Adaption-of-Thought: Learning Question Difficulty Improves Large Language Models for Reasoning

Mayi Xu[1], Yongqi Li[1], Ke Sun[1], Tieyun Qian[1,2,*]

[1] School of Computer Science, Wuhan University, China
[2] Intellectual Computing Laboratory for Cultural Heritage, Wuhan University, China
{xumayi,liyongqi,sunke1995,qty}@whu.edu.cn

Abstract

Large language models (LLMs) have shown excellent capability for solving reasoning problems. Existing approaches do not differentiate the question difficulty when designing prompting methods for them. Clearly, a simple method cannot elicit sufficient knowledge from LLMs to answer a hard question. Meanwhile, a sophisticated one will force the LLMs to generate redundant or even inaccurate intermediate steps for a simple question. Consequently, the performance of existing methods fluctuates among various questions.

In this work, we propose Adaption-of-Thought (ADOT), an adaptive method, to improve LLMs for the reasoning problem, which first measures the question difficulty and then tailors demonstration set construction and difficulty-adapted retrieval strategies for the adaptive demonstration construction. Experimental results on three reasoning tasks prove the superiority of our proposed method, showing an absolute improvement of up to 5.5% on arithmetic reasoning, 7.4% on symbolic reasoning, and 2.3% on commonsense reasoning.[1]

1 Introduction

Reasoning refers to the drawing of inferences through the use of reason. Solving reasoning tasks requires substantial knowledge and powerful analytical capabilities, which can narrow the gap between human and machine intelligence (Qiao et al., 2023). Previous small language models (SLMs) based methods often adopt the fine-tuning technique to enhance their performance. However, the reasoning ability of SLMs is naturally limited by the size of their parameters and fine-tuning requires abundant labor-intensive labeled data. Recently, large language models (LLMs) (Brown et al., 2020; Thoppilan et al., 2022; Chowdhery et al., 2023)

---

*Corresponding author.
[1]Our codes and implementation details are available at: https://github.com/NLPGM/AdoT
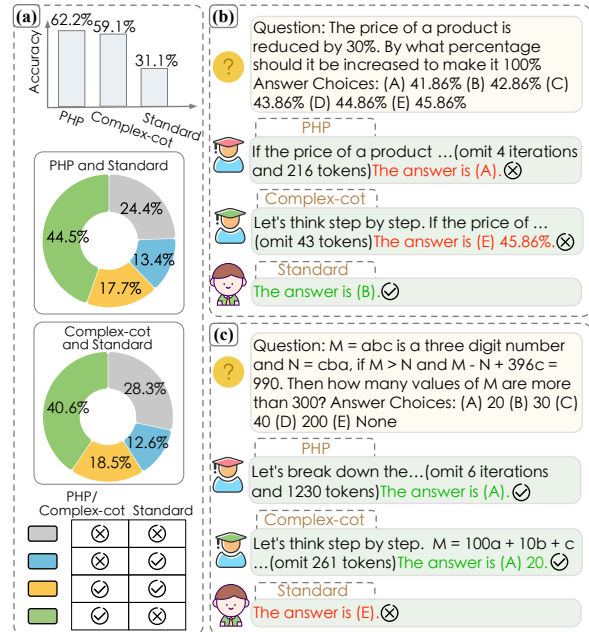

Figure 1: The results by different LLM based reasoning methods on the AQUA dataset.

have demonstrated significant efficacy in various tasks. Reasoning by prompting LLMs has also attracted wide attention. Typical methods include automatic demonstration construction (Zhang et al., 2023; Wan et al., 2023), rationale refinement (Fu et al., 2023), reasoning in code format (Chen et al., 2022), iteratively reasoning (Zheng et al., 2023; Xi et al., 2023), problem decomposition (Press et al., 2023; Zhou et al., 2023), problem planing (Wang et al., 2023a), and prompting ensemble (Wang et al., 2023b; Naik et al., 2023).

Despite the effectiveness of these methods, it is interesting that none of them can consistently perform the best on most datasets on three reasoning tasks as shown in Table 2. So what is the reason for such a phenomenon? Can we develop a method that performs generally well beyond the datasets and tasks? With these in mind, we choose two sophisticated prompting methods PHP (Zheng et al.,

2023) and COMPLEX-COT (Fu et al., 2023) and one simple method STANDARD (Brown et al., 2020) as examples to investigate their performance on the AQUA dataset. The results are shown in Fig. 1 (a).

Overall, PHP and COMPLEX-COT perform much better than STANDARD. However, the pie charts in Fig. 1 (a) show that there are up to 13.4% and 12.6% questions to which STANDARD can answer correctly but PHP and COMPLEX-COT cannot. A closer look reveals that the reason can be *the mismatch between the question difficulty and the complexity of the method*.

Given the easy question in Fig. 1 (b), PHP iterates four times and COMPLEX-COT builds the highly complex demonstration to interact with the LLM. The generated redundant information misleads the model to produce wrong answers. Meanwhile, STANDARD, which only contains the sample questions and their final answer in its demonstrations, gets the correct answer. On the contrary, given the hard question in Fig. 1 (c), PHP and COMPLEX-COT give correct answers, but STANDARD fails since its simple demonstration prevents it from drawing out sufficient knowledge from the LLM.

To address the mismatch problem, we propose a novel Adaption-of-Thought (ADOT) method to adaptively adjust the complexity of the prompting method to the question difficulty. Specifically, we first *present a difficulty measuring approach for questions* which computes the syntactic and semantic complexity of their rationales. We then *propose a demonstration set construction and a difficulty-adapted retrieval strategy* to adaptively construct reasonable demonstrations based on the difficulty of the questions. Extensive experimental results prove that our proposed ADOT method surpasses competitive baselines on all tasks. In summary, the main contributions of our work are as follows.

- We show the necessity of differentiating the question difficulty when solving the reasoning problems.

- We propose an ADOT method to adaptively adjust the complexity of the prompting so that we can elicit proper knowledge from LLMs based on the question difficulty.

- Our proposed method achieves the best performance on almost all datasets across three reasoning tasks.

2 Related Work

This section reviews the literature on reasoning.

Reasoning by Fine-tuning SLMs. SLMs-based reasoning methods (Rajani et al., 2019) need to fine-tune on task-oriented or domain-specific data to enhance their performance. For example, Talmor et al. (2020) trained RoBERTa (Liu et al., 2019) on both implicit pre-trained knowledge and explicit free-text statements to symbolic reasoning. Hendrycks et al. (2021) fine-tuned the GPT-2 (Radford et al., 2019) to generate full step-by-step solutions to arithmetic reasoning.

Though these approaches have shown better performance than traditional rule-based (Fletcher, 1985; Yuhui et al., 2010), symbolic-based (Liguda and Pfeiffer, 2012; Shi et al., 2015), and statistical-based (Koncel-Kedziorski et al., 2015a; Upadhyay et al., 2016) methods, their reasoning ability has been limited by the size of SLMs. Moreover, the fine-tuning of SLMs requires high-quality training data, which is quite labor-intensive.

Reasoning by Prompting LLMs. LLMs-based approaches have become prevalent in recent years and prompting is the main way to interact with LLMs. Many prompting methods have been proposed for reasoning problems (Chen et al., 2022; Wang et al., 2023a; Nathani et al., 2023; Wang et al., 2023b; Wan et al., 2023). The first attempt was made by Brown et al. (2020), which developed a ZERO-SHOT prompting method by adding a natural language description of the task in the prompt. Notably, Chain-of-Thought (COT) prompting (Wei et al., 2024), which performed the multi-step reasoning through step-by-step rationale examples, achieved excellent performances in reasoning. Kojima et al. (2022) found that adding instructions to encourage LLMs to automatically reason through some intermediate processes also contribute to improving performance. Subsequently, a series of studies optimized the COT prompting by decomposing the problem or refining the rationale and demonstration from various perspectives (Zheng et al., 2023; Zhang et al., 2023; Fu et al., 2023; Press et al., 2023; Zhou et al., 2023).

While these LLMs-based methods have made impressive progress, none of them has taken the difficulty of the question into consideration, leading to a mismatch between the problem and the solution. In contrast, our method can automatically adapt the demonstration to questions with varying difficulty.
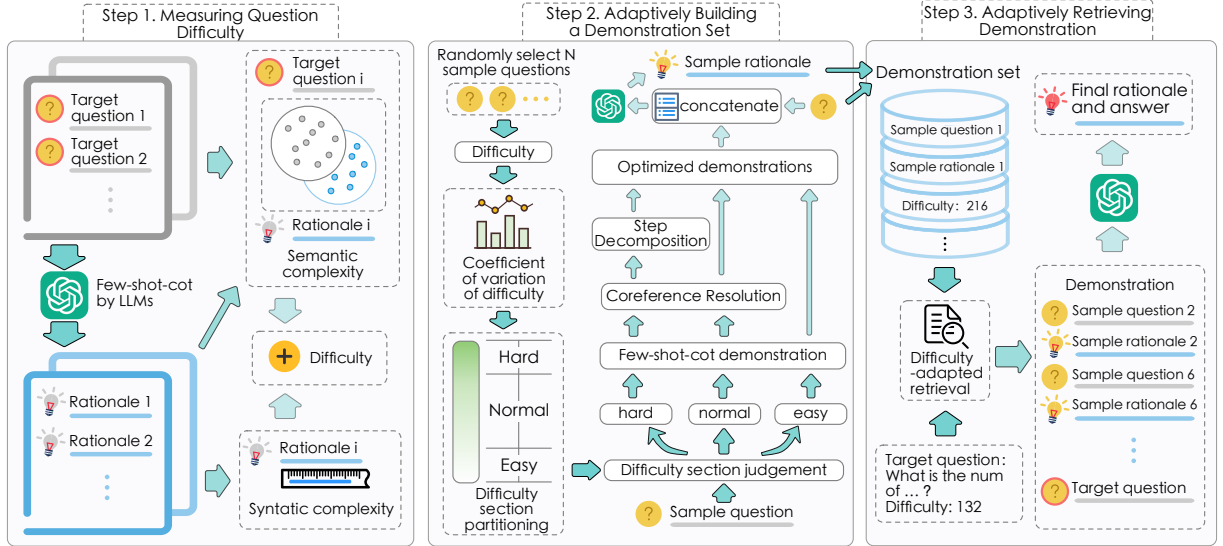
Figure 2: An overview of our proposed ADOT method.

3 Adaption-of-Thought Method

This section presents our proposed method.

3.1 Problem Definition

With the prevalence of LLMs, a popular paradigm to solve reasoning problems is inputting the prompt $\mathcal{T}$ and the target question q into an LLM $\mathcal{P}$. The probability to generate a rationale[2] r can be formulated as follows:

$$p(r \mid \mathcal{T}, q) = \prod_{i=1}^{|r|} \mathcal{P}\left(r^i \mid \mathcal{T}, q, r^{<i}\right) \qquad (1)$$

where $\mathcal{T}$ consists of demonstrations in the form of sample question-rationale pairs. r^i, $r^{<i}$, and $|r|$ denote the i-th token, the $i-1$ tokens before r^i, and the rationale length, respectively. The goal of our method is to automatically construct an adaptive $\mathcal{T}$ for each target question q based on its difficulty.

3.2 Method Overview

As analyzed in the introduction, a simple prompt cannot elicit sufficient knowledge from LLMs to answer a hard question, while a sophisticated one will force the LLM to generate redundant or even inaccurate intermediate steps for a simple question. To overcome the mismatch between the question difficulty and the prompting complexity, we propose an adaption-of-thought (ADOT) reasoning method, which consists of three modules. As shown in Fig. 2, ADOT first measures question

difficulty, and then adaptively builds a demonstration set based on the difficulty distribution of some randomly selected sample questions. Finally, it adaptively retrieves demonstrations from the set based on the difficulty of the target question.

3.3 Measuring Question Difficulty

The question difficulty often lies in how much effort needs to be made to answer it, since it is hard to measure the difficulty directly. In light of this, we propose to measure the question difficulty from its rationale. We first generate the rationales of different target questions using uniform demonstrations (Wei et al., 2024) to avoid the influence from different demonstrations. Then, we measure the difficulty of each question based on the complexity of its rationale, which can be further partitioned into syntactic complexity and semantic complexity. **Measuring syntactic complexity**. In applied linguistics, syntactic complexity refers to the range of forms that surface in language production and the degree of sophistication of such forms (Ortega, 2003). The higher syntactic complexity of a rationale indicates the more intricate articulation of the problem-solving process. Generally, the text length of a rationale is positively correlated with the syntactic complexity of the rationale (Ortega, 2003; Štajner and Saggion, 2013; Hwang and Polio, 2023). Hence we measure the syntactic complexity $Csyn_i$ of a rationale r_i by its text length as:

$$Csyn_i = length(r_i) \qquad (2)$$

Measuring semantic complexity. When reason-

---
[2]The rationales are also known as chain-of-thought or intermediate reasoning steps, which contain the final answers and the reasoning process towards them.

ing for a question, the question is hard to answer if a lot of extra knowledge needs to be introduced in its rationale. Generally, semantic words can be viewed as the carrier of knowledge. Hence, we measure the semantic complexity of a rationale r_i by calculating the newly added non-repetitive semantic words in r_i compared to q_i. The number of non-repetitive semantic words $sem(x)$ in the text x can be obtained through a simple procedure[3]. The semantic complexity $Csem_i$ of the i-th target question is then calculated as:

$$Csem_i = sem(r_i) - sem(q_i) \qquad (3)$$

Measuring comprehensive difficulty. The comprehensive difficulty d_i of the i-th target question q_i is measured by the combination of the syntactic complexity $Csyn_i$ and semantic complexity $Csem_i$ of its rationale r_i, which is shown as follows:

$$d_i = Csyn_i + Csem_i \qquad (4)$$

3.4 Adaptively Building a Demonstration Set

After getting the difficulty of target questions, we construct demonstrations with different complexity to match target questions.

Initially, we can take the original FEW-SHOT-COT demonstrations (Wei et al., 2024) to prompt LLMs to generate sample rationales for sample questions. However, we find that the sample rationales generated in this way are often not satisfying. Specifically, when the difficulty of sample questions gradually increases, the *coreference phenomena* and *over-complicated reasoning process* gradually appear in their sample rationales, which prevents LLMs from understanding sample rationales due to their weak semantics and confusing logic. Besides, even we have optimized the original demonstrations, LLMs cannot generate multiple sample rationales with significantly varying complexity for one sample question.

Based on the above observation, we propose to not only improve the quality of the original demonstrations in Wei et al. (2024) but also generate a high-quality demonstration set for a predefined N number of randomly selected sample questions[4]. This demonstration set will finally be used to solve the mismatch problem by providing LLMs with more choices of high-quality sample rationales

whose complexity is similar to the difficulty of the target questions. To this end, we first categorize the N sample questions into three difficulty sections (easy, normal, and hard). Then, we construct the optimized demonstrations for these questions in different sections. Our goal is to eliminate the coreference phenomena in both normal and hard sections and optimize the overcomplicated reasoning process in hard section.

