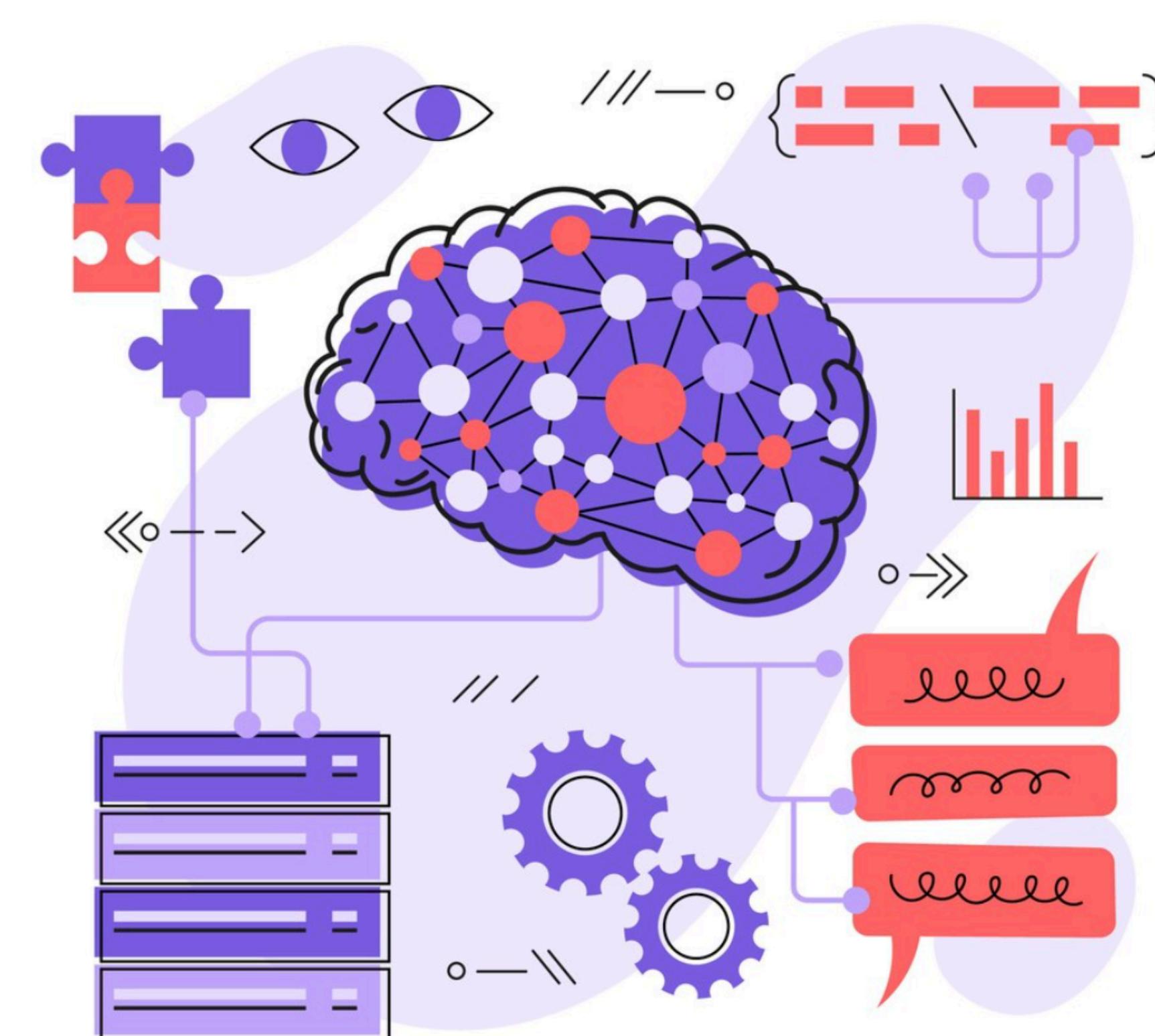


Large Language Models

Chain of Thought Prompting

M. Soleymani
Sharif University of Technology
Fall 2023

Chain of Thought Prompting for Large Language Model Reasoning



Hard Language Tasks: Reasoning

Tasks that require multiple steps of reasoning to solve

Reasoning Problems

Arithmetic Reasoning (AR)

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

Answer: The answer is **5**.

Symbolic Reasoning (SR)

Question: Take the last letters of the words in "Elon Musk" and concatenate them

Answer: The answer is **nk**.

Commonsense Reasoning (CR)

Question: What home entertainment equipment requires cable? Answer
Choices: (a) radio shack (b) substation (c) television (d) cabinet

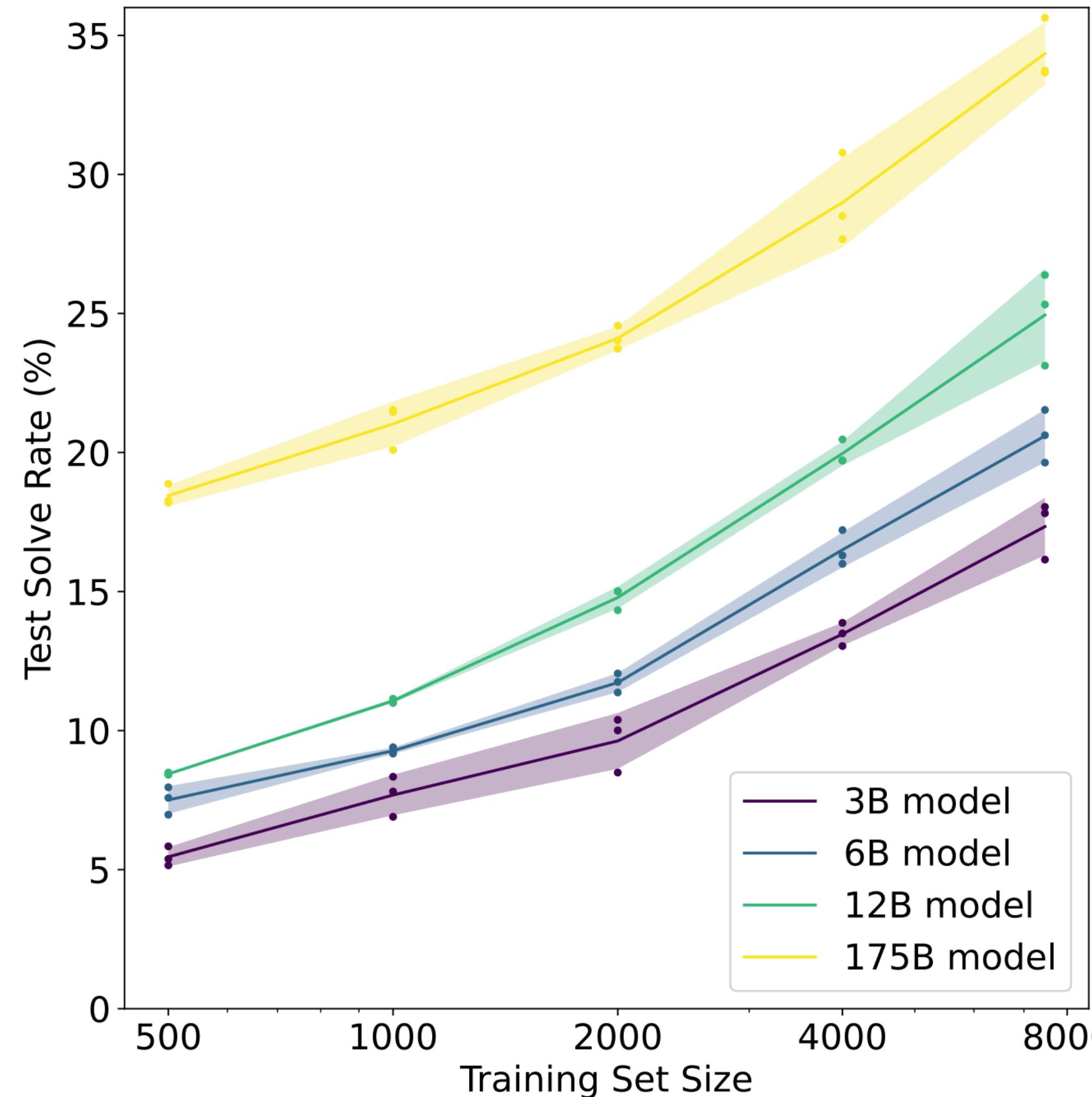
Answer: The answer is **(c)**.

Reasoning

- Multi-step reasoning is often seen as a weakness in NLP models
- There is former research on reasoning in small language models through fully supervised finetuning on specific datasets. However,
 - Creating a dataset containing explicit reasoning can be difficult and time-consuming
 - training on a specific dataset limits application to a specific domain
- Reasoning ability may emerge in language models at a certain scale, such as models with over 100 billion parameters (Wei et al., TMLR 2022)

Reasoning Problems

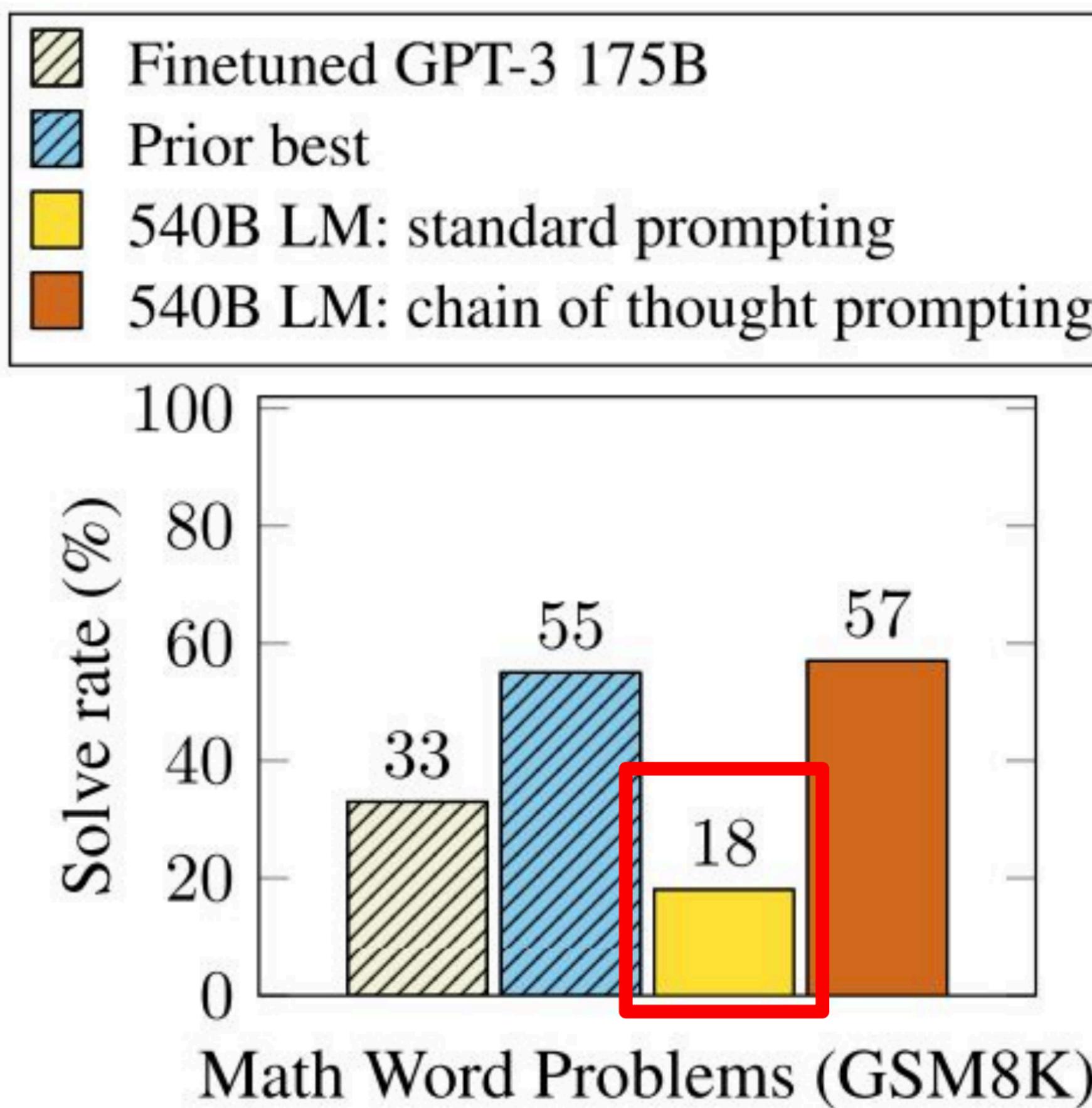
Fine-tune GPT-3 on GSM8K (arithmetic):



Conjecture: to achieve > 80%, needs 100 times more fine-tuning data for 175B model

Reasoning Problems

GSM8K (arithmetic):



Few-shot standard prompting with even larger model (PaLM 540B) also does not work well.

Reasoning Problems

Scaling up language model size does not **efficiently** achieve high performances,

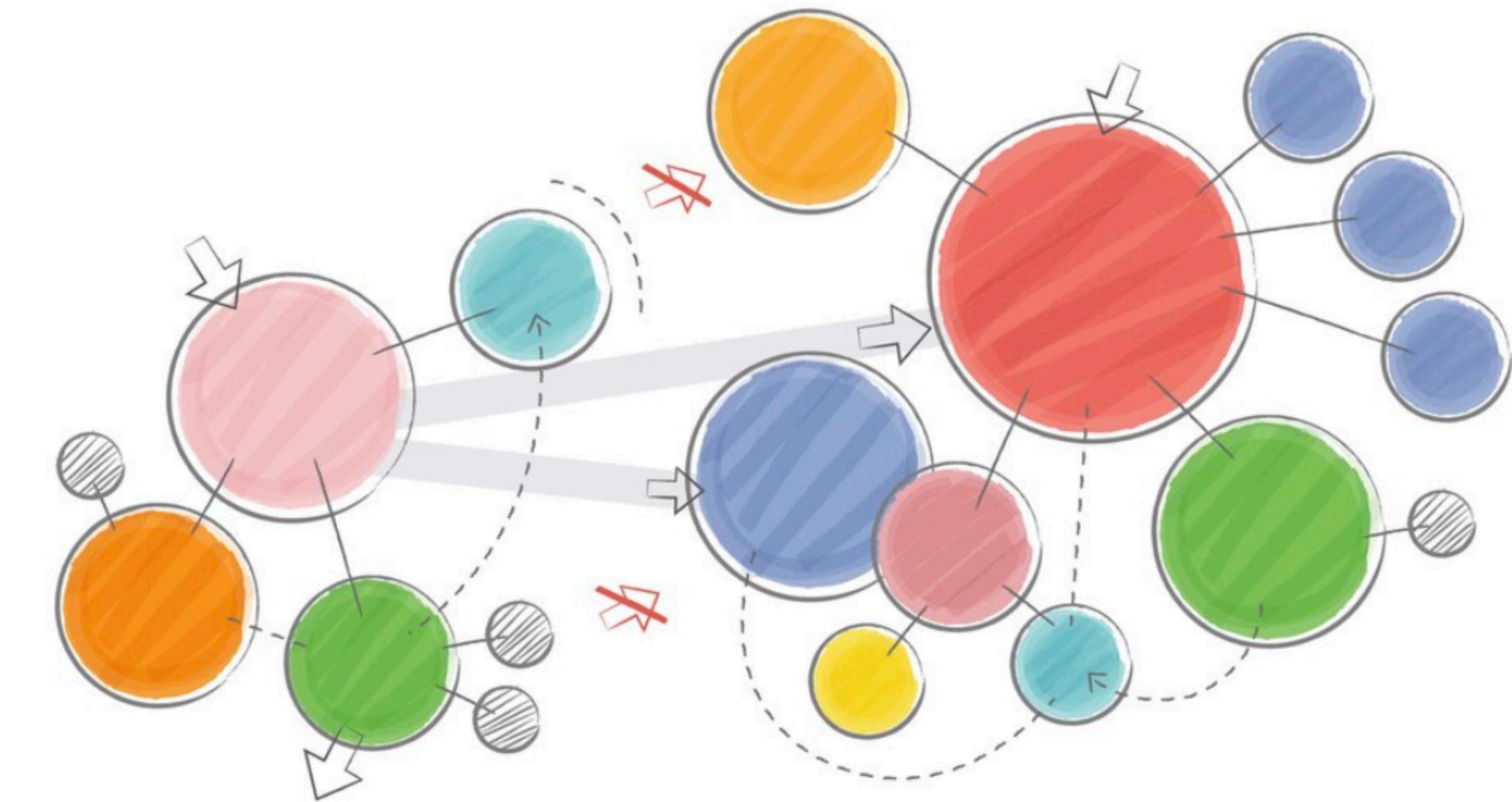
for Arithmetic Reasoning (AR), CommonSense Reasoning (CR) and Symbolic Reasoning (SR) tasks

Reasoning Problems

Scaling up language model size does not **efficiently** achieve high performances, for Arithmetic Reasoning (AR), CommonSense Reasoning (CR) and Symbolic Reasoning (SR) tasks

Proposed solution: **Chain of Thought (CoT) prompting**

Chain of Thought Prompting



Chain of Thought (CoT)

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

Few-Shot CoT

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma

Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Google Research, Brain Team
`{jasonwei,dennyzhou}@google.com`

Chain of Thought (CoT)

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

Few-Shot CoT

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma
Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Google Research, Brain Team
{jasonwei,dennyyzhou}@google.com

Both papers
will appear in
NeurIPS'22!

