

IAO prompting: Forcing Large Language Models to Show their Reasoning through an Input-Action-Output Template

Anonymous ACL submission

Abstract

The effectiveness of Large Language Models (LLMs) in tackling diverse reasoning problems is further improved by chain-of-thought prompting, which makes explicit the intermediate reasoning steps. Additionally, recent research has proved the importance of explicitly structuring the reasoning procedure. In this work, we introduce IAO (input-action-output) prompting, a straightforward template based prompting method that allows the complex reasoning process to be explicitly modelled in a structured manner. IAO autonomously breaks down problems into a series of simpler reasoning steps and then solves them in sequence, each with explicit input information, action applied, and intermediate output. The solved steps inform the subsequent steps, facilitating progressive reasoning. This explicit structure not only amplifies reasoning performance but also fosters enhanced interpretability and transparency. Extensive experiments across various reasoning tasks demonstrate IAO’s strong zero-shot capabilities, showcasing its effectiveness in unlocking and leveraging the true power of LLM reasoning.

1 Introduction

The recent progress in natural language processing (NLP) can be largely attributed to the proliferation of large language models (LLMs) (Vaswani et al., 2017; Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020; Chowdhery et al., 2023). Notably, these models excel at addressing diverse tasks, requiring minimal or even no explicit training data. This remarkable ability, named few-shot or zero-shot learning, allows LLMs to tackle challenges with none or just a handful of examples. The technique enabling this ability, *prompting* (Liu et al., 2023), has evolved into a pivotal area of exploration in NLP research garnering significant attention. The research has focus on creating effective prompts, both manually (Schick and Schütze, 2021;

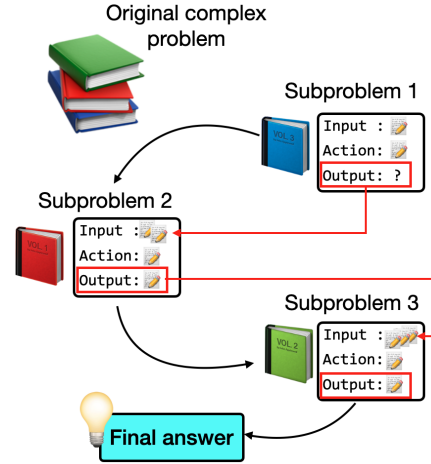


Figure 1: Illustration of IAO prompting with emphasis on the problem decomposition into Input-Action-Output. The intermediate output of each step is the input for subsequent reasoning steps.

Reynolds and McDonell, 2021) and through automated approaches (Gao et al., 2021; Shin et al., 2020).

However, recent research has highlighted the critical role of explicitly structuring the reasoning procedure. Chain-of-thought (CoT) prompting (Wei et al., 2022; Wang et al., 2022), for instance, achieved significant improvements by leveraging a step-by-step breakdown of tasks to enhance reasoning performance. This allows to guide LLM through a series of logical steps, akin to human reasoning. Loosely speaking, rather than simply presenting questions and expecting direct answers, this method involves breaking down complex tasks or problems into smaller, sequential steps. Each step builds upon the previous one, leading the model through a coherent chain of reasoning. This approach has opened a new wave of prompting methods for zero shot reasoning allowing to tackle complex and multi-step reasoning problems. Additionally, chain-of-thought can be expanded by providing demonstrations that typically consist of

a series of sequential steps that represent the desired reasoning process for a particular task or problem. These steps serve as scaffolding for the model, helping it to understand the logical flow of information and make informed decisions. Such an approach shows impressive performance improvements in reasoning tasks. Notably, in the zero-shot setting, it was shown that a simple prompt such as “let’s think step by step” could facilitate the step-by-step thinking process before answering the original question (Kojima et al., 2022). Such a task-agnostic method unveiled that LLMs can be good zero-shot reasoners.

However, the underlying reasoning steps of the reasoning chain occasionally lack in explicitness, posing challenges in comprehending the internal logic of the LLM. When the reasoning steps are not explicit enough, it becomes challenging to interpret how the model arrives at its conclusions. This opacity in the model’s decision-making process can hinder efforts to recognize errors, provide targeted feedback for improvement, or ensure the model behaves in a manner consistent with ethical or safety considerations. This may lead to issues of trust and accountability, especially in critical applications such as healthcare, or autonomous systems.

Addressing this challenge requires exploring techniques to enhance the transparency and interpretability of language models.

This work proposes IAO (Input-Action-Output) prompting, by introducing a simple yet powerful template. This prompting approach explicitly models the reasoning process in a structured manner, analogous to dissecting a complex problem into a series of well-defined, sequential subproblems. Each step towards the final answer meticulously outlines the information used, the action planned, and the intermediate output generated. This transparent breakdown not only maintains or improves the LLM’s reasoning abilities but IAO facilitates clearer problem decomposition, guiding the LLM to tackle complex tasks with greater accuracy and efficiency.

Despite the simplicity, IAO prompt successfully generates a plausible reasoning path in a zero-shot manner and reaches the correct answer in problems where the standard zero-shot-CoT approach fails or is not explicit enough.

We summarize our main contributions in this work as follows:

- We propose a new approach, IAO prompt,

(input-action-output) that employs a straightforward template to explicitly structure and model the LLM reasoning process step-by-step, while applying this structured format within a “chain-of-thought” paradigm.

- Extensive experiments across various reasoning tasks showcase IAO’s strong performance in the zero-shot setting. This suggests IAO potential as an improved alternative to existing chain-of-thought prompting methods.

2 Related Work

Prompting Chain-of-thought (CoT) prompting provides intermediate rationales in the reasoning chain for solving problems, leading to improved accuracy and interpretability in tasks requiring multi-step reasoning. The common instance of CoT is 0-shot CoT (Kojima et al., 2022) and few-shot CoT (Wei et al., 2022). The first one is the most relevant to our work. Loosely speaking, 0-shot CoT prompts the LLM with a general instruction such as “Let’s think step by step” and as a consequence the LLM produces intermediate reasoning steps. On the other hand, few-shot prompting requires manually constructed task-specific demonstrations, so it is more costly. Zhang et al. (2022) proposed an approach to automatically construct few-shot demonstrations based on zero-shot prompting to save the manual cost. Another categorization for prompting methods is based on whether they involve a single step or multiple steps. An example of multi-step prompting is the work by Zhou et al. (2022) which is discussed in the next paragraph.

Problem Decomposition and Subquestioning

Recent research shows that breaking down reasoning into smaller questions and answers improves problem-solving. Shridhar et al. (2022) even condensed LLMs into smaller models capable of reasoning. However, these studies mainly focused on the entire reasoning process, rather than its individual parts.

Similar research by (Zhou et al., 2022) found that breaking down reasoning tasks into sub-problems makes LLMs perform better, paving the way for our work. Their proposed approach, Least-to-Most prompting, is a prompting strategy that reduces a complex problem into a list of sub-questions and sequentially solves the sub-questions. Each sub-question is solved with the answer to previously solved sub-questions. Compared to zero-shot CoT,



