

# 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 (CoT) prompting, which makes the intermediate reasoning steps apparent. 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 improves reasoning performance but also interpretability and transparency. Experiments across various reasoning tasks demonstrate IAO’s strong zero-shot capabilities. Human evaluation validates the transparency and interpretability of IAO reasoning chains.

## 1 Introduction

The recent progress in natural language processing (NLP) can be largely attributed to the success 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 many tasks, requiring minimal or 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; Reynolds and McDonnell, 2021) and through automated approaches (Gao et al., 2021; Shin et al., 2020).

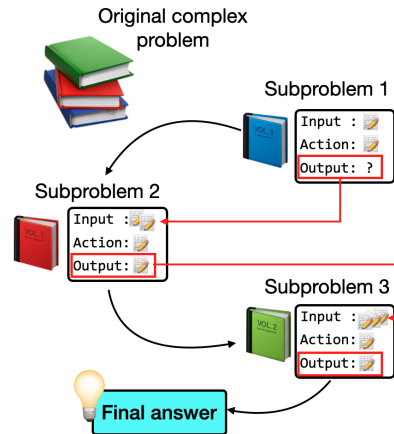


Figure 1: Illustration of IAO prompting with emphasis on the problem decomposition into Input-Action-Output. The intermediate output of each step is the input for subsequent reasoning steps.

One key technique to enhance LLM performance is Chain-of-Thought (CoT) prompting (Wei et al., 2022; Wang et al., 2022), which makes intermediate reasoning steps explicit and for this reason aids in complex decision-making processes. 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. 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 has unveiled that LLMs can be good zero-shot rea-

064	soners.		
065	Despite these advancements, a significant chal-	model the LLM reasoning process step-by-	115
066	lenge remains: ensuring the free-text reason-	step, while applying this structured format	116
067	ing chains of LLMs are interpretable and trans-	within a "chain-of-thought" paradigm.	117
068	parent. According to <a href="#">Singh et al. (2024)</a> , the interpretation		
069	of relevant knowledge from an LLM involves ex-	• We perform extensive experiments across var-	118
070	tracting useful knowledge concerning relationships	ious reasoning tasks and human evaluation of	119
071	either contained in the data or learned by the model.	the reasoning chains to assess the validity of	120
072	Relevant knowledge is knowledge that is useful for	IAO.	121
073	a particular problem and audience. Interpretabil-		
074	ity and transparency of the explanations are not	<b>2 Related Work</b>	122
075	just theoretical concerns but come from practical	<b>Prompting</b> Chain-of-thought (CoT) prompting	123
076	necessities.	provides intermediate rationales in the reasoning	124
077	Current methods, including CoT, improve reason-	chain for solving problems, leading to improved ac-	125
078	ing performance but often fall short of provid-	curacy and interpretability in tasks requiring multi-	126
079	ing a clear, structured mapping of how an output is	step reasoning. The common instance of CoT is	127
080	derived from the given input. In fact, while mak-	0-shot CoT ( <a href="#">Kojima et al., 2022</a> ) and few-shot	128
081	ing some intermediate steps explicit, CoT does not	CoT ( <a href="#">Wei et al., 2022</a> ). The first one is the most	129
082	always offer a detailed breakdown of the entire	relevant to our work. On the other hand, few-	130
083	reasoning process, leaving gaps in traceability.	shot prompting requires manually constructed task-	131
084	This opacity in the model’s decision-making pro-	specific demonstrations, so it is more costly. <a href="#">Zhang</a>	132
085	cess can hinder efforts to recognize errors, provide	<a href="#">et al. (2022)</a> proposed an approach to automatically	133
086	targeted feedback for improvement, or ensure the	construct few-shot demonstrations based on zero-	134
087	model behaves in a manner consistent with ethical	shot prompting to save the manual cost. Another	135
088	or safety considerations. Addressing this challenge	categorization for prompting methods is based on	136
089	requires exploring techniques to enhance the trans-	whether they involve a single step or multiple steps.	137
090	parency and interpretability of language models in	An example of multi-step prompting is the work	138
091	critical domains ( <a href="#">Chen et al., 2024</a> ).	by <a href="#">Zhou et al. (2022)</a> which is discussed in the	139
092	This work introduces IAO (Input-Action-Output)	next paragraph. There are several CoT prompts	140
093	prompting, a simple yet powerful prompt template	with structured reasoning chains in the form of	141
094	that explicitly models the reasoning process in a	programs ( <a href="#">Chen et al., 2023</a> ), graphs ( <a href="#">Besta et al.,</a>	142
095	structured manner, analogous to dissecting a com-	<a href="#">2024</a> ) or tables ( <a href="#">Ziqi and Lu, 2023</a> ). In this work	143
096	plex problem into a series of well-defined, sequen-	we focus on free-text reasoning chains.	144
097	tial subproblems. Each step towards the final an-	<b>Problem Decomposition and Subquestioning</b>	145
098	swer meticulously outlines the information used,	Breaking down complex reasoning into sub-	146
099	the action planned, and the intermediate output	questions enhances problem-solving, as supported	147
100	generated. This transparent breakdown not only	by research illustrating the efficiency of smaller,	148
101	maintains or improves the LLM’s reasoning abili-	focused LLMs ( <a href="#">Shridhar et al., 2022</a> ). <a href="#">Zhou et al.</a>	149
102	ties but IAO facilitates clearer problem decomposi-	<a href="#">(2022)</a> introduced Least-to-Most prompting, which	150
103	tion, guiding the LLM to tackle complex tasks with	sequentially solves a problem by structurally de-	151
104	greater accuracy and efficiency.	composing it into manageable sub-questions, pro-	152
105	Despite the simplicity, IAO prompt successfully	viding clearer reasoning steps compared to other	153
106	generates a plausible reasoning path in a zero-shot	methods like zero-shot CoT. This approach, albeit	154
107	manner and reaches the correct answer in problems	requiring precise prompt design, enhances LLM	155
108	where the standard zero-shot-CoT approach fails	performance. Another similar method, Tab-CoT	156
109	or is not explicit enough.	( <a href="#">Ziqi and Lu, 2023</a> ), utilizes a tabular format to	157
110	We summarize our main contributions in this	organize reasoning steps, offering clear advantages	158
111	work as follows:	in transparency and understanding, despite some-	159
112		times necessitating specific domain knowledge for	160
113	• We propose a new approach, IAO prompt,	prompt design. These methods contrast with IAO	161
114	(input-action-output) that employs a straight-	prompting, which autonomously decomposes prob-	162
	forward template to explicitly structure and	lems and explicitly outlines intermediate steps.	163

