An Experiment Towards Understanding Reasoning Capability of LMs through CoT Prompting

Suvendu Kar & Vishnu Dutt Jawalakar & Ashhar Zaman
Indian Institute of Science
Bengaluru, KA, India
{suvendukar,vishnudutt,ashharzaman}@iisc.ac.in

March 19, 2024

Abstract

We investigate how, in some situations, producing a chain of thought—a sequence of intermediate reasoning steps—significantly enhances the capacity of large language models to carry out complicated reasoning. Specifically, we demonstrate how these reasoning skills spontaneously arise in sufficiently large language models using a straightforward technique known as chain-of-thought prompting, in which a few examples of chain-of-thought demonstrations are given to aid in the prompting process. Research using the PaLM (text-bison-001) , and LLAMA(Llama-2-7B-Chat-GPTQ)language model demonstrates that a variety of arithmetic (mathematical + logical reasoning) reasoning tasks are performed better when chain-of-thought prompting is used. There can be considerable empirical gains. For example, on the GSM8K test of math word problems, prompting a PaLM (text-bison-001) with just eight chain-of-thought exemplars achieves state-of-the-art accuracy.

1 Introduction

Language models are a revolution in the field of NLP[5,6,7]. Enhancing language models' size has been demonstrated to yield several advantage$s^{2,6}$, including enhanced effectiveness and reduced sample size. But increasing the size of the model by itself hasn't shown to be enough to achieve great performance on difficult tasks like mathematical reasoning^3.

This work investigates how a straightforward technique driven by the notion that huge language models offer the intriguing possibility of in-context few-shot learning via prompting can unlock the reasoning power of these models. That is, one can "prompt" the model with a few input-output exemplars that illustrate the task, rather than fine-tuning a distinct language model checkpoint for every new task. It's amazing how well this has worked for a variety of easy question-answering activities.

A substantial collection of high-quality rationales must be created, which is more expensive for rationale-augmented training and fine-tuning techniques than for the straightforward input-output pairings employed in traditional machine learning. When it comes to activities requiring reasoning skills, the standard fewshot prompting strategy performs badly and frequently does not get much better as language model scale increases. In this study, we avoid the drawbacks of these two concepts while combining their merits. In particular, we investigate language models' few-shot prompting capabilities for reasoning tasks, given a triple-part prompt (input, chain of thinking, and outcome). A chain of thought is a series of intermediate natural language reasoning steps that lead to the final output, and we refer to this approach as chain-of-thought prompting.

Our empirical assessments of arithmetic reasoning benchmarks demonstrate that, occasionally to a remarkable extent, chain-of-thought prompting performs better than ordinary prompting. Because it doesn't require a big training dataset and can handle multiple jobs with no loss of generality, a prompting only technique is significant. This paper highlights how big language models may automatically learn the patterns underlying inputs and outputs via a large training dataset, based on a small number of samples with real language data.

2 Related Work

Research work was conducted on using intermediate steps to solve reasoning problems.A prior work^8 pioneer the idea of using natural language rationales to solve math word problems through a series of intermediate steps. Their work is a remarkable contrast to the literature using formal languages to reaso$n^{9,10}$. Cobbe et al. (2021)[11] extend Ling et al. (2017)[8] by creating a larger dataset and using it to finetune a pretrained language model rather than training a model from scratch.

Naturally, our work also relates closely to the large body of recent work on prompting. Since the popularization of few-shot prompting as given by Brown et al. (2020)[7], several general approaches have improved the prompting ability of models, such as automatically learning prompts (Lester et al.,2021)[12] or giving models instructions describing a task (Wei et al., 2022a^{13}; Sanh et al., 2022[14]). These approaches improve or augment the input part of the prompt (e.g., instructions that are prepended to inputs).

In our empirical experimental work, we tried different promoting techniques on few Math Work Problem Data set , and analyzed their performance against standard prompting[15], which is surely giving a fair idea about how promting techniques can differ the output of an language model.Also we tried empirically to find a "better" prompting technique specifically to boost up the performance for arithmetic reasoning.

3 Chain-of-Thought Prompting

Consider one's own thought process when solving a complicated reasoning task such as a multi-step math word problem. It is typical to decompose the problem into intermediate steps and solve each before giving the final answer: "After Jane gives 2 flowers to her mom she has 10 . . . then after she gives 3 to her dad she will have 7 . . . so the answer is 7." The goal of our work is to endow language models with the ability to generate a similar chain of thought—a coherent series of intermediate reasoning steps that lead to the final answer for a problem. We will show that sufficiently largelanguage models can generate chains of thought if demonstrations of chain-of-thought reasoning are provided in the exemplars for few-shot prompting. Through our experiments we observed that in few cases model producing a chain of thought to solve a math word problem correctly , but the same problem it answered wrongly when it was not producing stdep by step solution. The chain of thought in this case resembles a solution and, we opt to call it a chain of thought to better capture the idea that it mimics a step-by-step thought process for arriving at the answer (and also, solutions/explanations typically come after the final answer (Narang et al., 2020[16]; inter alia)). Chain-of-thought prompting has several attractive properties as an approach for facilitating reasoning in language models.

1. First, chain of thought, in principle, allows models to decompose multi-step problems into intermediate steps, which means that additional computation can be allocated to problems that require more reasoning steps.

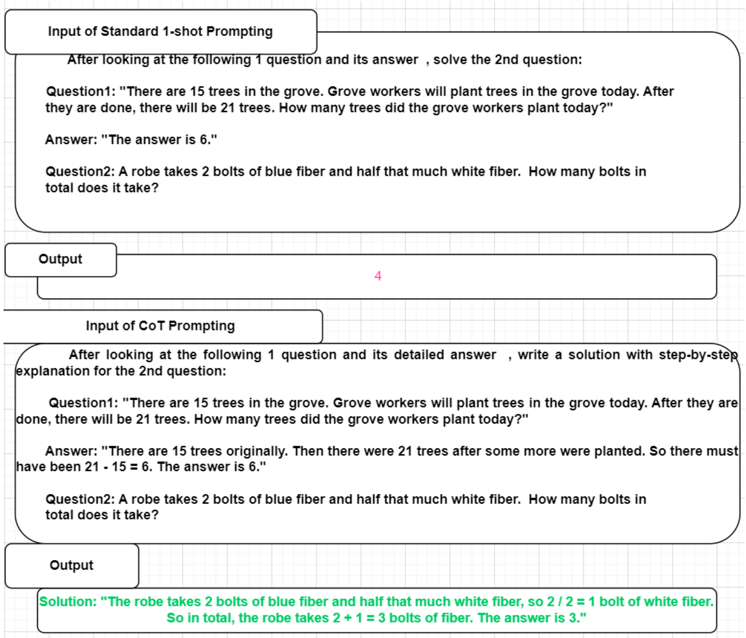
2. Second, a chain of thought provides an interpretable window into the behavior of the model, suggesting how it might have arrived at a particular answer and providing opportunities to debug where the reasoning path went wrong (This gives us the opportunity to analyze the reasoning capability of model,although fully characterizing a model's computations that support an answer remains an open question).

3. Third, chain-of-thought reasoning can be used for tasks such as math word problems, commonsense reasoning, and symbolic manipulation, and is potentially applicable (at least in principle) to any task that humans can solve via language.

4. Finally, chain-of-thought reasoning can be readily elicited in sufficiently large off-the-shelf language models simply by including examples of chain of thought sequences into the exemplars of few-shot prompting[15].

In empirical experiments, we will observe the utility of chain-of-thought prompting for arithmetic reasoning.

4 Methodology



Input of Standard 1-shot Prompting

After looking at the following 1 question and its answer , solve the 2nd question:

Question1: "There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?"

Answer: "The answer is 6."

Question2: A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

Output

4

Input of CoT Prompting

After looking at the following 1 question and its detailed answer , write a solution with step-by-step explanation for the 2nd question:

Question1: "There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?"

Answer: "There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6."

Question2: A robe takes 2 bolts of blue fiber and half that much white fiber. How many bolts in total does it take?

Output

Solution: "The robe takes 2 bolts of blue fiber and half that much white fiber, so 2 / 2 = 1 bolt of white fiber. So in total, the robe takes 2 + 1 = 3 bolts of fiber. The answer is 3."

Figure1:A 1 shot CoT, Standard Prompting Input and Their Model's Output Comparison

We begin by considering math word problems of the form in Figure 1, which measure the arithmetic reasoning ability of language models. Though simple for humans, arithmetic reasoning is a task where language models often struggle . Strikingly, chainof-thought prompting when used with PaLM(TEXT-BISON-001) language model performs comparably with task-specific finetuned models on several tasks, even achieving remarkable accuracy on the challenging GSM8K benchmark (Cobbe et al., 2021[11]).

4.1 Experimental Setup

We explore chain-of-thought prompting for PaLM(TEXT-BISON-001, with input-token-limit=8196,output-token-limit=1024,supported-generation-methods=['generateText', 'countTextTokens', 'createTunedTextModel'],temperature=0.5,top-p=0.95,top-k=40)).[Temperature controls the degree of randomness in token selection.Token limit determines the maximum amount of text output. Tokens are selected from most probable to least until the sum of their probabilities equals the top-p value. A top-k of 1 means the selected token is the most probable among all tokens.] model ,and LLAMA model(Llama-2-7b-Chat-GPTQ,in-features=4096, out-features=32000) on above mentioned benchmarks.With used a laptop with 8GB GEFORCE RTX 3070 Ti GPU, 512 GB SSD to run the tests for PaLM model.

Benchmark

We consider the following 3 math word problem benchmarks: (1) the GSM8K benchmark of math word problems (Cobbe et al., 2021[11]), (2) the SVAMP dataset of math word problems with varying structures (Patel et al., 2021[17]),and (3) the AQuA dataset of algebraic word problems.
For PaLM we used all 3 above, for LLAMA we tested with all except SVAMP.

Standard prompting.

For the baseline, we consider (a slightly modified as shown in figure 1)standard few-shot prompting, popularized by Brown et al. (2020[7]), in which a language model is given in-context exemplars of input–output pairs before outputting a prediction for a test-time example. Exemplars are formatted as questions(with question number) and answers. The model gives the answer directly, as shown in Figure 1 .

Chain of Thought Promting

Our proposed approach is to augment each exemplar in few-shot prompting with a chain of thought for an associated answer, as illustrated in Figure 1 .We manually composed a set of eight few-shot exemplars with chains of thought for prompting—Figure 1 shows one chain of thought exemplar(For GSM8K nad SVAMP data set), and the full set of exemplars is given in Appendix Table 1. (These particular exemplars did not undergo prompt engineering;)To investigate whether chain-of-thought prompting in this form can successfully elicit successful reasoning across a range of math word problems, we used this single set of eight chain of thought exemplars for all 3 benchmarks except AQuA, which is multiple choice instead of free response. For AQuA, we used four exemplars and solutions from the training set, as given in Appendix Table 3.Accross different data sets to have (k-1) shot evaluation we just delete kth example from the input prompting exemplars.From input in standard prompting we used Table 2 in Appendix Section for all data sets except AQuA, and for input regarding standard prompting on AQuA data set we used Table 4 in Appendix Section.For experiment over GSM8K and SVAMP data set we considered first 200 data points in respective test data sets, for AQuA we used full test data set.

