Enhancing Chain-of-Thoughts Prompting with Iterative Bootstrapping in Large Language Models

Jiashuo Sun¹* Yi Luo¹* Yeyun Gong² Chen Lin¹† Yelong Shen³ Jian Guo⁴ Nan Duan²
¹School of Informatics, Xiamen University ²Microsoft Research Asia ³Microsoft ⁴IDEA Research

Abstract

Large language models (LLMs) can achieve impressive performance on various reasoning tasks by incorporating chain-of-thought (CoT) prompting, where step-by-step reasoning is provided to guide LLMs to generate answers to questions, and the question-rationale-answer triplets are utilized as demonstration exemplars. However, the reasoning chains of demonstrations generated by LLMs are observed to be prone to errors, which can subsequently lead to incorrect reasoning during inference. Furthermore, inappropriate exemplars, e.g., overly simplistic or complex exemplars depending on the question’s difficulty level, can affect the LLM’s performance. To address these issues, we introduce Iter-CoT (Iterative bootstrapping in Chain-of-Thoughts prompting). Iter-CoT has two advantages: (1) it adopts iterative bootstrapping that enables LLMs to rectify errors autonomously, resulting in more precise and comprehensive reasoning chains. (2) it selects exemplars of challenging yet answerable (i.e., the LLM has the potential to answer correctly) questions, enhancing the LLMs’ generalizability to answer questions with varying difficulty levels. Experimental results exhibit Iter-CoT superior performance on three distinct reasoning tasks on ten datasets. Our code is publicly available at https://github.com/XMUDM/Iter-CoT.

1 Introduction

Chain-of-Thought (CoT) (Wei et al., 2022) prompting is a technique to enhance the reasoning abilities of large language models (LLMs) by generating a series of reasoning steps to obtain the answer, and the reasoning chains are utilized as exemplars to demonstrate the task and provide In-Context Learning (ICL) (Brown et al., 2020) to LLMs. Recently, LLMs (Chowdhery et al., 2022; Thoppilan et al., 2022; Rae et al., 2021; Smith et al., 2022; Scao et al., 2022) with CoT prompting have demonstrated remarkable performance in complex reasoning tasks, including arithmetic (Cobbe et al., 2021; Ling et al., 2017; Hosseini et al., 2014; Koncel-Kedziorski et al., 2015; Patel et al., 2021; Miao et al., 2020), commonsense (Talmor et al., 2019; Geva et al., 2021; Kojima et al., 2022; Wei et al., 2022), and symbolic reasoning (Wei et al., 2022).

Existing studies on CoT prompting can be classified into two categories. The first category is manually constructed CoT prompting((Wei et al.,...
where human annotators manually craft a collection of question-rationale-answer exemplars to guide the model’s reasoning process. However, human annotations’ inherent subjectivity and limitations make these approaches costly, sub-optimal, and highly inconsistent. The second category is automatically generated CoT prompting((Kojima et al., 2022; Zhang et al., 2022; Shum et al., 2023; Shao et al., 2023)), where LLM-generated CoT is utilized. In practice, reasoning chains generated by LLMs have demonstrated superior performance compared with human annotations.

However, three issues remain under-explored in the literature. (1) **Difficulty of questions.** It is regarded that questions of mediate difficulty level can best guide the LLMs (Diao et al., 2023). As shown in Figure 1, it is observed that simplistic examples (Simple-CoT) perform poorly in solving complex questions of more hops\(^1\), while excessively complex exemplars (Complex-CoT (Fu et al., 2022)) are unsatisfying on simpler questions. (2) **Correctness of demonstration.** Reasoning chains of demonstrations generated by LLMs are prone to errors (Zhang et al., 2022; Diao et al., 2023), which can significantly reduce overall performance. As shown in Figure 2, accuracy on various datasets decreases as incorrect exemplars increase. (3) **Missing contextual information.** Previous works merely combine the question and the “let’s think step by step” prompt (Kojima et al., 2022) during the generation of demonstrations without incorporating contextual information (such as incorrect reasoning chains and feedback answers generated by LLMs). Missing contextual information limits the LLM’s capability to learn from previous reasoning errors and avoid making similar errors.

In order to address the issues above, we propose Iter-CoT (Iterative bootstrapping in Chain-of-Thoughts Prompting). Iter-CoT allows LLMs to self-correct and summarize the more precise and comprehensive reasoning chains, which identify challenging yet answerable (i.e., LLM has the potential to answer correctly) questions as demonstrations in order to enhance the LLMs’ generalizability to answer questions with varying difficulty levels. We evaluate Iter-CoT on three distinct reasoning tasks (arithmetic, commonsense, and symbolic) across ten datasets. The experimental results show that Iter-CoT significantly outperforms existing prompting approaches.

Our contributions are summarized as follows: (1) We propose a new paradigm for CoT, which generates precise and comprehensive reasoning chains during interaction with LLMs. To our knowledge, our work is the first to illustrate the importance of iterative interaction with the LLMs to generate high-quality demonstrations. (2) We propose Iter-CoT, an approach that generates self-corrected and summarized reasoning chains on exemplars with intermediate difficulty levels, which are utilized as demonstrations to enhance the LLMs’ performance. (3) We implement Iter-CoT under both labeled and unlabeled conditions, achieving state-of-the-art (SOTA) results in both scenarios across ten datasets within three distinct tasks.

![Figure 3: Effect of re-answering the question based on the hint and previous rationales.](image)

---

\(^1\) Following (Shum et al., 2023; Fu et al., 2022), we measure the question’s difficulty by the number of hops in the rationale, with fewer hops indicating simpler questions and more hops indicating more complex questions.
in training set, we first apply the zero-shot-CoT to generate answers. For questions that are answered incorrectly, we prompt the LLM with a hint "the answer is incorrect" to re-answer. The process is repeated for six iterations until the number of correctly answered questions no longer increases.

As shown in Figure 3, the performance of the LLMs is improved (i.e., increasing from 4089 (54.7%) to 4898 (59.1%) after the first iteration, and the improvement sustains in subsequent iterations, ultimately reaching a peak of 5726 (76.6%). This observation suggests that LLMs can autonomously rectify errors with hints and contextual information.

### 2.2 The Value of Revised Examples

Examples containing erroneous rationales were ignored or screened out to prevent their adverse effects in previous studies (Zhang et al., 2022; Shum et al., 2023). However, inspired by the idea that students can improve their problem-solving abilities by learning from a collection of mistakes, we believe that allowing the model to learn from examples that have been answered incorrectly and then corrected can also effectively enhance the model’s inference performance. Figure 4 is a case study of the value of revised examples on the Date Understanding (Wei et al., 2022) dataset. It shows the different effects of two distinct demonstrations on the same test question. Using a randomly sampled exemplar as a demonstration is not beneficial (right side), even though it is already correctly answered. On the contrary, using the revised example’s reasoning chain (left side), where the reasoning chain is self-corrected by the LLM with the approach in Section 2.1, improves LLM’s reasoning ability.

### 3 Iter-CoT: Iterative Bootstrapping in Chain-of-Thought Prompting

Motivated by the observations in Section 2, we propose Iter-CoT (Iterative bootstrapping in Chain-of-Thoughts prompting), which generates the chain-of-thought demonstrations by guiding the LLM to rectify errors and summarize the reasoning chains on questions with appropriate difficulty levels. Following that, we put these exemplars into the demonstration pool. During the inference, we sampling and fixed the exemplars as the demonstration.

As shown in Figure 5, Iter-CoT consists of two stages, the construction stage of the demonstration pool and the inference stage. Moreover, the construction of the demonstration pool consists of three phases:

- **Initialization** The Zero-Shot-CoT (Kojima et al., 2022) method is employed on the training set to prompt the LLM to generate reasoning chains and answers. Error examples are recorded for the subsequent phases.

- **Bootstrapping** For each erroneous example, the Revise-Prompt ("Your answer is not right; can you think more carefully and give me the final answer?") is utilized to guide the LLM to self-correct until the correct answer is generated. In the absence of a prompted reference answer, the correct answer often corresponds to the correct reasoning chains.

- **Summarization** Once the correct answers are obtained, the Summary-Prompt ("Can you give me a complete solution reasoning process and final answer again?") is employed to guide the LLM in reviewing the previous rationales and summarizing the final reasoning chains. We reconfirm the cor-
First we need to find out how many kids from Riverside High were denied entry… 24 kids + 63 kids + 25 kids = ... = \(260 - 112 = 148\) kids. Final answer: 148.

### Inference

Correct CoT

Assistant: After thinking more carefully, I realized that I was not solving for the number of kids who got into the movie, but rather… 260 kids – 112 kids = 148 kids. The answer is: 148.

- **Zero-shot-CoT**
  - User: Q: Fern is checking IDs to get into an R-rated movie… How many kids got into the movie?
  - Assistant: Let’s think step by step.

- **Wrong CoT**
  - Assistant: First we need to find out how many kids from Riverside High were denied entry… 24 kids + 63 kids + 25 kids = ... = \(260 - 112 = 148\) kids. The answer is: 112.

- **Correct CoT**
  - Assistant: After thinking more carefully, I realized that I was not solving for the number of kids who got into the movie, but rather… 260 kids – 112 kids = 148 kids. The answer is: 148.

### Sampling N-Exemplars

**Demonstration**

Q: Fern is checking IDs to get into an R-rated movie… How many kids got into the movie?

A: Reasoning process:
1. We know that 20% of the 120 kids from…
2. We also know that 70% of the 90 kids…
   

**Fixed-Demonstrations**

Q1: Fern is checking IDs to get into an R-rated movie… How many kids got into the movie?

A: Reasoning process: Step 1: Calculate the... Qn: ...

- **Final CoT**
  - Assistant: Reasoning Process:
    1. We know that 20% of the 120 kids from...
    2. We also know that 70% of the 90 kids from...

### 4 Experiment

#### 4.1 Datasets and Evaluation Metrics

We evaluate our methods on ten datasets across three categories of different reasoning tasks, including (1) six arithmetic reasoning datasets: GSM8k (Cobbe et al., 2021), AQuA (Ling et al., 2017), AddSub (Hosseini et al., 2014), SingleEq (Koncel-Kedziorski et al., 2015), SVAMP (Patel et al., 2021) and ASDiv (Miao et al., 2020); (2) three commonsense reasoning datasets: CSQA (Talmor et al., 2019), StrategyQA (Geva et al., 2021) and Date Understanding (Wei et al., 2022); (3) one symbolic reasoning datasets: Letter Concatenation (Wei et al., 2022). Examples of each reasoning task and a detailed description of each dataset are shown in Table 7 and Table 8.