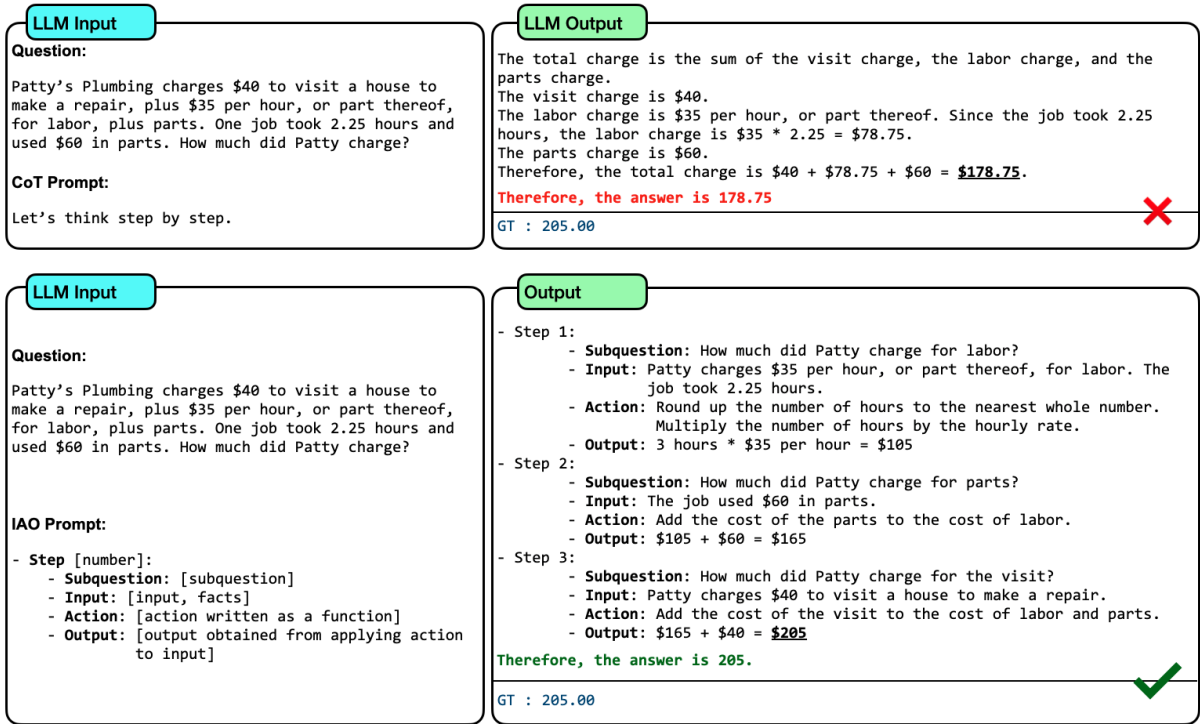


Figure 2: Example of IAO prompting compared to 0-shot CoT. This example is taken from GSM8k using PALM-2. CoT fails to find the correct answer due to overlooking some part of the input information.

### 3 IAO prompting

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

ity to understand eventual mistakes. In short, IAO prompting consists of three simple steps as show in Figure 2: (1) **Subquestion**: Instead of directly addressing the question, the LLM decomposes the problem into smaller problems by formulating subquestions, which it then attempts to answer one at a time in a sequential manner; (2) **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; (3) **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; (4) **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

### 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), Common-

Task	<i>Arithmetic Reasoning</i>		<i>Logical Reasoning</i>		<i>Commonsense Reasoning</i>		<i>Symbolic Reasoning</i>
	AQUA	GSM8K	Date Understanding	Object Tracking	StrategyQA	CommonsenseQA	Last Letter
0-shot CoT	<b>66.3</b>	78.2	86.2	63.1	74.4	80.1	77.2
0-shot IAO (ours)	63.9	<b>82.3</b>	<b>88.1</b>	<b>67.1</b>	<b>76.9</b>	<b>83.1</b>	<b>88.8</b>

Table 1: 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	<i>Arithmetic Reasoning</i>		<i>Logical Reasoning</i>		<i>Commonsense Reasoning</i>		<i>Symbolic Reasoning</i>
	AQUA	GSM8K	Date Understanding	Object Tracking	StrategyQA	CommonsenseQA	Last Letter
0-shot CoT	<b>70.3</b>	92.0	83.1	100	75.6	81.6	92.6
0-shot IAO (ours)	70.2	<b>94.2</b>	<b>83.2</b>	100	<b>76.3</b>	<b>84.8</b>	<b>94.7</b>

Table 2: 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.

senseQA (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 10 presents some dataset information and further details are in Table 9 and Appendix A.

## 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

### 5.1 Arithmetic Reasoning

**GPT-4** From Table 2, 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 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.

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 <b>(GT: (E))</b>			
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 <b>(D)</b> .			
	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 $x$ .	<b><math>x = 78.20</math></b>
GPT-4	CoT	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 <b>(E) \$78.20</b> .			
	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 $P$ .	<b><math>P = 78.2142857...</math></b>

Table 3: Output example from AQuA dataset. The final results are in **bold**.

**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 led 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.

**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, Output]	76.3
[Step, Subquestion, Input, Action]	59.0
[Step, Subquestion, Input, Action, Output]	80.9

Table 4: Performance in a 0-shot setting when one of the field of the prompt is removed.

## 5.5 Ablation studies

To better understand the proposed model and identify its key components, we perform ablation studies by removing parts of the template and assessing performance on specific tasks. Through this process, we aim to achieve three key objectives: (i) isolate the impact of each field in the proposed template, (ii) identify redundancies and (iii) enhance interpretability. The results are presented in Table 4 and Table 8 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 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 external tools to compute and produce results that can successfully be integrated as intermediate outputs in the chain of reasoning.

