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 *e**L**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 four complex commonsense reasoning benchmarks, our method outperforms SOTA baselines (up to **9.0%** improvement of accuracy). 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. PaLM2 (Anil et al., 2023), 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, 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).

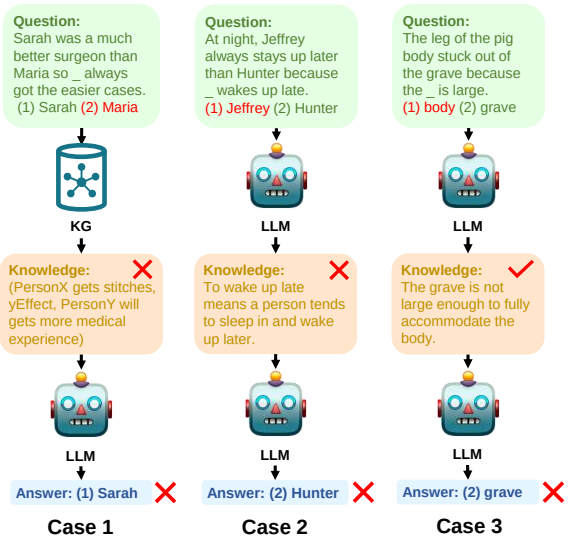


Figure 1: Some failed cases of traditional knowledge enhancement methods on complex commonsense reasoning tasks.

To further improve LLM's commonsense reasoning abilities, a series of works are proposed (Wang et al., 2023a; Wu et al., 2023; Li et al., 2024), which can be mainly divided into two different lines: **(1) Retrieval augmentation.** As shown in Case 1 of Figure 1, these methods retrieve knowledge corresponding to the question from knowledge graphs (KGs), then integrate it into the model's input as supplementary information (Chen et al., 2023; Wang et al., 2023a). **(2) Self-enhancement.** As illustrated in Case 2 and 3 of Figure 1, these methods employ 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., 2023). For the

former method, considering the limited coverage of commonsense knowledge by KGs and the fact that the retriever can only capture the semantic similarity of entities, it struggles to recall effective information in complex commonsense reasoning scenarios (e.g. event-based reasoning). As shown in Case 1 of Figure 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. Hence, the self-enhancement method becomes the dominant method for LLM augmentation in commonsense reasoning.

Our work follows the self-enhancement approach. Although these methods have made some progress, they still suffer from two main challenging problems: **(1) Noisy knowledge:** 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 Case 2 of Figure 1, 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 the LLM, it may still result in incorrect answers (Kojima et al., 2022; Lyu et al., 2023; Lanham et al., 2023). We define this situation as the 'invalid reasoning' issue. As illustrated in Case 3 of Figure 1, 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 experiments, the noisy knowledge issue accounts for 34% in all of the failure cases and the invalid reasoning issue accounts for 28%[1]. Hence, these two issues are not negligible for further improving the LLM's commonsense reasoning abilities.

In this paper, we propose a novel method named LINKED (*e**L**iciting, f**I**ltering and i**N**tegrating **K**nowledge in large languag**E** mo**D**el*) 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 the marginal consistent reasoning module 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 uniform. 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 extensive commonsense reasoning benchmarks. 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. Experimental results 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 issues.

(3) We conduct extensive experiments on two benchmarks, demonstrating that our method outperforms SOTA methods. Impressively, we observe up to **9.0%** accuracy improvement and **12.5%** EPS improvement. 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 Commonsense Reasoning Enhancement

Commonsense reasoning is a crucial capability that language models must master to progress toward AGI. However, since commonsense knowledge is rarely explicitly expressed in texts, models perform poorly on these tasks and require additional enhancement (Talmor et al., 2019; Sakaguchi et al.,

[1]In this experiment, we randomly choose 50 examples from failed cases on different benchmarks and analyze the corresponding error types.

