Guided Knowledge Generation with Language Models for Commonsense Reasoning

Xiao Wei[1*], **Haoran Chen**[1*], **Hang Yu**[1†], **Hao Fei**[2], **Qian Liu**[3]
[1]Shanghai University [2]National University of Singapore [3]University of Auckland
[1]{xwei,haoranchen,yuhang}@shu.edu.cn, [2]haofei37@nus.edu.sg,
[3]Liu.Qian@auckland.ac.nz

Abstract

Large Language Models (LLMs) have achieved notable success in commonsense reasoning tasks, benefiting from their rich world knowledge acquired through extensive pretraining. While approaches like Chain-of-Thought (CoT) have shown promise in enhancing LLMs' reasoning capabilities, mitigating the influence of inaccurate commonsense knowledge remains a challenge, particularly for small-scale LLMs (e.g., those with less than 10B parameters). In this work, we propose a novel method named Guided Knowledge Generation (GuideKG) to address these issues. It presents three advantages: (i) Employing LLMs to generate knowledge explanations and to automatically assign labels based on the probability of correct answers eliminates the need for costly manual annotation in subsequent training. (ii) Training a new module called the *Know-Filter*, which evaluates knowledge, with the introduction of a novel loss function to enhance its performance. (iii) Evaluating the effectiveness of knowledge fragments at the sentence level and fusing them allows for precise control over the generation process of LLMs. We evaluate our GuideKG on small-scale LLMs and show that it outperforms all baselines on four widely-used commonsense reasoning benchmarks. Moreover, our experiments reveal that, with proper guidance, small-scale LLMs can exhibit exceptional performance in commonsense reasoning. The code is publicly available[1].

1 Introduction

Commonsense reasoning abilities are crucial for achieving human-like intelligent systems, which encompass a comprehensive understanding of everyday world knowledge and the inference capacity to facilitate problem-solving and decision-making

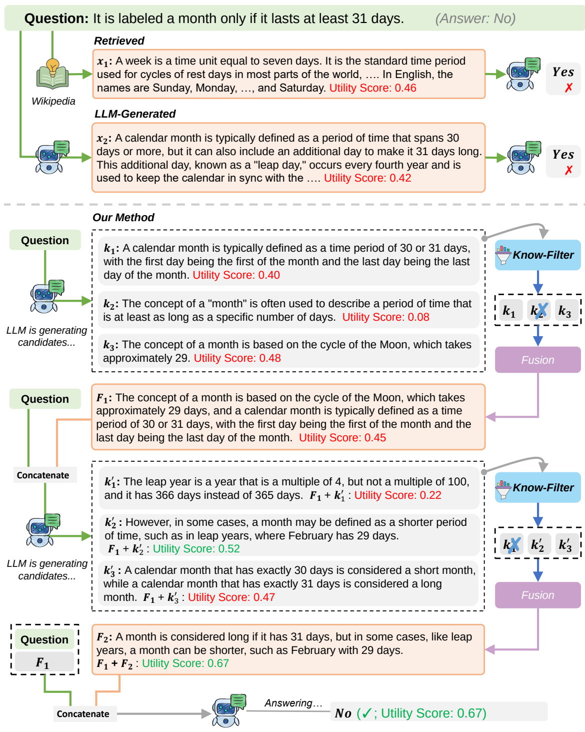

Figure 1: An illustrative example of commonsense reasoning. Using retrieved knowledge (x_1) and LLM-generated knowledge (x_2) resulted in incorrect answers. Our method guided LLM to generate high-quality knowledge (F_1+F_2), leading to correct answer.

across various daily scenarios (Sap et al., 2020; Fei et al., 2022b,a; Liu et al., 2023a; Hwang et al., 2023; Liu et al., 2023b).

Recently, the advent of LLMs (Qiao et al., 2023; Ouyang et al., 2022; Touvron et al., 2023a; Wu et al., 2024b) has significantly boosted various applications. They excel in storing and retrieving knowledge, largely due to extensive pretraining, and prompting techniques like CoT series (Kojima et al., 2022; Wang et al., 2023c; Yao et al., 2023; Besta et al., 2024; Zhang et al., 2023b) have enhanced their reasoning abilities. Despite these strides, numerous studies (Kojima et al., 2022; Wei et al., 2022) indicate that LLMs still show modest

* Equal contribution.
† Corresponding author.
[1]https://github.com/chenhaoran2018/GuideKG

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 1103–1136
November 12-16, 2024

improvements in commonsense reasoning, which relies more on prior knowledge, compared to other reasoning-intensive tasks like logical, symbolic, and arithmetic reasoning. This challenge is more pronounced in small-scale LLMs[2], constrained by limited model size and training data (Rejeleene et al., 2024). Consequently, despite the advancements brought by LLMs, achieving high-quality knowledge remains an open challenge in commonsense reasoning.

When LLMs directly respond to commonsense questions, they rely on implicit knowledge within their internal parameters to perform reasoning (Wang et al., 2023b). To provide useful context to support commonsense reasoning, one common approach is to retrieve knowledge from external knowledge bases, but this practice is limited by the scale and coverage of the knowledge bases and the performance of the retrieval system (Berchansky et al., 2023). For example, as shown in Fig. 1, knowledge retrieved from Wikipedia (x_1) related to "*week*" failed to answer the question regarding "*month*". Another emerging trend is to elicit related knowledge from LLMs (Liu et al., 2022a; Wang et al., 2023a). For example, Liu et al. (2022b) propose generating high-quality knowledge statements from large-scale LLMs (such as GPT-3 (Brown et al., 2020)) and then providing them to small-scale models. However, the reliability remains a concern; as depicted in Fig. 1, the generated statement x_2 misleads the small-scale inference model.

In this work, we discover that within the samplings for a given question, despite the predominance of incorrect knowledge generation, there exists a minority of samples that facilitate accurate model reasoning, even in small-scale LLMs. This benefit primarily arises from the stochastic sampling strategy inherent in auto-regressive models. Hence, the underlying principle of our method stems from making the correct parameter knowledge explicit, thus avoiding the shortcomings associated with relying solely on implicit parameter knowledge or external explicit knowledge.

To be specific, we propose External Guided Knowledge Generation (GuideKG), a cost-effective and robust framework to guide LLMs in the generation of valid knowledge, aiming to enhance their performance in commonsense reasoning tasks. Our approach initiates by gathering question-and-answer pairs, which serve as the training data for a new designed module, named Know-Filter. This module is used to evaluate the effectiveness of generated knowledge. We have introduced a new joint loss function called Utility-Weighted Classification Loss (UWC loss), which improves the evaluation performance of the Know-Filter by aligning it with the true utility of knowledge. To further amplify the efficacy of generated knowledge in reasoning, we propose a sentence-level generation strategy that integrates the Know-Filter into the model's auto-regressive generation process, rather than solely applying it post-generation. Concurrently, we propose a knowledge fusion mechanism to augment the robustness of the Know-Filter. In summary, our contributions are as follows:

- We propose a novel framework to guide LLMs in generating knowledge statements for solving commonsense reasoning questions, named GuideKG. Compared to existing reasoning strategies and external knowledge retrieval methods, our method is equally effective for smaller LLMs and provides knowledge that better facilitates LLM reasoning.
- We design a novel Know-Filter module that effectively reduces low-quality knowledge statements generated by LLMs. Moreover, it is trained on the automatically annotated data, eliminating the need for extra costs, and employs a unique sentence-level generation strategy and fusion mechanism to ensure effective knowledge generation.
- Experimental results verify that our GuideKG outperforms all baselines in four commonsense reasoning benchmarks, demonstrating its ability to guide LLMs of different scales in generating effective knowledge.

2 Related Work

Implicit World Knowledge in LLMs. During the pre-training process, LLMs store an extensive amount of world knowledge presented in the training data (Petroni et al., 2019). It has been indicated that implicit knowledge in LLMs is useful to enhance the performance of downstream tasks (Davison et al., 2019; Jiang et al., 2020; Marks and Tegmark, 2023; Liu et al., 2024a), such as commonsense knowledge (Liu et al., 2022b;

[2]In this work, small-scale LLMs refer to language models with parameter sizes below 10B.

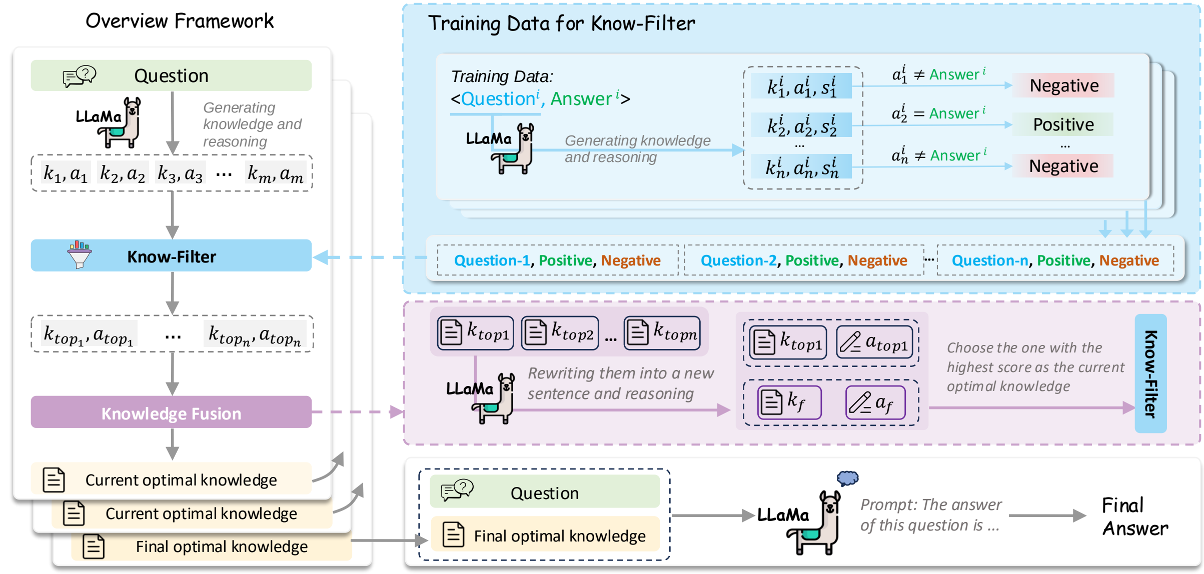

Figure 2: Overview of our proposed GuideKG. The processes of constructing knowledge filter datasets and training Know-Filter is indicated by blue arrows, while the process of Sentence-Level Fusion Generation is indicated by black arrows. The purple arrow is used to illustrate the process of the fusion mechanism.