Large Language Models are Zero-Shot Reasoners

Zero-Shot CoT

Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
Google Research*

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Chain of Thought (CoT)

Definition:

A chain of thought is a series of intermediate natural language reasoning steps that lead to the final output.

Chain of Thought (CoT)

Definition:

A chain of thought is a series of intermediate natural language reasoning steps that lead to the final output.

⟨input, output⟩ demonstrations are replaced with **⟨input, chain of thought, output⟩**

Compositionality of Language

- Compositionality of the languages
 - Compositional Out-of-Distribution generalization: ability to understand novel composition of known concepts
- Problem decomposition can help
 - Decompose multi-step reasoning into intermediate steps

Chain of Thought (CoT)

Definition:

A chain of thought is a series of intermediate natural language reasoning steps that lead to the final output.

use <input, **intermediate results**, output> triples

Benefits:

- Decomposition -> easier intermediate problems
- Interpretable
- More general than neural symbolic computing
- Leveraging prompting of LLM

Chain of Thought (CoT)

(Wei et al., 2022)

(a) Few-shot

Question: 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?

Answer: The answer is 11.

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Answer:

(Output) The answer is 8. ✗

Step-by-step Answer

(b) Few-shot-CoT

Question: 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?

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

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Answer:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

Chain of Thought (CoT)

Zero-shot

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Answer: The answer (arabic numerals) is

(Output) 8 ✗

Zero-shot-CoT

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

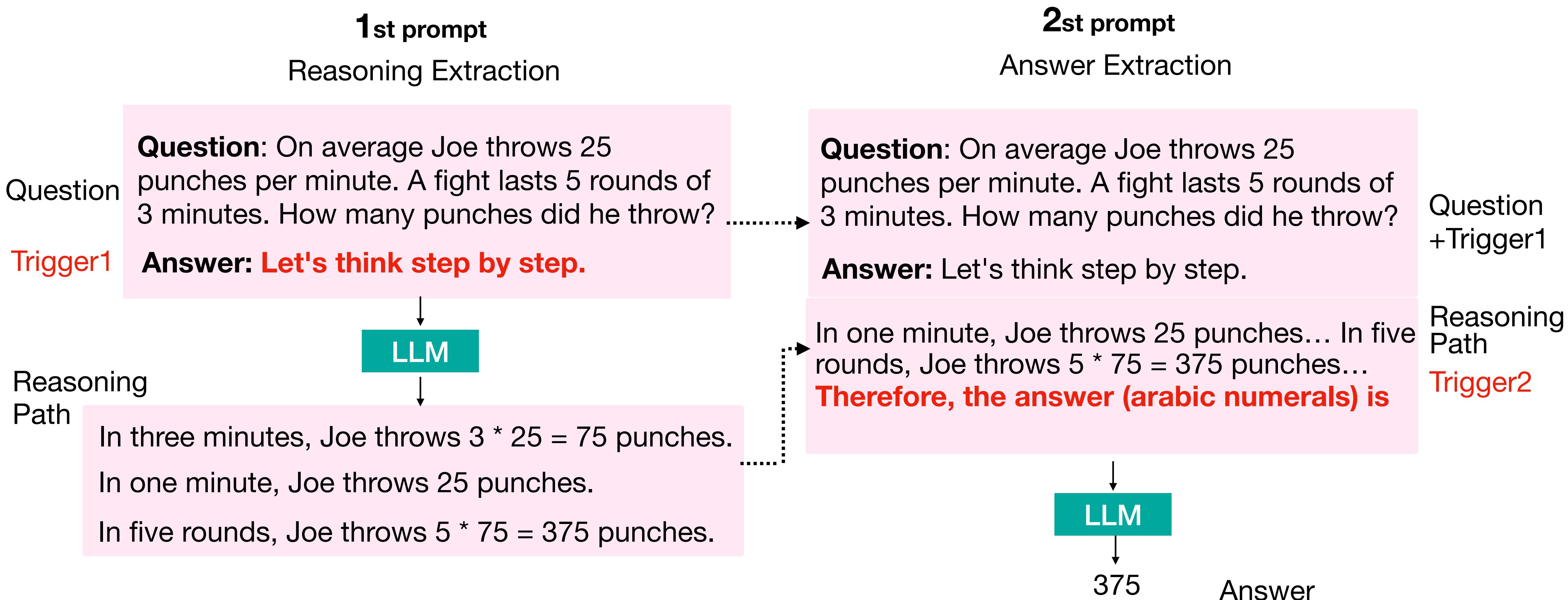
Answer: Let's think step by step.

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.* ✓

Two-stage Prompting
Step-by-step Answer

Zero-Shot Chain of Thought (CoT)

For zero-shot CoT, a **two-stage** prompting is applied:



Experiments



Models

Pre-trained LLMs:

- **Instruct GPT-3** (ada 350M, babbage 1.3B, curie 6.7B, and davinci 175B) (Ouyang et al., 2022)
- **PaLM** (8B, 62B, 540B) (Chowdhery et al., 2022)
- **LaMDA** (422M, 2B, 8B, 68B, 137B) (Thoppilan et al., 2022)
 - Dialogue-oriented LM.
 - Fine-tuned on human-annotated data.

Models

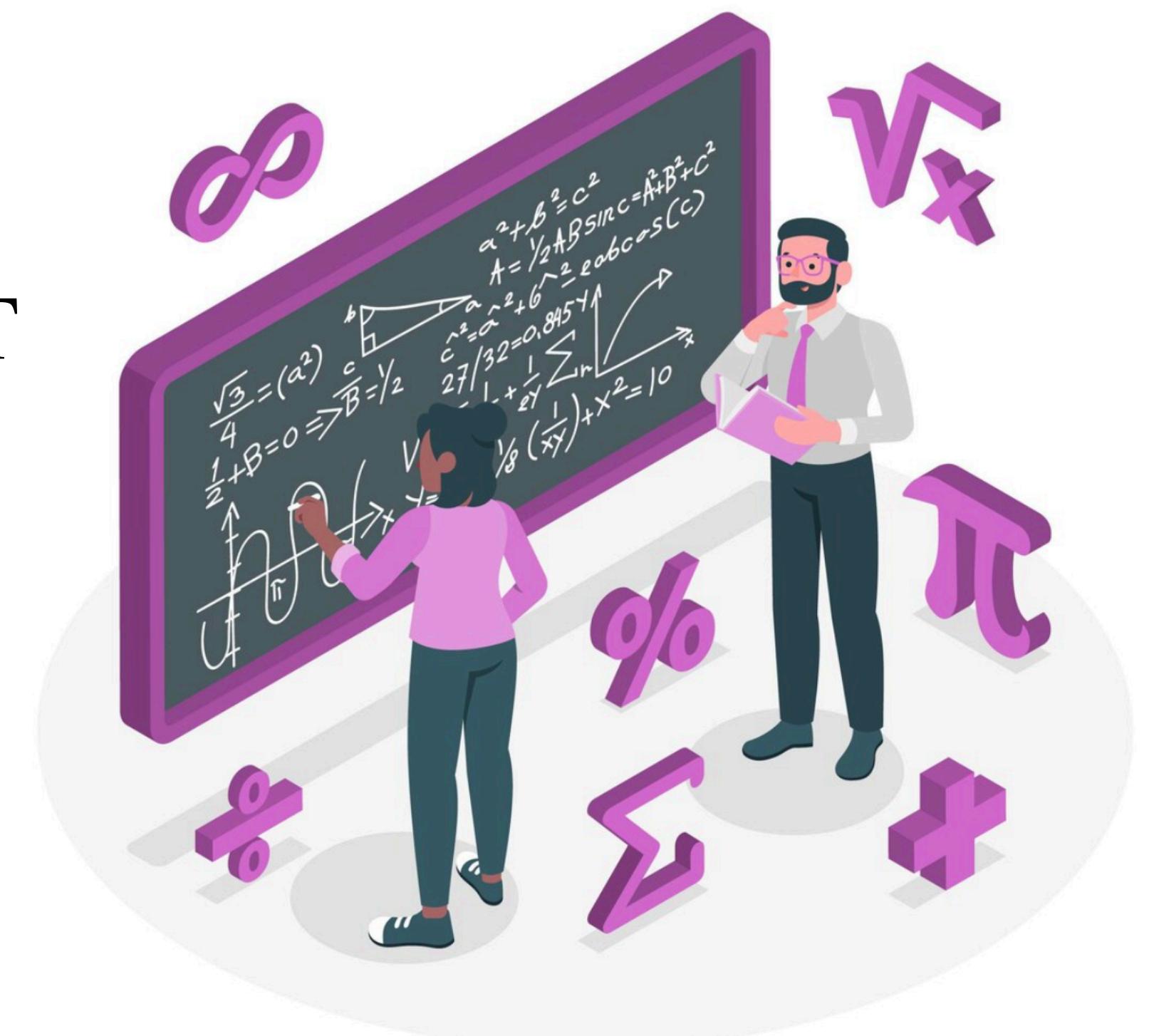
Pre-trained LLMs:

- **Instruct GPT-3** (ada 350M, babbage 1.3B, curie 6.7B, and davinci 175B) (Ouyang et al., 2022)
- **PaLM** (8B, 62B, 540B) (Chowdhery et al., 2022)
- **LaMDA** (422M, 2B, 8B, 68B, 137B) (Thoppilan et al., 2022)
- **GPT-3** (ada 350M, babbage 1.3B, curie 6.7B, davinci 175B)
- **GPT-2** (1.5B)
- **GPT-Neo** (2.7B), **GPT-J** (6B), **T0** (11B) (Sanh et al., 2022), **OPT** (13B) (Zhang et al., 2022)

Experiments

Arithmetic Reasoning

Prompting setups: zero-shot, few-shot, few-shot CoT



Free Response - Few-Shot CoT Prompt Exemplar

Question: 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.

Free Response - Few-Shot CoT Prompt Exemplar

Free Response

Question: 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.

Free Response

Question: 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 $5 \times 3 = 15$ dollars. So she has $23 - 15$ dollars left. $23 - 15$ is 8. The answer is 8.

You can have one or more equations.

Equations can be incomplete and combined math with words.

Free Response - Few-Shot CoT Prompt Exemplar

Free Response

Question: 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.

- Manually composed 8 exemplars
- All contains equations with flexible formats
- Benchmarked on:
 - **GSM8K** (Cobbe et al. 2021)
 - **SVAMP** (Patel et al., 2021)
 - **MAWPS** (Koncel-Kedziorski et al., 2016)

Multiple Choice - Few-Shot CoT Prompt Exemplar

Multiple Choice

Question: 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 $20 \text{ km/hr} * 2.5 \text{ hrs} = 50 \text{ km}$. The answer is (e).

GSM8K (Cobbe et al. 2021)

Multiple Choice

Question: 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).

The exemplars have various formats

Multiple Choice - Few-Shot CoT Prompt Exemplar

Multiple Choice

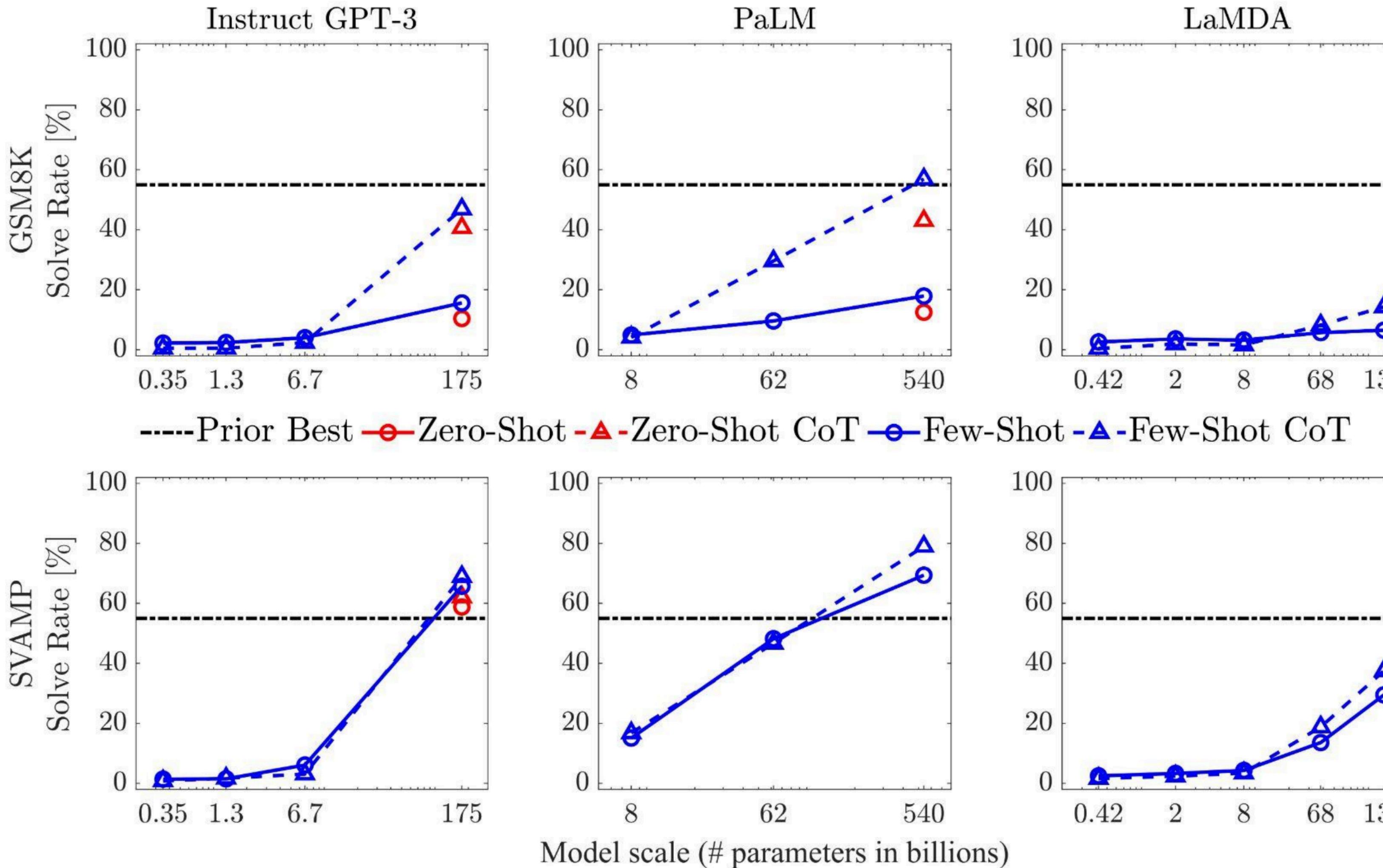
Question: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance? Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km