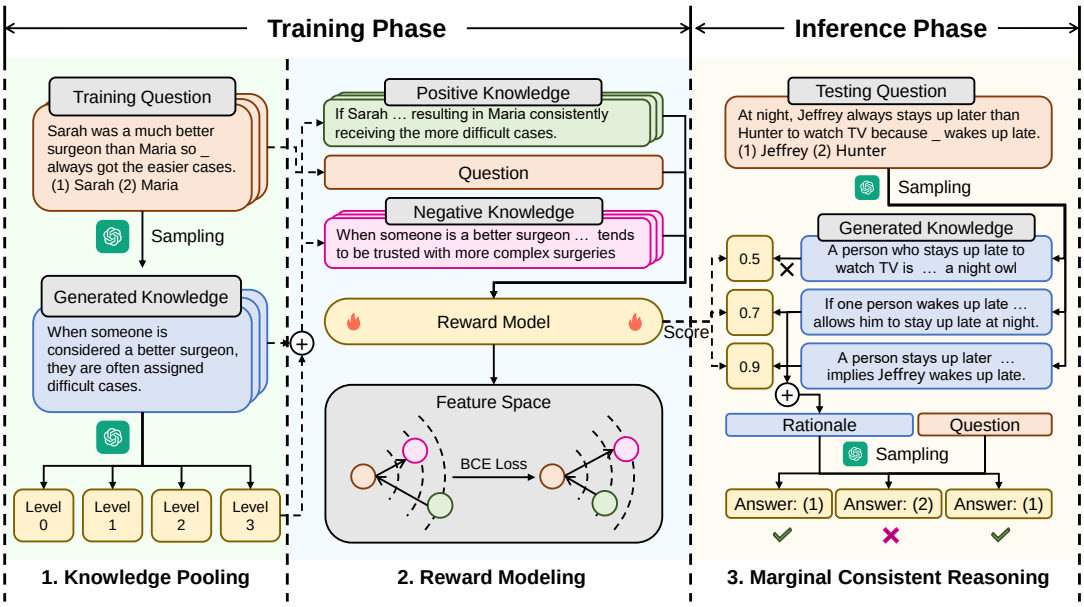


Figure 2: The main architecture of our proposed method LINKED.

2020; Wang et al., 2022). Traditional works usually fine-tune the model on synthetic commonsense datasets, but they incur high costs, and the models trained are difficult to apply to new commonsense reasoning tasks directly (Hwang et al., 2021; Khashabi et al., 2020; Lourie et al., 2021). Recently, the excellent in-context learning (ICL) capabilities of LLMs allow us to enhance their commonsense reasoning abilities without extra training. Specifically, we can supplement the additional commonsense knowledge through retrieval augmentation (Yu et al., 2022; Chen et al., 2023; Wang et al., 2023a) or self-enhancement methods (Wei et al., 2022; Wu et al., 2023; Li et al., 2023). Our work follows self-enhancement methods, while addressing the issues of noisy knowledge and invalid reasoning in previous methods.

2.2 Knowledge Enhancement for LLMs

LLMs have suffered from serious hallucination issues. To solve the problem, researchers retrieve related knowledge to enhance the models. Firstly, 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). Secondly, other works use KGs (such as ConceptNet (Speer et al., 2017)) as knowledge resources, they train a retriever, use it to get subgraphs 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). At last, researchers also elicit the knowledge inside LLMs to enhance themselves. They design new structures for the mid steps of reasoning (Yao et al., 2023a; Besta et al., 2023; Li et al., 2024) or generate higher quality rationales by referring to external knowledge sources or tools (Wang et al., 2023b; Yao et al., 2023b; Zhao et al., 2023). Our work aims to get high-quality commonsense knowledge from LLMs to further enhance their commonsense reasoning performances.

3 Methodology

Figure 2 demonstrates the main architecture of our 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 (§ 3.1). Then, we train the reward model using a ranking task based on the annotated data (§ 3.2). 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 (§ 3.3).

3.1 Knowledge Pool Construction

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:

$$r(q) = \mathcal{M}(q, P_d) \quad (1)$$
$$r(q, k) = \mathcal{M}(q, P_k, k) \quad (2)$$

Here, P_d is the prompt for LLMs to generate direct answer $r(q)$, while P_k 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):** $r(q) \neq a^* \ \wedge \ r(q, k) = a^*$
- ***Harmless*** **(Level 1):** $r(q) = a^* \wedge r(q, k) = a^*$
- ***Useless*** **(Level 2):** $r(q) \neq a^* \wedge r(q, k) \neq a^*$
- ***Harmful*** **(Level 3):** $r(q) = a^* \wedge r(q, k) \neq a^*$

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 knowledge level can gauge its effectiveness and harmfulness, offering supervised learning signals to train a reward model.

