Chain of Thought Prompting for Large Language Model Reasoning

Zihan Ding and Zixu Zhang

COS 597G - Fall 2022

Hard Language Tasks: 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: The answer is 5

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

A: The answer is **nk**.

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

A: The answer is (c).

Arithmetic Reasoning (AR) (+ - ×÷...) Symbolic Reasoning (SR)

Commonsense Reasoning (CR)

Fine-tune GPT-3 on GSM8K (arithmetic): (Cobbe et al. 2021)



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

GSM8K (arithmetic):



Prior best

- 540B LM: standard prompting
- 540B LM: chain of thought prompting



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

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

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 prompting

Chain of Thought Prompting

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 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

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

Machel Reid The University of Tokyo Yutaka Matsuo The University of Tokyo

Shixiang Shane Gu Google Research, Brain Team

> Yusuke Iwasawa The University of Tokyo

Definition:

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

Definition:

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

Intuition (from neural-symbolic computing):

use **<input**, **intermediate results**, **output>** triples, rather than simple **<input**, output> pairs

Definition:

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

Intuition (from neural-symbolic computing):

use <input, intermediate results, output> triples, rather than simple <input, output> pairs

Benefits:

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

(a) Few-shot

Examples

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: The answer is 11.

Q: 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?

A:

(Output) The answer is 8. X

(b) Few-shot-CoT (Wei et al., 2022) (a) Few-shot 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 Q: Roger has 5 tennis balls. He buys 2 more cans of tennis Examples balls. Each can has 3 tennis balls. How many tennis balls does he have now? **CoT Examples** A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 he have now? tennis balls. 5 + 6 = 11. The answer is 11. A: The answer is 11. Q: A juggler can juggle 16 balls. Half of the balls are golf balls, Q: 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 and half of the golf balls are blue. How many blue golf balls are there? there? A: A: (Output) The juggler can juggle 16 balls. Half of the balls are golf (Output) The answer is 8. X **Step-by-step Answer** 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.

(b) Few-shot-CoT (Wei et al., 2022) (a) Few-shot 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 Q: Roger has 5 tennis balls. He buys 2 more cans of tennis **Examples** balls. Each can has 3 tennis balls. How many tennis balls does he have now? **CoT Examples** A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 he have now? tennis balls. 5 + 6 = 11. The answer is 11. A: The answer is 11. Q: A juggler can juggle 16 balls. Half of the balls are golf balls, Q: 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 and half of the golf balls are blue. How many blue golf balls are there? there? A: A: (Output) The answer is 8. X (Output) The juggler can juggle 16 balls. Half of the balls are golf **Step-by-step Answer** 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.

(c) Zero-shot

Q: 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?

A: The answer (arabic numerals) is

(Output) 8 🗶

(0) 2010 01

	(a) Few-shot	(b) Few-shot-CoT (Wei et al., 2022)	
Examples	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: The answer is 11.	 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. 	CoT Examples
	Q: 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? A:	Q: 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? A:	
	(Output) The answer is 8. X	(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.	Step-by-step Answer
	(c) Zero-shot	(d) Zero-shot-CoT (KoJima et al., 2022))
	Q: 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? A: The answer (arabic numerals) is	 Q: 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? A: <i>Let's think step by step.</i> 	
	(Output) 8 🗙	(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:



Zero-Shot Chain of Thought (CoT)

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



Experiments

Pre-trained LLMs:

• Instruct GPT-3 (ada 350M, babbage 1.3B, curie 6.7B, and davinci 175B) (Ouyang et al., 2022)

- Instruct GPT-3 (ada 350M, babbage 1.3B, curie 6.7B, and davinci 175B) (Ouyang et al., 2022)
 - Not your familiar GPT-3 (Brown, et al., 2020)
 - Fine-tuned with human feedback
 - Stay tuned for the lecture on Nov. 14!!

- 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)
 - Only accessible to Googlers >.

- 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.

- 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)

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.





Train Verifier



(Cobbe et al. 2021)

Experiments Arithmetic Reasoning

Free Response

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.

• Manually composed 8 exemplars

(Wei et al., 2022)

Free Response

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.

- Manually composed 8 exemplars
- All contains equations with flexible formats

(Wei et al., 2022)

Free Response

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.

Free Response

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

(Wei et al., 2022)

(Wei et al., 2022)

You can have one or more equations.

Free Response

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.

Free Response

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

(Wei et al., 2022)

(Wei et al., 2022)

Equations can be incomplete and combined math with words.

Free Response

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.

