

A Details of Experimental Setup

A.1 Code, Prompts, Logs

All code is available at <https://anonymous.4open.science/r/cp-712E/README.md>.

All prompts are available at <https://anonymous.4open.science/r/cp-712E/main.py>.

Our experimental logs are available at <https://anonymous.4open.science/r/cp-712E/log/>.

A.2 Prompts For Answer Extraction

Table 7 summarizes the answer extraction prompt for each task used for the CP experiments.

A.3 Pre-training data examples

For instance, many web pages and books provide correct and incorrect answers to math reasoning^{4 5 6} and common sense reasoning^{7 8} questions.

B Additional Experimental Results

In this section, we provide a summary of additional example texts generated by Zero-shot-CP. gpt-35-turbo is used as the model if not specified. Table 6 illustrates example outputs of zero-shot prompting methods from CommonsenseQA. Figure 4–15 show a positive example and a negative example of Zero-shot-CP on each dataset. "GT" in the figures means "Ground Truth".

The 240 examples, along with our annotations, can be accessed at the following link: https://anonymous.4open.science/r/cp-712E/results/zero_shot_cp_gpt4_240_examples_labeled.txt.

Table 4 presents the comparison of Zero-shot-CP with Zero-shot and Zero-shot-CoT on all 12 datasets using gpt-35-turbo and GPT-4.

Table 8 presents the results of using various open-source LLMs: LLaMA3-8B, LLaMA3-70B, ChatGLM3-6B, and Qwen1.5-72B-Chat as base models.

Table 9 provides the categorization and counts of these 120 solved problems and 120 unsolved problems.

⁴ <https://prek-math-te.stanford.edu/operations/analyzing-thinking-underlying-wrong-answers>

⁵ <https://mathmistakes.org/category/elementary-school/>

⁶ <https://www.gutenberg.org/ebooks/38769>

⁷ https://www.proprofs.com/quiz-school/story.php?title=common-sense-quiz_1

⁸ <https://www.wikihow.com/Common-Sense-Quiz>

Method	Arithmetic					
	SingleEq	AddSub	MultiArith	GSM8K	AQUA	SVAMP
Zero-shot	90.6/81.7	92.4 /82.8	96.5/61.2	35.9/14.3	41.3/29.9	86.4/69.7
Zero-shot-CoT	91.7/91.1	89.6/86.6	97.7/94.8	90.9 / 75.1	70.1 / 55.9	90.4/81.9
Zero-shot-CP	91.7 / 91.7	91.6/ 90.6	97.8 / 95.2	88.8/73.2	62.2/40.2	91.5 / 83.2
	Common Sense		Other Reasoning Tasks		Symbolic Reasoning	
	Common SenseQA	Strategy QA	Date Understand	Shuffled Objects	Last Letter (4 words)	Coin Flip (4 times)
Zero-shot	82.9/71.3	64.8/65.0	73.2/40.4	40.7/33.9	5.0/4.2	36.6/49.6
Zero-shot-CoT	78.3/67.8	69.8/60.9	79.4 / 62.1	93.1 / 73.1	90.2 / 88.0	98.6 / 94.0
Zero-shot-CP	83.5 / 73.9	73.4 / 67.3	71.5/51.8	44.4/51.5	23.4/41.8	33.2/56.8

Table 4. Accuracy (in percentage) comparison of Zero-shot-CP with Zero-shot and Zero-shot-CoT on each dataset. The values on the left-hand side of each task represent the results obtained using GPT-4, while the values on the right-hand side represent the results obtained using gpt-35-turbo.

GPT-4	AQUA	GSM8K	AddSub	MultiArith
Let's give a correct and a wrong answer.	62.2	88.8	91.6	97.8
Let's first give a wrong answer, then give the correct answer.	69.3	86.1	90.9	95.0
Let's first give the correct answer, then give a wrong answer.	58.7	89.7	91.6	95.0
Let's give a correct and an incorrect answer.	66.5	88.7	91.6	97.7
Please give a correct and a wrong answer.	57.5	82.0	88.9	94.0
Let's give a correct answer.	71.7	75.9	89.4	97.0
Let's think step by step and give both a correct answer and a wrong answer.	71.3	89.5	91.4	97.2
Let's give a correct and a wrong answer. Let's also think step by step for the correct and the wrong answer.	52.8	88.9	89.4	96.7
Let's think step by step. (Zero-shot-CoT)	70.1	90.9	89.6	97.7

Table 5. Comparison prompting templates using accuracies (in percentage) on AQUA-RAT, GSM8K, AddSub and MultiArith in zero-shot setting. GPT-4 is used as the model. Bolded numbers indicate the best results within each block's column, while underlined numbers indicate the best results across the entire column.