<p>LLM Input</p> <p>Question:</p> <p>Patty's Plumbing charges \$40 to visit a house to make a repair, plus \$35 per hour, or part thereof, for labor, plus parts. One job took 2.25 hours and used \$60 in parts. How much did Patty charge?</p> <p>CoT Prompt:</p> <p>Let's think step by step.</p>	<p>LLM Output</p> <p>The total charge is the sum of the visit charge, the labor charge, and the parts charge. The visit charge is \$40. The labor charge is \$35 per hour, or part thereof. Since the job took 2.25 hours, the labor charge is $\\$35 \times 2.25 = \\78.75. The parts charge is \$60. Therefore, the total charge is $\\$40 + \\$78.75 + \\$60 = \underline{\\$178.75}$. Therefore, the answer is 178.75</p> <p>GT : 205.00 </p>
<p>LLM Input</p> <p>Question:</p> <p>Patty's Plumbing charges \$40 to visit a house to make a repair, plus \$35 per hour, or part thereof, for labor, plus parts. One job took 2.25 hours and used \$60 in parts. How much did Patty charge?</p> <p>IAO Prompt:</p> <p>- Step [number]: - Subquestion: [subquestion] - Input: [input, facts] - Action: [action written as a function] - Output: [output obtained from applying action to input]</p>	<p>Output</p> <p>- Step 1: - Subquestion: How much did Patty charge for labor? - Input: Patty charges \$35 per hour, or part thereof, for labor. The job took 2.25 hours. - Action: Round up the number of hours to the nearest whole number. Multiply the number of hours by the hourly rate. - Output: 3 hours * \$35 per hour = \$105 - Step 2: - Subquestion: How much did Patty charge for parts? - Input: The job used \$60 in parts. - Action: Add the cost of the parts to the cost of labor. - Output: \$105 + \$60 = \$165 - Step 3: - Subquestion: How much did Patty charge for the visit? - Input: Patty charges \$40 to visit a house to make a repair. - Action: Add the cost of the visit to the cost of labor and parts. - Output: \$165 + \$40 = \$205 Therefore, the answer is 205.</p> <p>GT : 205.00 </p>

Figure 2: Example of IAO prompting compared to 0-shot CoT. This example is taken from GSM8k using PALM-2. CoT fails to find the correct answer due to overlooking some part of the input information.

this method has more restrictions on the structure of reasoning by decomposing and sequentially answering. The main difference between our IAO prompting and Least-to-Most prompting is that the latter requires careful design of prompts for different tasks and it may not be as efficient as single-step approaches for some problems. In fact, Least-to-most prompting is a few-shot prompting method. On the other hand, IAO autonomously decomposes the problem and provide extra transparency by explicitly stating the intermediate input and output information as well as the planned action. Another approach that is close to IAO prompting is Tab-CoT (Ziqi and Lu, 2023). It presents a novel approach to prompting LLMs for complex reasoning tasks using a tabular format. This tabular structure allows for more explicit and organized reasoning steps. An advantage is that the tabular format makes the reasoning process more explicit and easier to understand. However, the design of the table prompt requires domain knowledge and careful consideration in some cases.

3 IAO prompting

We introduce IAO promoting, a new prompting technique that forces the LLM to decompose the problem into smaller ones, which it solves se-

quentially providing at each step the input, the action, and the intermediate output. This is inspired by procedural reasoning, the cognitive process of problem-solving and decision-making following a step-by-step procedure that involves breaking down a problem into smaller, manageable steps and executing a series of predefined operations or rules to reach a solution. This type of reasoning relies on explicit knowledge of procedures, rules, and algorithms rather than on implicit or intuitive understanding. The main motivation comes from the observation that often 0-shot CoT reasoning chains are incomplete or some intermediate steps are merged, which may lead to a wrong final answer. As shown in Figure 2, the LLM started reasoning directly on the question and missed an important aspect of the question, namely "or part thereof". By forcing the model to reason step by step within a structured framework, the reasoning chain and overall performance improves. Additionally, this prompting approach has the benefit of increased interpretability as it is clearer which input and action led to the intermediate result, and this improves the ability to understand eventual mistakes. In short, IAO prompting consists of three simple steps as show in Figure 2:

- **Subquestion:** Instead of directly addressing

the question, the LLM decomposes the problem into smaller problems by formulating sub-questions, which it then attempts to answer one at a time in a sequential manner.

- **Input:** we prompt the model to "think" about the input information available at that step, the facts and assumptions it needs or has up until that step in the reasoning chain.
- **Action:** the LLM "plans" the action it needs to perform in order to solve the subquestion. When possible and relevant, the question appears as a function or procedure.
- **Output:** the LLM outputs the result of applying the action to the input. This intermediate output is a foundational block for the next steps in the reasoning chain.

In the following sections, we present an empirical study of IAO prompting on a range of reasoning tasks and discuss the results obtained.

4 Experimental setup

Here, we define the tasks, models and baselines we use for the experimental validation of the proposed approach.

4.1 Tasks

We experiment with the following tasks: (a) **arithmetic reasoning** (GSM8k (Cobbe et al., 2021), AQuA (Ling et al., 2017)), (b) **commonsense reasoning** (StrategyQA (Geva et al., 2021), CommonsenseQA (Talmor et al., 2019)), (c) **symbolic reasoning** (Last Letter (Wei et al., 2022)) and (d) **logical reasoning** (Date Understanding & Shuffled Object Tracking (Srivastava et al., 2023)). Table 1 presents some dataset information and further details are in Table 8 and Appendix A.

Reasoning Type	Dataset	Size	Answer Type
<i>Arithmetic Reasoning</i>	AQUA	254	Multiple Choice
	GSM8k	1319	Numeral
<i>Commonsense Reasoning</i>	StrategyQA	2290	Yes/No
	CommonsenseQA	1221	Multiple Choice
<i>Other Reasoning</i>	Date Understanding	369	Multiple Choice
	Object Tracking	750	Multiple Choice
<i>Symbolic Reasoning</i>	Last Letter	500	String

Table 1: Tasks, data statistics and answer type.

4.2 Models

We use the following state of the art LLMs: PALM-2 (Anil et al., 2023) (text-unicorn) and GPT-4 (gpt-4-1106-preview) (Achiam et al., 2023). While GPT-4 is the more capable model, we find that PALM-2 provides an interesting comparison. We also experimented with GPT-3.5 models but found that their instruction-following capabilities were limited to support the template based prompt we propose. During generation, no task demonstration is provided and the models are prompted with the template and answer extraction prompt only. This highlights a core benefit of the template: the ability to precisely guide the model during generation without concrete demonstrations.

4.3 Baselines

As a baseline, we compare our approach to chain-of-thought (CoT) (Wei et al., 2022), in particular zero-shot-CoT (Kojima et al., 2022) with the prompt `Let's think step by step` appended to the question. We use two different settings: the single step where the reasoning prompt and answer extraction prompt are in a single API call; and the two-step setting.

Compute cost PaLM-2 (text-unicorn) has a cost of per API requests: \$0.0025/1K characters in input and \$0.0075/1K characters in output. GPT-4 has a cost \$0.01/1K tokens and \$0.03/1K tokens.

5 Results

Table 2 and Table 3 summarize the results for Palm-2 and GPT-4, respectively. From these results, we observe performance improvements consistent with our original hypothesis.

5.1 Arithmetic Reasoning

GPT-4 From Table 3, we see that IAO leads to mixed results in the arithmetic reasoning setting. IAO prompting maintains the performance for the AQuA dataset. On the other hand, there is an increase in performance for the GSM8k dataset that goes from 92.0% in the 0-shot CoT setting to 94.2%. We see that our approach does not significantly increase the models' performance which could be due to the inherent hardness and deep reasoning associated with these tasks. Moreover, we observe that most cases where the IAO prompt fails are due to inherent arithmetic calculation failures. Loosely speaking, while the "action" planned is

Task	Arithmetic Reasoning		Logical Reasoning		Commonsense Reasoning		Symbolic Reasoning
	AQUA	GSM8K	Date Understanding	Object Tracking	StrategyQA	CommonsenseQA	Last Letter
0-shot CoT	66.3	78.2	86.2	63.1	74.4	80.1	77.2
0-shot IAO (ours)	63.1	82.3	88.1	67.1	76.9	83.1	88.8

Table 2: Evaluation results for PALM-2. Bold denotes best result. All methods use the same answer extraction prompt in a single stage for fair comparison. All methods are evaluated under the zero-shot setting.

Task	Arithmetic Reasoning		Logical Reasoning		Commonsense Reasoning		Symbolic Reasoning
	AQUA	GSM8K	Date Understanding	Object Tracking	StrategyQA	CommonsenseQA	Last Letter
0-shot CoT	70.3	92.0	83.1	100	75.6	81.6	92.6
0-shot IAO (ours)	70.2	94.2	83.2	100	76.3	84.8	94.7

Table 3: Evaluation results for GPT-4. Bold denotes best result. All methods use the same answer extraction prompt in a single stage for fair comparison. All methods are evaluated under the zero-shot setting.

correct and the function is also correct, the LLM does calculation errors or fails to report the correct output as a final answer. We discuss this further in the following sections.