Jain et al., 2023; Liu et al., 2024b) and relational knowledge (Chen et al., 2022; Wan et al., 2023; Xu et al., 2024). However, pre-trained models are not flawless repositories of world knowledge. Factors like incorrect training data (Lee et al., 2022a; Zhang et al., 2024), algorithms with high uncertainty in decoding (Lee et al., 2022b; Zheng and Yang, 2021), and exposure bias (Wang and Sennrich, 2020; Wang et al., 2024) can result in the production of misleading knowledge (Zhang et al., 2023a). Considering these issues, we have designed a new module called Know-Filter to filter out irrelevant or harmful information. Additionally, by guiding the model's generation process, we enable the model to deliver more effective and confident knowledge.

Commonsense Reasoning Capabilities of LLMs. To improve the reasoning abilities of LLMs, a common approach is fine-tuning with data annotated by humans (Lightman et al., 2023) or extracted from the LLMs themselves (Zelikman et al., 2022; Huang et al., 2023; Magister et al., 2023; Ho et al., 2023; Fei et al., 2023; Fei et al.; Zheng et al., 2024). Though widely used, it may compromise LLMs' inherent generalizability (Kirkpatrick et al., 2016; Lin et al., 2023). Another approach is to prompt LLMs' to improve reasoning abilities (Zhou et al., 2022; Wei et al., 2022), which is a more efficient way. For example, Wang et al. (2023c) introduce a voting strategy to select the most consistent answer, and other strategies that simplify and address complex problems through problem decomposition (Jung et al., 2022; Wang et al., 2022; Press et al., 2023; Zhou et al., 2023). Optimizing the reasoning process avoids adjusting model parameters, but past efforts have often relied on the power brought by model scale. Recent works on retrieval-augmented generation encourage LLMs to solve problems based on explicit knowledge (Asai et al., 2023; Shao et al., 2023), however, these face conflicts between external knowledge and the knowledge embedded in model parameters (Wu et al., 2024a). In this work, we in the line to leverage the intrinsic knowledge of LLMs to enhance commonsense reasoning. Our method is to guide the model to generate the most effective knowledge statements that suit the following inference model, which is quite effective on small-scale LLMs.

3 Method

Task Definition. We focus on leveraging LLMs to solve the commonsense reasoning task. In this work, we formulate commonsense reasoning as a multiple-choice question answering problem. Formally, given a question q and a set of l candidate answers $\mathcal{A} = \{a_1, a_2, \cdots, a_l\}$, the task is to select the most appropriate answer $a^* \in \mathcal{A}$ to response q based on commonsense knowledge and reasoning.

Overall Framework. In this work, we address the commonsense reasoning task by guiding LLMs

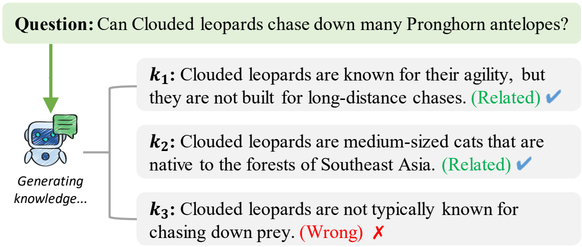

Figure 3: An example of sampled knowledge.

to generate effective knowledge to answer the question. Figure 2 shows the overview framework of the proposed method, GuideKG. First, we prompt the LLM to generate multiple knowledge statements for a given question. These statements undergo filtering by our newly designed Know-Filter module, which is trained with automatically annotated data using a novel UWC loss based on a small-scale model. The filtered, high-quality knowledge statements are then fused as a new context for the question. This context serves to guide the LLM in generating more diverse and useful statements in subsequent rounds. Through iterative generation, our approach effectively directs knowledge generation using LLM, providing accurate and useful knowledge for commonsense reasoning.

Below, we will provide a detailed description of our GuideKG, including how to automatically construct training dataset for Know-Filter, followed by the details of Know-Filter and Fusion Generation.

3.1 Knowledge Filter Dataset

Previous works (Thoppilan et al., 2022; Cobbe et al., 2021) predominantly employed human to annotate data, a practice that incurs considerable costs. Instead of relying on costly human annotations, we introduce a direct and cost-effective method for data collection based on LLMs.

Formally, we leverage an LLM to generate a set of knowledge statements for a given question q, denoting it as $\mathcal{K}_q = \{k_1, k_2, \cdots, k_m\}$, where $k_i \in \mathcal{K}_q$ represents a knowledge statement, which is a variable-length text. To achieve this goal, we instruct LLM to generate knowledge statements as follows: `Provide some knowledge related to the question`. In Figure 3, we show the examples of the knowledge statements[3].

It is observed that not all $k_i \in \mathcal{K}_q$ contribute to helping the LLM generate correct answers to q. It is necessary to distinguish between their effectiveness in supporting inferences. To achieve this goal, we use k_i as the context to obtain the LLM's response to the question q, denoted as a_i. We then assign positive or negative labels to k_i based on whether the LLM answers correctly. Finally, we obtain the probability corresponding to the correct answer a^* from the LLM's output probability distribution. This probability is denoted as s_i, which is used as the utility score[4] indicating that k_i is helpful in answering q.

To provide a clearer learning signal for Know-Filter, we choose knowledge with the highest utility as positive samples, and knowledge with the lowest utility as negative samples. Each sample consists of a question q, knowledge k_i, the LLM's response a_i and a utility score s_i.

Thus, we automatically form a labeled dataset, which indicates the usefulness of generated knowledge statements $\mathcal{K}_q$ for answering question q.

3.2 Know-Filter

Considering the instability of knowledge output from LLMs, we have developed a new model named Know-Filter. The model takes the question, candidate knowledge, and the LLM's answer as inputs, and it outputs a probability score to evaluate whether the knowledge can derive the correct answer. We use a smaller LM, MonoT5 (Nogueira et al., 2020), as our backbone model to reduce computational load and improve inference speed. During the training phase, we employ *true* and *false* tokens as labels for the loss function. During inference, the probability[5] of the *true* token is utilized as the evaluation score.

However, simply categorizing knowledge into two types is insufficient. To select a better starting point for generating subsequent knowledge, it is crucial to measure the effectiveness of knowledge at a finer granularity. Using the default cross-entropy loss can lead to overly polarized output scores. This polarization reduces sensitivity to the differences between knowledge fragments, which can adversely affect its generalizability during the evaluation phase. To address this issue, we use the probability of the correct answer from the LLM as the label to calculate the utility regularization

[3] All used instruct templates can be found in Appendix A

[4] It is noteworthy that when dealing with answers consisting of multiple tokens, we can determine s_i by calculating the average probability, instead of limiting ourselves to questions with single-token answers.

[5] Use the Softmax function during prediction to convert the probabilities of the *true* and *false* tokens into values between 0 and 1.

loss. Formally, let L_{ce} represent the cross-entropy loss associated with binary labels, and L_{cr} denote the regularization loss related to utility. The total loss L, named *Utility-Weighted Classification Loss*, is calculated as the weighted sum of these components:

$$L = \lambda \cdot L_{cr} + L_{ce}, \tag{1}$$

where λ denotes the weighting factor assigned to the utility regularization loss. For a batch size of m inputs with corresponding targets $t_1, t_2, \ldots, t_m$, the binary label loss is computed using cross-entropy as follows:

$$L_{ce} = -\frac{1}{m}\sum_{i=1}^{m} \log p(t_i), \tag{2}$$

where $p(t_i)$ represents the probability of the Know-Filter correctly predicting the label t_i. We denote the probability that the LLM produces the correct answer based on given knowledge as y_{true}. Our goal is to align the prediction probabilities of the Know-Filter as closely as possible with those of the LLM. To this end, we employ the L2 loss to compute the utility regularization loss L_{cr} as:

$$\begin{aligned} L_{cr} = -\frac{1}{m}\sum_{i=1}^{m} &||p(t_{true}) - y_{true}||^2 \\ &+ ||p(t_{false}) - (1 - y_{true})||^2, \end{aligned} \tag{3}$$

where $p(t_{true|false})$ denotes the probability of predicting *true* or *false* tokens by Know-Filter.

3.3 Sentence-Level Fusion Generation

The content generated by LLMs is typically influenced by preceding text, and sentences can serve as the minimal semantic units for evaluating knowledge effectiveness. Based on this assumption, we have deconstructed the auto-regressive process by using sentence terminators[6] in human language as signals for LLMs to pause generation. After pausing, we employ the Know-Filter to score the knowledge sentences and then select the top-N sentences for integration. Subsequently, the integrated result is used as the starting point to initiate the generation of the next knowledge sentence, continuing until the LLM outputs the default stop generation token <eos>. We refer to this process as Sentence-Level Fusion Generation (SLFG).

[6] Such as periods, exclamation marks, and question marks.

Formally, we sample m knowledge sentences based on the given q. After evaluation by Know-Filter, we obtain a ranked set of filtered knowledge statements $\mathcal{K}_q = \{k_1, \ldots, k_i, \ldots, k_m\}$, where a smaller i indicates that the Know-Filter considers k_i to be more effective.

However, Know-Filter is not a perfect scorer, hence k_1 is not always the optimal result and may lead to error propagation across multiple stages. Inspired by multi-chain reasoning (Yoran et al., 2023), we enhance the comprehensiveness of knowledge by integrating information from multiple knowledge statements to avoid missing critical details. Specifically, we select the top-n sentences from $\mathcal{K}_q$ and instruct the LLM to perform fusion:

Example:
<Instruction>
Rewriting the given knowledge into a new sentence requires retaining the part of the given knowledge that is relevant to the question.
<Candidate knowledge>
Clouded leopards are known for their agility, but they are not built for long-distance chases.
<Candidate knowledge>
Clouded leopards are medium-sized cats that are native to the forests of Southeast Asia.
<Question>
Can Clouded leopards chase down many Pronghorn antelopes?

We represent the fusion result as k_f. To avoid excessive noise that may arise from the fusion, the one with the highest Know-Filter score between k_f and k_1 is chosen as the optimal knowledge k_b. We refer to this process as one stage in SLFG. Next, k_b is appended to the end of the current stage's prompt, serving as the starting point for the next stage of generation. By repeating these steps until the LLM outputs <eos>, we complete the entire process and obtain the final knowledge statement k_c used to assist the LLM in answering the question q.

4 Experimental Setup

In this section, we will delineate the specific implementation details of our approach and explicate how it can be applied to various types of tasks.