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

GSM8K (Cobbe et al. 2021)

- 4 exemplars, whose questions, intermediate reasoning, and answers are from AQuA-RAT's **training set**
- Exemplars have flexible formats
- Benchmarked on **AQuA-RAT** (Ling et al., 2017)

Arithmetic Reasoning - Results



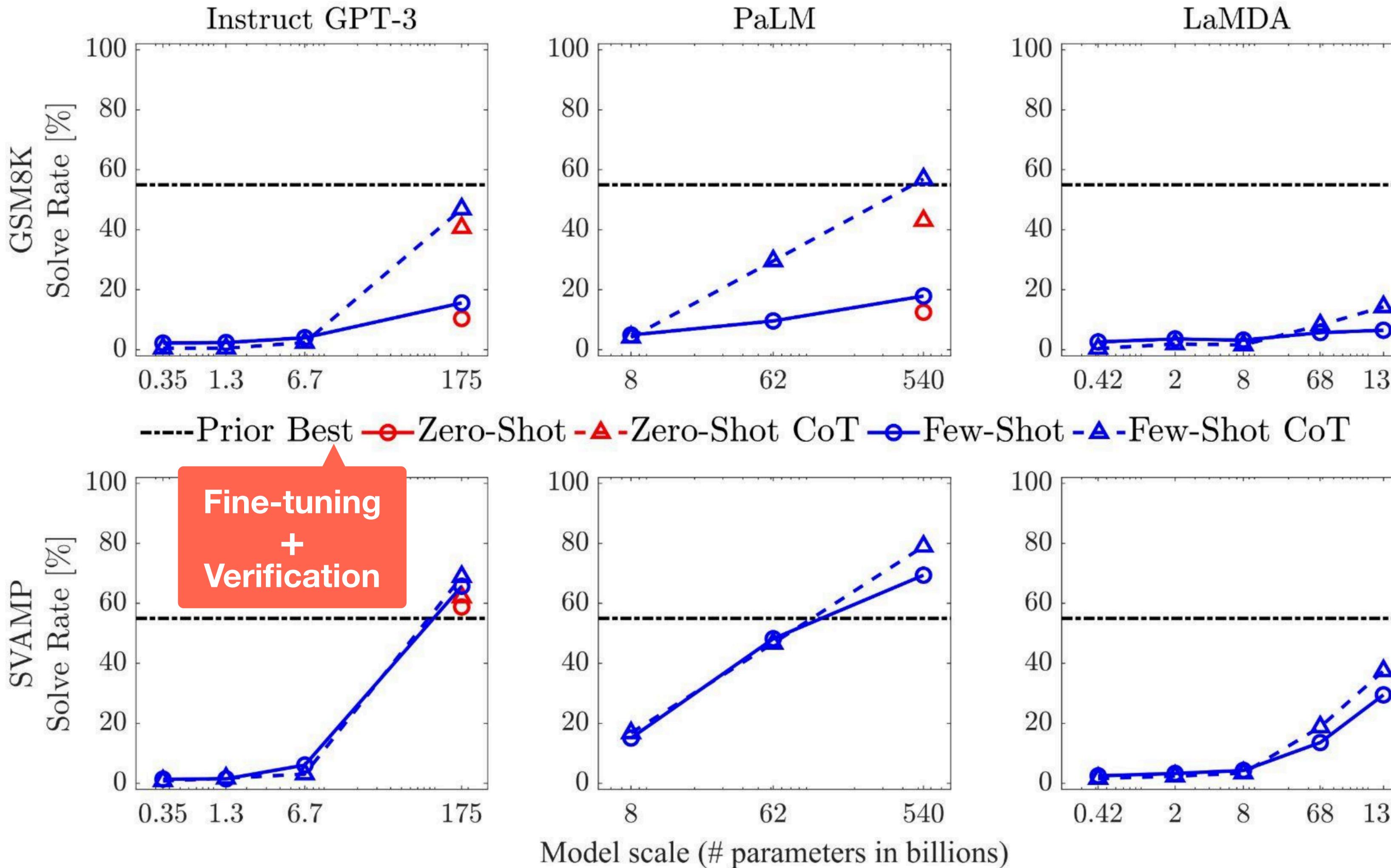
GSM8K

Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did he make?

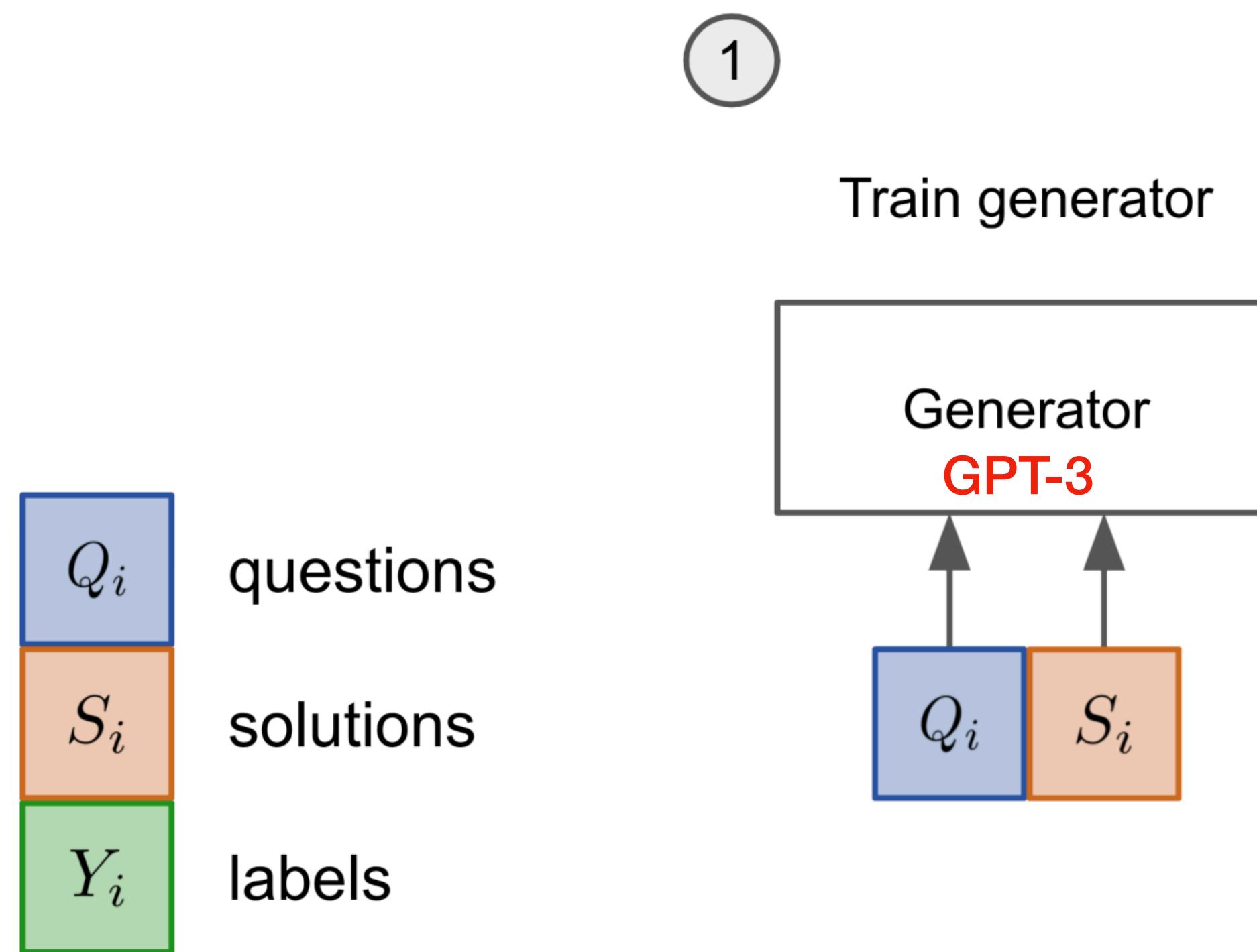
SVAMP

Each pack of dvds costs 76 dollars. If there is a discount of 25 dollars on each pack. How much do you have to pay to buy each pack?

Arithmetic Reasoning - Results

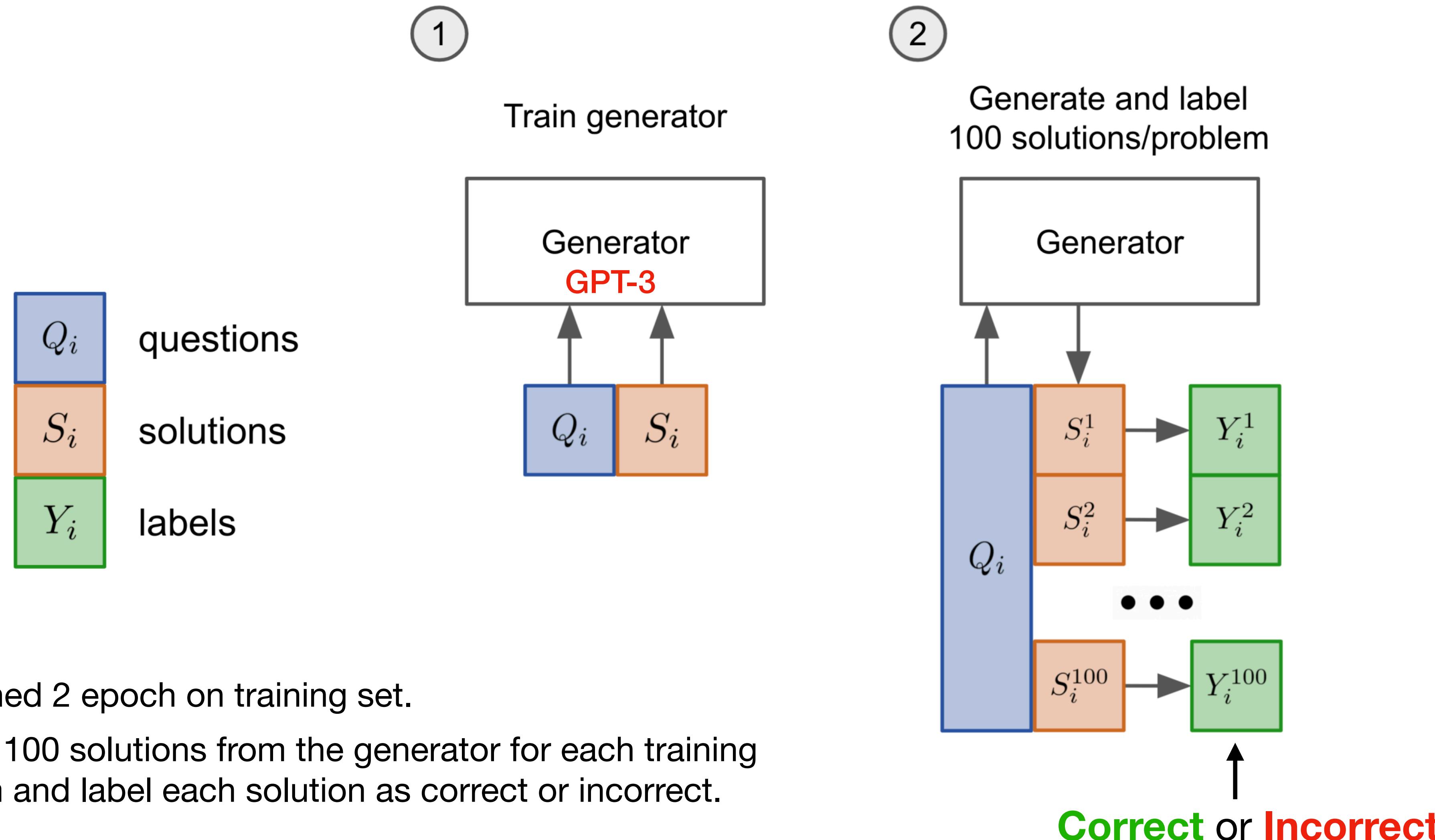


Prior Best – Fine-tuning + Verification



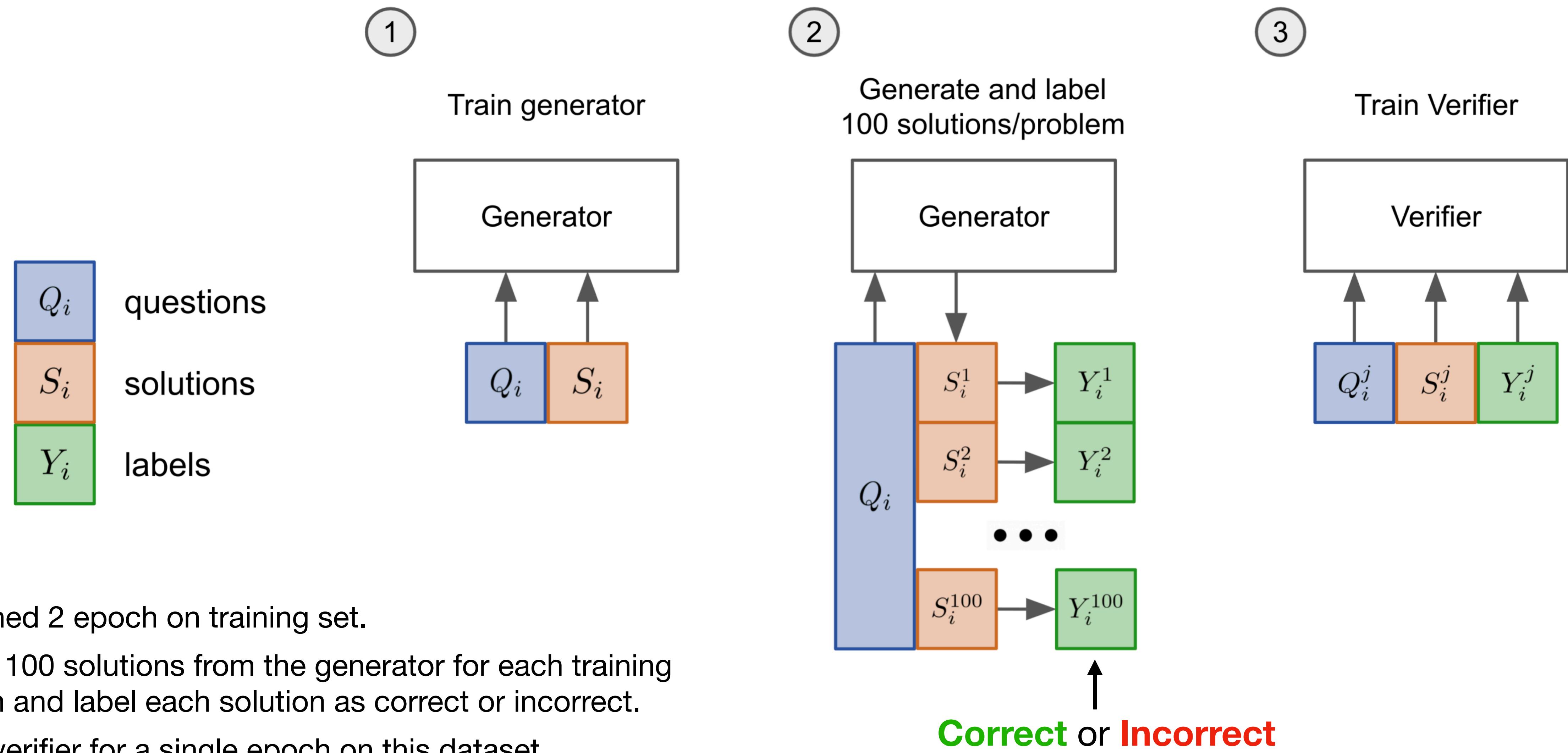
1. Fine-tuned 2 epoch on training set.

