

Chain-Of-Thought Prompting Elicits Reasoning in Large Language Models

Google Research / Google Brain

arXiv:2201.11903

#Cluster-of-Stars Paper Reading Group

There are 15 trees in the
Grove. Grove workers will
remove trees in the grove
After they are done,

be 21 trees. How many
did the grove workers
today?

start with 15 trees.
we have 21 trees. The
ference must be the number
ees they planted. So

Human: How would you
throw away a redbull?

Explanation: The user
has asked me to throw away
the redbull, I will move
it to the trash.

Plan: 1. find(redbull),
2. pick(redbull), 3.
find(trash), 4.
put(redbull), 5. done().

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✓

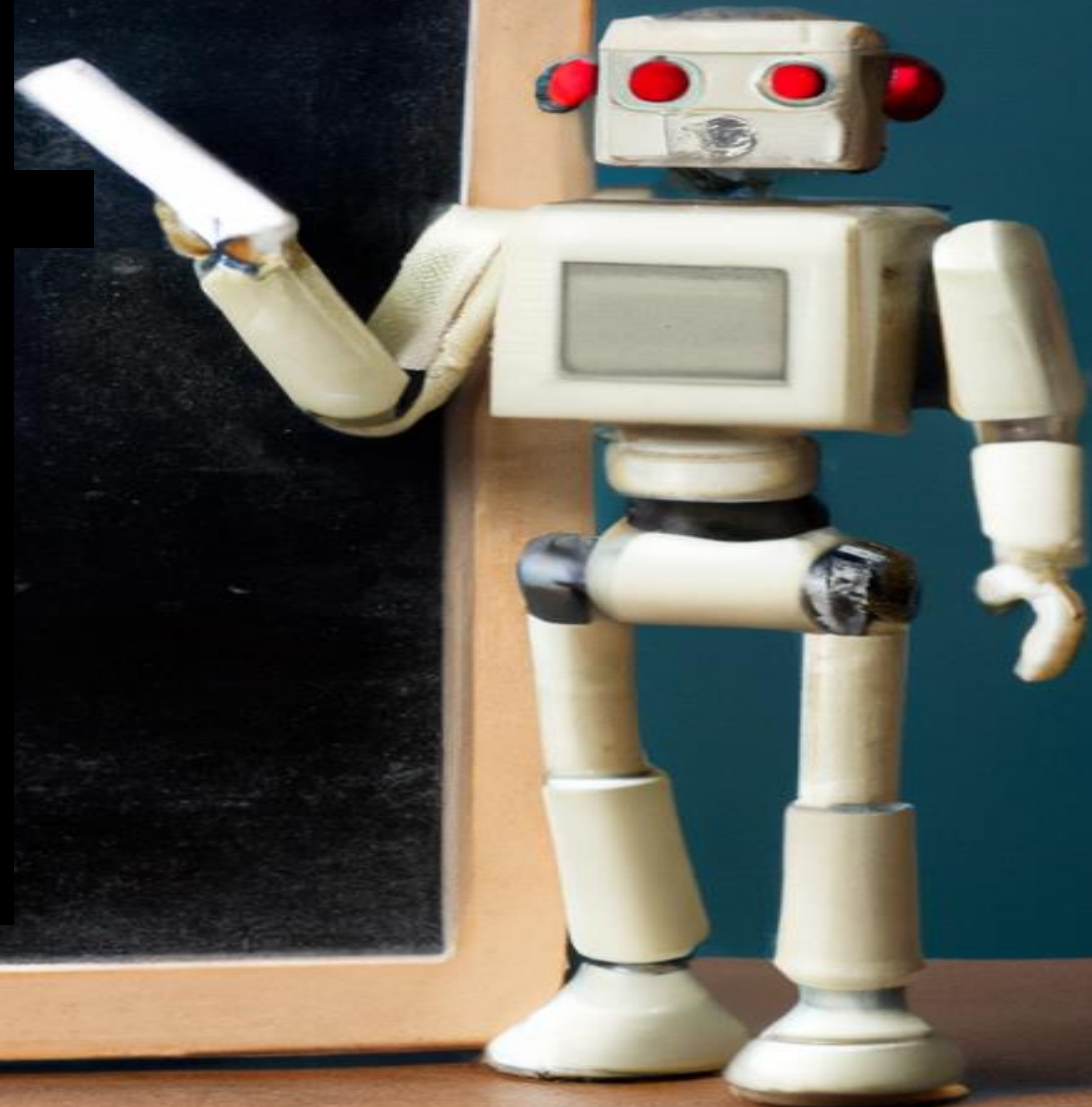
Conception flips the coin.
the coin. Is the coin still

by Conception. So the coin
ch is an odd number. The coin
ter an odd number of flips,
the answer is no.