5 Results

5.1 For PaLM Model

We measured the accuracy for CoT prompts by manually checking the answers with reasoning are correct or not.At the time of manual checking we found that (1)Sometime model gave advanced reasoning than the actual given answer in the test data(Table 7 under Appendix) (2)Model gave correct reasoning but no option has been chosen(Table 8 of Appendix) (3)Model gave wrong answer and told that no option was matching(Table 9 of Appendix) (4)Model chose correct answer in AQuA questions but wrong reasoning(Table 10 in Appendix) (5)Model derived correct answer with correct reasoning but gave final answer wrong in GSM8K data set(Table 11 of Appendix) (6)Model understood the question incorrectly (Table 12 of Appendix) (7)Model's reasoning diverted to wrong answer from initial correct approach(Table 13 of Appendix)

GSM8K and AQuA Data Set

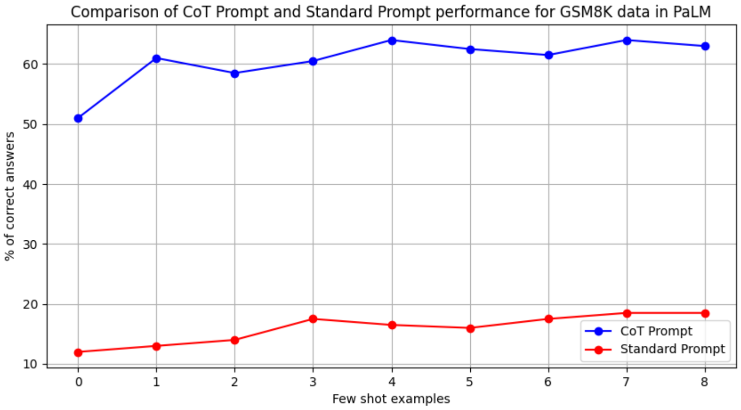


Figure : Full CoT prompt VS Standard Prompt for GSM8K

From figure 2 and Figure 3 in subsection A.1 under Appendix ,it is clear that CoT plays a big role in achieving tremendous success over standard prompt as shown in Figure 1.For AQuA data set full CoT method achieved on an average 15.88% more success that standard prompting, while this is 44.44% on GSM8K data set(We considered first 200 data from test set to testing).Just adding the extra sentence ”write a atep-by-step explanation ” along with question number over the set up as in Jason et. *al.*[15], we achieved 64% accuracy(with 7 shot full CoT) over GSM8K data set in PaLM text-bison-001 which is better than their finding on PaLM 540B model as well as our finding is better than prior state of the art best test report (from Cobbe et. al[11], which is $< 60\%$).More detailed outputs are listed in Table 1,2; plotted graph is at Figure 2 under A.1 of Appendix section

SVAMP Data

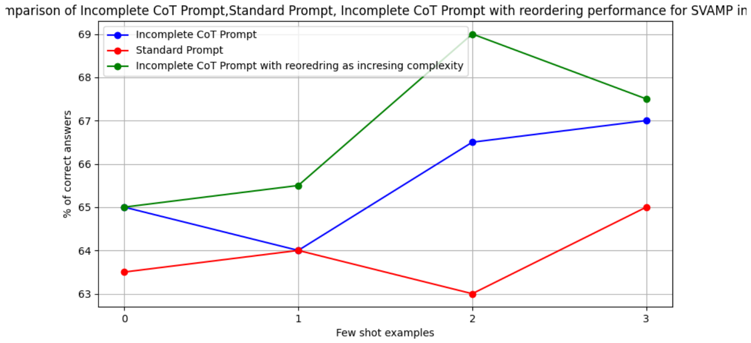


Figure A: Incomplete CoT VS Incomplete CoT+Reordering VS Standard Prompt for SVAMP

In our experiment we saw that over SVAMP data(We considered first 200 data from test set to testing) , few-shot standard prompt and few-shot full CoT prompt was performing more or less similar.To further explore model's behavior in terms of performance we firstly, performed 1,2,and 3 shot CoT prompting with incomplete reasoning in input exemplars (as shown in Table 5 in Appendix section), as well as secondly, performed CoT prompting with incomplete reasoning + reordered the input exemplars in terms their ascending complexity(in the sense that if more steps are there to solve a question then it is more complex than other with a solution contains relatively less step).We found that performance of incomplete resoning+reordering>performance with only reordering the exemplars, and in few cases they both were performing better than full CoT method.In each of those we achieved $> 63\%$ accuracy which is better than prior state of the art best finding(which was $< 50\%$) as mentioned in Jie et.*al.*[18]. Above Figure A plots a comparison. More detailed outputs are listed in Table 3,4; plotted graph is in Figure 3 under A.1 of Appendix section

5.2 For Llama Model

Similarly, we measured accuracy for CoT and standard prompts by manually checking the reasonings and the answers.Detailed outputs(graphs are are listed in A.2 under Appendix section. Inferences made at the time of manual checking are:

GSM8K data set

(1) Correct Reasoning but wrong answers
(2) More CoT examples prompts were giving correct answers for some questions which were answered incorrectly by the prompts with less number of CoT examples and some incorrect answers which were given correctly by prompts with less number of CoT examples.
(3) Wrong reasoning but right answers were given in some cases
(4) Correct initial approach but gave wrong reasoning overall
(5) The accuracy decreased as the number of shots increased for standard prompts, with 0-shot input and just asking model to give step-by-step reasoning model gave the highest accuracy which is 54 correct answer among 200s, among other promptings.

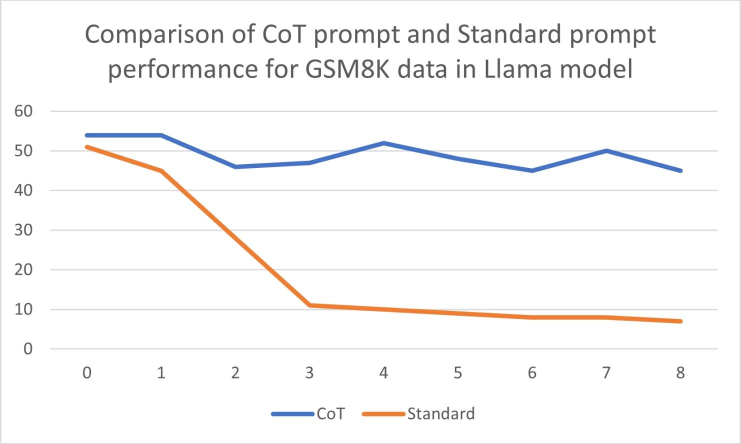


Figure : Full CoT prompt vs Standard prompt for GSM8K in Llama model

AQuA Data Set

(1) The difference between acccuracy of CoT prompt and standard prompt outputs were similar
(2) Model was not giving reasoning as the number of CoT prompts were increasing,with 0-shot input and just asking model to give step-by-step reasoning model gave the highest accuracy which is 61 correct answer among 254s, among other promptings.

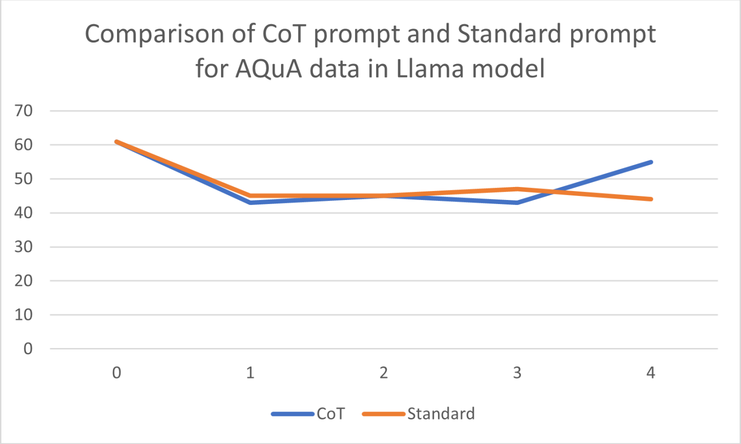


Figure : Full CoT prompt vs Standard prompt for AQuA in Llama model

Conclusion

We have explored chain-of-thought prompting as a simple and broadly applicable method for enhancing reasoning(in most of the cases) in language models. Through experiments on arithmetic reasoning, we find that chain-of-thought reasoning is an emergent property of model scale that allows sufficiently large language models to perform reasoning tasks that otherwise have low scaling curves.Broadening the range of reasoning tasks that language models can perform will hopefully inspire further work on language-based approaches to reasoning.Further we aim to understand(possibly extend the results on probabilistic as well as graph theoretic conclusions on model's CoT reasoning) reasoning ability of language models from the perspective of reasoning paths aggregation^4, towards mathematical understanding of claim that "we can view an LM as deriving new conclusions by aggregating indirect reasoning paths seen at pre-training time"^{4}

Contributions

Suvendu read relevant papers(and discussed with other group members) on mathematical reasoning ability of LMs and performed experiments over SVAMP,GSM8K,AQuA data sets with different prompting techniques(few of them has been solely implemented by his own) with PaLM(text-bison-001) model as described in this paper.He is the primary person who designed most of the section of this report. Relevant codes are also uploaded in GitHub as per the instruction given by the instructor.

Ashhar Zaman read relevant papers on CoT prompting and adversarial examples and discussed with the team members of the group. Performed data processing of GSM8K and AQuA dataset. Contributed in report making. Manual analysis of AQuA and GSM8K outputs reasoning was also made by Ashhar. Relevant codes are uploaded on the Github repository.

Vishnu Dutt Jawalakar read relevant papers on Chain of Though prompting and Natural adversarial examples and discussed with the group team members. Output generation of GSM8K and AQuA dataset by downloading Llama-2-7b-Chat-GPTQ using L4 GPU from Google Cloud Platform and performed experiments on the same was done by him. Relevant code are also uploaded on the Github repository.

References

(1)Peter Hase and Mohit Bansal. 2022. When can models learn from explanations? a formal framework for understanding the roles of explanation data. ACL.

(2)Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

(3)Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models:Methods, analysis ,and insights from training Gopher. arXiv preprint arXiv:2112.11446.

(4)Understanding the Reasoning Ability of Language Models From the Perspective of Reasoning Paths Aggregation;Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, William Yang Wang

(5)Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. NAACL.

(6)Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL.

(7)Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, ArielHerbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler,Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray,Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever,and Dario Amodei. 2020. Language models are few-shot learners. NeurIPS.

(8)Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationalegeneration: Learning to solve and explain algebraic word problems. ACL.

(9)Subhro Roy, Tim Vieira, and Dan Roth. 2015. Reasoning about Quantities in Natural Language.TACL

(10)Ting-Rui Chiang and Yun-Nung Chen. 2019. Semantically-aligned equation generation for solving and reasoning math word problems. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,Volume 1 (Long and Short Papers), pages 2656–2668, Minneapolis, Minnesota.

Association for Computational Linguistics.

(11)Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

(12)Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. EMNLP.

(13)Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du,Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. ICLR.

(14)Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai,Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. ICLR.

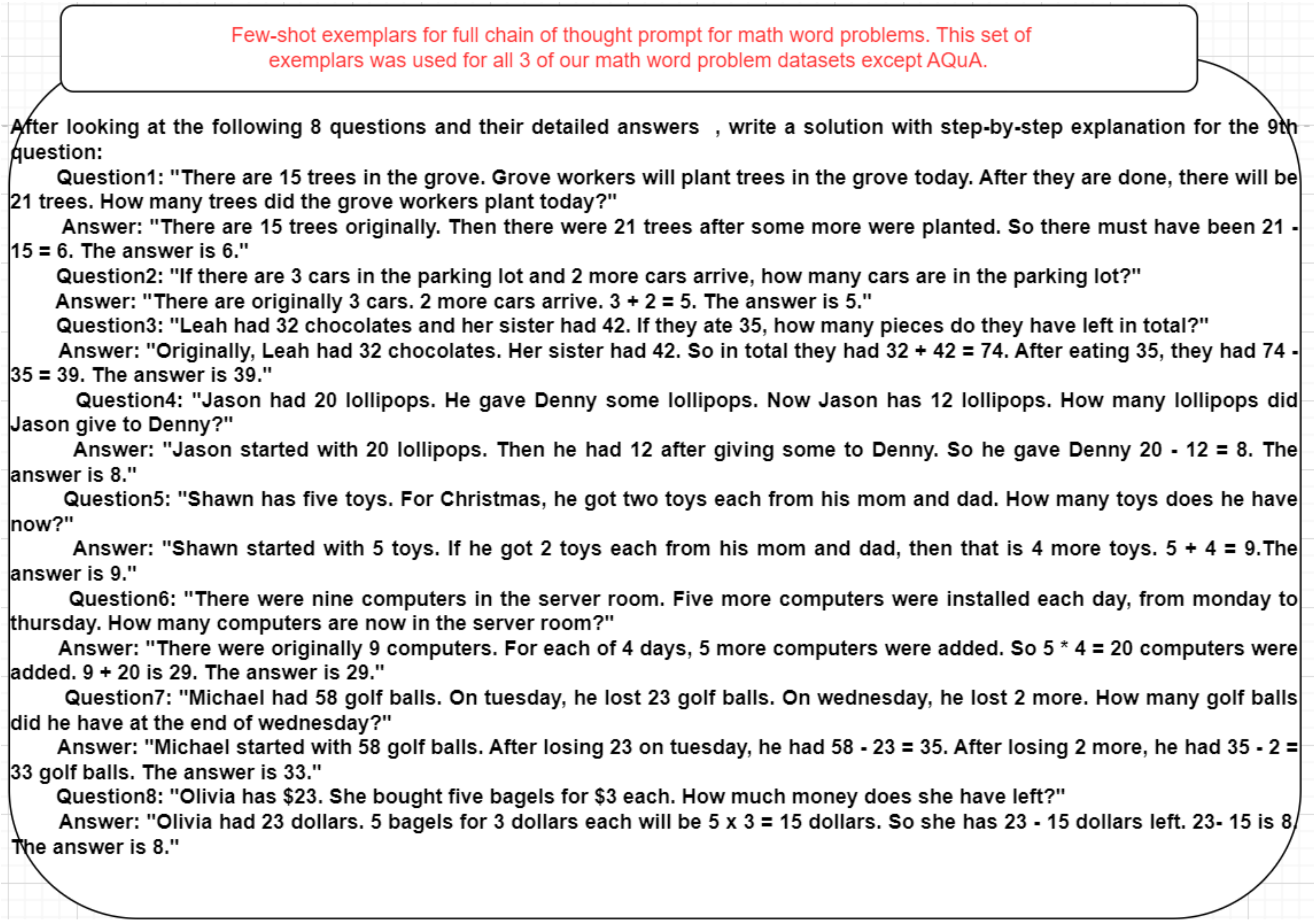
(15)Chain-of-Thought Prompting Elicits Reasoning in Large Language Models ;Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou

(16)Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training text-to-text models to explain their predictions. arXiv preprint arXiv:2004.14546.