Categorizing questions by difficulty. We first use the same approach described in Section 3.3 to measure the difficulty of the sample questions. In statistics, the coefficient of variation is a standardized measure of the dispersion of a distribution. Hence, the coefficient of variation v of all sample questions' difficulty can measure their dispersion, which is calculated as:

$$v = \frac{n}{\sum_{i=1}^{n} d_i} \times \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(d_i - \overline{d} \right)^2}, \qquad (5)$$

where $v \in [0,1]$, n and $\overline{d}$ are the number and the average difficulty of all sample questions, respectively. The higher v is, the more dispersed the distribution of difficulty will be. In other words, a higher v means more easy and hard questions are at the two ends of the distribution. Hence, we take half of the coefficient of variation, $p = v/2$, as the proportion of easy and hard questions. The rest $1\text{-}2 * p$ is the proportion of the normal question. Finally, the proportion of easy, normal, and hard questions is:

$$p : 1 - 2 * p : p \qquad (6)$$

Generating same rationales for easy sample questions. Generally, the sample rationales of easy sample questions are simple and easy for LLMs to understand. Hence, when generating rationales for them, we directly take the original FEW-SHOT-COT demonstrations.

Generating high-quality sample rationales for normal sample questions. For sample questions in the normal section, the coreference phenomena occur frequently in their rationales generated by the original FEW-SHOT-COT demonstrations. Coreference in linguistics refers to the phenomenon where two or more expressions in a text refer to the same person, place, thing, event, etc. For example, in the sentence "Tom met Harry in the park. He lent him his hat. He said that he liked it", expressions with the same color refer to the same person or thing.

Coreferences pose a big challenge on LLMs (Le and Ritter, 2023), especially for pronouns which

---

are hard to interpret due to their weak semantic meanings (Ehrlich, 1981; Zhang et al., 2019; Zhao et al., 2022). Coreference resolution aims to find the correct reference for a given anaphor, which is a key aspect of natural language understanding as it helps establish the relationship among different text components. The coreference resolution result for the former example is "Tom met Harry in the park. Tom lent Harry Tom's hat. Harry said that Harry liked Tom's hat", which is easy to understand.

To generate high-quality rationales for normal sample questions, we optimize the original FEW-SHOT-COT demonstrations through coreference resolution, which can be realized with the help of neuralcoref toolkit[5] and linguistic rules (Jurafsky and Martin, 2019). Coreference resolution maintains the form of rationales in FEW-SHOT-COT demonstrations and consolidates their logical clarity among the internal reasoning processes. In this way, LLMs will mimic the high-quality demonstrations and try to avoid using coreference in their generated content.

Generating high-quality sample rationales for hard sample questions. For sample questions in the hard section, over-complicated reasoning processes can often be observed in their sample rationales generated by the FEW-SHOT-COT demonstrations. Specifically, when the sample questions are hard, their sample rationales are often complex and challenging for LLMs to understand. The main reason is the lack of detailed connections among the reasoning steps in the over-complicated reasoning processes. Step decomposition aims to add steps in the reasoning process to bridge this gap.

Step 1: a / b = 3/4, then b = 4a / 3.	Step 1: a / b = 3/4, so b = 4a / 3.
Step 2: So 8a + 5(4a / 3) = 22.	Step 2: b = 4a / 3 and 8a + 5b = 22.
	Step 3: So 8a + 5(4a / 3) = 22.

Table 1: Step decomposition cases.

To generate high-quality rationales for hard sample questions, we optimize the original FEW-SHOT-COT demonstrations through both coreference resolution and step decomposition. As shown in Table 1, after adding Step 2 between Step 1 and Step 3, the logical clarity of the reasoning process is more clear. Step decomposition changes the form of rationales in FEW-SHOT-COT demonstrations and consolidates their logical clarity of the reasoning procedure. In this way, LLMs will mimic the high-quality demonstrations and try to generate a

---

[5]Avaiable at: https://github.com/huggingface/neuralcoref

more detailed reasoning process in their generated content for hard sample questions. More details and the whole optimized demonstrations are shown in Appendix C and Appendix I, respectively.

3.5 Adaptively Retrieving Demonstration

Although we have measured the difficulty of the question, it is hard for LLMs to grasp the subtle meaning of 'difficulty'. Therefore, we propose a difficulty-adapted retrieval method to show LLMs with several adaptive demonstrations whose complexity is similar to the difficulty of the target question. In this way, we not only elicit sufficient knowledge, but also avoid generating redundant or even inaccurate intermediate steps from LLMs.

Our difficulty-adapted retrieval method fetches adaptive demonstrations from the demonstration set constructed in Section 3.4. Remember that each demonstration corresponds to a sample question with a specific difficulty. Hence, we can leverage the difficulty of the sample question to enhance LLM's understanding of 'difficulty'. Specifically, we adaptively retrieve M demonstrations from the demonstration set. The difference between these demonstrations' complexity, i.e., their corresponding sample questions' difficulty, and the target question's difficulty is minimal. Then, we concatenate the target question behind the demonstrations and send them to the LLMs together. Therefore, LLMs will mimic the demonstrations and generate suitable rationales for the target questions, i.e., we finally solve the mismatch problem.

4 Experiment

4.1 Experimental Setup

Datasets: We adopt ten benchmark datasets across three reasoning tasks. **Arithmetic:** AQUA (Ling et al., 2017), GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021), MultiArith (Roy and Roth, 2015), AddSub (Hosseini et al., 2014), SingleEq (Koncel-Kedziorski et al., 2015b). **Symbolic:** Last Letters (Wei et al., 2024), Coin Flip (Wei et al., 2024).**Commonsense:** CSQA (Talmor et al., 2019), StrategyQA (Geva et al., 2021). Detailed statistics, including average difficulty and some parameters, are shown in Appendix D.

Baselines: We compare with thirteen state-of-the-art (sota) methods, including five *Zero-shot* setting method: ZERO-SHOT, ZERO-SHOT-COT (Kojima et al., 2022), PS+ (Wang et al., 2023a), AUTO-COT (Zhang et al., 2023), COSP (Wan et al., 2023)

Method	Arithmetic						Symbolic		Commonsense		Avg.
	AQuA	GSM8K	SVAMP	AddSub	MultiArith	SingleEq	Last Letters	Coin Flip	CSQA	StrategyQA	
ZERO-SHOT	55.1	76.4	81.9	92.9	95.0	95.7	57.6	38.4	72.9	58.4	72.4
ZERO-SHOT-COT	53.1	78.3	80.0	86.3	94.3	90.2	78.0	69.8	67.0	55.9	75.3
PS+	56.3	73.6	81.4	87.3	92.7	96.3	76.4	74.0	69.3	46.6	75.4
AUTO-COT	48.0	78.5	82.3	94.7	98.0	96.5	78.2	98.8	71.9	65.9	81.3
COSP	54.3	76.5	82.0	93.2	97.0	95.9	79.4	98.0	72.6	66.3	81.5
STANDARD	31.1	35.0	80.1	91.6	87.5	91.7	3.8	54.8	78.9	62.9	61.7
POT	54.1	71.6	85.2	-	-	-	-	-	-	-	-
FEW-SHOT-COT	57.1	78.2	83.9	94.9	97.7	96.3	77.4	98.2	77.7	68.3	83.0
COMPLEX-COT	59.1	79.3	79.8	90.9	98.5	97.6	85.2	96.0	73.5	62.8	82.2
SELF-ASK	58.3	74.5	85.3	94.4	96.2	97.0	71.2	99.2	77.1	71.4	82.5
LEAST-TO-MOST	53.5	75.5	84.6	93.9	96.3	94.3	81.6	97.8	78.3	70.5	82.7
SP	59.1	76.0	81.6	86.3	98.7	88.8	-	-	-	-	-
PHP	62.2	79.1	84.9	91.9	98.7	93.7	76.6	99.0	78.6	**71.8**	83.7
ADOT (Ours)	**67.7**	**79.8**	**86.4**	**95.7**	**99.3**	**97.8**	**92.6**	**100.0**	**81.2**	70.7	**87.1**

Table 2: Main experimental results. **Bold**: the best score. <u>Underline</u>: the second best score. "-": not applicable.

and eight *Few-shot* setting methods: STANDARD (Brown et al., 2020), FEW-SHOT-COT (Wei et al., 2024), POT (Chen et al., 2022), COMPLEX-COT (Fu et al., 2023), SELF-ASK (Press et al., 2023), LEAST-TO-MOST (Zhou et al., 2023), SP (Xi et al., 2023), PHP (Zheng et al., 2023). More details are shown in Appendix E. Methods like SELF-CONSISTENCY (Wang et al., 2023b), DIV-SE and IDIV-SE (Naik et al., 2023) must generate multiple rationales then ensemble them, i.e., their settings are different from our method and all baselines. Hence, we follow all previous works to exclude them for fair comparison.

Implementation details: For fair comparison, we adopt the `gpt-3.5-turbo-0613` engine of GPT-3.5 for all methods in the main experiment. The only exception is POT (Chen et al., 2022). We directly take its results since POT requires a code-davinci-002 engine, which is unavailable now. The temperature is set to 0 to ensure the stability and repeatability. We employ accuracy (%) as the metric. The demonstration set size N for all datasets is set to 50 uniformly without further tuning. A parameter study for N is given in section 4.6. The number of demonstrations M for different datasets is consistent with previous methods (Zhang et al., 2023; Wei et al., 2024; Press et al., 2023; Zhou et al., 2023; Xi et al., 2023; Zheng et al., 2023)

4.2 Main Experimental Results

Table 2 presents the main experimental results. Compared to the baselines, our proposed ADOT not only obtains the average sota performance on

all ten datasets of three tasks but also performs the best on almost each dataset. This demonstrates that our ADOT consistently performs better than all baselines, which cannot be achieved by any of the previous sota methods. We argue that the reason for this phenomenon is that previous methods are only good at answering questions in a specific difficulty range. We now analyze task by task.

(1) For the arithmetic task, the most challenging dataset is AQUA, with an average difficulty of 117.4. The representative complex methods, PHP and COMPLEX-COT, yield the previous best and second-best performance on it. The least challenging dataset for this task is AddSub (average difficulty: 36.2), and the simple method FEW-SHOT-COT produces the optimal baseline results on it. These results clearly demonstrate the importance of matching between prompting complexity and question difficulty.

(2) For the commonsense task, the simple STANDARD and the complex PHP perform the best on easy CSQA dataset (average difficulty: 36.1) and hard StrategyQA dataset (average difficulty: 49.0), which also proves the significance of matching.

(3) For the symbolic task, COMPLEX-COT and PHP perform well on Last Letter and Coin Flip, respectively. Meanwhile, simple STANDARD and ZERO-SHOT, perform the worst on them, respectively. The contrast performance between complex and simple methods shows that the mismatching will lead to terrible performance.

In summary, our ADOT is a widely-adaptive and practically applicable method with superior perfor-

Method	Arithmetic						Symbolic		Commonsense		AVG
	AQuA	GSM8K	SVAMP	AddSub	MultiArith	SingleEq	Last Letters	Coin Flip	CSQA	StrategyQA	
ADOT	**67.7**	**79.8**	**86.4**	95.7	**99.3**	**97.8**	**92.6**	**100.0**	**81.2**	**70.7**	**87.1**
w/o SynComp	63.4	79.4	86.0	**95.9**	98.8	96.7	91.2	98.6	80.2	70.5	86.1
w/o SemComp	62.6	79.0	84.6	95.2	98.0	97.6	90.8	98.4	79.7	69.8	85.6
w/o SD	65.1	79.2	85.1	95.4	98.5	**97.8**	90.6	98.4	80.6	69.6	86.0
w/o CR	63.0	78.7	84.9	94.9	97.8	97.2	81.6	99.0	79.1	68.8	84.5
w/o All	57.1	78.2	83.9	94.9	97.7	96.3	77.4	98.2	77.7	68.3	83.0

Table 3: Results for ablation study. "w/o": removing the corresponding strategy. **Bold**: the best score.
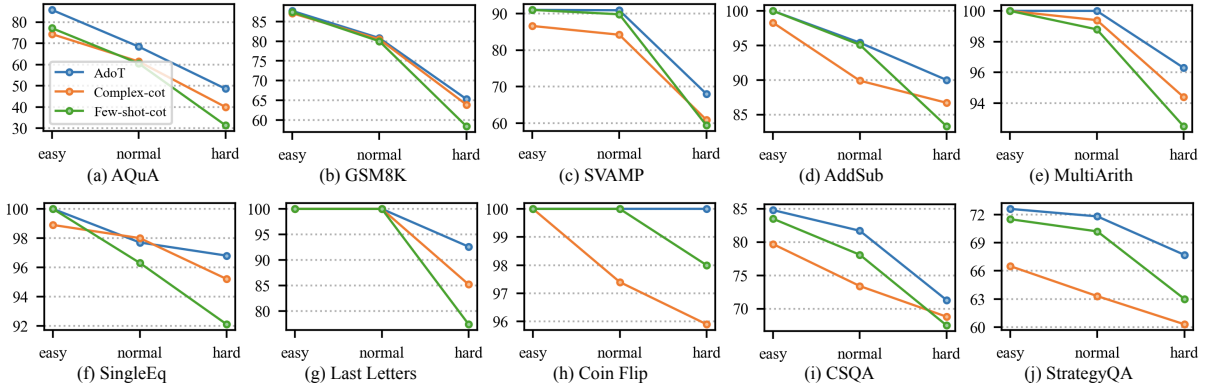

Figure 3: Accuracy (%) on different difficulty sections.

mance for all types of reasoning problems.

4.3 Ablation Study

We conduct an ablation study to examine how different components of our method influence the performance, including: the syntactic complexity (SynComp), semantic complexity (SemComp), step decomposition (SD), coreference resolution (CR), and all these strategies (All). The results are shown in Table 3. The "w/o CR" means removing both SD and CR strategies since the former is carried out based on the latter.

From Table 3, we can find that all components are essential to ADOT. Both SynComp and SemComp contribute to the performance improvement in three tasks, and the contribution of SemComp is more significant than SynComp. The main reason is that SemComp emphasizes the knowledge-level complexity and the reasoning problems are more sensitive to knowledge.

Remember that SD and CR strategies are proposed for constructing a high-quality demonstration set, and the results in Table 3 prove their effectiveness on all three tasks. Hence, high-quality demonstrations are critical to high performance.

After removing all strategies, the performance of the ' w/o All' variant dramatically drops by a percentage of 4.1 on average, showing that the com-

bination of these strategies is crucial to our method.

4.4 Adaptability for Different Difficulty

To examine the adaptability on questions with different difficulty, we investigate the accuracy of the typical simple method FEW-SHOT-COT, the typical complex method COMPLEX-COT, and ADOT in three difficulty sections[6]. Fig. 3 shows the results.

