Boosting Language Models Reasoning with Chain-of-Knowledge Prompting

Anonymous ACL submission

Abstract

Recently, Chain-of-Thought (CoT) prompting has delivered success on complex reasoning tasks, which aims at designing a simple prompt like "*Let's think step by step*" or multiple in-context exemplars with well-designed rationales to elicit Large Language Models (LLMs) to generate intermediate reasoning steps. However, the generated rationales often come with hallucinations, making unfactual and unfaithful reasoning chains. To mitigate this brittleness, we propose a novel Chain-of-Knowledge (CoK) prompting, where we aim at eliciting LLMs to generate explicit pieces of knowledge evidence in the form of structure triple. This is inspired by our human behaviors, i.e., we can draw a mind map or knowledge map as the reasoning evidence in the brain before answering a complex question. Benefiting from CoK, we additionally introduce a F^2-Verification method to estimate the reliability of the reasoning chains in terms of *factuality* and *faithfulness*. For the unreliable response, the wrong evidence can be indicated to prompt the LLM to rethink. Extensive experiments demonstrate that our method can further improve the performance of commonsense, factual, symbolic, and arithmetic reasoning tasks [1].

1 Introduction

Large Language Models (LLMs) have succeeded in advancing the state-of-the-arts for many Natural Language Processing (NLP) tasks (Brown et al., 2020; Rae et al., 2021; Thoppilan et al., 2022; Chowdhery et al., 2022; Scao et al., 2022; Zhang et al., 2022b; Bai et al., 2022; Touvron et al., 2023, *inter alia*), benefiting from the ultra-large-scale training corpora and computation resources. To unleash the LLMs' power of adaptation on unseen tasks without any parameter updates, in-context learning (ICL) has become one of the flourishing research topics, aiming at generating the prediction by conditioning on a few labeled exemplars (Figure 1 (a)) (Shin et al., 2022; Zhao et al., 2021; Liu et al., 2022; Lu et al., 2022; Dong et al., 2023).

A series of recent works have explored that LLMs can spontaneously decompose the complex multi-step problem into intermediate reasoning chains , elicited by a simple prompt like "*Let's think step by step*" or well-designed demonstrations with human-annotated rationales, which are called chain-of-thought (CoT) prompting (Figure 1 (b)) (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023c; Zhou et al., 2023; Yao et al., 2023b). This finding is intriguing and has been sensational because CoT may mainly specify an output space/-format that regularizes the model generation to look step-by-step while being in order and relevant to the query (Wang et al., 2023a; Min et al., 2022b).

Despite impressive performances, current LLMs are susceptible to generating hallucination (Ji et al., 2023; Zhang et al., 2023a), along with providing unfactual or unfaithful reasoning chains that inevitably lead to a wrong conclusion (Wang et al., 2023b). Take Figure 1 as an example. Given a query "*Is the following sentence plausible 'Derrick White backhanded a shot.'*" from StrategyQA (Geva et al., 2021), the standard ICL and CoT make a wrong answer. One of the reasoning steps "Derrick White is most likely a hockey" is fake (In fact, Derrick White is a basketball player), making the unfactual inference towards the question. In addition, the response may be unfaithful when the LLM generates logically sound reasoning chains while still providing an incorrect answer.

To address these concerns, we propose a novel **Chain-of-Knowledge (CoK)** prompting method to boost the LLM's reasoning capability by a series of exemplars that combine explicit structure knowledge evidence with textual explanations. To elaborate, CoK prompting consists of two compositions (Figure 1 (c)), i.e., evidence triples (CoK-ET)

[1]The code and data are available at `https://anonymous.4open.science/r/Chain-of-Knowledge-36EE`.

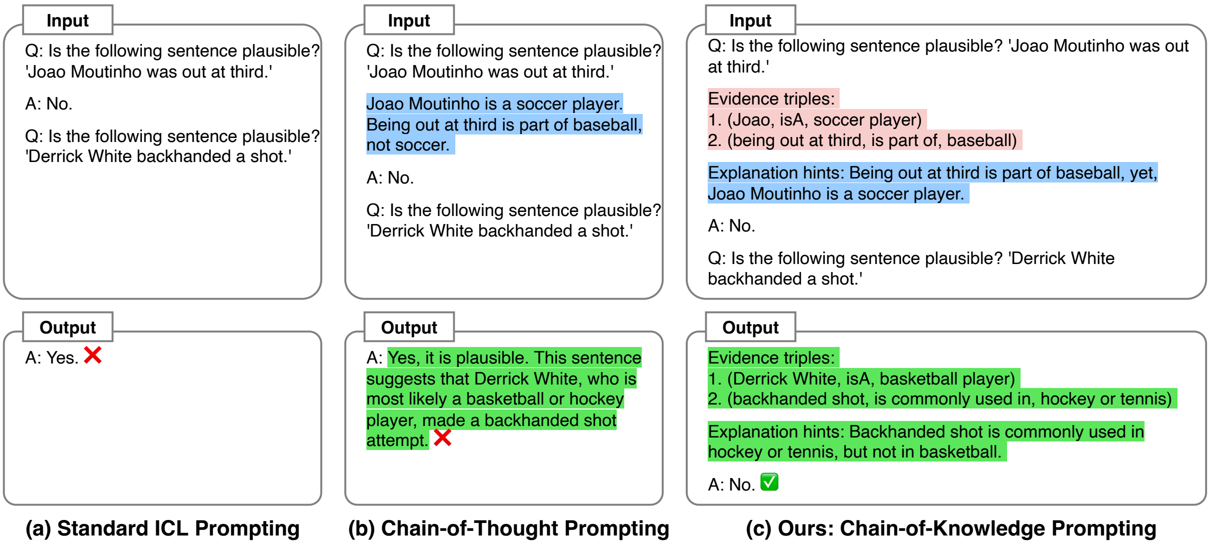

Figure 1: Comparison of three prompting methods: (a) ICL, (b) Chain-of-Thought (CoT), and (c) Chain-of-Knowledge (CoK) solving a StrategyQA question.

and explanation hints (CoK-EH), where CoK-ET is a list of structure triples can reflect the overall reasoning evidence from the query towards the answer and CoK-EH is the explanation of this evidence. To construct in-context exemplars with the CoK prompt, we first sample K labeled examples and each of them can be concatenated with a simple hint "*Let's think step by step*" to prompt the LLM to generate reasoning chains. Then, we retrieve some structure triples from the external knowledge base (KB) and judiciously manually annotate evidence triples to obtain a well-designed CoK prompt. Like standard ICL and CoT, the CoK prompt can be perceived as a rule that regularizes the output space/format and urges LLMs to generate explicit evidence instead of only attempting to generate vague textual reasoning chains. Furthermore, we also propose an F^2-Verification strategy to estimate the reliability of the reasoning chains in terms of *factuality* and *faithfulness*, where *factuality* is the quantification of the matching degree between reasoning evidence and ground-truth knowledge, and *faithfulness* is the consistency degree between reasoning evidence and the textual explanation with the final answer. Particularly for the unreliable response, the wrong pieces of evidence can be indicated to prompt the LLM to rethink this problem. We design a *rethinking algorithm* to reach this goal.

We have conducted empirical evaluations across various reasoning tasks (e.g., commonsense, factual, arithmetic, and symbolic), showing that CoK prompting with F^2-Verification can significantly outperform standard ICL and CoT prompting. We also integrate CoK prompting with some prevailing strategies, such as self-consistency. The results indicate that such CoK can serve as a plug-and-play module to further improve reasoning ability.

2 Related Work

Prompting for LLMs with in-context learning. In-Context Learning (ICL) is the task of causal language modeling, allowing LLMs to perform zero/few-shot learning with a well-designed text-based prompt (Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023; Thoppilan et al., 2022; Dong et al., 2023). ICL can bypass the model parameter update and achieve the salient performance by conditioning on a few labeled examples. Previous works have explored some impact facets of ICL. For example, the input-output mapping and the template format (Pan et al., 2023; Min et al., 2022b; Yoo et al., 2022), the different selection and permutations of the exemplars (Lu et al., 2022). To improve ICL's effectiveness, some novel methods have been proposed, involving meta-learning (Chen et al., 2022; Min et al., 2022a), prompt and exemplars engineering (Liu et al., 2022, 2023).

Chain-of-thought prompting elicits reasoning. Recently, CoT prompting has been presented to leverage reasoning and interpretable information to guide LLMs to generate reliable and explainable responses (Wei et al., 2022). A series of CoT-enhanced methods are proposed to further improve the reasoning ability (Kojima et al., 2022; Huang et al., 2022; Wang et al., 2023c; Si et al.,

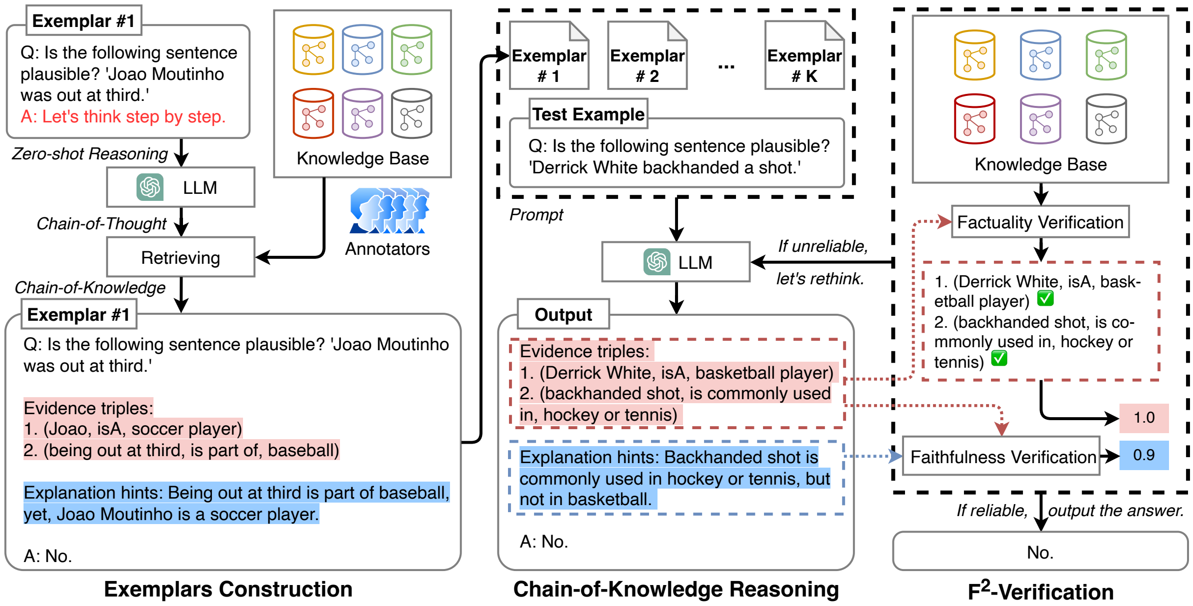

Figure 2: The proposed framework. We first construct exemplars with chain-of-knowledge (CoK) prompts. Then, the CoK prompts can be indicated to let the LLM generate reasoning chains, including evidence triples, explanation hints, and the final answer. At last, we estimate the reliability of reasoning chains in terms of *factuality* and *faithfulness*, and the unreliable ones will be rethought.