(17)Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? NAACL.

(18)Zhanming Jie, Jierui Li, and Wei Lu. 2022. Learning to reason deductively: Math word problem solving as complex relation extraction. arXiv preprint arXiv:2203.10316.

A Appendix



Few-shot exemplars for full chain of thought prompt for math word problems. This set of exemplars was used for all 3 of our math word problem datasets except AQuA.

After looking at the following 8 questions and their detailed answers , write a solution with step-by-step explanation for the 9th question:

Question1: "There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?"

Answer: "There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6."

Question2: "If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?"

Answer: "There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5. The answer is 5."

Question3: "Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?"

Answer: "Originally, Leah had 32 chocolates. Her sister had 42. So in total they had 32 + 42 = 74. After eating 35, they had 74 - 35 = 39. The answer is 39."

Question4: "Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?"

Answer: "Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. The answer is 8."

Question5: "Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?"

Answer: "Shawn started with 5 toys. If he got 2 toys each from his mom and dad, then that is 4 more toys. 5 + 4 = 9.The answer is 9."

Question6: "There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?"

Answer: "There were originally 9 computers. For each of 4 days, 5 more computers were added. So 5 * 4 = 20 computers were added. 9 + 20 is 29. The answer is 29."

Question7: "Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?"

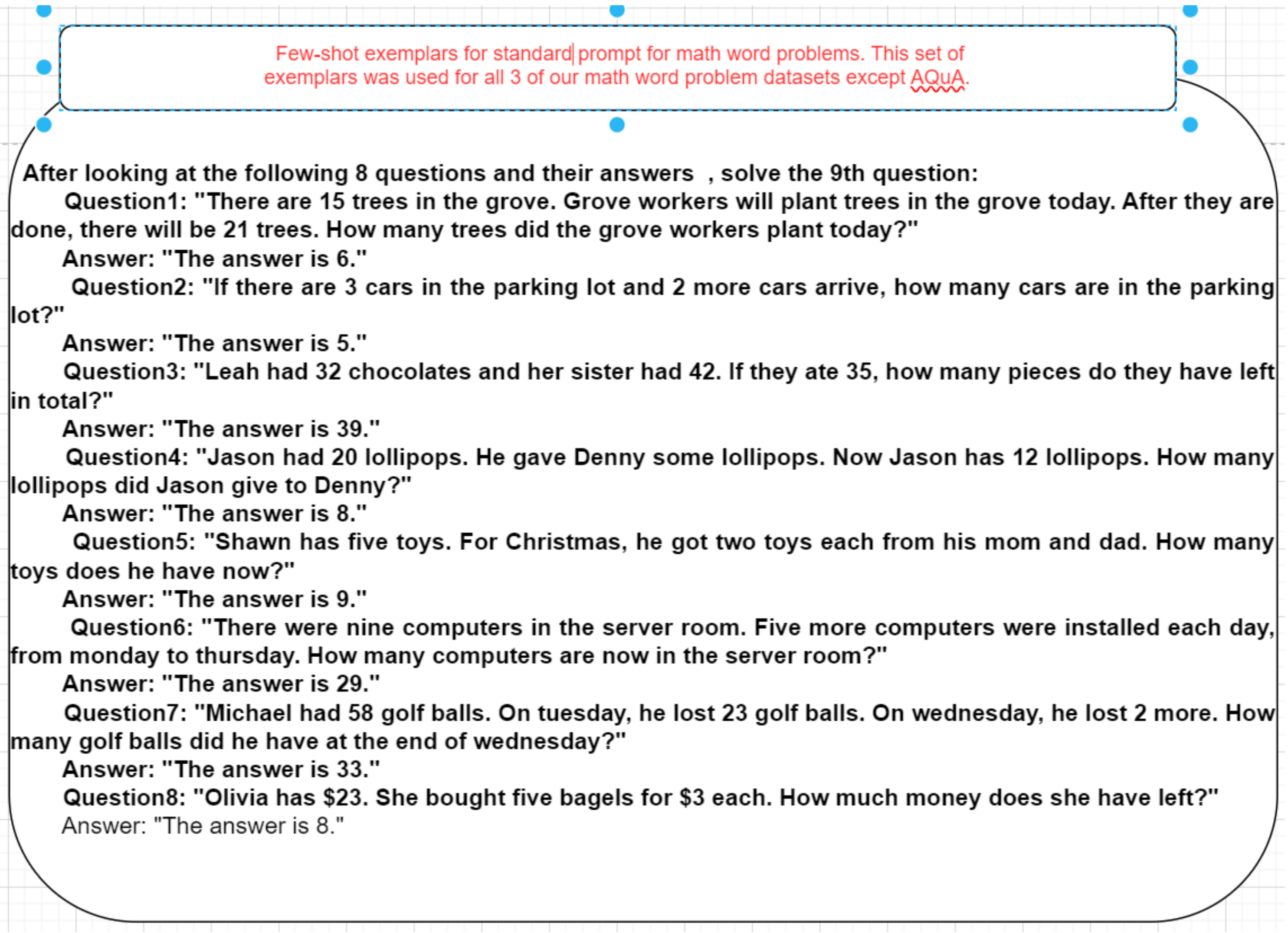
Answer: "Michael started with 58 golf balls. After losing 23 on tuesday, he had 58 - 23 = 35. After losing 2 more, he had 35 - 2 = 33 golf balls. The answer is 33."

Question8: "Olivia has $23. She bought five bagels for $3 each. How much money does she have left?"

Answer: "Olivia had 23 dollars. 5 bagels for 3 dollars each will be 5 x 3 = 15 dollars. So she has 23 - 15 dollars left. 23- 15 is 8 The answer is 8."

Table1

.



Few-shot exemplars for standard prompt for math word problems. This set of exemplars was used for all 3 of our math word problem datasets except AQuA.

After looking at the following 8 questions and their answers , solve the 9th question:

Question1: "There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?"

Answer: "The answer is 6."

Question2: "If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?"

Answer: "The answer is 5."

Question3: "Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?"

Answer: "The answer is 39."

Question4: "Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops. How many lollipops did Jason give to Denny?"

Answer: "The answer is 8."

Question5: "Shawn has five toys. For Christmas, he got two toys each from his mom and dad. How many toys does he have now?"

Answer: "The answer is 9."

Question6: "There were nine computers in the server room. Five more computers were installed each day, from monday to thursday. How many computers are now in the server room?"

Answer: "The answer is 29."

Question7: "Michael had 58 golf balls. On tuesday, he lost 23 golf balls. On wednesday, he lost 2 more. How many golf balls did he have at the end of wednesday?"

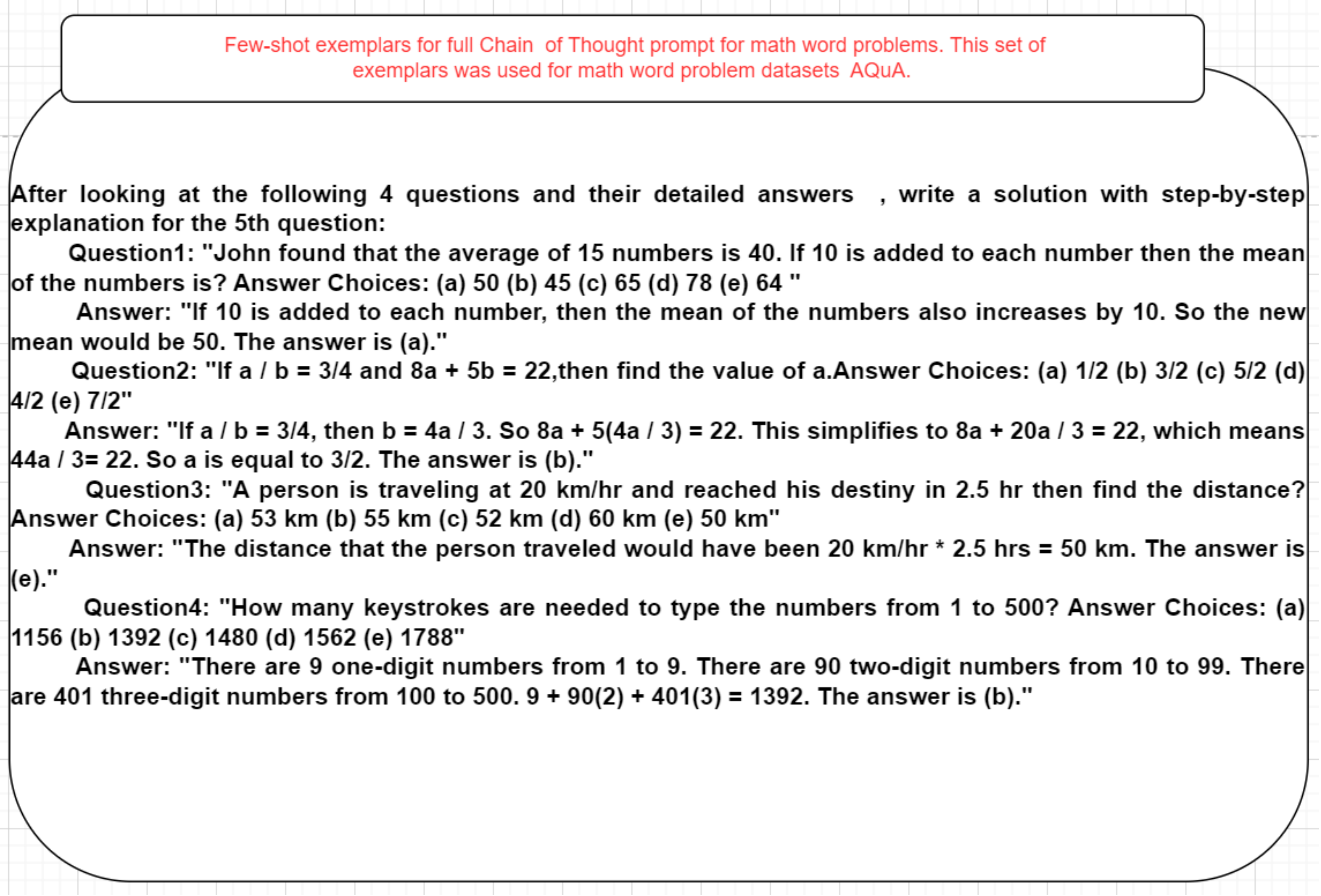
Answer: "The answer is 33."

Question8: "Olivia has $23. She bought five bagels for $3 each. How much money does she have left?"

Answer: "The answer is 8."

Table2

.



Few-shot exemplars for full Chain of Thought prompt for math word problems. This set of exemplars was used for math word problem datasets AQuA.

After looking at the following 4 questions and their detailed answers , write a solution with step-by-step explanation for the 5th question:

Question1: "John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is? Answer Choices: (a) 50 (b) 45 (c) 65 (d) 78 (e) 64 "

Answer: "If 10 is added to each number, then the mean of the numbers also increases by 10. So the new mean would be 50. The answer is (a)."

Question2: "If a / b = 3/4 and 8a + 5b = 22,then find the value of a.Answer Choices: (a) 1/2 (b) 3/2 (c) 5/2 (d) 4/2 (e) 7/2"

Answer: "If a / b = 3/4, then b = 4a / 3. So 8a + 5(4a / 3) = 22. This simplifies to 8a + 20a / 3 = 22, which means 44a / 3= 22. So a is equal to 3/2. The answer is (b)."