Question	Q: A bald eagle is likely to be found on what kind of work? Answer Choices: (A) rural area (B) book (C) canada (D) painting (E) aviary
Zero-shot	A: Among A through E, the answer is (E) aviary . However, it should be noted that bald eagles can also be found in rural areas (A), in Canada (C), and depicted in books (B) and paintings (D). The question seems to be asking where a bald eagle is likely to be found physically, and an aviary is a place where birds are kept.
Zero-shot-CoT	A: Let's think step by step. A bald eagle is a bird, so it could be found in an aviary (E), which is a place where birds are kept. It could also be found in a rural area (A), as these birds often live in areas with less human activity. A bald eagle could be depicted in a book (B) or a painting (D), as it is a popular symbol in various cultures, especially in the United States. Lastly, bald eagles are also found in Canada (C), as their habitat ranges across most of North America. Therefore, all of the options could be correct depending on the context of the question. Therefore, among A through E, the answer is (A) rural area , (B) book , (C) Canada , (D) painting , (E) aviary .
Zero-shot-CP	A: Let's give a correct and a wrong answer. Correct Answer: (D) painting Wrong Answer: (A) rural area Therefore, among A through E, the correct answer is (D) painting
Ground Truth	D

Table 6. An example question from CommonsenseQA and responses from Zero-shot, Zero-shot-CoT, Zero-shot-CP and Zero-shot-CoT-CP. We use GPT-4 as the base model.

No Task	CP answer extraction prompts
1 SingleEq	Therefore, the correct answer (arabic numerals) is
2 AddSub	Therefore, the correct answer (arabic numerals) is
3 MultiArith	Therefore, the correct answer (arabic numerals) is
4 GSM8K	Therefore, the correct answer (arabic numerals) is
5 AQUA-RAT	Therefore, among A through E, the correct answer is
6 SVAMP	Therefore, the correct answer (arabic numerals) is
7 CommonsenseQA	Therefore, among A through E, the correct answer is
8 StrategyQA	Therefore, the correct answer (Yes or No) is
9 Date Understanding	Therefore, among A through F, the correct answer is
10 Shuffled Objects	Therefore, among A through C, the correct answer is
11 Last Letters	Therefore, the correct answer is
12 Coin Flip	Therefore, the correct answer (Yes or No) is

Table 7. Answer extraction prompts used in our CP experiments.

	MultiArith	GSM8K	StrategyQA	AQUA	SVAMP
<i>LLaMA3-8B</i>					
Zero-shot	31.0	38.1	–	–	52.8
Zero-shot-CP	57.3	54.9	–	–	61.4
<i>LLaMA3-70B</i>					
Zero-shot	86.5	63.7	54.5	38.2	–
Zero-shot-CP	97.0	66.1	57.5	62.2	–
<i>ChatGLM3-6B</i>					
Zero-shot	5.3	4.3	–	–	–
Zero-shot-CP	67.0	40.0	–	–	–
<i>Qwen1.5-72B-Chat</i>					
Zero-shot	54.7	19.3	71.2	31.1	65.2
Zero-shot-CP	75.5	52.1	73.5	45.3	77.4

Table 8. Accuracy (in percentage) comparison of Zero-shot-CP with Zero-shot using open LLMs.

Category	# Examples
The given "correct" answer is the GT, and the given "wrong" answer is indeed incorrect.	112
The given "correct" answer is the GT, and the given "wrong" answer is also the GT.	4
The given "correct" answer is the GT, no "wrong" answer is given.	4
The given "correct" answer is incorrect, and the given "wrong" answer is the GT.	23
The given "correct" answer is incorrect, and the given "wrong" answer is also incorrect.	91
The given "correct" answer is incorrect, no "wrong" answer is given.	6

Table 9. Categorization results of Zero-shot-CP output (with GPT-4) on 240 problems. We manually annotated 10 solved problems and 10 unsolved problems for each of the 12 datasets. GT means Ground Truth. See Appendix B for the link of the examples.

Q: Mary found 2 seashells and Keith found 5 seashells on the beach . When they cleaned them , they discovered that 9 were cracked . How many seashells did they find together ?