4.1 Benchmarks

Four popular commonsense reasoning datasets are employed, categorized by the format of the questions into two types: 1) multiple-choice, i.e., Com-

LM	Methods	Extra Params	CommonsenseQA	StrategyQA	CommonsenseQA2	ARC-Challenge
Alpaca-7B	SC	-	51.5	<u>57.3</u>	49.1	50.4
	SK	-	49.6	55.7	51.1	48.1
	Verifier	880M	<u>61.8</u>	57.0	<u>52.6</u>	<u>56.7</u>
	Retrieval	108M	35.0	50.4	51.9	42.5
	Rainier	880M	52.5	48.0	47.6	42.1
	GuideKG	880M	**63.5**	**57.6**	**53.6**	**58.1**
Vicuna-7B	SC	-	62.2	<u>60.0</u>	54.9	<u>62.6</u>
	SK	-	60.6	58.0	55.3	58.5
	Verifier	880M	<u>68.3</u>	58.7	<u>58.3</u>	62.4
	Retrieval	108M	52.8	58.4	53.8	56.4
	Rainier	880M	58.6	52.5	51.8	53.9
	GuideKG	880M	**70.8**	**60.4**	**61.8**	**65.4**
Vicuna-13B	SC	-	67.8	<u>62.6</u>	63.7	73.4
	SK	-	63.9	60.3	63.5	70.3
	Verifier	880M	<u>70.7</u>	61.5	<u>64.3</u>	<u>73.6</u>
	Retrieval	108M	57.8	59.0	62.9	65.2
	Rainier	880M	65.6	52.5	55.9	64.2
	GuideKG	880M	**72.9**	**64.1**	**66.8**	**75.8**

Table 1: The accuracy of GuideKG and the baselines on benchmarks, respectively using Alpaca-7B, Vicuna-7B, and Vicuna-13B as the inference models. The term "Extra Params" refers to the size of the model parameters used by methods, excluding those of the inference model. The second best score is <u>underlined</u> and **bold** one is the best.

monsenseQA (Talmor et al., 2019) (CSQA) and ARC-Challenge (Geva et al., 2021) (ARC-c); and 2) true/false, i.e., CommonsenseQA2 (Talmor et al., 2021) (CSQA2) and StrategyQA (Geva et al., 2021) (SQA). These datasets encompass a broad range of commonsense questions, each introducing its own set of challenges. The primary metric for evaluation is accuracy, which measures the correctness of the answers. Appendix §B.2 provides additional details on the used datasets.

4.2 Models

Inference model. We select Alpaca (Taori et al., 2023) and Vicuna[7] (Zheng et al., 2023) as our experimental models. During the knowledge sampling phase, our generation configuration is as follows: temperature=1 and top_p=0.9. This high temperature setting is chosen to facilitate a broader spectrum of potential generative outcomes. In stages involving fusion and reasoning, greedy decoding is utilized as it yields the most confident responses from the model.

Know-Filter. It is derived by fine-tuning the mono-T5. This means that Know-Filter possesses a parameter size that is only 1% to 10% of the 7B inference model. We sampled two knowledge filter datasets from Alpaca-7B and Vicuna-7B for training. In the training phase, we retained the original hyperparameter settings of the mono-T5 fine-tuning script[8]. λ in Eq. (1) was set to 2. Appendix §B.2 extends more details of the training data of Know-Filter.

4.3 Baselines

We evaluated various baselines and compared them to GuideKG. Those based on internal knowledge include: 1) Self-Knowledge (SK; Liu et al. (2022b)), 2) Self-Consistency (SC; Wang et al. (2023c)), 3) Verifier (Li et al., 2023; Khalifa et al., 2023). Those based on external knowledge include: 1) Retrieval Li et al. (2023); Khalifa et al. (2023), 2) Rainier Liu et al. (2022a). Appendix §C provides additional details on the baselines used.

5 Experimental Results

In this section, we present the performance of GuideKG on commonsense reasoning and conduct a comparison with baseline models. Additionally, detailed ablation studies were carried out. Unless specifically stated otherwise, we sampled 10 knowledge sentences in each generation stage, with N set to 2 in the fusion mechanism.

5.1 Main results

As shown in Table 1, we report the average results of three runs on GuideKG and all baselines. GuideKG outperforms other baselines across all

[7]We used Vicuna 1.5 in our work, which is an adaptation finetuned from Llama2 (Touvron et al., 2023b).

[8]http://pygaggle.ai

	CSQA	SQA	CSQA2	ARC-c
GuideKG	**70.8**	**60.4**	**61.8**	**65.4**
w/o Know-Filter	63.0	58.6	55.7	61.2
w/o Guidance	69.1	59.2	59.9	62.7
w/o UWC-loss	70.0	58.8	59.3	63.0

Table 2: The ablation study of each component within GuideKG, utilizing Vicuna-7B as the inference model.

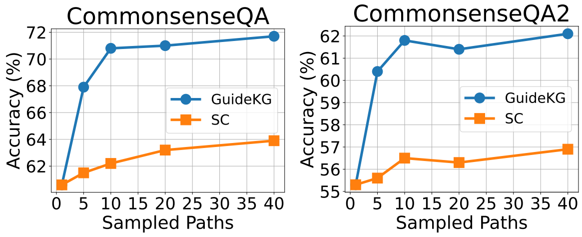

Figure 4: The impact of variations in sampling frequency on the performance of GuideKG and SC.

benchmarks. In addition to small-scale LLMs, we also conduct experiments on Vicuna-13B, the only difference being that we do not retrain Know-Filter but use the Know-Filter from Vicuna-7B. This is intended to determine whether the Know-Filter could retain its filtering capabilities on larger LLMs. The results indicate that GuideKG is equally applicable to large-scale LLMs, and even if the training data for Know-Filter do not originate from the LLM itself, the LLM can still benefit from GuideKG.

It can be observed that the enhancement in commonsense reasoning through SC is limited, aligning with findings from previous works on CoT (Chu et al., 2023). When faced with commonsense questions, LLMs prefer producing correct answers based on accurate knowledge rather than just detailed steps. LLMs' chains of thought may include factual inaccuracies or misleading information, which could hinder fact-based commonsense reasoning (Shaikh et al., 2023).

Additionally, retrieval-augmented generation also faces challenges. Research (Geva et al., 2021) has shown that relying solely on semantic matching is insufficient to retrieve the correct knowledge. This is due to the minimal overlap between commonsense reasoning problems and context, which reduces the likelihood of retrieval models exploiting shortcuts in problem language. Conversely, when LLMs generate knowledge, they can infer problem-solving strategies and provide appropriate knowledge (Valmeekam et al., 2023).

5.2 Component ablation

Then, we conducted detailed ablation experiments on GuideKG, with results shown in Table 2.

Initially, we removed Know-Filter and employed random shuffling as a replacement. Compared with the complete GuideKG, the absence of Know-Filter resulted in a significant decline in performance across all datasets. This result underscores the instability of sampling knowledge and the importance of knowledge filtering.

Next, we removed the guidance from the LLM generation process and only filtered the sampled knowledge using Know-Filter. Experimental results indicate that SLFG consistently improves inference performance compared to directly sampling complete knowledge.

Lastly, we compared the effect of training Know-Filter using only the cross-entropy loss function. The findings reveal that a utility-weighted classification loss significantly enhances reasoning performance across all datasets, particularly on the unseen SQA and ARC-c datasets.

5.3 Sampling Times

We selected Vicuna-7B as an experimental model and tested the impact of various sampling quantities on GuideKG across CSQA and CSQA2. The results, displayed in Figure 4, demonstrate that an increased number of sampling paths considerably enhances the performance of GuideKG. This supports the viewpoint that selecting the right knowledge, rather than following the majority, can prevent the oversight of correct reasoning paths due to the blind selection of the majority.

6 Analysis

We conduct theoretical experiments and case studies to investigate the performance ceiling and current performance of GuideKG.

6.1 Performance Upper Bound of GuideKG

GuideKG assumes that LLMs possess rich world knowledge and attempts to generate effective knowledge based on the model itself. As mentioned in 3.1, knowing the correct answer beforehand only requires one inference to get the probability of the correct answer under the given knowledge. Thus, we had an intriguing idea: selecting the knowledge that maximizes the probability of the correct answer at every stage of GuideKG, to simulate a perfect Know-Filter, thereby ascertaining the ideal performance of GuideKG on the dataset.

	Question: Is shrimp scampi definitely free of plastic?
Sampling	**K1**: Although shrimp scampi is often marketed as a sustainable, plastic-free option, there is actually a high-risk that it contains microplastics. Utility Score: 0.74 **K2**: Shrimp scampi is typically made with shrimp that are farm-raised, meaning they are not exposed to any plastic or pollutants. Utility Score: 0.33 **K3**: Shrimp scampi is commonly believed to be free of plastic due to its hard exoskeleton, which makes it less vulnerable to entanglement in plastic debris. Utility Score: 0.21
GuideKG	**Know-Filter outputs**: K1 score: 0.57, K3 score: 0.49, K2 score: 0.47 **Fusion K1 and K2**: Shrimp scampi is commonly believed to be free of plastic due to its hard exoskeleton, which makes it less vulnerable to entanglement in plastic debris. However, there is actually a high risk that it contains microplastics. Utility Score: 0.79 **Generate next sentence**: Shrimp scampi is commonly believed to be free of plastic due to its hard exoskeleton, which makes it less vulnerable to entanglement in plastic debris. However, there is actually a high risk that it contains microplastics. Scampi are filter feeders, meaning they ingest small particles from their environment, including plastic debris. Utility Score: 0.85 ↑ **Reasoning answer**: no
Self-Consistency	Since K2+K3>K1, the answer is yes

Table 3: An example from SQA. During the GuideKG process, Know-Filter and fusion mechanism determine the generation direction, while sentence-level generation further enhances the utility of the LLM.

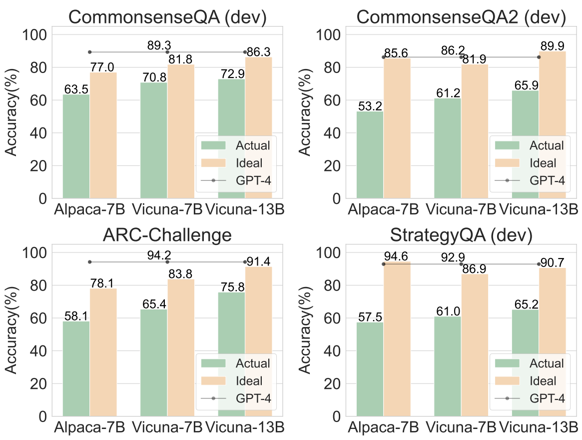

Figure 5: A comparison of the ideal versus actual performance of GuideKG. We provide GPT-4's (OpenAI, 2023) scores as a reference framework for performance.

The results are presented in Figure 5. Across four commonsense reasoning tasks, all models exhibit remarkable theoretical performance, approaching the current SOTA LLMs. This indicates that there is significant potential for improvement in small-scale LLMs, which is worth exploring. Additionally, enhancing the filtering performance of Know-Filter can notably boost LLM performance in commonsense reasoning. This also validates our previous hypothesis that aiding the model in generating better knowledge can enhance its commonsense reasoning capabilities.