3.2 Reward Model Design

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.

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.

Training Objective Here we encourage the reward model to give effective knowledge a higher score than the noisy one through the following objective function $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta) = -ylog(f(q, k; \theta)) - (1 - y)log(1 - f(q, k; \theta)) \quad (3)$$

where y represents the knowledge label and $f(\cdot; \theta)$ is the score predicted by the reward model. We use the 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. More training details and performances about our reward model are presented in Appendix A.

3.3 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:

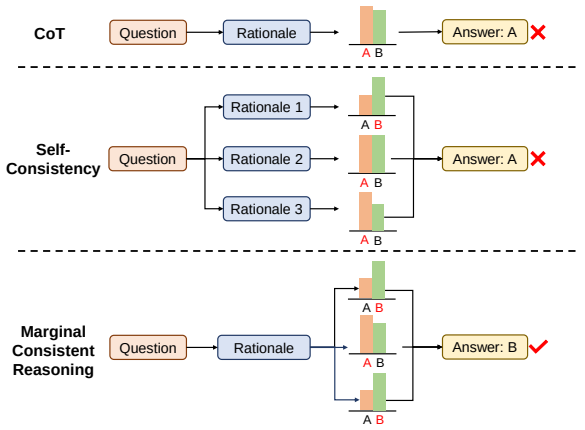


Figure 3: Comparison of different reasoning processes. The bars represent the probability distribution of options and **the option marked in red indicates the final prediction** in this sampling round.

$$\arg\max_a P(a|q) = \arg\max_a \sum_k P(a,k|q) \quad (4)$$

$$\sum_k P(a,k|q) \approx \frac{frequency(a)}{n} \propto frequency(a) \quad (5)$$

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 rationale is relatively random, leading to unstable answer distributions across different samplings based on them. Therefore, we cannot guarantee the '$\approx$' in the above equation to hold within a limited number of samplings. Like the Self-Consistency case in Figure 3, it is easy to select 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:

$$\arg\max_a P(a|q) \approx \arg\max_a P(a|k^*,q) \quad (6)$$

$$P(a|k^*,q) \approx \frac{frequency(a)}{n} \propto frequency(a) \quad (7)$$

Since it is unstable to continue to generate answers based on k in an auto-regressive manner, we use an effective rationale k^* as the condition to shift the calculation goal from joint probability $P(a,k|q)$ to marginal probability $P(a|k^*,q)$. 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 further decrease uncertainty during the sampling process. To make our method effective, we require a piece of k^* that supports the correct answer's generation, holding the first '$\approx$' in the equation. This is precisely the problem that is addressed in §3.2.

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 k^*. 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

Datasets We conduct experiments on four representative commonsense reasoning datasets: **WinoGrande (Wino)** (Sakaguchi et al., 2020), **HellaSwag (Hella)** (Zellers et al., 2019), **SocialIQA (SIQA)** (Sap et al., 2019) and **PIQA** (Bisk et al., 2020). For each dataset, we use 500 samples from the development set as our testing set. We present more details and discussions in Appendix B.1.

Baselines We include the following baselines in our experiments:

Few-shot. We prompt the LLM to directly answer questions in the test set through ICL.

Fine-tuning. We fine-tune the Roberta-large model (Liu et al., 2019) on the training data and use it to predict answers. Besides, we also apply two traditional SOTA methods: **UnifiedQA** (Khashabi et al., 2020) and **Unicorn** (Lourie et al., 2021).

Retrieval augmentation. For retrieval augmentation methods, we implement two baselines, **BM25** and **dense passage retrieval (DPR)** (Karpukhin et al., 2020), to retrieve additional commonsense knowledge from knowledge sources.

Self-enhancement. We implement several self-augmentation methods, including: **CoT** (Wei et al., 2022), **CoT-SC (SC)** (Wang et al., 2023c), **Self-Refine (SR)** (Madaan et al., 2023), **Least-to-Most (LtM)** (Zhou et al., 2023).

We illustrate the details and prompts when implementing these baselines in Appendix B.2.