(rectness of the answer and retain only the correct ones. This process enables the LLM to capture rich contextual information from multi-turn conversations, resulting in more precise and comprehensive reasoning chains.

Upon completion of the aforementioned process, the final generated reasoning chain is combined with the corresponding question and added to the demonstration pool until the sample size fulfills the requirements. The requirement is flexible, yet at least satisfies differences in various datasets shown in Table 7.

Our approach works in both label-available and non-available scenarios. We use a rule-based approach to determine the correctness of the answers when labels are available in the construction stage of the demonstration pool. In contrast, when labels are unavailable, we use a more powerful model (e.g., GPT-4 (OpenAI, 2023)) as an evaluator to assess the correctness of the answers.

**Inference**: During the inference stage, a random sampling approach (Iter-CoT can also use other sampling methods, which are shown in Session 4.8.) is used to select \( N \) exemplars from the demonstration pool, which served as fixed demonstrations for the entire test set.

4077
In the inference stage, we report the exact match accuracy as our evaluation metric following previous works (Wei et al., 2022; Kojima et al., 2022).

### 4.2 Baselines

We compare our methods with five baseline approaches: **Manual-CoT** (Wei et al., 2022), **Random-CoT**, **Complex-CoT** (Fu et al., 2022), **Auto-CoT** (Zhang et al., 2022) and **Self-Consistency (SC)** (Wang et al., 2022). Manual-CoT involves using manually constructed reasoning chains as exemplars, listed in the appendix of Wei et al. (2022). Random-CoT randomly selects \( n \) questions from the training set and generates chains using the "let’s think step by step" prompt. Complex-CoT selects most complex exemplars, such as exemplars with most complex rationales or longest questions from the training set, as demonstrations. Auto-CoT utilizes clustering techniques to sample questions and generate chains with the same approach. Specifically, we implement Auto-CoT by generating reasoning chains for the questions provided in their appendix as demonstrations. Self-Consistency generates multiple answers for a question and uses a majority voting mechanism to select the final answer.

### 4.3 Implementation Details

We implement Iter-CoT on GPT-3.5-turbo (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023), using the OpenAI API\(^2\). We implement Iter-CoT on open-source models using 8 A100-40Gs for inference using Llama-2-70B-Chat and Llama-2-70B (Touvron et al., 2023) without quantization in our experiments. During the construction stage of the demonstration pool, we utilize a temperature setting of 0.7, whereas during the inference stage, we fix the temperature to 0 for reproducibility. Moreover, we set temperature = 0.7 and \( n = 40 \) for evaluation under self-consistency as (Wang et al., 2022). We adopt the number of exemplars for each dataset based on the experimental configuration of prior work (Zhang et al., 2022; Fu et al., 2022). Specifically, for datasets lacking test sets and without comparable datasets for transfer (e.g., Date Understanding), we randomly select a small portion as the training set and reserve the remaining portion for evaluation as the test set. In addition, We conducted three trials and averages for each experiment requiring random sampling to obtain final results. The size of each dataset and the partitioning of train and test sets are shown in Table 7.

### 4.4 Main Results

As Iter-CoT can be applied with and without groundtruth labels, we implement two versions: Iter-CoT(w/ label) and Iter-CoT(w/o label). The latter is implemented with GPT-4 as the evaluator. The experimental results are presented in Table 1.

\( ^2\)https://platform.openai.com/
We have the following observations:

**Iter-CoT achieves superior performance on different tasks.** Without using annotations/labels, Iter-CoT achieves superior performance on different tasks, and its performance is comparable or even superior than methods with annotations/labels. When examining the results on the first five arithmetic reasoning tasks in Table 1, we observe that Iter-CoT(w/o label) outperforms Complex-CoT with its average scores surpassing those of Complex-CoT 0.4%. When labels are available, Iter-CoT can achieve the best average score (81.5%) on all ten datasets across three tasks with GPT-3.5-turbo, surpassing Manual-CoT by 3.8% and Random-CoT by 6.1%. Notably, on the Letter Concatenation dataset, Iter-CoT(w/ label) achieved remarkable improvements of 7% compared to the previous highest scores. On the first five arithmetic reasoning tasks in Table 1, Iter-CoT (w/ labels) continues to exhibit the best performance, at 83.8%, surpassing Complex-CoT, where the annotations of reasoning chains in demonstrations are needed, by 1.6%. In conclusion, our approach outperforms existing approaches and achieves state-of-the-art results across various tasks.

**Iter-CoT(w/o label) share comparable performance with Iter-CoT(w/ label).** Iter-CoT(w/o label) demonstrates a marginal superiority over Iter-CoT(w/ label) on the Singleeq and SVAMP datasets, with improvements of 0.3% and 0.7%, respectively. However, it registers slightly lower performance compared to Iter-CoT(w/ label) on all other datasets, resulting in the average score that is 0.6% lower than that of Iter-CoT(w/ label). The performance gap between Iter-CoT(w/o label) and Iter-CoT(w/ label) can be attributed to the inherent challenge of using GPT-4 for evaluating the correctness of responses. The errors generated during the evaluation would cause the selected demonstrations to be answered incorrectly initially or not with the correct reasoning chains. Nonetheless, the impact of these errors on the overall results remains acceptable. Statistically, GPT-4 demonstrates an 87.5% accuracy in determining the correctness of responses during the demonstration pool construction stage. Consequently, the proportion of non-compliant samples in the final selected demonstrations remains acceptable. Furthermore, the incorrectly evaluated demonstrations tend to be challenging, thereby offering valuable insights to LLMs.

**Self-consistency (SC) consistently augments the efficacy of all methodologies.** Notably, on the GSM8K and AQuA datasets, SC significantly improves model inference performance, resulting in respective enhancements of 5.9%, 7.2%, and 8.2% for the Manual-CoT, Iter-CoT(w/o label), and Iter-CoT(w/ label) methods on GSM8K, and 5.1%, 10.5%, and 10.2% on AQuA. On other datasets, SC has also demonstrated consistent improvements. Ultimately, across the ten datasets, the three methods exhibit average performance enhancements of 2.7%, 3.1%, and 3.4%, respectively. Moreover, With the inclusion of SC, Iter-CoT(w/ label) and Iter-CoT(w/o label) continue to exhibit consistent superiority over Manual-CoT, with an average performance of 80.9% and 71.3%, respectively.
4.5 Performance on Different Foundation Models

To validate the feasibility of our approach across various diverse models, we conduct experiments on GPT-4 and two open source models: Llama-2-70B and Llama-2-70B-Chat (Touvron et al., 2023), as shown in Table 2. When comparing the results of different methods within three distinct foundation models, we observe that our approach consistently outperforms Manual-CoT across varying models. Specifically, on the Llama-2-70B-Chat, Llama-2-70B, and GPT-4 models, the average improvement of Iter-CoT(w/ label) over Manual-CoT is 14.9%, 7.6%, and 3.6%, respectively. Furthermore, Iter-CoT(w/o label) exhibits performance closely aligned with Iter-CoT(w/ label) across diverse models, with an average score difference of merely 2.1%, 1%, and 2.5% within the three models.

4.6 Ablation Studies

During the construction stage of the demonstration pool, both bootstrapping and summarization phases play crucial roles in generating the final exemplars. We conduct a series of ablation experiments to investigate the impact of these two phases on the results. Explanations for all the methods employed in this section of ablation experiments can be found in the caption of Table 3.

In addition, for Iter-CoT(w/o label), we also investigate the accuracy of LLM evaluators and their impact on the results.

4.6.1 Impact of Bootstrapping and Summarization phase

Comparison of Methods with and without a Bootstrapping Phase  To investigate the impact of bootstrapping phase on model’s performance, we contrast Correct-CoT, the method only uses the bootstrapping process, with many methods that do not modify the reasoning chains when generating demonstrations. These methods include Init-Wrong-CoT, Random-CoT, Init-Correct-CoT and Best-of-N-CoT. The outcomes are presented in Table 3. The presence of incorrect examples exerts a deleterious influence on model reasoning, as evidenced by the performance decline observed in Init-Wrong-CoT (68.4%) and Random-CoT (70.1%). Moreover, a direct comparison between Init-Wrong-CoT, Best-of-N-CoT and Correct-CoT, wherein the examples share identical questions, accentuates the efficacy of the bootstrapping phase. The comparison between Init-Correct-CoT and Correct-CoT also illustrates that the bootstrapping phase facilitates the selection of questions with appropriate difficulty levels.

4.7 The Effect of Iterations in the Bootstrapping Phase

We posit that questions requiring multiple times of bootstrapping are more challenging. Consequently, we conduct experiments for Iter-CoT on three datasets with iterative bootstrapping, where the iterations are the number of times the bootstrapping phase is invoked. The results are depicted in Figure 7.

We observed that the performance of Iter-CoT(w/ label) tends to initially rise and then fall as the number of bootstrapping steps increases. In contrast, Iter-CoT(w/o label) shows a steady im-
Figure 7: The Iter-CoT’s overall performance with iterative bootstrapping on three datasets.

The Iter-CoT’s overall performance with iterative bootstrapping on three datasets. Improvement (except for GSM8K, which exhibits an initial increase followed by a decrease, likely due to the challenge and difficulty of the GSM8K compared to the other two datasets.). However, even as accuracy decreases with increasing iterations, the post-decline accuracy still outperforms most baselines.

In Table 1, for GSM8K, CSQA, Letter Concatenation and other datasets sharing the same exemplars with GSM8K (AddSub, SingleEq, SVAMP, and ASDiv), we utilize the best exemplars in this section. In future work, further iterations can be explored to generate exemplars of higher quality for other datasets.

Impact of Summarization phase To investigate the impact of the summarization phase on the model’s inference capability, similar to the previous section, we compare Iter-CoT with Correct-CoT, which excludes the utilization of the summarization phase. Table 3 demonstrates that the former outperforms the latter by 2.8%. We attribute this performance difference to the role of the summarization phase, which encourages the model to incorporate extensive contextual information, thereby facilitating the generation of more intricate and comprehensive reasoning chains.