- 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)

(Wei et al., 2022)

Multiple Choice

Q: 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

A: The distance that the person traveled would have been 20 km/hr * 2.5 hrs = 50 km. The answer is (e). • 4 exemplars, whose questions, intermediate reasoning, and answers are from AQuA-RAT's **training set**

Multiple Choice

Q: 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

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

Multiple Choice

Q: 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

A: 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

Q: 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

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

- 4 exemplars, whose questions, intermediate reasoning, and answers are from AQuA-RAT's **training set**
- Exemplars have flexible formats

Multiple Choice

Q: 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

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

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


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?



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?



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



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

• Both zero-shot and few-shot chain of thought promptings are emergent **abilities of model scale**.



- Both zero-shot and few-shot chain of thought promptings are emergent **abilities of model scale**.
- They do not positively impact performance for small models, but start to yield performance gains when used with models with more than ~100B parameters.



- Both zero-shot and few-shot chain of thought promptings are emergent **abilities of model scale**.
- They do not positively impact performance for small models, but 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.



- Both zero-shot and few-shot chain of thought promptings are emergent **abilities of model scale**.
- They do not positively impact performance for small models, but 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.
- Instruct GPT-3: text-davinci-002 achieves similar performance as PaLM 540B model



Experiments Symbolic Reasoning

Last letter concatenation

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

A: 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

Last letter concatenation

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

A: 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

Last letter concatenation

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

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

Last letter concatenation

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

A: The last letter of "Bill" is "I". The last letter of "Gates" is "s". Concatenating them is "Is". The answer is **Is**.

Symbolic Reasoning - Coin Flip

Coin Flip

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

A: 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

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

A: 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

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

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

Coin Flip

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

A: 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.

- 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.



*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.

• Standard prompting **fails out-of-domain** tests for both tasks.

- Standard prompting **fails out-of-domain** tests for both tasks.
- Both zero-shot and few-shot CoT promptings are emergent abilities of model scale.

- Standard prompting **fails out-of-domain** tests for both tasks.
- Few-shot CoT prompting is **emergent abilities of model scale.**
- CoT does not positively impact performance for small models, but start to yield performance gains when using models with more than ~100B parameters for both in-domain and out-of-domain tests.

- Standard prompting **fails out-of-domain** tests for both tasks.
- Few-shot CoT prompting is **emergent abilities of model scale.**
- CoT does not positively impact performance for small models, but start to yield performance gains when using models with more than ~100B parameters for both in-domain and out-of-domain tests.
- Zero-shot CoT using Instruct-GPT-3 175B achieves the similar performance as few-shot CoT in both tasks using 540B PaLM model.

Pre-Lecture Question 2

Q2: Wei et al., 2022 showed that CoT can improve out-of-domain performance. Can you state their results and why do you think it is the case (i.e., adding intermediate steps can improve robustness)?

Pre-Lecture Question 2

Q2: Wei et al., 2022 showed that CoT can improve out-of-domain performance. Can you state their results and why do you think it is the case (i.e., adding intermediate steps can improve robustness)?

While standard prompting fails out-of-domain tests for both tasks, large models with both zero-shot and few-shot CoT improve the performance of in-domain and out-of-domain tests. CoT prompt in this symbolic reasoning tasks really guides the LM to reason the process of mapping input to the output. Even if questions are *OOD* in the sense of "*how many words in the name*" and "*how many states to track*", the process of getting the output are still same, and can be resembled from exemplars. However, it is still unsure that CoT will improve other *OOD* scenarios with more complex reasoning processes.

Experiments CommonSense Reasoning

Commonsense Reasoning - Toy Problems

CSQA (Talmor et al., 2019)

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

A: The answer is (c).

Sport Understanding

Q: Is the following sentence plausible? "Jamel Murray was perfect from the line."

A: The answer is yes.

StrategyQA (Geva et al., 2021)

Q: Could Brooke Shields succeed at University of Pennsylvania?

A: The answer is yes.

Date Understanding

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

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

Commonsense Reasoning - Toy Problems

SayCan Robot Planning

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

Objects = [7up, apple, kettle chips, tea, multigrain chips, coke, lime soda, jalapeno chips, rice chips, orange, grapefruit soda, pepsi, redbull, energy bar, sponge, water].

The robot can pick up items with **pick(object)** and put down items with **put(object)** as well as find objects or locations with **find()**. The robot can only understand the explicit locations and objects listed.

Human: How would you throw away a redbull?