Generally, the simple FEW-SHOT-COT performs better than the complex COMPLEX-COT in the easy section, and the opposite result is found in the hard section. All these show that the mismatch problem is worth pondering when answering a question. Compared with the above methods, ADOT can outperform them in easy, normal, and hard sections in most cases due to the adaptive ability.

Moreover, the accuracy on easy section is also much higher than that on normal section. The same case exists in normal section and hard section. These results prove that our difficulty measuring and difficulty section partitioning strategies effectively build the connection between the question and the prompting method based on the difficulty.

---

[6]The difficulty sections of target questions are kept the same as that of sample questions measured in Section 3.4.
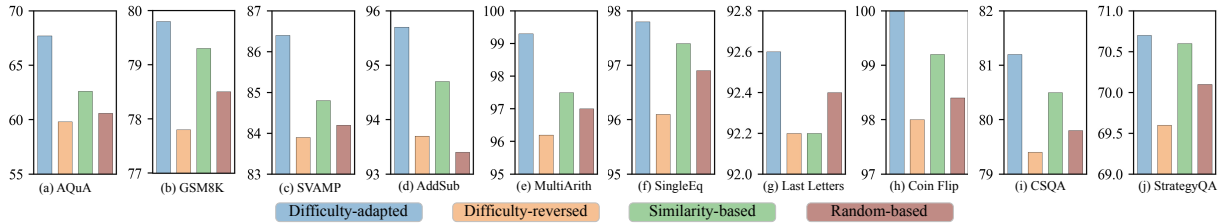
Figure 4: Accuracy (%) of different retrieval methods.

4.5 Effectiveness of Retrieval Method

To examine the effectiveness of our difficulty-adapted retrieval method, we compare it with three retrieval methods and show the results in Fig. 4. (1) The difficulty-reversed method retrieves M demonstrations where the sample questions have the biggest difficulty gap to the target questions. (2) The similarity-based method retrieves M demonstrations where the sample questions have the closest text similarity to the target questions. We adopt the ALL-MPNET-BASE-V2[7] to obtain the question embedding and leverage the cosine similarity to calculate the text similarity. (3) The random-based method randomly retrieves M demonstrations.

From Fig. 4, it is clear that our difficulty-adapted method achieves the best performance on all datasets. On the contrary, the difficulty-reversed method always performs worse than other methods, including the random-based one. These results prove that the mismatch between the question difficulty and the prompting complexity will hinder eliciting proper knowledge from LLMs. Furthermore, we believe that the idea of difficulty-adapted retrieval can be applied to more domains since the mismatch problem may generally exist.

4.6 The Parameter Study of Demonstration Set Size

We explore the effect of different demonstration set sizes on the largest arithmetic dataset, GSM8K. Fig. 5 shows the experimental results.

Small demonstration set sizes, such as 10, obviously prevent our ADOT method from performing well. When the size is small, the difficulty-adapted retrieval method of ADOT may not retrieve proper demonstrations whose complexity is similar to the target question. As the demonstration set size increases to no less than 30, the performance of our

method is stable and consistently better than that of the previous sota method.

In summary, our ADOT method will stably achieve high performance while setting a demonstration set size of no less than 30.
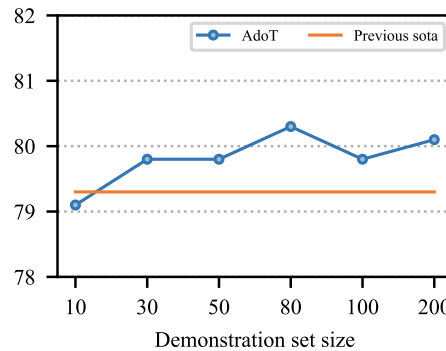
Figure 5: Accuracy (%) on different demonstration set size.

Figure 6: Results for case study. **Bold**: syntactic complexity. <u>Underline</u>: semantic complexity.

4.7 Case Study

We present an easy and a hard question for the case study, where PHP and COMPLEX-COT are sophisticated methods, and FEW-SHOT-COT and STANDARD are simple ones.

As Fig. 6 shows, simple methods, which contain sample questions with low syntactic and semantic

---

[7]Available at: https://huggingface.co/sentence-transformers/all-mpnet-base-v2. It provides the best quality on sentence embeddings and semantic search according to their report of Sentence-Transformers: https://www.sbert.net/index.html.

complexity, always generate straightforward reasoning processes. Meanwhile, sophisticated methods, which contain sample questions with high syntactic and semantic complexity, tend to generate complex reasoning processes. Hence, simple-/sophisticated methods fail to solve the hard/easy question due to the mismatch between the question difficulty and the prompting complexity. In contrast, our ADOT method can well address this mismatch problem by adaptively retrieving sample questions with proper syntactic and semantic complexity. Consequently, it can elicit proper knowledge no matter the question is easy or hard.

4.8   Supplement Experiments

We conduct more supplement experiments to provide a more detailed understanding of our method.

(1) The computational efficiency experiments and analysis are shown in Appendix F. In summary, our method improves reasoning performance and maintains high computational efficiency.

(2) To investigate the generalization of the methods on different LLMs with different sizes, we conduct experiments on the Llama2-chat family (Touvron et al., 2023). The detailed results and analysis are shown in Appendix G. In short, our ADOT can generally outperform the baselines regardless of the LLM sizes.

(3) We build multiple optimized demonstrations with different words and structures to analyze the sensitivity of our method to them in Appendix H. Overall, keeping the meaning of the sample rationales invariable and the answer format consistent are essential to maintain stable performance.

5   Conclusion

In this paper, we find the mismatch between the question difficulty and the prompting complexity will hinder eliciting proper knowledge from LLMs. To solve this problem, we propose a novel ADOT method to adaptively adjust the complexity of the promoting method based on the difficulty of the target question. We first propose a difficulty measurement strategy. Then, we construct a demonstration set through the optimized demonstrations. Finally, we present a difficulty-adapted retrieval strategy to adaptively retrieve reasonable demonstrations based on the difficulty of the target question. A series of experiments prove that the mismatch problem is critical in reasoning and our method is effective on this problem. Consequently, our

ADOT method achieves a new state-of-the-art performance on various reasoning tasks.

Limitations

In our ADOT method, adaptively constructing a demonstration set requires some unlabeled sample questions. Previous works (Zhang et al., 2023; Wan et al., 2023) obtain these unlabeled sample questions from the test set directly, which is not practical and reasonable as we illustrate in Appendix B. Hence, we obtain these unlabeled sample questions from the unlabeled question corpus. Although the unlabeled question corpus is easy to obtain from the collections of history queries, it may take some external time consumption to collect such unlabeled questions.

Ethics Statement

Our work is entirely at the methodological level, which means that there will not be any negative social impacts.

Acknowledgments

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Kate Ehrlich. 1981. Search and inference strategies in pronoun resolution: An experimental study. In *19th*

Annual Meeting of the Association for Computational Linguistics, pages 89–93.

Charles R Fletcher. 1985. Understanding and solving arithmetic word problems: A computer simulation. *Behavior Research Methods, Instruments, & Computers*, 17(5):565–571.

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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hyun-Bin Hwang and Charlene Polio. 2023. Text length effects on the reliability of syntactic complexity indices. *Research Methods in Applied Linguistics*, 2(3):100085.

Dan Jurafsky and James H Martin. 2019. Speech and language processing (3rd (draft) ed.).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015a. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015b. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, page 585–597.

Nghia T Le and Alan Ritter. 2023. Are large language models robust zero-shot coreference resolvers? *arXiv preprint arXiv:2305.14489*.

Christian Liguda and Thies Pfeiffer. 2012. Modeling math word problems with augmented semantic networks. In *International Conference on Application of Natural Language to Information Systems*, page 247–252. Springer.

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 (Volume 1: Long Papers)*, pages 158–167.

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

Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. 2023. Diversity of thought improves reasoning abilities of large language models. *arXiv preprint arXiv:2310.07088*.

Deepak Nathani, David Wang, Liangming Pan, and William Wang. 2023. Maf: Multi-aspect feedback for improving reasoning in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6591–6616.

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied linguistics*, 24(4):492–518.

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.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5368–5393.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shuming Shi, Yuehui Wang, Chin-Yew Lin, Xiaojiang Liu, and Yong Rui. 2015. Automatically solving number word problems by semantic parsing and reasoning. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, page 1132–1142.

Sanja Štajner and Horacio Saggion. 2013. Readability indices for automatic evaluation of text simplification systems: A feasibility study for spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 374–382.

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.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 33:20227–20237.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

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

Shyam Upadhyay, Ming-Wei Chang, Kai-Wei Chang, and Wen-tau Yih. 2016. Learning from explicit and implicit supervision jointly for algebra word problems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, page 297–306.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3493–3514.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2609–2634. Association for Computational Linguistics.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Jia Liu, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2023. Self-polish: Enhance reasoning in large language models via problem refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11383–11406.

Ma Yuhui, Zhou Ying, Cui Guangzuo, Ren Yun, and Huang Ronghuai. 2010. Frame-based calculus of solving arithmetic multi-step addition and subtraction word problems. In *2010 Second International Workshop on Education Technology and Computer Science*, volume 2, page 476–479. IEEE.

Hongming Zhang, Yan Song, Yangqiu Song, and Dong Yu. 2019. Knowledge-aware pronoun coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 867.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Xinran Zhao, Hongming Zhang, and Yangqiu Song. 2022. Pcr4all: A comprehensive evaluation benchmark for pronoun coreference resolution in english. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5963–5973.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*.

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

A Detailed Steps of Semantic Complexity Measurement

As mentioned in section 3.4, in order to describe the detailed steps of semantic complexity measurement, we list all the processing steps in Table 4, where the step 3-5 can be realized by nltk toolkit[8].

Step 1: Convert text to lowercase and remove extra spaces.
Step 2: Remove special characters and punctuation marks.
Step 3: Lemmatization.
Step 4: Tokenizing.
Step 5: Removing stop words.
Step 6: Remove blank characters.
Step 7: Remove numbers.
Step 8: Remove repetitive tokens.
Step 9: Calculate the number of left tokens.

Table 4: Detailed steps of semantic complexity measurement.

B Unlabeled Question Corpus

As mentioned in Section 3.4, an unlabeled question corpus is used to construct a demonstration set. Previous methods, such as AUTO-COT (Zhang et al., 2023) and COSP (Wan et al., 2023), directly take all unlabeled questions from test data as unlabeled question corpus. Obviously, this setting is not practical. First, we cannot obtain all test data immediately before testing them in actual deployment. Second, the distributions of test data may leak. Hence, we argue that taking unlabeled questions from outside of test data is more reasonable since unlabeled questions are typically cheap and easy to obtain, for example, via collections of history queries.

Next, we provide a detailed description of how to acquire unlabeled question corpus outside the test data, which will help future works follow this more sensible practice quickly. For AQuA[9], GSM8K[10], SVAMP[11], AddSub, MultiArith, SingleEq, StrategyQA[12] datasets, their unlabeled question corpus are derived from their train set. Note that the SVAMP, AddSub, MultiArith, and SingleEq dataset share a unlabeled question corpus since they are similar in format. The Last Letters dataset only contains test data, which is generated by randomly concatenating the first and last name from

the top one-thousand in name census data[13]. Hence, we construct its unlabeled question corpus by randomly concatenating 100 first and last name from the top two-thousand to top one-thousand. For the Coin Flip dataset, which only contains test data as well, we leverage `gpt-3.5-turbo-0613` to imitate the eight sample questions in Coin Flip to generate 100 unlabeled questions as unlabeled question corpus. For CSQA[14] dataset, previous works test on its train set since the labels of its test set cannot be downloaded. Hence, we adopt its unlabeled test set as an unlabeled question corpus.

Note that our method can also adopt the test data as an unlabeled question corpus, which will gain better performance. For instance, increasing 2.2% and 2.0% in AQuA (67.7% → 68.9%) and GSM8K (79.8% → 81.8%), respectively.

C Coreference Resolution and Step Decomposition

Conducting coreference resolution for FEW-SHOT-COT demonstrations is easy, which can be realized with the help of neuralcoref[15] toolkit—a tool which is proposed to conduct coreference resolution and widely adopted in the NLP community. Furthermore, there are some linguistic rules (Jurafsky and Martin, 2019) for coreference resolution, which is easy to follow.

The goal of the step decomposition is to let the LLM know how to decompose a complex rationale into an easy-to-understand rationale by adding more steps. To this end, we first show LLM a pair of (a compact rationale of the original rationale, the original rationale). The compact rationale is formed by the step deletion, which is the reverse process of step decomposition. Specifically, given a rationale with N steps, we randomly delete $\frac{N-1}{2}$ steps chosen from the first step to the $(N-1)$-th step, where the last step should be kept because it is often the final answer. By imitating the process in [a compact rationale of the original rationale $\overset{step\ decomposition}{\longrightarrow}$ the original rationale], LLMs will conduct [the original rationale $\overset{step\ decomposition}{\longrightarrow}$ a decomposed rationale of the original rationale]. Then, we manually modify or remove some unreasonable steps and make all demonstrations highly consistent in format.

D Dataset Statistics

As mentioned in Section 4.1, the detailed dataset statistics and the number of demonstrations M are shown in Table 5.

Datasets	Task	#Samples	Avg. diff	Answer	M
AQUA	Arithmetic	254	117.4	Option	4
GSM8K	Arithmetic	1319	79.8	Number	8
SVAMP	Arithmetic	1000	43.1	Number	8
AddSub	Arithmetic	395	36.2	Number	8
MultiArith	Arithmetic	600	48.9	Number	8
SingleEq	Arithmetic	508	38.1	Number	8
Last Letters	Symbolic	500	40.7	String	4
Coin Flip	Symbolic	500	50.0	Yes/No	8
CSQA	CS	1221	36.1	Option	7
StrategyQA	CS	2290	49.0	Yes/No	6

Table 5: Dataset statistics.

E Baselines

As mentioned in Section 4.1, this section presents the introduction of baselines.

- ZERO-SHOT (Brown et al., 2020) inputs the question to LLMs directly without any instructions and demonstrations.

- ZERO-SHOT-COT (Kojima et al., 2022) concatenates a simple but effective instruct "Let's think step by step" after question, which can activate the inherent multi-step reasoning capability of LLMs.

- PS+ (Wang et al., 2023a) devises a plan to divide the entire task into smaller subtasks, and then carries out the subtasks according to the plan.

- AUTO-COT (Zhang et al., 2023) samples questions with diversity and automatically generates rationales to construct demonstrations.

- COSP (Wan et al., 2023) constructs demonstrations from the LLM zero-shot outputs via carefully designed criteria that combine consistency, diversity and repetition.

- STANDARD (Brown et al., 2020) concatenates some sample questions and pure answers behind the target question.

- FEW-SHOT-COT (Wei et al., 2024) adopts some questions and manual chain-of-thought as demonstrations.