Q: What home entertainment
equipment requires cable?
Answer Choices: (a)
radio shack (b) substation
(c) television (d) cabinet
A: The answer must

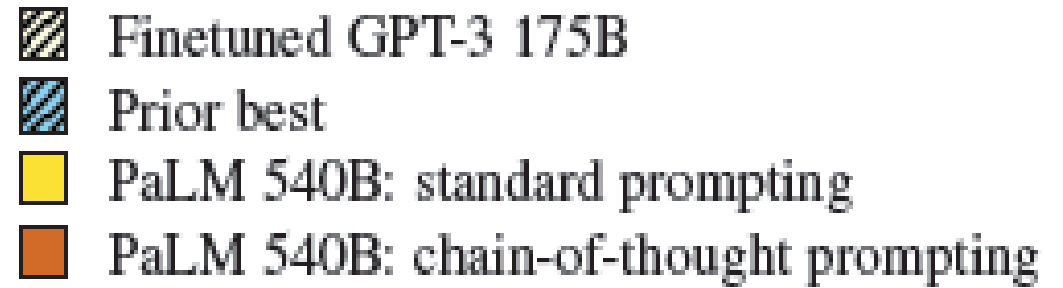
Definitions

- Few Shot Prompt/Exemplars: Providing multiple examples of question/answer pairs before asking the model a similar question
- Chain of Thought (CoT) Prompting: Including a series of intermediate reasoning steps in a few shot prompt
- Standard Prompting: The paper uses this term to describe a few shot prompt that only provides the answers, without the Chain of Thought reasoning of how the answer was generated.
- Ablation Study: Investigates the performance of an AI system by removing certain components to understand the contribution of the component to the overall system



Advantages of Chain of Thought Prompting

- Much better accuracy attained with CoT compared to standard prompting once a model has enough parameters.
- Relatively few annotations are needed compared to a finetuning approach.
- One pretrained model can perform many different tasks by just updating the prompt at inference time rather than expensive finetuning.
- If there are problems with the output, being able to view the model's reasoning facilitates debugging.

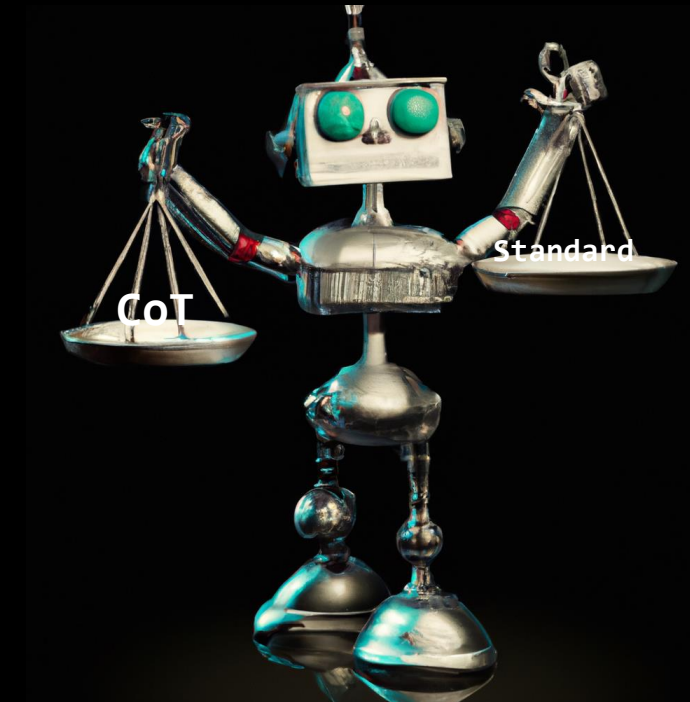


Disadvantages of Chain of Thought Prompting

- Worse performance than standard prompting on small models (< 100B parameters)
- If a large number of inference requests are planned, at some point the cost incurred from additional CoT few shot reasoning tokens prepended to the prompt during each query may be higher than an initial finetuning cost. (my thoughts)
- Causing the model to output additional text than the exact answer increases the chance of non-relevant text being included which could potentially cause parsing issues retrieving the answer