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

Commonsense Reasoning - Toy Problems

SayCan Robot Planning

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

Objects = [7up, apple, kettle chips, tea, multigrain chips, coke, lime soda, jalapeno chips, rice chips, orange, grapefruit soda, pepsi, redbull, energy bar, sponge, water].

The robot can pick up items with **pick(object)** and put down items with **put(object)** as well as find objects or locations with **find()**. The robot can only understand the explicit locations and objects listed.

Human: How would you throw away a redbull?

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

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

Commonsense Reasoning - Results



Commonsense Reasoning - Results



Commonsense Reasoning - Results



• For all tasks, scaling up **model size** improved the performance of standard prompting.

- For all tasks, scaling up **model size** improved the performance of standard prompting.
- CoT prompting led to further gains, with improvements appearing to be **largest for PaLM 540B**.

- For all tasks, scaling up **model size** improved the performance of standard prompting.
- CoT prompting led to further gains, with improvements appearing to be **largest for PaLM 540B**.
- CoT has **minimal** benefits on CSQA and StrategyQA tasks.

- For all tasks, scaling up **model size** improved the performance of standard prompting.
- CoT prompting led to further gains, with improvements appearing to be **largest for PaLM 540B**.
- CoT has **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.

Ablation Study - Variations of Few-Shot CoT

Change the types of CoT:

(b) Few-shot-CoT

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: 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? A:

(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

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: 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? A:

(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.

Standard prompting Equation only Variable compute only :: Reasoning after answer Chain of thought prompting **GSM8K** solve rate (%) 60 4020LaMDA PaLM (Wei et al., 2022)

Natural language in reasoning matters.

71

Ablation Study - Variations of Few-Shot CoT

Change the types of CoT:

blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4.

(b) Few-shot-CoT

 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 have new?
 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: 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?
 A:

 A:
 (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
 Fee and the golf balls are
Ablation Study - Variations of Few-Shot CoT



More intermediate computation does not help with the final answer.

(Wei et al., 2022)

PaLM

LaMDA

Ablation Study - Variations of **Few-Shot CoT**

Change the types of CoT:

(b) Few-shot-CoT 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 Reasoning after answer tennis balls. 5 + 6 = 11. The answer is 11. Q: 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? A: (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.

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

Ablation Study - Variations of Few-Shot CoT



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

Ablation Study - Robustness (Style of Exemplar)

Change the style of exemplar in few-shot CoT:

(b) Few-shot-CoT

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: 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?

A:

(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.



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

Ablation Study - Robustness (Trigger Sentence)

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

(d) Zero-shot-CoT (Ours)

Q: 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?

A: 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
а <u>т</u>	(Zero-shot)	17.7

(Kojima et al., 2022)

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

Ablation Study - Model Size

Change the model sizes in CoT prompting:

(Kojima et al., 2022)

	Original GPT-3 (0.3E	B/1.3B/6.7B/175B)	Instruct GPT-3 (0.3B/1.3B/6.7B/175B-1/175B-2)		
Zero-shot	p-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

Results on MultiArith AR task with different model sizes:

- Larger model, better reasoning
- CoT is effective only for larger models
- Few-shot better than zero-shot
- Instruct GPT-3 is much better than original GPT-3

Pre-Lecture Question 1

Q1: Describe how the two approaches from (Wei et al., 2022) and (Kojima et al., 2022) are different. Which one do you think is a more viable solution in terms of cost, performance and stability?

Wei's work uses few-shot setting with several demonstration examples required, and the CoT annotations need to be provided for each example, while Kojima's work uses LLM to generate the CoT with the two-stage prompting and no longer requires the annotated examples.

For most of benchmarks we have seen, few-shot CoT has better performance than zero-shot CoT, while zero-shot CoT does not require human annotations, which can be costly.

Although there is no direct comparison between zero-shot and few-shot CoT on stability, few-shot seems to be more robust as its performance does not vary significantly when changing the prompt annotations. On the other hand, zero-shot CoT has significant performance variance with different trigger sentences.

More Advances - Self-Consistency

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



Wang, Xuezhi, et al. "Self-consistency improves chain of thought reasoning in language models." arXiv preprint arXiv:2203.11171 (2022).

More Advances - Self-Consistency

Showcase results on AR, CR tasks:

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

Wang, Xuezhi, et al. "**Self-consistency improves chain of thought reasoning in language models**." arXiv preprint arXiv:2203.11171 (2022).

81

More Advances - Input-Rational Ensemble

Use model-generated rationale in few-shot CoT:



Wang, Xuezhi, et al. "Rationale-Augmented Ensembles in Language Models." arXiv preprint arXiv:2207.00747 (2022).