- POT (Chen et al., 2022) uses LLMs to generate Python programming language statements, and get the final answer by executing the generated program on a Python interpreter.

- COMPLEX-COT (Fu et al., 2023) adopts some demonstrations with higher reasoning complexity, i.e., chains with more reasoning steps.

- LEAST-TO-MOST (Zhou et al., 2023) breaks down a complex problem into a series of simpler subproblems and then solves them in sequence.

- SP (Xi et al., 2023) facilitates LLM's reasoning by guiding it to progressively refine the given problems to be more comprehensible and solvable.

- SELF-ASK (Press et al., 2023) asks itself (and answers) follow-up questions before answering the initial question, which will narrow the compositionality gap that models can correctly answer all sub-problems but not generate the overall solution.

- PHP (Zheng et al., 2023) enables automatic multiple interactions between users and LLMs by using previously generated answers as hints to progressively guide toward the correct answers.

F Computational Efficiency

As mentioned in section 4.8, we investigate the time-consumption and token-consumption of different methods to explore computational efficiency.

We compare our ADOT method with five best performing methods (PHP, FEW-SHOT-COT, LEAST-TO-MOST, SELF-ASK, COMPLEX-COT) in time-consumption. Closed-source LLMs (e.g., gpt-3.5-turbo-0613) cannot be downloaded, and interacting with them through API is susceptible to network conditions and other remote server-independent factors. To ensure a fair comparison, we investigate the time-consumption on the open-source LLMs llama2-7b-chat locally.

All the experiments are done in the same environment on Nvidia A800 GPU. The comparison results on three tasks are shown in Fig. 7, which demonstrates that the time-consumption of our method is close to some simple methods (such as FEW-SHOT-COT) and much less than some complex methods (such as PHP). For instance, our method is 1.27
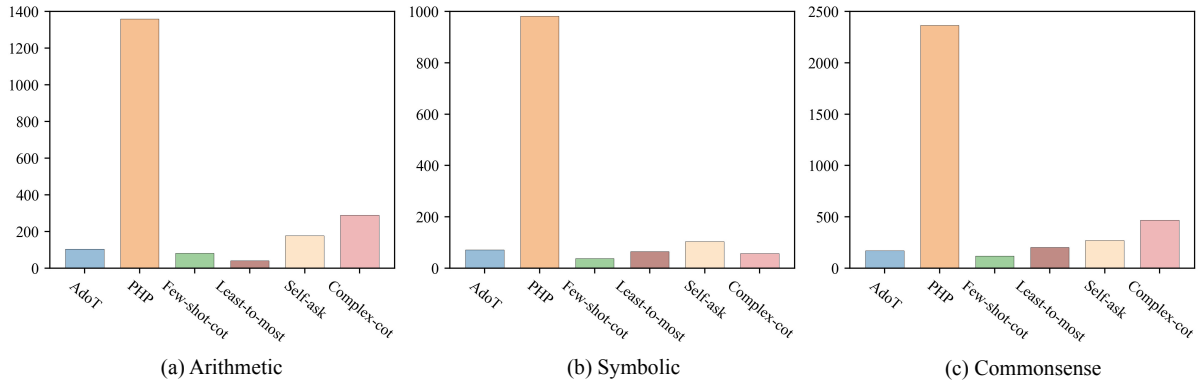
Figure 7: The time-consumption (seconds) of different methods.

Method	Arithmetic						Symbolic		Commonsense		AVG
	AQuA	GSM8K	SVAMP	AddSub	MultiArith	SingleEq	Last Letters	Coin Flip	CSQA	StrategyQA	
COSP	621.1	555.5	379.8	303.0	374.9	315.4	175.4	321.9	557.6	675.5	428.0
FEW-SHOT-COT	90.4	73.9	40.3	33.6	45.8	35.3	36.2	42.2	73.9	71.9	54.4
PHP	242.3	194.4	104.0	82.2	104.4	82.9	77.0	92.8	176.6	192.6	134.9
ADOT (Ours)	197.7	158.0	102.4	82.2	105.8	91.8	68.2	90.3	144.9	131.3	117.3

Table 6: The token-consumption of different methods.

times (104.7/82.2 seconds) less than FEW-SHOT-COT and 12.97 times (1358.3/104.7 seconds) faster than PHP on the average of six datasets in the arithmetic task.

There are two main reasons for the less time-consumption of our method. Frist, methods like PHP require iterative interaction with LLMs, while our method avoids this complex process. Second, the time-consumption is positively correlated with the length of input text. Our method only retrieves long demonstrations when the question is hard. However, the demonstrations of COMPLEX-COT and SELF-ASK are always long, so their time-consumption is usually much longer than ours.

From the perspective of token-consumption, we select three representative methods for comparison, including the previous best *Zero-shot* setting method COSP, the previous best *Few-shot* setting method PHP, and the vanilla chain-of-though pipeline FEW-SHOT-COT. The experimental results are shown in Table 6. Compared to FEW-SHOT-COT, our method produces 62.91 additional tokens (on average of all datesets). Compared to the COSP and PHP methods, our method will reduce 310.7 and 17.6 additional tokens, respectively.

Overall, our method improves reasoning performance and maintains high computational efficiency.

G Method Generalization on LLMs with Different Sizes

As mentioned in section 4.8, to explore whether our method is sufficiently generalizable on LLMs with different sizes or not, we experiment on `llama2` with three sizes. Table 7, Table 8, Table 9 shows the performance of our ADOT method and other baselines on `llama2-7b-chat`, `llama2-13b-chat`, and `llama2-70b-chat`, respectively[16]. Specifically, we do not reproduce the PoT on `llama2` since it requires powerful code-completion capability while `llama2` series LLMs are not good at.

Based on the experimental results, we can observe that ADOT can achieve sota performance in most cases regardless of datasets and LLM sizes. The average performance of ADOT improves 4.2%, 3.7%, and 6.8% on the previous best average performance. Hence, we argue that our ADOT method is effectively generalized on LLMs with different sizes. Furthermore, the results demonstrate that the performance of most methods increases as the LLM size increases. Generally, the larger the model is, the more knowledge it can store and the better reasoning ability it will achieve.

---

[16]The experiments are conduct on NVIDIA A800 GPU.

Method	Arithmetic						Symbolic		Commonsense		AVG
	AQuA	GSM8K	SVAMP	AddSub	MultiArith	SingleEq	Last Letters	Coin Flip	CSQA	StrategyQA	
ZERO-SHOT	23.6	17.1	46.0	70.9	36.7	70.5	0.0	32.4	56.3	55.7	40.9
ZERO-SHOT-COT	27.6	24.1	47.2	65.1	60.3	66.5	1.4	27.8	57.3	55.5	43.3
PS+	24.4	21.3	39.5	52.9	43.8	55.5	1.0	7.8	49.8	33.7	33.0
AUTO-COT	25.6	23.1	49.6	65.3	_72.8_	68.1	10.6	49.8	65.1	57.9	48.8
COSP	26.4	20.9	46.2	67.1	55.3	68.9	0.0	68.8	62.2	60.2	47.6
STANDARD	20.5	5.3	36.2	61.5	5.8	54.1	0.0	50.6	62.1	57.7	35.4
FEW-SHOT-COT	26.4	22.7	54.1	72.2	64.7	69.7	_16.8_	**74.8**	65.9	63.0	_53.0_
COMPLEX-COT	23.2	**28.3**	48.3	63.0	**76.7**	68.3	12.8	57.6	46.8	58.3	48.3
SELF-ASK	_28.0_	18.1	_54.6_	_76.2_	44.3	_75.2_	11.0	65.4	_67.1_	**63.8**	50.4
LEAST-TO-MOST	23.6	22.4	51.7	70.9	62.7	66.5	15.4	56.2	66.5	_62.8_	49.9
SP	23.2	20.3	46.7	67.1	58.2	69.7	-	-	-	-	-
PHP	26.4	14.2	46.4	64.3	53.5	67.7	0.0	69.4	60.2	60.4	46.3
ADOT (Ours)	**28.7**	_24.4_	**56.6**	**77.7**	64.5	**77.4**	**48.8**	_65.6_	**67.6**	60.9	**57.2**

Table 7: Main experimental results on llama2-7b-chat. **Bold**: the best score. Underline: the second best score.

Method	Arithmetic						Symbolic		Commonsense		AVG
	AQuA	GSM8K	SVAMP	AddSub	MultiArith	SingleEq	Last Letters	Coin Flip	CSQA	StrategyQA	
ZERO-SHOT	23.2	31.4	49.6	73.2	47.8	73.4	11.0	51.6	55.2	54.3	47.1
ZERO-SHOT-COT	_29.5_	35.2	47.9	68.6	73.2	68.7	29.8	47.8	54.5	53.8	50.9
PS+	20.1	32.9	47.4	60.3	62.5	62.8	17.8	8.6	50.6	32.7	39.6
AUTO-COT	26.8	37.1	57.6	67.6	_87.0_	76.4	18.2	85.4	61.3	59.6	57.7
COSP	28.0	15.9	56.8	79.0	76.5	80.9	_31.4_	49.2	66.2	64.6	54.9
STANDARD	20.5	4.9	22.6	53.9	9.0	50.6	0.0	50.6	65.8	63.7	34.2
FEW-SHOT-COT	26.8	37.1	60.8	77.2	86.0	**83.3**	24.8	77.2	_69.4_	_67.2_	_61.0_
COMPLEX-COT	25.6	**44.6**	57.4	66.1	85.2	77.2	22.4	**88.2**	54.1	42.7	56.4
SELF-ASK	28.0	32.3	_61.3_	_83.3_	60.7	82.5	15.2	_83.4_	68.1	**67.4**	58.2
LEAST-TO-MOST	25.2	34.9	60.8	77.0	72.3	80.3	13.0	77.4	68.4	66.8	57.6
SP	23.2	33.4	56.9	68.9	68.2	71.9	-	-	-	-	-
PHP	24.4	29.1	55.9	66.1	80.8	80.3	0.0	81.2	62.2	66.9	54.7
ADOT (Ours)	**31.1**	_39.5_	**62.3**	_79.5_	**89.0**	**85.2**	**46.0**	79.0	**69.5**	65.8	**64.7**

Table 8: Main experimental results on llama2-13b-chat. **Bold**: the best score. Underline: the second best score.

Method	Arithmetic						Symbolic		Commonsense		AVG
	AQuA	GSM8K	SVAMP	AddSub	MultiArith	SingleEq	Last Letters	Coin Flip	CSQA	StrategyQA	
ZERO-SHOT	33.1	49.0	67.4	58.2	18.8	73.0	18.0	55.2	50.8	59.1	48.3
ZERO-SHOT-COT	36.6	53.9	66.4	62.3	85.0	74.0	32.2	96.6	60.2	51.9	61.9
PS+	29.5	50.3	63.4	70.1	71.3	79.1	_52.2_	57.4	62.9	30.1	56.6
AUTO-COT	33.5	_57.5_	_75.2_	84.38	92.0	_83.5_	20.4	83.6	72.9	65.7	66.9
COSP	33.1	57.3	65.2	84.1	92.8	80.3	27.4	96.2	70.6	66.3	67.3
STANDARD	24.8	12.0	62.6	_87.6_	27.2	76.6	0.0	60.2	76.2	_72.3_	50.0
FEW-SHOT-COT	_37.4_	42.3	62.8	57.0	_93.7_	78.5	23.6	92.8	**77.1**	70.5	63.6
COMPLEX-COT	34.6	57.2	38.3	20.5	88.7	27.4	42.2	92.4	68.1	19.3	48.9
SELF-ASK	28.0	42.9	72.8	_87.6_	85.2	78.1	21.4	1.4	62.6	63.2	54.3
LEAST-TO-MOST	30.7	55.6	75.0	74.9	84.3	80.1	39.2	_96.8_	68.7	71.3	_67.7_
SP	22.1	46.7	74.3	81.8	86.7	82.9	-	-	-	-	-
PHP	33.5	44.9	70.5	80.8	92.7	81.1	0.0	**97.4**	68.1	72.1	64.1
ADOT (Ours)	**39.0**	**58.2**	**76.2**	**91.1**	**96.5**	**91.9**	**58.4**	88.0	_72.2_	**73.5**	**74.5**

Table 9: Main experimental results on llama2-70b-chat. **Bold**: the best score. Underline: the second best score.

H The Sensitivity Analysis of Demonstrations Words and Structures

As mentioned in section 4.8, we explore the sensitivity of our ADOT method to demonstrations with different words and structures. Compared with coreference resolution, the step decomposition is flexible since no rules exist. Hence, we explore the sensitivity of ADOT under different step decomposition with variant words and structures.

The step decomposition tries to add steps in the reasoning process to bridge the logical gap between reasoning steps. To explore whether adding steps in the reasoning process is effective, we design the ADOT (Variant 1), where the step decomposition is replaced with removing some steps in the reasoning process. Furthermore, we adopt the `gpt-3.5-turbo-0613` to rewrite the optimized demonstrations after step decomposition and coreference resolution. Specifically, we adopt a simple instruction "Given the question: **question**, please rewrite the reasoning process with different words and structures: **rationale**. Be careful not to change the meaning of the reasoning process." to generate ten variants, and then randomly pick three as ADOT (Variant 1-3). These variants are shown in Fig. 8 to Fig. 11. Experimental results on `llama2-7b-chat`, `llama2-13b-chat`, and `llama2-70b-chat` are shown in Table 10, Table 11, and Table 12, respectively.

First, after the step decomposition is replaced with removing some steps, the performances of Variant 1 significantly decline, especially in AQuA and Last Letters datasets. The main reason is that a clear logic among reasoning steps is important to arithmetic and symbolic tasks.

Second, the results show that the performance on AQuA and CSQA datasets is stable in different optimized demonstrations with different words and structures. However, the performance on Last Letters dataset fluctuates obviously. We argue that the main reason is that the arithmetic and commonsense tasks focus more on macro-level meaning, while the symbolic task concentrates more on micro-level words and structures. Hence, as long as the original meaning is not changed, step decomposition with different words and structures will produce similar results in arithmetic and commonsense tasks. While on symbolic task, more careful handcraft is needed to ensure that words and structures can express meaning precisely.

Third, the demonstrations rewritten by LLMs are not uniform in answer format (highlighted at the end of answers) in many cases, which may decrease performance. For instance, in the Last Letters dataset, the answer format of the rewritten demonstrations in ADOT (Variant 2, 3) is chaotic, while in ADOT (Variant 4) is uniform. Clearly, the performance of ADOT (Variant 4) is close to ADOT while ADOT (Variant 2 and 3) are far lower. Hence, adopting a uniform answer format is also a key point in keeping the performance stable.

Overall, the above sensitivity of demonstrations words and structures can provide a detailed understanding and guidance to the manual demonstration construction.