PALM-2 We observe the same trend when using PALM-2. There is a decrease in terms of percentage points (p.p), 3%, for the AQUA dataset (where the baseline accuracy is of 66.3%) but improves over the 0-shot CoT baseline for the GSM8k dataset (4.1 p.p). The same observations about the errors and failures for the IAO prompting apply to PALM-2 too.

5.2 Logical Reasoning

The datasets studied are Date Understanding and Object tracking from (Srivastava et al., 2023). The former asks the models to infer the date from a context. Tracking Shuffled Objects tests a model’s ability to infer the final state of objects given its initial state and a sequence of object shuffling.

GPT-4 GPT-4 achieved perfect scores for Object tracking task in both settings. However, that is not the case for the Date Understanding task where the baseline accuracy is of 83.1 and it is substantially matched by IAO prompting. The challenge in this task is to correctly interpret the question and understand the temporal setting. Even by forcing the LLM to interrogate itself about what it knows (*Input field*) it fails to solve this issue.

PALM-2 IAO prompting achieves 67.1 p.p (over a baseline of 63.1) for the Object Tracking dataset. On the other hand, it achieves an accuracy of 88.1 from a baseline of 86.2 corresponding to the 0-shot baseline for the Date Understanding task.

5.3 Commonsense Reasoning

CommonsenseQA asks questions with complex semantics that often require reasoning based on prior knowledge (Talmor et al., 2019). StrategyQA dataset have implicit multi-step questions requiring a wide-range of reasoning skills to answer them (Geva et al., 2021).

GPT-4 Overall, the improvements in accuracy are relatively modest for the StrategyQA dataset. This is likely because the dataset contains a number of ambiguous questions, which if read verbatim may have many plausible answers but the ground truth contains only one answer. GPT-4 will in such cases refrain from giving a final answer, which is counted as an incorrect answer. This lead to a task accuracy of 76.3 p.p in StrategyQA and 84.8 in CommonsenseQA where the baseline prompt accuracy are 75.6 and 81.6 respectively.

PALM-2 Here again, the increases in accuracy are relatively modest for the same reasons. For StrategyQA, we see an improvement from 74.4 p.p to 76.9. For CommonsenseQA, we observe an increase in accuracy over the 0-shot baseline of 3.0 p.p.

5.4 Symbolic Reasoning

We use the Last Letter Concatenation (Wei et al., 2022) dataset which contains questions asking the model to concatenate the last letters of each word.

GPT-4 The accuracy for this task has nearly plateaued, as GPT-4 reaches 92.6 p.p in the 0-shot setting. IAO prompting improves over the baseline reaching 94.7 p.p in accuracy.

Question	The original price of an item is discounted 22%. A customer buys the item at this discounted price using a \$20-off coupon. There is no tax on the item, and this was the only item the customer bought. If the customer paid \$1.90 more than half the original price of the item, what was the original price of the item? Answer Choices: (A) \$61 (B) \$65 (C) \$67.40 (D) \$70 (E) \$78.20 (GT: (E))					
PALM-2	CoT	Let x be the original price of the item. The discounted price is 0.78x. The customer paid 0.78x - 20. This is \$1.90 more than half the original price, which is 0.5x + 1.90. So, 0.78x - 20 = 0.5x + 1.90. Solving for x, we get x = 70. Therefore, among A through E, the answer is (D).				
	IAO	Subquestion	Input	Action Output		
		1. What is the discounted price of the item?	Original price = x, Discount = 22%	Discounted price = Original price - (Original price * Discount) Discounted price = x - (x * 0.22) = 0.78x		
		2. What is the price the customer paid for the item?	Discounted price = 0.78x, Coupon = \$20	Price paid = Discounted price - Coupon Price paid = 0.78x - 20		
		3. What is the equation that related the price paid to the original price?	3. Price paid = 0.78x - 20, Price paid = 0.5x + 1.90	Set the two expressions for the price paid equal to each other. 0.78x - 20 = 0.5x + 1.90		
		4. What is the original price of the item?	0.78x - 20 = 0.5x + 1.90	Solve for x. x = 78.20		
		Let's denote the original price of the item as P. The item is discounted by 22%, so the customer pays 78% of the original price. We can express this as 0.78P. Then, the customer uses a \$20-off coupon, so we subtract \$20 from the discounted price: 0.78P - 20. According to the problem, the customer paid \$1.90 more than half the original price of the item. Half the original price is $\frac{P}{2}$, so the amount the customer paid can also be expressed as $\frac{P}{2} + 1.90$. Now we can set up the equation: $0.78P - 20 = \frac{P}{2} + 1.90$ To solve for P, we first get rid of the fractions by multiplying every term by 2: $2(0.78P) - 2(20) = 2(\frac{P}{2}) + 2(1.90)$ $1.56P - 40 = P + 3.80$ Now, we subtract P from both sides to get: $0.56P - 40 = 3.80$ Next, we add 40 to both sides: $0.56P = 43.80$ Finally, we divide both sides by 0.56 to solve for P: $P = \frac{43.80}{0.56}$ $P = 78.2142857...$ The closest answer choice to \$78.21 is \$78.20. Therefore, among A through E, the answer is (E) \$78.20 .				
		CoT				
		GPT-4	IAO	Subquestion	Input	Action Output
				1. What is the equation that represents the relationship between the original price, the discount, the coupon, and the amount the customer paid?	Let the original price be P. The customer gets a 22% discount and then uses a \$20-off coupon. The customer paid \$1.90 more than half the original price.	Write the equation based on the given information. P - 0.22P - 20 = 0.5P + 1.90
		2. How can we simplify the equation from Step 1 to solve for P?	P - 0.22P - 20 = 0.5P + 1.90	Combine like terms and isolate P. 0.78P - 20 = 0.5P + 1.90		
		3. What is the value of P?	0.78P - 20 = 0.5P + 1.90	Solve for P. P = (20 + 1.90) / (0.78 - 0.5)		
		4. What is the numerical value of P?	(20 + 1.90) / (0.78 - 0.5)	Perform the arithmetic to find P. P = 21.90 / 0.28		
		5. What is the final calculation to determine the original price?	21.90 / 0.28	Calculate the value of P. P = 78.2142857...		

Table 4: Output example from AQuA dataset. The final results are in **bold**.

PALM-2 Here the gains are more substantial. IAO prompting improves the accuracy by 11.6 p.p reaching 88.8 p.p over a baseline of 77.2. From the examples we observe how dividing the problem into multiple subproblems and solving each one in a structured way increases the model’s reasoning abilities.

Prompt	Average
[Step, Input, Action, Output]	68.9
[Step, Subquestion, Action, Output]	77.7
[Step, Subquestion, Input, Action]	76.3
[Step, Subquestion, Input, Action]	59.0
[Step, Subquestion, Input, Action, Output]	80.9

Table 5: Performance in a 0-shot setting when one of the field of the prompt is removed.

5.5 Ablation studies

To gain a deeper understanding of the proposed model’s inner workings and identify its crucial components, we conduct a series of ablation studies by removing individual components of the template while evaluating its performance in the designated tasks. Through this process, we aim to achieve three key objectives: (i) isolate the impact of each field in the proposed template, (ii) iden-

tify redundancies and (iii) enhance interpretability. The results are presented in Table 5 and Table 10. First, we observe that the prompt with all fields ([Step, Subquestion, Input, Action, Output]) achieves the highest average performance (80.9 p.p), indicating that including all available information is beneficial. Including the Subquestion field consistently improves performance compared to excluding it, suggesting that it is important for the model to understand the context of the task. Removing the Output field from the prompt leads to a significant drop in performance (12.0 p.p). This suggests that the model is able to use the output information from previous steps to improve its performance on subsequent steps. Removing the Action field also results in a noticeable decrease in performance (3.2 p.p). This suggests that the model is able to use the action information to better understand the context of the task.