6.2 Case Study

Table 3 presents a running example of GuideKG. Know-Filter assigns higher scores to the knowledge that contributes to LLM's accurate answering of questions. To safeguard against filtering missteps, the fusion mechanism amalgamates the top two high-scoring knowledge fragments. As we continue generating along the chosen direction, we observe a significant boost in the model's utility in the correct answer. This indicates that the fusion mechanism retains crucial information within the knowledge and effectively ensures the LLM's accurate generation direction. Sentence-level generation further extends this trend. More examples are listed in Appendix E.

6.3 Further Analysis

We have also explored the impact of the scale of Know-Filter and the number of sentences fused on inferential performance, and visualized the reasoning outcomes of GuideKG. In addition, we tried various other evaluation models to screen knowledge, and the results show that our Know-Filter has the best performance. More details are provided in Appendix D.

7 Conclusion

We introduce GuideKG, a cost-effective and effective framework to enhance the commonsense reasoning performance of small-scale LLMs. Initially, we automatically collect training data and apply a novel Utility-Weighted Classification Loss to train a reliable Know-Filter. Subsequently, by integrating sentence-level generation and fusion strategies, GuideKG achieves significant improvements in commonsense reasoning benchmarks and demonstrates its efficacy across various datasets and models. We also showcase the impressive potential for improvements in commonsense reasoning tasks using small-scale LLMs.

8 Limitations

Computational Resources. Although small-scale LLMs have been selected as the inference models and Know-Filter, the process of multiple sampling decoding and externally guided knowledge generation still requires substantial computational overhead. This presents a key challenge for methods requiring multiple sampling (Yao et al., 2023; Li et al., 2023). In the future, we will try to use optimized decoding methods to overcome this issue.

Rich and High-Quality Training Data. To explore the generalization performance of the Know-Filter, the knowledge filter dataset was constructed solely based on CSQA and CSQA2, involving a limited range of commonsense question types. Therefore, as indicated in Table 1, the improvement brought about by the Know-Filter for SQA and ARC-c was relatively minor. Additionally, through sample analysis, several factors were identified that influence the evaluation capability of the Know-Filter. These factors can be addressed by enhancing data quality. Consequently, developing rich and high-quality training data is an important direction for further enhancing the performance of the Know-Filter.

9 Ethics Statement

The paper has proposed an externally guided generative approach designed to enhance the performance of LLM in commonsense reasoning tasks. Our Knowledge Filter Dataset, automatically constructed based on LLM outputs, has not undergone manual refinement, potentially incorporating erroneous information that could adversely affect the predictions of the Know-Filter. Besides, our approach rely on pre-trained language models, which are trained on large-scale web data that is known to contain biased or discriminatory content.

References

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL*, pages 41–46.

Moshe Berchansky, Peter Izsak, Avi Caciularu, Ido Dagan, and Moshe Wasserblat. 2023. Optimizing retrieval-augmented reader models via token elimination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1506–1524. Association for Computational Linguistics.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI*, pages 17682–17690.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *WWW '22: The ACM Web Conference 2022*, pages 2778–2788.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *CoRR*, abs/2309.15402.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

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

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 2: Short Papers), pages 1171–1182.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391.

Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2022a. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. pages 15460–15475.

Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. 2022b. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning*, pages 6373–6391.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 14852–14882.

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

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1051–1068.

EunJeong Hwang, Veronika Thost, Vered Shwartz, and Tengfei Ma. 2023. Knowledge graph compression enhances diverse commonsense generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 558–572.

Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do language models have a common sense regarding time? revisiting temporal commonsense reasoning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 6750–6774.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 1266–1279.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. GRACE: discriminator-guided chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 15299–15328.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

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

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022a. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 8424–8445.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022b. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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, ACL*, pages 5315–5333.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *CoRR*, abs/2305.20050.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and Tong Zhang. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *CoRR*, abs/2309.06256.

Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. 2022a. Rainier: Reinforced knowledge introspector for commonsense question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 8938–8958.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022b. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 3154–3169.

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

Qian Liu, Xiubo Geng, Yu Wang, Erik Cambria, and Daxin Jiang. 2024a. Disentangled retrieval and reasoning for implicit question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7804–7815.

Qian Liu, Sooji Han, Erik Cambria, Yang Li, and Kenneth Kwok. 2024b. Primenet: A framework for commonsense knowledge representation and reasoning based on conceptual primitives. *Cognitive Computation*, pages 1–28.

Xuebo Liu, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu, and Min Zhang. 2023b. Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 15536–15550.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adámek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL*, pages 1773–1781.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. In *Proceedings of the Natural Legal Language Processing Workshop, NLLP@EMNLP 2022*, pages 130–142.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Rodrigo Frassetto Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP*, volume EMNLP 2020 of *Findings of ACL*, pages 708–718.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 2463–2473.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 5687–5711.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. 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 5368–5393.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Rick Rejeleene, Xiaowei Xu, and John Talburt. 2024. Towards trustable language models: Investigating information quality of large language models. *CoRR*, abs/2401.13086.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics: Tutorial Abstracts, ACL, pages 27–33.

Omar Shaikh, Hongxin Zhang, William Held, Michael S. Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 4454–4470.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 9248–9274.

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*, pages 4149–4158.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of AI through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual.*

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

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models - A critical investigation. 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*.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: in-context learning for relation extraction using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 3534–3547.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 2714–2730.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 3544–3552.

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.

Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. 2024. Multiple-environment self-adaptive network for aerial-view geo-localization. *Pattern Recognition*, 152:110363.

Xingbo Wang, Renfei Huang, Zhihua Jin, Tianqing Fang, and Huamin Qu. 2023b. Commonsensevis: Visualizing and understanding commonsense reasoning capabilities of natural language models. *CoRR*, abs/2307.12382.

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*.

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 *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Kevin Wu, Eric Wu, and James Zou. 2024a. How faithful are RAG models? quantifying the tug-of-war between RAG and llms' internal prior. *CoRR*, abs/2404.10198.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024b. Next-gpt: Any-to-any multimodal llm. In *Proceedings of the International Conference on Machine Learning*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *CoRR*, abs/2304.01097.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. 2024. Faithful logical reasoning via symbolic chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics , ACL*, pages 13326–13365.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 5942–5966.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *NeurIPS*.

Xu Zhang, Zhedong Zheng, Linchao Zhu, and Yi Yang. 2024. Collaborative group: Composed image retrieval via consensus learning from noisy annotations. *Knowledge-Based Systems*, page 112135.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023a. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023*.

Li Zheng, Hao Fei, Fei Li, Bobo Li, Lizi Liao, Donghong Ji, and Chong Teng. 2024. Reverse multi-choice dialogue commonsense inference with graph-of-thought. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19688–19696.

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

Zhedong Zheng and Yi Yang. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120.

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*.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron C. Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *CoRR*, abs/2211.09066.

A Prompt Templates

In our approach, we utilized the official template as a guide to prompt the experimental model, using command-style prompts. These prompts comprise a commonsense question followed by an instruction descriptor. We refrained from incorporating demonstration examples into our prompts. Crafting demonstration examples for various tasks is both time-consuming and labor-intensive. Additionally, varying examples could introduce biases into the generated outcomes. To facilitate the collection of utility levels for the knowledge generated, and to simplify comparison with correct answers, we appended a left parenthesis "(" at the end of the knowledge integration reasoning prompts for the multiple-choice datasets CSQA and ARC-c. This strategy restricts the model's response format, preventing freeform generation that complicates extracting definitive answers.

A.1 Vicuna

We list Vicuna's prompt templates in Table 4, 5, 6.

A.2 Alpaca

We list Alpaca's prompt templates in Table 7, 8, 9.

B Datasets Information

B.1 Knowledge Filter Dataset

We collected samples on Alpaca-7B and Vicuna-7B and constructed two sets of Knowledge Filter Datasets. We sampled each question 20 times, ultimately retaining a pair of positive and negative samples. When sampling outcomes contained only positive or negative labels, we discarded those questions. For 80% of the questions, we collect positive and negative samples as the training set, and for the remaining 20% of the questions, we retain all sampling results, which are used as the validation set for selecting Know-Filter checkpoints. Table 9 and table 10 present examples of the knowledge filtering dataset.

B.2 Evaluation Datasets

Table 12 showcases the datasets leveraged in our experiment. The training sets of CSQA and CSQA2 were employed to gather data for knowledge filtering. For the evaluation of inferential capabilities, we utilized the officially partitioned test sets for CSQA2, SQA, and ARC-c, with the experimental outcomes for CSQA2 and SQA being provided through official online testing. Since the test set for CSQA is not public, we conducted our evaluation on the officially designated development set.

C Baseline Specification

Here are the details for each baseline:

Self-Knowledge (SK). Liu et al. (2022b) leverage knowledge generated by LLM to assist small-scale models in reasoning. We employed the same LLM to serve both as the knowledge generation model and the inference model, treating this setup as the vanilla baseline without GuideKG.

Self-Consistency (SC). Following the method described by Wang et al. (2023c), we sampled multiple chains from the model, selecting the most frequently occurring answer through a voting mechanism. We also utilized the default setting with temperature=0.7 for sampling chains.

Verifier. Li et al. (2023); Khalifa et al. (2023) re-rank the sampling results using a trained Verifier. For fairness, we finetuned a Verifier based on monoT5-large with the same knowledge filter dataset. The training process for the Verifier did not leverage utility regularization loss.

Retrieval. Lewis et al. (2020) propose using dense vector indexing of Wikipedia as a non-parametric memory for LLMs, thereby enhancing the accuracy and diversity of the generated text. We used the RAG implementation script provided by HuggingFace[9].

Rainier. Liu et al. (2022a) enable a small model to learn to generate knowledge related to the context in order to answer given questions. This method starts by imitating the knowledge generated by GPT-3, and then learns to generate its own knowledge through reinforcement learning. We use the knowledge it generates directly for reasoning, serving as a baseline for external sources of knowledge.

D Further Analysis

D.1 Visual Analysis of Operational Performance

We visualized the operation of GuideKG in Figure 6. For questions correctly inferred, we conducted an additional greedy decoding reasoning without adding knowledge, considering it as the baseline reasoning ability of LLM. Regarding incorrectly inferred questions, we retraced all knowledge sampled during the GuideKG process. If any piece of knowledge increased the utility of the correct answer above 0.5, we attributed the error to GuideKG's processing. Among the 864 questions correctly answered, GuideKG resolved an additional 210 questions compared to greedy decoding, achieving an absolute improvement of 17.2%. In 65.3% of the instances where reasoning was incorrect, Vicuna-7B failed to provide useful knowledge

[9]https://huggingface.co

Vicuna Prompt
Knowledge Generation Prompt
A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: Provide some knowledge related to the question and no less than 50 words. Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what? ASSISTANT:
Knowledge-Integrated Reasoning Prompt
A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: Choose the correct answer to the question based on knowledge. Knowledge: [replace_here] Question: A revolving door is convenient for two-direction travel, but it also serves as a security measure at a what? Answer Choices: (A) bank (B) library (C) department store (D) mall (E) New York ASSISTANT: (

Table 4: Vicuna-7B and Vicuna-13B's knowledge generation prompt and knowledge-integrated reasoning prompt used on CSQA and ARC-c tasks.