Method	Arithmetic	Symbolic	Commonsense
	AQuA	Last Letters	CSQA
ADOT (Variant 1)	25.2	12.8	66.1
ADOT (Variant 2)	26.0	25.0	67.0
ADOT (Variant 3)	27.6	27.6	<u>67.2</u>
ADOT (Variant 4)	<u>28.0</u>	<u>39.2</u>	67.0
ADOT (Ours)	**28.7**	**48.8**	**67.6**

Table 10: Variant performance on llama2-7b-chat. **Bold**: the best score. <u>Underline</u>: the second best score.

Method	Arithmetic	Symbolic	Commonsense
	AQuA	Last Letters	CSQA
ADOT (Variant 1)	28.7	15.0	67.3
ADOT (Variant 2)	29.5	40.0	**70.3**
ADOT (Variant 3)	29.1	20.0	<u>70.2</u>
ADOT (Variant 4)	**31.5**	**48.0**	<u>70.2</u>
ADOT (Ours)	<u>31.1</u>	<u>46.0</u>	69.5

Table 11: Variant performance on llama2-13b-chat. **Bold**: the best score. <u>Underline</u>: the second best score.

Method	Arithmetic	Symbolic	Commonsense
	AQuA	Last Letters	CSQA
ADOT (Variant 1)	37.4	44.4	69.9
ADOT (Variant 2)	38.2	53.2	72.3
ADOT (Variant 3)	<u>39.0</u>	52.8	**72.8**
ADOT (Variant 4)	**40.6**	<u>58.2</u>	<u>72.5</u>
ADOT (Ours)	<u>39.0</u>	**58.4**	72.2

Table 12: Variant performance on llama2-70b-chat. **Bold**: the best score. <u>Underline</u>: the second best score.

Variant 1

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?
Answer Choices: (A) 50 (B) 45 (C) 65 (D) 78 (E) 64
A: 10 is added to each number of the 15 numbers. The new mean of the 15 numbers would be 50. Therefore, the answer is (A) 50.

Q: If a / b = 3/4 and 8a + 5b = 22,then find the value of a.
Answer Choices: (A) 1/2 (B) 3/2 (C) 5/2 (D) 4/2 (E) 7/2
A: a / b = 3/4, so 8a + 5(4a / 3) = 22. 44a / 3 = 22, so a is equal to 3/2. Therefore, the answer is (B) 3/2.

Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (A) 53 km (B) 55 km (C) 52 km (D) 60 km (E) 50 km
A: The person traveling 20 km/hr and reached his destiny in 2.5 hr. So the distance = 50 km Therefore, the answer is (E) 50 km.

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (A) 1156 (B) 1392 (C) 1393 (D) 1562 (E) 1788
A: 1 to 500 have 9 one-digit numbers from 1 to 9. 1 to 500 have 90 two-digit numbers from 10 to 99. 1 to 500 have are 401 three-digit numbers from 100 to 500. Therefore, the answer is (B) 1392.

Variant 2

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?
Answer Choices: (A) 50 (B) 45 (C) 65 (D) 78 (E) 64
A: John observed that the mean of a set of 15 numbers is 40. Then, John added 10 to each number in the set. Consequently, the mean of the updated set is also increased by 10. To find the original sum of the 15 numbers, multiply the mean (40) by the number of values (15). Thus, the original sum is 600. Adding 10 to each number in the set results in a new sum of the 15 numbers. To calculate the new sum, add the original sum (600) to the product of 10 and the number of values (15). The new sum is 750. Since the new sum is 750, the new mean of the 15 numbers is obtained by dividing this sum by the number of values (15). The new mean is 50. Therefore, the correct answer is (A) 50.

Q: If a / b = 3/4 and 8a + 5b = 22,then find the value of a.
Answer Choices: (A) 1/2 (B) 3/2 (C) 5/2 (D) 4/2 (E) 7/2
A: Given the equation a / b = 3/4, we can solve for b by multiplying both sides of the equation by b. This gives us b = 4a / 3. Substituting this value of b into the second equation, 8a + 5b = 22, we get 8a + 5(4a / 3) = 22. Simplifying this equation further, we have 8a + 20a / 3 = 22. Combining like terms, we get 44a / 3 = 22. To solve for a, we can multiply both sides of the equation by 3/44. This gives us a = (3/44) * 22, which simplifies to 3/2. Therefore, the value of a is 3/2, and the correct answer choice is (B) 3/2.

Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (A) 53 km (B) 55 km (C) 52 km (D) 60 km (E) 50 km
A: Given that a person is traveling at a speed of 20 km/hr and reaches their destination in 2.5 hours, we can calculate the distance they traveled by multiplying their speed (20 km/hr) by the time it took (2.5 hours). Therefore, the distance traveled would be 20 km/hr * 2.5 hrs = 50 km. Hence, the answer is (E) 50 km.

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (A) 1156 (B) 1392 (C) 1393 (D) 1562 (E) 1788
A: To determine the number of keystrokes needed to type the numbers from 1 to 500, we can break down the different types of numbers. First, there are 9 single-digit numbers (1-9), and each of these requires 1 keystroke. So the total number of keystrokes for single-digit numbers is 9. Next, there are 90 two-digit numbers (10-99), and each digit in these numbers requires 2 keystrokes. Therefore, the total number of keystrokes for two-digit numbers is 90 * 2 = 180. Finally, there are 401 three-digit numbers (100-500), and each digit in these numbers requires 3 keystrokes. Thus, the total number of keystrokes for three-digit numbers is 401 * 3 = 1203. By adding up the number of keystrokes for single-digit numbers, two-digit numbers, and three-digit numbers, we get 9 + 180 + 1203 = 1392 keystrokes in total. Hence, the correct answer to the question is (B) 1392.

Variant 3

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?
Answer Choices: (A) 50 (B) 45 (C) 65 (D) 78 (E) 64
A: By adding 10 to each number of the 15 numbers, the mean of the numbers will also increase by 10. Since the average of the 15 numbers is 40, the original sum of the numbers is 15 multiplied by 40, which equals 600. When adding 10 to each number, the new sum of the numbers becomes 600 + (10 * 15) = 750. Consequently, the new mean of the 15 numbers is obtained by dividing the new sum, 750, by 15, resulting in 750/15 = 50. Therefore, the correct answer is (A) 50.

Q: If a / b = 3/4 and 8a + 5b = 22,then find the value of a.
Answer Choices: (A) 1/2 (B) 3/2 (C) 5/2 (D) 4/2 (E) 7/2
A: Given the equation a / b = 3/4 and 8a + 5b = 22, we need to find the value of a. To start, we can rewrite the first equation as b = 4a / 3. Substituting this expression for b into the second equation, we get 8a + 5(4a / 3) = 22. Simplifying further, we have 8a + (20a / 3) = 22. Combining like terms, we get 8a + (20a / 3) = 22. To eliminate the fraction, we multiply both sides of the equation by 3, resulting in 24a + 20a = 66. Simplifying further, we have 44a = 66. Dividing both sides by 44, we find that a = 66 / 44, which simplifies to 3 / 2. Therefore, the value of a is 3/2, and the correct answer is (B) 3/2.

Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (A) 53 km (B) 55 km (C) 52 km (D) 60 km (E) 50 km
A: The person traveled at a speed of 20 km/hr and reached their destination in 2.5 hours. To find the distance, we multiply the speed (20 km/hr) by the time (2.5 hours). This gives us a distance of 50 kilometers. Therefore, the answer is (E) 50 km.

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (A) 1156 (B) 1392 (C) 1393 (D) 1562 (E) 1788
A: To find the total number of keystrokes needed to type the numbers from 1 to 500, we can break it down into three categories: single-digit numbers, two-digit numbers, and three-digit numbers. First, let's consider the single-digit numbers (1-9). There are 9 single-digit numbers, and each number requires 1 keystroke. Therefore, the total number of keystrokes for the single-digit numbers is 9 * 1 = 9. Next, let's move on to the two-digit numbers (10-99). There are 90 two-digit numbers, and each number requires 2 keystrokes (one for each digit). Therefore, the total number of keystrokes for the two-digit numbers is 90 * 2 = 180. Finally, let's consider the three-digit numbers (100-500). There are 401 three-digit numbers, and each number requires 3 keystrokes (one for each digit). Therefore, the total number of keystrokes for the three-digit numbers is 401 * 3 = 1203. Adding up the number of keystrokes for the single-digit numbers, two-digit numbers, and three-digit numbers, we get 9 + 180 + 1203 = 1392. Therefore, the answer is (B) 1392.

Variant 4

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?
Answer Choices: (A) 50 (B) 45 (C) 65 (D) 78 (E) 64
A: By adding 10 to each number of the 15 numbers, the mean of the numbers will also increase by 10. Since the average of the 15 numbers is 40, the original sum of the numbers is 15 multiplied by 40, which equals 600. When adding 10 to each number, the new sum of the numbers becomes 600 + (10 * 15) = 750. Consequently, the new mean of the 15 numbers is obtained by dividing the new sum, 750, by 15, resulting in 750/15 = 50. Therefore, the correct answer is (A) 50.

Q: If a / b = 3/4 and 8a + 5b = 22,then find the value of a.
Answer Choices: (A) 1/2 (B) 3/2 (C) 5/2 (D) 4/2 (E) 7/2
A: Given the equation a / b = 3/4 and 8a + 5b = 22, we need to find the value of a. To start, we can rewrite the first equation as b = 4a / 3. Substituting this expression for b into the second equation, we get 8a + 5(4a / 3) = 22. Simplifying further, we have 8a + (20a / 3) = 22. Combining like terms, we get 8a + (20a / 3) = 22. To eliminate the fraction, we multiply both sides of the equation by 3, resulting in 24a + 20a = 66. Simplifying further, we have 44a = 66. Dividing both sides by 44, we find that a = 66 / 44, which simplifies to 3 / 2. Therefore, the value of a is 3/2, and the correct answer is (B) 3/2.

Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (A) 53 km (B) 55 km (C) 52 km (D) 60 km (E) 50 km
A: The person traveled at a speed of 20 km/hr and reached their destination in 2.5 hours. To find the distance, we multiply the speed (20 km/hr) by the time (2.5 hours). This gives us a distance of 50 kilometers. Therefore, the answer is (E) 50 km.

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (A) 1156 (B) 1392 (C) 1393 (D) 1562 (E) 1788
A: To find the total number of keystrokes needed to type the numbers from 1 to 500, we can break it down into three categories: single-digit numbers, two-digit numbers, and three-digit numbers. First, let's consider the single-digit numbers (1-9). There are 9 single-digit numbers, and each number requires 1 keystroke. Therefore, the total number of keystrokes for the single-digit numbers is 9 * 1 = 9. Next, let's move on to the two-digit numbers (10-99). There are 90 two-digit numbers, and each number requires 2 keystrokes (one for each digit). Therefore, the total number of keystrokes for the two-digit numbers is 90 * 2 = 180. Finally, let's consider the three-digit numbers (100-500). There are 401 three-digit numbers, and each number requires 3 keystrokes (one for each digit). Therefore, the total number of keystrokes for the three-digit numbers is 401 * 3 = 1203. Adding up the number of keystrokes for the single-digit numbers, two-digit numbers, and three-digit numbers, we get 9 + 180 + 1203 = 1392. Therefore, the answer is (B) 1392.

Figure 8: The variant 1-4 of demonstrations on AQuA dataset.

Variant 1

Q: Take the last letters of the words in "Elon Musk" and concatenate them.
A: The last letter of "Elon" is "n". The last letter of "Musk" is "k". The answer is nk.

Q: Take the last letters of the words in "Larry Page" and concatenate them.
A: The last letter of "Larry" is "y". The last letter of "Page" is "e". The answer is ye.

Q: Take the last letters of the words in "Sergey Brin" and concatenate them.
A: The last letter of "Sergey" is "y". The last letter of "Brin" is "n". The answer is yn.

Q: Take the last letters of the words in "Bill Gates" and concatenate them.
A: The last letter of "Bill" is "l". The last letter of "Gates" is "s". The answer is ls.

Variant 3

Q: Take the last letters of the words in "Elon Musk" and concatenate them.
A: Firstly, write down the initial word "Elon" one letter at a time: "E l o n". Extract the last letter from "E l o n", which is "n". Next, write down the second word "Musk" one letter at a time: "M u s k". Extract the last letter from "M u s k", which is "k". Merge the final letters "n" and "k" together. The merging of the final letters "n" and "k" is "n" + "k" = "nk". Hence, the response is nk, just as with the combined result "nk".

Q: Take the last letters of the words in "Larry Page" and concatenate them.
A: First, write down the two words "Larry Page" and separate the letters in each word: "L a r r y" and "P a g e". Next, identify the last letter of the first word, which is "y". Then, identify the last letter of the second word, which is "e". Finally, concatenate the last letters "y" and "e" together. The concatenation of the last letters "y" and "e" is "y" + "e" = "ye". Therefore, the answer is "ye".

Q: Take the last letters of the words in "Sergey Brin" and concatenate them.
A: First, isolate the letters of the initial word "Sergey" as follows: "S e r g e y". Identify the last letter of "S e r g e y", which is "y". Next, record the letters of the second word "Brin" individually: "B r i n". Determine the final letter of "B r i n", which is "n". Combine the last letters "y" and "n" together. The combination of the last letters "y" and "n" is "y" + "n" = "yn". Similarly, following the pattern seen in the result "yn", the answer is yn.

Q: Take the last letters of the words in "Bill Gates" and concatenate them.
A: Extract the individual letters from the first word "Bill" as "B", "i", "l", "l". Identify the last letter of "B", which is "l". Similarly, list out the letters of the second word "Gates" as "G", "a", "t", "e", "s". Determine the last letter of "s", which is "s". Combine the last letters "l" and "s". Concatenate the last letters "l" and "s" to get "ls". Hence, the final answer is "ls".

Variant 2

Q: Take the last letters of the words in "Elon Musk" and concatenate them.
A: Now, write down the first term "Elon" as individual characters: "E l o n". Extract the last character from "E l o n", which is "n". Next, write down the second term "Musk" as individual characters: "M u s k". Extract the last character from "M u s k", which is "k". Combine the last characters "n" and "k" together. The combination of the last characters "n" and "k" is "n" + "k" = "nk". Similarly, the answer is "nk" based on the combined result "nk".

Q: Take the last letters of the words in "Larry Page" and concatenate them.
A: 1. Separate the word 'Larry' into its individual letters: L - a - r - r - y. 2. Identify the last letter of 'Larry', which is 'y'. 3. Separate the word 'Page' into its individual letters: P - a - g - e. 4. Identify the last letter of 'Page', which is 'e'. 5. Combine the last letters 'y' and 'e' together. 6. The combination of 'y' and 'e' is 'ye'. 7. The answer is also 'ye'.

Q: Take the last letters of the words in "Sergey Brin" and concatenate them.
A: Begin by writing out the first word "Sergey" and separate each letter: "S e r g e y". Then, identify the last letter of "S e r g e y", which is "y". Moving on to the second word "Brin", write out each letter separately: "B r i n". Identify the last letter of "B r i n", which is "n". Next, combine the last letters "y" and "n" together. The final combination of the last letters "y" and "n" is "yn". Hence, the answer is yn.