A: Let's give a correct and a wrong answer. Correct answer: To find the total number of seashells they found together, we need to add the number of seashells Mary found (2) to the number of seashells Keith found (5).

$$2 + 5 = 7$$

Therefore, Mary and Keith found a total of 7 seashells together.

Wrong answer: To find the total number of seashells they found together, we can subtract the number of cracked seashells (9) from the sum of the seashells Mary and Keith found.

$$(2 + 5) - 9 = -2$$

This answer doesn't make sense because we can't have a negative number of seashells.

Therefore, the correct answer (arabic numerals) is 7. ✓

Q: Sam 's dog had puppies and 8 had spots . He gave 2 to his friends . He now has 6 puppies . How many puppies did he have to start with ?

A: Let's give a correct and a wrong answer. Correct answer:

To find out how many puppies Sam had to start with, we can work backwards from the information given. We know that Sam currently has 6 puppies after giving 2 to his friends. Before giving away the 2 puppies, he had 8 puppies with spots. Therefore, the total number of puppies he had to start with is $8 + 2 = 10$.

Wrong answer:

To find out how many puppies Sam had to start with, we can subtract the number of puppies he currently has from the number of puppies he gave away. Since he currently has 6 puppies and he gave away 2, we can subtract 6 from 2, which gives us -4. Therefore, Sam had -4 puppies to start with.

Therefore, the correct answer (arabic numerals) is The correct answer is 10. ✗ (GT: 8)

Fig. 4. Example outputs by Zero-shot-CP for AddSub.

Q: Machine A puts out a yo-yo every 6 minutes. Machine B puts out a yo-yo every 9 minutes. After how many minutes will they have produced 10 yo-yos? Answer Choices: (A) 24 minutes (B) 32 minutes (C) 36 minutes (D) 64 minutes (E) 72 minutes

A: Let's give a correct and a wrong answer. Correct answer: (C) 36 minutes

Explanation:

Machine A produces a yo-yo every 6 minutes, so in 36 minutes, it will produce $36/6 = 6$ yo-yos.

Machine B produces a yo-yo every 9 minutes, so in 36 minutes, it will produce $36/9 = 4$ yo-yos.

Therefore, in a total of 36 minutes, both machines will have produced $6 + 4 = 10$ yo-yos.

Wrong answer: (B) 32 minutes

Explanation:

Machine A produces a yo-yo every 6 minutes, so in 32 minutes, it will produce $32/6 = 5.33$ yo-yos (approximately).

Machine B produces a yo-yo every 9 minutes, so in 32 minutes, it will produce $32/9 = 3.56$ yo-yos (approximately).

Therefore, in a total of 32 minutes, both machines will have produced $5.33 + 3.56 = 8.89$ yo-yos (approximately), which is less than 10 yo-yos.

Therefore, among A through E, the correct answer is (C) 36 minutes. ☒

Q: On a map, the length of the road from Town F to Town G is measured to be 20 inches. On this map, 1/4 inch represents an actual distance of 10 miles. What is the actual distance, in miles, from Town F to Town G along this road? Answer Choices: (A) 800 (B) 720 (C) 960 (D) 1140 (E) 1160

A: Let's give a correct and a wrong answer. Correct Answer: (C) 960 miles

Explanation:

According to the map, 1/4 inch represents 10 miles. Therefore, 1 inch represents 4 times that, which is 40 miles.

Since the length of the road from Town F to Town G on the map is 20 inches, the actual distance in miles would be 20 inches multiplied by 40 miles per inch, which equals 800 miles.

Wrong Answer: (A) 800 miles

Explanation:

This answer is incorrect because it does not take into account the scale of the map. The scale of the map indicates that 1/4 inch represents 10 miles, so the actual distance from Town F to Town G is greater than 800 miles.

Therefore, among A through E, the correct answer is (C) 960 miles. ☐ (GT: A)

Fig. 5. Example outputs by Zero-shot-CP for AQUA-ART.

Q: Yesterday was April 30, 2021. What is the date one year ago from today in MM/DD/YYYY?
 Answer Choices: (A) 04/01/2020 (B) 05/08/2020 (C) 05/01/2020 (D) 05/15/2020 (E) 05/01/1971
 A: Let's give a correct and a wrong answer. Correct answer: (C) 05/01/2020

Wrong answer: (B) 05/08/2020

Therefore, among A through F, the correct answer is (C) 05/01/2020 ☒

Q: Yesterday was April 30, 2021. What is the date 10 days ago in MM/DD/YYYY? Answer Choices:
 (A) 01/21/2021 (B) 04/13/2021 (C) 04/21/1990 (D) 04/21/2021 (E) 04/18/2021 (F) 04/22/2021
 A: Let's give a correct and a wrong answer. Correct answer: (B) 04/13/2021

Wrong answer: (D) 04/21/2021

Therefore, among A through F, the correct answer is (B) 04/13/2021. ☒ (GT: D)

Fig. 6. Example outputs by Zero-shot-CP for Date Understanding.