QUESTION: If Ann is 9 years old and her brother is twice her age, how old will her brother be in 3 years?

MODEL ANSWER (CORRECT BUT UNRELATED STATEMENT): Ann is 9 years old. Her brother is twice her age. So her brother is 18 years old. In 3 years, she will be 12. So her brother will be $18 + 3 = 21$ years old. The answer is 21.



Arithmetic Reasoning

Datasets:

GSM8k – 8,000 grade school math questions answerable by a bright middle school student

SVAMP – math word problems with varying structures

MAWPS – online repository of Math Word Problems

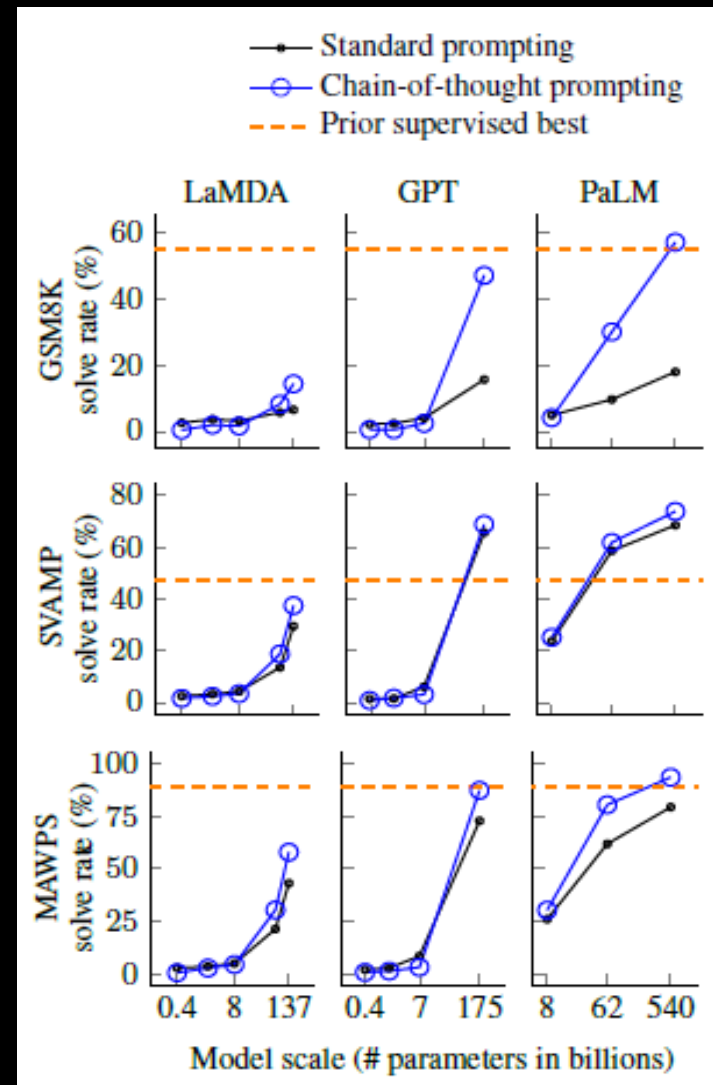
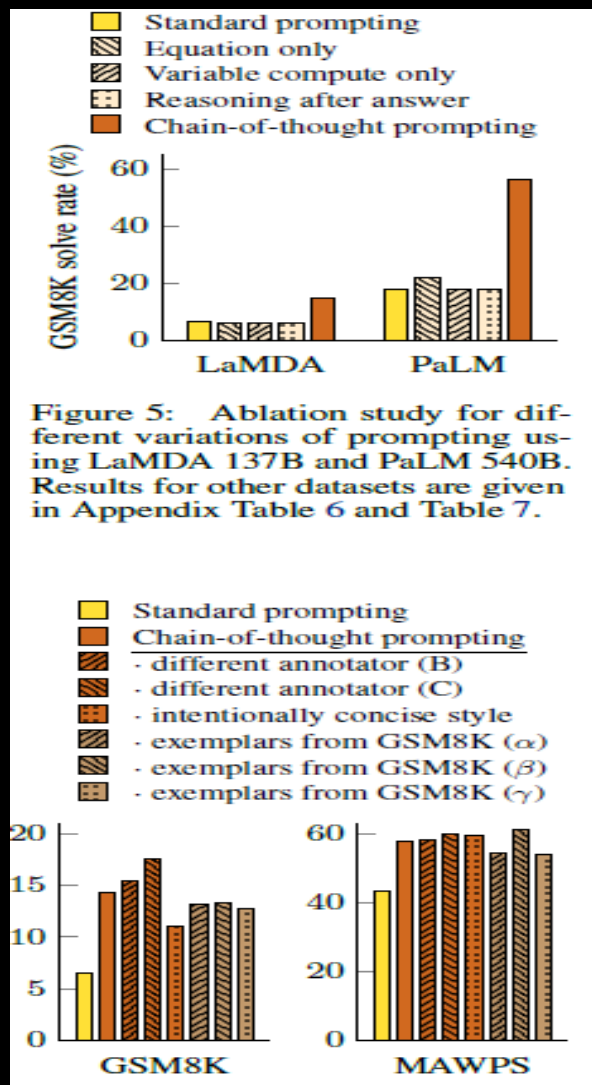
CoT increases the Solve Rate the most for the largest PaLM model while also improving LaMDA and GPT3 results.