Prior Best – Fine-tuning + Verification



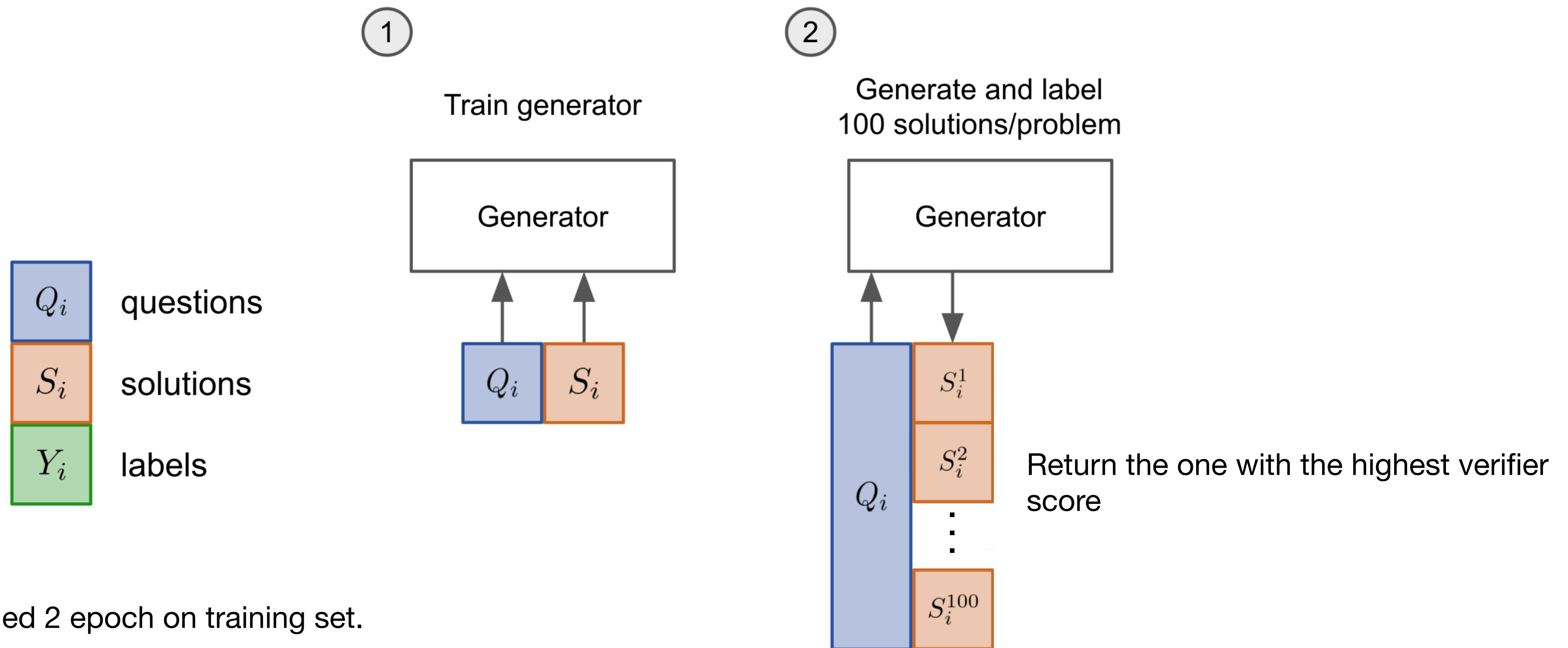
1. Fine-tuned 2 epoch on training set.
2. Sample 100 solutions from the generator for each training problem and label each solution as correct or incorrect.

Prior Best – Fine-tuning + Verification

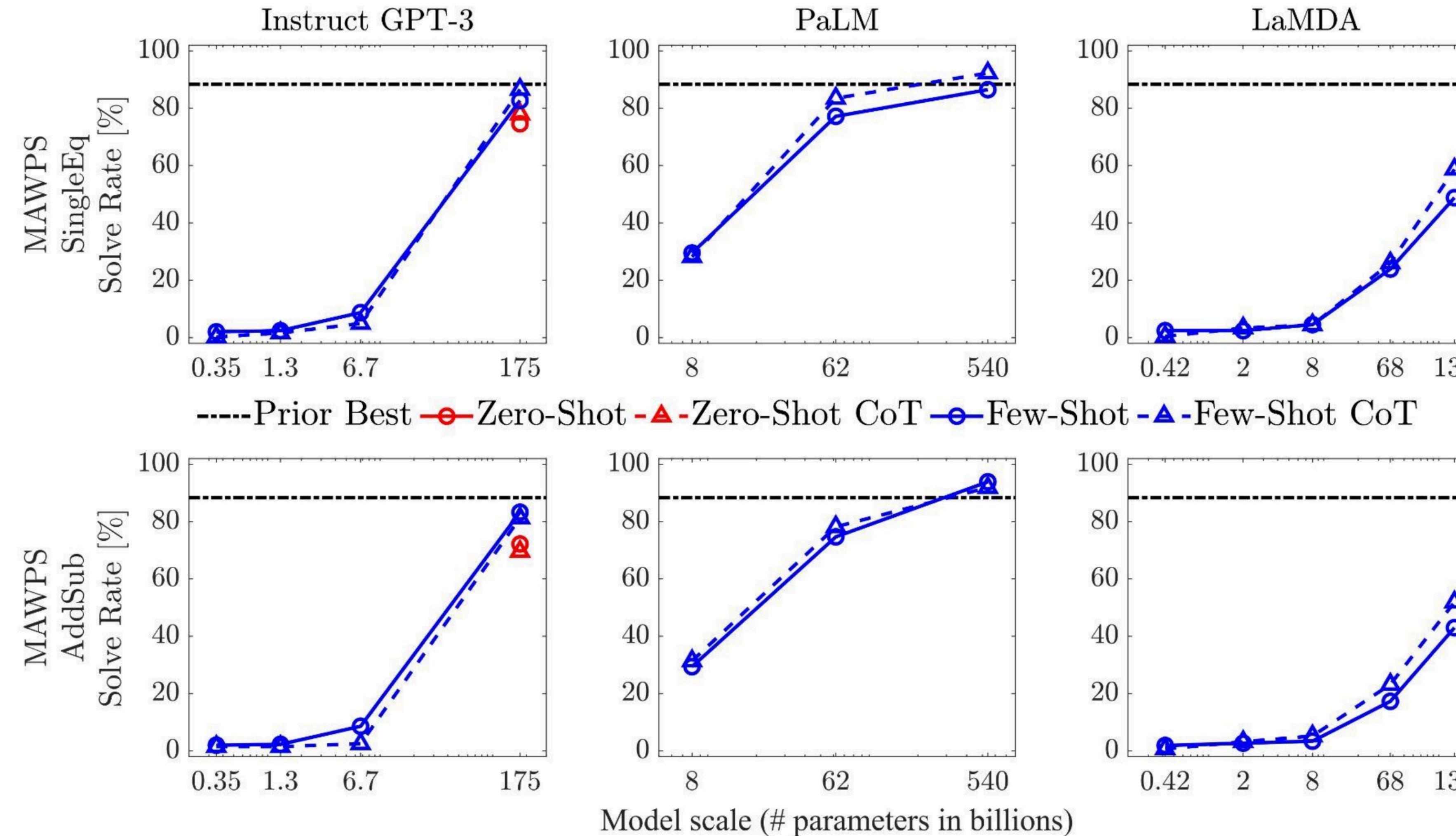


1. Fine-tuned 2 epoch on training set.
2. Sample 100 solutions from the generator for each training problem and label each solution as correct or incorrect.
3. Train a verifier for a single epoch on this dataset.

Prior Best – Fine-tuning + Verification

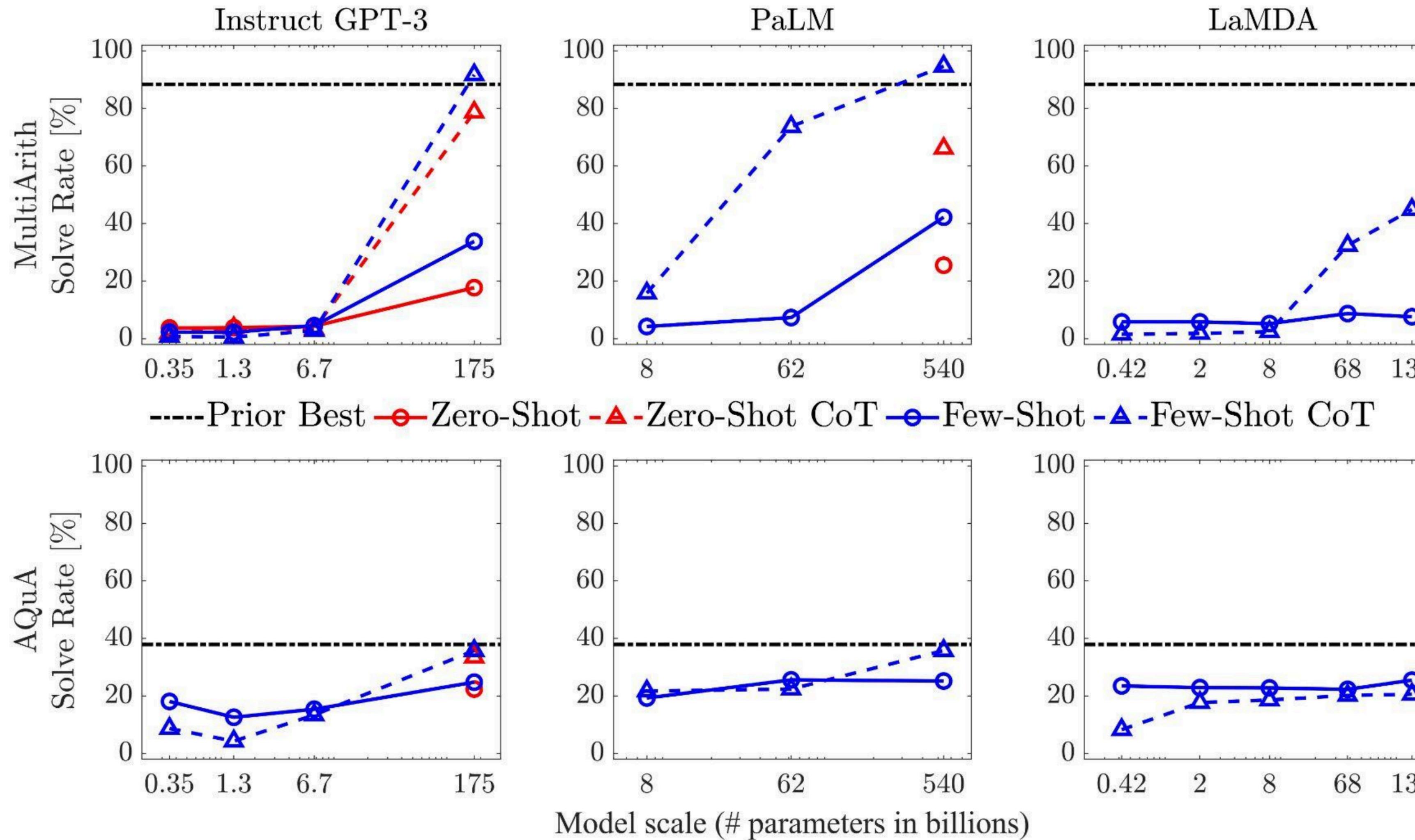


Arithmetic Reasoning - Results



For the easiest subset of MAWPS which only requires a single step to solve, performance improvements were either negative or very small

Arithmetic Reasoning - Results



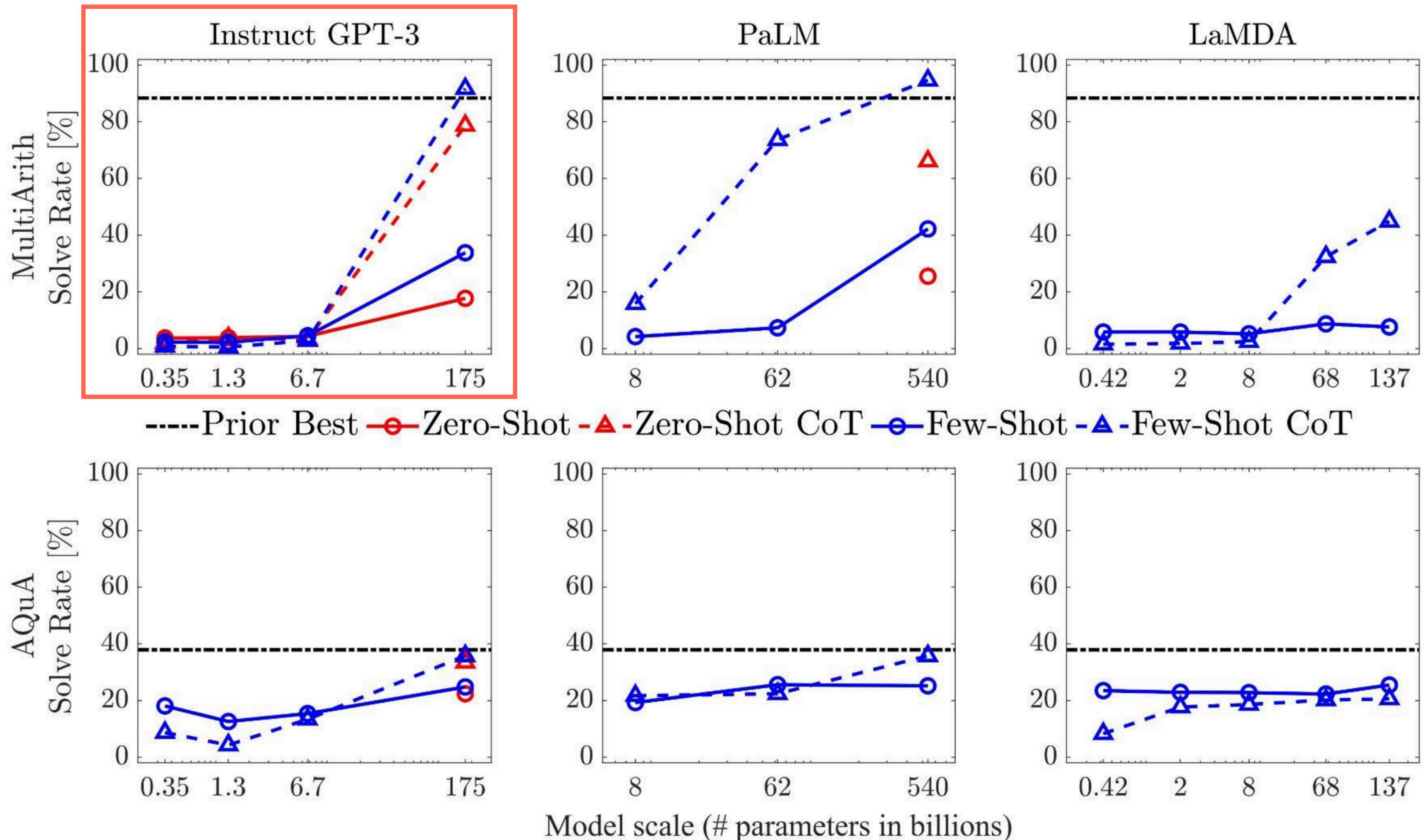
MAWPS - MultiArith

The school cafeteria ordered 42 red apples and 7 green apples for students lunches. But, if only 9 students wanted fruit, how many extra did the cafeteria end up with?