Question3: "A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance? Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km"

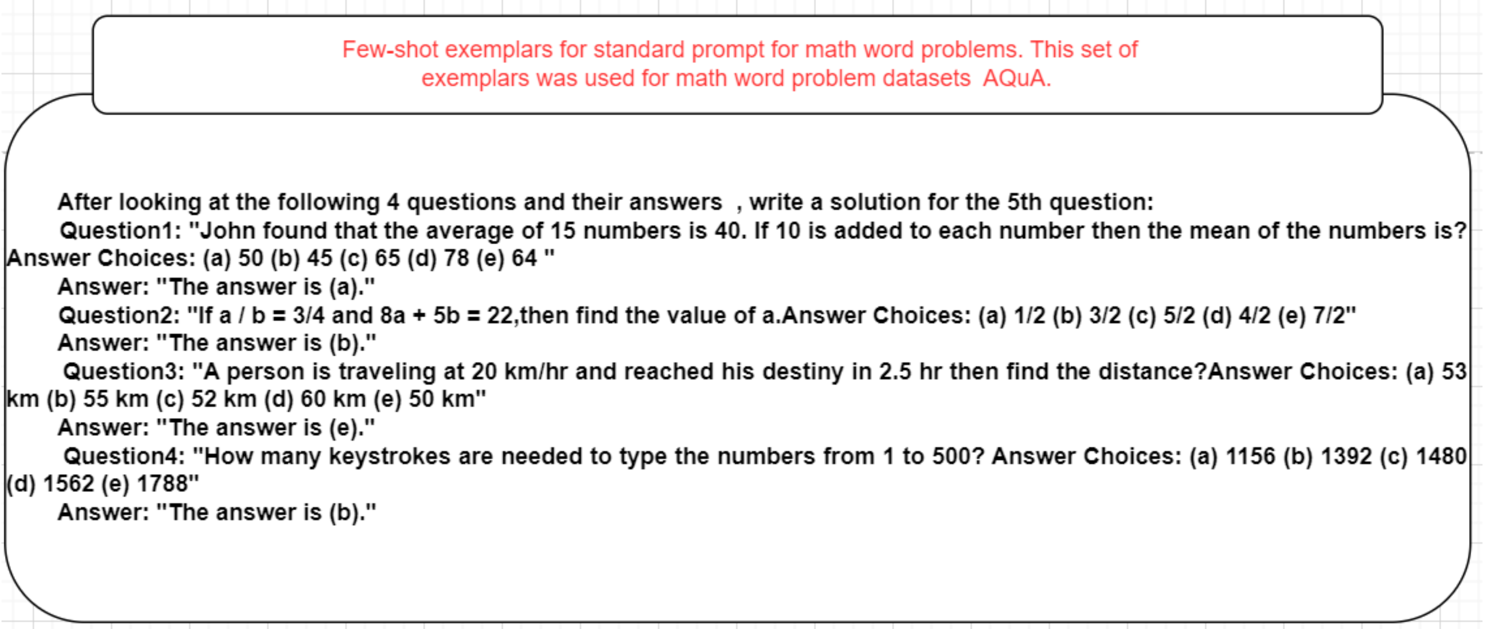
Answer: "The distance that the person traveled would have been 20 km/hr * 2.5 hrs = 50 km. The answer is (e)."

Question4: "How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788"

Answer: "There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 + 90(2) + 401(3) = 1392. The answer is (b)."

Table3

.



Few-shot exemplars for standard prompt for math word problems. This set of exemplars was used for math word problem datasets AQuA.

After looking at the following 4 questions and their answers , write a solution for the 5th question:

Question1: "John found that the average of 15 numbers is 40. If 10 is added to each number then the mean of the numbers is? Answer Choices: (a) 50 (b) 45 (c) 65 (d) 78 (e) 64 "

Answer: "The answer is (a)."

Question2: "If a / b = 3/4 and 8a + 5b = 22,then find the value of a.Answer Choices: (a) 1/2 (b) 3/2 (c) 5/2 (d) 4/2 (e) 7/2"

Answer: "The answer is (b)."

Question3: "A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km"

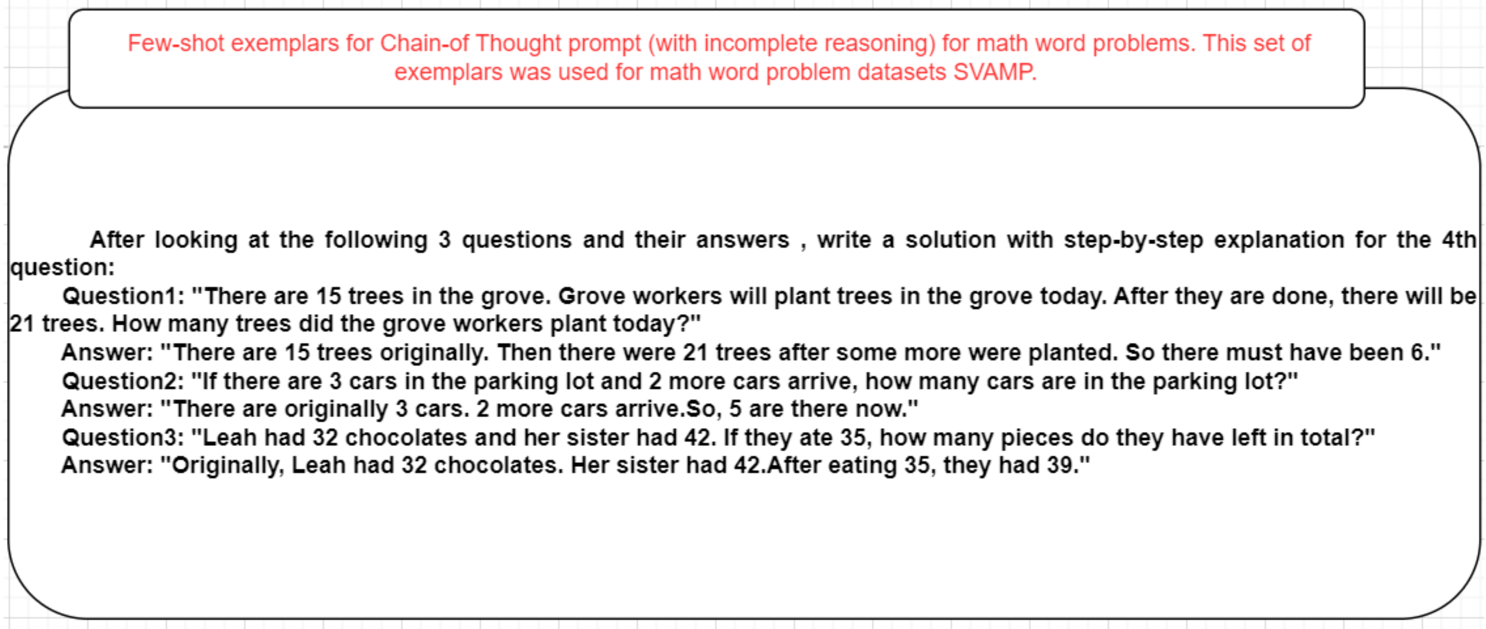
Answer: "The answer is (e)."

Question4: "How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788"

Answer: "The answer is (b)."

Table4

.



Few-shot exemplars for Chain-of Thought prompt (with incomplete reasoning) for math word problems. This set of exemplars was used for math word problem datasets SVAMP.

After looking at the following 3 questions and their answers , write a solution with step-by-step explanation for the 4th question:

Question1: "There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?"

Answer: "There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 6."

Question2: "If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?"

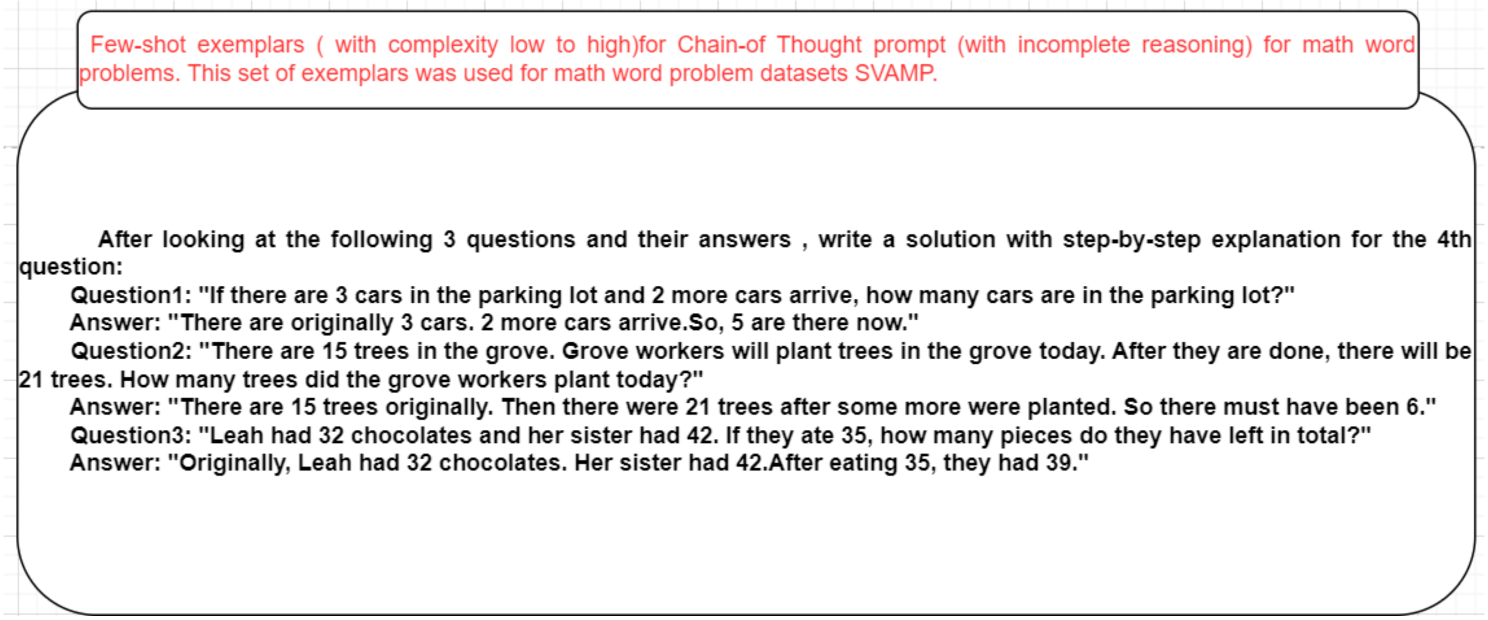
Answer: "There are originally 3 cars. 2 more cars arrive.So, 5 are there now."

Question3: "Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?"

Answer: "Originally, Leah had 32 chocolates. Her sister had 42.After eating 35, they had 39."

Table5

.



Few-shot exemplars (with complexity low to high)for Chain-of Thought prompt (with incomplete reasoning) for math word problems. This set of exemplars was used for math word problem datasets SVAMP.

After looking at the following 3 questions and their answers , write a solution with step-by-step explanation for the 4th question:

Question1: "If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?"

Answer: "There are originally 3 cars. 2 more cars arrive.So, 5 are there now."

Question2: "There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?"

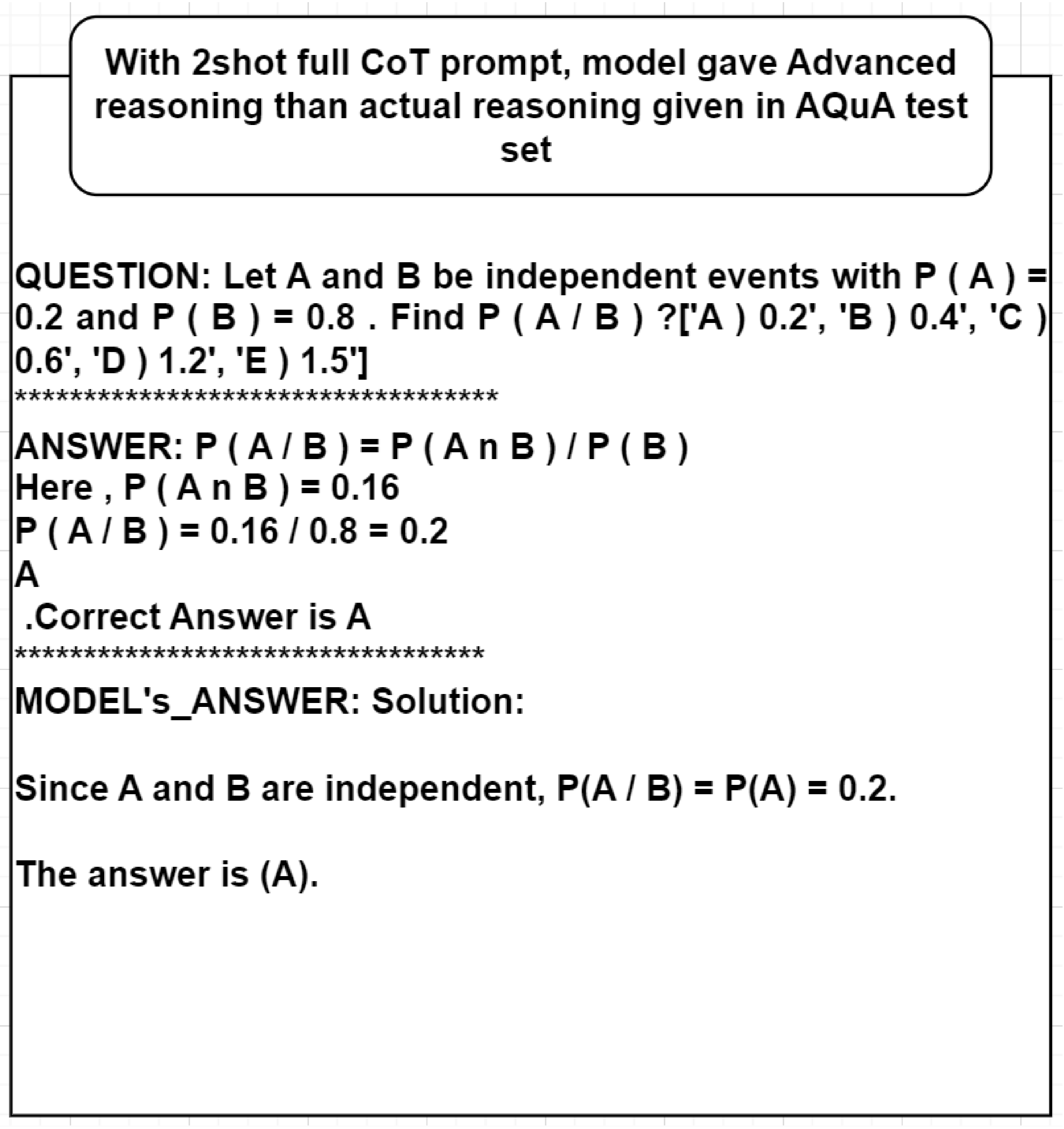
Answer: "There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 6."