Q: Take the last letters of the words in "Bill Gates" and concatenate them.
A: First, write each letter of the word "Bill" individually: "B i l l". Take note of the last letter of "B i l l", which is "l". Next, write each letter of the word "Gates" individually: "G a t e s". Take note of the last letter of "G a t e s", which is "s". Lastly, merge the last letters "l" and "s" together. By combining the last letters "l" and "s", we get "ls". Thus, the answer is "ls".

Variant 4

Q: Take the last letters of the words in "Elon Musk" and concatenate them.
A: Commence by writing down the first word "Elon" in separate letters: "E,l,o,n". Identify the last letter of "E,l,o,n" which is "n". Proceed to write down the second word "Musk" in separate letters: "M,u,s,k". Identify the last letter of "M,u,s,k" which is "k". Combine the last letters "n" and "k" together. The result of combining the last letters "n" and "k" is "nk". Therefore, the answer is "nk".

Q: Take the last letters of the words in "Larry Page" and concatenate them.
A: To begin, separate the first word "Larry" into its individual letters: "L a r r y". Select the last letter from "L a r r y", which is "y". Next, write down the second word "Page" as separate letters: "P a g e". Choose the last letter from "P a g e", which is "e". Combine the final letters "y" and "e" to form a single word. The combination of "y" and "e" is "ye". Therefore, the answer is "ye".

Q: Take the last letters of the words in "Sergey Brin" and concatenate them.
A: Start by listing the individual letters of the first word "Sergey" as follows: "S e r g e y". Identify the last letter of "S e r g e y", which is "y". Proceed to write down the separate letters of the second word "Brin": "B r i n". Identify the last letter of "B r i n", which is "n". Combine the last letters "y" and "n" together. The combination of the last letters "y" and "n" is "y" + "n" = "yn". Therefore, the answer is "yn".

Q: Take the last letters of the words in "Bill Gates" and concatenate them.
A: First, break down the word "Bill" into individual letters: "B i l l." Next, identify the last letter of "B i l l," which is "l." Then, break down the word "Gates" into individual letters: "G a t e s." Identify the last letter of "G a t e s," which is "s." Combine the last letters "l" and "s" together. The combination of the last letters "l" and "s" is "l" + "s" = "ls". Therefore, the answer is "ls".

Figure 9: The variant 1-4 of demonstrations on Last Letters dataset.

Variant 1

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: The answer of "What do people use to absorb extra ink from a fountain pen?" must be an item that can 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: The answer of "What home entertainment equipment requires cable?" must require 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: The answer of "The fox walked from the city into the forest, what was it looking for?" must be something in the forest. So the answer is (C).

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: The answer of "Sammy wanted to go to where the people were.Where might he go?" must be a place with 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: The answer of "Where do you put your grapes just before checking out?" should be the place where grocery items are placed before checking out. 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: The answer of "Google Maps and other highway and street GPS services have replaced what?" must be something that used to do what Google Maps and GPS services do, which is to give directions. So the answer is (D).

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices: (A) harder (B) anguish (C) bitterness (D) tears (E) sadness
A: The answer of "Before getting a divorce, what did the wife feel who was doing all the work?" should be the feeling of someone getting divorced who was doing all the work. So the answer is (C).

Variant 2

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: To find the answer to "What do people use to absorb extra ink from a fountain pen?" we need to choose an option that can soak up ink.
Let's consider the choices: (A) shirt pocket, (B) calligrapher's hand, (C) inkwell, (D) desk drawer, (E) blotter. Out of these options, (E) blotter stands out as the best choice. Blotters are commonly used to soak up excess ink from fountain pens. They help to avoid smudging and promote quick drying of the ink on the paper. Blotters are typically made of absorbent paper or cloth and are placed under the paper while writing. Consequently, to absorb extra ink from a fountain pen, people may employ (E) blotter. In conclusion, the correct answer is (E) blotter.

Q: What home entertainment equipment requires cable?
Answer Choices: (A) radio shack (B) substation (C) television (D) cabinet
A: To determine what home entertainment equipment requires cable, we need to consider the given options: (A) radio shack, (B) substation, (C) television, (D) cabinet. Out of these options, it is clear that televisions often require cables to connect to different devices such as cable or satellite boxes, DVD or Blu-ray players, gaming consoles, and streaming devices. These cables are necessary to transmit audio and video signals from the source device to the television screen. Thus, it can be concluded that the home entertainment equipment that requires cable is the television, making the answer (C) television.

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: Given the question: "The fox walked from the city into the forest, what was it searching for? Answer Choices: (A) beautiful flowers (B) hen house (C) natural habitat (D) storybook" Let's think through the question: "The fox walked from the city into the forest, what was it searching for?" It must be something that can be found in the forest. Let's consider the options provided: (A) beautiful flowers, (B) hen house, (C) natural habitat, (D) storybook. The most logical reason for a fox to leave the city and enter the forest is to find a suitable place to establish its (C) natural habitat. Foxes are wild animals that usually prefer wooded areas for shelter, hunting, and raising their young. Forests offer the necessary cover and resources for a fox's survival. Consequently, the fox was probably searching for its (C) natural habitat when it walked from the city into the forest. Therefore, the answer is (C) natural habitat.

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: Given the question: "Sammy desired to go to a place where people were present. To which place could he possibly go? Choose one: (A) crowded areas, (B) race track, (C) desert, (D) apartment, (E) roadblock." Let's rephrase the reasoning process using different words and structure: The solution to the question "Sammy desired to go to a place where people were present. To which place could he possibly go?" would be a location where individuals congregate or reside. Considering the provided choices: (A) crowded areas, (B) race track, (C) desert, (D) apartment, (E) roadblock. Under these circumstances, Sammy would most likely go to (A) crowded areas. Crowded areas are locales where there is a high concentration of individuals, such as towns, cities, or public spaces. These zones offer opportunities for socializing, engagement, and the existence of people. Hence, Sammy could go to (A) crowded areas to be in the presence of people. Thus, the answer is (A) crowded areas.

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: To determine where to put your grapes just before checking out, consider where one typically places items they intend to purchase before paying for them. The options given are (A) mouth, (B) grocery cart, (C) super market, (D) fruit basket, and (E) fruit market. Typically, before checking out at a store, you would place your items, including grapes, in the (B) grocery cart. The grocery cart is used to hold the items you have selected for purchase while you navigate through the store. This is a common practice when shopping in supermarkets or grocery stores. Therefore, the correct answer is (B) grocery cart.

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: To determine what Google Maps and other GPS services have replaced, let's analyze the available options: (A) united states, (B) mexico, (C) countryside, (D) atlas. These options represent possible answers to the question, "What have Google Maps and GPS services replaced?" The correct answer is (D) atlas. Before the advent of digital mapping services, people used physical books or collections of maps called atlases. Atlases were used for navigation and finding locations, including highways, streets, and geographic features. However, with the convenience of digital mapping services like Google Maps, people can now access detailed maps and navigation directions directly on their devices. Therefore, it can be concluded that Google Maps and other GPS services have replaced the need for traditional atlases. Hence, the answer is (D) atlas.

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices: (A) harder (B) anguish (C) bitterness (D) tears (E) sadness
A: The question is asking how the wife felt before getting a divorce, while she was doing all the work. We are given answer choices (A) harder, (B) anguish, (C) bitterness, (D) tears, and (E) sadness. Considering the given options, it is likely that the wife experienced a feeling of bitterness. When one partner carries the majority of the responsibilities and tasks in a marriage, it can lead to resentment and bitterness. This emotional response often arises when there is an unequal distribution of effort and workload, which can ultimately lead to marital dissatisfaction and divorce. Therefore, before the divorce, it is probable that the wife who was shouldering most of the work felt bitterness. The answer to the question is therefore (C) bitterness.

Figure 10: The variant 1-2 of demonstrations on CSQA dataset.

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: To find the solution for "What is used to absorb excess ink from a fountain pen?" we need to identify an item that has the ability to soak up ink. Considering the available options: (A) shirt pocket, (B) calligrapher's hand, (C) inkwell, (D) desk drawer, (E) blotter. A (E) blotter is specifically designed to absorb extra ink from a fountain pen after writing. It serves to prevent smudging and ensure speedy drying of the ink on the paper. Typically, made of absorbent paper or cloth, the blotter is placed beneath the paper while writing. Therefore, it is plausible to use (E) blotter to absorb the excess ink from a fountain pen. As a result, the correct answer is (E) blotter.

Q: What home entertainment equipment requires cable?
Answer Choices: (A) radio shack (B) substation (C) television (D) cabinet
A: To determine what home entertainment equipment requires cable, we need to examine the options: (A) radio shack, (B) substation, (C) television, and (D) cabinet. Out of these options, the equipment that typically requires cables to connect to other devices is the (C) television. Televisions often need cables to connect to devices like cable or satellite boxes, DVD or Blu-ray players, gaming consoles, and streaming devices. These cables are used to transmit audio and video signals from the source device to the television screen. Therefore, the home entertainment equipment that requires cable is the (C) television.

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: To determine what the fox was looking for when it walked from the city into the forest, we should consider the options given: (A) pretty flowers, (B) hen house, (C) natural habitat, and (D) storybook. Given that the fox is leaving the city and entering the forest, it is logical to assume that it is searching for something related to its natural environment. Foxes typically prefer wooded areas for shelter, hunting, and raising their young. This makes a (C) natural habitat the most likely choice. Therefore, the answer is (C) natural habitat.

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: The question is asking where Sammy might go if he wants to be where people are. Looking at the answer choices: (A) populated areas, (B) race track, (C) desert, (D) apartment, (E) roadblock. Based on this information, it is reasonable to assume that Sammy would go to (A) populated areas. Populated areas are places where there is a high concentration of people, such as cities or towns. These areas offer opportunities for social interaction and the presence of other people. Therefore, the answer is (A) populated areas.

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: To determine where to put your grapes before checking out, consider a place where you typically store items you plan to buy before paying for them. Review the given options: (A) mouth, (B) grocery cart, (C) super market, (D) fruit basket, (E) fruit market. Before completing your purchase at a store, it is customary to place your items, including grapes, in the (B) grocery cart. The grocery cart serves as a container for the items you have chosen to purchase while you continue shopping and make final selections. This is a common practice at supermarkets or grocery stores. Therefore, just before checking out, you would place your grapes in the (B) grocery cart. Consequently, the answer is (B) grocery cart.

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: The question is asking about what Google Maps and other GPS services have replaced. We are given multiple answer choices: (A) united states, (B) mexico, (C) countryside, and (D) atlas. To find the correct answer, we need to consider what navigational tool or system was commonly used before the introduction of digital mapping services. In this case, the correct answer is (D) atlas. Before the advent of Google Maps and GPS services, people used traditional atlases - physical books or collections of maps - to navigate and find locations. Atlases provided detailed maps of highways, streets, and geographical features. With the convenience of digital mapping services, people can now access these maps and navigation instructions directly on their devices. In conclusion, Google Maps and other GPS services have replaced the use of atlases. Therefore, the correct answer is (D) atlas.

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices: (A) harder (B) anguish (C) bitterness (D) tears (E) sadness
A: The question is asking about the wife's emotions before getting a divorce while bearing most of the workload. The answer choices are (A) harder, (B) anguish, (C) bitterness, (D) tears, and (E) sadness. Given the circumstances, it seems probable that the wife experienced feelings of bitterness. When one partner is responsible for the majority of the work and tasks in a relationship, it often leads to feelings of resentment and bitterness. This emotional reaction typically arises when there is an inequity in the distribution of effort and responsibilities. These circumstances can contribute to marital dissatisfaction and ultimately lead to divorce. Therefore, prior to getting a divorce, it is likely that the wife who was carrying most of the workload felt bitter. Thus, the correct answer is (C) bitterness.

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: In order to determine what people use to absorb extra ink from a fountain pen, we need to consider the options provided. These options are (A) shirt pocket, (B) calligrapher's hand, (C) inkwell, (D) desk drawer, and (E) blotter. We are looking for an item that has the ability to absorb ink. The (E) blotter is a logical choice for absorbing excess ink from a fountain pen. Blotters are typically made of absorbent paper or cloth and are placed underneath the paper while writing to prevent smudging and ensure quick drying of the ink. Using a blotter helps to absorb any extra ink that may be left on the paper. Therefore, the answer to the question is (E) blotter.

Q: What home entertainment equipment requires cable?
Answer Choices: (A) radio shack (B) substation (C) television (D) cabinet
A: To determine which home entertainment equipment requires cable, we need to consider the options provided: (A) radio shack, (B) substation, (C) television, (D) cabinet. Since the question is asking about home entertainment equipment that requires cable, we can eliminate options (A), (B), and (D), as they do not typically require cables for operation. Now, focusing on option (C) television, it is common for televisions to require cables to connect to various devices like cable or satellite boxes, DVD or Blu-ray players, gaming consoles, and streaming devices. These cables are necessary for transmitting audio and video signals from the source device to the television screen. Thus, we can conclude that the home entertainment equipment (C) television requires cable. Therefore, the correct answer is (C) television.

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: Given the question: "The fox ventured from the urban area into the woods, what was its intention? Choices: (A) beautiful flowers (B) chicken enclosure (C) native habitat (D) storybook" let's rephrase the reasoning process using different wording and structure: To determine the purpose of the fox's journey from the city into the forest, we will examine the available options: (A) pretty flowers, (B) hen house, (C) natural habitat, (D) storybook. The most plausible explanation for a fox leaving the urban area and entering the forest is to seek out its (C) native habitat. Foxes, being wild creatures, frequently gravitate towards wooded regions for safety, hunting, and raising their offspring. The forest offers the necessary concealment and resources for a fox's survival. Consequently, it is highly likely that the fox was searching for its (C) native habitat when it traveled from the city into the forest. Thus, the correct answer is (C) native habitat. Make sure not to alter the meaning of the reasoning process.

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: Considering the question "Sammy desired to be in the vicinity of individuals. Where might he choose to go?", the logical deduction would be a location where people commonly gather or reside. Examining the available answer choices: (A) populated areas, (B) race track, (C) desert, (D) apartment, (E) roadblock. In this particular circumstance, it is highly likely that Sammy would opt for (A) populated areas. Populated areas are defined as regions with a high concentration of individuals, such as cities, towns, or public spaces. These areas facilitate social interaction, engagement, and the presence of people. Thus, Sammy would probably venture to (A) populated areas to join the others. Hence, the correct answer is (A) populated areas.

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: The solution is to identify the place where grapes are usually kept before being paid for while checking out. The options include the mouth, grocery cart, supermarket, fruit basket, and fruit market. In this scenario, considering typical shopping practices, the items selected for purchase are placed in the grocery cart just before checking out. The grocery cart is commonly used to hold items while moving through the store and making final choices. Therefore, when checking out, the grapes would be placed in the grocery cart. Based on this reasoning, the answer to the question is option (B) grocery cart.