AQuA-RAT

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

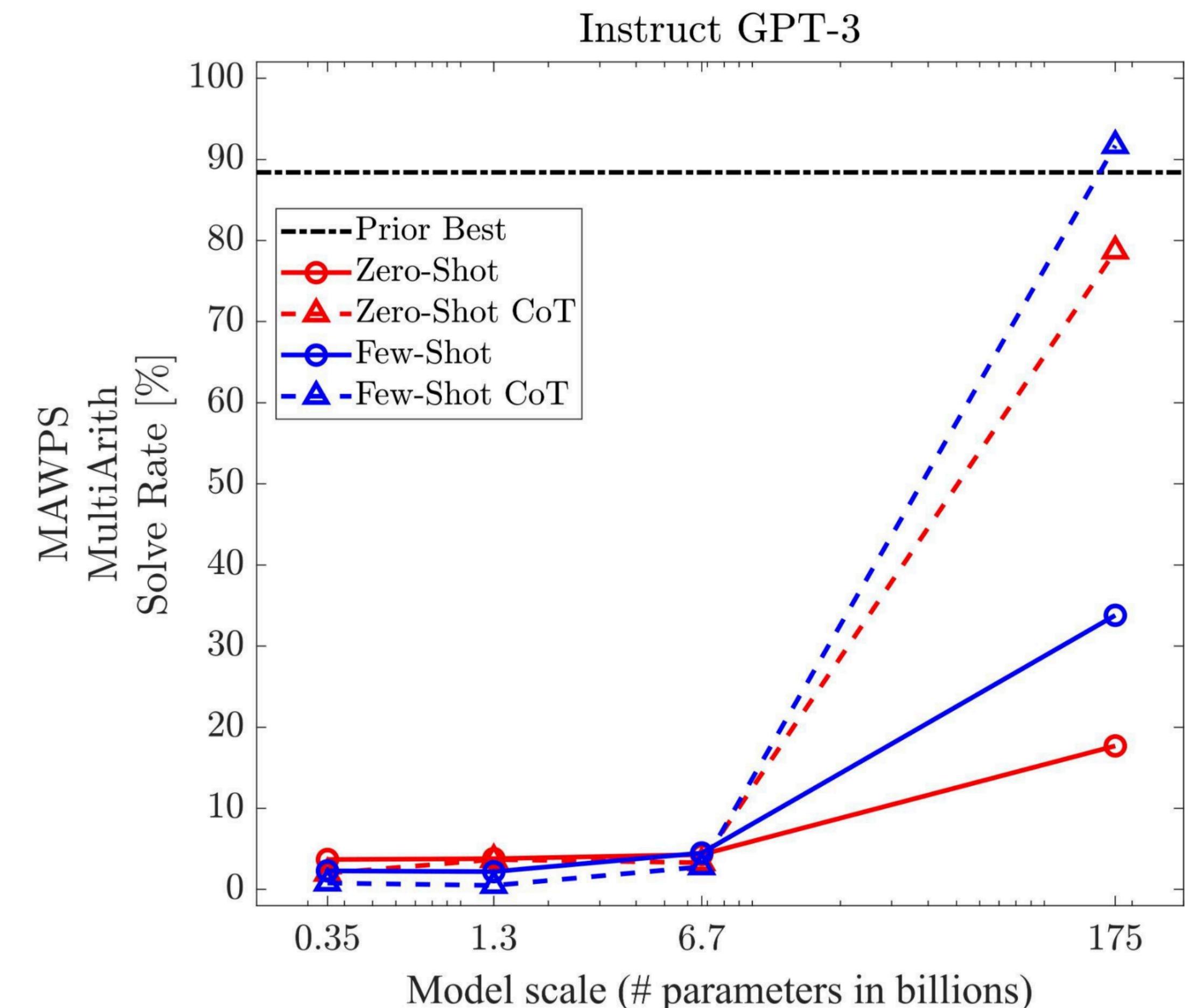
Arithmetic Reasoning - Results



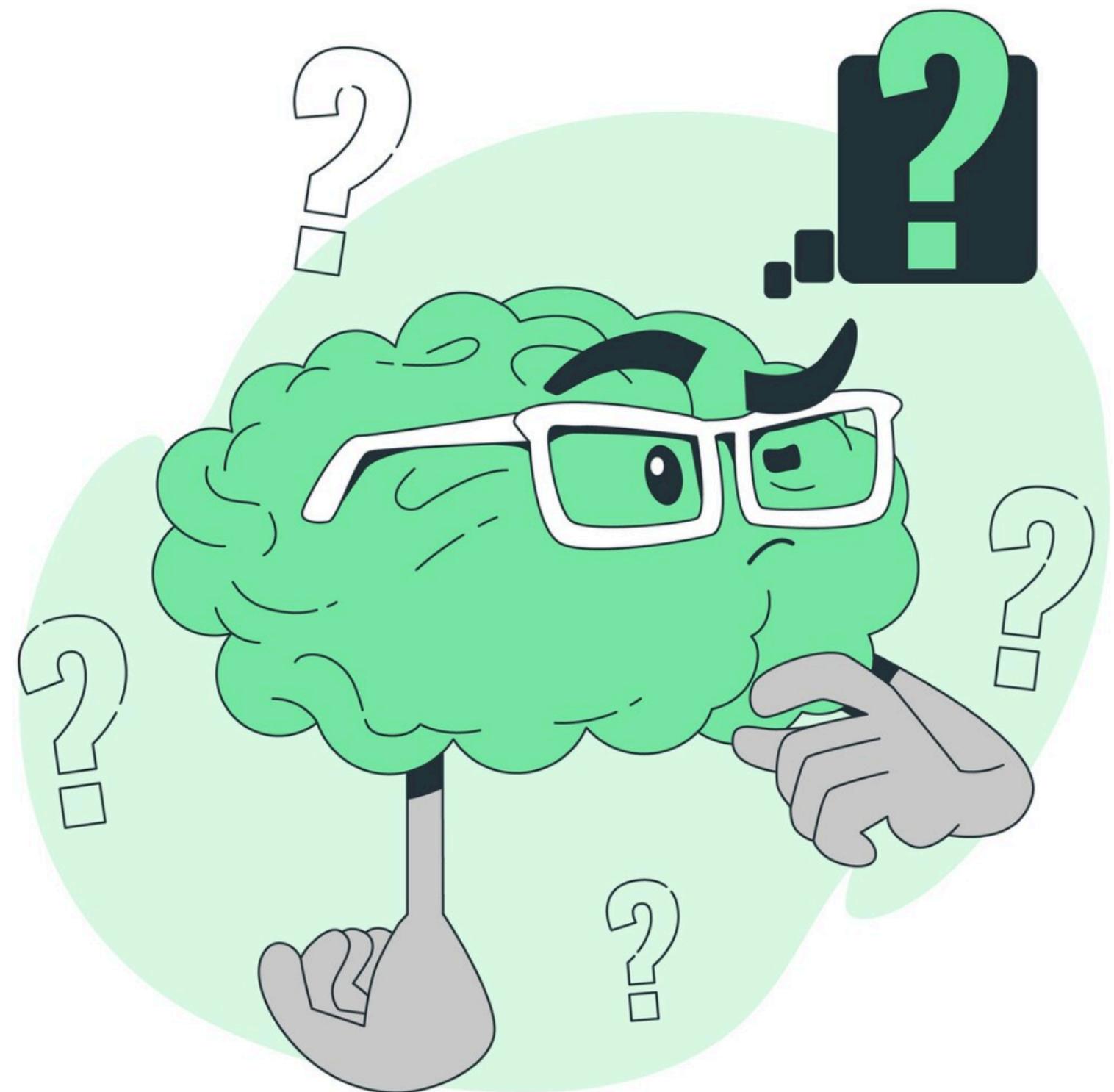
Instruct GPT-3: text-davinci-002 achieves similar performance as PaLM 540B model

Arithmetic Reasoning - Observations

- Both zero-shot and few-shot CoT promptings are emergent **abilities of model scale**.
- Do not positively impact performance for small models
 - start to yield performance gains when used with models with more than ~100B parameters.
- Few-shot CoT achieves **better** performance on LLM than zero-shot CoT.



Experiments Symbolic Reasoning



Symbolic Reasoning - Last Letter Concatenation

Last letter concatenation

Question: Take the last letters of the words in "Elon Musk" and concatenate them

Answer: The last letter of "Elon" is "n".

The last letter of "Musk" is "k".

Concatenating them is "nk".

The answer is nk.

- Generate full names by **randomly concatenating** names from the **top one-thousand first and last** names from name census data
- 4 exemplars with **strict** format

Symbolic Reasoning - Coin Flip

Coin Flip

Question: A coin is heads up. Tom does not flip the coin. Mike does not flip the coin. Is the coin still heads up?

Answer: The coin was flipped by no one. So the coin was flipped 0 times. The coin started heads up, and it was not flipped, so it is still heads up. So the answer is yes.

Coin Flip

Question: A coin is heads up. Jamey flips the coin. Teressa flips the coin. Is the coin still heads up?

Answer: The coin was flipped by Jamey and Teressa. So the coin was flipped 2 times, which is an even number. The coin started heads up, so after an even number of flips, it will still be heads up. So the answer is yes.

8 exemplars with strict format.

Symbolic Reasoning - In & Out-of-domain Test

Last letter concatenation

Question: Take the last letters of the words in "Elon Musk" and concatenate them

Answer: The last letter of "Elon" is "n". The last letter of "Musk" is "k". Concatenating them is "nk". The answer is **nk**.

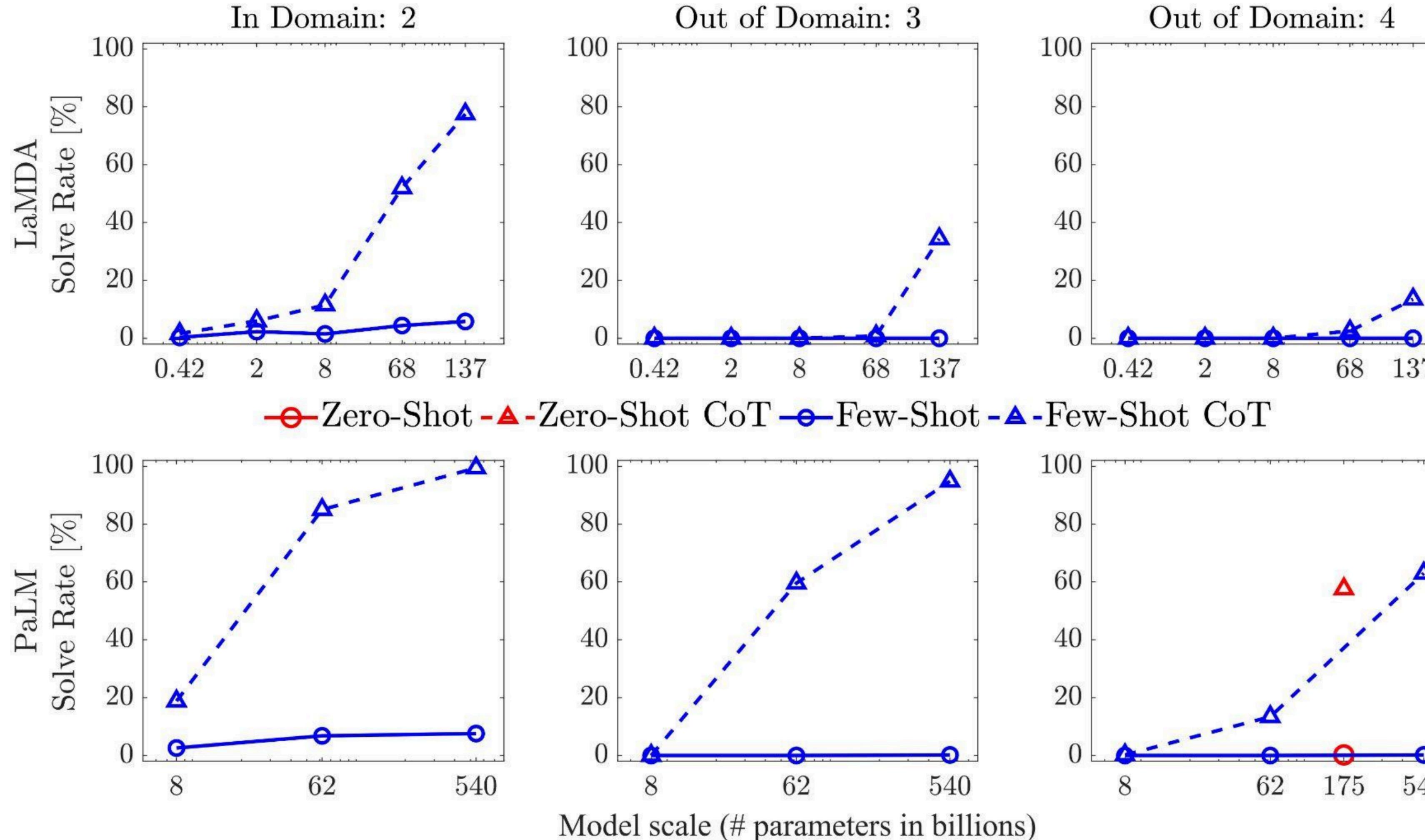
Coin Flip

Question: A coin is heads up. **Jamey** flips the coin. **Teressa** flips the coin. Is the coin still heads up?

Answer: **The coin was flipped by Jamey and Teressa. So the coin was flipped 2 times, which is an even number. The coin started heads up, so after an even number of flips, it will still be heads up.** So the answer is **yes**.

- **In-domain test set:** examples had the same number of steps as the few-shot exemplars
- **Out-of-domain (OOD) test set:** examples had more steps than those in the exemplars.