Methods	WinoGrande		HellaSwag		SocialIQA		PIQA	
	ACC	EPS	ACC	EPS	ACC	EPS	ACC	EPS
Few-shot	70.6	0.0	67.8	0.0	71.9	0.0	78.2	0.0
Fine-Tuning Method								
Roberta-large	64.0	-	68.6	-	73.1	-	62.2	-
Unified QA	62.0	-	34.4	-	63.0	-	78.6	-
Unicorn	72.6	-	27.2	-	**74.8**	-	78.2	-
Retrieval Augmentation Method								
BM25 + LLM	64.0	25.7	45.6	38.2	55.0	36.3	59.0	21.1
DPR + LLM	65.6	55.9	60.6	37.6	67.2	48.6	73.2	57.1
Self-Enhancement Method								
CoT	69.2	57.8	64.4	42.0	67.1	40.1	82.8	63.8
CoT-SC	71.8	49.7	65.8	40.1	72.3	48.5	85.4	67.5
Self-Refine	61.4	55.0	49.0	35.3	69.0	47.8	80.4	67.9
Least-to-Most	70.2	63.3	47.2	37.6	72.6	51.3	82.2	64.4
LINKED	**81.6 (+9.0)**	**75.8 (+12.5)**	**71.0 (+2.4)**	**48.0 (+6.0)**	73.5 (-1.3)	**55.3 (+4.0)**	**86.0 (+0.6)**	**69.8 (+1.9)**

Table 2: Comparison of LINKED performance with some strong baselines on GPT-3.5. The best results are highlighted in **bold**, while the second-best results are underlined. '-' indicates the method applies different models thus can not compute EPS.

Metrics In traditional 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 measure how many wrong answers are corrected by the method (effectiveness) and how many correct answers are made incorrect (harmfulness). To make up for the issue, we design a novel metric called **effectiveness-preservation score (EPS)** as follows:

$$ES = \frac{|\{q|r(q,k) = a^* \wedge q \in \mathcal{Q}_{false}\}|}{|\mathcal{Q}_{false}|} \quad (8)$$

$$PS = 1 - \frac{|\{q|r(q,k) \neq a^* \wedge q \in \mathcal{Q}_{true}\}|}{|\mathcal{Q}_{true}|} \quad (9)$$

$$EPS = \frac{2 * ES * PS}{ES + PS} \quad (10)$$

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. The ES quantifies the method's effectiveness in improving the model's performance on previously unanswered questions, while the PS measures the method's detrimental impact on questions the model initially answered correctly. Our EPS metric provides a measurement of the impact on both aspects.

Implementation Details In this work, we utilize `gpt-3.5-turbo-0613` provided by OpenAI as the LLM and `Deberta-v3-large` as the backbone of our reward model. For generation parameters, 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. 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 two key conclusions: **(1) Our method effectively enhances the LLM's commonsense reasoning performance.** For different datasets, our work significantly surpasses most existing SOTA methods. Impressively, on WinoGrande, our method exhibits a significant **9.0%** improvement in accuracy. **(2) Our method maintains a good balance between effectiveness and harmfulness.** On average, we improve EPS by **5.4%**, demonstrating that our method can introduce effective knowledge while avoiding damage to the LLM's original reasoning capabilities. We validate the robustness and generalizability of the results in Appendix C.

4.3 Ablation Study

To 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) Both modules are effective.** After we remove any

Method	Wino	Hella	SIQA	PIQA
LINKED	**81.6**	**71.0**	**73.5**	**86.0**
-w/o RM	78.0	68.6	71.9	82.0
-w/o MCR	80.0	69.2	71.7	85.6
-w/o both	78.4	69.4	72.7	82.2

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

Question	Knowledge	Ranking	Human
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	✓
	If a person ... suggests that Hunter, in this case, wakes up late and consequently stays up later than Jeffrey to watch TV.	5	✗

Table 4: Examples on WinoGrande. The correct answer to the question is **bolded**, the noisy statement is marked in red, and the correct statement is marked in blue.

of the two modules, the accuracy decreases, which indicates both the RM and MCR can successfully improve commonsense reasoning performance. **(2) The reward model plays important roles.** In most cases, removing the reward model results in the greatest performance decline. This indicates that high-quality knowledge assumes a prominent role in LLMs' commonsense reasoning.