## 6.1 Case studies

Table 3 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 5 showcases an example of a real-world reasoning problem from the StrategyQA dataset. The task requires determining whether the daily

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			
		Subquestion	Input	Action	Output
	IAO	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.	The number of people who pass through 30th Street Station every day is 120,000.	120,000
		2. What is the capacity of Dorton Arena?	Dorton Arena is a multi-purpose arena located in Raleigh, North Carolina. It has a seating capacity of 10,000 people.	The capacity of Dorton Arena is 10,000 people.	10,000
		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.	<b>No</b>
GPT-4	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:				
	CoT	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.			
	<b>The model fails to choose an answer.</b>				
		Subquestion	Input	Action	Output
	IAO	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.	Let's say the daily passenger traffic at 30th Street Station is 100,000 people.
		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.	Let's assume Dorton Arena has a seating capacity of 7,610 people.
		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).	100,000 is much greater than 7,610.
	<b>Therefore, the answer is No.</b>				

Table 5: Output example from StrategyQA dataset. The final results are in **bold**.

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 16, Table 13,

Table 12, Table 15 and Table 14 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 6 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 are derived from a two-stage prompt approach where the first stage assists the

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

Table 6: One stage vs two stage prompting comparison.

LLM in breaking down the problem, improving comprehension, and the second stage focuses on extracting the answer. This sequential prompting enhances accuracy but demands more computational resources than single-stage prompting. The choice between them hinges on the trade-off between accuracy and resource availability. For simpler tasks, single-stage prompting may be adequate, while complex tasks justify the extra computational expense of two results in improved accuracy.

### 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. We test 3-shot CoT and 3-shot IAO for the AQuA dataset, the most challenging for both LLMs. 3-shot IAO improves the accuracy by 5 points (76.4 from 71.9) for GPT-4 and by 0.8 in PALM-2 (69.2 from 68.4).

### 6.4 Human evaluation

We supplement the evaluation protocol with human evaluation. We asked 120 human crowdworkers from CloudConnect Research questions aimed at gauging the interpretability and transparency of IAO compared to CoT. For doing this, we sample 10 questions that both IAO and CoT responded correctly in terms of accuracy and 10 questions were both answers were wrong. This is done for the most challenging dataset for each reasoning type, hence AQuA, StrategyQA and Date Understanding. For each question, we present the annotators with the IAO and CoT reasoning chains. We ask three questions: (1a) *if the answers are correct, which reasoning text is more useful?* (1b) *if the answers are wrong, which reasoning text do you prefer to spot the reasoning mistake?* (2) *which reasoning text is more transparent?* (3) *which reasoning text is easier to interpret?* Table 7 presents the key

findings.

Question	AQuA	StrategyQA	Date
(1a) Useful	12	9	27
(1b) Efficiency	86	73	87
(2) Transparency	64	76	89
(3) Interpretability	74	71	83

Table 7: Human evaluation of IAO and CoT reasoning chains. The values reported are percentages of times IAO is chosen over CoT.

In general, when the answer is correct, IAO is perceived as notably less useful, with only 12% and 9% preference, in AQuA and StrategyQA respectively. This is to be expected, as IAO prompts generates longer reasoning chains compared to CoT and favours redundancy of information. However, when the answer is wrong, IAO is much more favored, with preferences being: 86% for AQuA, 73% for StrategyQA, and 87% for Date. This suggests that IAO is considered more efficient compared to CoT for determining errors in the reasoning process. In general, IAO is preferred in terms of transparency and interpretability of the reasoning text as reported by the third and fourth row, indicating a clearer and understandable presentation in IAO compared to CoT reasoning chains.

## 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 shows domain independence, requiring minimal adaptation to cater to different problem types. Additional human evaluation validates IAO utility in finding errors in free-text reasoning chains.

### Limitations

While IAO prompting demonstrates compelling advantages in terms of performance and transparency, it is crucial to acknowledge potential limitations:



553 **Output Length** : The structured nature of IAO  
554 prompts might lead to lengthier responses com-  
555 pared to baseline models. This can have implica-  
556 tions on computational costs and real-time applica-  
557 bility, particularly in resource-constrained settings.  
558 Future work could explore techniques for compress-  
559 ing the output or developing domain-specific adap-  
560 tations to mitigate this limitation.

561 **Interpretability Trade-off** : While the struc-  
562 tured output enhances interpretability, it is impor-  
563 tant to consider that some users might prefer more  
564 concise summaries. The human evaluation study  
565 supports this. In fact, when the answer is correct,  
566 evaluators heavily preferred CoT reasoning chains  
567 over IAO's. This suggests a potential trade-off be-  
568 tween detailed explanations and user preferences.  
569 Future work could explore methods to balance the  
570 level of detail and provide tailored interpretations  
571 based on individual user needs.

## 572 Ethical Considerations

573 While IAO holds promise for improved reasoning,  
574 concerns arise regarding potential bias amplifica-  
575 tion and misuse as per any prompting method for  
576 LLMs. Breaking down complex tasks into smaller  
577 steps could inadvertently magnify existing biases  
578 in the LLM or training data, leading to biased fi-  
579 nal outputs. This necessitates careful bias detec-  
580 tion and mitigation. Additionally, the structured  
581 nature could be exploited to "trick" the model, gen-  
582 erating harmful or misleading outputs. Safeguards  
583 like fact-checking and verification become crucial,  
584 especially in sensitive domains. Responsible de-  
585 ployment and use are fundamental to ensure LLMs  
586 benefits are harnessed ethically and its risks are  
587 minimized.

## 588 References

589 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama  
590 Ahmad, Ilge Akkaya, Florencia Leoni Aleman,  
591 Diogo Almeida, Janko Altenschmidt, Sam Altman,  
592 Shyamal Anadkat, et al. 2023. Gpt-4 technical report.  
593 *arXiv preprint arXiv:2303.08774*.

594 Rohan Anil, Andrew M Dai, Orhan Firat, Melvin John-  
595 son, Dmitry Lepikhin, Alexandre Passos, Siamak  
596 Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng  
597 Chen, et al. 2023. Palm 2 technical report. *arXiv*  
598 *preprint arXiv:2305.10403*.