Q: A coin is heads up. Emanuel does not flip the coin. Cheyenne does not flip the coin. Perez flips the coin. Yvette does not flip the coin. Is the coin still heads up? Note that "flip" here means "reverse".

A: Let's give a correct and a wrong answer. Correct answer: No, the coin is not still heads up. Since Perez flips the coin, it means that the coin is reversed or flipped, resulting in a different side facing up.

Wrong answer: Yes, the coin is still heads up. Since only Perez flips the coin and the others do not, the initial state of the coin remains unchanged.

Therefore, the correct answer (Yes or No) is No, the coin is not still heads up. ☒

Q: A coin is heads up. Mya flips the coin. Fernando flips the coin. Bubba does not flip the coin. Tommy does not flip the coin. Is the coin still heads up? Note that "flip" here means "reverse".

A: Let's give a correct and a wrong answer. Correct answer: No, the coin is not still heads up. When Mya flips the coin, it will be reversed, so it will be tails up. The subsequent actions of Fernando, Bubba, and Tommy are irrelevant because the coin has already been flipped and its state has changed.

Wrong answer: Yes, the coin is still heads up. Since Bubba and Tommy did not flip the coin, its original state remains unchanged.

Therefore, the correct answer (Yes or No) is No, the coin is not still heads up. ☒ (GT: Yes)

Fig. 7. Example outputs by Zero-shot-CP for Coin Flip.

Q: The man closed his eyes as the music played, what effect did the music have? Answer Choices: (A) coma (B) enjoyable (C) soothing (D) universal (E) good or bad

A: Let's give a correct and a wrong answer. Correct answer: (C) soothing

The music had a soothing effect on the man, causing him to close his eyes and relax.

Wrong answer: (A) coma

The music did not cause the man to enter a coma.

Therefore, among A through E, the correct answer is (C) soothing. ☒

Q: As he looked out the window, he knew the landing was happening soon, and it made him nervous, but where would he be soon? Answer Choices: (A) apartment building (B) disembark (C) stairwell (D) deplane (E) airport

A: Let's give a correct and a wrong answer. Correct answer: (D) deplane

Wrong answer: (C) stairwell

Therefore, among A through E, the correct answer is (D) deplane. ☐ (GT: E)

Fig. 8. Example outputs by Zero-shot-CP for CommonsenseQA.

Q: Jeff's work is 3 miles away. He walks there and back each day he works. How many miles does he walk if he has to work 5 times a week?

A: Let's give a correct and a wrong answer. Correct answer: Jeff walks 3 miles to work and 3 miles back, so he walks a total of 6 miles each day he works. If he works 5 times a week, he would walk a total of 6 miles/day * 5 days/week = 30 miles in a week.

Wrong answer: Jeff walks 3 miles to work and 3 miles back, so he walks a total of 3 miles each day he works. If he works 5 times a week, he would walk a total of 3 miles/day * 5 days/week = 15 miles in a week.

Therefore, the correct answer (arabic numerals) is 30 miles. ☒

Q: If one person can paint half a house in 5 days, how many hours would it take for 5 people to paint a whole house?

A: Let's give a correct and a wrong answer. Correct answer: If one person can paint half a house in 5 days, it means that one person can paint one whole house in 10 days. Since there are 5 people working together, we can divide the time by 5. Therefore, it would take 5 people 2 days to paint a whole house.

Wrong answer: If one person can paint half a house in 5 days, it means that one person can paint one whole house in 10 days. Since there are 5 people working together, we can divide the time by 5. Therefore, it would take 5 people 2 hours to paint a whole house.

Therefore, the correct answer (arabic numerals) is 2 days. ☐ (GT: 48)

Fig. 9. Example outputs by Zero-shot-CP for GSM8K.

Q: Take the last letters of each words in "Marlen Sonja Anita Dale" and concatenate them.

A: Let's give a correct and a wrong answer. Correct answer: The last letters of each word in "Marlen Sonja Anita Dale" are "n a a e." Therefore, when concatenated, they form the word "naae."