More Advances - Input-Rational Ensemble

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

	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,

Pre-Lecture Question 3 and Discussion

Q3: Do you think the CoT method can be useful to other NLP tasks that we have seen in the previous lectures (standard NLP tasks that are beyond the arithmetic/logic reasoning tasks that these papers evaluated on)? Do you have any ideas about how we can collect the CoT data?

Reference

- 1. Wei, Jason, et al. "Chain of thought prompting elicits reasoning in large language models." *arXiv preprint arXiv:2201.11903* (2022).
- 2. Kojima, Takeshi, et al. "Large Language Models are Zero-Shot Reasoners." *arXiv preprint arXiv:2205.11916* (2022).
- 3. Cobbe, Karl, et al. "Training verifiers to solve math word problems." *arXiv preprint arXiv:2110.14168* (2021).
- 4. Patel, Arkil, Satwik Bhattamishra, and Navin Goyal. "Are NLP Models really able to Solve Simple Math Word Problems?." *NAACL 2021*.
- 5. Miao, Shen-Yun, Chao-Chun Liang, and Keh-Yih Su. "A diverse corpus for evaluating and developing English math word problem solvers." ACL 2020 (2020).
- 6. Koncel-Kedziorski, Rik, et al. "MAWPS: A math word problem repository." *NAACL 2016*.
- 7. Ling, Wang, et al. "Program induction by rationale generation: Learning to solve and explain algebraic word problems." *arXiv preprint arXiv:1705.04146* (2017).
- 8. Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. NAACL.
- 9. Ahn, Michael, et al. "Do as i can, not as i say: Grounding language in robotic affordances." arXiv preprint arXiv:2204.01691 (2022).
- 10. Wang, Xuezhi, et al. "Self-consistency improves chain of thought reasoning in language models." arXiv preprint arXiv:2203.11171 (2022).
- 11. Wang, Xuezhi, et al. "Rationale-Augmented Ensembles in Language Models." *arXiv preprint arXiv:2207.00747* (2022).

Discussion (This part will not be in presentation)

Summary of Arithmetic Reasoning Benchmark

Dataset	N	Example problem
GSM8K	1,319	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	1,000	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?
ASDiv	2,096	Ellen has six more balls than Marin. Marin has nine balls. How many balls does Ellen have?
AQuA	254	A car is being driven, in a straight line and at a uniform speed, towards the base of a vertical tower. The top of the tower is observed from the car and, in the process, it takes 10 minutes for the angle of elevation to change from 45° to 60° . After how much more time will this car reach the base of the tower? Answer Choices: (a) $5\sqrt{3} + 1$ (b) $6\sqrt{3} + \sqrt{2}$ (c) $7\sqrt{3} - 1$ (d) $8\sqrt{3} - 2$ (e) None of these
MAWPS: SingleOp	562	If there are 7 bottle caps in a box and Linda puts 7 more bottle caps inside, how many bottle caps are in the box?
MAWPS: SingleEq	508	Benny bought a soft drink for 2 dollars and 5 candy bars. He spent a total of 27 dollars. How much did each candy bar cost?
MAWPS: AddSub	395	There were 6 roses in the vase. Mary cut some roses from her flower garden. There are now 16 roses in the vase. How many roses did she cut?
MAWPS: MultiArith	600	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?

Arithmetic Reasoning - Few-Shot CoT MAWPS Results



Arithmetic Reasoning - Zero-Shot CoT Additional Results

Table 3: Model scale study. Evaluation metric is accuracy on MultiArith dataset. 175B-1 : textdavinci-001, 175B-2 : text-davinci-002. It is verified that CoT is effective when the model is larger, such as Instruct GPT-3 (175B parameters; text-davinci-001 and text-davinci-002) and Original GPT-3 (175B parameters; davinci). In this experiment, the order of performance (ascending) is Zero-shot, Few-shot (8samples), Zero-shot-CoT, and Few-shot-CoT (8samples) for davinci and text-davinci-002.

×.	Original GPT-3 (0.3B	8/1.3B/6.7B/175B)	Instruct GPT-3 (0.3B/1.3B/6.7B/175B-1/175B-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	4.3/18/63/44.3		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

Prior Best – Fine-tuning + Verification



1. Fine-tuned 2 epoch on training set.

Prior Best – Fine-tuning + Verification



Finding the one with the highest score

 S_i^{100}

Prior Best – Fine-tuning + Verification



2. Sample 100 solutions from the generator for each training problem and label each solution as correct or incorrect.



 Y_i^1

 Y^2_i

 Y_{i}^{100}

 S_i^{100}

Arithmetic Reasoning - Results



MAWPS - SingleEq If there are 7 bottle caps in a box and Linda puts 7 more bottle caps inside, how many bottle caps are in the box?