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: The question asks what Google Maps and other GPS services have replaced. The options given are (A) united states, (B) mexico, (C) countryside, and (D) atlas. To answer this question, we need to examine the purpose of Google Maps and GPS services. These digital mapping services provide detailed maps and navigation instructions for highways, streets, and geographical features. Before the advent of digital mapping, people relied on a navigational tool called an atlas. Atlases were physical books or collections of maps that helped people find locations and navigate their way. With this in mind, we can determine that Google Maps and GPS services have replaced the use of traditional atlases. These digital services offer the same functionality as atlases but in a more convenient and accessible format. Instead of carrying around a physical book, people can now access maps and navigation instructions directly on their devices. Therefore, the answer to the question is (D) atlas.

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices: (A) harder (B) anguish (C) bitterness (D) tears (E) sadness
A: Analyzing the question "What were the wife's feelings before seeking a divorce as she carried the majority of the responsibilities?" The options are (A) harder, (B) anguish, (C) bitterness, (D) tears, and (E) sadness. In this circumstance, it is probable that the wife experienced a feeling of (C) bitterness. When one partner assumes the majority of the tasks and obligations in a marriage, it can lead to resentment and bitterness. Such emotional reactions usually arise as a result of an imbalance in the distribution of effort and workload, factors that contribute to marital dissatisfaction and ultimately divorce. Hence, before deciding on divorce, the wife, who shouldered most of the responsibilities, likely felt (C) bitterness . Therefore, (C) bitterness is the correct answer.

Figure 11: The variant 3-4 of demonstrations on CSQA dataset.

I Optimized demonstrations

As mentioned in Section 3.4, Fig. 12-Fig. 17 show the optimized demonstrations for the ten datasets. The original FEW-SHOT-COT demonstration is proposed by Wei et al. (2024).

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?
Answer Choices: (A) 50 (B) 45 (C) 65 (D) 78 (E) 64
A: If 10 is added to each number, then the mean of the numbers also increases by 10. So the new mean would be 50. The answer is (A).

Q: If a / b = 3/4 and 8a + 5b = 22,then find the value of a.
Answer Choices: (A) 1/2 (B) 3/2 (C) 5/2 (D) 4/2 (E) 7/2
A: a / b = 3/4, then b = 4a / 3. So 8a + 5(4a / 3) = 22. This simplifies to 8a + 20a / 3 = 22, which means 44a / 3 = 22. So a is equal to 3/2. The answer is (B).

Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (A) 53 km (B) 55 km (C) 52 km (D) 60 km (E) 50 km
A: The distance that the person traveled would have been 20 km/hr * 2.5 hrs = 50 km. The answer is (E).

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (A) 1156 (B) 1392 (C) 1393 (D) 1562 (E) 1788
A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (B).

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?
Answer Choices: (A) 50 (B) 45 (C) 65 (D) 78 (E) 64
A: If 10 is added to each number of the 15 numbers, then the mean of the 15 numbers also increases by 10. So the new mean of the 15 numbers would be 50. The answer is (A).

Q: If a / b = 3/4 and 8a + 5b = 22,then find the value of a.
Answer Choices: (A) 1/2 (B) 3/2 (C) 5/2 (D) 4/2 (E) 7/2
A: a / b = 3/4, then b = 4a / 3. So 8a + 5(4a / 3) = 22. 8a + 5(4a / 3) = 22 simplifies to 8a + 20a / 3 = 22, 8a + 20a / 3 = 22 means 44a / 3 = 22. So a is equal to 3/2. The answer is (B).

Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (A) 53 km (B) 55 km (C) 52 km (D) 60 km (E) 50 km
A: The person traveling 20 km/hr and reached his destiny in 2.5 hr would have been 20 km/hr * 2.5 hrs = 50 km. The answer is (E).

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (A) 1156 (B) 1392 (C) 1393 (D) 1562 (E) 1788
A: 1 to 500 have 9 one-digit numbers from 1 to 9. 1 to 500 have 90 two-digit numbers from 10 to 99. 1 to 500 have are 401 three-digit numbers from 100 to 500. The 9 one-digit numbers * 1 keystroke + The 90 two-digit numbers * 2 keystrokes + The 401 three-digit numbers * 3 keystrokes = 1392 keystrokes. The answer is (B).

Q: John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is?
Answer Choices: (A) 50 (B) 45 (C) 65 (D) 78 (E) 64
A: 10 is added to each number of the 15 numbers, so the mean of the 15 numbers also increases by 10. The average of 15 numbers is 40, so the original sum of the 15 numbers is 15 * 40 = 600. Adding 10 to each number of the 15 numbers, so the new sum of the 15 numbers is 600 + 10 * 15 = 750. The new sum of the 15 numbers is 75, so the new mean of the 15 numbers is 750 / 15 = 50. Therefore, the answer is (A) 50.

Q: If a / b = 3/4 and 8a + 5b = 22,then find the value of a.
Answer Choices: (A) 1/2 (B) 3/2 (C) 5/2 (D) 4/2 (E) 7/2
A: a / b = 3/4, so b = 4a / 3. b = 4a / 3 and 8a + 5b = 22, so 8a + 5(4a / 3) = 22. 8a + 5(4a / 3) = 22, so 8a + 20a / 3 = 22. 8a + 20a / 3 = 22, so 44a / 3 = 22. 44a / 3 = 22, so a is equal to 3/2. Therefore, the answer is (B) 3/2.

Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (A) 53 km (B) 55 km (C) 52 km (D) 60 km (E) 50 km
A: The person traveling 20 km/hr and reached his destiny in 2.5 hr, so the distance have been 20 km/hr * 2.5 hrs. The distance have been 20 km/hr * 2.5 hrs, so the distance = 50 km Therefore, the answer is (E) 50 km.

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (A) 1156 (B) 1392 (C) 1393 (D) 1562 (E) 1788
A: There are 9 single-digit numbers (1-9), which require 1 keystrokes for each digit, so the number of keystrokes for single-digit numbers is 9 * 1 = 9. There are 90 two-digit numbers (10-99), which require 2 keystrokes for each digit, so the number of keystrokes for two-digit numbers is 90 * 2 = 180. There are 401 three-digit numbers (100-500), which require 3 keystrokes for each digit, so the number of keystrokes for three-digit numbers is 401 * 3 = 1203. The number of keystrokes for single-digit numbers, two-digit numbers, and three-digit numbers is 9, 180, and 1203, respectively, so the total number of keystrokes is 9 + 180 + 1203 = 1392. Therefore, the answer is (B) 1392.

Figure 12: The demonstrations for the AQuA dataset.

Original Few-shot-cot demonstration

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: There are 15 trees originally. Then there were 21 trees after the Grove workers planted some more. So there must have been 21 - 15 = 6 trees that were planted. 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: There are originally 3 cars. Then 2 more cars arrive. Now 3 + 2 = 5 cars are in the parking lot. 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: Originally, Leah had 32 chocolates and her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39 pieces left in total. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 20 lollipops originally. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8 lollipops. 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: Shawn started with 5 toys. He then got 2 toys each from his mom and dad. So he got 2 * 2 = 4 more toys. Now he has 5 + 4 = 9 toys. 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: There were originally 9 computers. For each day from monday to thursday, 5 more computers were installed. So 4 * 5 = 20 computers were added. Now 9 + 20 = 29 computers are now in the server room. The answer is 29.

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?
A: Michael started with 58 golf balls. He lost 23 on Tuesday, and lost 2 more on wednesday. So he had 58 - 23 = 35 at the end of Tuesday, and 35 - 2 = 33 at the end of wednesday. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 23 dollars. She bought 5 bagels for 3 dollars each. So she spent 5 * 3 = 15 dollars. Now she has 23 - 15 = 8 dollars left. The answer is 8.

Few-shot-cot demonstration optimized by coreference resolution

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: There are 15 trees in the grove originally. Then there were 21 trees in the grove after the Grove workers planted some more trees in the grove. So there must have been 21 current trees - 15 original trees = 6 trees that were planted by grove workers in the grove. 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: There are originally 3 cars in the parking lot. Then 2 more cars arrive at the parking lot. Now 3 original cars + 2 newly arrived cars = 5 cars are in the parking lot. 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: Originally, Leah had 32 chocolates and Leah's sister had 42 chocolates. So in total Leah and Leah's sister had 32 chocolates + 42 chocolates = 74 chocolates. After Leah and Leah's sister eating 35 chocolates, Leah and Leah's sister had 74 chocolates - 35 chocolates = 39 chocolates pieces left in total. The answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 20 lollipops originally. Then Jason had 12 lollipops after Jason giving some lollipops to Denny. So Jason gave Denny 20 lollipops - 12 lollipops = 8 lollipops. 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: Shawn started with 5 toys. Shawn then got 2 toys each from Shawn's mom and dad. So Shawn got 2 * 2 toys = 4 more toys. Now Shawn has 5 toys + 4 toys = 9 toys. 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: There were originally 9 computers in the server room. For each day from monday to thursday, 5 more computers were installed in the server room. So 4 * 5 computers = 20 computers were added in the server room. Now 9 computers + 20 computers = 29 computers are now in the server room. The answer is 29.

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?
A: Michael started with 58 golf balls. Michael lost 23 golf balls on Tuesday, and Michael lost 2 golf balls more on wednesday. So Michael had 58 golf balls - 23 golf balls = 35 golf balls at the end of Tuesday, and 35 golf balls - 2 golf balls = 33 golf balls at the end of wednesday. The answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 23 dollars. Olivia bought 5 bagels for 3 dollars each. So Olivia spent 5 * 3 dollars = 15 dollars. Now Olivia has 23 dollars - 15 dollars = 8 dollars left. The answer is 8.

Few-shot-cot demonstration optimized by coreference resolution and step decomposition

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: There are 15 trees in the grove originally.
Then there were 21 trees in the grove after the Grove workers planted some more trees in the grove.
The num of current trees are 21.
The num of original trees are 15.
So there must have been 21 current trees - 15 original trees = 6 trees that were planted by grove workers in the grove.
Therefore, 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: There are originally 3 cars in the parking lot.
Then 2 more cars arrive at the parking lot.
So, there are 3 original cars and 2 newly arrived cars in the parking lot.
Now 3 original cars + 2 newly arrived cars = 5 cars are in the parking lot.
Therefore, 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: Originally, Leah had 32 chocolates and Leah's sister had 42 chocolates.
So in total Leah and Leah's sister had 32 chocolates + 42 chocolates = 74 chocolates.
Leah and Leah's sister eating 35 chocolates of the 74 chocolates.
The num of chocolates Leah and Leah's sister originally had are 74.
The num of chocolates Leah and Leah's sister eat are 35.
So, Leah and Leah's sister had 74 chocolates - 35 chocolates = 39 chocolates pieces left in total.
Therefore, the answer is 39.

Q: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?
A: Jason had 20 lollipops originally.
Jason giving some lollipops to Denny.
Then Jason had 12 lollipops after Jason giving some lollipops to Denny.
So Jason gave Denny 20 lollipops - 12 lollipops = 8 lollipops.
Therefore, 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: Shawn started with 5 toys.
Shawn then got 2 toys from Shawn's mom.
Now Shawn has 5 toys + 2 toys = 7 toys.
Shawn then also got 2 toys from Shawn's dad.
Now Shawn has 7 toys + 2 toys = 9 toys.
Therefore, 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: There were originally 9 computers in the server room.
For each day from monday to thursday, 5 more computers were installed in the server room.
From monday to thursday, there are 4 days: Monday, Tuesday, Wednesday, Thursday.
So 4 * 5 computers = 20 computers were added in the server room.
The num of original computers are 9.
The num of newly added computers are 20.
So, now 9 computers + 20 computers = 29 computers are now in the server room.
Therefore, the answer is 29.

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?
A: Michael started with 58 golf balls.
Michael lost 23 golf balls on Tuesday.
The num of original golf balls are 58.
The lost num of golf balls on Tuesday are 23.
So, Michael had 58 golf balls - 23 golf balls = 35 golf balls at the end of Tuesday.
Michael lost 2 golf balls more on wednesday.
The num of golf balls at the end of Tuesday are 35.
The lost num of golf balls on wednesday are 2.
So, Michael had 35 golf balls - 2 golf balls = 33 golf balls at the end of wednesday.
Therefore, the answer is 33.

Q: Olivia has $23. She bought five bagels for $3 each. How much money does she have left?
A: Olivia had 23 dollars.
Olivia bought 5 bagels for 3 dollars each.
So Olivia spent 5 * 3 dollars = 15 dollars.
The num of orginal dollars Olivia had are 23.
The num of dollars Olivia spent are 15.
Now Olivia has 23 dollars - 15 dollars = 8 dollars left.
Therefore, the answer is 8.

Figure 13: The demonstrations for the GSM8K, SVAMP, AddSub, MultiArith, and SingleEq datasets.

Q: Take the last letters of the words in "Elon Musk" and concatenate them.
A: The last letter of "Elon" is "n". The last letter of "Musk" is "k". Concatenating them is "nk". The answer is nk.

Q: Take the last letters of the words in "Larry Page" and concatenate them.
A: The last letter of "Larry" is "y". The last letter of "Page" is "e". Concatenating them is "ye". The answer is ye.

Q: Take the last letters of the words in "Sergey Brin" and concatenate them.
A: The last letter of "Sergey" is "y". The last letter of "Brin" is "n". Concatenating them is "yn". The answer is yn.

Q: Take the last letters of the words in "Bill Gates" and concatenate them.
A: The last letter of "Bill" is "l". The last letter of "Gates" is "s". Concatenating them is "ls". The answer is ls.

Q: Take the last letters of the words in "Elon Musk" and concatenate them.
A: The last letter of "Elon" is "n". The last letter of "Musk" is "k". Concatenating "n" and "k" is "nk". The answer is nk.

Q: Take the last letters of the words in "Larry Page" and concatenate them.
A: The last letter of "Larry" is "y". The last letter of "Page" is "e". Concatenating "y and "e" is "ye". The answer is ye.

Q: Take the last letters of the words in "Sergey Brin" and concatenate them.
A: The last letter of "Sergey" is "y". The last letter of "Brin" is "n". Concatenating "y" and "n" is "yn". The answer is yn.

Q: Take the last letters of the words in "Bill Gates" and concatenate them.
A: The last letter of "Bill" is "l". The last letter of "Gates" is "s". Concatenating "l" and "s" is "ls". The answer is ls.

Q: Take the last letters of the words in "Elon Musk" and concatenate them.
A: Write down the fisrt word "Elon" in separate letters: "E l o n".
Take the last letter of "E l o n", which is "n".
Write down the Second word "Musk" in separate letters: "M u s k".
Take the last letter of "M u s k", which is "k".
Concatenate the last letters "n" and "k" together.
The concatenation of the last letters "n" and "k" is "n" + "k" = "nk".
Same with the concatenation result "nk", the answer is nk.