Vicuna Prompt
Knowledge Generation Prompt
A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: Provide some knowledge related to the question and no less than 50 words. Question: Are more people today related to Genghis Khan than Julius Caesar? ASSISTANT:
Knowledge-Integrated Reasoning Prompt
A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: Answer the question based on knowledge. Answer 'Yes' or 'No'. Question: Are more people today related to Genghis Khan than Julius Caesar? Knowledge: [replace_here] ASSISTANT:

Table 5: Vicuna-7B and Vicuna-13B's knowledge generation prompt and knowledge-integrated reasoning prompt used on CSQA2 and SQA tasks.

Vicuna Prompt
Sentence Fusion Prompt
A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: Rewriting the given knowledge into a new sentence requires retaining the part of the given knowledge that is relevant to the question and correct. [Unfused sentences] Question: Are more people today related to Genghis Khan than Julius Caesar? ASSISTANT:

Table 6: Vicuna-7B and Vicuna-13B's sentence fusion prompt were used on four commonsense reasoning tasks.

Alpaca Prompt
Knowledge Generation Prompt
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: Provide some knowledge related to the question and no less than 50 words. ### Input: Question: A revolving door is convenient for two-direction travel, but it also serves as a security measure at a what? ### Response:
Knowledge-Integrated Reasoning Prompt
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: Choose the correct answer to the question based on knowledge. ### Input: Knowledge: [replace_here] Question: A revolving door is convenient for two direction travel, but it also serves as a security measure at a what? Answer Choices: (A) bank (B) library (C) department store (D) mall (E) new york ### Response: (

Table 7: Alpaca-7B's knowledge generation prompts and knowledge-integrated reasoning prompts used on CSQA and ARC-c tasks.

Alpaca Prompt
Knowledge Generation Prompt
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: Provide some knowledge related to the question, avoid answering the question directly, and no less than 50 words. ### Input: Question: Are more people today related to Genghis Khan than Julius Caesar? ### Response:
Knowledge-Integrated Reasoning Prompt
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: Answer the question based on knowledge. Answer 'Yes' or 'No'. ### Input: Question: Are more people today related to Genghis Khan than Julius Caesar? Knowledge: [replace_here] ### Response:

Table 8: Alpaca-7B's knowledge generation prompt and knowledge-integrated reasoning prompt used on CSQA2 and SQA tasks.

Alpaca Prompt
Sentence Fusion Prompt
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request. ### Instruction: Rewriting the given knowledge into a new sentence requires retaining the part of the given knowledge that is relevant to the question and correct. [Unfused sentences] ### Input: Question: A revolving door is convenient for two-direction travel, but it also serves as a security measure at a what? ### Response:

Table 9: Alpaca-7B's sentence fusion prompt was used on four commonsense reasoning tasks.

Train		
From	Num	example
CSQA	4529	{"question": "If I am suffering from boredom, and I am also hungry, what should I do?", "pos": "If you are suffering from boredom and hunger, it’s important to take action to address both issues. First, try to find a way to alleviate your boredom by engaging in a new activity or hobby. This could be anything from reading a book to taking up a new sport. Additionally, make sure to eat regular meals to satisfy your hunger and provide your body with the nutrients it needs. It’s also important to stay hydrated by drinking plenty of water.", "neg": "If you are suffering from boredom and hunger, it may be a good idea to find something to do that is both enjoyable and fulfilling.", "pos_score": 0.98, "neg_score": 0.01, "pos_pred": "(C) grocery shop", "neg_pred": "(D) see art"}
CSQA2	3500	{"question": "Can you use your phone to call 911 after you have removed your sim card and data plan with your provider?", "pos": "You can still call 911 on your phone even if you have removed your SIM card and data plan with your provider. The phone’s built-in emergency call feature is designed to work without a SIM card or data plan, and it will automatically use a temporary phone number to connect to the emergency services. However, it’s important to note that this feature may not work in all countries or regions, and the availability of emergency services may be limited in some areas. Additionally, if you have a physical SIM card inserted in your phone, you may still be able to make calls and use data services, but you will need to contact your provider to re-activate your service.", "neg": "You cannot use your phone to call 911 after removing the SIM card and data plan with your provider.", "pos_score": 0.99, "neg_score": 0.03, "pos_pred": "yes", "neg_pred": "no"}

Table 10: Information and examples in the training set of knowledge filter datasets.

Dev		
From	Num	example
CSQA	787	{"question": "What is part of a republic like the USA?", "knowledge": [["A republic is a form of government in which the people have the power to elect representatives to make decisions on their behalf.", 0.17, "(D) democracy"], ["A republic is a form of government in which the people have the power to elect representatives to make laws and decisions on their behalf.", 0.23, "(D) democracy"], ["A republic is a form of government in which the people have the power to elect representatives to govern them.", 0.23, "(D) democracy"], ["A republic is a form of government in which the people have the power to elect representatives to govern on their behalf. The United States of America is a republic because its citizens elect representatives to the federal government, which is responsible for making laws and policies that affect the entire country.", 0.58, "(A) state"], ["A republic is a form of government in which the people have the power to elect representatives to make laws and decisions on their behalf. The United States is a republic, and as such, it has a system of checks and balances in place to ensure that no one branch of government has too much power.", 0.61, "(A) state"], , ["The United States is a federal republic, which means that it is a country made up of a central government and constituent political units, such as states or provinces.", 0.98, "(A) state"]]}
CSQA2	1102	{"question": "Most people do not have any capillary filled areas on their noses.", "knowledge": [["The human nose is covered in small blood vessels called capillaries. These capillaries are responsible for delivering oxygen and nutrients to the skin cells and helping to maintain the health and appearance of the skin.", 0.13, "yes"], ["The human nose is a complex structure made up of bone, cartilage, and soft tissue.", 0.21, "yes"], ["Capillary filled areas on the nose are small, thin lines or dots that are commonly seen on the bridge or tip of the nose.", 0.31, "yes"], ["Capillary refill is a test used to evaluate skin hydration. It involves pinching the skin on the back of the hand and observing how quickly the capillaries refill. Normal capillary refill time is less than two seconds. Dermatitis, eczema, and other skin conditions can affect capillary refill time.", 0.43, "yes"], ["Capillary refill is a test used to evaluate skin hydration. It involves pinching the skin on the back of the hand and observing how quickly the capillaries refill.", 0.50, "no"], ["Capillary refill is a test used to check for dehydration. When a person’s nose is pinched, the normal capillary filled area appears within a few seconds.", 0.59, "no"], ["Capillary refill is a test used to assess the health of the skin and the functioning of the capillaries. It involves pressing a piece of cotton wool or a lint-free cloth onto the skin and then observing the time it takes for the capillaries to refill. Most people have capillary-filled areas on their noses, particularly on the tip and the alar regions.", 0.76, "no"]]}

Table 11: Information and examples in the development set of knowledge filter dataset.

Dataset	# Train	# Dev	# Test
CommonsenseQA	9741	1221	-
StrategyQA	-	-	490
CommonsenseQA2	9264	2541	2473
ARC-Challenge	-	-	1172

Table 12: The number of samples per segment within each evaluation benchmark utilized in our experiments.

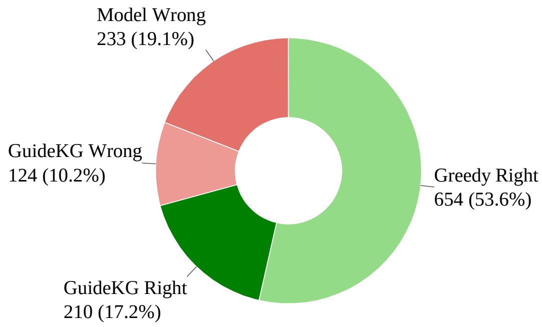

Figure 6: Outcome Distribution of Vicuna-7B's Reasoning on CSQA via GuideKG.

in 10 sampling attempts. For the remaining inaccurately inferred questions, we believe the failure was due to GuideKG's inability to filter knowledge accurately. Below, we detailedly analyze the reasons for these outcomes through specific examples.

Appendix E showcases GuideKG's operation cases across various benchmarks.

In the correct instances, the Know-Filter accurately filters the optimal knowledge and gradually increases the model's utility in the correct answer through iterative generation. Table 14 presents an example where the integration mechanism functions effectively, indicating its corrective role when the Know-Filter selects inaccurate knowledge. In incorrect instances, we observed that as the length of knowledge increases, the average score given by Know-Filter gradually rises while the probability of the correct answer decreases. The bias introduced by text length causes longer knowledge at the same stage to dominate. Moreover, analyzing the knowledge in these instances revealed that content incorrect but highly similar to the question sometimes scores higher than effective knowledge, suggesting that Know-Filter's scoring may also be influenced by text similarity.

Besides the evaluation errors of Know-Filter, another part of the reason for incorrect reasoning stems from the model's sampling results not including knowledge helpful for correctly answering the question. This may result from the model not having a deep enough memory of certain knowledge, which could be mitigated by increasing the sampling frequency. Another possibility is that the model's pre-training data is not comprehensive or contains incorrect knowledge, leading to a lack of necessary knowledge reserves for answering questions, marking a primary reason for theoretical performance limitations.

D.2 Know-Filter Size

We trained three Know-Filter models of different parameter sizes (Small, Base, and Large) to investigate the impact of scaling the Know-Filter model size on our approach's effectiveness. The results presented in Figure 7 demonstrate that enlarging the parameter size of the Know-Filter contributes to enhanced performance across the majority of datasets.

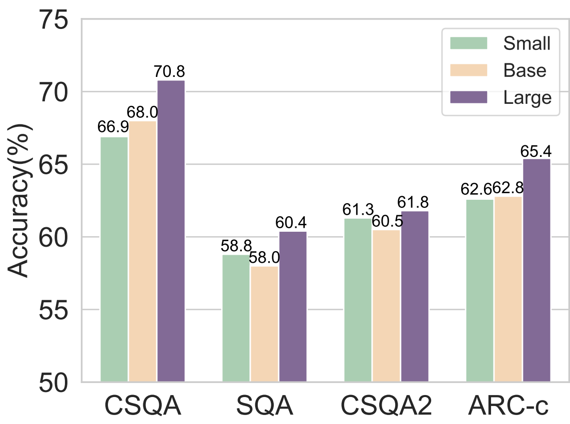

Figure 7: The performance of Know-Filters of various scales on four commonsense reasoning benchmarks, using Vicuna-7B as the inference model.