MAWPS - AddSub There were 6 roses in the vase. Mary cut some roses from her flower garden. There are now 16 roses in the vase. How many roses did she cut?

Commonsense Reasoning - CSQA CoT Prompt

CSQA (Talmor et al., 2019)

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. The answer is (c). • 7 exemplars from training dataset

Commonsense Reasoning - CSQA CoT Prompt

CSQA (Talmor et al., 2019)

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. The answer is (c).

- 7 exemplars from training dataset
- Manually composed intermediate reasoning with strict format:

The answer must _____. *Of the above choices, only* _____.

Commonsense Reasoning - Strategy QA CoT Prompt

StrategyQA (Geva et al., 2021)

Q: Could Brooke Shields succeed at University of Pennsylvania?

A: Brooke Shields went to Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. The answer is yes. • 6 exemplars from training dataset

Commonsense Reasoning - Strategy QA CoT Prompt

StrategyQA (Geva et al., 2021)

Q: Could Brooke Shields succeed at University of Pennsylvania?

A: Brooke Shields went to Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. The answer is yes.

- 6 exemplars from training dataset
- Manually composed intermediate reasoning with flexible format

Commonsense Reasoning - Strategy QA CoT Prompt

StrategyQA (Geva et al., 2021)

Q: Could Brooke Shields succeed at University of Pennsylvania?

A: Brooke Shields went to Princeton University. Princeton University is about as academically rigorous as the University of Pennsylvania. Thus, Brooke Shields could also succeed at the University of Pennsylvania. The answer is yes.

StrategyQA (Geva et al., 2021)

Q: Yes or no: Is it common to see frost during some college commencements?

A: College commencement ceremonies can happen in December, May, and June. December is in the winter, so there can be frost. Thus, there could be frost at some commencements. The answer is yes.

Commonsense Reasoning - CoT Prompt

Date Understanding

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

A: If 2015 is coming in 36 hours, then it is coming in 2 days. 2 days before 01/01/2015 is 12/30/2014, so today is 12/30/2014. So one week from today will be 01/05/2015. So the answer is 01/05/2015.

Sport Understanding

Q: Is the following sentence plausible? "Jamel Murray was perfect from the line."

A: Jamal Murray is a basketball player. Being perfect from the line is part of basketball. The answer is yes.

6 exemplars for Date Understanding

8 exemplars for Sport Understanding

Commonsense Reasoning - CoT Prompt

SayCan Robot Planning (Ahn et al., 2022)

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

Objects = [7up, apple, kettle chips, tea, multigrain chips, coke, lime soda, jalapeno chips, rice chips, orange, grapefruit soda, pepsi, redbull, energy bar, sponge, water].

The robot can pick up items with **pick(object)** and put down items with **put(object)** as well as find objects or locations with **find()**. The robot can only understand the explicit locations and objects listed.

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().

Ablation Study - Robustness (Example Distribution)

Change the examples in **few-shot CoT**:

(b) Few-shot-CoT

Q: Roger has 5 tennis b balls. Each can has 3 ten he have now? A: Roger started with 5 ba tennis balls. 5 + 6 = 11. Th	alls. He buys 2 more cans inis balls. How many tennis b alls. 2 cans of 3 tennis balls he answer is 11.	of tennis valls does each is 6			
Q: A juggler can juggle 16 and half of the golf balls a there? A:	6 balls. Half of the balls are g are blue. How many blue golf	golf balls, balls are			
(Output) The juggler can juballs. So there are 16 / 2 blue. So there are 8 / 2 = 4	uggle 16 balls. Half of the ball = 8 golf balls. Half of the golf 4 blue golf balls. The answer i	Is are golf t balls are is 4. Chan	ge examples from i	n-domain to out-o	f-domair
	Zero-shot	Few-shot-CoT [†]	Zero-shot-CoT	Few-shot-CoT	Г
AQUA-RAT MultiArith	22.4 17.7	$\frac{31.9}{27.0}$	33.5 78.7	39.0 88.2	-

Results for **few-shot** on two **AR** tasks with examples[†] from a CR task (CommonsenseQA): across-domain examples with the same format have minor performance degradation.