Ablation Experiments

- Variable Compute Only
- Equation Only
- Reasoning After Answer

Robustness Checks

- Different Annotators
- Alternate Exemplars
- Intentionally Concise Style



Reproducing results with GPT3 text-curie-001

Standard few shot prompt without chain of thought reasoning:

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

A: The answer is 6.

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

A: The answer is 5.

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

A: The answer is 39.

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

A: The answer is 8.

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

A: The answer is 9.

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

A: The answer is 29.

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

A: The answer is 33.

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

A: The answer is 8.

Q: Diva has 2 milkbones. Jack buys her 2 boxes which each contain 15 milkbones. How many milkbones does Diva currently have?

A: The answer is 30.



Wrong!

Reproducing results with GPT3 text-curie-001

Few shot prompt with chain of thought reasoning:

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

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

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

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

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

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

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

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

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

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

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

A: Michael started with 58 golf balls. 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.

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

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

Q: Diva has 2 milkbones. Jack buys her 2 boxes which each contain 15 milkbones. How many milkbones does Diva currently have?

A: Diva started with 2 milkbones. After Jack bought her 2 boxes which each contain 15 milkbones, she has $5 + 15 = 30$ milkbones. $30 + 2 = 32$ milkbones. The answer is 32.



Correct answer, incorrect reasoning

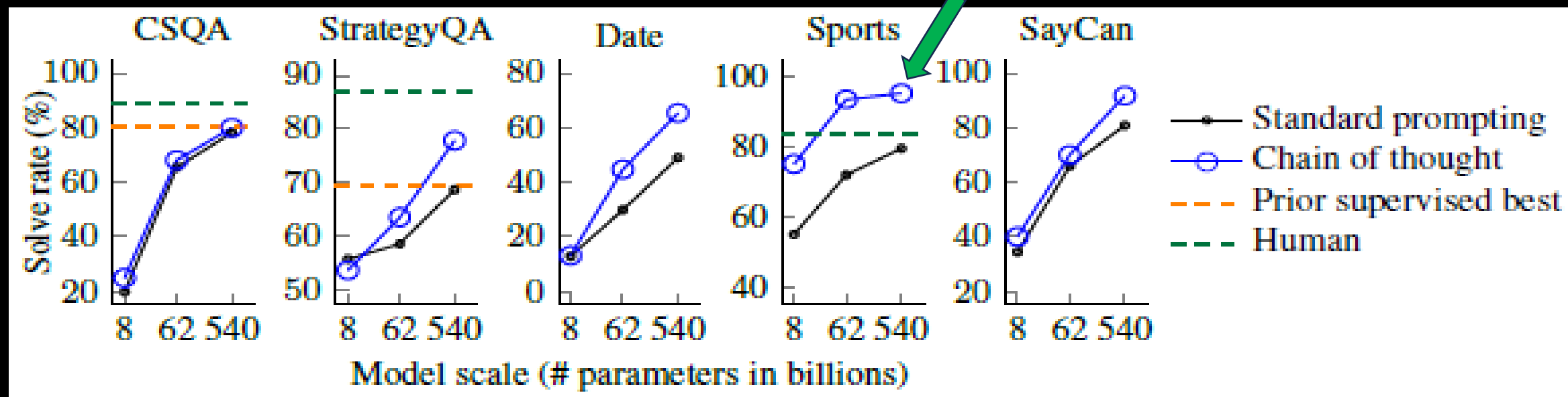
Common Sense Reasoning

Tasks

- CSQA: Common sense questions about the world
- StrategyQA: Questions requiring a multi step strategy to answer
- Date: Inferring a date from a given context
- Sports: Determine if a statement about a sport is plausible
- SayCan: Natural language to make a robot perform a task



CoT prompting outperformed an unaided human sports enthusiast on sports understanding (95.4% vs 84%)!



Reproducing Common Sense results with GPT3 text-babbage-001

Standard few shot prompt without chain of thought reasoning:

Q: What do people use to absorb extra ink from a fountain pen? Answer Choices: (a) shirt pocket (b) calligrapher's hand (c) inkwell (d) desk drawer (e) blotter
A: The answer is (e).

Q: What home entertainment equipment requires cable?
Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet
A: The answer is (c).

Q: The fox walked from the city into the forest, what was it looking for? Answer Choices: (a) pretty flowers (b) hen house (c) natural habitat (d) storybook
A: The answer is (b).

Q: Sammy wanted to go to where the people were. Where might he go? Answer Choices: (a) populated areas (b) race track (c) desert (d) apartment (e) roadblock
A: The answer is (a).

Q: Where do you put your grapes just before checking out? Answer Choices: (a) mouth (b) grocery cart (c)super market (d) fruit basket (e) fruit market
A: The answer is (b).

Q: Google Maps and other highway and street GPS services have replaced what? Answer Choices: (a) united states (b) mexico (c) countryside (d) atlas
A: The answer is (d).

Q: Before getting a divorce, what did the wife feel who was doing all the work? Answer Choices: (a) harder (b) anguish (c) bitterness (d) tears (e) sadness
A: The answer is (c).

Q: I have accidentally left my debit card at a restaurant and nobody can find it, what should I do? Answer Choices: (a) eat leftovers (b) put on sunscreen (c) call bank (d) take a bath
A: The answer is (b).

 **Wrong!**

Reproducing Common Sense results with GPT3 text-babbage-001

Few shot prompt with chain of thought reasoning:

So the answer is (e).

Q: What home entertainment equipment requires cable?

Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet

A: The answer must require cable. Of the above choices, only television requires cable. So the answer is (c).

Q: The fox walked from the city into the forest, what was it looking for? Answer Choices: (a) pretty flowers (b)

hen house (c) natural habitat (d) storybook

A: The answer must be something in the forest. Of the above choices, only natural habitat is in the forest. So the answer is (b).

Q: Sammy wanted to go to where the people were. Where might he go? Answer Choices: (a) populated areas

(b) race track (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Of the above choices, only populated areas have a lot of people. So the answer is (a).

Q: Where do you put your grapes just before checking out? Answer Choices: (a) mouth (b) grocery cart (c)super market (d) fruit basket (e) fruit market

A: The answer should be the place where grocery items are placed before checking out. Of the above choices, grocery cart makes the most sense for holding grocery items. So the answer is (b).

Q: Google Maps and other highway and street GPS services have replaced what? Answer Choices: (a) united states (b) mexico (c) countryside (d) atlas

A: The answer must be something that used to do what Google Maps and GPS services do, which is to give directions. Of the above choices, only atlases are used to give directions. So the answer is (d).