599 Maciej Besta, Nils Blach, Ales Kubicek, Robert Gersten-  
600 berger, Michal Podstawski, Lukas Gianinazzi, Joanna

Gajda, Tomasz Lehmann, Hubert Niewiadomski, Pi-  
otr Nyczyk, et al. 2024. Graph of thoughts: Solving  
elaborate problems with large language models. In  
*Proceedings of the AAAI Conference on Artificial*  
*Intelligence*, volume 38, pages 17682–17690. 601  
602  
603  
604  
605

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. 606  
607  
608  
609  
610  
611

Wenhu Chen, Xueguang Ma, Xinyi Wang, and  
William W Cohen. 2023. Program of thoughts  
prompting: Disentangling computation from reason-  
ing for numerical reasoning tasks. *Transactions on*  
*Machine Learning Research*. 612  
613  
614  
615  
616

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao,  
An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian  
McAuley, Linda Petzold, and William Yang Wang.  
2024. A survey on large language models for critical  
societal domains: Finance, healthcare, and law.  
*arXiv preprint arXiv:2405.01769*. 617  
618  
619  
620  
621  
622

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin,  
Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul  
Barham, Hyung Won Chung, Charles Sutton, Sebas-  
tian Gehrmann, et al. 2023. Palm: Scaling language  
modeling with pathways. *Journal of Machine Learn-*  
*ing Research*, 24(240):1–113. 623  
624  
625  
626  
627  
628

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*. 629  
630  
631  
632  
633

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and  
Kristina Toutanova. 2019. [BERT: Pre-training of](#)  
[Deep Bidirectional Transformers for Language Un-](#)  
[derstanding](#). In *Proceedings of the 2019 Conference*  
*of the North American Chapter of the Association for*  
*Computational Linguistics: Human Language Tech-*  
*nologies, Volume 1 (Long and Short Papers)*, pages  
4171–4186, Minneapolis, Minnesota. Association for  
Computational Linguistics. 634  
635  
636  
637  
638  
639  
640  
641  
642

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon,  
Pengfei Liu, Yiming Yang, Jamie Callan, and Gra-  
ham Neubig. 2023. Pal: Program-aided language  
models. In *International Conference on Machine*  
*Learning*, pages 10764–10799. PMLR. 643  
644  
645  
646  
647

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021.  
[Making pre-trained language models better few-shot](#)  
[learners](#). In *Proceedings of the 59th Annual Meet-*  
*ing of the Association for Computational Linguistics*  
*and the 11th International Joint Conference on Natu-*  
*ral Language Processing (Volume 1: Long Papers)*,  
pages 3816–3830, Online. Association for Computa-  
tional Linguistics. 648  
649  
650  
651  
652  
653  
654  
655

656	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. <i>Transactions of the Association for Computational Linguistics</i> , 9:346–361.	Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. <i>Transactions on Machine Learning Research</i> .	713 714 715 716 717 718 719
662	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. <i>Advances in neural information processing systems</i> , 35:22199–22213.	Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In <i>Proceedings of NAACL-HLT</i> , pages 4149–4158.	720 721 722 723 724
667	Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In <i>Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 158–167.	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. <b>Attention is All you Need</b> . In <i>Advances in Neural Information Processing Systems</i> , volume 30. Curran Associates, Inc.	725 726 727 728 729
673	Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. <i>ACM Computing Surveys</i> , 55(9):1–35.	Boshi Wang, Xiang Deng, and Huan Sun. 2022. <b>Iteratively prompt pre-trained language models for chain of thought</b> . In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 2714–2730, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	730 731 732 733 734 735
678	Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. <i>The Journal of Machine Learning Research</i> , 21(1):5485–5551.	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in Neural Information Processing Systems</i> , 35:24824–24837.	736 737 738 739 740
684	Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In <i>Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems</i> , pages 1–7.	Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. 2023. Large language models as analogical reasoners. <i>arXiv preprint arXiv:2310.01714</i> .	741 742 743 744
689	Timo Schick and Hinrich Schütze. 2021. <b>It’s not just size that matters: Small language models are also few-shot learners</b> . In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 2339–2352, Online. Association for Computational Linguistics.	Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. In <i>The Eleventh International Conference on Learning Representations</i> .	745 746 747 748
696	Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. <b>AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts</b> . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 4222–4235, Online. Association for Computational Linguistics.	Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In <i>The Eleventh International Conference on Learning Representations</i> .	749 750 751 752 753 754
703	Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 4136–4149.	Jin Ziqi and Wei Lu. 2023. <b>Tab-CoT: Zero-shot tabular chain of thought</b> . In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 10259–10277, Toronto, Canada. Association for Computational Linguistics.	755 756 757 758 759
709	Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. <i>arXiv preprint arXiv:2402.01761</i> .		
		<b>A Datasets</b>	760
		All included datasets are in English. We note that AQuA-RAT, Date Understanding, Object Shuffling are under the Apache License, Version 2.0. GSM8K and StrategyQA are under the MIT License. We also note that the datasets may include	761 762 763 764 765

Prompt	Date Understanding	Last Letter	AQuA	CommonsenseQA	Average
[Step, Input, Action, Output]	82.4	46.0	64.6	82.7	68.9
[Step, Subquestion, Action, 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 8: Performance in a 0-shot setting when one of the field of the prompt is removed.

names of individuals collected from the internet, i.e., publicly available facts about a person but not in an offensive way. The following list shows the sources of data we used for this study:

- AQuA-RAT: <https://github.com/google-deepmind/AQuA>
- GSM8K: <https://github.com/openai/grade-school-math>
- StrategyQA: [https://github.com/google/BIGbench/tree/main/bigbench/benchmark\\_tasks/strategyqa](https://github.com/google/BIGbench/tree/main/bigbench/benchmark_tasks/strategyqa)
- Last Letter: [https://github.com/kojima-takeshi188/zero\\_shot\\_cot/tree/main/dataset/last\\_letters](https://github.com/kojima-takeshi188/zero_shot_cot/tree/main/dataset/last_letters)
- Date Understanding: from BIG-Bench (BIG-bench collaboration, 2021): <https://github.com/google/BIG-bench/blob/main/>
- 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 9: Datasets statistics and files used. D.U: Date Understanding, O.T: Object Tracking, L.L: Last Letters Concatenation, C.QA: CommonsenseQA, S.QA: StrategyQA

## B Ablation studies

The table 8 reveals how removing each field impacts the model’s accuracy on different tasks. This helps understand the importance and relevance of

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 10: Tasks, data statistics and answer type.

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

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.