6 Discussion

Our experimental results confirmed the effectiveness of our proposed prompting approach under the

Question	Could all of the people who pass through 30th Street Station every day fit in Dorton Arena? (GT: (No))			
PALM-2	CoT	30th Street Station is the busiest train station in Philadelphia, with an average of 120,000 passengers passing through each day. Dorton Arena is a multi-purpose arena in Raleigh, North Carolina, with a seating capacity of 5,000 people. Therefore, the answer is No		
	IAO	Subquestion	Input	Action
		1. How many people pass through 30th Street Station every day?	30th Street Station is the busiest train station in Philadelphia, with an average of 120,000 passengers passing through each day. Dorton Arena is a multi-purpose arena located in Raleigh, North Carolina. It has a seating capacity of 10,000 people.	The number of people who pass through 30th Street Station every day is 120,000.
		2. What is the capacity of Dorton Arena?	The number of people who pass through 30th Street Station every day is 120,000. The capacity of Dorton Arena is 10,000 people.	The capacity of Dorton Arena is 10,000 people.
GPT-4	CoT	3. Can all of the people who pass through 30th Street Station every day fit in Dorton Arena?	The number of people who pass through 30th Street Station every day is 120,000. The capacity of Dorton Arena is 10,000 people.	The number of people who pass through 30th Street Station every day is greater than the capacity of Dorton Arena.
		To determine if all of the people who pass through 30th Street Station every day could fit in Dorton Arena, we need to know two key pieces of information: 1. The average daily number of people passing through 30th Street Station. 2. The seating capacity of Dorton Arena. Once we have these two figures, we can compare them to see if the arena can accommodate the daily passengers of 30th Street Station. Let's assume we have the necessary data: - The average daily number of people passing through 30th Street Station (we'll call this number X). - The seating capacity of Dorton Arena (we'll call this number Y). If X is less than or equal to Y, then the answer is Yes, all of the people can fit in Dorton Arena. If X is greater than Y, then the answer is No, they cannot all fit. Since I don't have the current data, you would need to provide the average daily passenger count for 30th Street Station and the seating capacity of Dorton Arena for me to complete this comparison.		
		The model fails to choose an answer.		
	IAO	Subquestion	Input	Action
		1. What is the daily passenger traffic at 30th Street Station?	We need to find or estimate the number of people passing through 30th Street Station every day.	Research or estimate the daily passenger traffic at 30th Street Station.
		2. What is the seating capacity of Dorton Arena?	We need to find the seating capacity of Dorton Arena.	Research the seating capacity of Dorton Arena.
		3. Can the daily passenger traffic at 30th Street Station fit in Dorton Arena?	[Output from Step 1] and [Output from Step 2]	Compare 100,000 (daily passenger traffic) with 7,610 (seating capacity of Dorton Arena).
		Therefore, the answer is No.		

Table 6: Output example from StrategyQA dataset. The final results are in **bold**.

zero-shot setting. We summarize the advantages of our method compared to conventional chain-of-thought method as follows: (i) **Clarity and structure**: IAO requires each step of the reasoning process to be explicit making the reasoning process clear and easy to follow. This forces the LLMs to articulate their thoughts and reasoning explicitly. On the other hand, in the conventional zero-shot-CoT steps might be implicit or unclear, requiring more back-and-forth to understand. (ii) **Focus and Guidance**: IAO subquestions help focus the LLM on specific aspects of the problem, leading to more relevant outputs. Contrarily, zero-shot-CoT can contain open-ended discussions, which may lead to tangent or irrelevant information or can be prone to ambiguity and implicit assumptions. (iii) **Transparency**: IAO prompting makes the thought process visible, understandable and easily interpretable. In fact, sometimes it occurs that 0-shot-CoT produces a wrong reasoning chain but finds the correct answers. In this case, it is difficult to understand the process that led to that conclusion. (iv) **Integration with tools**: Open-ended discussions are less suited for integration with external tools (like calculator and python shell) that can further aid the math computation within the arithmetic domain (Gao et al., 2023). On the contrary, the structured format of IAO makes it easier to call ex-

ternal tools to compute and produce results that can successfully be integrated as intermediate outputs in the chain of reasoning.

6.1 Case studies

Table 4 compares the solutions of two large language models (LLMs), PALM-2 and GPT-4, to a math word problem from the AQuA dataset. Using the 0-shot-CoT setting, PALM-2 makes an error in the final calculation, resulting in an incorrect answer of \$70. However, it reaches the correct result using the IAO prompt. This is an example of a point raised earlier in the paper discussing the low scores for the AQuA dataset. Arithmetic errors made in early intermediate steps may lead to incorrect conclusions. On the other hand, GPT-4 appears to be less prone to such mistakes. GPT-4 achieved the correct answer of \$78.20 by systematically breaking down the problem, setting up the equation, and performing calculations in both 0-shot CoT and IAO settings.

Table 6 showcases an example of a real-world reasoning problem from the StrategyQA dataset. The task requires determining whether the daily passenger traffic at 30th Street Station could fit within the seating capacity of Dorton Arena, with the correct answer being "No." This analysis highlights the potential benefits of IAO in guiding

LLMs towards accurate solutions. In the 0-shot-CoT setting, PALM-2 achieved the correct answer, leveraging its access to the relevant information ("120,000 passengers" and "5,000 seating capacity"). However, its approach lacked explicit reasoning steps. However, GPT-4 failed to choose an answer without additional information. While its initial analysis outlined the necessary comparison, it couldn't access or process the specific data points. On the other hand, in the case of PALM-2, the reasoning chain doesn't explicitly show IAO intervention, but PALM-2 likely accessed relevant data based on the prompt's context. However, GPT-4 through IAO prompting effectively identified the necessary information (daily passengers, arena capacity), estimated values for the missing data (100,000 passengers, 7,610 capacity) and compared the values, and correctly concluded that the arena wouldn't hold all passengers. Loosely speaking, the IAO prompts provided GPT-4 with a clear roadmap to break down the problem, gather information, and perform the comparison, leading to a well-reasoned, transparent and verifiable solution. More examples are shown in Table 15, Table 11, Table 12, Table 14 and Table 13 in the Appendix.

6.2 Two-stage IAO prompting

In all previous experiments, we utilized a single API call per sample. While effective, a critical question remains: can employing a two-stage API call per sample further enhance IAO results? To answer this, we propose a focused investigation. We chose to test this hypothesis on the dataset with the lowest gains in terms of performance and test it using PALM-2.

Table 7 presents the results of comparing one-stage and two-stage IAO prompting. The table shows the performance of both approaches on two datasets, AQuA and GSM8k, for arithmetic reasoning, as well as the average performance across both datasets. Overall, two-stage IAO achieved the best average performance (73.7%), outperforming both one-stage 0-shot CoT (72.3%) and one-stage IAO (72.7%). On the individual datasets, two-stage IAO achieved the highest score on GSM8k (83.2%), while one-stage 0-shot CoT achieved the highest score on AQuA (66.3%). These results suggest that two-stage IAO prompting can be an effective strategy for improving the performance of language models on various tasks.

These findings stem from employing a two-stage

	AQuA	GSM8k	Average
One-stage 0-shot CoT	66.3	78.2	72.3
One-stage IAO	63.1	83.2	72.7
Two-stage IAO	63.5	83.9	73.7

Table 7: One stage vs two stage prompting comparison.

prompt setup. Initially, the prompt leads the language model (LLM) in dissecting the problem into manageable parts, enhancing its comprehension of the core issue. Following this, a separate prompt guides the extraction of the answer, ensuring focused attention at each stage of the reasoning process.

