinese commonsense knowledge from pretrained language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 9253–9265. 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. CoRR, abs/2306.06427.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023b. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. CoRR, abs/2308.13259.

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

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

Dingjun Wu, Jing Zhang, and Xinmei Huang. 2023. Chain of thought prompting elicits knowledge augmentation. In Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, pages 6519–6534. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. CoRR, abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, pages 535–546. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 4364–4377. Association for Computational Linguistics.

Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Robert Jankowski, Yanghua Xiao, and Deqing Yang. 2023. Distilling script knowledge from large language models for constrained language planning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 4303–4325. Association for Computational Linguistics.

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

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 5823–5840. Association for Computational Linguistics.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

A Baseline Implementation Details

A.1 Fine-tune

We train `roberta-large` model on 5,000 QA pairs, of which we divide 500 samples as the validation set. For the hyper-parameters in training, we set the batch size to 64, epochs to 2, learning rate to 5×10^{-5}, gradient accumulation steps to 16, and warm-up steps to 300.

A.2 Retrival Augmentation

We use the relevant data provided in Yu et al. (2022)'s work for the corpus and training set. Besides, we use `bert-base-uncased` as the base model to train the retriever. When training, we set the batch size to 16, learning rate to 2×10^{-5}, linear warm-up steps to 1237 and epochs to 20.

A.3 Self-enhancement

We use 3-shot prompts for Few-shot, CoT, CoT-SC and 5-shot prompts for Self-Refine, Least-to-Most. Figure 7,8, 9 and 10 show parts of the prompts on WinoGrande. We will release all of the prompts if the paper is accepted.

System Instruction: You are a helpful assistant that use your own knowledge to choose the correct answer to the question. Use your commonsense knowledge to choose correct answer for some questions. Your response should be in this form:
'Answer: ({option}) {answer}'
If there is not proper option, you can give 'Answer: None'.
Now answer the following questions:

Question: The test was hard for Samuel but a breeze for Randy , since _ had failed to study for it.
(1) Samuel (2) Randy
Answer: (1) Samuel.

Question: Kyle slowly wormed their way into the life of Derrick, because _ was good and manipulating people.
(1) Kyle (2) Derrick
Answer: (1) Kyle.

Question: Donald was very grounded but Michael often got lost in their daydreams. _ was very capricious all the time.
(1) Donald (2) Michael
Answer: (2) Michael.

Question: {}

Figure 7: Prompts for Few-shot.

System Instruction: You are a helpful assistant that break down the question step by step and choose the correct answer to the question**.** Use your commonsense knowledge to choice correct answer for some questions and give the reasoning process. Your response should be in this form:
'{Reasoning_content}
So the answer is: ({option}) {answer}'
Now answer the following questions:

Question: The test was hard for Samuel but a breeze for Randy , since _ had failed to study for it.
(1) Samuel (2) Randy
Answer: To pass a test, a person need to study for it. If a person feel the test like a breeze, it means the test is easy for him. A person feels the test easy, because he studies hard for it. Since we know that Samuel feel the test very hard, she may fail to study for it.
So the answer is: (1) Samuel.

Question: Kyle slowly wormed their way into the life of Derrick, because _ was good and manipulating people.
(1) Kyle (2) Derrick
Answer: A person wormes his way into other's life, because he is friendly and approachable. A friendly person is considered good. A person is seen as manipulating people, that means he like to interact with others and others like him. Since Kyle slowly wormed their way into the life of Derric, Kyle will be seen as good and manipulating people.
So the answer is: (1) Kyle.

Question: Donald was very grounded but Michael often got lost in their daydreams. _ was very capricious all the time.
(1) Donald (2) Michael
Answer: A person is grounded means he works hard and does not like to fantasize. A person often gets lost in his daydreams, he is seen as unrealistic and egocentric. A person is capricious all the time means he does everything only according to his own ideas. Since Michael often gets lost in their daydreams but Donald does not, Michael is seen as very capricious all the time.
So the answer is: (2) Michael.