4.4 Human Evaluation

In this section, we explore whether our method effectively solves the two issues found in previous work and whether our metric is effective through manual evaluation.

Method Evaluation We manually verify whether our method truly resolves the two issues mentioned in §1. Firstly, for the noisy knowledge issue, we conduct the case study, comparing the first and last knowledge ranked by the reward model (see Table 4). As we can see, the knowledge ranked 1st contains the key evidence that leads to the correct answer, while the knowledge ranked 5th contains the wrong statement without any evidence to support it. Therefore, our method can effectively mitigate noisy knowledge by assigning it a lower score. Secondly, for the invalid reasoning issue, we manually annotate and compute the occurrence rates of the issue under different methods (see Table 5). It demonstrates that our method can reduce the rate across different datasets, mitigating this issue.

Method	WinoGrande	HellaSwag
CoT	25.0	35.0
CoT-SC	20.0	30.0
LINKED	**15.0**	**10.0**

Table 5: The ratios of the invalid reasoning issue across different methods and datasets.

Method	WinoGrande		HellaSwag	
	ES	PS	ES	PS
DPR	0.58	0.87	0.80	0.52
CoT	0.95	0.95	0.94	0.87
LINKED	0.90	1.00	0.87	0.82

Table 6: Pearson's correlations of our metrics vs. human judgments.

Metrics Evaluation We compare the correlations of the ES and PS with the human evaluation scores separately. The intuition is that a good evaluation metric should assign a good score to a good method (i.e. effective or harmless). Thus, we manually evaluate the effectiveness and harmfulness of the injected knowledge generated by different methods (DPR, CoT, Ours), calculating Pearson's correlations under different cases (see results in Table 6). In most cases, our metrics show a high positive correlation with human evaluations ($\geq$ **0.80**), indicating the effectiveness of these two scores. Since the EPS metric is the average of them, we can further prove its validity and reliability.

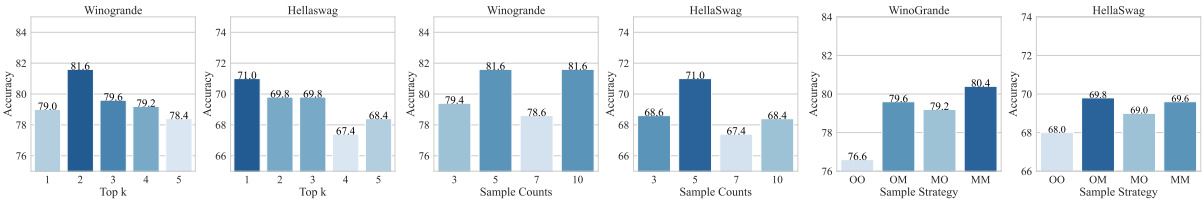
We present additional evaluation results and detailed experimental setups in Appendix D.

4.5 Experimental Factors Analysis

In our experiments, various factors can influence the performance, here we aim to draw general conclusions by observing the effects of them.

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 4a. We find that the optimal value for top-k is no more than 2. Compared to the introduction of a large volume of relevant knowledge, the filtration of knowledge is more crucial for LLMs.

Sampling Counts We change the numbers of generated knowledge to figure out whether more sampling counts make it more likely to bring effective knowledge. As illustrated in Figure 4b,



(a) Top-k knowledge (b) Sampling counts (c) Sampling strategy

Figure 4: Comparison of the impact of different experimental factors on performance.

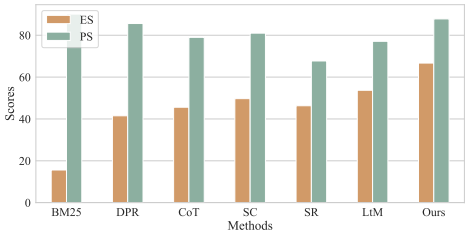


Figure 5: Comparison of ES and PS on WinoGrande.

Method	Wino	Hella	SIQA	PIQA	Avg
CoT	0.77k	0.99k	0.96k	0.64k	0.84k
SC	0.97k	1.17k	1.25k	0.85k	1.06k
SR	2.89k	3.75k	2.89k	2.41k	2.99k
LtM	3.35k	3.02k	2.52k	2.43k	2.83k
Ours	1.39k	1.85k	1.92k	1.31k	1.62k

Table 7: Token consumption comparison.

the number of effective knowledge produced by a model does not directly correlate with the sampling count. LLMs exhibit significant quality fluctuations between multiple rounds of generation.