## C Experimental setting

All results reported for PALM-2 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>

- 810 • PALM-2 (text-unicorn):  
811 [https://cloud.google.com/  
812 vertex-ai/docs/generative-ai/  
813 model-reference/text](https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text)

## 814 **D Additional Examples**

	Question	Take the last letters of each words in "Pauline Kerry Jeannette Hope" and concatenate them. (GT: "eyee")			
PALM-2	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 <b>eyeee</b> .			
	IAO	Subquestion	Input	Action	Output
		1. What are the last letters of each word in "Pauline Kerry Jeannette Hope"?	"Pauline Kerry Jeannette Hope"	Get the last letters of each word	"e y e e"
		2. Concatenate the last letters of each word.	"e y e e"	Concatenate the letters.	<b>"eyee"</b>
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 <b>"eyee"</b>			
	IAO	Subquestion	Input	Action	Output
		1. What is the last letter of the first word "Pauline"?	The word "Pauline"	Identify the last letter of the word "Pauline".	The last letter is "e".
		2. What is the last letter of the second word "Kerry"?	The word "Kerry"	Identify the last letter of the word "Kerry".	The last letter is "y".
		3. What is the last letter of the third word "Jeannette"?	The word "Jeannette"	Identify the last letter of the word "Jeannette".	The last letter is "e".
		4. What is the last letter of the fourth word "Hope"?	The word "Hope"	Identify the last letter of the word "Hope".	The last letter is "e".
		3. Concatenate the last letters of each word.	The last letters "e", "y", "e", "e"	Concatenate the letters "e", "y", "e", "e" in the order they were obtained.	The concatenated result is <b>"eyee"</b> .

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

	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))			
PALM-2	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 <b>D</b>			
	IAO	Subquestion	Input	Action	Output
		1. What is the date 10 days ago?	10/16/1924	subtract 10 days from the given date	10/06/1924
		2. What is the answer choice that matches the date 10 days ago?	10/06/1924	compare the date 10 days ago with the answer choices	<b>(D) 10/06/1924</b>
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 <b>(D) 10/06/1924</b> .			
	IAO	Subquestion	Input	Action	Output
		1. What is the date 10 days before 10/16/1924?	The given date is 10/16/1924.	Subtract 10 days from the given date.	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 <b>(D) 10/06/1924</b>			

Table 13: Output examples from Date Understanding dataset. The final results are in **bold**.

	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. <span style="float: right;">GT: (B) The Odyssey</span>			
PALM-2	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.			
	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 <b>The Great Gatsby</b> .
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>(B) The Odyssey</b> .			
	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 <b>The Odyssey</b> , Claire now has The Great Gatsby.

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

	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)																													
PALM-2	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 <math>A + 8</math> apples. We also know that Andrea has half as many bananas as Jamal, so Andrea has <math>B/2</math> bananas.</p> <p>We are given that Andrea has 52 apples, so we can write the equation <math>A + 8 = 52</math>. Solving for A, we get <math>A = 44</math>.</p> <p>We are also given that Jamal has 4 more bananas than apples, so we can write the equation <math>B = A + 4</math>.</p> <p>Substituting the value of A, we get <math>B = 44 + 4 = 48</math>.</p> <p>Now we can calculate the total number of fruits. Andrea has <math>A + 8</math> apples and <math>B/2</math> bananas, so she has <math>44 + 8 + 48/2 = 72</math> fruits.</p> <p>Jamal has A apples and B bananas, so he has <math>44 + 48 = 92</math> fruits.</p> <p>Therefore, the total number of fruits is <math>72 + 92 = 164</math>.</p> <p>Therefore, the answer is <b>164</b>.</p>																													
	IAO	<table border="1"> <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 <math>52 - 8 = 44</math> 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 <math>44 + 4 = 48</math> 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 <math>48 / 2 = 24</math> 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 <math>52 + 24 + 44 + 48 = 168</math> fruits in total.</td> <td>There are <b>168</b> 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 <b>168</b> 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 <b>168</b> 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 Jamal's apples = <math>52 - 8</math> 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 Jamal's bananas = <math>44 + 4</math> 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 Andrea's bananas = <math>48 / 2</math> 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 Total fruits = <math>52 + 24 + 44 + 48</math> Total fruits = <math>76 + 92</math> Total fruits = 168</p> <p>Therefore, the answer is 168.</p>																													
	IAO	<table border="1"> <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 <math>52 - 8 = 44</math> 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 <math>44 + 4 = 48</math> 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 <math>48 / 2 = 24</math> 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 <math>52 + 24 = 76</math> 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 <math>44 + 48 = 92</math> 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 <math>76 + 92 = 168</math> 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 15: 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 <span style="float: right;">(GT: (A))</span>			
	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? 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. The question asks what the man did not do. 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 <b>(D)</b></p>			
GPT-4	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.
		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 <b>(A) initiate</b> .
	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. (B) ignore - Ignoring a conversation would not cause difficulty in keeping track; it would mean he is not participating at all. (C) question - Not questioning might make it harder to understand, but it doesn't necessarily mean he would lose track of the conversation. (D) answer - If he doesn't answer, it could be a result of losing track, not a cause. (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 <b>(A) initiate</b></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 <b>(A) initiate</b> .			

Table 16: Output examples from CommonsenseQA dataset. The final results are in **bold**.



## E Errors from AQua

**Note:** Several questions where IAO gave the wrong answers were due to the output exceeding the maximum number of tokens.

### E.1 Example 1

*Question:* A grocer makes a 25% profit on the selling price for each bag of flour it sells. If he sells each bag for \$100 and makes \$3,000 in profit, how many bags did he sell?

*Answer Choices:* (A) 12 (B) 16 (C) 24 (D) 30 (E) 40"

*GT:* "C"

#### IAO : Step 1:

- *Subquestion:* What is the profit made on one bag of flour?
- *Input:* 25% profit on the selling price of \$100 per bag
- *Action:* Calculate the profit per bag by multiplying the selling price by the profit percentage.
- *Output:* \$25 profit per bag (since 25% of \$100 is \$25)

#### Step 2:

- *Subquestion:* How many bags of flour were sold to make \$3,000 in profit?
- *Input:* \$3,000 total profit, \$25 profit per bag
- *Action:* Divide the total profit by the profit per bag to find the number of bags sold.
- *Output:* 120 bags sold (since  $\$3,000 / \$25$  per bag = 120 bags)

#### Step 3:

- *Subquestion:* Which answer choice corresponds to the number of bags sold?
- *Input:* 120 bags sold, Answer Choices: (A) 12 (B) 16 (C) 24 (D) 30 (E) 40
- *Action:* Identify the correct answer choice that matches the number of bags sold.
- *Output:* None of the answer choices match the number of bags sold.