2022; Wang et al., 2022; Zhou et al., 2023; Zhang et al., 2023b; Fu et al., 2023; Besta et al., 2023). For example, Wang et al. (2023c) introduce *Self-consistency* to suppress the wrong rationales problem by marginalizing out the sampled reasoning paths to find the most consistent answer. Fu et al. (2023) and Besta et al. (2023) proposed logical thinking graph to let LLMs better reasoning. (Lyu et al., 2023) translates the complex problem into interleaved natural language or programming language to make the reasoning chains faithful. (Li et al., 2023) and (Yao et al., 2023a) introduce coarse or fine-grained labels to verify the reasoning chains. Differently, we focus on alleviating the hallucination in terms of the factuality and faithfulness of the reasoning chains.

3 Methodology

The generated reasoning chains elicited by CoT prompting sometimes come with mistakes, ultimately leading to hallucinated answers. We attribute this challenge to the textual reasoning chain: *LLMs may forcibly generate a textual rationale that conforms to the prompt format of CoT but is logically ambiguous and reaches the wrong answer.* To address this challenge, we provide our specific solution on how to boost LLM's reasoning ability in two corners: elicitation format of the prompt and knowledge-enhanced post-verification. The overview of the framework is shown in Figure 2.

3.1 Chain-of-Knowledge Prompting

It is widely recognized that reasoning can be modeled as induction and deduction on the existing knowledge system (Goswami, 2002). This is inspired by human behaviors that draw a mind map or knowledge map to analyze the question and find the correct path to the answer. Fortunately, we can adopt the concept of the triple in the KB, which can be viewed as a "(subject, relation, object)", to formalize the explicit evidence of the reasoning chains. To elaborate, we propose **Chain-of-Knowledge** (CoK) prompting to facilitate a better elicitation prompt for LLMs, which consists of two main ingredients, i.e., evidence triples (CoK-ET) and explanation hints (CoK-EH). CoK-ET represents a list of multiple triples and each of them represents the knowledge evidence probed from LLMs to support the step-by-step thinking process. CoK-EH denotes the explanation of the reasoning chain, which is similar to CoT. Take Figure 1 as an example, we can urge the LLM to generate explicit shreds of evidence to support the final answer.

3.2 Exemplars Construction

Building upon the insights of previous studies (Zhang et al., 2023b; Min et al., 2022b; Wang et al., 2023c), the performance of ICL hinges on the annotated rationale. This indicates that the key challenge of CoK prompting lies in constructing textual rationales with their structure evidence triples. As shown in Figure 2, we first perform exemplars construction to obtain a well-designed task-specific prompt. Specifically, we follow (Wei et al., 2022; Wang et al., 2023c) to randomly select K questions as the basic demonstrations. To automatically obtain CoK-EH, we follow (Kojima et al., 2022) to generate a textual rationale for each question via zero-shot CoT with a simple prompt "*Let's think step by step*". Another challenge is how to obtain annotated CoK-ET that better expresses the textual rationale. To figure it out, we first follow (Pan et al., 2022) to construct a KB $\mathcal{K}$ from six domains, involving *dictionary*, *commonsense*, *entity*, *event*, *script*, and *causality*, which are in the form of triple. We then directly use the retrieving tool proposed by (Pan et al., 2022) to retrieve some candidate triples. Finally, we invite 5 people as professional annotators to manually design the corresponding CoK-ET based on the retrieved triples [2]. More details can be found in Appendix D.1.

3.3 F^2-Verification

After the exemplars construction, we can obtain K annotated data $\mathcal{E} = \{(Q_i, T_i, H_i, A_i)\}_{i=1}^{K}$. Symbolically, Q_i, H_i and A_i represent the input query, the explanation hint, and the final answer of the i-th exemplar, respectively; each of them is the token sequence. T_i denotes the list of evidence triples, which contains multiple knowledge triples, i.e., $T_i = \{(s_{ij}, r_{ij}, o_{ij})\}_j$, where s_{ij}, r_{ij} and o_{ij} are subject, relation and object, respectively. Given a test query input $\hat{Q}_i$, we can directly choose one permutation of $\mathcal{E}$ and concatenate them with this test query into a linear sequence $\hat{I}_i = [\mathcal{E}; \hat{Q}_i]$ to prompt the LLM to generate the prediction, i.e., $\hat{y}_{ik} = \arg\max_{\sigma(\hat{y}_{i(\leq k)})} P(y|\hat{y}_{i(<k)}, \mathcal{E}, \hat{Q}_i)$, where $P(y|\cdot)$ is the prediction distribution, $\hat{y}_{ik}$ is the k-th token, $\sigma(\cdot)$ denotes the decoding strategy (e.g., temperature sampling, beam search, and nucleus sampling), $[\cdot;\cdot]$ is the concatenation operation.

Due to the well-designed format of the demonstrations, the final prediction derived from the LLM $\hat{y}_i$ consists of a list of evidence triples $\hat{T}_i$, a sequence of explanation hints $\hat{H}_i$ and the final answer $\hat{A}_i$. However, LLMs may generate hallucinated rationales, so the final answer can not be guaranteed. We attribute this problem to two factors: 1) some steps in the rationale may not correspond to the fact, contributing to the wrongness, and 2) the relation between the final answer and the reasoning chains is still ambiguous, making the response unfaithful. To alleviate these drawbacks, we propose F^2-Verification to estimate the answer reliability towards both **F**actuality and **F**aithfulness [3].

Factuality Verification. We first verify the factuality, which can be viewed as the matching degree between each generated evidence triple and the ground-truth knowledge from KBs [4]. Concretely, we first define a function $f_v(\hat{r}_{ij}|\hat{s}_{ij}, \hat{o}_{ij}, \mathcal{K})$ to represent the factuality of each evidence. We design two different strategies of f_v. 1) Exact verification. We can retrieve all relevant triples based on the subject $\hat{s}_{ij}$ and object $\hat{o}_{ij}$, and then find whether the generated relation $\hat{r}_{ij}$. i.e., $f_v(\hat{r}_{ij}|\hat{s}_{ij}, \hat{o}_{ij}, \mathcal{K}) = \mathbb{I}((\hat{s}_{ij}, \hat{r}_{ij}, \hat{o}_{ij}) \in \mathcal{K})$ exists. 2) Implicit verification. For a triple that does not exist in KB, it could be corrected. Thus, we can transform the factuality verification into a graph completion task that predicts whether the triple is true. For simplicity, we use TransR (Lin et al., 2015) to pretrain the KB $\mathcal{K}$ and use the off-the-shelf energy function to assign a score for each evidence triple, i.e., $f_v(\hat{r}_{ij}|\hat{s}_{ij}, \hat{o}_{ij}, \mathcal{K}) = ||\mathbf{s}_{ij}^{(r,c)} + \mathbf{r}^c - \mathbf{o}_{ij}^{(r,c)}||_2^2 + \alpha||\mathbf{r}^c - \mathbf{r}_{ij}||_2^2$, where $\alpha > 0$ is the balancing factor, $||\cdot||_2^2$ is Frobenius norm. $\mathbf{s}_{ij}^{(r,c)} = \mathbf{s}_{ij}\mathbf{M}_r$ and $\mathbf{o}_{ij}^{(r,c)} = \mathbf{o}_{ij}\mathbf{M}_r$ denote the projection representations of the subject s_{ij} and object o_{ij} in the relation space r, respectively. $\mathbf{s}_{ij}, \mathbf{r}_{ij}$, and $\mathbf{o}_{ij} \in \mathbb{R}^d$ are the d-dimension embeddings of $\hat{s}_{ij}$, $\hat{r}_{ij}$ and $\hat{o}_{ij}$, respectively. $\mathbf{r}^c \in \mathbb{R}^d$ is the prototype embeddings of the relation r. $\mathbf{M}_r \in \mathbb{R}^{d\times d}$ is the trainable projection matrix of relation r. We join the two strategies in our framework. If the evidence triple exists in $\mathcal{K}$, we will use an exact verification strategy to assign a score, or we use an implicit verification strategy.

Faithfulness Verification. As defined by (Jacovi and Goldberg, 2020; Lyu et al., 2023), *if the reasoning process derived from the model can accurately be expressed by an explanation, we call it*

[2] In fact, during the exemplars construction, the generated textual reasoning chains and retrieved triples could have some mistakes. Fortunately, we found that there is no strong connection between the reasoning validity of both CoK-ET and CoK-EH and the performance of the model predictions, which is similar to findings in (Wang et al., 2023a). We will bring detailed discussions at Section 4.3.

[3] We find that (He et al., 2022) also proposed rethinking and retrieving processes to reduce wrongness, different from them, we focus on fine-grained detection and injection with the proposed CoK prompt.

[4] We assume that the knowledge from the KB is correct and up-to-date.

faithful. Previous works based on chain-of-thought prompting are hard to verify faithfulness due to the lack of sufficient evidence to understand the relation between the explanation and the answer (Ye and Durrett, 2022). So, we propose a faithfulness verification method to find out these cases. Specifically, given one test query $\hat{Q}_i$, a list of evidence triples $\hat{T}_i$ and the final answer $\hat{A}_i$, we directly concatenate them as a new sequence $\hat{H}'_i$. We leverage the pre-built sentence encoder SimCSE (Gao et al., 2021) to calculate the similarity between $\hat{H}'_i$ and $\hat{H}_i$. We denote this function as $f_u(\hat{H}_i|\hat{H}'_i = [\hat{Q}_i; \hat{T}_i; \hat{A}_i]) = SimCSE(\hat{H}_i, \hat{H}'_i)$.