Q: Before getting a divorce, what did the wife feel who was doing all the work? Answer Choices: (a) harder (b) anguish (c) bitterness (d) tears (e) sadness

A: The answer should be the feeling of someone getting divorced who was doing all the work. Of the above choices, the closest feeling is bitterness. So the answer is (c).

Q: I have accidentally left my debit card at a restaurant and nobody can find it, what should I do? Answer Choices: (a) eat leftovers (b) put on sunscreen (c) call bank (d) take a bath

A: The answer is (c).

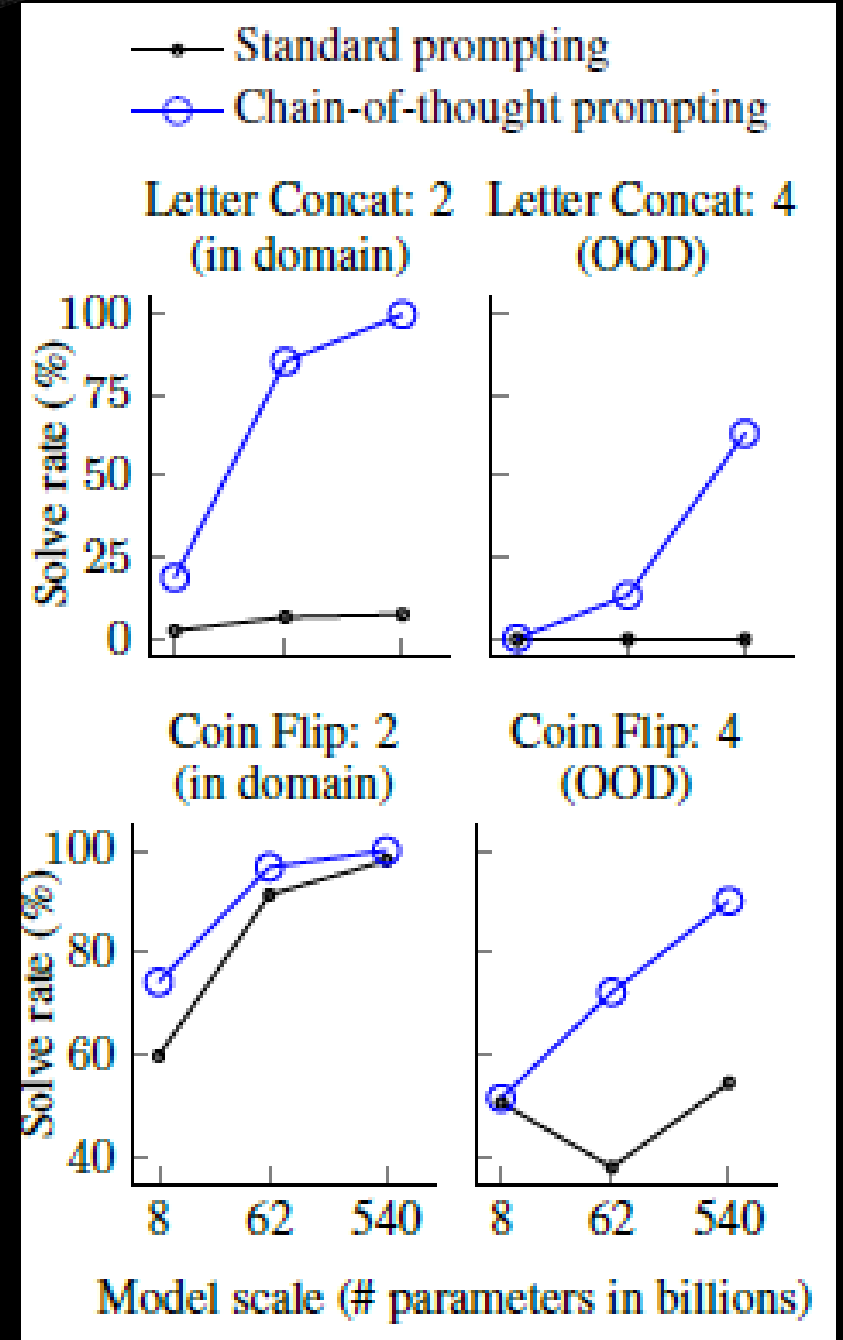
 **Correct, no reasoning outputted**

Symbolic Reasoning

Task #1 - Coin Flip heads or tails?