Symbolic Reasoning - Last Letter Concatenation



In-Domain

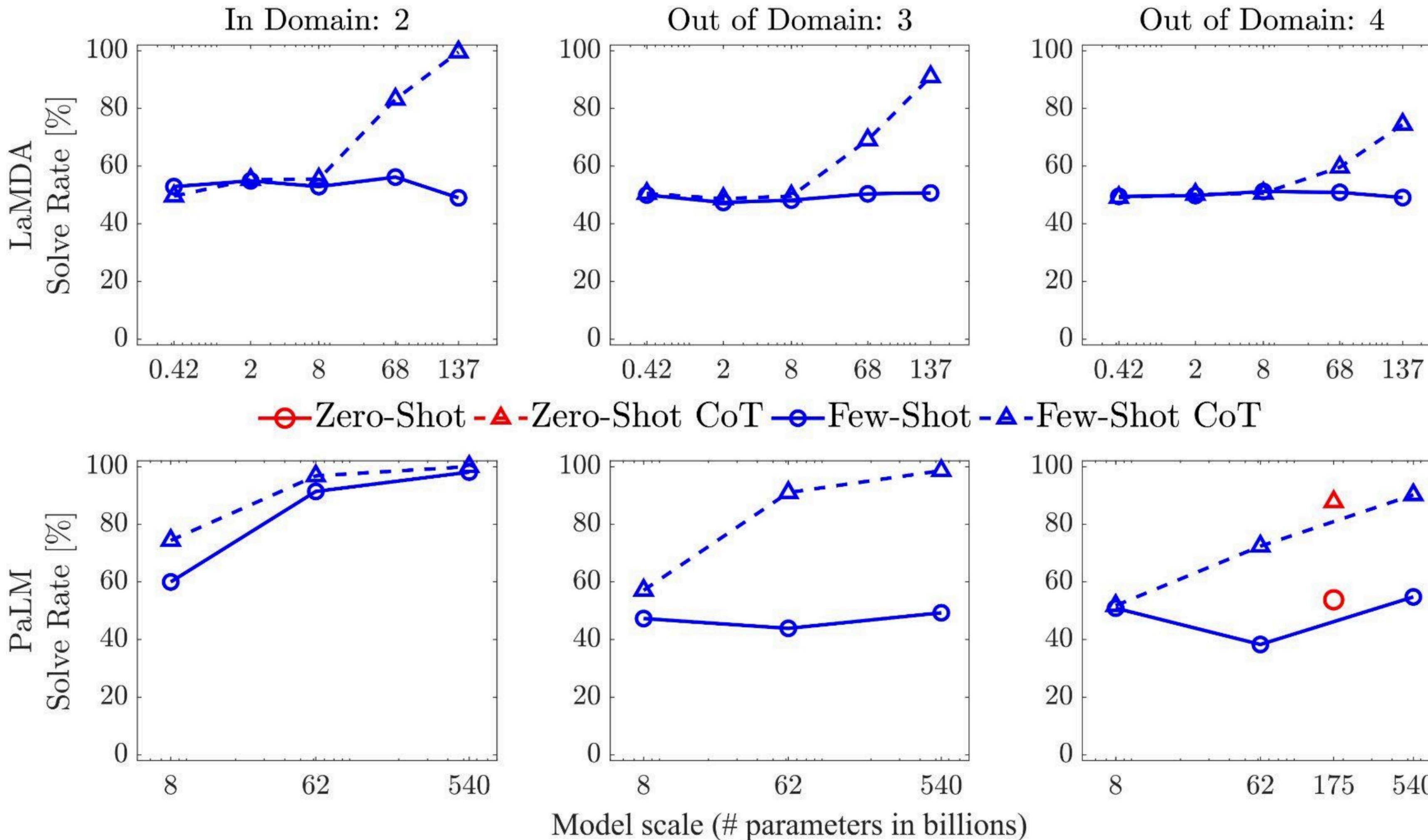
Take the last letters of the words in "**Elon Musk**" and concatenate them.

Out-of-Domain

Take the last letters of the words in "**Johann Sebastian Bach**" and concatenate them.

Zero-Shot results use **Instruct-GPT-3 175B** text-davinci-002 model.

Symbolic Reasoning - Coin Flip



In-Domain

A coin is heads up. **Tom does not flip the coin. Mike does not flip the coin.** Is the coin still heads up?

Out-of-Domain

A coin is heads up. **Tom does not flip the coin. Mike does not flip the coin. Jake flips the coin.** Is the coin still heads up?

Zero-Shot results use **Instruct-GPT-3 175B** text-davinci-002 model.

Symbolic Reasoning - Observations

- CoT promptings are emergent **abilities of model scale**
- Standard prompting **fails out-of-domain** tests for both tasks.
- **Zero-shot** CoT using **Instruct-GPT-3 175B** achieves the similar performance as **few-shot** CoT in both tasks using **540B PaLM model**.

Experiments

CommonSense Reasoning



Commonsense Reasoning - Toy Problems

CSQA (Talmor et al., 2019)

Question: What home entertainment equipment requires cable? Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet

Answer: The answer is **(c)**.

StrategyQA (Geva et al., 2021)

Question: Could Brooke Shields succeed at University of Pennsylvania?

Answer: The answer is **yes**.

Sport Understanding

Question: Is the following sentence plausible?
“Jamel Murray was perfect from the line.”

Answer: The answer is **yes**.

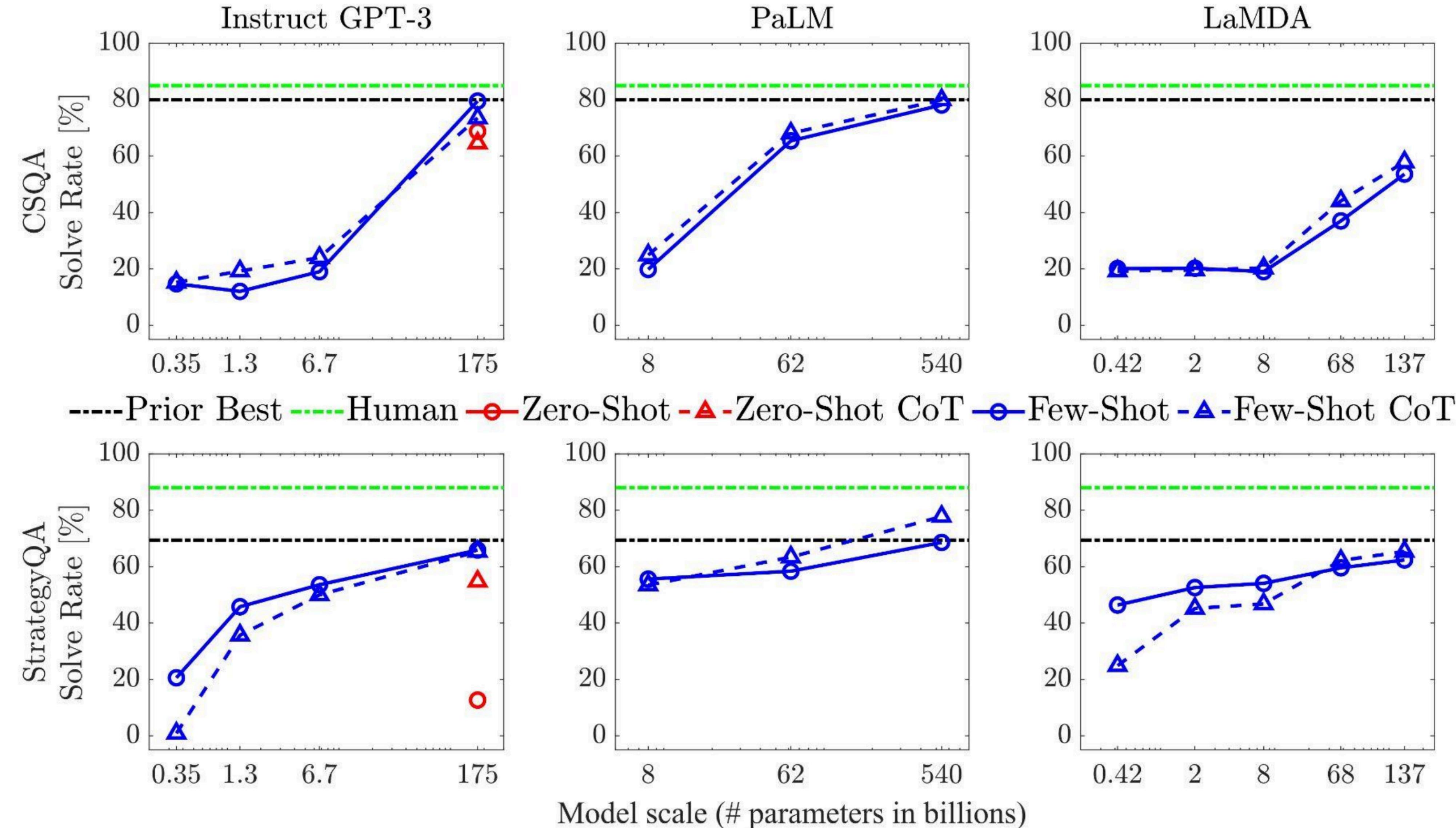
BIG-bench (Srivastava et al., 2022)

Date Understanding

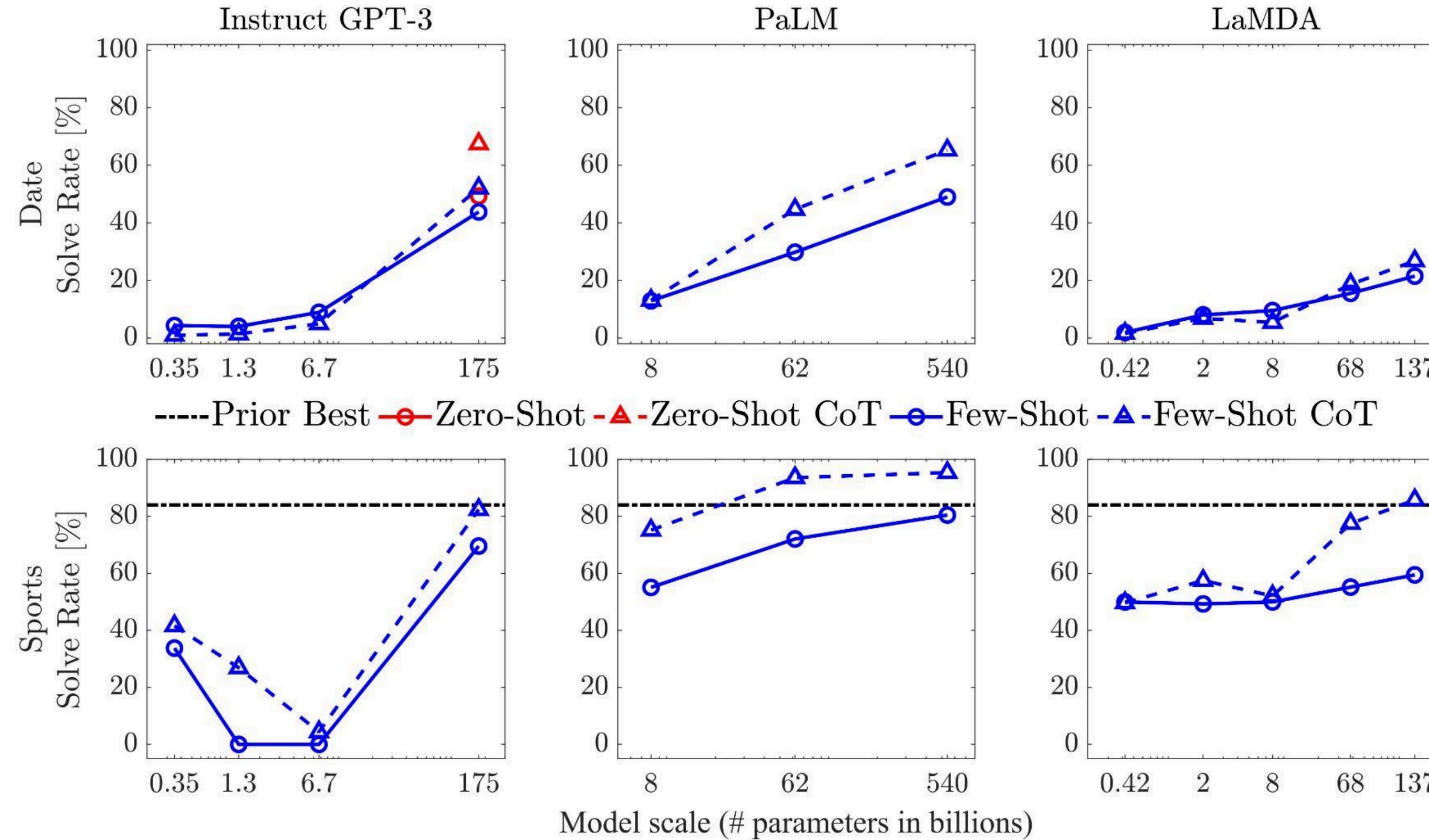
Question: 2015 is coming in 36 hours. What is the date one week from today in MM/DD/YYYY

Answer: So the answer is **01/05/2015**.

Commonsense Reasoning - Results



Commonsense Reasoning - Results



Commonsense Reasoning - Toy Problems

(Ahn et al.,2022)

SayCan Robot Planning

Locations = [counter, table, user, trash, bowl].

Objects = [cup, apple, kettle chips, tea, multigrain chips, coke, lime soda, jalapeno chips, rice chips, orange, grapefruit soda, pepsi, redbull, energy bar, sponge, water].

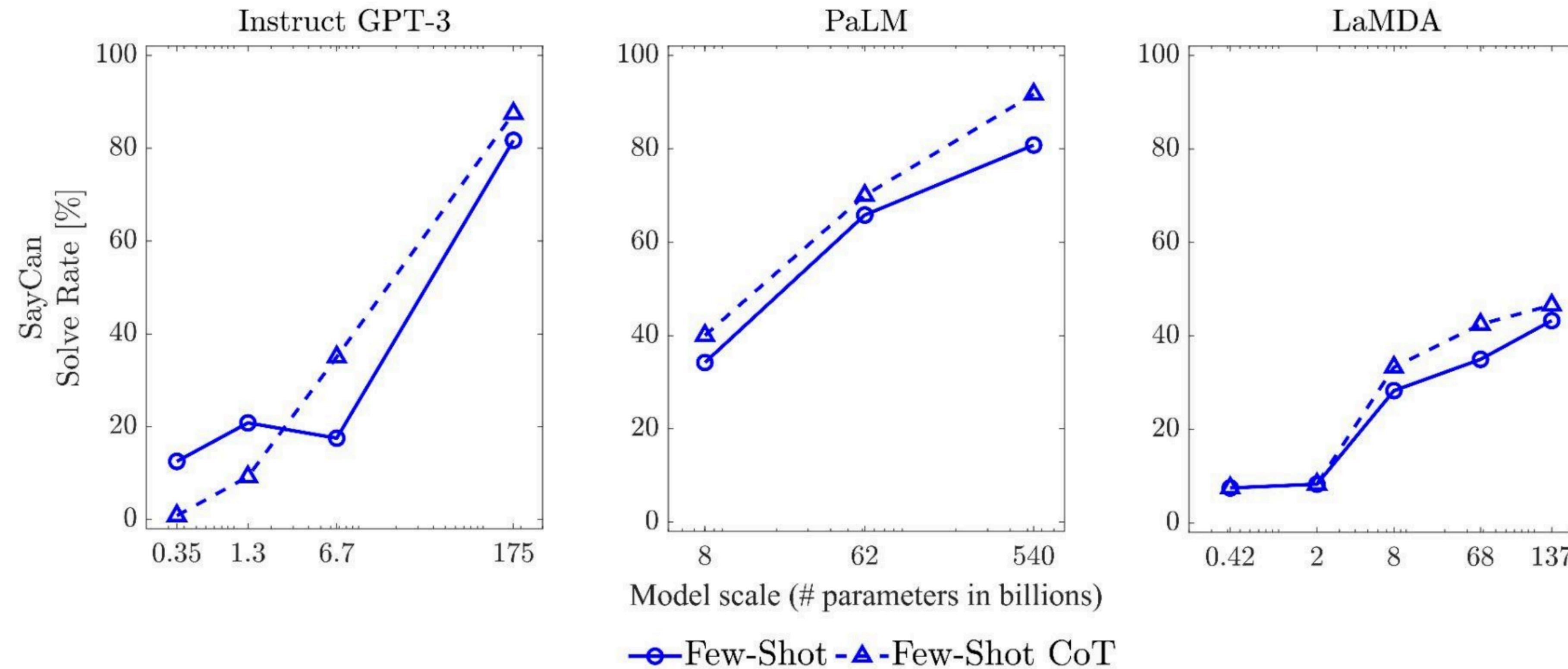
Actions: **pick(object), put(object), find(object), find(location).**

Human: How would you throw away a cup?