4.8 Sampling Strategies

After establishing the demonstration pool, we can employ various sampling methods to select examples for inference. We utilized three sampling techniques: random sampling, similarity-based sampling (retrieves the most similar examples according to BM25 similarity), and complexity-based sampling (selects the examples with the most complex reasoning steps). While other sampling methods are also applicable, they are not the subject of discussion here. The performance is shown in Table 4.

The results show that employing effective sampling methods can further enhance performance. However, introducing complexity-based sampling incurs additional overhead (as it requires manually annotated reasoning chains), thus in our main experiment, we report results based on random sampling, which represents the simplest sampling method, requiring no additional expenditure.

4.8.1 Impact of LLM evaluators’ Accuracy in Iter-CoT

In Section 4.4, we mentioned the potential errors when using GPT-4 as an evaluator, which could impact the results. To investigate the influence of evaluators’ accuracy on model inference capability in Iter-CoT(w/o label), we select three distinct LLMs and employ them as both the foundation model and evaluator. For each experiment, we compute the evaluator’s accuracy along with the final inference performance of the foundation model.

The experimental results are presented in Figure 6. All three subplots exhibit a common trend: the performance of the foundation model improves as the evaluator’s accuracy increases. We attribute this performance difference to the role of the summarization phase, which encourages the model to incorporate extensive contextual information, thereby facilitating the generation of more intricate and comprehensive reasoning chains.

<table>
<thead>
<tr>
<th>Method</th>
<th>GSM8K</th>
<th>Date</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>80.9</td>
<td>78.6</td>
<td>79.8</td>
</tr>
<tr>
<td>Similarity</td>
<td>79.8</td>
<td>79.1</td>
<td>79.5</td>
</tr>
<tr>
<td>Complexity</td>
<td>81.3</td>
<td>79.7</td>
<td>80.5</td>
</tr>
</tbody>
</table>

Table 4: Performance of Iter-CoT(w/ label) utilizing various sampling methods.
other methods, and the length of generated reasoning chains are not included in this section. Details of these experiments and analysis are discussed in Section B.1, A.2, A.3, A.4, A.5 and A.6.

5 Related Work

5.1 Chain-of-thought Prompting

5.1.1 Manually Constructed CoT Prompts

Wei et al. (2022) proposed Manual-CoT, an approach that employs manually-crafted demonstrations as prompts. Subsequently, Wang et al. (2022) proposed "Self-Consistency," a decoding strategy that aggregates multiple answers from LLMs through majority voting. Li et al. (2022) increased the randomness of the prompts to enhance the diversity of generated reasoning paths (Huang et al., 2024; Li et al., 2023; Liu et al., 2020). Diao et al. (2023) annotated the reasoning chain manually for the most uncertain questions. Despite their performance in enhancing reasoning, these approaches are costly, suboptimal, and sensitive.

5.1.2 Automatically Generated CoT Prompts

Kojima et al. (2022) proposed "Let’s think step by step" prompt that guides LLMs to generate reasoning steps without manually constructed demonstrations. Following this work, Zhang et al. (2022) and Shum et al. (2023) employed zero-shot-cot Kojima et al. (2022) to generate the reasoning process. In contrast, Shao et al. (2023) employed seed demonstrations to synthesize examples by automatically repeating forward and backward processes.

We propose a novel approach to generate reasoning chains by allowing LLMs to retrace their reasoning process after inferring the answer.

5.2 In-Context Learning

In-Context Learning (ICL) enables LLMs to perform target tasks during inference with a few task-specific examples, avoiding parameter modification (Shao et al., 2023; Brown et al., 2020; Zhang et al., 2023b, 2021, 2023a). According to Zhao et al. (2021), the accuracy of LLMs in ICL relies heavily on exemplar selection and permutation, promoting significant efforts to develop effective few-shot demonstration selection methods.

Zhang et al. (2022) adopted a clustering-based method to select demonstrations. Fu et al. (2022) selected the demonstrations with the most reasoning steps. Similarly, Diao et al. (2023) chose the demonstrations with most uncertain questions. Additionally, Shum et al. (2023) added the demonstrations with the correct answer to the samples pool and sampled the exemplars with a trained model. These studies all strive to minimize the use of incorrect exemplars. Contrarily, Zelikman et al. (2022) handled erroneous examples by hinting the model with the correct answers to generate results again. We conduct a comparative analysis with their approach, which is presented in Section A.5.

Through iterative bootstrapping, our approach selects challenging yet answerable exemplars, enhancing the LLMs' generalizability across varying difficulty levels.

6 Conclusion

This paper proposes Iter-CoT, an iterative bootstrapping in chain-of-thoughts prompting for LLM reasoning. Unlike previous work, our method prompts LLMs to self-correct their errors in reasoning chains by leveraging iterative bootstrapping and obtaining more precise and comprehensive reasoning chains. Experimental results on ten reasoning datasets among three different tasks demonstrate that our approach significantly outperforms the previous methods, achieving new state-of-the-art.

Acknowledgements

Chen Lin is the corresponding author supported by the Natural Science Foundation of China (No.62372390) and Xiamen Key Laboratory of Intelligent Storage and Computing, School of Informatics, Xiamen University.

Limitations

Cost of Iter-CoT Iter-CoT incurs the same cost during the inference stage as other baselines, as all additional expenses are only incurred during the construction phase of the demonstration pool. The demonstration pool has both a maximum and minimum size. The maximum size is obtained by applying Iter-CoT on the entire training set, while the minimum size corresponds to the required exemplars during inference.

Accuracy of Evaluator In Iter-CoT (w/o label), the model's performance relies significantly on the accuracy of the evaluator. Therefore, a more powerful and robust model than the primary one is required to serve as the evaluator. Moreover, the introduction of the evaluator also results in increased overhead.
References


Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yu-taka Matsuo, and Yusuke Iwasawa. 2022. **Large language models are zero-shot reasoners.** *arXiv Preprint.*


Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. **On the advance of making language models better reasoners.** *arXiv Preprint.*


Hugo Touvron, Louis Martin, Kevin Stone, Peter Al- bert, Amjad Almahairi, Yasmine Babaei, Nikolay


A Analysis for Iter-CoT

A.1 Performance Across Different Levels of Difficulty

We investigate the generalization ability of Iter-CoT on questions with varying difficulty levels. We follow the same hop-based criterion as previous work (Fu et al., 2022) to measure query difficulty. We sort the test set of GSM8K according to the number of hops of the annotated reasoning chains and conduct experiments using Iter-CoT and other baselines, as shown in Figure 9. Our results indicate that Iter-CoT is comparable to other methods for questions with few hops, whereas its performance is significantly better than other methods for questions with more hops. Iter-CoT performs the same as Simple-CoT on 2-hop questions (1% higher), while it is on par with Complex-CoT on 8-hop questions (2% lower) and is substantially ahead of the other methods (about 20%). This suggests that with Iter-CoT, we can select exemplars with intermediate difficulty levels, which can greatly improve the performance of LLM among questions of varying difficulty. Moreover, the effect of Iter-CoT(w/o label) is even superior to Iter-CoT(w/ label) on 7-hop and 8-hop questions, which shows the robustness of our proposed methods.

A.2 Effective of Different Numbers of Seed Examples

In order to investigate the sensitivity of our approaches and conventional CoT to the seed examples, we conducted an experiment on the GSM8K dataset as shown in Figure 10. It demonstrates that both of our approaches outperform CoT, and are more stable as the number of examples increases. Additionally, our experiment also shows that the overall performance is not determined by the quantity of increasing exemplars. For instance, the Iter-CoT peak occurs at five exemplars, while the Iter-CoT(w/o label) and Random-CoT peaks at four exemplars.

A.3 Comparison with CoT

As Section 3 mentions, Iter-CoT can generate more precise and comprehensive reasoning chains than zero-shot-CoT. We conduct inference on three distinct reasoning datasets (GSM8K, Letter(4) and Date Understanding) utilizing both Iter-CoT’s first stage and Zero-Shot-CoT with the same questions, which are shown in Table 5. We use the same LLMs and temperature to generate reasoning chains and answers. We observe that the Final CoT generated after the Iter-CoT’s first stage is naturally more precise and comprehensive compared to CoT generated by zero-shot-CoT, resulting in higher quality demonstrations.

A.4 Comparison with Verify-and-Correct Methods

Our work parallels certain existing methods that improve LLMs reasoning abilities through error correction strategies, such as ReAct (Yao et al., 2023), Self-Refine (Madaan et al., 2023), and Self-Ask (Press et al., 2023), all of which fall under the category of LLM-Agent approaches. The key difference of our Iter-CoT method from these lies in its unique mechanism: it learns from errors to create moderately challenging examples with comprehensive, detailed reasoning chains, facilitating in-context learning for the model. In contrast, methods like ReAct, Self-Refine, and Self-Ask employ LLMs more as agents for planning and decision-making. For example, Self-ask decomposes problems into sub-problems (planning), then utilizes a search engine to find answers for each sub-problem (decision-making), and then compiles the results. However, these methods typically involve iterative calls to LLMs for each query during the inference phase, leading to extended processing times and increased costs. In contrast, our approach builds a high-quality demonstration pool in advance, eliminating additional computational demands during the inference stage.

A.5 Comparison with STaR-CoT

In concurrent work closely related to ours, Zelikman et al. (2022) applies a similar methodology to ours in regenerating the reasoning chains for incorrect examples (we denote STaR-CoT in this paper). Although our work shares similar ideas to their approach, we differ in several key respects. First, their approach does not involve any contextual information from the previous step. Moreover, their approach provides the correct answer directly to the model, rather than using the correctness of the answer or error reasoning steps as bootstrapping information. It is more prone to generating erroneous reasoning chains despite arriving at the correct answer. Moreover, they continue fine-tuning the model using the corrected reasoning chains, which is impossible for LLMs with 175B parameters (such as ChatGPT (Ouyang et al., 2022)). We conduct inference on one arithmetic
reasoning dataset GSM8K utilizing both Iter-CoT’s first stage and STaR-CoT with the same questions shown in Table 6. We use the same LLMs and temperature to generate reasoning chains and answers. It is observed that although STaR-CoT generates the correct answer, the rationales are wrong, leading to confusion of the entire reasoning chain and reducing the overall performance (80.9% compared to 76.3%).

A.6 Average Length of Reasoning Chains

Figure 8: Average length of reasoning chains generated by different methods on GSM8K, CSQA and Last Letter Concatenation.