Task #2 - Name Last letter concatenation "Jack Youngblood" = "kd"

- Performed less well on "out of domain" tasks where the number of coin flips or words differed from the training data. However, in all cases including Chain of Thought prompting met or exceeded the solve rate from Standard Prompting.



Reproducing Symbolic Reasoning results with Cohere

Standard Few shot prompt without chain of thought reasoning:

INPUT

Q: Take the last letters of the words in "Sergey Brin" and concatenate them.

A: The answer is yn.

Q: Take the last letters of the words in "Bill Gates" and concatenate them.

A: The answer is ls.

Q: Take the last letters of the words in "Jeremy Howard" and concatenate them.

A:

OUTPUT

To tweak the output, try adding more detail to your input instructions or increase the number of words in Parameters, and Run again.

The answer is d.

 Wrong!

Reproducing Symbolic Reasoning results with Cohere

Few shot prompt with chain of thought reasoning:

INPUT

A: The last letter of "Sergey" is "y". The last letter of "Brin" is "n". Concatenating them is "yn". The answer is yn.

Q: Take the last letters of the words in "Bill Gates" and concatenate them.

A: The last letter of "Bill" is "l". The last letter of "Gates" is "s". Concatenating them is "ls". The answer is ls.

Q: Take the last letters of the words in "Jeremy Howard" and concatenate them.

A:

OUTPUT

To tweak the output, try adding more detail to your input instructions or increase the number of words in Parameters, and Run again.

The last letter of "Jeremy" is "y". The last letter of "Howard" is "d". Concatenating them is "yd". The answer is yd.