Wrong answer: The last letters of each word in "Marlen Sonja Anita Dale" are "n j a e." Therefore, when concatenated, they form the word "njae."

Therefore, the correct answer is "naae." ✓

Q: Take the last letters of each words in "Rich Parker Edward Milagros" and concatenate them.

A: Let's give a correct and a wrong answer. Correct answer: The last letters of each word in "Rich Parker Edward Milagros" are "h r d s s". Therefore, when concatenated, they form the word "hrdss".

Wrong answer: The last letters of each word in "Rich Parker Edward Milagros" are "r r d s s". Therefore, when concatenated, they form the word "rrdss".

Therefore, the correct answer is "hrdss". ✗ (GT: hrdss)

Fig. 10. Example outputs by Zero-shot-CP for Last Letter Concatenation.

Q: Roger is at the library helping put away books. There are 14 book to put away total but a librarian takes 2 of them and leaves Roger with the rest. If he can fit 3 books on a shelf, how many shelves will he need?

A: Let's give a correct and a wrong answer. Correct answer:

If there are 14 books in total and the librarian takes 2 of them, then Roger is left with $14 - 2 = 12$ books to put away. Since he can fit 3 books on a shelf, he will need $12 / 3 = 4$ shelves.

Wrong answer:

If there are 14 books in total and the librarian takes 2 of them, then Roger is left with $14 - 2 = 12$ books to put away. Since he can fit 3 books on a shelf, he will need $12 / 3 = 3$ shelves.

Therefore, the correct answer (arabic numerals) is 4 shelves. ✓

Q: John earned 8 dollars for each lawn he mowed. If he had 15 lawns to mow, but forgot to mow 7 of them, how much money did he actually earn?

A: Let's give a correct and a wrong answer. Correct answer: John earned 8 dollars for each lawn he mowed. Since he mowed 15 lawns, he would have earned $8 * 15 = 120$ dollars.

Wrong answer: John earned 8 dollars for each lawn he mowed. Since he forgot to mow 7 lawns, he actually earned $8 * (15 - 7) = 8 * 8 = 64$ dollars.

Therefore, the correct answer (arabic numerals) is 120 dollars. ✗ (GT: 64)

Fig. 11. Example outputs by Zero-shot-CP for MultiArith.

Q: Alice, Bob, and Claire are dancers at a square dance. At the start of a song, they each have a partner: Alice is dancing with Karl, Bob is dancing with Rodrigo, and Claire is dancing with Helga.

Throughout the song, the dancers often trade partners. First, Bob and Alice switch partners. Then, Alice and Claire switch partners. Finally, Alice and Bob switch partners. At the end of the dance, Bob is dancing with

Which choice is true ? Answer Choices: (A) Karl. (B) Rodrigo. (C) Helga.

A: Let's give a correct and a wrong answer. Correct answer: (C) Helga.

Explanation:

At the start of the dance, Bob is dancing with Rodrigo. After the first partner switch, Bob and Alice switch partners, so Bob is now dancing with Karl. Then, Alice and Claire switch partners, so Bob is still dancing with Karl. Finally, Alice and Bob switch partners, so Bob is now dancing with Helga. Therefore, at the end of the dance, Bob is dancing with Helga.

Therefore, among A through C, the correct answer is (C) Helga. ☒

Q: Alice, Bob, and Claire are dancers at a square dance. At the start of a song, they each have a partner: Alice is dancing with Sam, Bob is dancing with Karl, and Claire is dancing with Helga.

Throughout the song, the dancers often trade partners. First, Alice and Claire switch partners. Then, Alice and Bob switch partners. Finally, Claire and Bob switch partners. At the end of the dance, Claire is dancing with

Which choice is true ? Answer Choices: (A) Sam. (B) Karl. (C) Helga.

A: Let's give a correct and a wrong answer. Correct answer: (B) Karl.

Explanation:

At the start of the dance, Alice is dancing with Sam, Bob is dancing with Karl, and Claire is dancing with Helga.

First, Alice and Claire switch partners. After this switch, Alice is now dancing with Helga, and Claire is dancing with Sam. Bob's partner remains the same (Karl).

Next, Alice and Bob switch partners. After this switch, Alice is now dancing with Karl, and Bob is dancing with Helga. Claire's partner remains the same (Sam).

Finally, Claire and Bob switch partners. After this switch, Claire is now dancing with Karl, and Bob is dancing with Sam. Alice's partner remains the same (Helga).