Two-stage prompting offers higher accuracy but requires more computational resources than single-stage approaches. Choosing between them depends on balancing accuracy with available resources. For simpler tasks emphasizing efficiency, single-stage prompting may suffice. However, for more complex tasks where sacrificing some computational power leads to significantly better accuracy, two-stage prompting is preferred.

6.3 Few-shot extension

Providing the model a handful of examples substantially improves the reasoning abilities of LLMs. IAO can be paired with methods to automatically generate examples such as Auto-CoT (Zhang et al., 2022) or (Yasunaga et al., 2023) to automatically generate structured IAO-type answers to be provided as demonstrations.

7 Conclusion

In this paper, we propose IAO prompting, a novel template-based approach that guides Large Language Models (LLMs) towards decomposing complex problems into manageable steps. By explicitly structuring input information, planned actions, and intermediate outputs, IAO facilitates sequential problem-solving. Our evaluations across diverse reasoning tasks demonstrate performance gains, increased clarity, improved structural coherence, and enhanced transparency within the reasoning chain. Moreover, IAO boasts remarkable domain independence, requiring minimal adaptation to cater to different problem types. This translates to a versatile tool to unlock the full potential of LLMs in tackling intricate reasoning challenges.

Limitations

While IAO prompting demonstrates compelling advantages in terms of performance and transparency, it is crucial to acknowledge potential limitations: **Output Length:** The structured nature of IAO prompts might lead to lengthier responses compared to baseline models. This can have implications on computational costs and real-time applicability, particularly in resource-constrained settings. Future work could explore techniques for compressing the output or developing domain-specific adaptations to mitigate this limitation. **Interpretability Trade-off:** While the structured output enhances interpretability, it is important to consider that some users might prefer more concise summaries. This suggests a potential trade-off between detailed explanations and user preferences. Future work could explore methods to balance the level of detail and provide tailored interpretations based on individual user needs.

Ethical Considerations

While IAO holds promise for improved reasoning, concerns arise regarding potential bias amplification and misuse as per any prompting method for LLMs. Breaking down complex tasks into smaller steps could inadvertently magnify existing biases in the LLM or training data, leading to biased final outputs. This necessitates careful bias detection and mitigation. Additionally, the structured nature could be exploited to "trick" the model, generating harmful or misleading outputs. Safeguards like fact-checking and verification become crucial, especially in sensitive domains. Responsible deployment and use are fundamental to ensure LLMs benefits are harnessed ethically and its risks are minimized.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- 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.
- 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*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.
- Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. **Making pre-trained language models better few-shot learners**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.
- 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.
- 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.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of

655	prompting methods in natural language processing.	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	711
656	<i>ACM Computing Surveys</i> , 55(9):1–35.	Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,	712
657	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine	et al. 2022. Chain-of-thought prompting elicits rea-	713
658	Lee, Sharan Narang, Michael Matena, Yanqi Zhou,	soning in large language models. <i>Advances in Neural</i>	714
659	Wei Li, and Peter J Liu. 2020. Exploring the limits	<i>Information Processing Systems</i> , 35:24824–24837.	715
660	of transfer learning with a unified text-to-text trans-		
661	former. <i>The Journal of Machine Learning Research</i> ,	Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong	716
662	21(1):5485–5551.	Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and	717
663	Laria Reynolds and Kyle McDonell. 2021. Prompt pro-	Denny Zhou. 2023. Large language models as ana-	718
664	gramming for large language models: Beyond the	logical reasoners. <i>arXiv preprint arXiv:2310.01714</i> .	719
665	few-shot paradigm. In <i>Extended Abstracts of the</i>		
666	<i>2021 CHI Conference on Human Factors in Comput-</i>	Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex	720
667	<i>ing Systems</i> , pages 1–7.	Smola. 2022. Automatic chain of thought prompting	721
668		in large language models. In <i>The Eleventh Interna-</i>	722
669		<i>tional Conference on Learning Representations</i> .	723
670	Timo Schick and Hinrich Schütze. 2021. <i>It’s not just</i>		
671	<i>size that matters: Small language models are also few-</i>	Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei,	724
672	<i>shot learners</i> . In <i>Proceedings of the 2021 Conference</i>	Nathan Scales, Xuezhi Wang, Dale Schuurmans,	725
673	<i>of the North American Chapter of the Association</i>	Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022.	726
674	<i>for Computational Linguistics: Human Language</i>	Least-to-most prompting enables complex reasoning	727
675	<i>Technologies</i> , pages 2339–2352, Online. Association	in large language models. In <i>The Eleventh Interna-</i>	728
676	for Computational Linguistics.	<i>tional Conference on Learning Representations</i> .	729
677			
678	Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric	Jin Ziqi and Wei Lu. 2023. <i>Tab-CoT: Zero-shot tabular</i>	730
679	Wallace, and Sameer Singh. 2020. <i>AutoPrompt: Elic-</i>	<i>chain of thought</i> . In <i>Findings of the Association for</i>	731
680	<i>iting Knowledge from Language Models with Auto-</i>	<i>Computational Linguistics: ACL 2023</i> , pages 10259–	732
681	<i>matically Generated Prompts</i> . In <i>Proceedings of the</i>	10277, Toronto, Canada. Association for Computa-	733
682	<i>2020 Conference on Empirical Methods in Natural</i>	tional Linguistics.	734
683	<i>Language Processing (EMNLP)</i> , pages 4222–4235,		
684	Online. Association for Computational Linguistics.	A Datasets	735
685			
686	Kumar Shridhar, Jakub Macina, Mennatallah El-Assady,	All included datasets are in English. We note	736
687	Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan.	that AQuA-RAT, Date Understanding, Object Shuf-	737
688	2022. Automatic generation of socratic subquestions	fling are under the Apache License, Version 2.0.	738
689	for teaching math word problems. In <i>Proceedings</i>	GSM8K and StrategyQA are under the MIT Li-	739
690	<i>of the 2022 Conference on Empirical Methods in</i>	cense. We also note that the datasets may include	740
691	<i>Natural Language Processing</i> , pages 4136–4149.	names of individuals collected from the internet,	741
692		i.e., publicly available facts about a person but not	742
693	Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao,	in an offensive way. The following list shows the	743
694	Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch,	sources of data we used for this study:	744
695	Adam R Brown, Adam Santoro, Aditya Gupta, Adrià		
696	Garriga-Alonso, et al. 2023. Beyond the imitation	• AQuA-RAT: https://github.com/	745
697	game: Quantifying and extrapolating the capabili-	google-deepmind/AQuA	746
698	ties of language models. <i>Transactions on Machine</i>		
699	<i>Learning Research</i> .	• GSM8K: https://github.com/	747
700	Alon Talmor, Jonathan Herzig, Nicholas Lourie, and	openai/grade-school-math	748
701	Jonathan Berant. 2019. Commonsenseqa: A question		
702	answering challenge targeting commonsense knowl-	• StrategyQA: https://github.com/	749
703	edge. In <i>Proceedings of NAACL-HLT</i> , pages 4149–	google/BIGbench/tree/main/	750
704	4158.	bigbench/benchmark_tasks/	751
705	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob	strategyqa	752
706	Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz		
707	Kaiser, and Illia Polosukhin. 2017. <i>Attention is All</i>	• Last Letter: https://github.com/	753
708	<i>you Need</i> . In <i>Advances in Neural Information Pro-</i>	kojima-takeshi188/zero_shot_	754
709	<i>cessing Systems</i> , volume 30. Curran Associates, Inc.	cot/tree/main/dataset/last_	755
710		letters	756
711	Boshi Wang, Xiang Deng, and Huan Sun. 2022. <i>Itera-</i>		
712	<i>tively prompt pre-trained language models for chain</i>	• Date Understanding: from BIG-Bench	757
713	<i>of thought</i> . In <i>Proceedings of the 2022 Conference on</i>	(BIG-bench collaboration, 2021):	758
714	<i>Empirical Methods in Natural Language Processing</i> ,	https://github.com/google/	759
715	pages 2714–2730, Abu Dhabi, United Arab Emirates.	BIG-bench/blob/main/	760
716	Association for Computational Linguistics.		