Finally, for each query $\hat{Q}_i$ we can obtain a score $\mathcal{C}_i$ $(0 < \mathcal{C}_i < 1)$ that represents whether the rationale is reliable towards the answer:

$$\begin{aligned} \mathcal{C}_i = \gamma \frac{1}{|\hat{T}_i|} \sum_{j=1}^{|\hat{T}_i|} f_v(\hat{s}_{ij}|\hat{r}_{ij}, \hat{o}_{ij}, \mathcal{K}) \\ + (1-\gamma) f_u(\hat{H}_i|\hat{H}'_i = [\hat{Q}_i; \hat{T}_i; \hat{A}_i]), \end{aligned} \tag{1}$$

where $0 < \gamma < 1$ is the balancing factor and set to be 0.5 as default, $|\hat{T}_i|$ is the number of triples.

3.4 Rethinking Process

F^2-Verification facilitates us to ensure the factuality and faithfulness of the triples and explanations generated by the model. Moreover, beyond the scope of verification, we can boost the performance of LLMs even further by employing a *rethinking process*. The algorithm pseudo-code Algorithm 1 can be found in Appendix A, we initialize a reliability threshold θ $(0 < \theta < 1)$, iteration number N, and an unreliability set U. All the queries in the testing set $\mathcal{D}_{test}$ are unreliable at first. For each query $\hat{Q}_i \in U$ in the n-th iteration, we obtain CoK prompt $\hat{I}_i^{(n)}$ by combining the demonstrations $\mathcal{E}$ and the query $\hat{Q}_i$. The prompt can be used to elicit the LLM to generate a list of evidence triples $\hat{T}_i^{(n)}$, explanation hints $\hat{H}_i^{(n)}$, and the final answer $\hat{A}_i^{(n)}$.

The proposed rethinking algorithm will allow the LLMs to assess the reliability of the rationales (i.e., $\hat{T}_i^{(n)}$ and $\hat{H}_i^{(n)}$) via calculating the score in Eq. 1. An entry is no longer considered unreliable if $\mathcal{C}_i^{(n)}$ is not below the threshold θ, which subsequently leads to the final answer $\hat{A}_i$. Conversely, if $\mathcal{C}_i^{(n)}$ fails to reach θ, we can select the evidence triples with lower scores and inject the corresponding correct knowledge triples from the KB into the CoK prompt $\hat{I}_i^{(n+1)}$ in the next iteration (Line 12, Algorithm 1) [5]. This dynamic generate-evaluate procedure continues until all entries in U are considered reliable or the maximum number of iterations N is reached. For cases where the maximum number of iterations is reached without any triples' reliability score surpassing θ, triples with the highest reliability scores will be selected for inference (Line 15-17, Algorithm 1).

4 Experiments

4.1 Experimental Setups

Tasks and Datasets. During experiments, we choose five different kinds of tasks to evaluate the performance of our method. The datasets and corresponding implementation details are shown in the following. 1) **Commonsense & factual reasoning**. We select CommonsenseQA (CSQA) (Talmor et al., 2019), StrategyQA (Geva et al., 2021), OpenBookQA (Mihaylov et al., 2018) the AI2 Reasoning Challenge (ARC-c) (Clark et al., 2018), sports understanding from the BIG-Bench benchmark (bench collaboration., 2022), and BoolQ (Clark et al., 2019) for evaluating CoK on commonsense and factual reasoning. 2) **Symbolic reasoning.** Two symbolic reasoning tasks are evaluated in our experiments, specifically, Last Letter Concatenation and Coin Flip tasks (Wei et al., 2022). 3) **Arithmetic reasoning.** We use grade school math problems GSM8K (Cobbe et al., 2021), a challenging dataset over math word problems SVAMP (Patel et al., 2021), and two others AQuA (Ling et al., 2017), MultiArith (Roy and Roth, 2015) for math problem solving tasks.

Implementation Details. For the LLM, we employ the publicly accessible GPT-3 (Brown et al., 2020) models, namely, *gpt-3.5-turbo* and *text-davinci-002* with 175B parameters unless otherwise stated. We use greedy decoding with temperature 0 and max output length 512, keeping consistent with baselines for a fair comparison. For the datasets from commonsense reasoning and factual reasoning, the KBs we choose are a combination of Wiktionary [6], ConceptNet (Speer et al., 2017), Wikidata5M (Wang et al., 2021), ATOMIC (Sap et al., 2019), GLUCOSE (Mostafazadeh et al., 2020), ASER (Zhang et al., 2020, 2022a), and CausalBank (Li et al., 2020). For the Last Letter

[5] This is similar to correcting wrong reasoning paths, we ensure that the label is not leaked to the model.

[6] https://en.wiktionary.org/wiki/Wiktionary.

Model	*Commonsense & Factual*						*Symbolic*		*Arithmetic*			
	Common Sense QA	Strategy QA	OpenBook QA	ARC-c	Sports	BoolQ	Letter	Coin	GSM8K	SVAMP	AQuA	MultiArith
Fine-tuning	91.2	73.9	91.0	75.0	-	92.4	-	-	55.0	57.4	37.9	-
text-davinci-002 reasoning results												
Zero-Shot SP	68.8	12.7	44.7	46.8	38.1	50.2	0.2	12.8	10.4	58.8	22.4	17.7
Zero-Shot CoT	64.6	54.8	68.4	64.7	77.5	52.7	57.6	91.4	40.7	62.1	33.5	78.7
Few-Shot SP	**79.5**	65.9	**76.6**	68.2	69.6	53.6	0.0	49.1	15.6	65.7	24.8	33.8
Manual CoT	73.5	65.4	73.0	69.9	82.4	55.0	59.0	74.5	46.9	68.9	35.8	91.7
Auto-CoT	74.4	65.4	-	-	-	-	59.7	**99.9**	47.9	69.5	36.5	92.0
CoK	75.4	66.6	73.9	71.1	83.2	56.8	59.4	97.4	**51.2**	**69.9**	**37.8**	**94.6**
CoK + F^2-V	77.3	**67.9**	74.8	**73.0**	**84.1**	**59.9**	**61.1**	-	-	-	-	-
gpt-3.5-turbo reasoning results												
Manual CoT	76.5	62.6	82.6	84.9	84.0	65.1	73.0	97.4	79.1	79.5	55.1	97.3
Manual CoT + SC	78.2	63.7	85.0	86.5	86.5	66.6	**74.5**	99.0	87.6	85.0	66.8	98.8
ComplexCoT	75.4	62.2	-	-	-	-	-	-	79.3	77.7	56.5	95.4
ComplexCoT + SC	76.0	63.0	-	-	-	-	-	-	**89.2**	85.6	65.0	98.23
CoK	77.1	63.8	83.5	85.7	85.9	67.9	63.1	98.0	83.2	81.4	60.2	99.0
CoK + SC	78.9	65.0	86.1	**87.5**	87.4	69.4	68.3	**99.2**	88.2	**86.0**	**69.7**	**99.3**
CoK + F^2-V	77.8	64.5	85.0	86.6	87.0	69.2	65.4	-	-	-	-	-
CoK + SC + F^2-V	**79.3**	**66.6**	**87.0**	87.4	**87.9**	**69.9**	69.7	-	-	-	-	-

Table 1: Accuracy of *gpt-3.5-turbo* model over commonsense, factual, symbolic, and arithmetic reasoning tasks.

Connection task in Symbolic Reasoning, we manually construct a dictionary KB for each word in Wiktionary. For example, the triple of the word "system" is "(system, last letter, m)". For the rest datasets (e.g., arithmetic reasoning and coin dataset), we do not perform F^2-Verification because we can not find any KBs for these tasks. More details is shown in Appendix D.

Baselines. In our experiments, we first consider few-shot/zero-shot standard prompting (SP) popularized by Brown et al. (2020) as the naive baselines, and then some prevailing methods serve as strong baselines. 1) Chain-of-thought prompting (Few-Shot CoT & Manual CoT) (Wei et al., 2022), 2) Zero-Shot CoT (Kojima et al., 2022), 3) Auto-CoT (Zhang et al., 2023b), 4) Complexity-based prompting (ComplexCoT) (Fu et al., 2023). We also integrate Self-Consistency (SC) into Manual CoT, ComplexCoT, and our CoK when validating the *gpt-3.5-turbo* model. The number of sampled reasoning paths is 10.

4.2 Competitive Performance of CoK

We first evaluate on commonsense and factual reasoning. As shown in Table 1, we make the following observations: 1) CoK prompting steadily exceeds the performance of the previous CoT strategies. Specifically, our method respectively achieves 1.9%, 1.2%, 0.9%, 1.5%, 0.8%, and 1.8% improvement with *text-davinvi-002*, and respectively achieves 0.6%, 1.2%, 0.9%, 0.8%, 1.9%, and 2.8% improvement with *gpt-3.5-turbo*. This demonstrates that the combination of explicit evidence triples and explanation can boost the LLMs' reasoning ability. This also suggests that a better elicitation format is critical for prompt-based learning. 2) Based on F^2-Verification, the performances can be further enhanced across tasks. In particular, the performance of CoK + F^2-Verification nearly approaches fine-tuning on the StrategyQA and ARC-c tasks. This indicates that conducting post-verification and correcting false evidence triples by injecting ground-truth knowledge is crucial to the reasoning. 3) CoK steadily outperforms the ComplexCoT method by a significant margin, which requires a much higher computational cost than our approach.

We also explore how can CoK prompting adapt to non-knowledge-intensive tasks, such as symbolic and arithmetic reasoning. Results in Table 1 suggest that CoK prompting can also make high improvements on these tasks, indicating that decomposing the reasoning chains into explicit triples is helpful for LLMs to understand complex tasks.

Finally, we compare some ensemble baselines with *self-consistency* (SC), and we find 1) self-consistency can substantially improve the accuracy

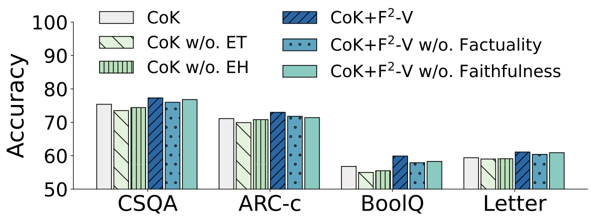

Figure 3: Ablation study results: accuracy when we remove different components.

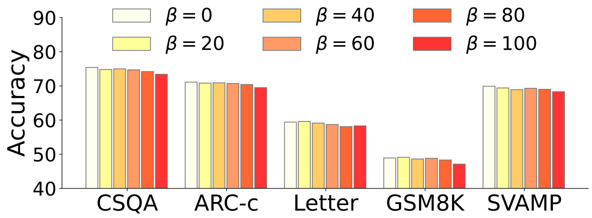

Figure 4: Effect of wrong demonstrations with $\beta\%$ wrong evidence triples.