Therefore, at the end of the dance, Claire is dancing with Karl. So the correct answer is (B) Karl.

Therefore, among A through C, the correct answer is (B) Karl. ☐ (GT: C)

Fig. 12. Example outputs by Zero-shot-CP for Tracking Shuffled Object.

Q: A pet supply store has 600 bags of dog food and 327 bags of cat food. How many more bags of dog food are there than cat food?

A: Let's give a correct and a wrong answer. Correct answer: To find out how many more bags of dog food there are than cat food, we subtract the number of bags of cat food from the number of bags of dog food.

Number of bags of dog food - Number of bags of cat food = $600 - 327 = 273$

Therefore, there are 273 more bags of dog food than cat food.

Wrong answer: To find out how many more bags of dog food there are than cat food, we add the number of bags of dog food and cat food together.

Number of bags of dog food + Number of bags of cat food = $600 + 327 = 927$

Therefore, there are 927 more bags of dog food than cat food.

Therefore, the correct answer (arabic numerals) is 273. ✓

Q: Harry Hound had a terrible earache yesterday. When I peered into his ears yesterday, I found 36 frisky fleas having a party in his right ear and 85 baby fleas sleeping peacefully in his left ear. I cleaned out Harry Hound's ears. How many fleas perished?

A: Let's give a correct and a wrong answer. Correct answer: If you found 36 fleas in Harry Hound's right ear and 85 fleas in his left ear, and you cleaned out his ears, it can be assumed that all the fleas were removed and none perished.

Wrong answer: If we assume that all the fleas perished, then the total number of fleas that perished would be $36 + 85 = 121$ fleas.

Therefore, the correct answer (arabic numerals) is 0 fleas perished. ✗ (GT: 121)

Fig. 13. Example outputs by Zero-shot-CP for SingleEq.

Q: Is it unusual to play Happy hardcore music at a funeral?

A: Let's give a correct and a wrong answer. Correct answer: Yes, it is highly unusual to play Happy hardcore music at a funeral. Funerals are typically solemn occasions where mourners gather to pay their respects and honor the life of the deceased. The music played at funerals is usually chosen to reflect the somber and reflective atmosphere of the event. Genres such as classical music, hymns, or soft instrumental pieces are more commonly associated with funerals.

Wrong answer: No, it is not unusual to play Happy hardcore music at a funeral. Funerals are evolving events, and people have different ways of expressing their grief and celebrating the life of the deceased. Some individuals may have had a strong connection to Happy hardcore music, and playing it at the funeral could be a way to honor their memory and bring a sense of joy and energy to the ceremony. Ultimately, the choice of music at a funeral is a personal decision, and if it reflects the wishes and personality of the deceased, then it can be considered appropriate.

Therefore, the correct answer (Yes or No) is Yes, it is highly unusual to play Happy hardcore music at a funeral. ✓

Q: Would a pear sink in water?

A: Let's give a correct and a wrong answer. Correct answer: Yes, a pear would sink in water. Pears have a density greater than that of water, which means they are denser and will sink when placed in water.

Wrong answer: No, a pear would not sink in water. Pears are known for their buoyancy and ability to float in water due to their unique composition.

Therefore, the correct answer (Yes or No) is Yes, a pear would sink in water. ✗ (GT: No)

Fig. 14. Example outputs by Zero-shot-CP for StrategyQA.

Q: Mary is baking a cake. The recipe calls for 7 cups of sugar and 10 cups of flour. She already put in 4 cups of sugar. How many more cups of sugar does she need to add?

A: Let's give a correct and a wrong answer. Correct answer: Mary needs to add 3 more cups of sugar.

Wrong answer: Mary needs to add 6 more cups of sugar.

Therefore, the correct answer (arabic numerals) is 3. ✓

Q: The grasshopper and the frog had a jumping contest. The grasshopper jumped 13 inches. The grasshopper jumped 2 inches farther than the grasshopper. How far did the frog jump?

A: Let's give a correct and a wrong answer. Correct answer: The grasshopper jumped 13 inches, and the frog jumped 2 inches farther than the grasshopper. Therefore, the frog jumped $13 + 2 = 15$ inches.

Wrong answer: The grasshopper jumped 13 inches, and the frog jumped 2 inches farther than the grasshopper. Therefore, the frog jumped $13 + 2 = 15$ inches.

Therefore, the correct answer (arabic numerals) is 15 inches. ✗ (GT: 11)

Fig. 15. Example outputs by Zero-shot-CP for SVAMP.