- Object Tracking: from BIG-Bench (BIG-bench collaboration, 2021): <https://github.com/google/BIG-bench/blob/main/>
- CommonsenseQA: <https://www.tau-nlp.sites.tau.ac.il/commonsenseqa>

Dataset	Avg words	Filename
AQuA	51.9	test.jsonl
GSM8k	46.9	test.jsonl
D.U.	35.0	task.json
O.T.	91.1	three_objects/task.json
L.L	15.0	last_letters.json
C.QA	27.8	dev_rand_split.jsonl
S.QA	9.6	task.json

Table 8: Datasets statistics and files used. D.U: Date Understanding, O.T: Object Tracking, L.L: Last Letters Concatenation, C.QA: CommonsenseQA, S.QA: StrategyQA

Dataset	Answer prompt
AQuA	Therefore, among A through E, the answer is
GSM8k	Therefore, the answer (arabic numerals) is
D.U.	Therefore, among A through F, the answer is
O.T.	Therefore, among A through C, the answer is
L.L.	Therefore, the answer is
C.QA	Therefore, among A through E, the answer is
S.QA	The answer (Yes or No) is

Table 9: Answer extraction prompts used. D.U: Date Understanding, O.T: Object Tracking, L.L: Last Letters Concatenation, C.QA: CommonsenseQA, S.QA: StrategyQA

B Experimental setting

All results reported for PALM-4 are the results of average over 3 runs. The results reported for GPT-4 are over a single run due to higher costs. For each model, the temperature was set to 0, the maximum number of output tokens was set to 1024.

- gpt-4-1106-preview: <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo>
- PALM-2 (text-unicorn): <https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text>

C Ablation studies

The table 10 reveals how removing each field impacts the model’s accuracy on different tasks. This helps understand the importance and relevance of each field in guiding the model towards successful completion of tasks. For example, removing the "Subquestion" field seems to have a significant negative impact on the "Date Understanding" accuracy, suggesting that subquestions might provide crucial context for understanding temporal information. The Last Letter task seems most affected by any field removal, highlighting its dependence on full prompt context.

D Additional Examples

Prompt	Date Understanding	Last Letter	AQuA	CommonsenseQA	Average
[Step, Input, Action, Output]	82.4	46.0	64.6	82.7	68.9
[Step, Subquestion, Input, Output]	81.8	84.8	63.0	81.2	77.7
[Step, Subquestion, Input, Output]	85.9	76.0	61.0	82.5	76.3
[Step, Subquestion, Input, Action]	86.2	4.4	62.6	82.9	59.0
[Step, Subquestion, Input, Action, Output]	88.1	88.8	63.9	83.1	80.9

Table 10: Performance in a 0-shot setting when one of the field of the prompt is removed.

PALM-2	Question	Today's meeting is rescheduled to 11 am tomorrow, 10/16/1924. What is the date 10 days ago in MM/DD/YYYY? Answer Choices: (A) 10/05/1924 (B) 05/05/1924 (C) 10/12/1924 (D) 10/06/1924 (E) 11/02/1924 (F) 10/04/1924 (GT: (A))			
	CoT	The date of the meeting is 10/16/1924. 10 days ago from 10/16/1924 is 10/06/1924. Therefore, among A through F, the answer is D			
PALM-2	IAO	Subquestion	Input	Action	Output
	IAO	1. What is the date 10 days ago? 2. What is the answer choice that matches the date 10 days ago?	10/16/1924 10/06/1924	subtract 10 days from the given date compare the date 10 days ago with the answer choices	10/06/1924 (D) 10/06/1924
GPT-4	CoT	To find the date 10 days before 10/16/1924, we need to subtract 10 days from that date. Starting with 10/16/1924: 1. If we subtract 10 days from the 16th, we would be in the previous month, October, since October has 31 days. 2. To find the exact date, we subtract 10 from 16, which gives us 6. Therefore, the date 10 days before 10/16/1924 is 10/06/1924. Among the answer choices provided: (A) 10/05/1924 (B) 05/05/1924 (C) 10/12/1924 (D) 10/06/1924 (E) 11/02/1924 (F) 10/04/1924 The correct answer is (D) 10/06/1924.			
	IAO	Subquestion	Input	Action	Output
GPT-4	IAO	1. What is the date 10 days before 10/16/1924? To find the date 10 days before 10/16/1924, we need to count backwards day by day. Since October has 31 days, subtracting 10 days from the 16th will not take us into the previous month. Therefore, we simply subtract 10 from 16. 16 - 10 = 6 The month (October) and the year (1924) remain the same. Therefore, the date 10 days before 10/16/1924 is 10/06/1924. Therefore, among A through F, the answer is (D) 10/06/1924	The given date is 10/16/1924.	Subtract 10 days from the given date.	The date 10 days before 10/16/1924.

Table 11: Output examples from Date Understanding dataset. The final results are in **bold**.

PALM-2	Question	Take the last letters of each words in "Pauline Kerry Jeannette Hope" and concatenate them. (GT: "eyee")			
	CoT	The last letters of each word in "Pauline Kerry Jeannette Hope" are "e", "y", "e", and "e". Concatenating them, we get "eyeee". Therefore, the answer is eyeee .			
PALM-2	IAO	Subquestion	Input	Action	Output
	IAO	1. What are the last letters of each word in "Pauline Kerry Jeannette Hope"? 2. Concatenate the last letters of each word.	"Pauline Kerry Jeannette Hope" "e y e e"	Get the last letters of each word Concatenate the letters.	"e y e e" "eyee"
GPT-4	CoT	To solve this, we need to take the last letter of each word in the phrase "Pauline Kerry Jeannette Hope" and then concatenate them, which means to link them together in a sequence. Here are the last letters of each word: - Pauline: e - Kerry: y - Jeannette: e - Hope: e Now, we concatenate these letters: e + y + e + e Therefore, the answer is "eyee"			
	IAO	Subquestion	Input	Action	Output
GPT-4	IAO	1. What is the last letter of the first word "Pauline"? 2. What is the last letter of the second word "Kerry"? 3. What is the last letter of the third word "Jeannette"? 4. What is the last letter of the fourth word "Hope"? 3. Concatenate the last letters of each word.	The word "Pauline" The word "Kerry" The word "Jeannette" The word "Hope" The last letters "e", "y", "e", "e"	Identify the last letter of the word "Pauline". Identify the last letter of the word "Kerry". Identify the last letter of the word "Jeannette". Identify the last letter of the word "Hope". Concatenate the letters "e", "y", "e", "e" in the order they were obtained.	The last letter is "e". The last letter is "y". The last letter is "e". The last letter is "e". The concatenated result is "eyee" .

Table 12: Output examples from Last Letters dataset. The final results are in **bold**.