Domain d	CSQA	StrategyQA	OpenBookQA	ARC-c
$\#d \rightarrow \#d$	75.4	66.6	59.4	48.9
GSM8K$\rightarrow \#d$	73.6	61.4	55.0	47.3

Table 2: Domain adaptation results (%). $\#d$ means the domain of exemplars, $\#d \rightarrow \#d$ means the examples sampled from the current domain, while GSM8K$\rightarrow \#d$ means the current task use GSM8K exemplars.

on Manual CoT, ComplexCoT and CoK, 2) CoK + SC + F^2-V achieve the best performances on most tasks, where SC boosts reasoning by assembling all reasoning paths at each rethinking iteration and F^2-V boosts reasoning by assembling all iterations.

4.3 Discussions

Ablation Study. In this section, we aim to explore how much each part of the component contributes to the performance. We perform an ablation study to see how the performance changes. We conduct the experiments on four tasks, including CSQA, ARC-c, BoolQ, and the Last Letter Connection. The ablation settings are shown in Appendix D.3. Results in Figure 3 demonstrate that the performance drops when removing each component, which shows the significance of all components. For CoK, we can see that the performance of the variant CoK w/o. ET is lower than CoK w/o. EH on all the tasks, which suggests that urging the LLM to generate explicit evidence triples is the most important contribution to the performance. In addition, both the evidence triples and explanation hints can be fully utilized in the rethinking process because they can guide the LLM to verify the reasoning chains via either factuality or faithfulness.

Effect of Wrong Demonstrations. Recall the discussion in Section 3.2 about the demonstrations that may have some mistakes. To see if chain-of-knowledge prompting has a similar phenomenon to previous works (Wang et al., 2023a) that there is no strong connection between the validity of reasoning chains and the performance of model prediction. We perform negative random replacement when constructing exemplars. Specifically, we choose $\beta\%$ evidence triples in each in-context example and replace them randomly from the KB to form a wrong reasoning path. We choose five tasks and draw some bar charts in Figure 4. Results illustrate that the accuracy rate drops slightly as the value of β increases from 0 to 100. However, even if the evidence triplets are all wrong, the performance will not decrease significantly. This phenomenon is counter-intuitive, yet, in line with the expected situation we considered before.

Domain Adaptation of Demonstrations. To investigate the adaptation of CoK prompting, for each exemplar, we replace the prompt with an alternative exemplar from other domains. Specifically, we choose CSQA, StrategyQA, OpenBookQA, and ARC-c tasks from commonsense reasoning, and for each task, we select the demonstrations from the GSM8K task, which is very different from them. The results for the domain adaptation settings are shown in Table 2. The performance is exciting because we find the LLM can easily know how to follow the chain-of-knowledge paradigm to solve a new problem, although the given prompt is completely irrelevant.

Model Effectiveness To investigate the effectiveness of CoK when applying different backbones, we extra choose *gpt-4* to evaluate CSQA and GSM8K. As shown in Figure 5, by comparing with CoT and CoT+SC, CoK and CoK+SC achieve average performance improvements of 1.6% on *text-davinci-002*, 1.4% on *gpt-3.5-turbo* and 1.0% on *gpt-4*, indicating that CoK is adaptable to various LLMs and effectively boosts performance across different backbones.

Prompt Engineering. In Figure 6, we analyze the effectiveness of different prompt engineering strategies. "Manual" denotes constructing prompt via human annotation, while "Auto" means to use zero-shot CoT and KB to build prompt. Results

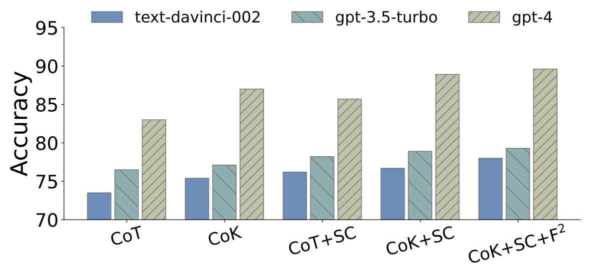

Figure 5: Comparison of CoT, CoT+SC, CoK, CoK+SC and CoK+SC+F^2-V over CSQA when using different backbone.

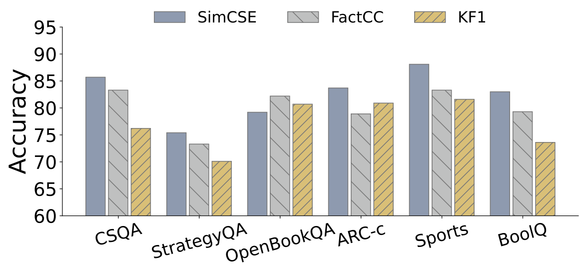

Figure 7: Hallucination Evaluation of different faithfulness score $f_u(\cdot)$.

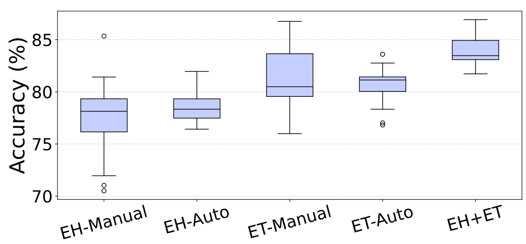

Figure 6: Performance of *gpt-3.5-turbo* over GSM8K with different prompt.

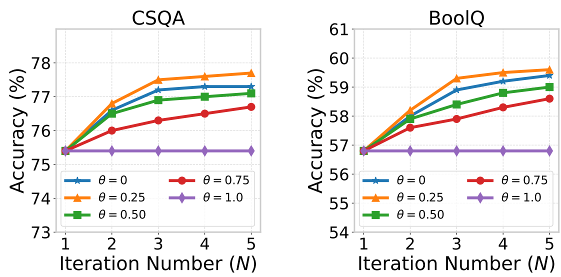

Figure 8: Effectiveness of different rethinking iteration N and reliability threshold θ.

demonstrate that leveraging zero-shot CoT with KB can reduce the variance and improve the accuracy.

Hallucination Evaluation. To investigate the hallucination, we choose different faithfulness scores $f_u(\cdot)$ to make the comparison. Apart from SimCSE, we also choose FactCC (Kryscinski et al., 2020) and Knowledge F1 (KF1) (Shuster et al., 2021). As shown in Figure 7, we choose all tasks from commonsense reasoning to make an evaluation. We find that our framework can achieve the highest accuracy on most of the tasks when using SimCSE as a faithfulness score, which indicates the effectiveness of hallucination reduction.

Effectiveness of Rethinking Process. Recall the *rethinking process*, when the reasoning chains generated by the LLMs fail to pass verifications and the reliability score is below the threshold θ, we provide them with additional opportunities to regenerate in the rethinking stage. Figure 8 demonstrates the effectiveness over three tasks with different combinations of rethinking iteration number $N \in \{1, 2, 3, 4, 5\}$ and threshold $\theta \in \{0, 0.25, 0.5, 0.75, 1.0\}$. From the analysis, we can draw some following suggestions. 1) In most cases, the accuracy increases a lot when the LLM rethinks step-by-step in the first 3 iterations. 2) The rethinking process converges faster when using a smaller threshold. It is not difficult to understand that when the threshold is small, almost all testing queries will be injected with knowledge and regenerated, which is similar to the method of combining *self-consistency* and F^2-Verification. Interestingly, we observe that the performance may decrease by about 2% when $\theta < 0.25$, we blame it on an over-injection problem because it may inject some irrelevant or inconsistent information.

5 Conclusion

We propose chain-of-knowledge prompting, which aims to decompose the reasoning chains derived from the LLMs into multiple evidence triples and explanation hints, to further improve the reasoning capabilities. Based on the chain-of-knowledge prompt, we introduce F^2-Verification and fully exploit external knowledge bases to perform post-verification for the generated reasoning chains in terms of factuality and faithfulness. A rethinking process then be used to inject knowledge to correct the false evidence triples and elicit the LLM to regenerate the answer. Our extensive results show that it outperforms other prompt methods over multiple reasoning tasks. In the future, we will 1) further improve the performance of other scale LLMs, 2) extend the KB to search engines to realize real-time verification, and 3) perform interpretability analysis on LLMs' reasoning.

Limitations

Our work is based on prompting methods for large language models and achieves outstanding performance across several benchmarks. However, it still carries the following limitations: (1) The evidence triples in knowledge bases are finite, which might not ensure comprehensive coverage of the model's requirements for all questions. (2) In light of the integration of the re-thinking algorithm, CoK might require more API calls compared to vanilla CoT methods.

Social Impact and Ethics

In terms of social impact, the knowledge bases we utilize are all from publicly available data sources. Infusing factual knowledge into the model's reasoning process will not introduce additional bias. Moreover, it can to some extent prevent the model from providing irresponsible and harmful answers.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

BIG bench collaboration. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

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

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 *NeurIPS*.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 719–730. Association for Computational Linguistics.

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

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *CoRR*, abs/2301.00234.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

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. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Usha Goswami. 2002. Inductive and deductive reasoning. *Blackwell handbook of childhood cognitive development*, pages 282–302.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3887–3896. PMLR.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *CoRR*, abs/2210.11610.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4198–4205. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9332–9346. Association for Computational Linguistics.

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

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2020. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3629–3636. International Joint Conferences on Artificial Intelligence Organization. Main track.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2181–2187. AAAI Press.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 158–167. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

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

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

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2791–2809. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, Toronto, Canada. Association for Computational Linguistics.

Xiaoman Pan, Wenlin Yao, Hongming Zhang, Dian Yu, Dong Yu, and Jianshu Chen. 2022. Knowledge-in-context: Towards knowledgeable semi-parametric language models. *CoRR*, abs/2210.16433.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1743–1752. The Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *ArXiv*, abs/1811.00146.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv preprint*, abs/2211.05100.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woo-Myoung Park, Jung-Woo Ha, and Nako Sung. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5168–5186. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2022. Prompting GPT-3 to be reliable. *CoRR*, abs/2210.09150.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen 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 Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

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. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023b. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

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*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *CoRR*, abs/2207.00747.

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*.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2422–2437. Association for Computational Linguistics.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022a. ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artificial Intelligence*, 309:103740.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *WWW*, pages 201–211.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. Opt: Open pre-trained transformer language models.