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

These tasks not only require **multi-steps reasoning**, but also need **priori knowledge** to understand complex semantics.

Commonsense Reasoning - Results



Commonsense Reasoning - Observations

- Scaling up **model size** improved the performance of standard prompting.
- CoT prompting made further gains
 - **largest improvement for PaLM 540B.**
- CoT show **minimal** benefits on CSQA and StrategyQA tasks
- **Few-shot** achieves better performance than **Zero-shot** CoT on 175B GPT-3 model for CSQA and Strategy QA tasks, but **Zero-shot** CoT shows significant improvement for **Date understanding** task.

Math Word Problems (free response)

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.

Math Word Problems (multiple choice)

Q: 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

A: 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).

CSQA (commonsense)

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

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm³, which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

Ablation Study: Variations of Few-Shot CoT

Change the types of CoT:

(b) Few-shot-CoT

Question: 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?

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

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Answer:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

Equation only



5+6=11. The answer is 11.

Ablation Study: Variations of Few-Shot CoT

Change the types of CoT:

(b) Few-shot-CoT

Question: 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?

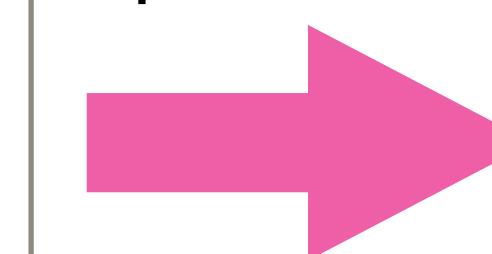
Answer: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Answer:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

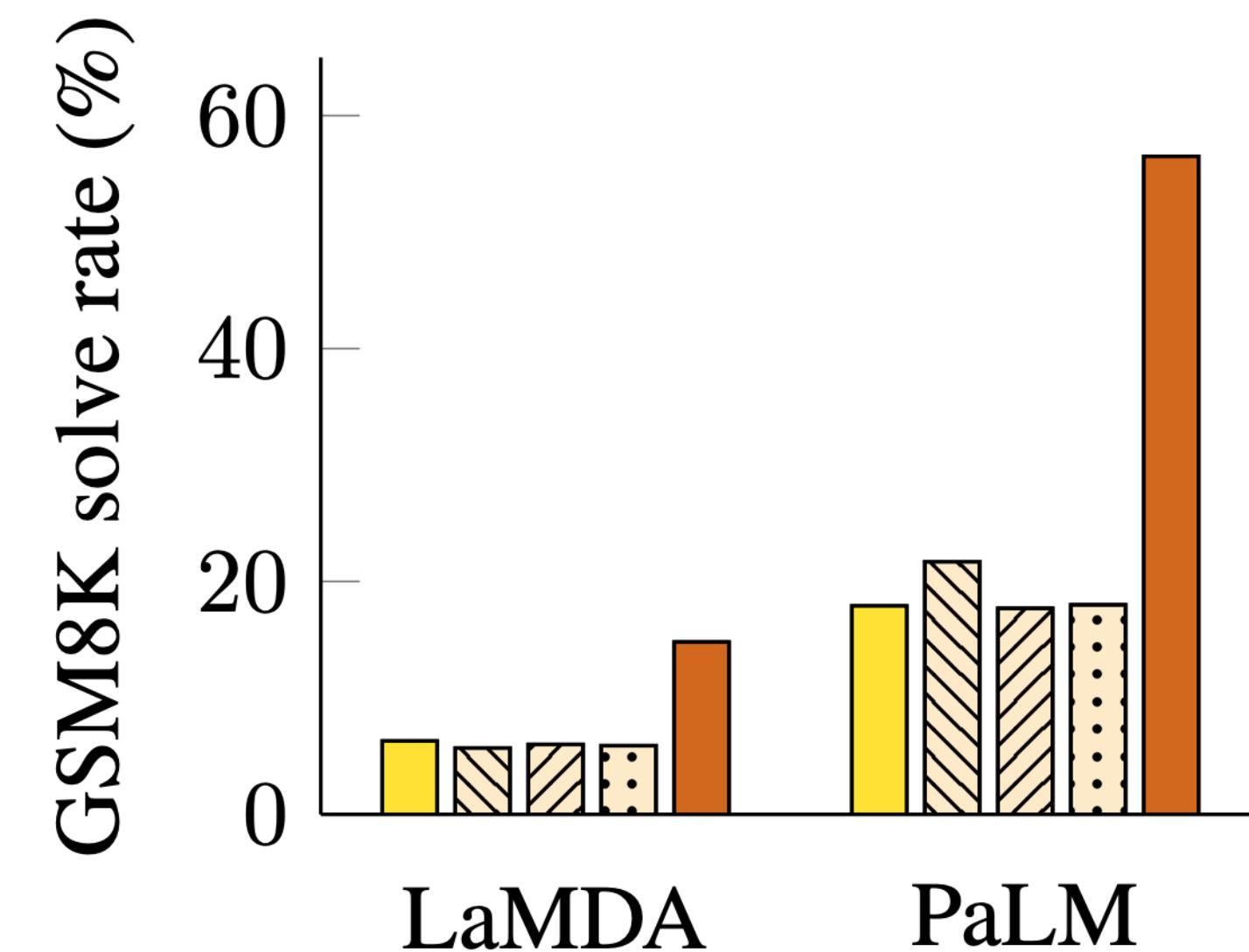
Equation only



5+6=11. The answer is 11.

Natural language in reasoning matters.

- █ Standard prompting
- █ Equation only
- █ Variable compute only
- █ Reasoning after answer
- █ Chain-of-thought prompting



Ablation Study: Variations of Few-Shot CoT

Change the types of CoT:

(b) Few-shot-CoT

Question: 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?

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

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Answer:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

Variable compute only



..... **The answer is 11.**

Ablation Study: Variations of Few-Shot CoT

Change the types of CoT:

(b) Few-shot-CoT

Question: 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?

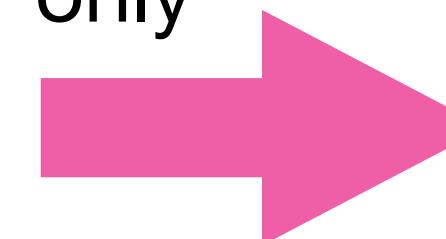
Answer: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Answer:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

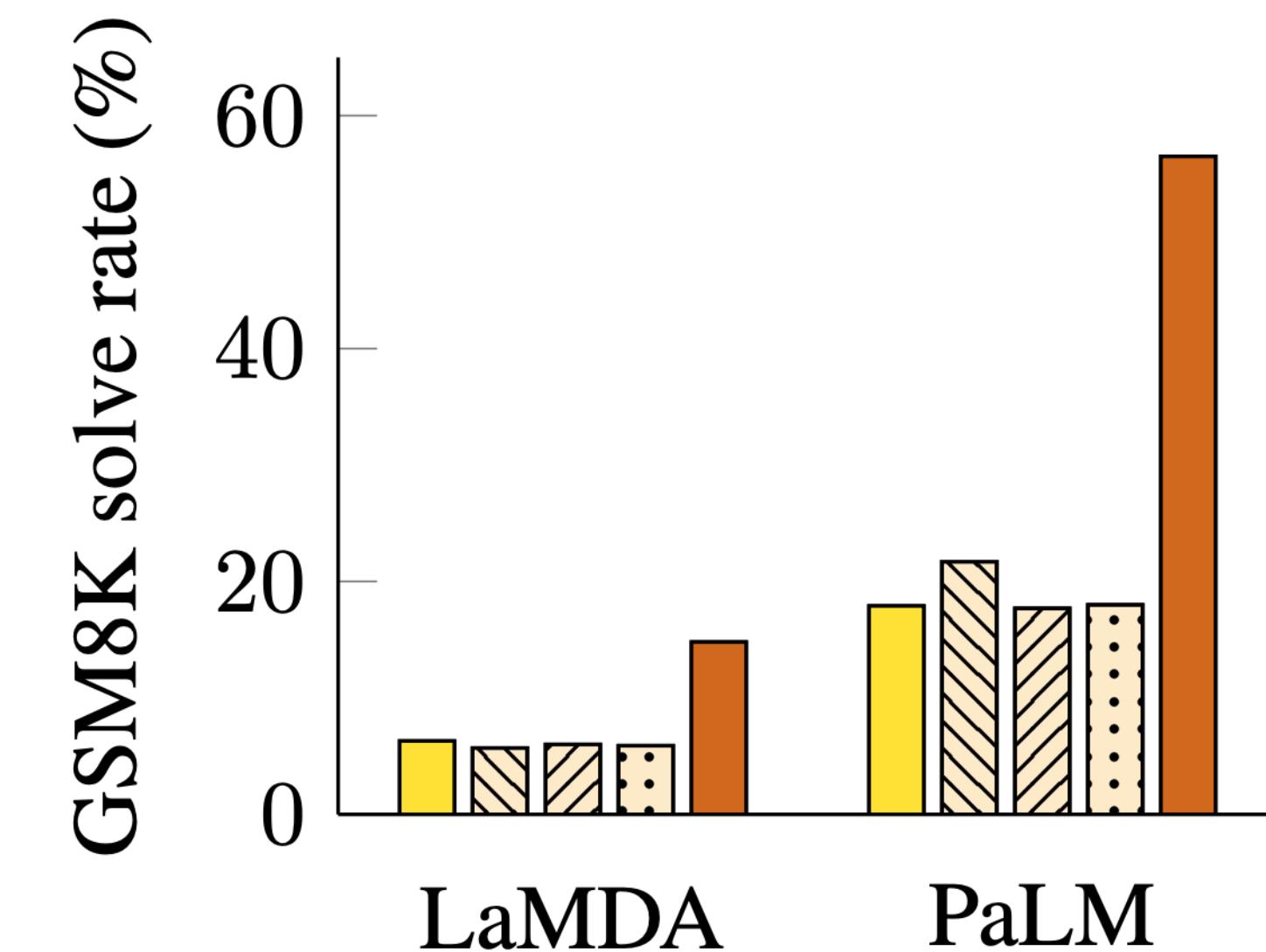
Variable
compute
only



*The
answer is 11.*

More intermediate computation does not help with the final answer.

- Standard prompting
- Equation only
- Variable compute only
- Reasoning after answer
- Chain-of-thought prompting



Ablation Study: Variations of Few-Shot CoT

Change the types of CoT:

(b) Few-shot-CoT

Question: 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?

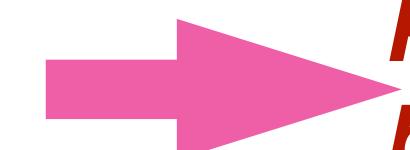
Answer: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Answer:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

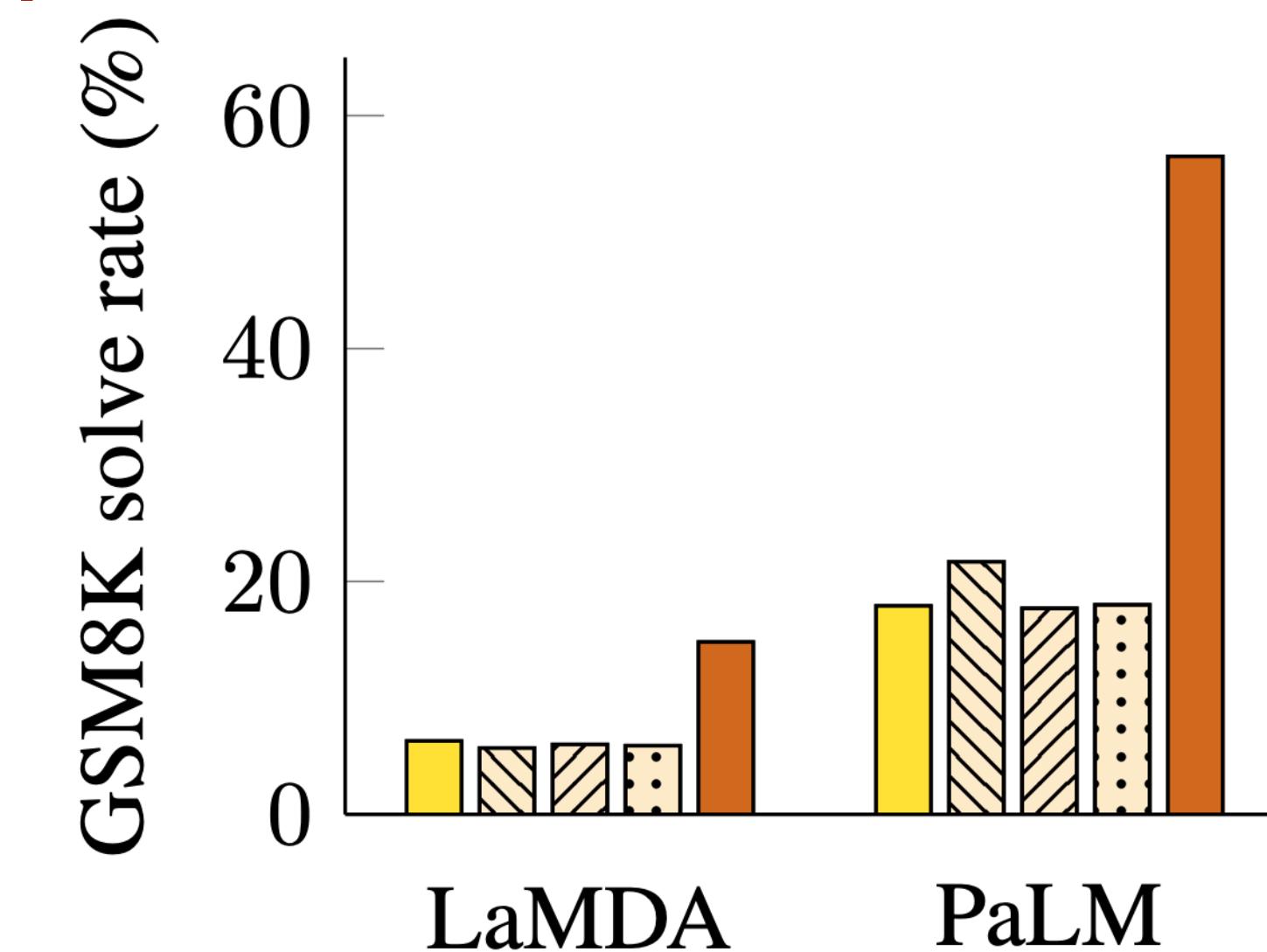
Reasoning
after
answer



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

- █ Standard prompting
- █ Equation only
- █ Variable compute only
- █ Reasoning after answer
- █ Chain-of-thought prompting