We compare the average length of reasoning chains generated by different methods, which are demonstrated in Figure 8. The average length of the reasoning chains demonstrated by Iter-CoT is significantly higher than other methods on the three datasets (GSM8K, CSQA and Last Letter Concatenation). These results provide solid evidence that the reasoning chains demonstrated by Iter-CoT are more comprehensive than those by other alternative methods.

B Experiment Details

B.1 Datasets and Tasks

We evaluate Iter-CoT using ten datasets from three different categories of reasoning tasks. The specific descriptions, divisions, and references of each dataset are shown in Table 7. The examples of each reasoning task are shown in Table 8.

C Full Exemplars Generated by Iter-CoT
Figure 9: The performance on GSM8K across different numbers of hops.

Figure 10: Efficiency of Different Numbers of Seed Examples on GSM8K.
Q: Four classmates were comparing their ages based on their birth month. They found out that Jolyn is 2 months older than Therese while Therese is 5 months older than Aivo. Then, Leon is 2 months older than Aivo. How much older in months is Jolyn than Leon? (GSM8K)

A: Reasoning process: We are given that Jolyn is 2 months older than Therese and Therese is 5 months older than Aivo. Let the ages of Therese and Aivo be $T$ and $A$, respectively. We can write the ages of Jolyn, Therese, and Aivo as $J = T+2$, $T = A+5$, and $A = A$. Leon is 2 months older than Aivo, so $L = A+2$. We want to find the difference between Jolyn's age and Leon's age, so $J - L = (T+2) - (A+2)$. Substituting for $T$ and $A$ using the equations above gives us $(A+7) - (A+2) = 5$. Therefore, Jolyn is 5 months older than Leon. Final answer: 5.

A: Let's think step by step. First let's say the ages of Therese and Aivo is $T$ and $A$. $T=A+5$ $J=T+2$ $L=A+2$ $J-L=T+2-(A+2)$ $J-L=2$ The correct answer is: 2. ✗

Q: Take the last letters of the words in "Steve Sweeney" and concatenate them. (Letter(4))

A: Reasoning process: First, we need to split the sentence "Steve Sweeney" into individual words by looking for the spaces. This gives two words: "Steve" and "Sweeney". Next, we need to take last letter of each word, which are "e" and "y". Finally, we need to concatenate these letters together to get the final answer. Final answer: ey.

A: Let's think step by step. First we take the last letters of the words in "Steve Sweeney", which are "ete eeney". Since we need to concatenate them, we will end up with: "ete eeney" Therefore, the correct answer is: "ete eeney". ✓

Q: May 6, 1992 is like yesterday to Jane, but that is actually ten years ago. What is the date one week from today in MM/DD/YYYY? Choices: A.05/13/2044 B.06/10/2002 C.05/13/2002 D.03/13/2003 E.11/13/2001 (Date Understanding)

A: Reasoning process: 1. May 6, 1992 is ten years ago from today, which means today is May 6, 2002. 2. The date one week from today is the date one week from May 6, 2002. 3. The date one week from May 6, 2002 is May 13, 2002. 4. The date one week from today in MM/DD/YYYY is 05/13/2002. Final answer: C.05/13/2002. ✓

A: Let's think step by step. The date one week from today is the date one week from Monday May 13, 2002. Thus the correct answer is: B.06/10/2002. ✗

Table 5: This is an example of Iter-CoT compared to CoT on three different reasoning datasets.
Q: Tina makes $18.00 an hour. If she works more than 8 hours per shift, she is eligible for overtime, which is paid by your hourly wage + 1/2 your hourly wage. If she works 10 hours every day for 5 days, how much money does she make? (GSM8K)

A: Reasoning process: First, let’s figure out how many hours Tina works per week. She works 10 hours per day and 5 days per week, so 10*5 = 50 hours. Next, let’s figure out how many of those hours are overtime. She is eligible for overtime if she works more than 8 hours per shift and she works 10 hours per shift, so 10-8 = 2 hours of overtime per shift. And since Tina works 5 shifts and 2 hours of overtime per shift, she works a total of 5*2 = 10 hours of overtime. Now we can figure out how much Tina makes in regular pay. She works 50-10 = 40 hours per week at her regular wage of $18 per hour, so she makes 40*18 = $720 in regular pay. Now we can figure out how much Tina makes in overtime pay. She gets time and a half for overtime, so 1.5 * $18 per hour = $27 per hour. And since she works 10 hours of overtime and makes $27 per hour, she makes 10*27 = $270 in overtime pay. So in total, Tina makes 720 + 270 = $990 per week.

✔

A: The correct answer is: 990. Let’s think step by step. Tina works 10 hours per day for 5 days, so she works 10*5 = 50 hours. She is eligible for overtime after 8 hours per shift, so she works 50-8 = 42 hours of overtime. She makes $18.00 per hour, so she makes 18*42 = 756 dollars of overtime pay. Overtime pay is her hourly wage + 1/2 her hourly wage, so she makes 18+(18/2) = $27.00 per hour of overtime. She works 42 hours of overtime, so she makes 27*42 = 1134 dollars of overtime pay. She makes $18.00 per hour for her regular hours, so she makes 18*8 = $144 per hour of regular pay. She makes a total of 144+1134 = 1278 dollars. Therefore, the correct answer is: 990. ✔

✔

Table 6: This is an example of Final CoT compared to STaR-CoT on GSM8K. The result shows that STaR-CoT is prone to generate erroneous reasoning chains despite arriving at the correct answer.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Answer Format</th>
<th>Examples</th>
<th>Train</th>
<th>Test</th>
<th>Licence</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM8K</td>
<td>Number</td>
<td>8</td>
<td>7473</td>
<td>1319</td>
<td>MIT License</td>
</tr>
<tr>
<td>AQuA</td>
<td>Multiple choice</td>
<td>4</td>
<td>97467</td>
<td>254</td>
<td>Apache-2.0</td>
</tr>
<tr>
<td>AddSub</td>
<td>Number</td>
<td>8</td>
<td>GSM8K*</td>
<td>395</td>
<td>-</td>
</tr>
<tr>
<td>SingleEq</td>
<td>Number</td>
<td>8</td>
<td>GSM8K*</td>
<td>508</td>
<td>-</td>
</tr>
<tr>
<td>SVAMP</td>
<td>Number</td>
<td>8</td>
<td>GSM8K*</td>
<td>1000</td>
<td>MIT License</td>
</tr>
<tr>
<td>ASDiv</td>
<td>Number</td>
<td>8</td>
<td>GSM8K*</td>
<td>2096</td>
<td>-</td>
</tr>
<tr>
<td>CSQA</td>
<td>Multiple choice</td>
<td>7</td>
<td>9741</td>
<td>1221</td>
<td>-</td>
</tr>
<tr>
<td>StrategyQA</td>
<td>Binary</td>
<td>6</td>
<td>2821</td>
<td>1880</td>
<td>Apache-2.0</td>
</tr>
<tr>
<td>Date</td>
<td>Multiple choice</td>
<td>8</td>
<td>69</td>
<td>300</td>
<td>Apache-2.0</td>
</tr>
<tr>
<td>Letter(4)</td>
<td>String</td>
<td>4</td>
<td>500(2*)</td>
<td>500(4*)</td>
<td>Apache-2.0</td>
</tr>
</tbody>
</table>

Table 7: The statistics of the datasets used in this paper. Examples are the number of examples demonstrations for each dataset. GSM8K* denotes constructed the training set using the GSM8K, cause no available training set for the current dataset. 2* and 4* in the "Letter(4)" row refers to using 2 letters in the training set while using 4 letters in the test set (Out-of-Domain).
### Reasoning Tasks Example

| Arithmetic | Marissa is hiking a 12-mile trail. She took 1 hour to walk the first 4 miles, then another hour to walk the next two miles. If she wants her average speed to be 4 miles per hour, what speed (in miles per hour) does she need to walk the remaining distance? |
| Commonsense | Sammy wanted to go to where the people were. Where might he go? Choices: A. race track B. populated areas C. the desert D. apartment E. roadblock |
| Symbolic | Take the last letters of each word in “Geo Kody Isaiai Giovanni” and concatenate them. |