Question: {}

Figure 8: Prompts for CoT and CoT-SC.

System: You are a helpful, respectful and honest assistant. You should use your reasoning abilities to give a feedback to the given rational. Your response should be in this form: 'Feedback: {feedback}'.
Question: Kyle slowly wormed their way into the life of Derrick, because _ was good and manipulating people.\n(1) Kyle (2) Derrick\nRational: A person wormes his way into other's life, because he is friendly and approachable. A person is seen as manipulating people, that means he like to interact with others. Since Kyle slowly wormed their way into the life of Derric, Derric will be seen as good and manipulating people. So the answer is: (2) Derric.
Answer: Feedback: The rational is wrong. Since a person wormes his way into other's life is friendly, it's Kyle who was good, not Derrick.
…
Question: {}

System: You are a helpful, respectful and honest assistant. You should use your reasoning abilities, the given rational and feedback to update your answer to the given questions in reasoning tasks. You should reply the correct rationales and the answer. Your response should be in this form: '{reason} So the answer is: ({option}) {answer}'. If you don't know the answer to a question, please reply 'Answer: None'.
Question: Kyle slowly wormed their way into the life of Derrick, because _ was good and manipulating people.\n(1) Kyle (2) Derrick\nRational: A person wormes his way into other's life, because he is friendly and approachable. A person is seen as manipulating people, that means he like to interact with others. Since Kyle slowly wormed their way into the life of Derric, Derric will be seen as good and manipulating people. So the answer is: (2) Derric.
Feedback: The rational is wrong. Since a person wormes his way into other's life is friendly, it's Kyle who was good, not Derrick.
Answer: A person wormes his way into other's life, because he is friendly and approachable. A friendly person is considered good. A person is seen as manipulating people, that means he like to interact with others and others like him. Since Kyle slowly wormed their way into the life of Derric, Kyle will be seen as good and manipulating people. So the answer is: (1) Kyle.
…
Question: {}

Figure 9: Prompts for Self-Refine.

System: You are a helpful, respectful and honest assistant. You should use your reasoning abilities to break down the questions into subquestions. You should reply the correct subquestions. Your response should be in this form: 'To solve the question, we need to solve these subquestions:\nQuestion 1:{subquestion}\nQuestion 2:{subquestion}.
Question: The test was hard for Samuel but a breeze for Randy , since _ had failed to study for it.\n(1) Samuel (2) Randy
Answer: To solve the question, we need to solve these subquestions:\nQuestion 1: Why the test is hard for Samul?\nQuestion 2:Why the test is a breeze for Randy?\nQuestion 3:Who had fail to study for the test?
…
Question: {}

System: You are a helpful, respectful and honest assistant. You should use your reasoning abilities to answer the given subquestion in reasoning tasks. You should reply the correct answer to the subquestion. Your response should be in this form: 'Answer: {answer}'.
Question: The test was hard for Samuel but a breeze for Randy , since _ had failed to study for it.\n(1) Samuel (2) Randy\nQuestion 1: Why the test is hard for Samul?
Answer: Answer: If the test is hard for Samul, he may not study for it.
…
Question: {}

System: You are a helpful, respectful and honest assistant. You should use your reasoning abilities and the given context to answer the given questions in reasoning tasks. You should reply the answer. Your response should be in this form: 'So the answer is: ({option}) {answer}'. If you don't know the answer to a question, please reply 'Answer: None'.
Question: The test was hard for Samuel but a breeze for Randy , since _ had failed to study for it.\n(1) Samuel (2) Randy\nQuestion 1: Why the test is hard for Samul? Answer: If the test is hard for Samul, he may not study for it.\nQuestion 2:Why the test is a breeze for Randy? Answer: If Randy feel the test like a breeze, the test is easy for her. In that case, she may study hard for it.\nQuestion 3:Who had fail to study for the test? Answer: Since Samul does not study for the test, Samul fails to study for it.
Answer: So the answer is: (1) Samuel.
…
Question: {}

Figure 10: Prompts for Least-to-Most.