PALM-2	Question	Alice, Bob, and Claire are friends and avid readers who occasionally trade books. At the start of the semester, they each buy one new book: Alice gets The Great Gatsby, Bob gets The Odyssey, and Claire gets Lolita. As the semester proceeds, they start trading around the new books. First, Alice and Bob swap books. Then, Alice and Claire swap books. Finally, Bob and Claire swap books. At the end of the semester, Bob has Which choice is true ? Answer Choices: (A) The Great Gatsby. (B) The Odyssey. (C) Lolita. <div>GT: (B) The Odyssey</div>			
	CoT	At the start of the semester, Alice has The Great Gatsby, Bob has The Odyssey, and Claire has Lolita. After Alice and Bob swap books, Alice has The Odyssey and Bob has The Great Gatsby. After Alice and Claire swap books, Alice has Lolita and Claire has The Odyssey. After Bob and Claire swap books, Bob has Lolita and Claire has The Great Gatsby. Therefore, among A through C, the answer is C.			
GPT-4	IAO	Subquestion	Input	Action	Output
		1. What book does Bob have after the first swap?	Alice has The Great Gatsby, Bob has The Odyssey.	Bob and Alice swap books.	Bob has The Great Gatsby.
		2. What book does Bob have after the second swap?	Bob has The Great Gatsby, Claire has Lolita.	Bob and Claire swap books.	Bob has The Great Gatsby.
		3. What book does Bob have at the end of the semester?	Bob has The Great Gatsby.	No more swaps.	Bob has The Great Gatsby .
GPT-4	CoT	Let's track the movement of the books step by step: 1. Initially: - Alice has The Great Gatsby. - Bob has The Odyssey. - Claire has Lolita. 2. After Alice and Bob swap books: - Alice has The Odyssey. - Bob has The Great Gatsby. - Claire still has Lolita. 3. Then Alice and Claire swap books: - Alice has Lolita. - Bob still has The Great Gatsby. - Claire has The Odyssey. 4. Finally, Bob and Claire swap books: - Alice still has Lolita. - Bob has The Odyssey. - Claire has The Great Gatsby. At the end of the semester, Bob has The Odyssey. Therefore, among A through C, the answer is (B) The Odyssey .			
	IAO	Subquestion	Input	Action	Output
		1. What book does Bob have after he and Alice swap books?	Alice has The Great Gatsby, Bob has The Odyssey.	Swap the books between Alice and Bob.	Alice now has The Odyssey, Bob now has The Great Gatsby.
		2. What book does Alice have after she and Claire swap books?	Alice has The Odyssey, Claire has Lolita.	Swap the books between Alice and Claire.	Alice now has Lolita, Claire now has The Odyssey.
	3. What book does Bob have after he and Claire swap books?	Bob has The Great Gatsby, Claire has The Odyssey.	Swap the books between Bob and Claire.	Bob now has The Odyssey , Claire now has The Great Gatsby.	

Table 13: Output examples from Object Tracking dataset. The final results are in **bold**.

PALM-2	Question	Andrea has 8 more apples than Jamal and half as many bananas as Jamal. Jamal has 4 more bananas than apples. How many fruits are there if Andrea has 52 apples? (GT: 168)																													
	CoT	<p>Let's denote the number of apples Jamal has as A and the number of bananas Jamal has as B.</p> <p>From the given information, we know that Andrea has 8 more apples than Jamal, so Andrea has $A + 8$ apples. We also know that Andrea has half as many bananas as Jamal, so Andrea has $B/2$ bananas.</p> <p>We are given that Andrea has 52 apples, so we can write the equation $A + 8 = 52$. Solving for A, we get $A = 44$.</p> <p>We are also given that Jamal has 4 more bananas than apples, so we can write the equation $B = A + 4$.</p> <p>Substituting the value of A, we get $B = 44 + 4 = 48$.</p> <p>Now we can calculate the total number of fruits. Andrea has $A + 8$ apples and $B/2$ bananas, so she has $44 + 8 + 48/2 = 72$ fruits.</p> <p>Jamal has A apples and B bananas, so he has $44 + 48 = 92$ fruits.</p> <p>Therefore, the total number of fruits is $72 + 92 = 164$.</p> <p>Therefore, the answer is 164.</p>																													
	IAO	<table><thead><tr><th>Subquestion</th><th>Input</th><th>Action</th><th>Output</th></tr></thead><tbody><tr><td>1. How many apples does Jamal have?</td><td>Andrea has 8 more apples than Jamal. Andrea has 52 apples.</td><td>Jamal has $52 - 8 = 44$ apples.</td><td>Jamal has 44 apples.</td></tr><tr><td>2. How many bananas does Jamal have?</td><td>Jamal has 4 more bananas than apples. Jamal has 44 apples.</td><td>Jamal has $44 + 4 = 48$ bananas.</td><td>Jamal has 48 bananas.</td></tr><tr><td>3. How many bananas does Andrea have?</td><td>Andrea has half as many bananas as Jamal. Jamal has 48 bananas.</td><td>Andrea has $48 / 2 = 24$ bananas.</td><td>Andrea has 24 bananas.</td></tr><tr><td>4. How many fruits are there in total?</td><td>Andrea has 52 apples and 24 bananas. Jamal has 44 apples and 48 bananas.</td><td>There are $52 + 24 + 44 + 48 = 168$ fruits in total.</td><td>There are 168 fruits in total.</td></tr></tbody></table>	Subquestion	Input	Action	Output	1. How many apples does Jamal have?	Andrea has 8 more apples than Jamal. Andrea has 52 apples.	Jamal has $52 - 8 = 44$ apples.	Jamal has 44 apples.	2. How many bananas does Jamal have?	Jamal has 4 more bananas than apples. Jamal has 44 apples.	Jamal has $44 + 4 = 48$ bananas.	Jamal has 48 bananas.	3. How many bananas does Andrea have?	Andrea has half as many bananas as Jamal. Jamal has 48 bananas.	Andrea has $48 / 2 = 24$ bananas.	Andrea has 24 bananas.	4. How many fruits are there in total?	Andrea has 52 apples and 24 bananas. Jamal has 44 apples and 48 bananas.	There are $52 + 24 + 44 + 48 = 168$ fruits in total.	There are 168 fruits in total.									
	Subquestion	Input	Action	Output																											
1. How many apples does Jamal have?	Andrea has 8 more apples than Jamal. Andrea has 52 apples.	Jamal has $52 - 8 = 44$ apples.	Jamal has 44 apples.																												
2. How many bananas does Jamal have?	Jamal has 4 more bananas than apples. Jamal has 44 apples.	Jamal has $44 + 4 = 48$ bananas.	Jamal has 48 bananas.																												
3. How many bananas does Andrea have?	Andrea has half as many bananas as Jamal. Jamal has 48 bananas.	Andrea has $48 / 2 = 24$ bananas.	Andrea has 24 bananas.																												
4. How many fruits are there in total?	Andrea has 52 apples and 24 bananas. Jamal has 44 apples and 48 bananas.	There are $52 + 24 + 44 + 48 = 168$ fruits in total.	There are 168 fruits in total.																												
GPT-4	CoT	<p>First, let's determine the number of apples and bananas Jamal has.</p> <p>According to the information given, Andrea has 8 more apples than Jamal. Since Andrea has 52 apples, we can calculate the number of apples Jamal has by subtracting 8 from Andrea's total:</p> <p>Jamal's apples = Andrea's apples - 8</p> <p>Jamal's apples = $52 - 8$</p> <p>Jamal's apples = 44</p> <p>Now we know Jamal has 44 apples.</p> <p>Next, we are told that Jamal has 4 more bananas than apples. Since we know Jamal has 44 apples, we can find out how many bananas he has:</p> <p>Jamal's bananas = Jamal's apples + 4</p> <p>Jamal's bananas = $44 + 4$</p> <p>Jamal's bananas = 48</p> <p>So, Jamal has 48 bananas.</p> <p>Now, let's find out how many bananas Andrea has. We are told that Andrea has half as many bananas as Jamal. So we can calculate this as follows:</p> <p>Andrea's bananas = Jamal's bananas / 2</p> <p>Andrea's bananas = $48 / 2$</p> <p>Andrea's bananas = 24</p> <p>Now we can calculate the total number of fruits by adding the number of apples and bananas both Andrea and Jamal have:</p> <p>Total fruits = Andrea's apples + Andrea's bananas + Jamal's apples + Jamal's bananas</p> <p>Total fruits = $52 + 24 + 44 + 48$</p> <p>Total fruits = $76 + 92$</p> <p>Total fruits = 168</p> <p>Therefore, the answer is 168.</p>																													
	IAO	<table><thead><tr><th>Subquestion</th><th>Input</th><th>Action</th><th>Output</th></tr></thead><tbody><tr><td>How many apples does Jamal have?</td><td>Andrea has 8 more apples than Jamal. Andrea has 52 apples.</td><td>Subtract 8 from the number of apples Andrea has to find out how many apples Jamal has.</td><td>Jamal has $52 - 8 = 44$ apples.</td></tr><tr><td>How many bananas does Jamal have?</td><td>Jamal has 4 more bananas than apples. Jamal has 44 apples.</td><td>Add 4 to the number of apples Jamal has to find out how many bananas he has.</td><td>Jamal has $44 + 4 = 48$ bananas.</td></tr><tr><td>How many bananas does Andrea have?</td><td>Andrea has half as many bananas as Jamal. Jamal has 48 bananas.</td><td>Divide the number of bananas Jamal has by 2 to find out how many bananas Andrea has.</td><td>Andrea has $48 / 2 = 24$ bananas.</td></tr><tr><td>How many fruits does Andrea have in total?</td><td>Andrea has 52 apples and 24 bananas.</td><td>Add the number of apples and bananas Andrea has.</td><td>Andrea has $52 + 24 = 76$ fruits.</td></tr><tr><td>How many fruits does Jamal have in total?</td><td>Jamal has 44 apples and 48 bananas.</td><td>Add the number of apples and bananas Jamal has.</td><td>Jamal has $44 + 48 = 92$ fruits.</td></tr><tr><td>How many fruits are there in total?</td><td>Andrea has 76 fruits, and Jamal has 92 fruits.</td><td>Add the number of fruits Andrea has to the number of fruits Jamal has.</td><td>There are $76 + 92 = 168$ fruits in total.</td></tr></tbody></table>	Subquestion	Input	Action	Output	How many apples does Jamal have?	Andrea has 8 more apples than Jamal. Andrea has 52 apples.	Subtract 8 from the number of apples Andrea has to find out how many apples Jamal has.	Jamal has $52 - 8 = 44$ apples.	How many bananas does Jamal have?	Jamal has 4 more bananas than apples. Jamal has 44 apples.	Add 4 to the number of apples Jamal has to find out how many bananas he has.	Jamal has $44 + 4 = 48$ bananas.	How many bananas does Andrea have?	Andrea has half as many bananas as Jamal. Jamal has 48 bananas.	Divide the number of bananas Jamal has by 2 to find out how many bananas Andrea has.	Andrea has $48 / 2 = 24$ bananas.	How many fruits does Andrea have in total?	Andrea has 52 apples and 24 bananas.	Add the number of apples and bananas Andrea has.	Andrea has $52 + 24 = 76$ fruits.	How many fruits does Jamal have in total?	Jamal has 44 apples and 48 bananas.	Add the number of apples and bananas Jamal has.	Jamal has $44 + 48 = 92$ fruits.	How many fruits are there in total?	Andrea has 76 fruits, and Jamal has 92 fruits.	Add the number of fruits Andrea has to the number of fruits Jamal has.	There are $76 + 92 = 168$ fruits in total.	
Subquestion	Input	Action	Output																												
How many apples does Jamal have?	Andrea has 8 more apples than Jamal. Andrea has 52 apples.	Subtract 8 from the number of apples Andrea has to find out how many apples Jamal has.	Jamal has $52 - 8 = 44$ apples.																												
How many bananas does Jamal have?	Jamal has 4 more bananas than apples. Jamal has 44 apples.	Add 4 to the number of apples Jamal has to find out how many bananas he has.	Jamal has $44 + 4 = 48$ bananas.																												
How many bananas does Andrea have?	Andrea has half as many bananas as Jamal. Jamal has 48 bananas.	Divide the number of bananas Jamal has by 2 to find out how many bananas Andrea has.	Andrea has $48 / 2 = 24$ bananas.																												
How many fruits does Andrea have in total?	Andrea has 52 apples and 24 bananas.	Add the number of apples and bananas Andrea has.	Andrea has $52 + 24 = 76$ fruits.																												
How many fruits does Jamal have in total?	Jamal has 44 apples and 48 bananas.	Add the number of apples and bananas Jamal has.	Jamal has $44 + 48 = 92$ fruits.																												
How many fruits are there in total?	Andrea has 76 fruits, and Jamal has 92 fruits.	Add the number of fruits Andrea has to the number of fruits Jamal has.	There are $76 + 92 = 168$ fruits in total.																												