Table 8: The examples are sampled from the test sets of three typical reasoning datasets, GSM8K (Cobbe et al., 2021), CommonSenseQA (Talmor et al., 2019) and Last Letter Concatation (Wei et al., 2022), respectively.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>Iter-CoT Exemplars</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM8K</td>
<td>Q: Sheila, Purity, and Rose want to rent a house. Sheila has offered to pay five times Purity’s share of the rent. Rose can only afford thrice what Purity pays. If Rose’s share is $1,800, what is the total house rent? A: Reasoning Process: We are given that Rose’s share of the rent is $1800, and that Rose can only afford thrice what Purity pays. Therefore, Purity’s share of the rent is 1800 / 3 = $600. We are also given that Sheila has offered to pay five times Purity’s share of the rent. Therefore, Sheila’s share is 5 * 600 = $3000. Finally, we have Rose’s share of $1800, Purity’s share of $600 and Sheila’s share of $3000. Therefore, the total rent is 1800 + 600 + 3000 = $5400. Final answer: 5400.</td>
</tr>
<tr>
<td></td>
<td>Q: Padma is trading cards with Robert. Padma started with 75 cards and traded 2 or her valuable ones for 10 of Robert’s cards. Robert started with 88 of his own cards and traded another 8 of his cards for 15 of Padma’s cards. How many cards were traded between both Padma and Robert? A: Reasoning Process: Padma started with 75 cards. She traded 2 cards for 10 cards, so she now has 75-2+10=83 cards. Robert started with 88 cards. He traded 8 cards for 15 cards, so he now has 88-8+15=95 cards. The total number of cards traded between both Padma and Robert is 2+8+10+15=35 cards. This includes the 2 cards that Padma traded for 10 cards, the 8 cards that Robert traded for 15 cards, and the 10 and 15 cards that were traded between them. Final answer: 35.</td>
</tr>
<tr>
<td></td>
<td>Q: Juanita goes through 1 bottle of sunscreen a month. She likes to stock up for the entire year when a certain website offers 30% off her entire purchase. If each bottle is $30.00 how much will all of the sunscreen cost after the discount? A: Reasoning Process: Juanita needs 1 bottle of sunscreen a month, and there are 12 months in a year, so she needs 12 bottles for the year. The website is offering a 30% discount on her entire purchase. Each bottle of sunscreen costs $30.00. To find the total cost of the sunscreen after the discount, we first need to find the discount amount. This is done by multiplying the price per bottle by the discount rate and the number of bottles: $30.00 * 0.3 * 12 = $108.00. Next, we need to find the price before the discount. This is done by multiplying the price per bottle by the number of bottles: $30.00 * 12 = $360.00. Finally, we subtract the discount amount from the total price before the discount to find the final price after the discount: $360.00 - $108.00 = $252.00. Final answer: 252.</td>
</tr>
<tr>
<td></td>
<td>Q: On a three-day trip, Wendy drove 125 miles on the first day, and 223 miles on the second day. How many miles did she drive on the third day, if the total miles that Wendy drove for the trip is 493 miles? A: Reasoning Process: 1. Wendy drove 125 miles on the first day and 223 miles on the second day, so the total miles she drove on the first two days is 125+223 = 348 miles. 2. The total miles Wendy drove on the trip is 493 miles. 3. Therefore, the remaining miles she drove on the third day is 493-348 = 145 miles. Final answer: 145.</td>
</tr>
<tr>
<td></td>
<td>Q: Mrs. Lopez and her family are going to the movie theatre. Adult tickets are $11. Children’s tickets (ages 3-12) are $8. Senior citizen’s tickets (ages 60+) are $9. Mrs. Lopez needs to buy movie tickets for her husband, herself, her parents (ages 72 and 75), and her three children (ages 7, 10, 14). How much money does she need to buy everyone’s tickets? A: Reasoning Process: 1. Mrs. Lopez needs to buy tickets for her husband and herself. The cost for these tickets will be $11<em>2=$22. 2. Mrs. Lopez needs to buy tickets for her parents, who are senior citizens. The cost for these tickets will be $9</em>2=$18. 3. Mrs. Lopez needs to buy tickets for her three children, who are between the ages of 3-12. The cost for these tickets will be $8*3=$24. The total cost of all the tickets will be $22+$18+$24=$64. Final answer: 64.</td>
</tr>
<tr>
<td></td>
<td>Q: Mark has 30 candies, Peter has 25 candies, and John has 35 candies. They decided to combine their candies together and share them equally. How many candies will each one of them have? A: Reasoning Process: 1. Mark has 30 candies, Peter has 25 candies, and John has 35 candies. 2. We need to add the number of candies that each person has to find the total number of candies they have. 3. So, 30 + 25 + 35 = 90 candies. 4. Then, we need to divide the total number of candies by the number of people, which is 3. So, 90 / 3 = 30 candies. Final answer: 30.</td>
</tr>
</tbody>
</table>

Table 9: The exemplars are selected on GSM8K train set. They are transferred to AddSub, SingleEq, SVAMP and ASDiv.
Q: Ben’s potato gun can launch a potato 6 football fields. If a football field is 200 yards long and Ben’s dog can run 400 feet/minute, how many minutes will it take his dog to fetch a potato he launches?
A: Reasoning Process: To find the total distance that the potato can travel, we need to multiply the number of football fields by the number of feet per football field. Since there are 3 feet per yard, and there are 200 yards per football field, we have 200×3 = 600 feet per football field. Therefore, the total distance that the potato can travel is 6×600 = 3600 feet. Since Ben’s dog can run at a rate of 400 feet per minute, it will take the dog 3600/400 = 9 minutes to fetch the potato. Final answer: 9.

Q: Brittany has 3 gerbils: Puffy, Muffy, and Scruffy. Puffy weighs 5 ounces more than Muffy. Muffy weighs 3 ounces less than Scruffy. If Scruffy weighs 12 ounces, how much would the scale indicate, in ounces, if Brittany put Puffy and Muffy on the scale?
A: Reasoning Process: We are given that Scruffy weighs 12 ounces and that Muffy weighs 3 ounces less than Scruffy. We are also told that Puffy weighs 5 ounces more than Muffy. Using this information, we can calculate the weight of Muffy by subtracting 3 ounces from Scruffy’s weight of 12 ounces, giving us a weight of 12 - 3 = 9 ounces for Muffy. We can then calculate the weight of Puffy by adding 5 ounces to Muffy’s weight of 9 ounces, giving us a weight of 9 + 5 = 14 ounces for Puffy. If Brittany puts Puffy and Muffy on the scale, the scale would indicate a weight of 14 + 9 = 23 ounces. Final answer: 23.

Table 10: (Cont.) The exemplars are selected on GSM8K train set. They are transferred to AddSub, SingleEq, SVAMP and ASDiv.

Q: What number has a 5:1 ratio to the number 11? Options: A:22 B:50 C:55 D:52 E:12
A: Reasoning Process: In order to find the number that has a 5:1 ratio to 11, we first need to understand what a ratio means. A ratio is a comparison of two numbers, and it is usually written in the form of a:b. In this case, the ratio is 5:1, which means that for every 5 units of the first number, there is 1 unit of the second number. Therefore, if we want to find
the number that has a 5:1 ratio to 11, we need to find a number that is 5 times larger than 11. Then, we can check each of the given options to see if any of them are 5 times larger than 11. The only option that fits this description is C:55. Final answer: C.

Q: The average of seven numbers is 18. The average of first three numbers is 14 and the average of last three numbers is 23. What is the middle number? Options: A:25 B:27 C:15 D:32 E:34
A: Reasoning Process: 1. The average of the seven numbers is 18, so the sum of the seven numbers is 18 × 7 = 126
2. The average of the first three numbers is 14, so the sum of the first three numbers is 14 × 3 = 42
3. The average of the last three numbers is 23, so the sum of the last three numbers is 23 × 3 = 69
4. The sum of the first three numbers and the last three numbers is 42 + 69 = 111
5. The sum of the middle number and the sum of the first three numbers plus the last three numbers is equal to the sum of the seven numbers, in other words, 111 + x = 126.
6. The middle number is x = 15 Final answer: C.

Q: Average score for Virat Kohli in a series of 10 matches is 38.9 runs. If the average for first six matches comes out to be 42 what is his average in the last 4 matches of the series? Options: A:34.25 B:34.28 C:24.252 D:64.28 E:34.21
A: Reasoning Process: 1) To find the average score for Kohli in the last 4 matches, we need to find the total number of runs he scored in these matches. 2) We know that his average score in the series of 10 matches is 38.9 runs. Therefore, he scored a total of 38.9 × 10 = 389 runs in these matches. 3) We are given that his average score in the first 6 matches is 42 runs. Therefore, he scored a total of 42 × 6 = 252 runs in these matches. 4) The total number of runs he scored in the last 4 matches is 389 - 252 = 137 runs. 5) Therefore, his average score in the last 4 matches is 137 / 4 = 34.25 runs. Final answer: A.

Q: A train runs at the speed of 72 kmh and crosses a 250 metre long platform in 26 seconds. What is the length of the train ? Options: A:220 metre B:230 metre C:250 metre D:260 metre E:270 metre
A: Reasoning Process: We know that the train crosses a 250 metre long platform in 26 seconds at the speed of 72 kmph. To find the length of the train, we need to first convert the speed from kmph to m/s, so we use the formula: \[ V = \frac{V_{kmph}}{3.6} \]. Plugging in the values, we get: \[ V = \frac{72}{3.6} = 20 \text{ m/s} \]. Now that we have the speed in m/s, we can use the formula: \[ D = V \times t \] to find the distance the train covers in 26 seconds. Plugging in the values, we get: \[ D = 20 \times 26 = 520 \text{ m} \]. Since the length of the platform is 250 metre, and the train covers 520 metre in 26 seconds, the length of the train must be: \[ 520 - 250 = 270 \text{ m} \]. Final answer: E.

Table 11: The exemplars are selected on AQuA train set.
<table>
<thead>
<tr>
<th>Q</th>
<th>CSQA</th>
<th>Iter-CoT Exemplars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where could a fungus grow and not be disturbed by sunlight? Choices: A. under rocks B. Manhattan C. toenails D. grocery store E. fallen tree</td>
<td>A. Reasoning process: 1. Fungi need moisture and shade to grow. 2. Rocks can provide both moisture and shade. 3. Therefore, rocks can be a good place for fungi to grow without being disturbed by sunlight. Final answer: A.</td>
<td></td>
</tr>
<tr>
<td>The man wanted to take a bus, but all the chairs were full of people doing what? Choices: A. sitting down B. refusing C. saving money D. sitting down E. going outdoors</td>
<td>A. Reasoning process: 1. Eliminate answer C (&quot;saving money&quot;) because it doesn’t have anything to do with taking a bus. 2. Eliminate answer E (&quot;going outdoors&quot;) because it doesn’t make sense for people to be on a bus if they’re already outside. 3. Choose answer A (&quot;sitting down&quot;) because it’s the only remaining answer that makes sense given the context clues in the question. Final answer: A.</td>
<td></td>
</tr>
<tr>
<td>A beaver can destroy a machine’s functionality if they build their dam in this? Choices: A. strip club B. pocket C. millpond D. Ontario E. lake or river</td>
<td>A. Reasoning process: 1. A beaver’s dam is built in a body of water. 2. A millpond is a man-made pond used for a watermill. 3. Therefore, if a beaver builds its dam in a millpond, it can potentially destroy the functionality of the watermill. Final answer: C.</td>
<td></td>
</tr>
<tr>
<td>How might a person address someone who is leaving? Choices: A. further cause B. wave goodbye C. attempt suicide D. insulting them E. thank god</td>
<td>A. Reasoning process: 1. We know that the person is leaving, so we can eliminate answer choices A and C because they don’t make sense in context. 2. We can further eliminate answer choice D because it’s rude to insult someone who is leaving. 3. That leaves us with answer choices B and E. 4. Out of these two, the more appropriate option would be B because it’s a common way to show polite acknowledgement of someone leaving. Final answer: B.</td>
<td></td>
</tr>
<tr>
<td>What is a characteristic of thin glass? Choices: A. break easily B. shattering C. melt D. bend E. hold water</td>
<td>A. Reasoning process: A. Break easily - This fits the characteristic of thin glass, as it is known for its fragility and tendency to break under pressure. B. Shattering - This could be a possible characteristic of thin glass, but it is not mentioned in the given choices, so we cannot select it as the answer. C. Melt - This does not fit the characteristic of thin glass, as it does not need to be heated in order to bend or shape it. D. Bend - This does not fit the characteristic of thin glass, as it is too fragile to be bent or shaped without breaking. E. Hold water - This does not fit the characteristic of thin glass, as it is not a property typically associated with glass. Final answer: A.</td>
<td></td>
</tr>
<tr>
<td>What might be the result of a season of successful skiing? Choices: A. finish line B. broken bones C. broken legs D. chapped lips E. healthy body</td>
<td>A. Reasoning process: 1. A season of successful skiing could result in a healthy body if the person is exercising regularly and taking good care of their physical health while skiing. 2. None of the other options, such as a finish line, broken bones, or chapped lips, are related to the potential health benefits of skiing, so they cannot be the correct answer. 3. Therefore, the correct answer is E. healthy body. Final answer: E.</td>
<td></td>
</tr>
<tr>
<td>Glass that hasn’t been treated to be extra strong is what? Choices: A. weak B. fragile C. forceless D. regular E. flimsy</td>
<td>A. Reasoning process: 1. The question asks about glass that hasn’t been treated to be extra strong. This means that the glass has not undergone any special processes or treatments to make it stronger than usual. 2. The answer choices are: A. weak, B. fragile, C. forceless, D. regular, and E. flimsy. 3. We need to find the answer choice that best describes glass that hasn’t been treated to be extra strong. 4. &quot;A. weak&quot; means not strong. &quot;B. fragile&quot; means easily broken. &quot;C. forceless&quot; means without strength or power. &quot;D. regular&quot; means normal or usual. &quot;E. flimsy&quot; means weak and easily broken. 5. Based on these definitions, the correct answer is &quot;B. fragile&quot;. Glass that hasn’t been treated to be extra strong is likely to be fragile, or easily broken. This is because it has not undergone any special processes to make it stronger, so it will be more susceptible to breakage and damage. Other words that may be used to describe this type of glass include &quot;delicate&quot; or &quot;brittle&quot;. Final answer: B.</td>
<td></td>
</tr>
</tbody>
</table>