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*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

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*.

A Rethinking Algorithm

The rethinking algorithm is shown in 1.

Algorithm 1 Rethinking Process

Require: Exemplars $\mathcal{E}$, testing query set $\mathcal{D}_{test} \leftarrow \{\hat{Q}_i\}_{i=1}^{M}$, KB $\mathcal{K}$, iterator number $N(\geq 1)$, reliability threshold $0 < \theta < 1$.

Initialize an unreliability set $U \leftarrow \mathcal{D}_{test}$.
for each iteration $n \leftarrow 1, \cdots, N$ **do**
  for each query $\hat{Q}_i$ in U **do**
    Obtain a CoK prompt $\hat{I}_i^{(n)}$. If n is 1, $\hat{I}_i^{(n)} \leftarrow [\mathcal{E}; \hat{Q}_i]$.
    Generate evidence triple $\hat{T}_i^{(n)}$, explanation hint $\hat{H}_i^{(n)}$ and answer $\hat{A}_i^{(n)}$ from the LLM.
    Calculate reliability score $\mathcal{C}_i^{(n)}$ in Eq. 1.
    if $\mathcal{C}_i^{(n)} \geq \theta$ **then**
      Obtain final answer $\hat{A}_i \leftarrow \hat{A}_i^{(n)}$.
      Remove $\hat{Q}_i$ from U.
      continue
    end if
    For the evidence triples that $f_v(\hat{r}_{ij}^{(n)}|\hat{s}_{ij}^{(n)}, \hat{o}_{ij}^{(n)}, \mathcal{K}) < \theta$, inject the corresponding correct knowledge triples $\hat{T}_i'$ into the prompt, i.e., $\hat{I}_i^{(n+1)} \leftarrow [\hat{I}_i^{(n)}; \hat{T}_i']$.
  end for
end for
for each query $\hat{Q}_i$ in U **do**
  Obtain the final answer $\hat{A}_i \leftarrow \arg\max_{\hat{A}_i^{(n)}} \mathcal{C}_i^{(n)}$.
end for
return all the answers $\{\hat{A}_i\}_{i=1}^{M}$.

B Case Study

We end this section with a case study to show the effectiveness of our proposed chain-of-knowledge prompting and the rethinking process with F^2-Verification. We randomly choose two examples from CSQA and Last Letter Connection tasks, and the results are listed in Table 3. We can see that our proposed method can effectively generate explicit evidence triples with corresponding explanation hints, and the wrong triples can be detected through the proposed F^2-Verification. During the rethinking process, the LLM can be guided with a new prompt with injected knowledge and then reach the correct answer. We also find the LLM can spontaneously check whether the injected knowledge is useful for the question solution (as shown in Appendix E.1), which indicates that our method can achieve meticulous injection and avoid injecting useless knowledge.

C Statistics of Datasets

We evaluate CoK on 12 publicly available benchmark datasets that cover arithmetic reasoning, commonsense reasoning, symbolic reasoning and natural language understanding tasks. The statistics of the datasets are shown in Table 4.

The detailed information of each dataset is shown in the follow:

Commonsense & Factual reasoning

- CSQA (Talmor et al., 2019): it is a commonsense QA task and the answer format is multi-choice. The homepage is `https://www.tau-nlp.org/commonsenseqa`, and `https://github.com/jonathanherzig/commonsenseqa`.

- StrategyQA (Geva et al., 2021): it is a commonsense QA task with Yes or No answer format. We use the open-domain setting (question-only set) from (bench collaboration., 2022): `https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/strategyqa`. The original dataset is from `https://github.com/eladsegal/strategyqa`, MIT license: `https://github.com/eladsegal/strategyqa/blob/main/LICENSE`.

- OpenBookQA (Mihaylov et al., 2018): it is a multi-choice QA task to evaluate commonsense knowledge. The original dataset is from `https://allenai.org/data/open-book-qa`.

- ARC-c (Clark et al., 2018): it is a multip-choice commonsense QA task. The original dataset is from

Query	Evidences Triples & Explanations	Injected Knowledge Prompt	Rethinking
[K **exemplars**]\n\n What island country is ferret popular? Answer Choices: (A) own home (B) north carolina (C) great britain (D) hutch (E) outdoors	A: Evidence triples: 1. (ferret, isA, animal)\n 2. (country, isA, place)\n 3. (ferret, popular, Britain)\n 4. (north carolina, located in, Britain)\n Explanation hints: The answer must be a place where ferrets are popular. Of the above choices, only North Carolina is a place where ferrets are popular.\n So the answer is (D).	[K **exemplars**]\n\n What island country is ferret popular? Answer Choices: (A) own home (B) north carolina (C) great britain (D) hutch (E) outdoors\n (North Carolina, located in, US.)\n(Great Britain, located in, Britain)	A: Evidence triples: 1. (ferret, isA, animal)\n 2. (country, isA, place)\n 3. (ferret, popular, Great Britain)\n 4. (Great Britain, located in, Britain)\n Explanation hints: The answer must be a place where ferrets are popular. Of the above choices, only North Carolina is a place where ferrets are popular.\n So the answer is (C).
[K **exemplars**]\n\n Take the last letters of each words in "Prince Rene Vishal Patrick" and concatenate them.	A: Evidence triples:\n1. (Prince, last latter, e) \n2. (Rene, last latter, e) \n3. (Vishal, last latter, i) \n4. (Patrick, last latter, k) \n5. (final answer, is, eelk) \n Explanation hints: The last letter of "Prince" is "e". The last letter of "Rene" is "e". The last letter of "Vishal" is "i". The last letter of "Patrick" is "k". Concatenating them is "eeik". \n So the answer is eeik.	[K **exemplars**]\n\n Take the last letters of each words in "Prince Rene Vishal Patrick" and concatenate them.\n (Vishal, last latter, l)	A: Evidence triples:\n1. (Prince, last latter, e) \n2. (Rene, last latter, e) \n3. (Vishal, last latter, l) \n4. (Patrick, last latter, k) \n5. (final answer, is, eelk) \n Explanation hints: The last letter of "Prince" is "e". The last letter of "Rene" is "e". The last letter of "Vishal" is "l". The last letter of "Patrick" is "k". Concatenating them is "eelk". \n So the answer is eelk.

Table 3: Case study on CSQA and Last Letter Connection: the chain-of-knowledge prompting with the rethinking process. The tokens in red, blue, and green respectively denote the wrong rationales, the injected knowledge, and the corrected rationales.

Dataset	Number of samples	Average words	Answer Format	Licence
CSQA	1,221	27.8	Multi-choice	Unspecified
StrategyQA	2,290	9.6	Yes or No	Apache-2.0
OpenBookQA	500	27.6	Multi-choice	Unspecified
ARC-c	1,172	47.5	Multi-choice	CC BY SA-4.0
Sports	1,000	7.0	Yes or No	Apache-2.0
BoolQ	3,270	8.7	Yes or No	CC BY SA-3.0
Last Letters	500	15.0	String	Unspecified
Coin Flip	500	37.0	Yes or No	Unspecified
GSM8K	1,319	46.9	Number	MIT License
SVAMP	1,000	31.8	Number	MIT License
AQuA	254	51.9	Multi-choice	Apache-2.0
MultiArith	600	31.8	Number	CC BY SA-4.0

Table 4: Dataset Descriptions.

https://allenai.org/data/arc. CC BY SA-4.0 license: https://creativecommons.org/licenses/by-sa/4.0/.

- Sports understanding from BIG-Bench (bench collaboration., 2022): the answer format is Yes or No. Apache License v.2: https://github.com/google/BIG-bench/blob/main/LICENSE.

- BoolQ (Clark et al., 2019): it is a knowledge-intensive task and the format is Yes or No. The original dataset is from https://github.com/google-research-datasets/boolean-questions. CC BY SA-3.0 license: https://creativecommons.org/licenses/by-sa/3.0/.

Symbolic & Arithmetic reasoning

- Last Letters & Coin Flip (Wei et al., 2022) are novel benchmarks to evaluate whether the LLM can solve a simple symbolic reasoning problem. The last letters dataset is from https://huggingface.co/datasets/ChilleD/LastLetterConcat. The coin flip dataset is from https://huggingface.co/datasets/skrishna/coin_flip.

- GSM8K (Cobbe et al., 2021): https://github.com/openai/grade-school-math, MIT license: https://github.com/openai/grade-school-math/blob/master/LICENSE.

- SVAMP (Patel et al., 2021): https://github.com/arkilpatel/SVAMP, MIT license: https://github.com/arkilpatel/SVAMP/blob/main/LICENSE.

- AQuA (Ling et al., 2017): https://github.com/deepmind/AQuA, license: https://github.com/deepmind/AQuA/blob/master/LICENSE.

- Math Word Problem Repository MultiArith (Roy and Roth, 2015), license: CC BY 4.0, dataset: `https://huggingface.co/datasets/ChilleD/MultiArith`.

D Implementation Details

D.1 CoK Construction

For each dataset, we aim to construct demonstrations with multiple well-designed exemplars. The prompt example of each dataset is shown in Appendix F.

During the prompt construction, we first randomly select multiple labeled examples from the training set. For a fair comparison, we directly choose the selected labeled data from (Wei et al., 2022; Wang et al., 2023c; Kojima et al., 2022). Specifically, we choose 8 labeled data for Coin Flip, ARC-c, AQuA, GSM8K, MUltiArith, CSQA, SVAMP, OpenBookQA; 4 labeled data for Last Letter Connection; 6 labeled data for Sports, BoolQ, StrategyQA.

For each label data, we first use zero-shot CoT (Kojima et al., 2022) to perform textual reasoning chain generation. We directly connect a simple prompt “Let’s think step by step.” after the input query to elicit the LLM to generate rationale and the final answer [7]. We then remove this prompt and rebuild the input query by concatenating the input query and the generated textual reasoning chain.

To construct evidence triples, we aim to retrieve some relevant knowledge triples from the pre-built knowledge base (as shown in Appendix D.2). During retrieval, given a textual reasoning chain (e.g., CoK-EH), we encode it with the basic sentence encoder model (e.g., BERT) and then retrieve the most relevant knowledge triple using the maximum inner product search tool SCaNN (Guo et al., 2020). Due to the retrieved knowledge triples may consist of noises and redundant information. To improve the reliability of the evidence triples, we have invited five domain experts (including volunteer professors and Ph.D. students from diverse research areas) to manually annotate the evidence triples based on the retrieved knowledge triples.