D.3 The Number of Fused Sentences

The fusion mechanism implemented in this study enhances single sentences with a higher density of knowledge, thereby mitigating the risk of overlooking critical information due to erroneous assessments, which in turn boosts the overall performance. However, it is important to recognize the potential introduction of incorrect, misleading noise information during the fusion process. To quantitatively assess the impacts of varying sentence quantities on this mechanism, we conducted experiments with different sentence count settings (1, 2, 3, 4, 5). As depicted in Figure 8, the optimum outcome was achieved by fusing the first two sen-

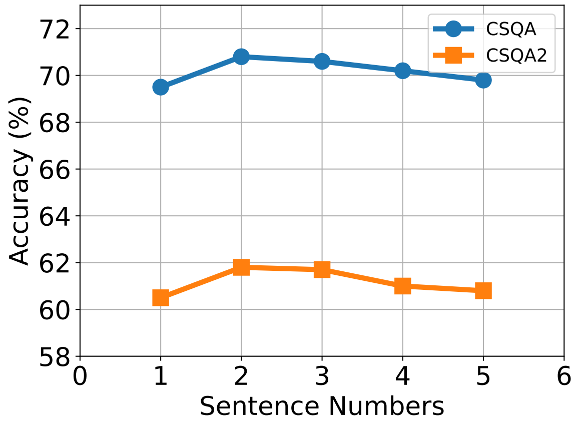

Figure 8: The impact of fusing sentence quantity on reasoning performance, utilizing Vicuna-7B as the inference model.

tences when the sampled knowledge quantity was set to 10; beyond this sentence count, a decline in the model's inferential performance was observed.

D.4 Different Evaluation Models

We selected three different types of evaluation models as baselines, namely the bidirectional encoder (Bi-Encoder), VERA (Liu et al., 2023a), and an LLM fine-tuned with LoRA (Hu et al., 2022).

First, we employ a BERT-based bi-encoder to compute the semantic relevance between questions and knowledge. Here, we selected the currently best-performing pre-trained sentence embedding model from Sentence-Transformers (Reimers and Gurevych, 2020) (i.e., all-mpnet-base-v2), and fine-tuned it on the Know-Filter dataset. Then, we generated sentence embeddings for questions and knowledge separately, and calculated their dot product similarity, which was passed through a sigmoid function to serve as the relevance score. Experimental results showed that this sentence embedding model exhibited reasonable filtering performance, but still fell short of our method (Know-Filter), resulting in an average performance drop of 3.55%.

Next, we employ VERA as our evaluation model. VERA is a general-purpose commonsense statement validation model, designed to estimate the plausibility of declarative natural language statements based on commonsense knowledge. Due to not having been fine-tuned, VERA's evaluation performance on CSQA is even lower than that of the bi-encoder.

Finally, we fine-tuned Llama-7B by adding LORA modules, enabling it to perform the binary classification task of true/false judgment. Experimental results demonstrated that the LORA fine-tuning method achieved performance close to our Know-Filter, with an average performance drop of 2.0%.

	Params	CSQA	CSQA2
Bi-Encoder	110M	66.7(-4.1)	59.7(-2.1)
VERA	4700M	62.5(-8.3)	59.9(-1.9)
LLaMA-LoRA	6700M	69.2(-1.6)	60.5(-1.3)
GuideKG	880M	**70.8**	**61.8**

Table 13: Comparison results between additional evaluation model and our Know-Filter, using Vicuna-7B as the inference model.

Considering the parameter scale factor, our original Know-Filter still maintains an advantage. Since the base model (mono-T5) of our original Know-Filter had previously undergone training for similar tasks, this result also indicates that rich, high-quality training data can enhance the Know-Filter's performance. In our future work, further improving the Know-Filter's capabilities while maintaining a low parameter count will be an important aspect.

D.5 Generalization Analysis

We have provided a cost-effective method for generating prompts that can efficiently evaluate the effectiveness of knowledge and enhance its utility. This approach is universally applicable across multiple reasoning tasks. When expanding to other languages (such as Chinese, Japanese, Arabic, etc.), we can employ models that are either specifically designed for those languages or are multilingual in capability as the reasoning models and Know-Filters. This adaptation does not affect the operational mechanism of our method but enables its effective execution across different languages. When extending to specific domains (such as law, finance, healthcare, etc.), the vast general knowledge incorporated within modern LLMs, which often includes a substantial amount of domain-specific knowledge, usually does not significantly affect performance. However, replacing the reasoning model and Know-Filter with models specifically trained for certain domains (like legalBERT (Mamakas et al., 2022) , DoctorGLM (Xiong et al., 2023) , etc.) could greatly benefit our method. It would facilitate the generation of valid domain-specific knowledge, thereby enhancing performance in specific areas.

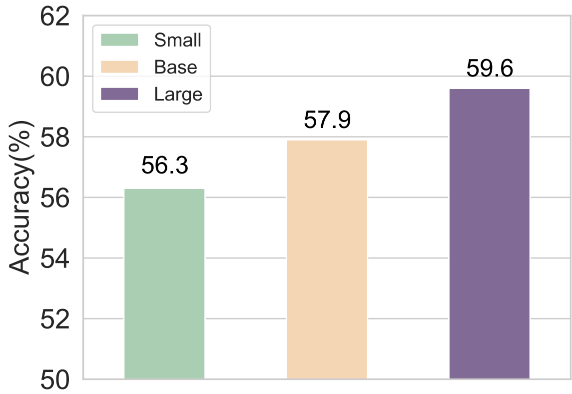

Figure 9: Internal evaluation metrics of knowledge filters with different sizes.

D.6 Internal Evaluation of the Know-Filter

The Know-Filter serves GuideKG, and its primary concern is whether it can select the best knowledge. Therefore, common document ranking metrics such as MRR, MAP, and NDCG are not suitable for the internal evaluation of the Know-Filter.Below, we provide a detailed explanation of the computational process for the internal evaluation criteria of Know-Filter.

Our evaluation method measures the predictions of the Know-Filter from two aspects to determine if they pass. First, if the predicted optimal knowledge matches the actual optimal knowledge, the data point passes the test. This is similar to Precision@1. Second, if the utility label of the predicted optimal knowledge is greater than 0.5, the test passes. This is because for both iterative generation and final inference results, knowledge with a utility label greater than 0.5 is acceptable. We also believe this approach may help select a more generalizable Know-Filter. It is important to note that these two aspects have an "or" relationship; satisfying either one means the test passes. In Figure 9, we present the best internal evaluation results for training Know-Filter of different sizes on the knowledge filtering dataset collected from Vicuna.

E Running Cases

We demonstrate representative case studies of GuideKG across each benchmark, derived from the operations of Vicuna-7B. We categorize these into concurrent and non-concurrent generation phases to comprehensively depict GuideKG's capabilities. Each category includes typical positive and negative cases for an in-depth sample analysis. We present a "Filter score" (assigned by Know-Filter) and a "Label score" (the probability of LLM producing the correct answer under knowledge prompt) for samples from CSQA and ARC, along with the response ("Answer") provided by LLM under knowledge prompts. Due to the unavailability of test sets for CSQA2 and SQA, we cannot offer Label scores. To maintain brevity, the knowledge statements in the concurrent generation phase examples are exclusively those produced in the first stage of SLGF.

E.1 CSQA

We list examples on CSQA by GuideKG in Table 14, 15, 16, 17.

E.2 ARC-c

We list examples on ARC-c by GuideKG in Table 18, 19, 20, 21.

E.3 SQA

We list examples on SQA by GuideKG in Table 22, 23.

E.4 CSQA2

We list examples on CSQA2 by GuideKG in Table 24, 25.

QUESTION: A revolving door is convenient for two-direction travel, but it also serves as a security measure at a what?			
Sampling knowledge	Filter score	Label score	Answer
A revolving door is a convenient feature for allowing two-direction travel in and out of a building, such as a shopping mall or airport.	0.56	0.012	mall
A revolving door is a convenient feature for allowing two-directional travel in and out of a building, but it also serves as a security measure at entrances to some establishments, such as banks, government buildings, and corporate offices.	0.41	0.95	bank
A revolving door is a convenient design for two-direction travel, but it also serves as a security measure at entrances to buildings, such as government buildings, banks, and office buildings.	0.36	0.69	bank
A revolving door is a convenient feature for allowing people to enter and exit a building in both directions.	0.35	0.18	mall
A revolving door allows people to travel in both directions through a narrow opening, such as a doorway.	0.17	0.22	mall
FUSION: A revolving door is convenient for two-directional travel and also serves as a security measure at entrances to some establishments, such as banks, government buildings, and corporate offices.	0.58	0.96	bank
QUESTION: If you are awaking multiple times throughout the night because a lot is on your mind, what is a likely cause?			
Sampling knowledge	Filter score	Label score	Answer
Multiple awakenings throughout the night can be a sign of various causes, including anxiety, stress, depression, medication side effects, or underlying health conditions like sleep apnea or restless leg syndrome.	0.35	0.95	depression
There are several causes for waking up multiple times throughout the night due to a buzzing in the ear.	0.27	0.02	getting out of bed
It is common to wake up multiple times throughout the night if a lot is on your mind.	0.26	0.001	getting out of bed
There could be several reasons why you are awakening multiple times throughout the night.	0.22	0.001	getting out of bed
It is common to wake up multiple times throughout the night if there is a lot on your mind.	0.12	0.001	getting out of bed
FUSION: If you are awaking multiple times throughout the night due to a buzzing in the ear, it is likely caused by anxiety, stress, depression, medication side effects, or underlying health conditions like sleep apnea or restless leg syndrome.	0.65	0.95	depression

Table 14: Correct examples in the same generation stage on the CSQA dataset by GuideKG. Each example contains a question, five sampled knowledges, and one fusion result. Shown from high to low according to the Filter score, meaning the second row's knowledge is considered the optimal knowledge by the Know-Filter.

QUESTION: Where would you find a monkey in the wild?			
Sampling knowledge	Filter score	Label score	Answer
Monkeys are found in the wild in a variety of habitats, including forests, woodlands, and grasslands.	0.62	0.28	zoo
Monkeys can be found in the wild in a variety of habitats, including forests, jungles, and grasslands.	0.62	0.37	zoo
Monkeys are found in the wild in various parts of the world, including Africa, Asia, and South America.	0.58	0.20	zoo
Monkeys are primates that are found in various habitats across the world, including tropical forests, grasslands, and wetlands.	0.58	0.20	zoo
Monkeys are found in the wild in tropical and subtropical regions around the world, including Africa, Asia, and South America.	0.56	0.19	zoo
FUSION: You can find a monkey in the wild in a variety of habitats, including forests, woodlands, and grasslands.	0.65	0.44	zoo
QUESTION: What type of keyboard is made up of one or more pipe divisions?			
Sampling knowledge	Filter score	Label score	Answer
A keyboard made up of one or more pipe divisions is called a "modular keyboard." It consists of individual modules or sections that can be connected to form a complete keyboard layout.	0.86	0.17	modular keyboard
A keyboard made up of one or more pipe divisions is called a "pipe organ" or simply "organ." It is a musical instrument that produces sound through the use of pipes and air pressure.	0.85	0.97	organ
A keyboard that is made up of one or more pipe divisions is called a mechanical keyboard.	0.82	0.05	mechanical keyboard
A pipe organ is a type of keyboard instrument that is made up of one or more pipe divisions.	0.76	0.98	organ
A type of keyboard known as the "Pipe Organ" is made up of one or more pipe divisions.	0.73	0.97	organ
FUSION: A modular keyboard is made up of one or more pipe divisions, which are individual modules or sections that can be connected to form a complete keyboard layout.	0.01	0.18	modular keyboard

Table 15: Incorrect examples in the same generation stage on the CSQA dataset by GuideKG. Each example contains a question, five sampled knowledges, and one fusion result. Shown from high to low according to the Filter score, meaning the second row's knowledge is considered the optimal knowledge by the Know-Filter.