Table 12: The exemplars are selected on CSQA train set.
<table>
<thead>
<tr>
<th>DATASET</th>
<th>Iter-CoT Exemplars</th>
</tr>
</thead>
</table>
| StrategyQA | Q: Does a Toyota Prius require an anchor in order to park?  
A: Reasoning process: A Toyota Prius is a car, and any car needs to be anchored to the ground in order to park, unless it is parked on a slope. Final answer: Yes.  
Q: Did the husband of Elizabeth Hamilton work with a lieutenant?  
A: Reasoning process: 1. First, we need to know who Elizabeth Hamilton is. Elizabeth Hamilton was the wife of Alexander Hamilton.  
2. Then, we need to determine if Alexander Hamilton worked with a lieutenant. According to historical records, Alexander Hamilton did work with a lieutenant named John Laurens during the American Revolutionary War.  
3. Therefore, the answer to the question "Did the husband of Elizabeth Hamilton work with a lieutenant?" is Yes. Final answer: Yes.  
Q: Would Constantinople survive a hypothetical invasion by Charlemagne?  
A: Reasoning process: 1. Constantinople was the capital of the Byzantine Empire, and Charlemagne was the ruler of the Carolingian Empire, which was in Europe.  
2. Constantinople was a well-fortified city with strong walls and defenses.  
3. Charlemagne was known for his military prowess and conquest.  
4. The Byzantine Empire was a major power in its time, and Constantinople was a key strategic city for the empire.  
5. However, the Carolingian Empire was also a major power and Charlemagne was a skilled and experienced military leader.  
6. In a hypothetical situation, Constantinople may put up a strong resistance against Charlemagne’s forces, but it would not be able to withstand a sustained attack. Final answer: Yes.  
Q: Would King Henry V hypothetically prevail against Charlemagne at Agincourt?  
A: Reasoning process: 1. First, we need to understand the question and identify the key information: "Did John Lennon collaborate with the filmmaker behind Walden?"  
2. Next, we need to research what Walden is and who the filmmaker behind it is: Walden is a documentary film by Alan Watts.  
3. With this information, we can now search for evidence of John Lennon collaborating with Alan Watts.  
4. After conducting research, it is clear that John Lennon did collaborate with Alan Watts on a film project called "The Touch of Zen". Final answer: Yes.  
Q: Did John Lennon collaborate with the filmmaker behind Walden?  
A: Reasoning process: Monks are members of religious orders that usually take vows of poverty, chastity, and obedience. These vows are typically associated with a life of spiritual contemplation and non-violent pursuits. However, there are examples of monks engaging in warfare, such as the Knights Templar during the Crusades and Buddhist Shaolin monks who have defended temples during conquests. These examples suggest that it is not always the case that monks are forbidden from engaging in warfare. Final answer: Yes.  
Q: Could Saint Augustine have read Hamlet?  
A: Reasoning process: Saint Augustine was an early Christian saint who lived about 1,600 years ago. Hamlet is a play by Shakespeare which is believed to have been written in the late 16th century. It is possible that the play was written earlier than that and performed for Saint Augustine. Final answer: Yes. |

Table 13: The exemplars are selected on StrategyQA train set.
Q: Yesterday was 12/31/1929. Today could not be 12/32/1929 because December has only 31 days. What is the date yesterday in MM/DD/YYYY? Choices: A. 12/31/1929 B. 12/31/1856 C. 01/10/1930 D. 04/01/1930 E. 12/31/1838
A: Reasoning process: 1. We know that yesterday was 12/31/1929.
2. Today cannot be 12/32/1929 because December only has 31 days.
3. Therefore, the date yesterday must be 12/31/1929; Final answer: A.

Q: The deadline is Jun 1, 2021, which is 2 days away from now. What is the date one week from today in MM/DD/YYYY? Choices: A. 05/23/2044 B. 06/10/2002 C. 05/13/2002 D. 03/13/2003 E. 11/13/2001
A: Reasoning process: When we are given that 2015 is coming in 36 hours, we first need to determine what the current date is. We know that 36 hours is equal to 1.5 days, so we need to subtract 1.5 days from the date when 2015 arrives, which is 12/31/2014. This means that the current date is 12/29/2014; Final answer: E.

Q: May 6, 1992 is like yesterday to Jane, but that is actually ten years ago. What is the date one week from today in MM/DD/YYYY? Choices: A. 05/13/2044 B. 06/10/2002 C. 05/13/2002 D. 03/13/2003 E. 11/13/2001
A: Reasoning process: 1. May 6, 1992 is ten years ago from today, which means today is May 6, 2002.
2. The date one week from today is the date one week from May 6, 2002.
3. The date one week from May 6, 2002 is May 13, 2002.
4. The date one week from today in MM/DD/YYYY is 05/13/2002. Final answer: C.

Q: The current local time is 3:02 pm of 5/4/2004. What is the date tomorrow in MM/DD/YYYY? Choices: A. 05/05/1915 B. 05/06/2004 C. 05/05/2005 D. 05/04/2004 E. 02/15/2004 F. 05/05/2004
A: Reasoning process: The current date is 5/4/2004, and tomorrow is the next day, which is 5/5/2004. This is because the current month is May and the current day is the 4th, so the next day is the 5th. The year remains the same because we are only looking at the date for tomorrow and not a longer time period; Final answer: F.

A: Reasoning process: * The question is asking for the date today in MM/DD/YYYY format. * We know that today's date is September 9, which is 09. * The next step is to find the correct day. September has 30 days, so the date today is 09/30/1909. * However, this is not correct because the year is 1909, not 1939. * Therefore, the correct date today is 09/09/1909; Final answer: C.

Q: Today is the palindrome day of 2020, because the MMDDYYYY format of the date is the same backwards as forwards. What is the date 24 hours later in MM/DD/YYYY? Choices: A. 02/03/1982 B. 02/03/2100 C. 02/03/2020 D. 02/04/2020 E. 02/03/2094 F. 01/02/2020
A: Reasoning process: The palindrome date is of the form MM/DD/YYYY, and it is the same when read backwards. The date of today is 02/02/2020 and if we read it backwards it is still the same, so it is the palindrome date. The next day is the 02/03/2020 and the palindrome date of the day after tomorrow is the 02/04/2020; Final answer: C.

Q: The deadline is Jun 1, 2021, which is 2 days away from now. What is the date one week ago from today in MM/DD/YYYY? Choices: A. 05/23/2021 B. 05/24/2021 C. 04/13/2021 D. 05/23/1940 E. 05/23/2021 F. 05/23/2020
A: Reasoning process: If the deadline is Jun 1, 2021, which is 2 days away, then today is May 30, 2021. If today is May 30, 2021, then one week ago was May 23, 2021; Final answer: E.