To improve the annotation efficiency, we also employ the idea of self-training, which aims to generate annotated data based on very few data. Specifically, we can first manually annotate two labeled data with evidence triples and explanation hints to form a 2-shot CoK prompt. Then, each rest labeled data is concatenated with this 2-shot CoK prompt and the LLM can generate the corresponding rationale and answer. Thus, we can invite these experts to verify and correct them.

Finally, we obtain five different annotated demonstrations. To select the best one for each dataset, before the self-training process, we randomly choose some examples from the training set to form a validation set and use it to perform an evaluation. The prompt which has the best accuracy value is chosen.

D.2 Knowledge Bases

We follow (Pan et al., 2022) to select six knowledge bases from different domains.

Dictionary We utilize lexical knowledge, which comprises definitions and example sentences of English words (e.g., < *“apple”, definition, “A common, round fruit ...”* >) from Wiktionary [8]. Specifically, we use the Wiktionary dump dated April 30, 2022 that contains 1.3M word definitions and 470K example sentences for 1M words/phrases.

Commonsense We incorporate commonsense knowledge from ConceptNet (Speer et al., 2017), which encompasses a wide range of general daily life knowledge, for instance, < *“rabbit”, CapableOf, “move fast”* >. We follow previous works (Zhang et al., 2020) to include the core 600K high-quality triplets.

Causality The external knowledge we utilize for causality is derived from the auto-extracted causal knowledge source, CausalBank (Li et al., 2020). This database consists of large-scale English sentences that express cause-effect relationships, such as < *“babies cry”, because, “they are lonely”* >).

Entity We integrate named entity knowledge in Wikidata5M (Wang et al., 2021). For each entity, the properties are collected in the format of < *“Kepler space telescope”, Named after, “Johannes Kepler”* >). Here, related sentences refer to the sentences from an entity’s own article, or the sentences of other articles that link to this entity.

Event We take into account knowledge of everyday events, utilizing both human-constructed event

[7] To alleviate the generation bias problem, we also use the self-consistency decoding (Wang et al., 2023c) to sample one rationale.

[8] `https://en.wiktionary.org/wiki/Wiktionary:Main_Page`.

Input	Response
Vanilla: *Knowledge Injection before Reasoning*	
Q: Who plays mary jane in spiderman 3? \n Knowledge: (Kirsten Dunst, (actors, jk simmons), (award received, Milos Foreman)), (Kirsten Dunst, actors, Kirsten Dunst, occupation, female model)	The answer to the question is Kirsten Dunst. To explain our reasoning with the knowledge you provided:\n - Kirsten Dunst played Mary Jane in Spiderman 3. To confirm this information, we can look at actors who were in Spiderman 3, and we know that J.K. Simmons was also in the film. However, this information does not answer our question.\n - We can also look at awards received, and we know that Milos Forman received awards for a film he directed called "One Flew Over the Cuckoo's Nest" which Kirsten Dunst was not in.\n - We can look at Kirsten Dunst's occupation and know she is a female model, which does not help answer our question.\n - Therefore, we can conclude that Kirsten Dunst played Mary Jane in Spiderman 3.
Ours: *Knowledge Injection after F^2-Verification*	
Q: Who plays mary jane in spiderman 3? \n Knowledge: (spiderman 3, stars, kirsten dunst)	First, we need to recall which film we're talking about, which is Spiderman 3. \n Next, we need to think about the key players in the film, and who played Mary Jane in that particular movie. The actress who portrayed Mary Jane in Spiderman 3 is Kirsten Dunst.

Table 5: Performance of different knowledge injection.

knowledge graphs such as ATOMIC (Sap et al., 2019) and GLUCOSE (Mostafazadeh et al., 2020), as well as automatically extracted event knowledge graphs like ASER (Zhang et al., 2020, 2022a).

Scripts Specifically, we use 325K triples that are in the form of < verbal information, context, nonverbal information >, where verbal information is an utterance, nonverbal information can be body movements, vocal tones, or facial expressions, etc., and context is the entire text of the scene from which the verbal-nonverbal pair is extracted.

D.3 Ablation Settings

For the ablation study, we design six variants: 1) **CoK** is the basic prompt without the rethinking process; 2) **CoK w/o. ET** consists of only explanation hints (is the same as CoT); 3) **CoK w/o. EH** only has evidence triples; 4) **CoK + F^2-V** denotes the whole framework; 5) **CoK + F^2-V w/o. Factuality** aims to only use faithfulness score in the rethinking process; and 6) **CoK + F^2-V w/o. Faithfulness** aims to only use a factuality score.

E Analysis

E.1 Why is verification useful?

We demonstrate some cases to show how the LLM verifies the usefulness of the injected knowledge. We use *gpt-3.5-turbo* model because it can solve a problem through conversation. We define two following settings. 1) Vanilla: knowledge injection before reasoning. We directly concatenate the related knowledge triples with the input query, and prompt the LLM to think step by step. 2) Ours: knowledge injection after F^2-Verification. We first use a CoK prompt to elicit the LLM to generate evidence triples and find error triples. Then, we correct them into corresponding ground truth triples and concatenate them with the input query.

As shown in Table 5, we can see that the LLM can spontaneously detect and analyze each injected triple. For example, if the knowledge triple is useless or has no contribution to the reasoning, the LLM can talk to me about which and why are useless (the red text in Table 5). This indicates that the knowledge provided by the traditional knowledge injection method is not completely useful. In contrast, our approach can accurately locate the false reasoning evidence derived from the LLM after the verification stage, so as to perform targeted knowledge injection.

F Exemplars with Chain-of-Knowledge Prompts

The details of our prompts are shown below.

Q: Take the last letters of the words in "Elon Musk" and concatenate them. **A:** Evidence triples: 1. (Elon, last latter, n) 2. (Musk, last latter, k) 3. (final answer, is, nk) Explanation hints: The last letter of "Elon" is "n". The last letter of "Musk" is "k". Concatenating them is "nk". So the answer is nk.
Q: Take the last letters of the words in "Larry Page" and concatenate them. **A:** Evidence triples: 1. (Larry, last latter, y) 2. (Page, last latter, e) 3. (final answer, is, ye) Explanation hints: The last letter of "Larry" is "y". The last letter of "Page" is "e". Concatenating them is "ye". So the answer is ye.
Q: Take the last letters of the words in "Sergey Brin" and concatenate them. **A:** Evidence triples: 1. (Sergey, last latter, y) 2. (Brin, last latter, n) 3. (final answer, is, yn) Explanation hints: The last letter of "Sergey" is "y". The last letter of "Brin" is "n". Concatenating them is "yn". So the answer is yn.
Q: Take the last letters of the words in "Bill Gates" and concatenate them. **A:** Evidence triples: 1. (Bill, last latter, l) 2. (Gates, last latter, s) 3. (final answer, is, ls) Explanation hints: The last letter of "Bill" is "l". The last letter of "Gates" is "s". Concatenating them is "ls". So the answer is ls.

Table 6: Examples Chain-of-Knowledge prompting for the letter concatenation task.

Q: A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up? **A:** Evidence triples: 1. (coin, start with, head up) 2. (coin, flips, flipped) 3. (coin, not flips, flipped) Explanation hints:The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is yes.
Q: A coin is heads up. Jamey flips the coin. Teressa flips the coin. Is the coin still heads up? **A:** Evidence triples: 1. (coin, start, head up) 2. (coin, flips, flipped) 3. (coin, flips, head up) Explanation hints: The coin was flipped by Jamey and Teressa. So the coin was flipped 2 times, which is an even number. The coin started heads up, so after an even number of flips, it will still be heads up. So the answer is yes.
Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up? **A:** Evidence triples: 1. (coin, start with, head up) 2. (coin, flips, flipped) 3. (coin, not flips, flipped) Explanation hints: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.
Q: A coin is heads up. Millicent does not flip the coin. Conception flips the coin. Is the coin still heads up? **A:** Evidence triples: 1. (coin, start with, head up) 2. (coin, flips, flipped) 3. (coin, not flips, flipped) Explanation hints: The coin was flipped by Conception. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.
(The rest demonstration examples are shown in Table 8.)

Table 7: Examples Chain-of-Knowledge prompting for the coinflip task.

(The front demonstration examples are shown in Table 7.)
Q: A coin is heads up. Sal flips the coin. Raymond does not flip the coin. Is the coin still heads up? **A:** Evidence triples: 1. (coin, start with, head up) 2. (coin, flips, flipped) 3. (coin, not flips, flipped) Explanation hints: The coin was flipped by Sal. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.
Q: A coin is heads up. Conception flips the coin. Kristian does not flip the coin. Is the coin still heads up? **A:** Evidence triples: 1. (coin, start with, head up) 2. (coin, flips, flipped) 3. (coin, not flips, flipped) Explanation hints: The coin was flipped by Conception. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.
Q: A coin is heads up. Inga does not flip the coin. Elanor does not flip the coin. Is the coin still heads up? **A:** Evidence triples: 1. (coin, start with, head up) 2. (coin, not flips, head up) 3. (coin, not flips, head up) Explanation hints: The coin was flipped by no one. So the coin was flipped 0 times. The coin started heads up, and it was not flipped, so it is still heads up. So the answer is yes.

Table 8: (Following Table 7) Examples Chain-of-Knowledge prompting for the coinflip task.

Q: Is the following sentence plausible? ’Fernando Tatis Jr. walked on ball four.’ **A:** Evidence triples: 1. (Fernando Tatis Jr., isA, baseball player) 2. (walking on ball four, is part of, baseball) Explanation hints: A walk is when the batter reaches first base without hitting the ball. Reaching first base can happen a few ways, but one of them is by the pitcher throwing four balls. So the answer is yes.
Q: Is the following sentence plausible? ’Jayson Tatum was caught offsides in the Champions Leage Semifinal.’ **A:** Evidence triples: 1. (Jayson Tatum, isA, basketball player) 2. (being caught offsides, is part of, soccer) Explanation hints: Being caught offsides is part of soccer, yet, Jayson Tatum is a basketball player. So the answer is no.
Q: Is the following sentence plausible? ’Marouane Fellaini scored in extra time.’ **A:** Evidence triples: 1. (Marouane Fellaini, isA, soccer player) 2. (scoring in extra time, is part of, soccer) So, the answer is yes.
Q: Is the following sentence plausible? ’Gerrit Cole committed a handball in the European Cup.’ **A:** Evidence triples: 1. (Gerrit Cole, isA, soccer player) 2. (committing a handball, is part of, soccer) Explanation hints: Committing a handball is part of soccer, yet, Gerrit Cole is a soccer player. So the answer is no.
Q: Is the following sentence plausible? ’Klaas Jan Huntelaar scored the easy layup.’ **A:** Evidence triples: 1. (Klaas Jan Huntelaar, isA, soccer player) 2. (scoring a layup, is part of, basketball) Explanation hints: Scoring a layup is part of basketball, yet, Klaas Jan Huntelaar is a soccer player. So the answer is no.
Q: Is the following sentence plausible? ’Mario Gomez earned a direct kick.’ **A:** Evidence triples: 1. (Mario, isA, soccer player) 2. (earning a direct kick, is part of, soccer) Explanation hints: Earning a direct kick is part of soccer, and Mario Gomez is a soccer player. So the answer is yes.