QUESTION: Where would you find a monkey in the wild?				
Stage	Final knowledge	Filter score	Label score	Answer
1	You can find a monkey in the wild in a variety of habitats, including forests, woodlands, and grasslands.	0.65	0.44	zoo
2	You can find a monkey in the wild in a variety of habitats, including forests, woodlands, and grasslands. You can find monkeys in the wild in different parts of the world, such as Africa, Asia, and South America, where they are native to tropical and subtropical regions. They are often found in groups, led by a dominant male.	0.69	0.56	thailand
3	You can find a monkey in the wild in a variety of habitats, including forests, woodlands, and grasslands. You can find monkeys in the wild in different parts of the world, such as Africa, Asia, and South America, where they are native to tropical and subtropical regions. They are often found in groups, led by a dominant male. You would find a monkey in the wild in trees, where they can be seen swinging from branch to branch and eating a variety of foods, including fruit, leaves, and insects.	0.72	0.62	thailand
QUESTION: Obstructing justice is sometimes an excuse used for police brutality which causes what in people?				
Stage	Sampling knowledge	Filter score	Label score	Answer
1	Obstructing justice is a criminal offense that occurs when someone interferes with the administration of justice or impedes an investigation. This can include actions such as interfering with an investigation or trial, or impeding the ability of law enforcement to perform their duties. Obstructing justice is sometimes used as an excuse for police brutality, which can have a variety of negative effects on people, including physical harm, emotional trauma, and damage to their rights and freedoms.	0.64	0.34	attack
2	Obstructing justice is a criminal offense that occurs when someone interferes with the administration of justice or impedes an investigation. This can include actions such as interfering with an investigation or trial, or impeding the ability of law enforcement to perform their duties. Obstructing justice is sometimes used as an excuse for police brutality, which can have a variety of negative effects on people, including physical harm, emotional trauma, and damage to their rights and freedoms. Police brutality is the use of excessive force by law enforcement, and it is a serious violation of people's rights.	0.32	0.62	getting hurt

Table 16: Correct examples in the different generation stages on the CSQA dataset by GuideKG.

QUESTION: If you take the risk of buying a used car, you still hope it can what?				
Stage	Final knowledge	Filter score	Label score	Answer
1	If you take the risk of buying a used car, you still hope that it will run smoothly and reliably for you.	0.83	0.36	go fast
2	If you take the risk of buying a used car, you still hope that it will run smoothly and reliably for you. You may also hope that the car will be in good condition and not have any major mechanical issues that will require expensive repairs.	0.84	0.21	go fast
QUESTION: Who might wear dark glasses indoors?				
Stage	Final knowledge	Filter score	Label score	Answer
1	Individuals who work or spend time in environments with high levels of artificial light, such as photographers, video editors, or people who use computer screens for extended periods, may wear dark glasses indoors to reduce glare and protect their eyes from the strain of bright light.	0.56	0.50	blind person
2	Individuals who work or spend time in environments with high levels of artificial light, such as photographers, video editors, or people who use computer screens for extended periods, may wear dark glasses indoors to reduce glare and protect their eyes from the strain of bright light. People who have eye conditions like cataracts, glaucoma, or retinal diseases, or who have certain medical conditions such as photophobia, may wear dark glasses indoors to filter out excessive light and reduce eye strain.	0.68	0.18	ray charles
2	Obstructing justice is a criminal offense that occurs when someone interferes with the administration of justice or impedes an investigation. This can include actions such as interfering with an investigation or trial, or impeding the ability of law enforcement to perform their duties. Obstructing justice is sometimes used as an excuse for police brutality, which can have a variety of negative effects on people, including physical harm, emotional trauma, and damage to their rights and freedoms. Police brutality is the use of excessive force by law enforcement, and it is a serious violation of people's rights.	0.32	0.62	getting hurt

Table 17: Incorrect examples in the different generation stages on the CSQA dataset by GuideKG.

QUESTION: The end result in the process of photosynthesis is the production of sugar and oxygen. Which step signals the beginning of photosynthesis?			
Sampling knowledge	Filter score	Label score	Answer
The beginning of photosynthesis is signaled by the conversion of light energy into chemical energy through the process of photolysis.	0.69	0.001	light energy is converted to chemical energy
The beginning of photosynthesis is signaled by the absorption of light energy by chlorophyll, a pigment found in plant cells.	0.68	0.99	chlorophyll in the leaf captures light energy
The beginning of photosynthesis is signaled by the absorption of light energy by chlorophyll pigments in the chloroplasts of plant cells.	0.68	0.99	chlorophyll in the leaf captures light energy
The beginning of photosynthesis is signaled by the absorption of light energy by pigment molecules, specifically chlorophyll, in the thylakoid membranes of the chloroplasts in plant cells.	0.67	0.99	chlorophyll in the leaf captures light energy
The process of photosynthesis is initiated when a plant's chloroplasts absorb light energy from the sun.	0.65	0.99	chlorophyll in the leaf captures light energy
FUSION: The beginning of photosynthesis is signaled by the absorption of light energy by pigment molecules, specifically chlorophyll, in the thylakoid membranes of the chloroplasts in plant cells.	0.63	0.99	chlorophyll in the leaf captures light energy
QUESTION: Petrified palm trees are found in sedimentary rock near glaciers. The presence of the petrified palm trees most likely provides evidence for which statement?			
Sampling knowledge	Filter score	Label score	Answer
Petrified palm trees are found in sedimentary rock near glaciers, providing evidence for the past presence of tropical or subtropical climates in the area.	0.82	0.99	the climate in the area was once tropical
Petrified palm trees are often found in sedimentary rock near glaciers, which suggests that they were once part of a tropical forest during a time when the Earth's climate was warmer and wetter.	0.79	0.99	the climate in the area was once tropical
Petrified palm trees are found in sedimentary rock near glaciers, indicating that they were once submerged in water.	0.78	0.001	there was once more water in the area
Petrified palm trees are often found in sedimentary rock near glaciers, and their presence is evidence of a warmer, wetter climate in the past.	0.76	0.01	there was once more water in the area
Petrified palm trees are fossilized remains of palm trees that were once alive.	0.67	0.27	there was once more water in the area
FUSION: The presence of petrified palm trees in sedimentary rock near glaciers indicates that they were once part of a tropical rainforest that existed in the area millions of years ago, providing evidence for the past presence of tropical or subtropical climates.	0.81	0.99	the climate in the area was once tropical

Table 18: Correct examples in the same generation stage on the ARC-c dataset by GuideKG.

QUESTION: Which of these gases is the most abundant greenhouse gas in the lower atmosphere of Earth?			
Sampling knowledge	Filter score	Label score	Answer
The most abundant greenhouse gas in the lower atmosphere of Earth is carbon dioxide (CO_2). It is a colorless and odorless gas that is naturally present in the atmosphere.	0.79	0.001	carbon dioxide
The most abundant greenhouse gas in the lower atmosphere of Earth is water vapor, which accounts for about 95% of the total greenhouse effect.	0.76	0.99	water vapor
Water vapor is the most abundant greenhouse gas in the lower atmosphere of Earth, followed by carbon dioxide and methane.	0.72	0.99	water vapor
The most abundant greenhouse gas in the lower atmosphere of Earth is water vapor.	0.71	0.99	water vapor
The most abundant greenhouse gas in the lower atmosphere of Earth is water vapor.	0.71	0.99	water vapor
FUSION: Carbon dioxide (CO_2) is the most abundant greenhouse gas in the lower atmosphere of Earth, followed by water vapor and methane.	0.66	0.001	carbon dioxide
QUESTION: On August 21, a flash flood warning was issued for the Las Vegas area. Which statement best describes this warning in terms of weather and climate?			
Sampling knowledge	Filter score	Label score	Answer
The flash flood warning issued for the Las Vegas area on August 21 indicated a severe weather event that posed a significant risk to the area.	0.35	0.04	It is a rare event inconsistent with local climate and weather
The flash flood warning issued for the Las Vegas area on August 21 indicates a severe weather event, typically associated with heavy rainfall, that can result in sudden and rapid flooding of low-lying areas.	0.30	0.17	It is a rare event inconsistent with local climate and weather
The flash flood warning issued for the Las Vegas area on August 21 indicates a severe weather event that is likely to occur due to heavy rainfall in the region.	0.29	0.02	It is a rare event inconsistent with local climate and weather
A flash flood warning is a severe weather warning issued by the National Weather Service when rainfall is expected to cause flooding in a specific area.	0.23	0.25	It is a rare event inconsistent with local climate and weather
The flash flood warning for the Las Vegas area on August 21 indicated that a severe thunderstorm with heavy rainfall was expected to impact the region.	0.22	0.31	It is a rare event inconsistent with local climate and weather
FUSION: The flash flood warning issued for the Las Vegas area on August 21 indicates a severe weather event, typically associated with heavy rainfall, that can result in sudden and rapid flooding of low-lying areas.	0.29	0.17	It is a rare event inconsistent with local climate and weather

Table 19: Incorrect examples in the same generation stage on the ARC-c dataset by GuideKG.