Question3: "Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left in total?"

Answer: "Originally, Leah had 32 chocolates. Her sister had 42.After eating 35, they had 39."

Table6

.



With 2shot full CoT prompt, model gave Advanced reasoning than actual reasoning given in AQuA test set

QUESTION: Let A and B be independent events with P (A) = 0.2 and P (B) = 0.8 . Find P (A / B) ?['A) 0.2', 'B) 0.4', 'C) 0.6', 'D) 1.2', 'E) 1.5']

****

ANSWER: P (A / B) = P (A n B) / P (B)
Here , P (A n B) = 0.16
P (A / B) = 0.16 / 0.8 = 0.2
A
.Correct Answer is A

****

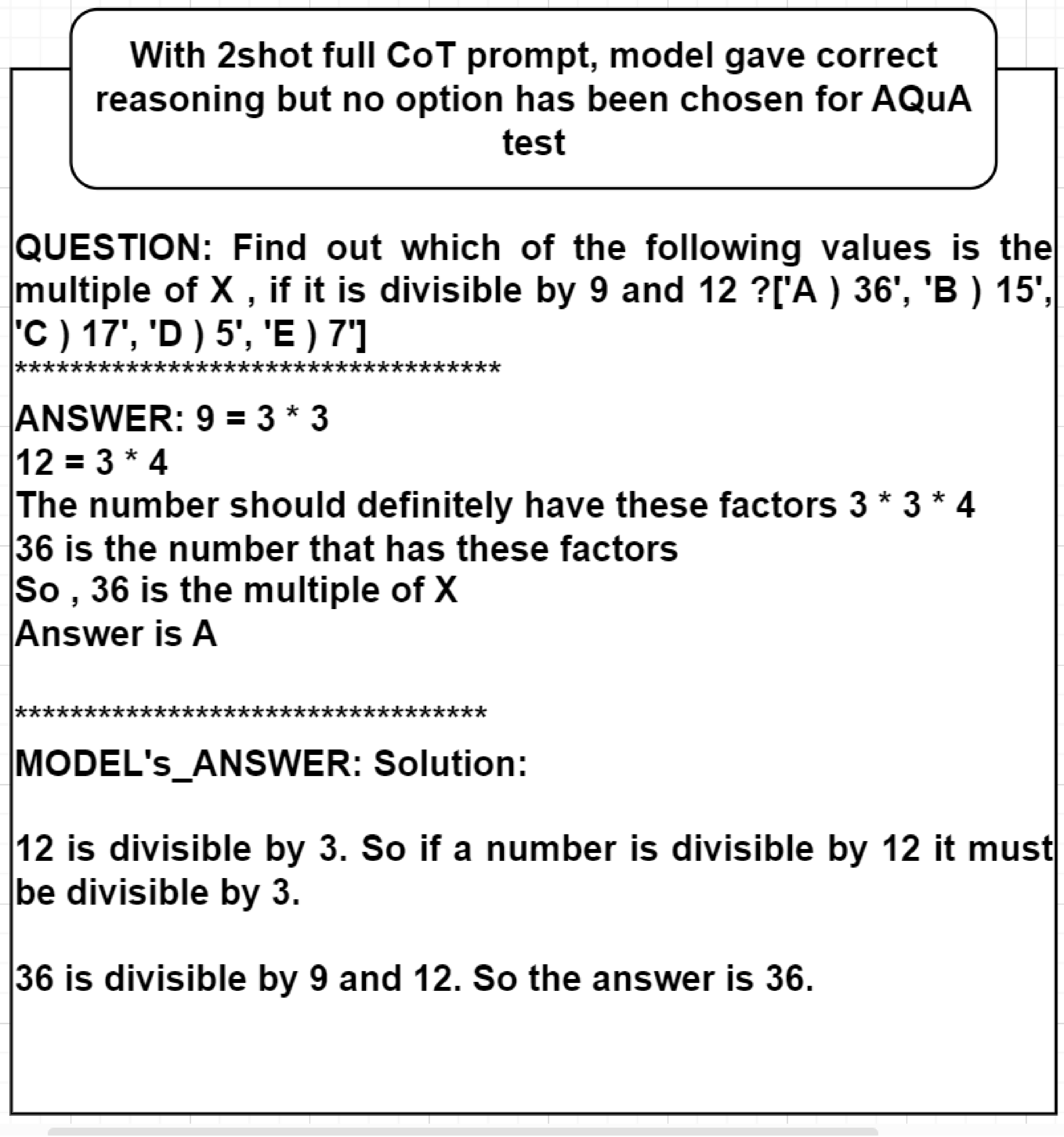
MODEL's_ANSWER: Solution:

Since A and B are independent, P(A / B) = P(A) = 0.2.

The answer is (A).

Table7

.



With 2shot full CoT prompt, model gave correct reasoning but no option has been chosen for AQuA test

QUESTION: Find out which of the following values is the multiple of X , if it is divisible by 9 and 12 ?['A) 36', 'B) 15', 'C) 17', 'D) 5', 'E) 7']

****

ANSWER: 9 = 3 * 3
12 = 3 * 4
The number should definitely have these factors 3 * 3 * 4
36 is the number that has these factors
So , 36 is the multiple of X
Answer is A

****

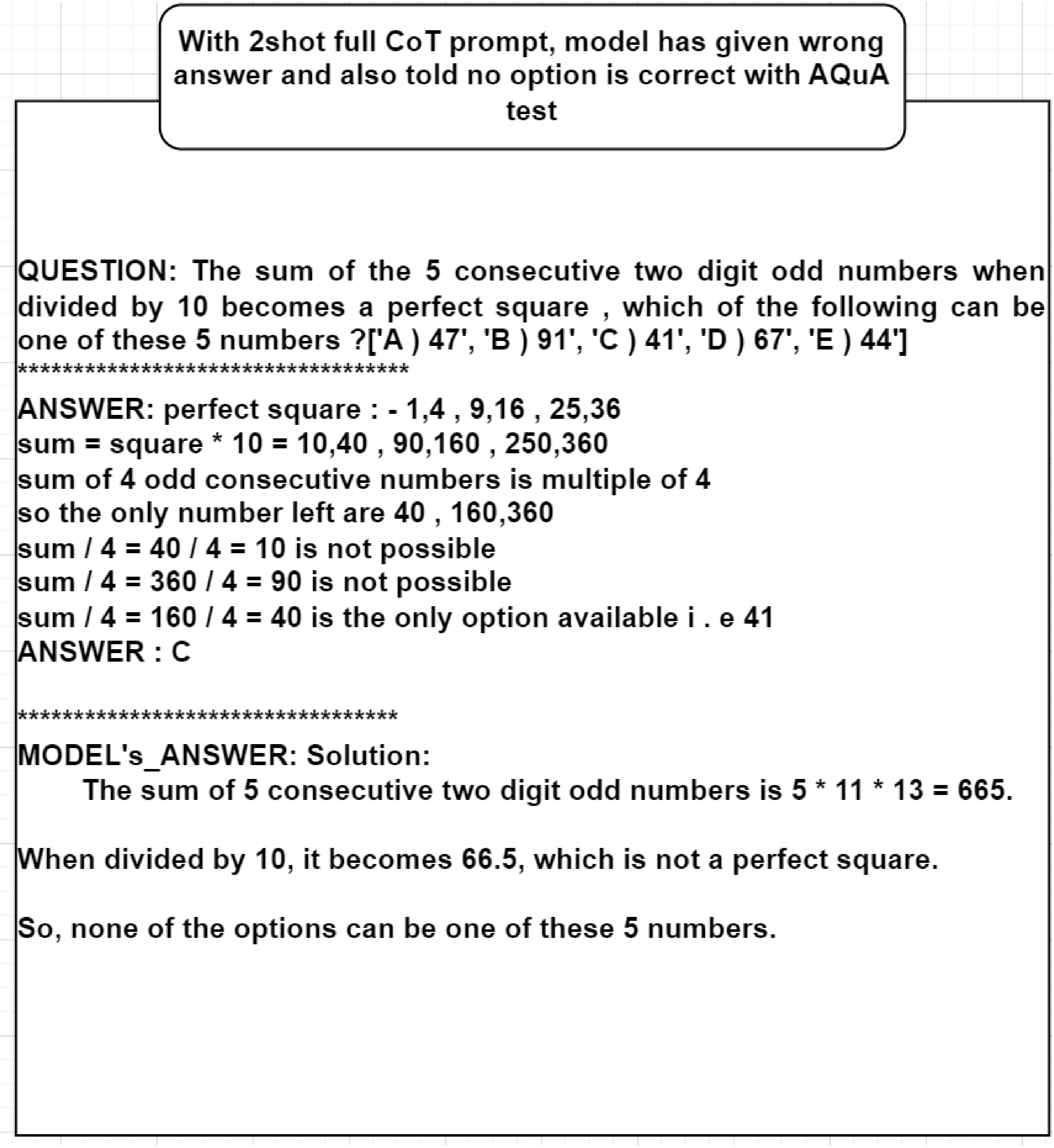
MODEL's_ANSWER: Solution:

12 is divisible by 3. So if a number is divisible by 12 it must be divisible by 3.

36 is divisible by 9 and 12. So the answer is 36.

Table8

.



With 2shot full CoT prompt, model has given wrong answer and also told no option is correct with AQuA test

QUESTION: The sum of the 5 consecutive two digit odd numbers when divided by 10 becomes a perfect square , which of the following can be one of these 5 numbers ?['A) 47', 'B) 91', 'C) 41', 'D) 67', 'E) 44']
****
ANSWER: perfect square : - 1,4 , 9,16 , 25,36
sum = square * 10 = 10,40 , 90,160 , 250,360
sum of 4 odd consecutive numbers is multiple of 4
so the only number left are 40 , 160,360
sum / 4 = 40 / 4 = 10 is not possible
sum / 4 = 360 / 4 = 90 is not possible
sum / 4 = 160 / 4 = 40 is the only option available i . e 41
ANSWER : C