Table 9: Examples Chain-of-Knowledge prompting for the sports understanding task.

Q: Putting a cardboard box in a bin instead of the trash can? Answer Choices: (A) conserve energy for later (B) save wild animal species (C) keep it from the trash (D) reduce the height of landfills **A:** Evidence triples: 1. (cardboard box, isA, item) 2. (bin, isA, container) 3. (cardboard box, usage, container) 4. (trash can, usage, container) 5. (bin, usage, container) Explanation hints: The answer must be an environmental reason. Of the above choices, only (D) are environmental reasons. So the answer is (D).
Q: When the needle of a compass lines up with Earth 's magnetic poles , the needle is? Answer Choices: (A) pointing east (B) pointing west (C) pointing at arctic (D) pointing to space **A:** Evidence triples: 1. (compass, is affected by, Earth's magnetic field) 2. (magnetic field, go, from the south pole to the north pole) Explanation hints: The needle of the compass is affected by Earth's magnetic field. The magnetic field lines go from the south pole to the north pole. So, when the needle of the compass is lined up with the Earth's magnetic poles, the needle is pointing north. So the answer is (C).
Q: Which uses gills to breathe? Answer Choices: (A) hermit crab (B) human (C) blue whale (D) bluebird **A:** Evidence triples: 1. (gills, isA, respiratory organ) 2. (respiratory organ, function, breathing) 3. (hermit crab, hasA, gills) 4. (blue whale, hasA, gills) 5. (bluebird, doesNotHave, gills) Explanation hints: The answer must be an item that has gills. Gills are respiratory organs that help with breathing. Of the above choices, only hermit crabs and blue whales have gills. So the answer is (A).
Q: Which is the most accurate statement? Answer Choices: (A) finches with fat beaks can eat ants in a small anthill (B) finches with narrow beaks can eat only human meat (C) finches with narrow beaks can eat ants in a small anthill (D) finches with fat beaks can eat only human meat **A:** Evidence triples: 1. (finches, isA, bird) 2. (bird, hasA, beak) 3. (finches, hasA, fat beak) 4. (finches, hasA, narrow beak) 5. (finches, eats, ants) 6. (finches, eats, human meat) Explanation hints: The answer must be a statement about what finches can eat. Of the above choices, only finches with narrow beaks can eat ants in a small anthill. So the answer is (C).

Table 10: Examples Chain-of-Knowledge prompting for the ARC-c task.

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today? Answer Choices: (A) 12 (B) 6 (C) 8 (D) 14 **A:** Evidence triples: 1. (grove, originally have, 15 trees) 2. (grove, now have, 21 trees) 3. (21 - 15, equal, 6) Explanation hints: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. So the answer is (B).
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot? Answer Choices: (A) 3 (B) 2 (C) 5 (D) 1 A: Evidence triples: 1. (parking lot, originally have, 3 cars) 2. (parking lot, add, 2 cars) 3. (parking lot, now have, 3 + 2 = 5 cars) Explanation hints: There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. So the answer is (C).
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total? Answer Choices: (A) 39 (B) 74 (C) 109 (D) 40 **A:** Evidence triples: 1. (Leah, have, 32 chocolates) 2. (her sister, have, 42 chocolates) 3. (they, have, 32 + 42 = 74 chocolates) 4. (they, eat, 35 chocolates) 5. (thay, now have, 74 - 35 = 39 chocolates) Explanation hints: Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. So the answer is (A).
Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny? Answer Choices: (A) 32 (B) 16 (C) 9 (D) 8 **A:** Evidence triples: 1. (Jason, originally have, 20 lollipops) 2. (Jason, now have, 12 lollipops) 3. (Jason, give, 20 - 12 = 8 lollipops) Explanation hints: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. So the answer is (D).

Table 12: Examples Chain-of-Knowledge prompting for the AQuA, GSM8K and MultiArith task.

Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now? Answer Choices: (A) 5 (B) 20 (C) 9 (D) 1 **A:** Evidence triples: 1. (Shawn, have, 5 toys) 2. (his mon, give him, 2 toys) 3. (his dad, give him, 2 toys) 4. (Shawn, now have, 5 + 2 + 2 = 9 toys) Explanation hints: Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9. So the answer is (C).
Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room? Answer Choices: (A) 20 (B) 29 (C) 11 (D) 18 **A:** Evidence triples: 1. (server room, originally have, 9 computers) 2. (each day, installed, 5 computers) 3. (each from monday to thursday, have, 4 days) 4. (5 * 4, equal, 20) 5. (server room, now have, 9 + 20 = 29) Explanation hints: There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 is 29. So the answer is (B).
Q: Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday? Answer Choices: (A) 33 (B) 35 (C) 81 (D) 83 **A:** Evidence triples: 1. (Michael, have, 58 golf balls) 2. (Michael, lost, 23 golf balls) 3. (Michael, lost, 2 golf balls) 4. (Michael, now have, 58 - 23 - 2 = 33 golf balls) Explanation hints: Michael started with 58 golf balls. After losing 23 on tuesday, he had 58 - 23 = 35. After losing 2 more, he had 35 - 2 = 33 golf balls. So the answer is (A).
Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left? Answer Choices: (A) 37 (B) 8 (C) 15 (D) 10 **A:** Evidence triples: 1. (Olivia, have, 23 dollars) 2. (Olivia, buy, 5 * 3 = 15 dollars) 3. (Olivia, now have, 23 - 15 = 8 dollars) Explanation hints: Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 x 3 = 15 dollars. So she has 23 - 15 dollars left. 23 - 15 is 8. So the answer is (B).

Table 13: Examples Chain-of-Knowledge prompting for the AQuA, GSM8K and MultiArith task.

Q: Do hamsters provide food for any animals? **A:** Evidence triples: 1. (Hamsters, isA, prey animals) 2. (Prey, is food for, predators) 3. (hamsters, provide food, animals) Explanation hints: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals. So the answer is yes.
Q: Could Brooke Shields succeed at University of Pennsylvania? **A:** Evidence triples: 1. (Brooke Shields, isA, student) 2. (student, could succeed, at University of Pennsylvania) 3. (Brooke Shields, could succeed, at University of Pennsylvania) Explanation hints: Brooke Shields is a student. Students could succeed at University of Pennsylvania. Thus, Brooke Shields could succeed at University of Pennsylvania. So the answer is yes.
Q: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls? **A:** Evidence triples: 1. (Hydrogen, has atomic number, 1) 2. (1, squared is, 1) 3. (1, exceeds, 5) 4. (5, is the number of, Spice Girls) Explanation hints: Hydrogen has an atomic number of 1. 1 squared is 1. There are 5 Spice Girls. Thus, Hydrogen's atomic number squared is less than 5. So the answer is no.
Q: Yes or no: Is it common to see frost during some college commencements? **A:** Evidence triples: 1. (Frost, isA, weather condition) 2. (Weather condition, is, common during some college commencements) 3. (Frost, is common, during some college commencements) Explanation hints: Frost is a weather condition. Weather conditions are common during some college commencements. Thus, frost is common during some college commencements. So the answer is yes.
Q: Yes or no: Could a llama birth twice during War in Vietnam (1945-46)? **A:** Evidence triples: 1. (Llama, isA, mammal) 2. (Mammal, gives birth, live young) 3. (Llama, could give birth, during War in Vietnam) Explanation hints: Llamas are mammals. Mammals give birth to live young. Therefore, it is possible for a llama to give birth during the War in Vietnam. So the answer is no.
Q: Yes or no: Would a pear sink in water? **A:** Evidence triples: 1. (pear, density, 0.6g/cm3) 2. (water, density, 1.0g/cm3) 3. (1.0g/cm3, is larger than, 0.6g/cm3) 4. (pear, can not sink in, water) Explanation hints: The density of a pear is about 0.6g/cm3, which is less than water. Objects less dense than water float. Thus, a pear would float. So the answer is no.

Table 14: Examples Chain-of-Knowledge prompting for the BoolQ task.

Q: What do people use to absorb extra ink from a fountain pen? Answer Choices: (A) shirt pocket (B) calligrapher's hand (C) inkwell (D) desk drawer (E) blotter **A:** Evidence triples: 1. (fountain pen, isA, item) 2. (ink, isA, liquid) 3. (fountain pen, carrier of, liquid) 4. (fountain pen, usage, writing) 5. (blotters, usage, writing) 6. (blotters, absorb, liquid) Explanation hints: The answer must be an item that can absorb ink. A fountain pen which is full of liquid can writing on a blotter. Of the above choices, only blotters are used to absorb ink. So the answer is (E).
Q: What home entertainment equipment requires cable? Answer Choices: (A) radio shack (B) substation (C) television (D) cabinet **A:** Evidence triples: 1. (home entertainment equipment, isA, item) 2. (cable, isA, item) 3. (home entertainment equipment, requires, cable) 4. (television, isA, home entertainment equipment) 5. (television, requires, cable) Explanation hints: The answer must be an item of home entertainment equipment that requires cable. Of the above choices, only television requires cable. So the answer is (C).
Q: The fox walked from the city into the forest, what was it looking for? Answer Choices: (A) pretty flowers (B) hen house (C) natural habitat (D) storybook **A:** Evidence triples: 1. (fox, isA, animal) 2. (forest, place of residence, animal) 3. (natural habitat, located in, forest) Explanation hints: The answer must be something in the forest. Of the above choices, only natural habitat is in the forest which is the living place for fox animal. So the answer is (C).
(The rest demonstration examples are shown in Table 16.)

Table 15: Examples of Chain-of-Knowledge prompting for the CSQA task.