Table 14: Output examples from GSM8k dataset. The final results are in **bold**.

PALM-2	Question	The man tried to reply to the woman, but he had difficulty keeping track of conversations that he didn't do what to? Answer Choices: (A) initiate (B) ignore (C) question (D) answer (E) ask (GT: (A))			
	CoT	<p>The man tried to reply to the woman, but he had difficulty keeping track of conversations that he didn't do what to?</p> <p>The man tried to reply to the woman, but he had difficulty keeping track of conversations. This means that he was not able to follow the conversation properly. He was not able to understand what the woman was saying. This could be because he was not paying attention or because he was not able to understand the language.</p> <p>The question asks what the man did not do.</p> <p>The answer choices are: (A) initiate (B) ignore (C) question (D) answer (E) ask.</p> <p>The man did not initiate the conversation. He was replying to the woman. He did not ignore the woman. He was trying to reply to her. He did not question the woman. He was trying to understand what she was saying. He did not answer the woman. He was trying to reply to her. He did not ask the woman anything. He was trying to understand what she was saying.</p> <p>Therefore, among A through E, the answer is (D)</p>			
	IAO	Subquestion	Input	Action	Output
		1. What is the main idea of the question?	The man tried to reply to the woman, but he had difficulty keeping track of conversations that he didn't do what to?	Identify the main idea of the question.	The main idea of the question is that the man had difficulty keeping track of conversations that he didn't do something to.
GPT-4		2. What is the missing word in the question?	The man tried to reply to the woman, but he had difficulty keeping track of conversations that he didn't do what to?	Identify the missing word in the question.	The missing word in the question is "initiate".
		3. What is the answer to the question?	The man tried to reply to the woman, but he had difficulty keeping track of conversations that he didn't do what to?	Choose the answer that best fits the question.	The answer to the question is (A) initiate.
	CoT	<p>To solve this reasoning problem, we need to determine which action would make it difficult for the man to keep track of conversations. If he has difficulty keeping track of conversations, it is likely because he is not actively engaged in starting or controlling the flow of the conversation.</p> <p>Let's look at the options:</p> <p>(A) initiate - If he doesn't initiate the conversation, he might have trouble keeping track because he is not setting the topic or pace of the conversation.</p> <p>(B) ignore - Ignoring a conversation would not cause difficulty in keeping track; it would mean he is not participating at all.</p> <p>(C) question - Not questioning might make it harder to understand, but it doesn't necessarily mean he would lose track of the conversation.</p> <p>(D) answer - If he doesn't answer, it could be a result of losing track, not a cause.</p> <p>(E) ask - Similar to initiate, if he doesn't ask questions, he might not be engaged enough to follow the conversation.</p> <p>Between initiating and asking, initiating a conversation is a more fundamental action for controlling the flow of the conversation. If he doesn't initiate conversations, he is more likely to be a passive participant and thus have difficulty keeping track.</p> <p>Therefore, among A through E, the answer is (A) initiate</p>			
	IAO	Subquestion	Input	Action	Output
	1. What does the phrase "difficulty keeping track of conversations" imply about the man's ability to engage in conversations?	The man had difficulty keeping track of conversations.	Analyze the phrase to determine what aspect of conversation the man struggles with.	The man struggles with following along or maintaining focus in conversations that he is not actively leading or starting.	
	2. Which answer choice is related to leading or starting a conversation?	Answer choices (A) initiate (B) ignore (C) question (D) answer (E) ask	Identify the answer choice that means to begin or lead a conversation.	(A) initiate	
	3. Does the difficulty in keeping track of conversations imply that the man did not do what?	The man had difficulty keeping track of conversations; the man tried to reply to the woman.	Determine which action from the answer choices would be most difficult for someone who cannot keep track of conversations.	The man would have difficulty replying (or keeping track) if he did not initiate the conversation.	
	Therefore, among A through E, the answer is (A) initiate.				

Table 15: Output examples from CommonsenseQA dataset. The final results are in **bold**.