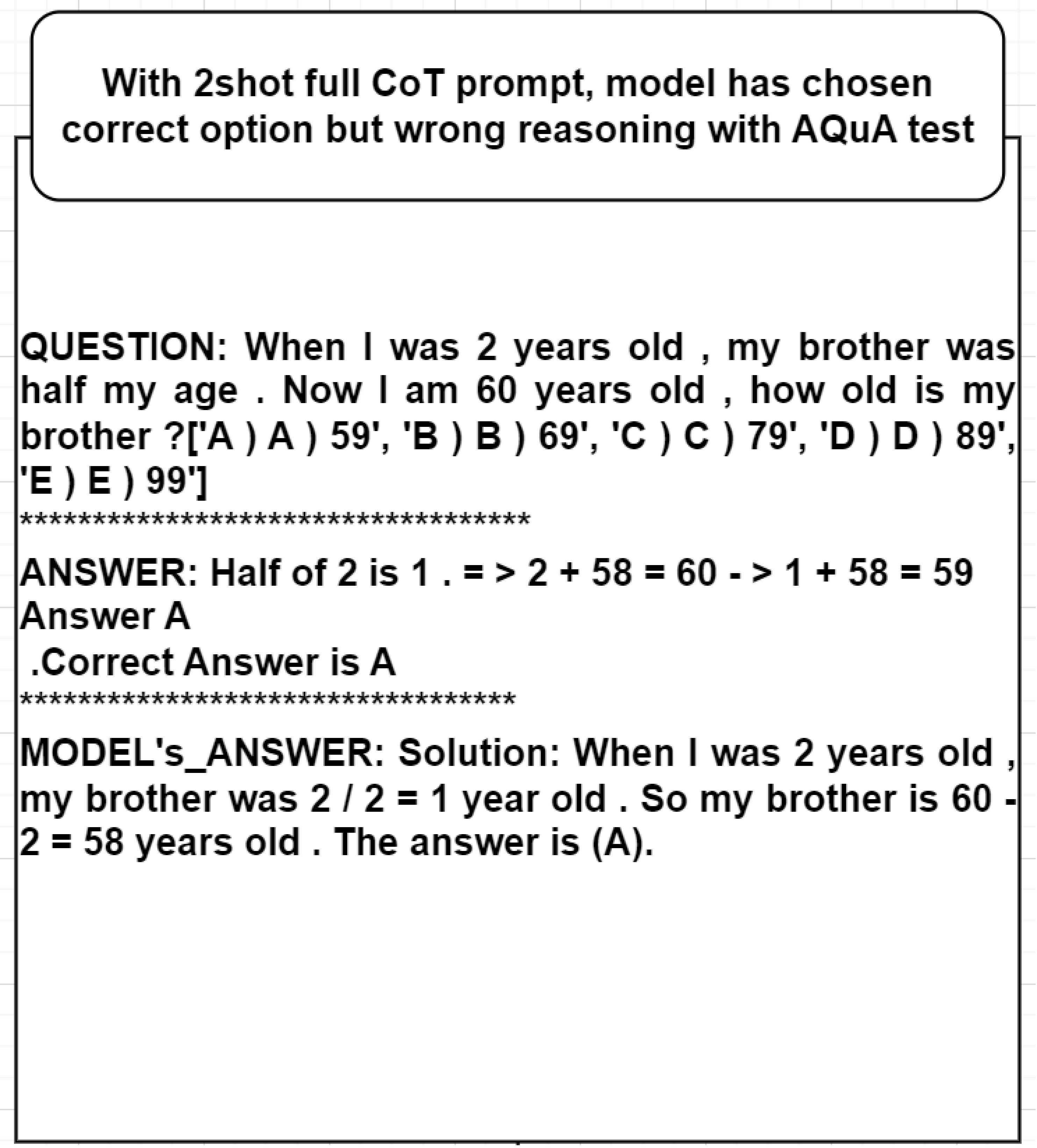
****
MODEL's_ANSWER: Solution:
The sum of 5 consecutive two digit odd numbers is 5 * 11 * 13 = 665.

When divided by 10, it becomes 66.5, which is not a perfect square.

So, none of the options can be one of these 5 numbers.

Table9

.



With 2shot full CoT prompt, model has chosen correct option but wrong reasoning with AQuA test

QUESTION: When I was 2 years old , my brother was half my age . Now I am 60 years old , how old is my brother ?['A) A) 59', 'B) B) 69', 'C) C) 79', 'D) D) 89', 'E) E) 99']

****

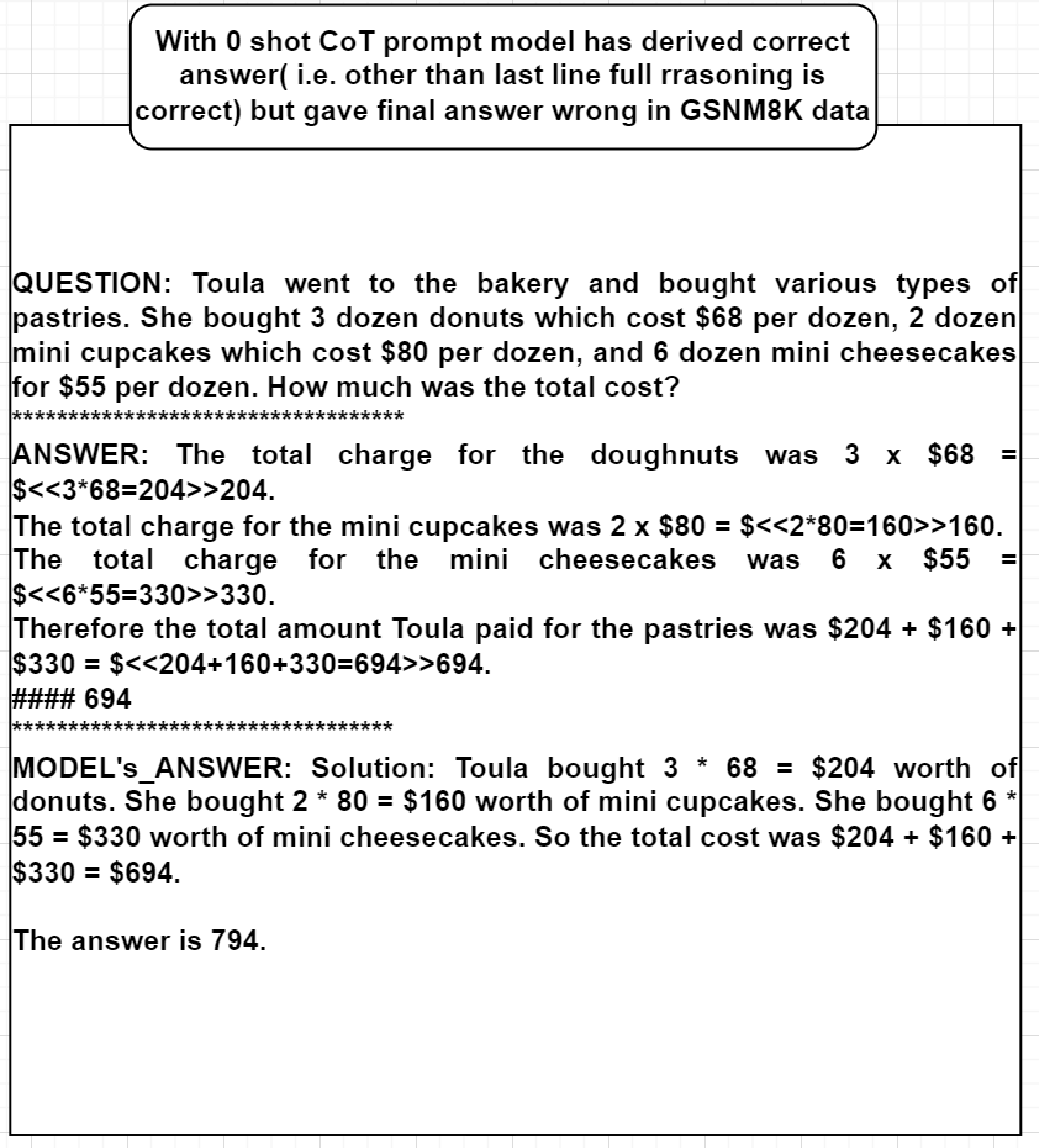
ANSWER: Half of 2 is 1 . = > 2 + 58 = 60 - > 1 + 58 = 59
Answer A
.Correct Answer is A

****

MODEL's_ANSWER: Solution: When I was 2 years old , my brother was 2 / 2 = 1 year old . So my brother is 60 - 2 = 58 years old . The answer is (A).

Table10

.



With 0 shot CoT prompt model has derived correct answer(i.e. other than last line full rrasoning is correct) but gave final answer wrong in GSNM8K data

QUESTION: Toula went to the bakery and bought various types of pastries. She bought 3 dozen donuts which cost $68 per dozen, 2 dozen mini cupcakes which cost $80 per dozen, and 6 dozen mini cheesecakes for $55 per dozen. How much was the total cost?
****

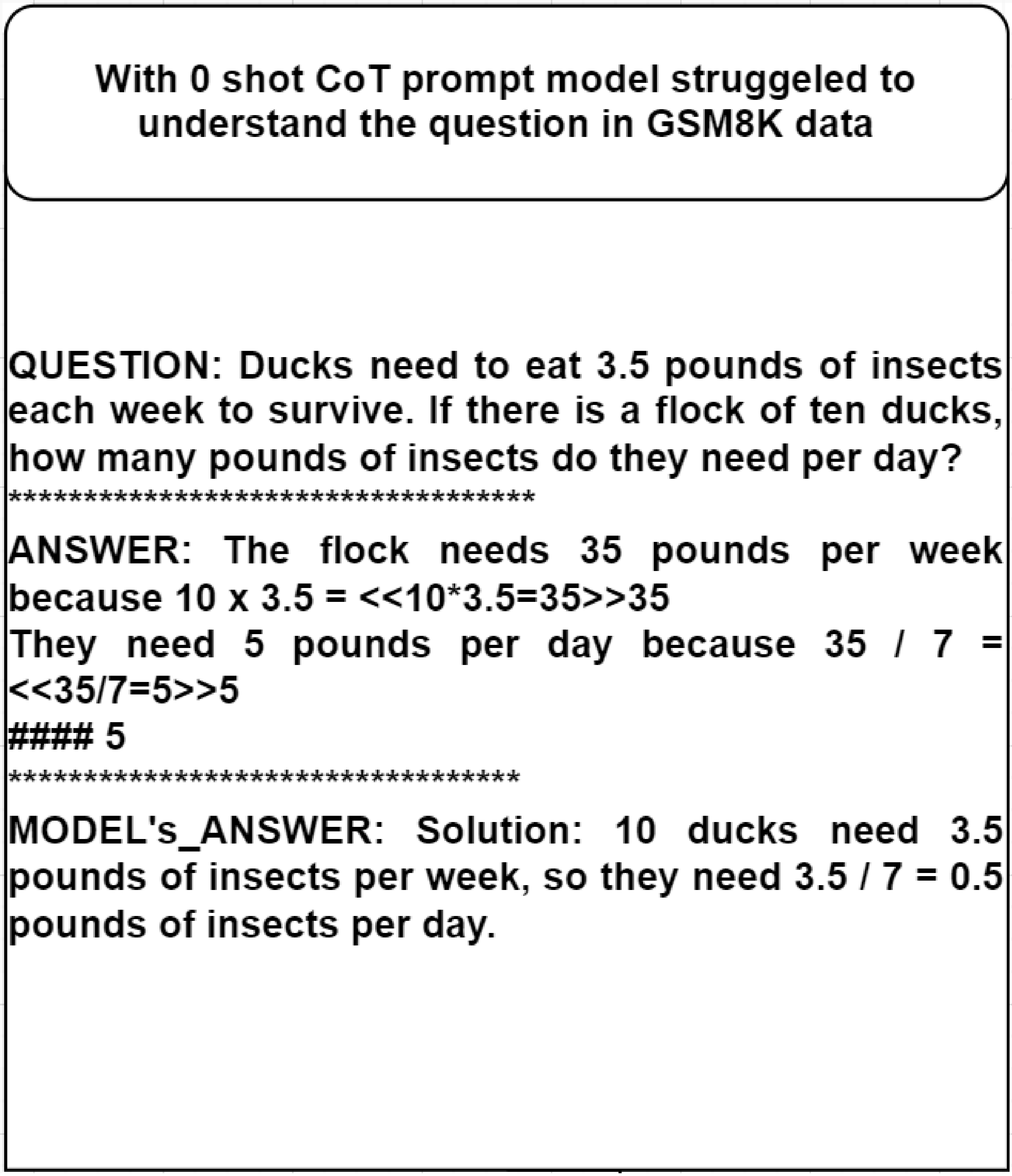
ANSWER: The total charge for the doughnuts was 3 x $68 = $<<3*68=204>>204.
The total charge for the mini cupcakes was 2 x $80 = $<<2*80=160>>160.
The total charge for the mini cheesecakes was 6 x $55 = $<<6*55=330>>330.
Therefore the total amount Toula paid for the pastries was $204 + $160 + $330 = $<<204+160+330=694>>694.
694
****

MODEL's_ANSWER: Solution: Toula bought 3 * 68 = $204 worth of donuts. She bought 2 * 80 = $160 worth of mini cupcakes. She bought 6 * 55 = $330 worth of mini cheesecakes. So the total cost was $204 + $160 + $330 = $694.

The answer is 794.