Q: Take the last letters of the words in "Larry Page" and concatenate them.
A: Write down the fisrt word "Larry" in separate letters: "L a r r y".
Take the last letter of "L a r r y", which is "y".
Write down the Second word "Page" in separate letters: "P a g e".
Take the last letter of "P a g e", which is "e".
Concatenate the last letters "y" and "e" together.
The concatenation of the last letters "y" and "e" is "y" + "e" = "ye".
Same with the concatenation result "ye", the answer is ye.

Q: Take the last letters of the words in "Sergey Brin" and concatenate them.
A: Write down the fisrt word "Sergey" in separate letters: "S e r g e y".
Take the last letter of "S e r g e y", which is "y".
Write down the Second word "Brin" in separate letters: "B r i n".
Take the last letter of "B r i n", which is "n".
Concatenate the last letters "y" and "n" together.
The concatenation of the last letters "y" and "n" is "y" + "n" ="yn".
Same with the concatenation result "yn", the answer is yn.

Q: Take the last letters of the words in "Bill Gates" and concatenate them.
A: Write down the fisrt word "Bill" in separate letters: "B i l l".
Take the last letter of "B i l l", which is "l".
Write down the Second word "Gates" in separate letters: "G a t e s".
Take the last letter of "G a t e s", which is "s".
Concatenate the last letters "l" and "s" together.
The concatenation of the last letters "l" and "s" is "l" + "s" = "ls".
Same with the concatenation result "ls", the answer is ls.

Figure 14: The demonstrations for the Last Letters datasets.

Original Few-shot-cot demonstration	Few-shot-cot demonstration optimized by coreference resolution	Few-shot-cot demonstration optimized by coreference resolution and step decomposition

Original Few-shot-cot demonstration

Q: A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up?
A: The coin was flipped by Ka and Sherrie. 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. Jamey flips the coin. Teressa flips the coin. Is the coin still heads up?
A: 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: 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: 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. Sal flips the coin. Raymond does not flip the coin. Is the coin still heads up?
A: 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: 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: 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.

Q: A coin is heads up. Ryan flips the coin. Shaunda flips the coin. Is the coin still heads up?
A: The coin was flipped by Ryan and Shaunda. 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.

Few-shot-cot demonstration optimized by coreference resolution

Q: A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up?
A: The coin was flipped by Ka and Sherrie. So the coin was flipped 2 times. The coin flipped times is an even number. The coin started heads up, so after an even number of flips, the coin will still be heads 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: The coin was flipped by Jamey and Teressa. So the coin was flipped 2 times. The coin flipped times is an even number. The coin started heads up, so after an even number of flips, the coin 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: The coin was flipped by Maybelle. So the coin was flipped 1 time. The coin flipped times is an odd number. The coin started heads up, so after an odd number of flips, the coin 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: The coin was flipped by Conception. So the coin was flipped 1 time. The coin flipped times is an odd number. The coin started heads up, so after an odd number of flips, the coin will be tails up. So the answer is no.

Q: A coin is heads up. Sal flips the coin. Raymond does not flip the coin. Is the coin still heads up?
A: The coin was flipped by Sal. So the coin was flipped 1 time. The coin flipped times is an odd number. The coin started heads up, so after an odd number of flips, the coin 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: The coin was flipped by Conception. So the coin was flipped 1 time. The coin flipped times is an odd number. The coin started heads up, so after an odd number of flips, the coin 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: The coin was flipped by no one. So the coin was flipped 0 times. The coin started heads up, and the coin was not flipped, so the coin is still heads up. So the answer is yes.

Q: A coin is heads up. Ryan flips the coin. Shaunda flips the coin. Is the coin still heads up?
A: The coin was flipped by Ryan and Shaunda. So the coin was flipped 2 times. The coin flipped times is an even number. The coin started heads up, so after an even number of flips, the coin will still be heads up. So the answer is yes.

Few-shot-cot demonstration optimized by coreference resolution and step decomposition

Q: A coin is heads up. Ka flips the coin. Sherrie flips the coin. Is the coin still heads up?
A: The coin was flipped by Ka. The coin was flipped by Sherrie. So the coin was flipped 2 times. The coin flipped times is an even number. The coin started heads up, so after an even number of flips, the coin will still be heads 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: The coin was flipped by Jamey. The coin was flipped by Teressa. So the coin was flipped 2 times. The coin flipped times is an even number. The coin started heads up, so after an even number of flips, the coin 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: The coin was flipped by Maybelle. So the coin was flipped 1 time. The coin flipped times is an odd number. The coin started heads up, so after an odd number of flips, the coin 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: The coin was flipped by Conception. So the coin was flipped 1 time. The coin flipped times is an odd number. The coin started heads up, so after an odd number of flips, the coin will be tails up. So the answer is no.

Q: A coin is heads up. Sal flips the coin. Raymond does not flip the coin. Is the coin still heads up?
A: The coin was flipped by Sal. So the coin was flipped 1 time. The coin flipped times is an odd number. The coin started heads up, so after an odd number of flips, the coin 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: The coin was flipped by Conception. So the coin was flipped 1 time. The coin flipped times is an odd number. The coin started heads up, so after an odd number of flips, the coin 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: The coin was flipped by no one. So the coin was flipped 0 times. The coin started heads up, and the coin was not flipped, so the coin is still heads up. So the answer is yes.

Q: A coin is heads up. Ryan flips the coin. Shaunda flips the coin. Is the coin still heads up?
A: The coin was flipped by Ryan. The coin was flipped by Shaunda. So the coin was flipped 2 times. The coin flipped times is an even number. The coin started heads up, so after an even number of flips, the coin will still be heads up. So the answer is yes.

Figure 15: The demonstrations for the Coin Flip datasets.

Original Few-shot-cot demonstration

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: The answer must be an item that can absorb ink. 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: The answer must require 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: The answer must be something in the forest. Of the above choices, only natural habitat is in the forest. So the answer is (C).

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: 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: 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: 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 atlas are used to give directions. So the answer is (D).

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices: (A) harder (B) anguish (C) bitterness (D) tears (E) sadness
A: The answer should be the feeling of someone getting divorced who was doing all the work. Of the above choices, the closest feeling is bitterness. So the answer is (C).

Few-shot-cot demonstration optimized by coreference resolution

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: The answer of "What do people use to absorb extra ink from a fountain pen?" must be an item that can absorb ink. Of the choices: "(A) shirt pocket (B) calligrapher's hand (C) inkwell (D) desk drawer (E) blotter", only (E) 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: The answer of "What home entertainment equipment requires cable?" must require cable. Of the choices: "(A) radio shack (B) substation (C) television (D) cabinet", only (C) 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: The answer of "The fox walked from the city into the forest, what was it looking for?" must be something in the forest. Of the choices: "(A) pretty flowers (B) hen house (C) natural habitat (D) storybook", only (C) natural habitat is in the forest. So the answer is (C).

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: The answer of "Sammy wanted to go to where the people were. Where might he go?" must be a place with a lot of people. Of the choices: "(A) populated areas (B) race track (C) desert (D) apartment (E) roadblock", only (A) 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: The answer of "Where do you put your grapes just before checking out?" should be the place where grocery items are placed before checking out. Of the choices: "(A) mouth (B) grocery cart (C)super market (D) fruit basket (E) fruit market", (B) 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: The answer of "Google Maps and other highway and street GPS services have replaced what?" must be something that used to do what Google Maps and GPS services do, which is to give directions. Of the choices: "(A) united states (B) mexico (C) countryside (D) atlas", only (D) atlas are used to give directions. So the answer is (D).

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices: (A) harder (B) anguish (C) bitterness (D) tears (E) sadness
A: The answer of "Before getting a divorce, what did the wife feel who was doing all the work?" should be the feeling of someone getting divorced who was doing all the work. Of the choices: "(A) harder (B) anguish (C) bitterness (D) tears (E) sadness", the closest feeling is (C) bitterness. So the answer is (C).

Few-shot-cot demonstration optimized by coreference resolution and step decomposition

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: The answer of "What do people use to absorb extra ink from a fountain pen?" must be an item that can absorb ink. Consider the options given: (A) shirt pocket, (B) calligrapher's hand, (C) inkwell, (D) desk drawer, (E) blotter. The (E) blotter are used to absorb excess ink from a fountain pen after writing, helping to prevent smudging and to ensure that the ink dries quickly on the paper. They are typically made of absorbent paper or cloth and are placed underneath the paper as you write. Hence, people may use (E) blotter to absorb extra ink from a fountain pen. Therefore, the answer is (E) blotter.

Q: What home entertainment equipment requires cable?
Answer Choices: (A) radio shack (B) substation (C) television (D) cabinet
A: The answer of "What home entertainment equipment requires cable?" must require cable. Consider the options given: (A) radio shack, (B) substation, (C) television, (D) cabinet. The (C) television sets often require cables to connect to various devices such as cable or satellite boxes, DVD or Blu-ray players, gaming consoles, streaming devices, and more. These cables are used to transmit audio and video signals from the source device to the television screen. Hence, the home entertainment equipment (C) television requires cable. Therefore, the answer is (C) television.

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: The answer of "The fox walked from the city into the forest, what was it looking for?" must be something in the forest. Consider the options given: (A) pretty flowers, (B) hen house, (C) natural habitat, (D) storybook. The most likely reason for a fox to leave the city and enter the forest is to find a suitable place to establish its (C) natural habitat. Foxes are wild animals that often prefer wooded areas for shelter, hunting, and raising their young. Forests provide the necessary cover and resources for a fox's survival. Hence, the fox was likely looking for its (C) natural habitat when it walked from the city into the forest. Therefore, the answer is (C) natural habitat.

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: The answer to "Sammy wanted to go to where the people were. Where might he go?" would be a location where people gather or reside. Consider the options given: (A) populated areas, (B) race track, (C) desert, (D) apartment, (E) roadblock. In this scenario, Sammy would most likely go to (A) populated areas. Populated areas are locations where there is a concentration of people, such as towns, cities, or public spaces. These areas provide opportunities for social interaction, engagement, and the presence of people. Hence, Sammy might go to (A) populated areas to be where the people are. Therefore, the answer is (A) populated areas.

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: The answer to "Where do you put your grapes just before checking out?" would be a place where you typically keep items you intend to purchase before paying for them. Consider the options given: (A) mouth, (B) grocery cart, (C) super market, (D) fruit basket, (E) fruit market. Just before checking out at a store, you would typically place your items, including grapes, in the (B) grocery cart. The grocery cart is used to hold the items you've selected for purchase while you move through the store and make your final selections. This is a common practice when shopping in supermarkets or grocery stores. Hence, you would put your grapes in the (B) grocery cart just before checking out. Therefore, the answer is (B) grocery cart.

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: The answer to "Google Maps and other highway and street GPS services have replaced what?" would be a navigational tool that was commonly used before the advent of digital mapping services. Consider the options given: (A) united states, (B) mexico, (C) countryside, (D) atlas. Google Maps and GPS services have replaced the need for traditional (D) atlases. Atlases were physical books or collections of maps that people used to navigate and find locations, including highways, streets, and geographical features. With the convenience of digital mapping services, people can now access detailed maps and navigation instructions directly on their devices. Hence, Google Maps and other GPS services have replaced the use of (D) atlases. Therefore, the answer is (D) atlas.

Q: Before getting a divorce, what did the wife feel who was doing all the work?
Answer Choices: (A) harder (B) anguish (C) bitterness (D) tears (E) sadness
A: The question "Before getting a divorce, what did the wife feel who was doing all the work?" is inquiring about the emotional state of the wife before the divorce. Consider the options given: (A) harder, (B) anguish, (C) bitterness, (D) tears, (E) sadness. Given the context, it is likely that the wife felt a sense of (C) bitterness. When one partner feels burdened with most of the responsibilities and tasks in a relationship, it can lead to resentment and bitterness. This emotional response often arises when there is an imbalance in the distribution of effort and workload, which can contribute to marital dissatisfaction and, ultimately, divorce. Hence, before getting a divorce, the wife who was shouldering most of the work might have felt (C) bitterness. Therefore, the answer is (C) bitterness.

Figure 16: The demonstrations for the CSQA datasets.

Q: Do hamsters provide food for any animals?
A: 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: Brooke Shields went to Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. So the answer is yes.

Q: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls?
A: 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: College commencement ceremonies can happen in December, May, and June. December is in the winter, so there can be frost. Thus, there could be frost at some commencements. So the answer is yes.

Q: Yes or no: Could a llama birth twice during War in Vietnam (1945-46)?
A: The War in Vietnam was 6 months. The gestation period for a llama is 11 months, which is more than 6 months. Thus, a llama could not give birth twice during the War in Vietnam. So the answer is no.

Q: Yes or no: Would a pear sink in water?
A: 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.

Q: Do hamsters provide food for any animals?
A: Hamsters are prey animals. Prey animals 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: Brooke Shields went to Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. So the answer is yes.

Q: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls?
A: 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: College commencement ceremonies can happen in December, May, and June. December is in the winter, so there can be frost. Thus, there could be frost at some college commencements. So the answer is yes.

Q: Yes or no: Could a llama birth twice during War in Vietnam (1945-46)?
A: The War in Vietnam was 6 months. The gestation period for a llama is 11 months. The gestation period for a llama is more than 6 months. Thus, a llama could not give birth twice during the War in Vietnam. So the answer is no.

Q: Yes or no: Would a pear sink in water?
A: The density of a pear is about 0.6g/cm3. The density of a pear is less than water. Objects less dense than water float. Thus, a pear would float. So the answer is no.

Q: Do hamsters provide food for any animals?
A: Hamsters are prey animals. Prey are food for predators. So Hamsters 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: Brooke Shields went to Princeton University. Princeton University is academically rigorous. Hence, Brooke Shields could succeed at the Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. So the answer is yes.

Q: Yes or no: Hydrogen's atomic number squared exceeds number of Spice Girls?
A: Hydrogen has an atomic number of 1. 1 squared is 1. There are 5 Spice Girls. The atomic number of Hydrogen 1 does not exceed the number of Spice Girls 5. 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: College commencement ceremonies can happen in December, May, and June. December is in the winter, winter can see frost. So December can see frost. Hence, College commencement can see frost. Thus, there could be frost at some college commencements. So the answer is yes.

Q: Yes or no: Could a llama birth twice during War in Vietnam (1945-46)?
A: The War in Vietnam was 6 months. The gestation period for a llama is 11 months. The gestation period for a llama is more than 6 months. A llama could not give birth twice in 6 months. Thus, a llama could not give birth twice during the War in Vietnam. So the answer is no.

Q: Yes or no: Would a pear sink in water?
A: The density of a pear is about 0.6g/cm3. The density of a pear is about 1g/cm3. The density of a pear is less than water. Objects less dense than water float. Thus, a pear would float. So the answer is no.

Figure 17: The demonstrations for the StrategyQA datasets.