Therefore, among answer choices A through E, the answer is none of the provided options. There seems to be a mistake as none of the answer choices correspond to the correct number of bags sold based on the given information.

**CoT :** To solve this problem, we need to determine how much profit the grocer makes on each bag of flour and then use that information to find out how many bags were sold to make a total profit of \$3,000.

#### Step 1: Calculate the profit per bag.

Since the grocer makes a 25% profit on the selling price, we can calculate the profit per bag as follows:

$$\text{Profit per bag} = 25\% \text{ of selling price}$$

$$\text{Profit per bag} = 0.25 \times \$100$$

$$\text{Profit per bag} = \$25$$

#### Step 2: Calculate the number of bags sold.

Now that we know the profit per bag is \$25, we can calculate the number of bags sold to make a total profit of \$3,000.

$$\text{Total profit} = \text{Profit per bag} \times \text{Number of bags sold}$$

$$\$3,000 = \$25 \times \text{Number of bags sold}$$

To find the number of bags sold, we divide the total profit by the profit per bag:

$$\text{Number of bags sold} = \frac{\$3,000}{\$25}$$

$$\text{Number of bags sold} = 120$$

However, 120 is not one of the answer choices provided, which means we need to re-evaluate our calculations.

Let's go back to the statement that the grocer makes a 25% profit on the selling price. This means that the selling price includes both the cost price and the profit. If the selling price is \$100 and the profit is 25% of this selling price, then the profit per bag is indeed \$25, as we calculated.

Now, let's re-calculate the number of bags sold using the correct profit per bag:

$$\text{Number of bags sold} = \frac{\$3,000}{\$25}$$

$$\text{Number of bags sold} = 120$$

Since 120 is not among the answer choices, we need to check if we've made a mistake in interpreting the answer choices or the question itself. If the question and calculations are correct, then there might be an error in the answer choices provided. If this is the case, none of the answer choices (A) through (E) is correct.

Let's review the question one more time to ensure we haven't missed any details. The question states that the grocer makes a 25% profit on the selling price for each bag of flour and sells each bag for \$100. Our calculation of the profit per bag at \$25 is correct, and so is our calculation of the number of bags sold to make a \$3,000 profit.

Therefore, it seems there is a discrepancy with the answer choices provided, as none of them match our calculation of 120 bags. If we must choose from the given options, we would have to conclude that none of them is correct. However, since this is not a standard outcome for such problems, it's possible that there is a misunderstanding or an error in the question or answer choices.

**E.2 Example 2**

*Question:* Find  $25/12 \times 5$   
*Answer Choices:* (A) 2.5498 (B) 0.4167 (C) 3.3987 (D) 8.5497 (E) 5.6312  
*GT:* B

**IAO : Step 1:**

- *Subquestion:* What is the product of  $(\frac{25}{12})$  and 5?
- *Input:*  $(\frac{25}{12}), 5$
- *Action:* Multiply  $(\frac{25}{12})$  by 5.
- *Output:*  $\frac{25}{12} \times 5 = \frac{25 \times 5}{12} = \frac{125}{12}$

**Step 2:**

- *Subquestion:* What is the decimal equivalent of 125/12?
- *Input:* 125/12
- *Action:* Divide 125 by 12.
- *Output:* 10.4167

**Step 3:**

- *Subquestion:* Which answer choice matches the output from Step 2?

- *Input:* 10.4167, Answer Choices: (A) 2.5498 (B) 0.4167 (C) 3.3987 (D) 8.5497 (E) 5.6312
- *Action:* Compare the output with the given answer choices.
- *Output:* None of the answer choices match the output from Step 2.

It seems there might have been a mistake in the calculation or in the provided answer choices. Let's re-evaluate the calculation in Step 2 to ensure accuracy.

**Step 4:**

- *Subquestion:* What is the correct decimal equivalent of 125/12?
- *Input:* 125/12
- *Action:* Divide 125 by 12 using long division or a calculator.
- *Output:* 10.4167 (confirmed)

Upon re-evaluation, the calculation is confirmed to be correct, but the output does not match any of the provided answer choices. It's possible that there is an error in the question or the answer choices. However, based on the calculation we've done, none of the answer choices (A) through (E) is correct.

Therefore, among A through E, the answer is none of the provided choices.

**COT : To solve the expression**

$$\frac{25}{12} \times 5$$

, follow these steps:

**Step 1: Simplify the expression by performing the multiplication.**

$$\frac{25}{12} \times 5 = \frac{25 \times 5}{12}$$

**Step 2: Multiply the numerator.**

$$\frac{25 \times 5}{12} = \frac{125}{12}$$

**Step 3: Divide 125 by 12 to get the decimal form.**

$$125 \div 12 = 10.4167$$

**Step 4: Compare the result with the answer choices.**

The result, 10.4167, is not listed among the answer choices. It seems there might have been a mistake in the calculation. Let's try again.

**Step 1: Simplify the expression by performing the multiplication.**

$$\frac{25}{12} \times 5 = \frac{25 \times 5}{12}$$

**Step 2: Multiply the numerator.**

$$\frac{25 \times 5}{12} = \frac{125}{12}$$

**Step 3: Divide 125 by 12 to get the decimal form.**

$$125 \div 12 = 10 \text{ remainder } 5$$

To express the remainder as a decimal, we divide 5 by 12:

$$5 \div 12 = 0.4167$$

Adding this to the whole number part (10), we get:

$$10 + 0.4167 = 10.4167$$

This result still does not match any of the answer choices. However, upon re-evaluating the answer choices, it seems that none of them are correct if the original expression was indeed

$$\frac{25}{12} \times 5$$

. It's possible that there might be a typo in the question or the answer choices.

Given the calculations we've performed, none of the answer choices (A) through (E) match the correct result of 10.4167. Therefore, it appears there is an error in the provided information.

## F Errors from Date Understanding

### F.1 Example 1

*Question:* 2015 is coming in 36 hours. What is the date 24 hours later in MM/DD/YYYY?