Sampling Strategy In our MCR module, we only construct one rationale and sample multiple answers. Here, we explore the performance of integrating other sampling strategies. Concretely, we compare the accuracy under four settings: one rationale + one answer (OO), one rationale + multi-answer (OM), multi-rationale + one answer (MO), and multi-rationale + multi-answer (MM). We set the top-k value to 3 and the sampling count to 3. As we can get from 4c, OM and MM perform the best among all, but considering the higher cost of the latter, our MCR module adopts the former.

4.6 Effects Analysis of Different Methods

We evaluate the effect of different methods on the model's performance using ES and PS scores. The results are shown in Figure 5, from which we get the following findings: **(1) Retrieval augmentation methods have low harmfulness but also low effectiveness.** From the results, we can see that the BM25 and DPR methods get higher PS and lower ES among all the methods, proving that these methods struggle to retrieve effective information. **(2) Self-enhancement method can cause significant harm to the model's commonsense reasoning.** As for self-enhancement methods (i.e. CoT, SC, SR, LtM), they have a relatively higher ES but lower PS as well, highlighting the serious noisy knowledge issues in these methods. Our method performs well in both effectiveness and harmfulness (high ES and high PS).

4.7 Cost Analysis

To demonstrate the efficiency and practicality of our method, we calculate its average token cost per example and compare it with other methods (see Table 7). As we can see, compared to other self-enhancement methods (e.g. SR, LtM), our method uses significantly fewer tokens, averaging only twice the number used by the basic CoT method. This indicates that our method can achieve high performance in commonsense reasoning with fewer computational resources during downstream inference. Our method is also cost-efficient when training, which we discuss in Appendix A.3.

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 four representative commonsense reasoning 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. We leave these limitations as our future work to explore.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. Palm 2 technical report. CoRR, abs/2305.10403.

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.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 7432–7439. AAAI Press.

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

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.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single QA system. In Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of Findings of ACL, pages 1896–1907. 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.

Jiachun Li, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. 2024. Focus on your question! interpreting and mitigating toxic cot problems in commonsense reasoning. CoRR, abs/2402.18344.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. 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.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark. 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 13480–13488. AAAI Press.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21:140:1–140:67.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8732–8740. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. CoRR, abs/1904.09728.

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 Reward Models Training Details

A.1 Training Settings

For each dataset, we generate 5,000 <q,k,l> triples to train the reward model and randomly choose 500 samples from these triples as the validation set. During training, we set the learning rate as 1×10^{-5}, the batch size to 16, the epochs to 3, and the warm-up steps to 50. We choose DeBERTa as the backbone since it performs well on natural language inference tasks (He et al., 2023). When distinguishing between different qualities of knowledge, it is crucial for the model to possess this capability.

A.2 Training Performance

We use MRR@10 to evaluate whether the model can rank positive knowledge among the top positions and report the performance of our reward model on the validation set (see Table 8). The results indicate that our reward model can effectively distinguish between good knowledge and noisy knowledge.

	Wino	Hella	SIQA	PIQA
MRR@10	0.81	0.89	0.96	0.93

Table 8: The performance on the validation set.

A.3 Training Cost

Compared to other training methods, our reward model requires minimal training to achieve high performance. For the volume of training data, we use only 2,000 training examples per dataset, while other training methods in our work used at least 5,000 samples. For the time cost of training, on average, each epoch of training our reward model takes **56** seconds, significantly less than the 1,182 seconds required to train Roberta-large. Although we can not obtain the specific training time costs for the UnifiedQA and Unicorn methods, given their large training data volumes (Khashabi et al., 2020; Lourie et al., 2021), we can reasonably infer that our time cost is also significantly lower than these methods. In conclusion, the results demonstrate the cost-efficiency of our method during the training phase.