 **Correct!**

Large Language Models are Zero-Shot Reasoners

arXiv:2205.11916

Google Research, Google Brain, University of Tokyo

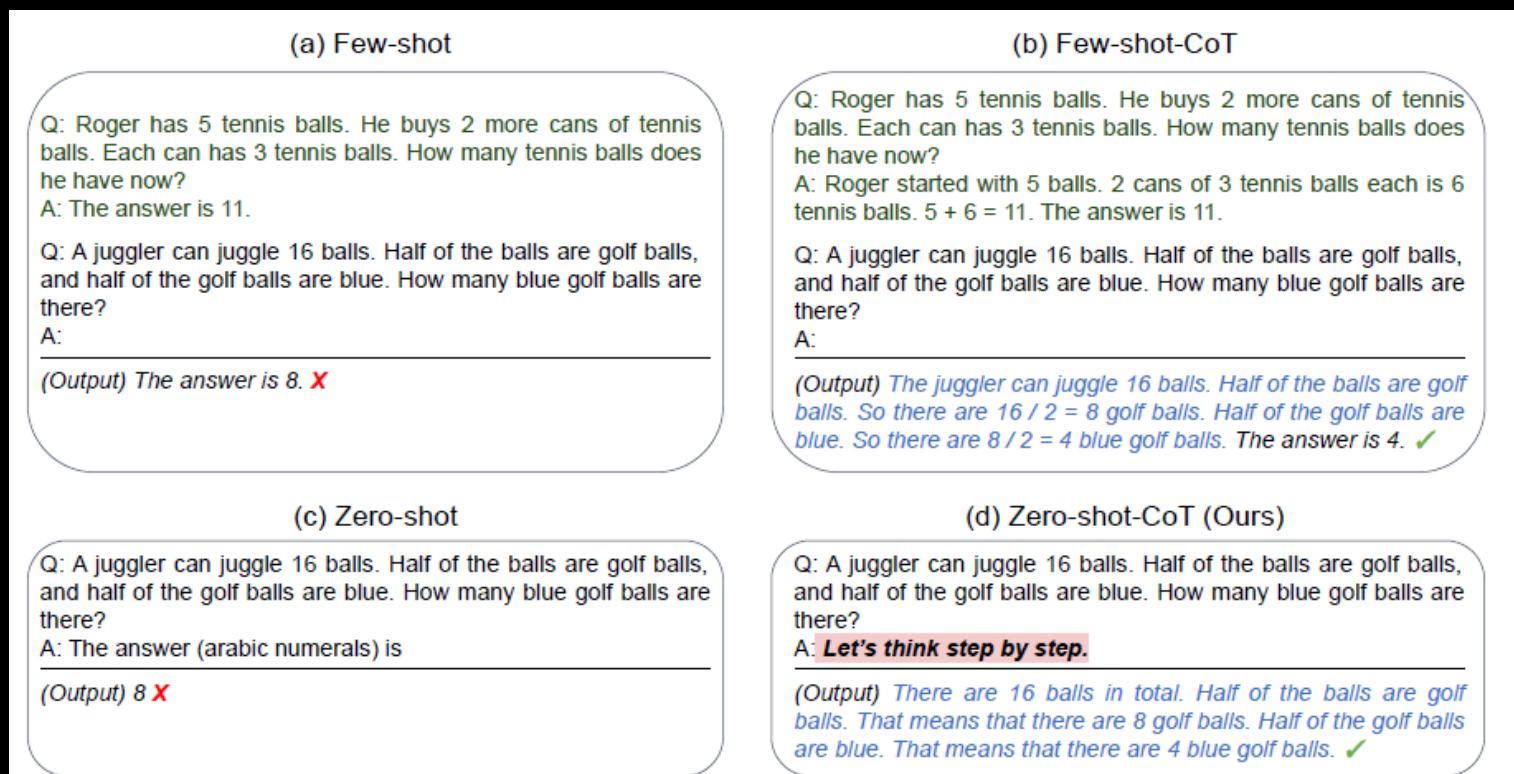


Figure 1: Example inputs and outputs of GPT-3 with (a) standard Few-shot ([Brown et al., 2020]), (b) Few-shot-CoT ([Wei et al., 2022]), (c) standard Zero-shot, and (d) ours (Zero-shot-CoT). Similar to Few-shot-CoT, Zero-shot-CoT facilitates multi-step reasoning (blue text) and reach correct answer where standard prompting fails. Unlike Few-shot-CoT using step-by-step reasoning examples **per task**, ours does not need any examples and just uses the same prompt “Let’s think step by step” *across all tasks* (arithmetic, symbolic, commonsense, and other logical reasoning tasks).

Reproducing results with GPT3 text-davinci-003

Standard zero shot prompt without zero shot chain of thought reasoning:

Playground

Q: Diva has 2 milkbones. Jack buys her 2 boxes which each contain 15 milkbones. How many milkbones does Diva currently have?

A: Diva currently has 17 milkbones.


Wrong!

Zero shot chain of thought reasoning:

Playground Your presets

Q: Diva has 2 milkbones. Jack buys her 2 boxes which each contain 15 milkbones. How many milkbones does Diva currently have?

A: Let's think step by step. Diva has 2 milkbones. Jack buys her 2 boxes, which each contain 15 milkbones. Therefore, Diva now has $2 + (15 + 15) = 32$ milkbones.


Correct!

Table 2: Comparison with baseline methods using accuracies on MultiArith and GSM8K. text-davinci-002 is used as the model if not specified. We used the same 8 examples as described in [Wei et al., 2022] for Few-shot and Few-shot-CoT settings. (*1) To verify the variance of changing examples, we report two results for 4-shot-cot by splitting the eight examples into two groups. (*2) We insert “Let’s think step by step.” at the beginning of answer part of each exemplars for Few-shot-CoT to test performance gains. Further experiment results with PaLM are found at Appendix D

	MultiArith	GSM8K
Zero-Shot	17.7	10.4
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	78.7	40.7
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	92.8	51.5
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55
PaLM 540B: Zero-Shot	25.5	12.5
PaLM 540B: Zero-Shot-CoT	66.1	43.0
PaLM 540B: Zero-Shot-CoT + self consistency	89.0	70.1
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4

Conclusion

Chain of Thought prompting helps the most under the following scenarios:

- A language model has a large number of parameters
- The task requires multistep reasoning
- The scaling curve is relatively flat (adding additional parameters doesn't significantly improve the results).