Table 14: The exemplars are selected on Date Understanding train set.
Table 15: The exemplars are selected on Last Letter Concatenation train set.
<table>
<thead>
<tr>
<th>DATASET</th>
<th>Iter-CoT(wo label) Exemplars</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM8K</td>
<td>Q: In a house, there are 16 cats. Two of them are white, and 25% of them are black. The rest of the cats are grey. How many grey cats are in this house? A: Reasoning Process: Given: Total number of cats = 16 Number of white cats = 2 Percentage of black cats = 25% Step 1: Calculate the number of black cats. Number of black cats = (25/100) * 16 = 4 Step 2: Calculate the number of grey cats. Number of black cats = Total number of cats - (Number of white cats + Number of black cats) Number of grey cats = 16 - (2 + 4) = 10 Final answer: 10.</td>
</tr>
<tr>
<td></td>
<td>Q: Ben starts each day with $50, spends $15 every day and saves the rest. After a few days, his mom doubled his total savings and his dad gave him an additional $10. If he now has $500, How many days have elapsed? A: Reasoning Process: We know that Ben starts each day with $50 and spends $15 every day. So, his daily savings can be calculated as: Daily savings = $50 - $15 = $35 Let's denote the number of days that have elapsed as &quot;d&quot;. Therefore, the total savings before his mom doubled it and his dad gave him an additional $10 can be calculated as: Total savings before mom doubled it and dad gave him an additional $10 = ($35 * d) + $10 According to the question, his total savings after his mom doubled it and his dad gave him an additional $10 is $500. So, we can write the equation: ($35 * d) + 2 + $10 = $500 Simplifying the equation, we get: ($37 * d) + $10 = $500 Subtracting $10 from both sides, we get: ($37 * d) = $490 Dividing both sides by $70, we get: d = $490 / $70 Calculating the value of d, we get: d = 7 Therefore, approximately 7 days have elapsed. Final answer: 7.</td>
</tr>
<tr>
<td></td>
<td>Q: There were 600 people in the stadium when the football game started. Before the game was over, one-fourth of the boys and one-eighth of the girls left early. How many people remained to see the end of the game if there were 240 girls at the beginning of the game? A: Reasoning Process: To find out how many people remained to see the end of the game, we need to subtract the number of boys who left early and the number of girls who left early from the total number of people at the beginning of the game. Given: Total people at the beginning of the game = 600 Number of girls at the beginning of the game = 240 First, let's calculate the number of boys at the beginning of the game: Total people at the beginning of the game - Number of girls at the beginning of the game = Number of boys at the beginning of the game = 600 - 240 = 360 Next, let's calculate how many boys left early: (1/4) * Number of boys at the beginning of the game = (1/4) * 360 = 90 boys Then, let's calculate how many girls left early: (1/8) * Number of girls at the beginning of the game = (1/8) * 240 = 30 girls Finally, to find out how many people remained to see the end of the game, subtract the number of boys who left early and the number of girls who left early from the total number of people at the beginning of the game: Total people at the beginning of the game - Number of boys who left early - Number of girls who left early = Number of people who remained at the end of the game 600 - 90 - 30 = 480 Final answer: 480.</td>
</tr>
<tr>
<td></td>
<td>Q: A basketball team played 40 games and won 70% of the games. It still had 10 games to play. How many games can they lose to win 60% of their games? A: Reasoning Process: To solve the problem, we need to determine the maximum number of games the basketball team can lose to maintain a 60% win rate. Step 1: Calculate the total number of games played by adding the games already played (40) to the games remaining (10): Total number of games = 40 + 10 = 50 games Step 2: Calculate the number of games the team needs to win to achieve a 60% win rate: Number of games to win = 50 x 60% = 30 games Step 3: Calculate the number of games already won: Number of games already won = 40 x 70% = 28 games Step 4: Calculate the maximum number of games the team can lose: Maximum number of games that can be lost = Number of games to win - Number of games already won = 30 - 28 = 2 games. Final answer: 2.</td>
</tr>
<tr>
<td></td>
<td>Q: Jack will have ten times more handball trophies than Michael has right now in three years. If Michael has 30 trophies right now, and the number of his trophies increases by 100 in three years, what's the total number of trophies they'll have altogether after three years? A: Reasoning Process: Given: - Michael currently has 30 handball trophies. - In three years, the number of Michael's trophies will increase by 100. To find the total number of trophies they'll have altogether after three years, we need to calculate the number of trophies Jack will have in three years and then add it to the number of trophies Michael will have in three years. Since Jack will have ten times more handball trophies than Michael in three years, we can calculate Jack's number of trophies as follows: Jack's trophies in three years = 10 * (Michael's trophies in three years) = 10 * (30 + 100) = 10 * 130 = 1300. Now, let's calculate the total number of trophies they'll have altogether after three years: Total trophies in three years = Michael's trophies in three years + Jack's trophies in three years = 130 + 1300 = 1430. Final answer: 1430.</td>
</tr>
<tr>
<td></td>
<td>Q: John decides to install a ramp in his house. He needs to get permits which cost $250. After that, he hires a contractor which costs $150 an hour and the guy works for 3 days at 5 hours per day. He also pays an inspector 80% less to make sure it is OK. How much was the total cost? A: Reasoning Process: The cost of permits is $250. The contractor works for 3 days, with 5 hours per day, resulting in a total of 3 * 5 = 15 hours. The contractor charges $150 per hour, so the total cost of hiring the contractor is 15 * $150 = $2250. The inspector is paid 80% less, which means he is paid only 20% of the original cost. So, the cost of the inspector is 20/100 * $2250 = $450. To calculate the total cost, we need to add the cost of permits, contractor, and inspector: Total cost = Cost of permits + Cost of contractor + Cost of inspector = $250 + $2250 + $450 = $2950. Final answer: 2950.</td>
</tr>
</tbody>
</table>

Table 16: The exemplars are selected on GSM8K train set. This set of exemplars is used by GSM8K, ASDiv, SVAMP, and SingleEq.
GSM8K
Q: Ignatius owns 4 bicycles. A friend of his owns different types of cycles, which have three times as many tires as Ignatius’s bikes have. He has one unicycle, a tricycle, and the rest are bikes. How many bicycles does the friend own? A: Reasoning Process: Ignatius owns 4 bicycles, and each bicycle has 2 tires. So Ignatius has a total of 4 bicycles * 2 tires per bicycle = 8 tires. The friend’s cycles have three times as many tires as Ignatius’s bikes. Therefore, the friend’s cycles have 3 * 8 tires = 24 tires. We also know that the friend has one unicycle (1 tire) and one tricycle (3 tires). To find out how many bicycles the friend owns, we subtract the number of tires from the unicycle and tricycle from the total number of tires the friend’s cycles have: 24 tires - 1 tire (unicycle) - 3 tires (tricycle) = 20 tires Since each bicycle has 2 tires, the number of bicycles the friend owns is 20 tires / 2 tires per bicycle = 10 bicycles. Therefore, the friend owns 10 bicycles. Final answer: 10.

Q: James spends 3 hours vacuuming and 3 times as long on the rest of his chores. How long does he spend on his chores? A: Reasoning Process: To find out how long James spends on his chores, we need to add the time he spends vacuuming (3 hours) to the time he spends on the rest of his chores. The rest of his chores take 3 times as long as vacuuming, so we can calculate that by multiplying the vacuuming time (3 hours) by 3. 3 hours (vacuuming) x 3 = 9 hours (rest of chores) Now, we add the vacuuming time and the rest of the chores time together: 3 hours (vacuuming) + 9 hours (rest of chores) = 12 hours Therefore, James spends a total of 12 hours on his chores. Final answer: 12.

Table 17: (Cont.) The exemplars are selected on GSM8K train set. This set of exemplars is used by GSM8K, ASDiv, SVAMP, and SingleEq.

AQuA
Q: There are a lot of houses such that the numbers of their doorplates are odd numbers and the first number of the doorplates is 545, the last number of the doorplates is 805. How many houses are there? Options: A)125 B)111 C)101 D)121 E)131 A: Reasoning process: To find the count of houses with odd doorplate numbers between 545 and 805, we need to find the difference between the two numbers and divide it by 2 (since we are incrementing by 2 for odd numbers) and then add 1 to include the first number. The difference between 805 and 545 is 260. Dividing 260 by 2 gives us 130. Adding 1 to include the first number gives us a total of 131 houses. Final answer: E.

Q: The ratio of the radius of two circles is 2: 3, and then the ratio of their areas is? Options: A)1: 8 B)4:9 C)1: 9 D)1: 3 E)1: 2 A: Reasoning process: To find the ratio of the areas of two circles, we need to square the ratio of their radii. Given that the ratio of the radii is 2:3, the ratio of their areas would be $(2^2):(3^2)$ which simplifies to 4:9. Final answer: B.

Q: A train passes a station platform in 36 seconds and a man standing on the platform in 20 seconds. If the speed of the train is 54 km/hr, what is the length of the platform? Options: A)388 B)(240 C)88 D)66 E)221 A: Reasoning process: Given: - Speed of the train = 54 km/hr - Train passes a station platform in 36 seconds - Train passes a man standing on the platform in 20 seconds Let's denote: - Length of the train as 'L' - Length of the platform as 'P' When the train passes the platform, it covers a distance equal to the length of the train plus the length of the platform. Distance covered when passing the platform = L + P The train covers this distance in 36 seconds. Speed = Distance / Time Therefore, (L + P) / 36 = 54 km/hr To convert km/hr to m/s, we multiply by (5/18): (L + P) / 36 = 54 * (5/18) m/s Simplifying, we have: (L + P) / 36 = 15 m/s Similarly, when the train passes the man, it covers a distance equal to the length of the train. Distance covered when passing the man = L. The train covers this distance in 20 seconds. Speed = Distance / Time Therefore, L / 20 = 54 * (5/18) m/s Simplifying, we have: L / 20 = 15 m/s Now, we have two equations: 1) (L + P) / 36 = 15 2) L / 20 = 15 From equation 2, we can express L in terms of P: L = 20 * 15 = 300 Substituting this value of L in equation 1: (300 + P) / 36 = 15 Simplifying, we have: 300 + P = 15 * 36 300 + P = 540 P = 540 - 300 P = 240 Therefore, the length of the platform is 240 meters. Final answer: B.

Q: A fruit seller had some apples. He sells 40% apples and still has 420 apples. Originally, he had Options: A)588 apples B)600 apples C)672 apples D)700 apples E)None A: Reasoning process: The fruit seller sells 40% of his apples and still has 420 apples remaining. Let’s denote the original number of apples as x. After selling 40% of x, the remaining number of apples is 0.6x. According to the given information, 0.6x = 420. To find the value of x, we divide both sides of the equation by 0.6: x = 420 / 0.6 = 700. Final answer: D.