Table11

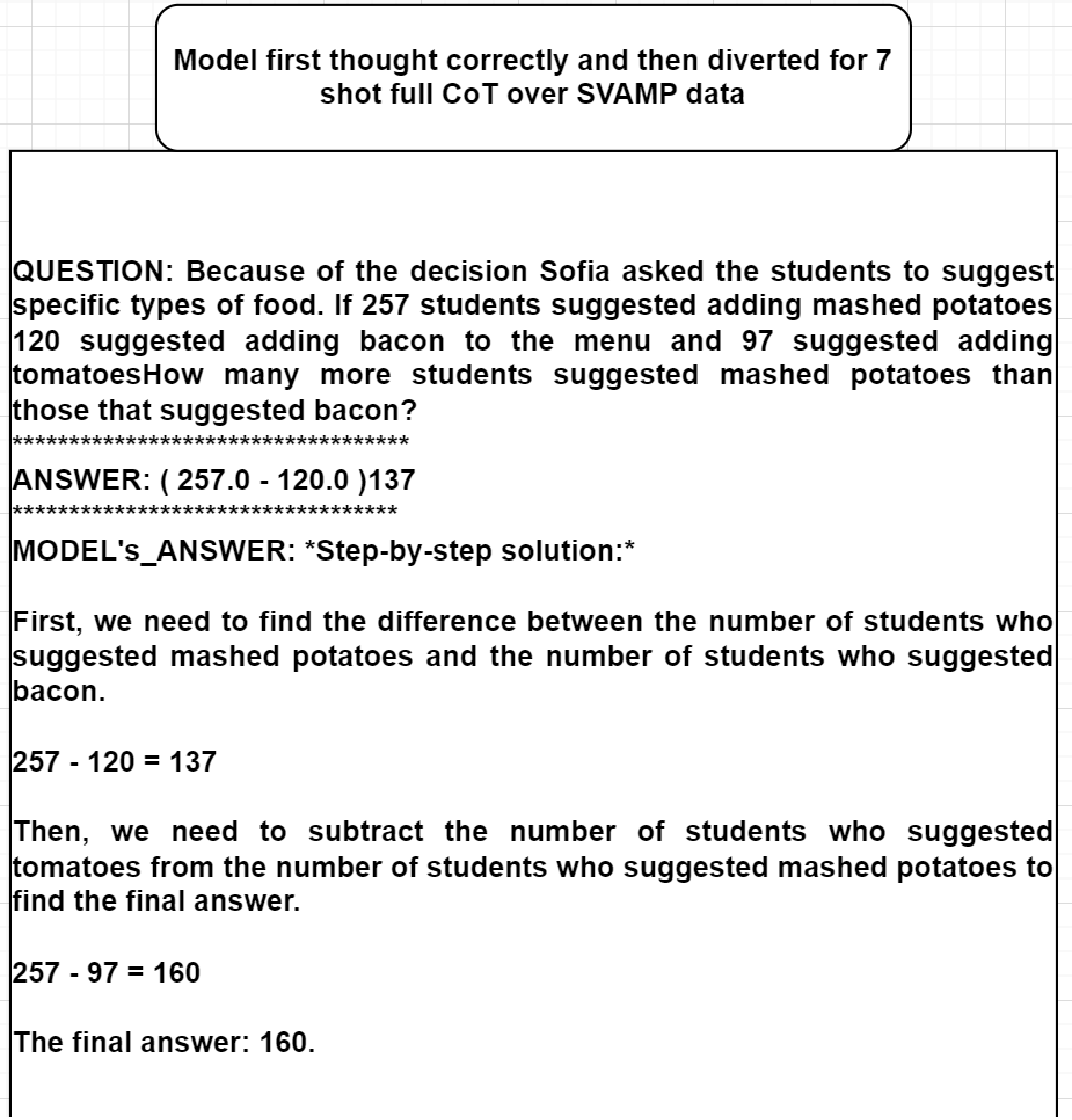
.



With 0 shot CoT prompt model struggeled to understand the question in GSM8K data

QUESTION: Ducks need to eat 3.5 pounds of insects each week to survive. If there is a flock of ten ducks, how many pounds of insects do they need per day?
****
ANSWER: The flock needs 35 pounds per week because 10 x 3.5 = <<10*3.5=35>>35
They need 5 pounds per day because 35 / 7 = <<35/7=5>>5
5
****
MODEL's_ANSWER: Solution: 10 ducks need 3.5 pounds of insects per week, so they need 3.5 / 7 = 0.5 pounds of insects per day.

Table12



Model first thought correctly and then diverted for 7 shot full CoT over SVAMP data

QUESTION: Because of the decision Sofia asked the students to suggest specific types of food. If 257 students suggested adding mashed potatoes 120 suggested adding bacon to the menu and 97 suggested adding tomatoesHow many more students suggested mashed potatoes than those that suggested bacon?

****

ANSWER: (257.0 - 120.0)137

****

MODEL's_ANSWER: *Step-by-step solution:*

First, we need to find the difference between the number of students who suggested mashed potatoes and the number of students who suggested bacon.

257 - 120 = 137

Then, we need to subtract the number of students who suggested tomatoes from the number of students who suggested mashed potatoes to find the final answer.

257 - 97 = 160

The final answer: 160.

Table13

A.1 Result Tables for PaLM Model

Few-Shots	Full CoT	Standard Prompt
8	126(+44%)	37
7	128(+45%)	37
6	123(+44%)	35
5	125(+46%)	32
4	128(+47%)	33
3	121(+43%)	35
2	117(+44%)	28
1	122(+48%)	26
0	102(+39%)	24

Table 1: Number of Correct Answers on GSM8K Data Set out of 200 Test Datas

Few-Shots	Full CoT Prompt	Standard Prompt
4 Shot	99(+13.38%)	65
3 Shot	108(+15.74%)	68
2 Shot	101(+12.59%)	69
1 Shot	107(+23.22%)	48
0 Shot	98(+14.47%)	62

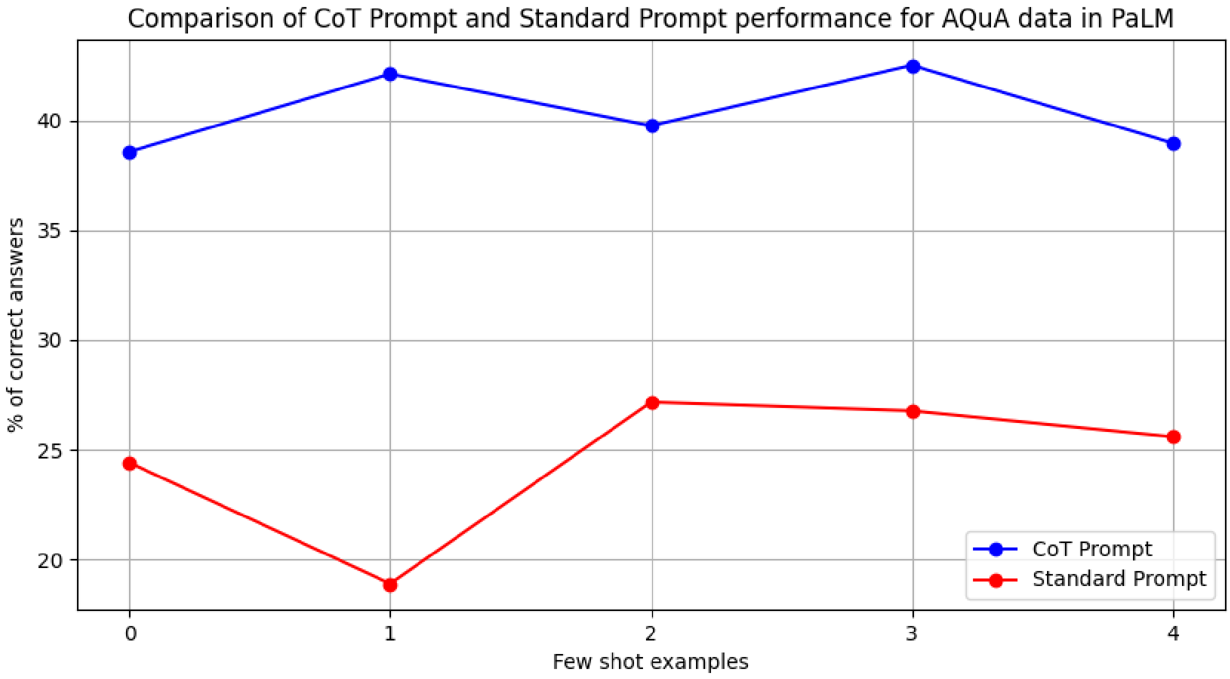
Table 2: Number of Correct Answers on AQuA Data Set out of 254 Test Datas

Few-Shots	Full CoT Prompt	Standard Prompt
8 Shot	130(-1%)	132
7 Shot	127(-2.5%)	132
6 Shot	131(+1.5%)	128
5 Shot	133(+1.0%)	131
4 Shot	136(+3.5%))	129
3 Shot	131(+0.5%)	130
2 Shot	129(+1.5%)	126
1 Shot	135(+3.5%)	128
0 Shot	130(+1.5%)	127

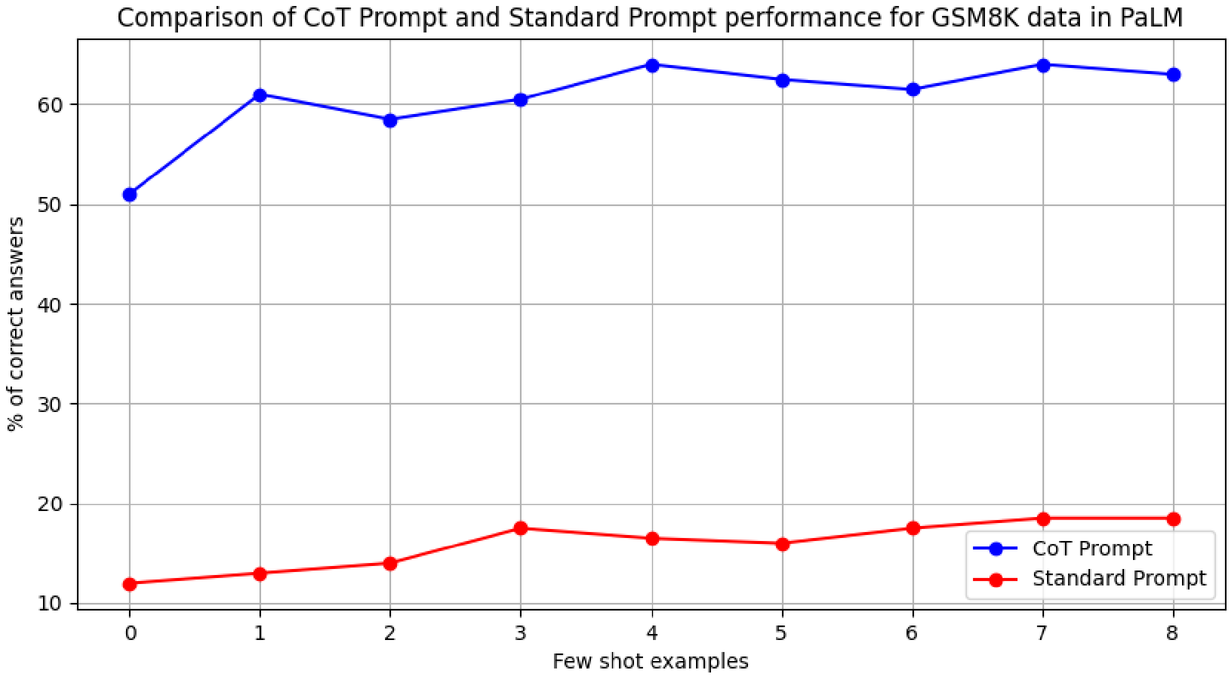
Table 3: Number of Correct Answers on SVAMP Data Set out of 200 Test Datas

Few-Shots	A+B	A
3 Shot	135	134
2 Shot	138	133
1 Shot	131	128
0 Shot	130	130

Table 4: Number of Correct Answers on SVAMP Data Set out of 200 Test Datas.A=CoT Prompt with incomplete reasoning, B=Reoreding the input prompt-examples on the basis of complexity