B Main Experiment Details

B.1 Datasets Selection

Here, we discuss the reasons for choosing these four datasets to evaluate our method. As we have mentioned in §1, retrieval augmentation methods struggle to recall effective information in complex commonsense reasoning scenarios. For representative benchmarks like CSQA (Talmor et al., 2019), since it focuses on relatively simple entity-based knowledge, LLMs have already shown high performance on it (>90%) (Anil et al., 2023) and can be effectively augmented using retrieval-augmented methods (Yu et al., 2022). Hence, our work does not extend to this dataset and selects harder tasks. Following former works (Anil et al., 2023; Touvron et al., 2023; OpenAI, 2023), we select these four benchmarks for evaluating the commonsense reasoning ability.

B.2 Baseline Implementation Details

We report the implementation details of baselines in the main experiment:

Few-shot We use 3-shot prompts for the few-shot, which are presented in Figure 7.

Roberta-large For each dataset, we train the `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 3×10^{-5}, and cosine warm-up steps to 500.

UnifiedQA & Unicorn Both methods train the T5 model (Raffel et al., 2020) on multiple commonsense question-answering datasets to obtain generalized commonsense reasoning capabilities.

BM25 + LLM We apply the BM25 algorithm to retrieve the top 3 most relevant knowledge triples from ATOMIC-2020 for each test question.

DPR + LLM 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.

Self-enhancement We use 3-shot prompts for CoT, CoT-SC and 5-shot prompts for Self-Refine, Least-to-Most. Figure 8, 9 and 10 show parts of

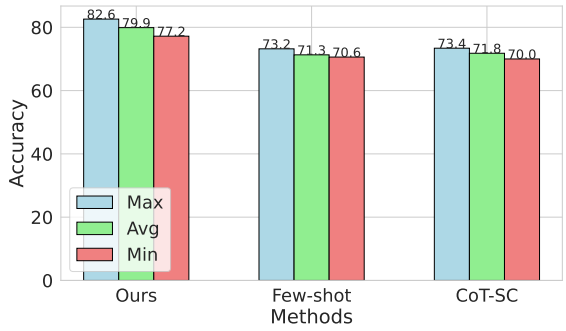


Figure 6: The robustness experiment.

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

Table 9: 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.

the prompts on WinoGrande. We also demonstrate the prompts of our method in Figure 11.

C More Results for Main Exeperiment

C.1 The Robustness of Our Method

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).

C.2 The Generalization of Our Method

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 to generate in our pilot experiment (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 9). We can get that the accuracy increases by **15.2%** after integrating filtered knowledge, which is **4.6** points higher than the injection of normal knowledge. This indicates that the filtered knowledge has generalization across different models in knowledge enhancement scenarios, highlighting the critical value of our work in downstream applications.

D Human Evaluation Details

D.1 Method Evaluation Details

Noisy Knowledge Issue We report the full experimental results of our case study on the noisy knowledge issue (see Table 10). we 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 10). 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.

Invalid Reasoning Issue We randomly select 20 answers from the results of different methods. If the knowledge in the answer is correct but the final prediction is incorrect, then the case is marked as invalid reasoning.

D.2 Metric Evaluation Details

For each piece of knowledge, we manually classify it into one of five categories: effective, relatively effective, neural, relatively harmful, and harmful. Then, we assign corresponding scores of 1, 0.5, 0, -0.5, and -1 to each category of knowledge, respectively. We randomly select 20 samples and calculate the Pearson's correlation between this score and both ES and PS.

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.

System Instruction: You are a helpful assistant that generate knowledge according to the question.
Use your commonsense knowledge to generate knowledge for some questions. Your response should be in this form:
'Knowledge: {knowledge}'
Remember you cannot directly answer the question as your knowledge.
Now the question is as follows:
Question: The test was hard for Samuel but a breeze for Randy , since _ had failed to study for it
Answer: Knowledge: 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.

…

Question: {}

System Instruction: You are a helpful assistant that choose the correct answer to the question based on the given knowledge.
Use the provided knowledge and your own commonsense knowledge to choice 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:
Knowledge: 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.
Question: The test was hard for Samuel but a breeze for Randy , since _ had failed to study for it.
(1) Samuel (2) Randy

…

Knowledge: {}
Question: {}

Figure 11: Prompts for our method.

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 10: Examples in case study. The correct answer to the question is **bolded**, some noisy knowledge statement is marked in red, and some correct knowledge statement is marked in blue.