*Answer Choices:* (A) 12/30/2014 (B) 01/01/2015 (C) 01/04/2015 (D) 12/30/2059 (E) 12/31/2014 (F) 01/09/2015

*GT:* "A"

### IAO : Step 1:

• *Subquestion:* What is the current date if 2015 is coming in 36 hours?

• *Input:* 2015 is coming in 36 hours.

• *Action:* Subtract 36 hours from the start of 2015 to find the current date.

• *Output:* The current date is December 30, 2014.

### Step 2:

• *Subquestion:* What is the date 24 hours after the current date?

• *Input:* The current date is December 30, 2014.

• *Action:* Add 24 hours to the current date to find the date 24 hours later.

• *Output:* The date 24 hours later is December 31, 2014.

Therefore, among the answer choices A through F, the answer is (E) December 31, 2014.

**COT** : If 2015 is coming in 36 hours, we can determine the current date by subtracting 36 hours from the start of 2015.

1. Since there are 24 hours in a day, 36 hours is equivalent to 1 day + 12 hours.

2. The start of 2015 is at 00 : 00 on January 1, 2015.

3. Subtracting 1 day from January 1, 2015, takes us back to December 31, 2014.

4. Since we need to subtract 12 more hours, and we are currently at 00 : 00 on December 31, 2014, we would still be on December 31, 2014, because we are not going back an entire day, just 12 hours into the previous day.

So, the current date and time, 36 hours before 2015, is 12 : 00 (noon) on December 31, 2014.

Now, we need to determine the date 24 hours later:

1. Adding 24 hours to 12 : 00 on December 31, 2014, will bring us to 12 : 00 on January 1, 2015.

Therefore, among A through F, the answer is (B) 01/01/2015.

## F.2 Example 2

*Question:* Tomorrow is 11/12/2019. What is the date 10 days ago in MM/DD/YYYY?

*Answer Choices:* (A) 10/23/2019 (B) 11/02/2019 (C) 10/27/2019 (D) 11/01/2048 (E) 11/01/2019

*GT:* "E"

### IAO : Step 1:

- *Subquestion:*  
What is the date 10 days before 11/12/2019?
- *Input:* The given date is 11/12/2019.
- *Action:* Subtract 10 days from the given date.
- *Output:* The date 10 days before 11/12/2019 is 11/02/2019.

Therefore, among A through E, the answer is (B) 11/02/2019.

**COT :** To find the date 10 days before 11/12/2019, we need to subtract 10 days from that date.

**Step 1: Start with the given date:** 11/12/2019 (which is in MM/DD/YYYY format, meaning November 12, 2019).

**Step 2: Subtract 10 days from the 12th of November.** Since  $12 - 10 = 2$ , if November had more than 2 days, we would end up in November itself. However, since we're subtracting more days than there are in the month up to the 12th, we need to go back to the previous month, which is October.

**Step 3: October has 31 days.** To find out what date it would be 10 days before November 12, we need to count back from October 31.

**Step 4: Since we need to go back 10 days from November 12, and there are only 1 day of November included in this count (November 12 itself), we subtract the remaining 9 days from the total number of days in October.**

**Step 5:**  $31$  (total days in October)  $- 9$  (days to subtract)  $= 22$ .

**Step 6: Therefore, 10 days before November 12, 2019, would be October 22, 2019.**

**Step 7: Convert this date into MM/DD/YYYY format, which gives us 10/22/2019.**

However, none of the answer choices match this date. It seems there might be a mistake in the answer choices provided or in the calculation. Let's double-check the calculation:

**Given date:** 11/12/2019

**Subtract 10 days:**

\* November has 12 days up to and including the 12th. \* We need to subtract 10 days. \*  $12 - 10 = 2$ , so we would be at November 2nd if we didn't need to go into the previous month. \* But since we're subtracting 10 days, we do need to go into the previous month. \* October has 31 days, so the day before November 1st is October 31st. \* We've already accounted for 1 day in November, so we subtract 1 from 10, leaving us with 9 days to subtract from October 31st. \*  $31 - 9 = 22$ , so we end up on October 22nd.

Upon re-evaluating the calculation, the correct date is indeed October 22, 2019, which is not listed in the answer choices. It seems there may be a typo in the question or the answer choices. However, based on the calculation, the correct answer is not present in the options provided.

**Therefore, among A through E, the answer is none of the above.**

## G Errors from StrategyQA

### G.1 Example 1

*Question:* Would John Muir not likely have a vitamin D deficiency?

*GT:* "yes"

**IAO :** To determine whether John Muir would likely have a vitamin D deficiency, we need to consider factors that influence vitamin D levels, such as exposure to sunlight, diet, skin pigmentation, age, geographic location, and lifestyle. Since we don't have specific information about John Muir's personal details, we'll have to rely on general knowledge about his life and activities.

John Muir (1838–1914) was a Scottish-American naturalist, author, environmental philosopher, and early advocate for the preservation of wilderness in the United States. He spent a significant amount of time outdoors, particularly in places like Yosemite and other parts of the American West.

#### 1. Step 1:

- *Subquestion:* Did John Muir spend a significant amount of time outdoors?
- *Input:* Historical knowledge about John Muir's life and activities.
- *Action:* Assess the likelihood of outdoor activity based on known biographical information.