QUESTION: A scientist maps a long region in which earthquakes originate and determines this region is a transform plate boundary. Which evidence would cause the scientist to reevaluate this determination?				
Stage	Final knowledge	Filter score	Label score	Answer
1	A transform plate boundary is a type of plate boundary where two tectonic plates move past each other along a fault, resulting in earthquakes.	0.72	0.10	The region shows extensive faulting of sediments.
2	A transform plate boundary is a type of plate boundary where two tectonic plates move past each other along a fault, resulting in earthquakes. The evidence that would cause a scientist to reevaluate the determination of a transform plate boundary in a long region could include the presence of volcanic activity, the absence of a seismic gap, or the existence of normal faults instead of transform faults.	0.67	0.39	The region shows extensive faulting of sediments.
3	A transform plate boundary is a type of plate boundary where two tectonic plates move past each other along a fault, resulting in earthquakes. The evidence that would cause a scientist to reevaluate the determination of a transform plate boundary in a long region could include the presence of volcanic activity, the absence of a seismic gap, or the existence of normal faults instead of transform faults. These findings would suggest that the region may be a convergent plate boundary, where two plates are colliding, or a divergent plate boundary, where two plates are moving apart.	0.72	0.51	Volcanism also characterizes the region.
QUESTION: What is the primary cause of rainstorms?				
Stage	Sampling knowledge	Filter score	Label score	Answer
1	The primary cause of rainstorms is the upward movement of air in the atmosphere, which creates convection.	0.58	0.01	Mountains direct moist air upward.
2	The primary cause of rainstorms is the upward movement of air in the atmosphere, which creates convection. This is caused by the heating of the surface of the Earth by the sun, which causes the air near the surface to become warmer and more buoyant.	0.71	0.99	Earth is heated by the Sun.
3	The primary cause of rainstorms is the upward movement of air in the atmosphere, which creates convection. This is caused by the heating of the surface of the Earth by the sun, which causes the air near the surface to become warmer and more buoyant. As this warmer, more buoyant air rises, it cools and the moisture in the air condenses, forming clouds and ultimately leading to rainfall.	0.72	0.96	Earth is heated by the Sun.

Table 20: Correct examples in the different generation stages on the ARC-c dataset by GuideKG.

QUESTION: Garden plants on Earth require four resources to stay alive: soil, air, water, and sunlight. How many of these resources are necessary for life to exist on the moon or another planet?				
Stage	Final knowledge	Filter score	Label score	Answer
1	Only three of the four resources are necessary for life to exist on the moon or another planet: soil, air, and water.	0.58	0.03	3
2	Only three of the four resources are necessary for life to exist on the moon or another planet: soil, air, and water. To survive on the moon or another planet, garden plants require soil, air, water, and sunlight, just like on Earth.	0.59	0.34	3
3	Only three of the four resources are necessary for life to exist on the moon or another planet: soil, air, and water. To survive on the moon or another planet, garden plants require soil, air, water, and sunlight, just like on Earth. However, the absence of a thick atmosphere and the extreme temperature fluctuations on the moon make it difficult for plants to grow.	0.61	0.12	3

QUESTION: What can most likely be seen without the aid of a telescope on a clear night?				
Stage	Sampling knowledge	Filter score	Label score	Answer
1	On a clear night, the most likely thing that can be seen without the aid of a telescope is the naked eye.	0.68	0.97	the Milky Way galaxy
2	On a clear night, the most likely thing that can be seen without the aid of a telescope is the naked eye. The naked eye can see a wide range of objects, including stars, planets, and meteors.	0.61	0.91	the Milky Way galaxy
3	On a clear night, the most likely thing that can be seen without the aid of a telescope is the naked eye. The naked eye can see a wide range of objects, including stars, planets, and meteors. The brightness of the object determines how easily it can be seen.	0.57	0.80	the Milky Way galaxy
4	On a clear night, the most likely thing that can be seen without the aid of a telescope is the naked eye. The naked eye can see a wide range of objects, including stars, planets, and meteors. The brightness of the object determines how easily it can be seen. For example, the brightest stars can be seen from far away, while the dimmer objects, like meteors, can only be seen for a short time.	0.68	0.49	asteroids

Table 21: Incorrect examples in the different generation stages on the ARC-c dataset by GuideKG.

QUESTION: Is a mustache related to a dictator?		
Sampling knowledge	Filter score	Answer
A mustache is not necessarily related to a dictator.	0.47	no
A mustache is not inherently related to a dictator, as many individuals, including heads of state, have worn mustaches for various reasons.	0.47	no
The association between a mustache and a dictator is a stereotype that has been perpetuated in popular culture.	0.39	yes
A mustache is a facial hairstyle that is often associated with power and authority, which may be why it is sometimes associated with dictators.	0.36	yes
Mustaches have been associated with dictators and authoritarian figures throughout history.	0.35	yes
FUSION: A mustache is not inherently related to a dictator, but it can be associated with certain historical figures who were dictators.	0.37	yes
QUESTION: Would a monocle be appropriate for a cyclops?		
Sampling knowledge	Filter score	Answer
A monocle would not be appropriate for a cyclops, as a cyclops is a mythical creature with only one eye in the center of its forehead, while a monocle is a type of eyepiece that is typically worn by humans to correct vision problems or improve visual acuity.	0.55	no
A monocle would not be appropriate for a cyclops, as the term "cyclops" refers to a mythical creature with a single eye in the center of its forehead.	0.53	no
A monocle would not be appropriate for a cyclops, as a cyclops is a mythological creature with a single eye in the center of its forehead, while a monocle is a single lens spectacle that is typically worn on the bridge of the nose.	0.53	no
A monocle is an optical device that is worn on the eye to correct vision.	0.30	yes
A monocle is a single-lens spectacle that is placed on the eye to correct vision problems or to add a touch of style.	0.27	yes
FUSION: No, a monocle would not be appropriate for a cyclops because it is a mythical creature with only one eye in the center of its forehead, while a monocle is a type of eyepiece that is typically worn by humans to correct vision problems or improve visual acuity.	0.56	no

Table 22: Examples in the same generation stage on the SQA dataset by GuideKG.

QUESTION: Did Metroid have a link with the studio of the DOA series?			
Stage	Final knowledge	Filter score	Answer
1	Metroid is a popular video game series developed by Nintendo, while the DOA (Dead or Alive) series is developed by Team Ninja, a different studio altogether.	0.55	no
2	Metroid is a popular action-adventure video game series developed by Nintendo, and while it has no direct link with the studio of the DOA series, it is developed by a different studio altogether. The DOA series, on the other hand, is developed by EA Tiburon and has a completely different gameplay style.	0.58	no
3	Metroid is a popular action-adventure video game series developed by Nintendo, and while it has no direct link with the studio of the DOA series, it is developed by a different studio altogether. The DOA series, on the other hand, is developed by EA Tiburon and has a completely different gameplay style. Metroid is known for its unique blend of exploration and shooting mechanics, while the DOA series focuses more on fast-paced combat and sports elements.	0.60	no
QUESTION: Did Ferdinand Porsche hold a triple citizenship?			
Stage	Sampling knowledge	Filter score	Answer
1	Ferdinand Porsche was an Austrian-born automotive engineer who became a naturalized citizen of Germany in 1938.	0.52	no
2	Ferdinand Porsche was an Austrian-born automotive engineer who became a naturalized citizen of Germany in 1938. Ferdinand Porsche held dual citizenship with Austria and Germany, but he did not hold triple citizenship. He was known for his design and development of the iconic Porsche sports car and was a key figure in the automotive industry.	0.56	no

Table 23: Examples in the different generation stage on the SQA dataset by GuideKG.

QUESTION: The paper on which American currency is printed has essentially the same ingredients as writing paper?		
Sampling knowledge	Filter score	Answer
The paper used for printing American currency is called "Federal Reserve Note Paper." It is made from a combination of cotton and linen fibers, which gives it the strength and durability needed to withstand the handling and transportation of currency.	0.64	no
The paper used for printing American currency is made from a combination of cotton and linen fibers, which gives it the necessary strength and durability to withstand frequent handling and exposure to moisture.	0.63	no
The paper used for printing American currency is made from a blend of cotton and linen fibers, which gives it a slightly rough texture and makes it durable enough to withstand frequent handling.	0.62	no
The paper used for printing American currency is made from wood pulp and has a similar composition to standard writing paper.	0.36	yes
The paper used for printing American currency is primarily composed of wood pulp, which is derived from wood chips that have been processed in a pulping mill.	0.29	yes
FUSION: The paper used for printing American currency, also known as banknote paper, is made from a combination of cotton and linen fibers to make it resistant to water, tearing, and fading.	0.68	no
QUESTION: University teacher always earns more than fast food worker?		
Sampling knowledge	Filter score	Answer
The average salary of a university teacher varies depending on factors such as the teacher's experience, education level, and the institution they work for.	0.63	no
The average salary of a university teacher in the United States is significantly higher than that of a fast food worker.	0.38	yes
The income disparity between university teachers and fast food workers is due to a variety of factors, including the education and qualifications required for each profession, the level of skill involved in each job, and the demand for each profession.	0.38	yes
The income disparity between university teachers and fast food workers is due to various factors, including education, experience, job responsibilities, and industry demands.	0.37	yes
University teachers typically have a higher level of education and expertise compared to fast food workers, which often leads to higher pay.	0.34	yes
FUSION: University teachers earn more than fast food workers, but the exact amount varies based on factors such as experience, education level, and institution.	0.35	yes

Table 24: Examples in the same generation stage on the CSQA2 dataset by GuideKG.

QUESTION: Those who hold the majority of the house and senate has the ability to pass laws without negotiating with the minority party.			
Stage	Final knowledge	Filter score	Answer
1	In the United States, the majority party in the House of Representatives and the Senate has the power to pass legislation without the support of the minority party, as long as they have a majority of seats in each chamber.	0.40	yes
2	The majority party in the House of Representatives and the Senate in the United States has the power to pass legislation without the support of the minority party, as long as they have a majority of seats in each chamber. This is known as "passing a bill with a simple majority." In the House, a majority is defined as a majority of representatives present and voting, while in the Senate, a majority is defined as a majority of Senators present and voting.	0.42	yes
3	The majority party in the House of Representatives and the Senate in the United States has the power to pass legislation without the support of the minority party, as long as they have a majority of seats in each chamber. The majority party in the House and Senate has the ability to pass laws without negotiating with the minority party, as long as they have the necessary votes. This is known as "passing a bill with a majority vote." The minority party can try to block or delay the passage of a bill using various procedural tactics, but ultimately, the majority party can still pass a bill if they have the necessary votes. This system is known as "majoritarianism" and is designed to ensure that the majority of the population can have its voice heard through the democratic process.	0.43	yes
QUESTION: Dracula is out only during the night because the day sun will kill him.			
Stage	Sampling knowledge	Filter score	Answer
1	Dracula is a fictional character in Bram Stoker's novel "Dracula." In the story, Dracula is a vampire who can only be out during the night because the sun's rays will cause him to weaken and eventually die.	0.38	yes
2	Dracula is a fictional character in Bram Stoker's novel "Dracula." In the story, Dracula is a vampire who can only be out during the night because the sun's rays will cause him to weaken and eventually die. This is because vampires in popular culture are often depicted as being weakened or killed by the sunlight.	0.40	yes
3	Dracula is a fictional character in Bram Stoker's novel "Dracula." In the story, Dracula is a vampire who can only be out during the night because the sun's rays will cause him to weaken and eventually die. This is because vampires in popular culture are often depicted as being weakened or killed by the sunlight. In reality, there is no scientific basis for this belief, and the concept of vampires is purely fictional.	0.69	no

Table 25: Examples in the different generation stage on the CSQA2 dataset by GuideKG.