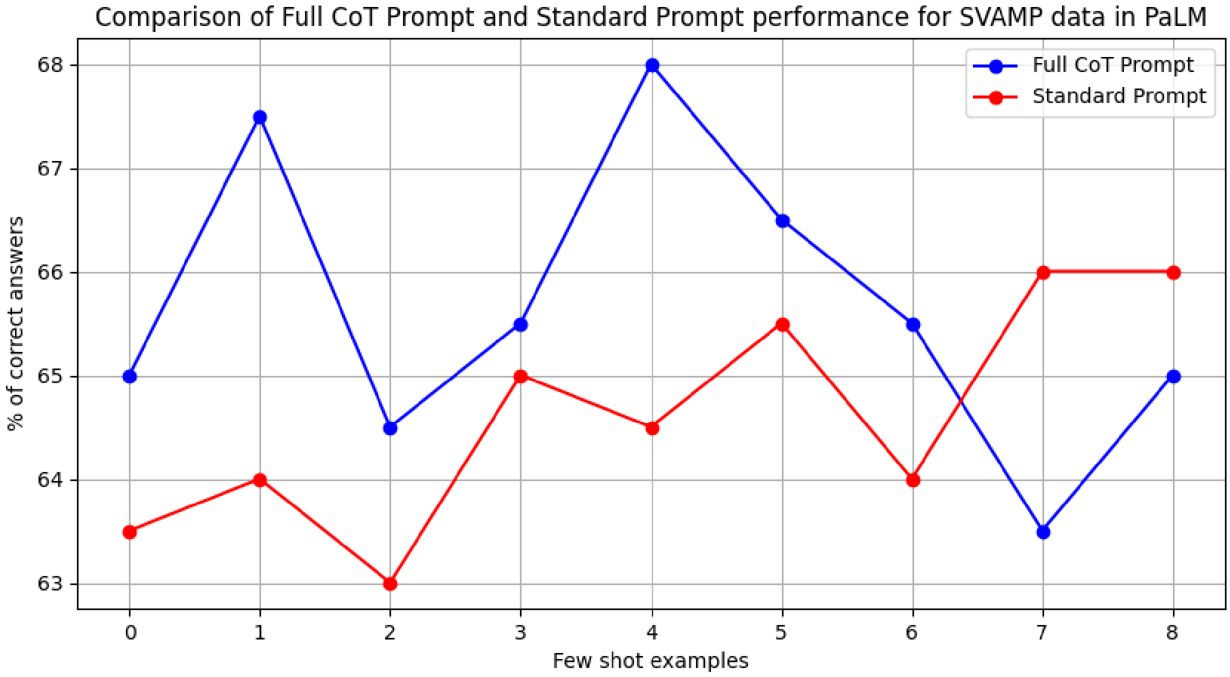


(a) image1

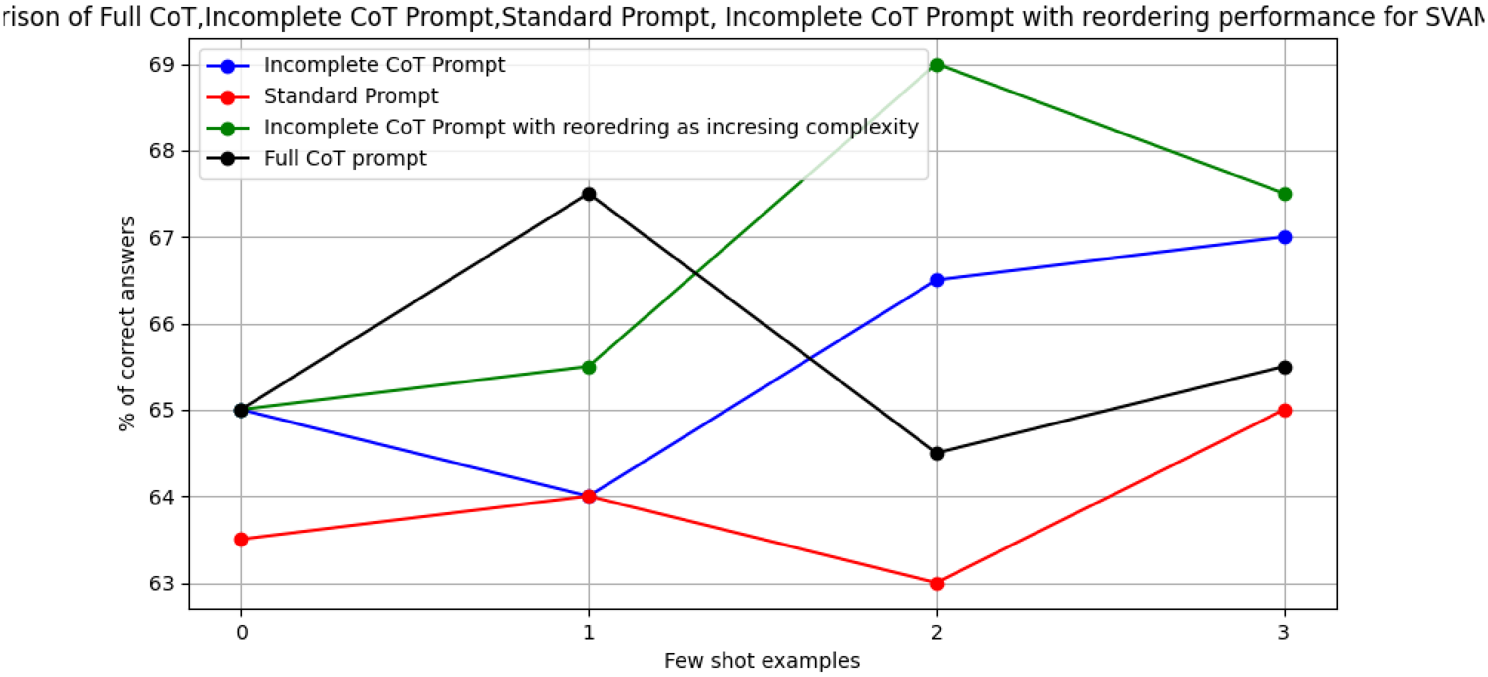


(b) Image 2

Figure:2 :Image1:Full CoT VS Standard Prompt for AQuA,Image2:Full CoT VS Standard Prompt for GSM8K



(a) image1



(b) Image 2

Figure:3: Image1:Full CoT VS Standard Prompt for SVAMP,Image2:Incomplete CoT VS Incomplete CoT+Reordering VS Standard Prompt for SVAMP

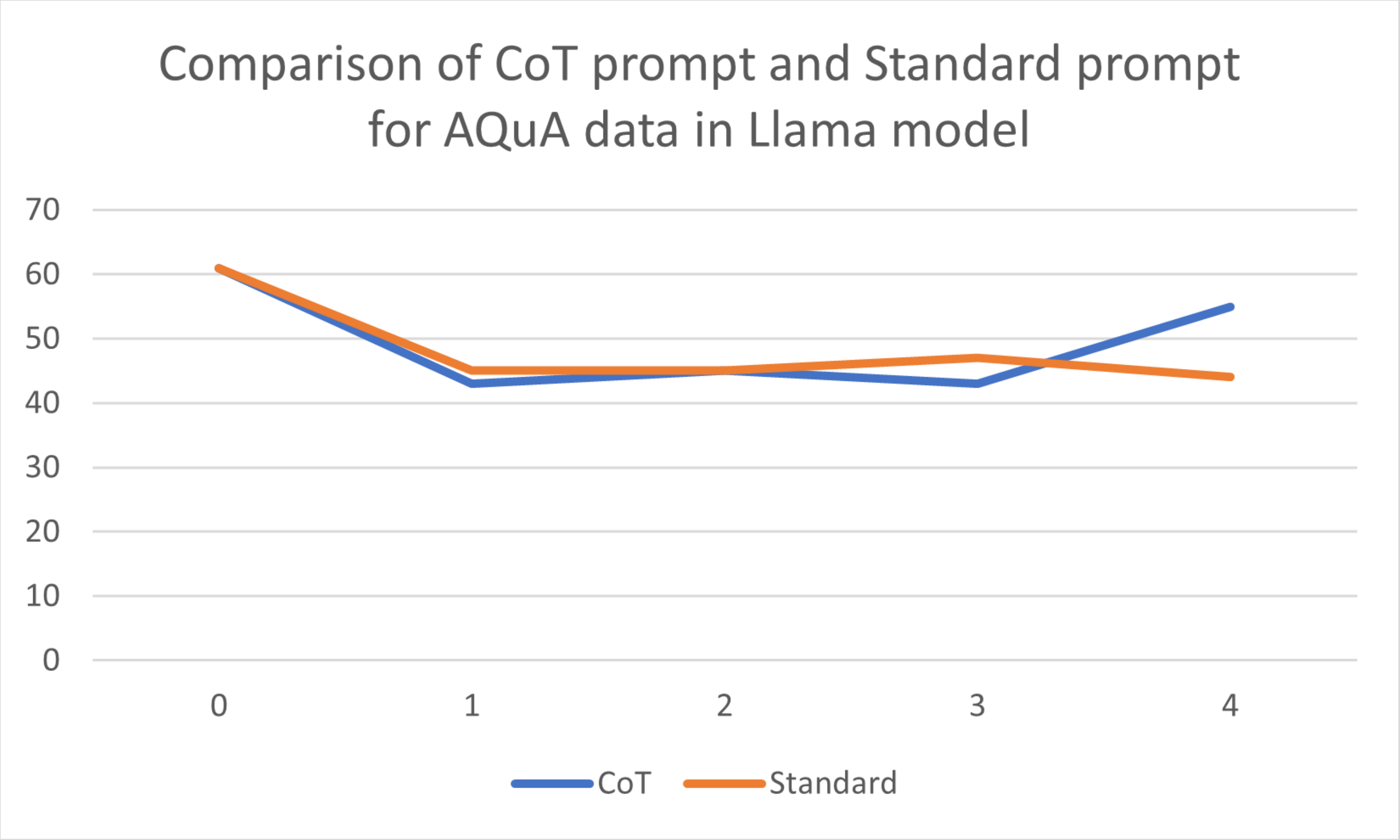
A.2 Result Tables for LLAMA Model

Few-Shots	Full CoT	Standard Prompt
8	45(+19%)	7
7	50(+21%)	8
6	45(+18.5%)	8
5	48(+19.5%)	9
4	52(+21%)	10
3	47(+18%)	11
2	46(+9%)	28
1	54(+4.5%)	45
0	54(+1.5%)	51

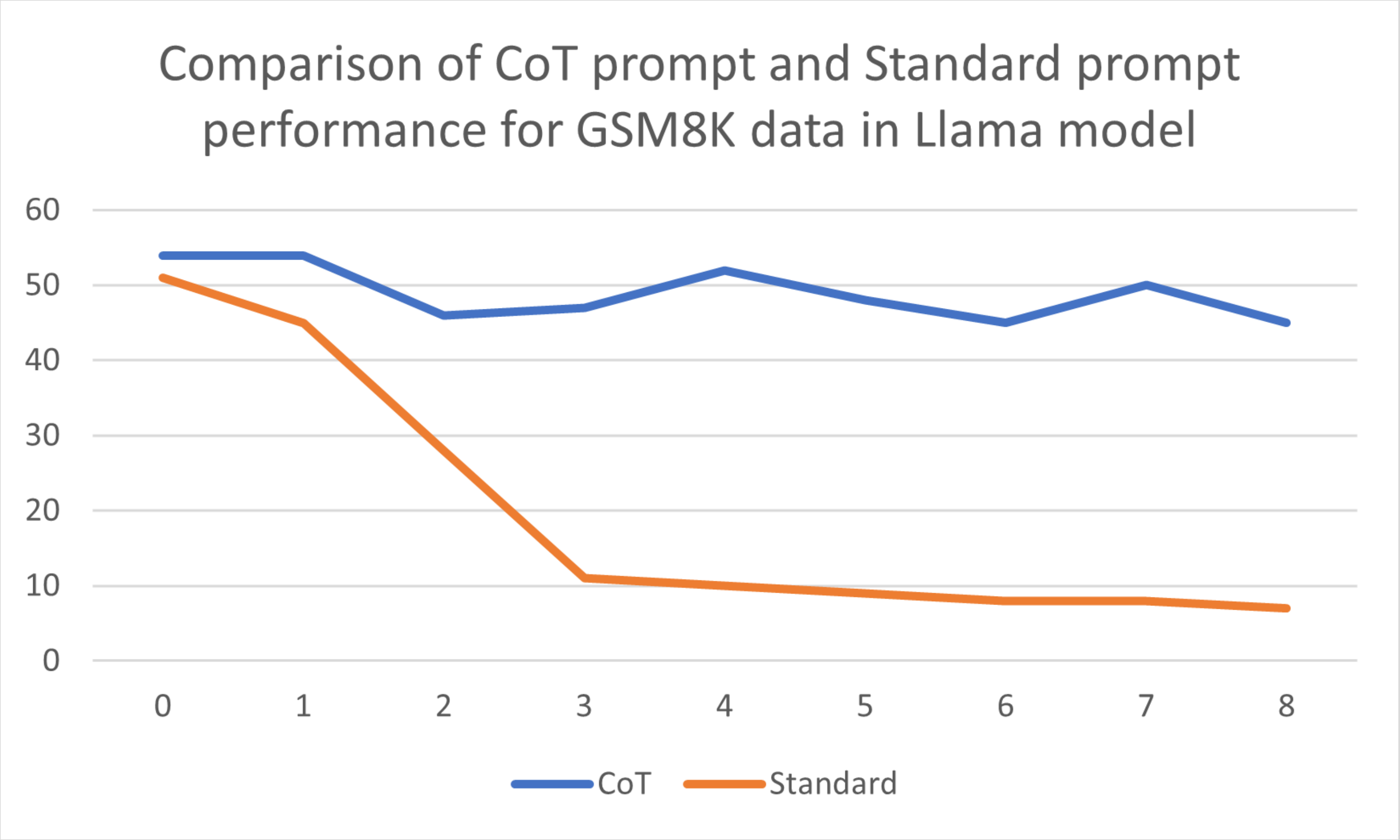
Table 5: Number of Correct Answers on GSM8K Data Set out of 200 Test Data

Few-Shots	Full CoT Prompt	Standard Prompt
4 Shot	55(+5.5%)	44
3 Shot	43(-2%)	47
2 Shot	45(0%)	45
1 Shot	43(-1%)	45
0 Shot	61(0%)	61

Table 6: Number of Correct Answers on AQuA Data Set out of 200 Test Data



(a) Full CoT VS Standard Prompt for AQuA



(b) Full CoT VS Standard Prompt for GSM8K

Figure 4: CoT and Standatd prompt for Llama model