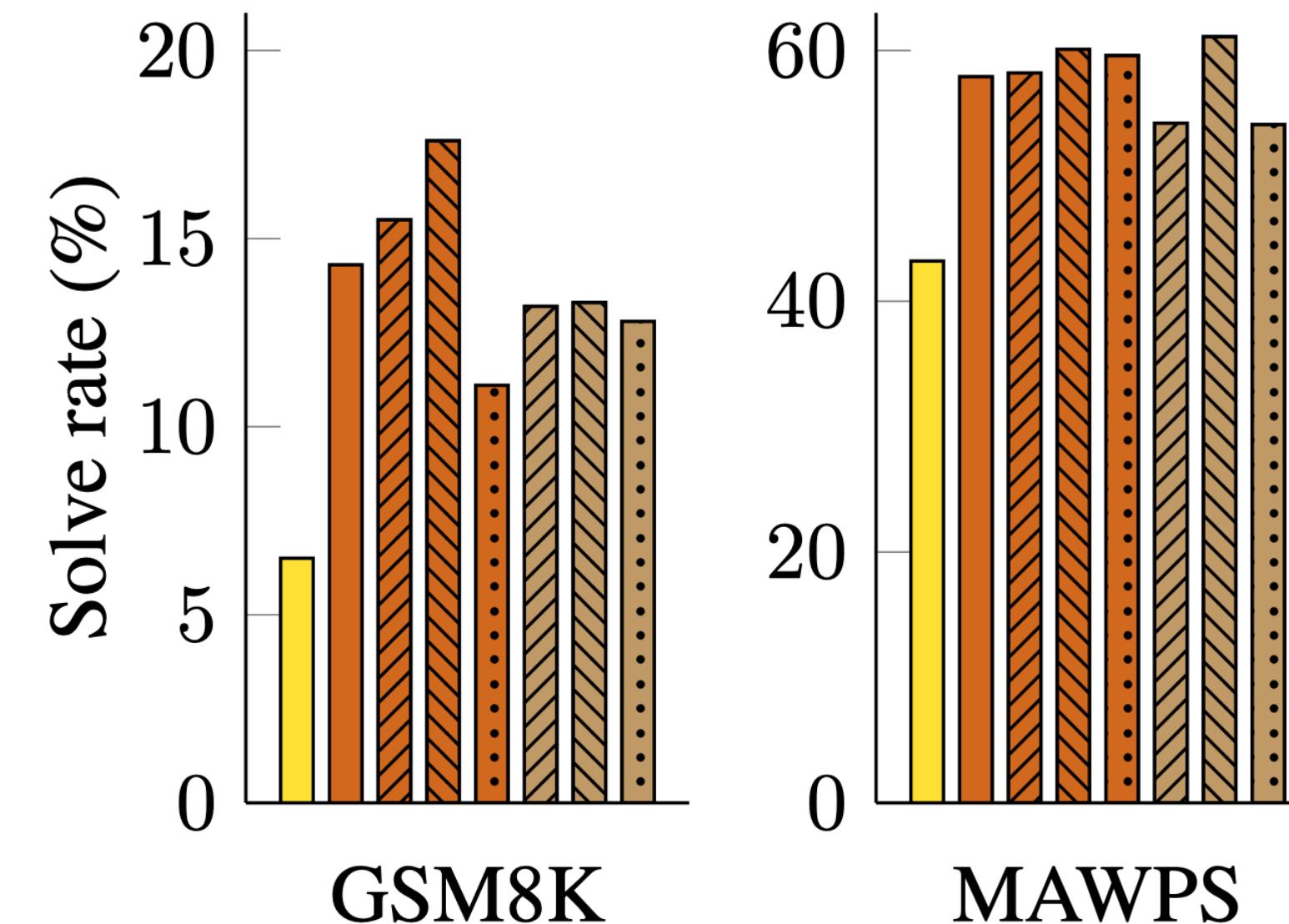
CoT is not just activating knowledge seen in pre-training.



Ablation Study: Robustness to Exemplars

Change the style of exemplar in **few-shot CoT**:

- Standard prompting
- Chain-of-thought prompting
- different annotator (B)
- different annotator (C)
- intentionally concise style
- exemplars from GSM8K (α)
- exemplars from GSM8K (β)
- exemplars from GSM8K (γ)



Results for **few-shot** LaMDA 137B on two **AR** tasks: have variance, but CoT still outperforms standard prompting, **robust against linguistic styles, different exemplars**.

Zero-shot Ablation Study: Robustness to Trigger Sentence

Change the template (trigger sentence) in **zero-shot CoT**:

Zero-shot CoT

Question: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

Answer: Let's think step by step.

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

No.	Template	Accuracy
1	Let's think step by step.	78.7
2	First, (*1)	77.3
3	Let's think about this logically.	74.5
4	Let's solve this problem by splitting it into steps. (*2)	72.2
5	Let's be realistic and think step by step.	70.8
6	Let's think like a detective step by step.	70.3
7	Let's think	57.5
8	Before we dive into the answer,	55.7
9	The answer is after the proof.	45.7
-	(Zero-shot)	17.7

Results for **zero-shot** GPT3 (davinci-002) 175B on MultiArith **AR** task: different templates encourage the model to express reasoning quite differently

Zero-shot Ablation Study: Model Size

Different model sizes in CoT prompting on MultiArith AR task:

	Original GPT-3 (0.3B / 1.3B / 6.7B / 175B)	Instruct GPT-3 (S / M / L / XL-1 / XL-2)
Zero-shot	2.0 / 1.3 / 1.5 / 3.3	3.7 / 3.8 / 4.3 / 8.0 / 17.7
Few-shot	5.2 / 5.2 / 4.0 / 8.1	3.0 / 2.2 / 4.8 / 14.0 / 33.7
Zero-shot-CoT	1.7 / 2.2 / 2.3 / 19.0	2.0 / 3.7 / 3.3 / 47.8 / 78.7
Few-shot-CoT	4.3 / 1.8 / 6.3 / 44.3	2.5 / 2.5 / 3.8 / 36.8 / 93.0

	GPT-2 (1.5B)	GPT-Neo (2.7B)	GPT-J (6B)	T0 (11B)	OPT (13B)
Zero-shot	3.2	3.0	2.7	2.8	3.7
Zero-shot-CoT	2.2	1.3	2.5	3.2	2.2

- Larger model, better reasoning
- CoT is effective only for larger models
- Few-shot better than zero-shot
- Instruction tuning is important to achieve high performance from CoT

More Advances - Self-Consistency

Change greedy decode (single-path) to self-consistency (multi-path) in few-shot CoT:

Prompt with example chains of thought

Q: Shawn has five toys. He gets two more each from his mom and dad. How many toys does he have now?

A: Shawn started with 5 toys. 2 toys each from his mom and dad is 4 more toys. The final answer is $5+4=9$. The answer is 9.

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for \$2 per egg. How much does she make every day?

A:

Language model

Sample decode with diverse reasoning paths

She has $16 - 3 - 4 = 9$ eggs left. So she makes $\$2 * 9 = \18 per day. **The answer is \$18.**

This means she uses $3 + 4 = 7$ eggs every day. So in total she sells $7 * \$2 = \14 per day. **The answer is \$14.**

She eats 3 for breakfast, so she has $16 - 3 = 13$ left. Then she bakes muffins, so she has $13 - 4 = 9$ eggs left. So she has $9 * \$2 = \18 . **The answer is \$18.**

Majority vote

The answer is \$18.

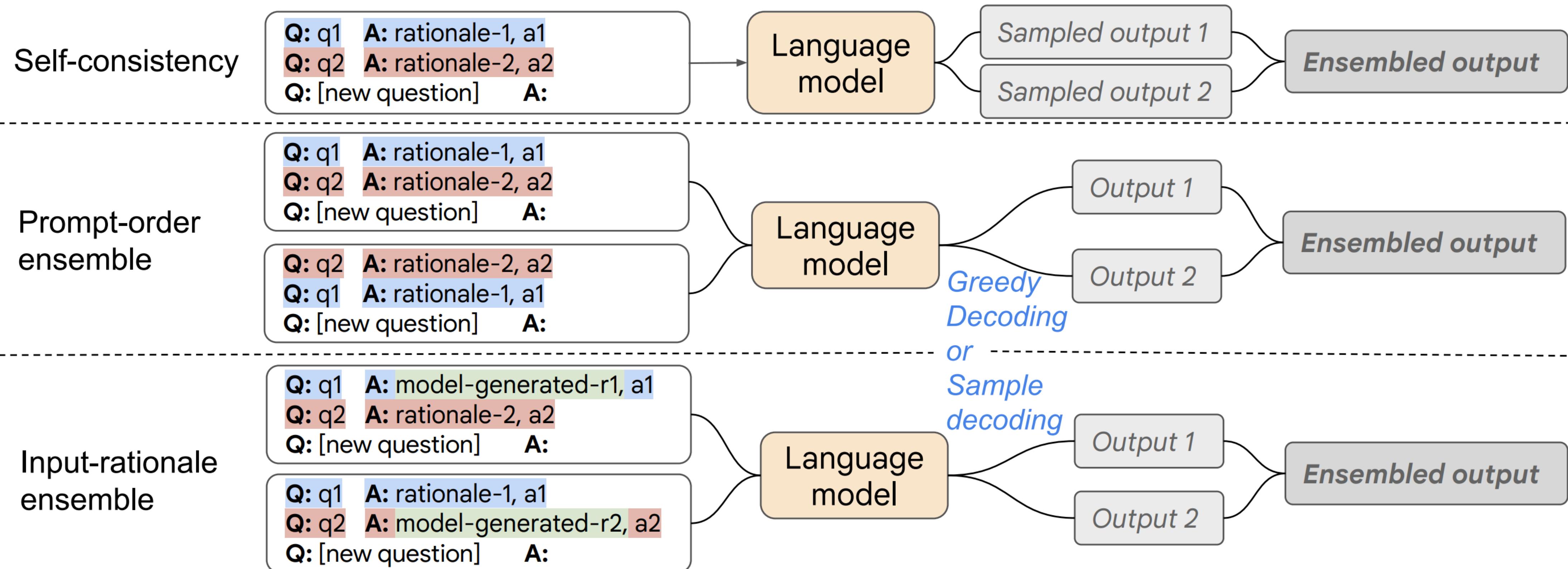
More Advances - Self-Consistency

Showcase results on **AR**, **CR** tasks:

	Method	GSM8K	CommonsenseQA
	Previous SoTA	$35^e / 57^g$	91.2^a
LaMDA (137B)	Greedy decode (Single-path)	17.1	57.9
	Self-Consistency (Multi-path)	27.7 (+10.6)	63.1 (+5.2)
PaLM (540B)	Greedy decode (Single-path)	56.5	79.0
	Self-Consistency (Multi-path)	74.4 (+17.9)	80.7 (+1.7)

More Advances - Input-Rational Ensemble

Use model-generated rationale in few-shot CoT:



More Advances - Input-Rational Ensemble

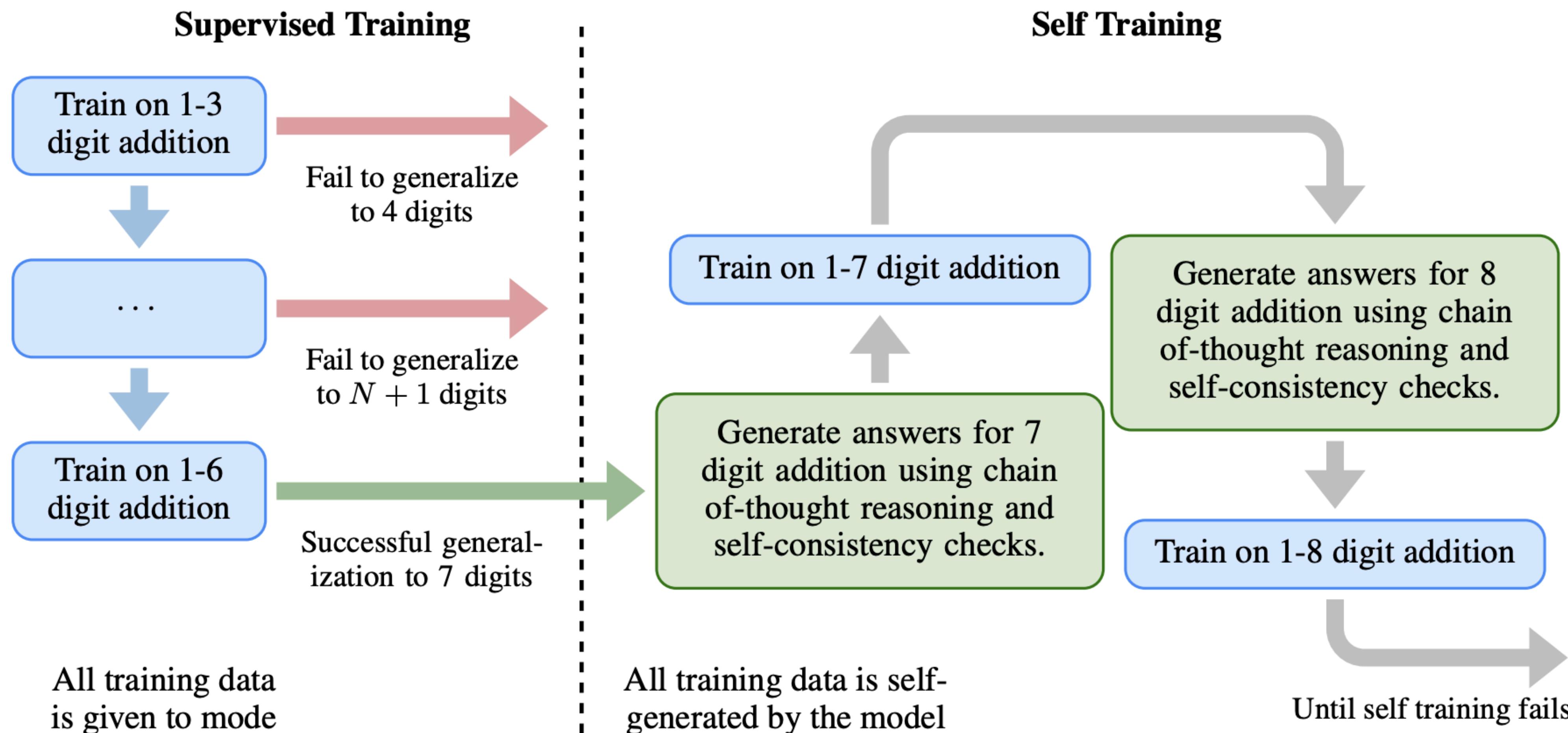
Showcase performance for AR reasoning tasks (PaLM-540B):

Method	GSM8K
Standard-prompting	17.9
Few-shot CoT (Wei et al. 2022)	56.5
Zero-shot CoT (Kojima et al. 2022)	43.0
Self-consistency (Wang et al. 2022)	74.4
Prompt-order ensemble	75.4
Input-rationale ensemble	73.8

Performance improvement on reasoning is great over previous CoT, but not significant against self-consistency,

More Advances - Self-Education (SECToR)

- Self-Education: LLMs can teach themselves new skills using CoT



More Advances: Self-Education (SECToR)

- A supervised fine-tuning that includes training with curriculum learning is required.
- Self-education process:
 - Use CoT to generate solutions to problems that it could not otherwise solve
 - Then, the model is fine-tuned to generate these exact solutions without using CoT
 - This obtained model can now directly solve problems without using CoT
 - The self-learning process can continue accordingly

LLMs are still unskilled at complex reasoning

- Benchmarks may be too simple to accurately gauge the true reasoning abilities of LLMs
- “LLMs are still far from achieving acceptable performance on common planning/ reasoning tasks which pose no issues for humans to do” (Valmeekam et al. 2022)
- Early experiments on GPT-4 showed signs of limitations on reasoning tasks requiring planning and backtracking (Bubeck et al. 2023)
- Autoregressive nature of LLMs may prevent them from planning and backtracking, two abilities necessary for complex reasoning (Gendron et al., 2023)

Questions