(The front demonstration examples are shown in Table 15.)
Q: Sammy wanted to go to where the people were. Where might he go? Answer Choices: (A) populated areas (B) race track (C) desert (D) apartment (E) roadblock **A:** Evidence triples: 1. (Sammy, isA, person) 2. (populated areas, place of residence, people) Explanation hints: Sammy is a person, so that the answer must be a place with a lot of people. Of the above choices, only populated areas have a lot of people. So the answer is (A).
Q: Where do you put your grapes just before checking out? Answer Choices: (A) mouth (B) grocery cart (C) super market (D) fruit basket (E) fruit market **A:** Evidence triples: 1. (checking out, isA, action) 2. (checking out, place of take place, mall) 3. (grapes, isA, merchandise) 4. (mall, sell, fruit) 5. (grocery cart, usage, hold merchandise before checking out) Explanation hints: The answer should be the place where grocery items are placed before checking out. Of the above choices, grocery cart makes the most sense for holding grocery items. So the answer is (B).
Q: Google Maps and other highway and street GPS services have replaced what? Answer Choices: (A) united states (B) mexico (C) countryside (D) atlas **A:** Evidence triples: 1. (Google Maps, isA, webapp) 2. (GPS, isA, navigation systems) 3. (atlas, usage, navigation) Explanation hints: The answer must be something that used to do what Google Maps and GPS services do, which is to give directions. Of the above choices, only atlases are used to give directions. So the answer is (D).

Table 16: (Following Table 15) Examples of Chain-of-Knowledge prompting for the CSQA task.

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today? **A:** Explanation hints: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. Evidence triples: 1. (grove, originally have, 15 trees) 2. (grove, now have, 21 trees) 3. (21 - 15, equal, 6) So the answer is 6.
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot? **A:** Explanation hints: There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. Evidence triples: 1. (parking lot, originally have, 3 cars) 2. (parking lot, add, 2 cars) 3. (parking lot, now have, 3 + 2 = 5 cars) So the answer is 5.
Q: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total? **A:** Explanation hints: Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. Evidence triples: 1. (Leah, have, 32 chocolates) 2. (her sister, have, 42 chocolates) 3. (they, have, 32 + 42 = 74 chocolates) 4. (they, eat, 35 chocolates) 5. (thay, now have, 74 - 35 = 39 chocolates) So the answer is 39.
(The rest demonstration examples are shown in Table 18.)

Table 17: Examples Chain-of-Knowledge prompting for the SVAMP task.

(The front demonstration examples are shown in Table 17.)
Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny? **A:** Explanation hints: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. v Evidence triples: 1. (Jason, originally have, 20 lollipops) 2. (Jason, now have, 12 lollipops) 3. (Jason, give, 20 - 12 = 8 lollipops) So the answer is 8.
Q: Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now? **A:** Explanation hints: Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9. Evidence triples: 1. (Shawn, have, 5 toys) 2. (his mon, give him, 2 toys) 3. (his dad, give him, 2 toys) 4. (Shawn, now have, 5 + 2 + 2 = 9 toys) So the answer is 9.
Q: There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room? **A:** Explanation hints: There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 is 29. Evidence triples: 1. (server room, originally have, 9 computers) 2. (each day, installed, 5 computers) 3. (each from monday to thursday, have, 4 days) 4. (5 * 4, equal, 20) 5. (server room, now have, 9 + 20 = 29) So the answer is 29.

Table 18: (Follow by Table 17) Examples Chain-of-Knowledge prompting for the SVAMP task.

Q: Putting a cardboard box in a bin instead of the trash can?
Answer Choices: (A) conserve energy for later (B) save wild animal species (C) keep it from the trash (D) reduce the height of landfills
A: Evidence triples:
1. (cardboard box, isA, item)
2. (bin, isA, container)
3. (cardboard box, usage, container)
4. (trash can, usage, container)
5. (bin, usage, container)
Explanation hints: The answer must be an environmental reason. Of the above choices, only (D) are environmental reasons.
So the answer is (D).

Q: When the needle of a compass lines up with Earth 's magnetic poles , the needle is?
Answer Choices: (A) pointing east (B) pointing west (C) pointing at arctic (D) pointing to space
A: Evidence triples:
1. (compass, is affected by, Earth's magnetic field)
2. (magnetic field, go, from the south pole to the north pole)
Explanation hints: The needle of the compass is affected by Earth's magnetic field. The magnetic field lines go from the south pole to the north pole. So, when the needle of the compass is lined up with the Earth's magnetic poles, the needle is pointing north.
So the answer is (C).

Q: Which uses gills to breathe?
Answer Choices: (A) hermit crab (B) human (C) blue whale (D) bluebird
A: Evidence triples:
1. (gills, isA, respiratory organ)
2. (respiratory organ, function, breathing)
3. (hermit crab, hasA, gills)
4. (blue whale, hasA, gills)
5. (bluebird, doesNotHave, gills)
Explanation hints: The answer must be an item that has gills. Gills are respiratory organs that help with breathing. Of the above choices, only hermit crabs and blue whales have gills.
So the answer is (A).

Q: Which is the most accurate statement?
Answer Choices: (A) finches with fat beaks can eat ants in a small anthill (B) finches with narrow beaks can eat only human meat (C) finches with narrow beaks can eat ants in a small anthill (D) finches with fat beaks can eat only human meat
A: Evidence triples:
1. (finches, isA, bird)
2. (bird, hasA, beak)
3. (finches, hasA, fat beak)
4. (finches, hasA, narrow beak)
5. (finches, eats, ants)
6. (finches, eats, human meat)
Explanation hints: The answer must be a statement about what finches can eat. Of the above choices, only finches with narrow beaks can eat ants in a small anthill.
So the answer is (C).

(The rest demonstration examples are shown in Table 20).

Table 19: Examples of Chain-of-Knowledge prompting for the OpenbookQA task.

(Other demonstration examples are shown in Table 19.)

Q: What type of useful product can be made from the moving winds?
Answer Choices: (A) metal (B) wood (C) bananas (D) electricity
A: Evidence triples:
1. (wind, isA, moving air)
2. (electricity, isA, moving electrons)
3. (wind, can generate, electricity)
Explanation hints: The answer must be a type of energy. Of the above choices, only electricity can be made from the moving winds.
So the answer is (D).

Q: Harriet wants to know the area of a rectangular sheet of paper. Its size is unknown to her. Which one of these will be the only tool that she needs?
Answer Choices: (A) a ruler (B) a compass (C) a calculator (D) a protractor
A: Evidence triples:
1. (area, isA, measurement)
2. (rectangular sheet of paper, isA, two-dimensional object)
3. (area, measurement of, two-dimensional object)
4. (ruler, usage, measuring)
5. (ruler, usage, measuring length)
6. (ruler, usage, measuring width)
7. (area, calculation of, length x width)
Explanation hints: The answer must be an item that can help Harriet find the area of the rectangular sheet of paper. Of the above choices, only a ruler can help Harriet find the area of the rectangular sheet of paper.
So the answer is (A).

Q: What uses eyes to see by sensing light?
Answer Choices: (A) trees (B) algae (C) worms (D) elephants
A: Evidence triples:
1. (eyes, isA, organ)
2. (eyes, usage, seeing)
3. (eyes, sensing, light)
4. (trees, usage, photosynthesis)
5. (algae, usage, photosynthesis)
6. (worms, usage, burrowing)
7. (elephants, usage, trampling)
Explanation hints: The answer must be an entity that uses eyes to see by sensing light. Of the above choices, only trees, algae, and worms use eyes to see by sensing light.
So the answer is (D).

Q: The cell of a plant when photosynthesizing may most require?
Answer Choices: (A) access to roofless space (B) a complete vacuum in space (C) a room with blackened windows (D) dark basement living quarters
A: Evidence triples:
1. (cell, isA, plant)
2. (photosynthesis, isA, process)
3. (cell, requires, photosynthesis)
4. (photosynthesis, requires, light)
5. (dark basement living quarters, blocks, light)
Explanation hints: The cell of a plant when photosynthesizing may most require access to sunlight. Of the above choices, only (A) provides access to sunlight.
So the answer is (A).

Table 20: (Follow by Table 19) Examples of Chain-of-Knowledge prompting for the OpenbookQA task.

Q: Do hamsters provide food for any animals? **A:** Evidence triples: 1. (Hamsters, isA, prey animals) 2. (Prey, is food for, predators) 3. (hamsters, provide food, animals) Explanation hints: Hamsters are prey animals. Prey are food for predators. Thus, hamsters provide food for some animals. So the answer is yes.
Q: Could Brooke Shields succeed at University of Pennsylvania? **A:** Evidence triples: 1. (Brooke Shields, isA, student) 2. (student, could succeed, at University of Pennsylvania) 3. (Brooke Shields, could succeed, at University of Pennsylvania) Explanation hints: Brooke Shields is a student. Students could succeed at University of Pennsylvania. Thus, Brooke Shields could succeed at University of Pennsylvania. So the answer is yes.
Q: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls? **A:** Evidence triples: 1. (Hydrogen, has atomic number, 1) 2. (1, squared is, 1) 3. (1, exceeds, 5) 4. (5, is the number of, Spice Girls) Explanation hints: Hydrogen has an atomic number of 1. 1 squared is 1. There are 5 Spice Girls. Thus, Hydrogen's atomic number squared is less than 5. So the answer is no.
Q: Yes or no: Is it common to see frost during some college commencements? **A:** Evidence triples: 1. (Frost, isA, weather condition) 2. (Weather condition, is, common during some college commencements) 3. (Frost, is common, during some college commencements) Explanation hints: Frost is a weather condition. Weather conditions are common during some college commencements. Thus, frost is common during some college commencements. So the answer is yes.
Q: Yes or no: Could a llama birth twice during War in Vietnam (1945-46)? **A:** Evidence triples: 1. (Llama, isA, mammal) 2. (Mammal, gives birth, live young) 3. (Llama, could give birth, during War in Vietnam) Explanation hints: Llamas are mammals. Mammals give birth to live young. Therefore, it is possible for a llama to give birth during the War in Vietnam. So the answer is no.
Q: Yes or no: Would a pear sink in water? **A:** Evidence triples: 1. (pear, density, 0.6g/cm3) 2. (water, density, 1.0g/cm3) 3. (1.0g/cm3, is larger than, 0.6g/cm3) 4. (pear, can not sink in, water) Explanation hints: The density of a pear is about 0.6g/cm3, which is less than water. Objects less dense than water float. Thus, a pear would float. So the answer is no.

Table 21: Examples of Chain-of-Knowledge prompting for the StrategyQA task.