Table 18: The exemplars are selected on AQuA train set.
<table>
<thead>
<tr>
<th>DATASET</th>
<th>Iter-CoT(w/o label) Exemplars</th>
</tr>
</thead>
</table>
| CSQA    | Q: The Coast Guard is a police force that often does what? Choices: A. board vessels B. direct traffic C. ocean D. force law enforcement E. arrest  
A. Reasoning process: The Coast Guard is a branch of the military that is responsible for maritime law enforcement and search and rescue operations. They are often involved in activities related to the sea and vessels. Among the options provided, the most suitable answer is A. board vessels. The Coast Guard frequently boards vessels for inspections, investigations, and enforcement of maritime laws and regulations. Final answer: A.  
A. Reasoning process: To determine the best option for storing folding chairs for religious experiences, we need to consider the specific requirements and context of religious activities. Option A: Closet - This could be a viable choice if there is a dedicated closet within the religious establishment specifically for storing folding chairs. It would ensure that the chairs are easily accessible and organized. Option B: Church - While a church may have designated areas for storing chairs, such as a storage room or basement, it may not be the most suitable choice if the religious experiences are held in other locations outside of the church. Option C: Room - This option is quite generic and lacks specificity. It could potentially be a suitable choice if there is a dedicated room within the religious establishment specifically for storing folding chairs. Option D: Basement - Similar to option B, a basement could be a suitable choice if the religious establishment has a dedicated basement area for storing chairs. However, if the religious experiences are held in other locations, this may not be the most practical option. Option E: Garage - Storing folding chairs in a garage may not be ideal for religious experiences as it is typically used for parking vehicles and may not provide a suitable and organized space for the chairs. Considering the various options, the most suitable choice for storing folding chairs for religious experiences would be A. closet. This option ensures that the chairs are easily accessible and organized within the religious establishment. Final answer: A.  
Q: What might people do to protect their legs from getting dirty on the farm? Choices: A. wear jeans B. milk cow C. believe in god D. avoid mud E. plant flowers  
A. Reasoning process: To protect their legs from getting dirty on the farm, people can take several measures. Let's analyze the options provided: A. wear jeans - This option seems like a practical choice as jeans are known for their durability and ability to provide protection against dirt and mud. B. milk cow - While milking a cow is a common farm activity, it does not directly address the question of protecting legs from getting dirty. C. believe in god - This option does not appear to be relevant to protecting legs from getting dirty on the farm. D. avoid mud - Avoiding mud can be a preventive measure, but it may not always be possible or practical on a farm. E. plant flowers - Planting flowers does not address the question of protecting legs from getting dirty. Final answer: A.  
Q: What may jogging make you feel like? Choices: A. increased heart rate B. improved health C. frightened D. shin splints E. being tired  
A. Reasoning process: When you go jogging, your heart rate typically increases as your body works harder to meet the demands of the exercise. So, option A (increased heart rate) seems like a plausible choice. Jogging is also known to have numerous health benefits, such as improving cardiovascular fitness, strengthening muscles, and aiding in weight loss. Therefore, option B (improved health) is also a likely option. However, jogging is not typically associated with feelings of being frightened or experiencing shin splints. Shin splints are a common overuse injury that can occur during running or high-impact activities, but they are not directly related to jogging itself. Lastly, jogging can often lead to feelings of tiredness, especially if you push yourself and exert a lot of energy during the activity. So, option E (being tired) is a reasonable choice as well. Final answer: E.  
Q: The best way to cure boredom at a high elevation with no buildings around is to do what? Choices: A. do crossword puzzle B. go skiing C. play games D. hang out at bar E. sleep  
A. Reasoning process: In a high elevation with no buildings around, the best way to cure boredom would be to go skiing. Skiing is a popular recreational activity that can be enjoyed in mountainous areas, especially at high elevations. It provides an exhilarating experience and allows you to make the most of the natural surroundings. Final answer: B.  
Q: Where is an inexpensive place to have a reception? Choices: A. outside B. wedding party C. restaurant D. church basement E. country club  
A. Reasoning process: Based on the given choices, the most inexpensive option for having a reception would likely be a church basement. Church basements are often available for a lower cost or even for free for members of the church or community. This makes them a budget-friendly option for hosting events such as wedding receptions. Final answer: B.  
Q: Where does the red cross respond to a storm? Choices: A. summer B. tropics C. blue shield D. weather report E. America  
A. Reasoning process: The correct answer is E. America. The Red Cross responds to storms in various locations, including the United States, where they provide assistance and support during natural disasters such as storms and hurricanes. Final answer: E.  

Table 19: The exemplars are selected on CommonSenseQA train set.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Iter-CoT (w/o label) Exemplars</th>
</tr>
</thead>
</table>
| **STQA** | Q: Is the Antarctic Peninsula south of the Equator?  
A: Reasoning process: The Antarctic Peninsula is located at the southernmost part of the Earth, which means it is in the southern hemisphere. Therefore, since the Antarctic Peninsula is located in the southern hemisphere, it is south of the Equator. Final answer: Yes. |
| | Q: Is lunch on the beach a good activity to spot the full circle of a rainbow?  
A: Reasoning process: First, to see a rainbow, you need to be outside in the rain or near a source of water. If you are on the beach, you are likely to be near a source of water, so that criterion is satisfied. Second, you need to have the sun shining. If you are having lunch on the beach, you are likely to be there during the daytime, when the sun is shining, so that criterion is also satisfied. Third, you need to have the right angle between the sun, your eyes, and the water droplets in the air. This is a bit harder to predict, but if the conditions are right, you should be able to see a rainbow. However, as mentioned in the previous answer, the full circle of a rainbow cannot usually be seen from ground level, and lunch on the beach occurs at midday when the sun is likely high in the sky. Therefore, it is unlikely that you would be able to see the full circle of a rainbow from the beach during lunch. Final answer: No. |
| | Q: Were any of the characters in Macbeth played by an Oscar-winning actor?  
A: Reasoning process: To answer the question, we need to research the actors who have played the characters in Macbeth and find out if any of them have won an Oscar. Some of the characters in Macbeth have been played by Oscar-winning actors. For instance, the character of Macbeth has been played by actors such as Laurence Olivier, Orson Welles, and Michael Fassbender, all of whom have won Oscars for their performances in other roles. Final answer: Yes. |
| | Q: Are monks forbidden from engaging in warfare?  
A: Reasoning process: Monks are members of religious orders that usually take vows of poverty, chastity, and obedience. These vows are typically associated with a life of spiritual contemplation and non-violent pursuits. However, there are examples of monks engaging in warfare, such as the Knights Templar during the Crusades and Buddhist Shaolin monks who have defended temples during conquests. These examples suggest that it is not always the case that monks are forbidden from engaging in warfare. Final answer: No. |
| | Q: Does Post Malone have a fear of needles?  
A: Reasoning process: We know that Post Malone is a person because the question states that he is a person. People can have fears because it is a common human trait to have fears of various things. The fear of needles is a common fear among people because needles can cause pain, discomfort, and anxiety for many people. However, the fact that Post Malone has many tattoos on his body suggests that he is comfortable with needles and therefore does not have a fear of needles. Additionally, the fact that the most common method of tattooing involves electric tattoo machines and multiple needles suggests that Post Malone has been exposed to needles many times and has not shown any fear or discomfort. Therefore, it is unlikely that Post Malone has a fear of needles. Final answer: No. |
| | Q: Can you get negative effects from a placebo?  
A: Reasoning process: Placebos are substances or treatments that have no therapeutic effect, and are often used in medical trials as a control group to compare the effects of a real treatment. Since placebos themselves have no active ingredients, it is unlikely for them to cause negative effects on their own. However, it is possible for a person to experience negative effects from a placebo due to the nocebo effect. This occurs when a person believes a treatment to be harmful, even though it has no active ingredients, and they experience negative side effects as a result of their expectations and beliefs about the treatment. It is possible for a person to experience negative effects from a placebo due to the nocebo effect. Final answer: Yes. |

Table 20: The exemplars are selected on StrategyQA train set.
**Table 21:** The exemplars are selected on Date Understanding train set.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>Iter-CoT(w/o label) Exemplars</th>
</tr>
</thead>
</table>
| Date Understanding | Q: This is the last day of 1899. What is the date one week ago from today in MM/DD/YYYY? Choices: A.12/24/1802 B.12/24/1899 C.12/17/1899 D.01/02/1900 E.03/12/1900 F.05/24/1900  
A: Reasoning process: To find the date one week ago from today, we need to subtract 7 days from the last day of 1899, which is December 31, 1899. Subtracting 7 days from December 31, 1899, we get December 24, 1899. Final answer: B.  
Q: Today is the last day of the first quarter of 2008. What is the date a month ago in MM/DD/YYYY? Choices: A.06/17/2018 B.06/19/2018 C.06/13/2018 D.05/23/2018 E.07/09/2018 F.06/18/2018  
A: Reasoning process: To find the date one month ago from the last day of the first quarter of 2008, we need to subtract one month from March 31, 2008. Since February only has 29 days in a leap year, we can calculate the answer is 02/29/2008 Final answer: D.  
Q: Jane and John married on Jan 2, 1958. Today is their golden wedding anniversary. What is the date yesterday in MM/DD/YYYY? Choices: A.01/01/2008 B.05/06/2004 C.01/05/2005 D.05/04/2004 E.02/15/2004 F.05/05/2004  
A: Reasoning process: To find the date one year ago from today, we need to subtract 1 year from the given date. Given that today is Jan 2, 2008, the correct answer is: 01/01/2008. Final answer: E.  
A: Reasoning process: To find the date one day ago from today, we need to subtract 1 day from the given date, which is July 9th, 1972.  
|
| | Final answer: A.  
A: Reasoning process: The US Thanksgiving in 2001 was on November 22nd. Adding one week to that date, we get November 29th Final answer: E.  
Q: The current local time is 3:02 pm of 5/4/2004. What is the date tomorrow in MM/DD/YYYY? Choices: A.05/05/1915 B.05/06/2004 C.01/05/2005 D.05/04/2004 E.02/15/2004 F.05/05/2004  
A: Reasoning process: The current local time is 3:02 pm of 5/4/2004. To find the date tomorrow, we need to add one day to the current date. Since it is already 3:02 pm, tomorrow’s date will be the same as the current date until 3:02 pm. After 3:02 pm, it will be the next day. Final answer: F.  
|

**Table 22:** The exemplars are selected on Last Letter Concatation(4) train set.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>Iter-CoT(w/o label) Exemplars</th>
</tr>
</thead>
</table>
| Letter(4) | Q: Take the last letters of the words in "Keith Ruiz" and concatenate them.  
A: Reasoning process: 1) The last letters of "Keith Ruiz" are "h" and "z". 2) Concatenating these letters gives us "hz". Therefore, the correct answer is: hz. Final answer: hz.  
Q: Take the last letters of the words in "Saul Moss" and concatenate them.  
A: Reasoning process: Let’s take the last letters of each word in "Saul Moss": - The last letter of "Saul" is "l". - The last letter of "Moss" is "s". Now, let’s concatenate them: - "]l + "s" = "ls" Final answer: ls.  
Q: Take the last letters of the words in "Morton Stanton" and concatenate them.  
A: Reasoning process: Taking the last letters of each word in "Morton Stanton" and concatenating them, we have: n (from Morton) n (from Stanton) Final answer: nn.  
Q: Take the last letters of the words in "Drew Barnett" and concatenate them.  
A: Reasoning process: Taking the last letters of the words "Drew" and "Barnett", we have "w" and "t" respectively. Concatenating them gives us "wt". Final answer: wt.  
|