1145	• <i>Output:</i> Yes, John Muir spent a significant amount of time outdoors.	• Lack of exposure to sunlight	1190
1146		• Inadequate dietary intake of vitamin D	1191
1147	<b>2. Step 2:</b>	• Certain medical conditions	1192
1148	• <i>Subquestion:</i> Does spending a significant amount of time outdoors increase exposure to sunlight?	• The body's inability to absorb or convert vitamin D to its active form	1193
1149			1194
1150			
1151	• <i>Input:</i> General knowledge about outdoor activities and sunlight exposure.	<b>Step 2: Consider John Muir's lifestyle.</b>	1195
1152		John Muir was a naturalist, explorer, and advocate for the preservation of wilderness. His lifestyle involved spending a significant amount of time outdoors, often in sunny environments like California's Sierra Nevada mountains.	1196
1153	• <i>Action:</i> Correlate outdoor activity with sunlight exposure.		1197
1154			1198
1155	• <i>Output:</i> Yes, spending a significant amount of time outdoors generally increases exposure to sunlight.		1199
1156			1200
1157			1201
1158	<b>3. Step 3:</b>	<b>Step 3: Assess the likelihood of vitamin D deficiency based on lifestyle.</b>	1202
1159	• <i>Subquestion:</i> Does increased exposure to sunlight affect vitamin D levels?	Given that John Muir spent a lot of time outdoors, he would have been exposed to sunlight regularly. Sunlight exposure is a primary source of vitamin D, as the body synthesizes vitamin D when the skin is exposed to UV rays from the sun.	1203
1160			1204
1161	• <i>Input:</i> Scientific knowledge about vitamin D synthesis.		1205
1162			1206
1163	• <i>Action:</i> Relate sunlight exposure to vitamin D synthesis.	<b>Step 4: Draw a conclusion based on the information.</b>	1207
1164		Since John Muir was frequently outdoors and exposed to sunlight, it is <i>less likely</i> that he would have suffered from vitamin D deficiency due to lack of sunlight exposure.	1208
1165	• <i>Output:</i> Yes, increased exposure to sunlight typically leads to higher vitamin D levels due to the skin's production of vitamin D in response to UVB rays.	<b>Therefore, the answer (Yes or No) is No.</b>	1209
1166		<i>Question:</i> Can French Defence initial move defend against four move checkmate?	1210
1167		<i>GT:</i> "no"	1211
1168		<b>IAO</b> : To determine if the French Defence can defend against the four-move checkmate (also known as the Scholar's Mate), we need to analyze the moves involved in both the French Defence and the Scholar's Mate.	1212
1169	<b>4. Step 4:</b>		1213
1170	• <i>Subquestion:</i> Based on John Muir's outdoor activities and the relationship between sunlight exposure and vitamin D synthesis, is it likely that John Muir had a vitamin D deficiency?		1214
1171			1215
1172	• <i>Input:</i> Outputs from Step 1, Step 2, and Step 3.		1216
1173			1217
1174	• <i>Action:</i> Infer the likelihood of vitamin D deficiency based on the correlation between outdoor activity, sunlight exposure, and vitamin D synthesis.		1218
1175			1219
1176	• <i>Output:</i> It is unlikely that John Muir had a vitamin D deficiency given his significant time spent outdoors and the resulting increased exposure to sunlight.		1220
1177			1221
1178			1222
1179			1223
1180			1224
1181			1225
1182			1226
1183			1227
1184			1228
1185	Therefore, the answer (Yes or No) is No.		1229
1186	<b>COT</b> : <b>Step 1: Identify the factors that contribute to vitamin D deficiency.</b>		1230
1187	Vitamin D deficiency can be caused by several factors, including:		1231
1188			1232
1189			1233

1235	– <i>Input:</i> The four-move checkmate (Scholar’s Mate) sequence.	2. Qh5 Nc6	1281
1236			
1237	– <i>Action:</i> Describe the typical sequence of moves that lead to the four-move checkmate.	3. Bc4 Nf6	1282
1238			
1239		4. Qxf7#	1283
1240	– <i>Output:</i> The four-move checkmate typically involves the moves 1.e4 e5 2.Qh5 Nc6 3.Bc4 Nf6 4.Qxf7#.	In this sequence, White aims to attack the f7 square, which is a weak point in Black’s position because it is only defended by the king at the beginning of the game.	1284
1241			1285
1242			1286
1243	• <b>Step 3:</b>	The French Defense, on the other hand, is a chess opening that begins with the moves:	1287
1244	– <i>Subquestion:</i> Can the opening move of the French Defence prevent the specific sequence of the four-move checkmate?	1. e4 e6	1288
1245			1289
1246	– <i>Input:</i> The opening move of the French Defence (1.e4 e6) and the sequence of the four-move checkmate (1.e4 e5 2.Qh3 Nc6 3.Bc4 Nf6 4.Qxf7#).	The idea behind the French Defense is to solidify Black’s position and create a strong pawn chain that can later be used to counterattack White’s center. The French Defense does not immediately contest the e5 square, which is a key component of setting up the Scholar’s Mate.	1290
1247			1291
1248	– <i>Action:</i> Compare the opening moves of the French Defence with the moves required for the four-move checkmate to determine if the French Defence can prevent the checkmate sequence.	Let’s see if the French Defense can prevent the Four-Move Checkmate:	1292
1249			1293
1250		1. e4 e6	1294
1251	– <i>Output:</i> The French Defence does not play 1...e5, which is a necessary move for the four-move checkmate. Instead, it plays 1...e6, which already deviates from the Scholar’s Mate sequence.	2. d4 d5	1295
1252			1296
1253		With the second move d5, Black challenges White’s center and opens lines for the bishop and queen, which can help in defending against early threats.	1297
1254		If White tries to set up the Scholar’s Mate anyway:	1298
1255			1299
1256		1. 3. Qh5	1300
1257			1301
1258		Black can now defend easily because the move ...e6 has already prevented the bishop from coming to c4 to attack the f7 square. Black has several reasonable moves here, such as ...Nf6, which attacks the invading queen and prepares to castle, or even ...g6, which would force the queen to move again.	1302
1259		Since the French Defense does not allow White to establish the same attacking pattern as in the Scholar’s Mate, it effectively prevents the quick checkmate.	1303
1260		<b>Therefore, the answer (Yes or No) is Yes.</b>	1304
1261	• <b>Step 4:</b>		1305
1262	– <i>Subquestion:</i> Does the deviation from the four-move checkmate sequence inherently prevent the checkmate from occurring?		1306
1263			1307
1264	– <i>Input:</i> The deviation in the opening move of the French Defence from the Scholar’s Mate sequence.		1308
1265			1309
1266	– <i>Action:</i> Analyze the impact of the deviation on the possibility of the four-move checkmate occurring.		1310
1267			1311
1268	– <i>Output:</i> Since the French Defence does not follow the move 1...e5, it inherently prevents the specific sequence of the four-move checkmate from occurring.		1312
1269			1313
1270			1314
1271			1315
1272			1316
1273			1317
1274			1318
1275			
1276	Therefore, the answer (Yes or No) is Yes.		
1277	<b>COT</b> : The Four-Move Checkmate, also known as the Scholar’s Mate, is a quick checkmate that White can deliver after the following moves:		
1278			
1279			
1280	1. e4 e5		