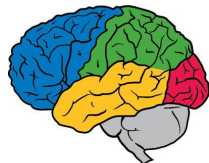


Emergence and reasoning in large language models

Jason Wei
Senior Research Scientist
Google Brain



Outline

- **Emergent abilities** of large language models (Wei et al., 2022).
- **Chain-of-thought prompting** elicits reasoning in large language models (Wei et al., 2022).
- Additional work
 - **Self-consistency** improves chain-of-thought reasoning in language models (Wang et al., 2022).
 - **Least-to-most prompting** enables complex reasoning in large language models (Zhou et al., 2022).
 - Language models are **multilingual chain-of-thought** reasoners (Shi et al., 2022).
 - **Challenging BIG-Bench tasks** and whether chain-of-thought can solve them (Suzgun et al., 2022).
- Feel free to interrupt with questions :)

Emergent Abilities of Large Language Models

Jason Wei¹

jasonwei@google.com

Yi Tay¹

yitay@google.com

Rishi Bommasani²

nlprishi@stanford.edu

Colin Raffel³

crffel@gmail.com

Barret Zoph¹

barretzoph@google.com

Sebastian Borgeaud⁴

sborgeaud@deepmind.com

Dani Yogatama⁴

dyogatama@deepmind.com

Maarten Bosma¹

bosma@google.com

Denny Zhou¹

dennyzhou@google.com

Donald Metzler¹

metzler@google.com

Ed H. Chi¹

edchi@google.com

Tatsunori Hashimoto²

thashim@stanford.edu

Oriol Vinyals⁴

vinyals@deepmind.com

Percy Liang²

pliang@stanford.edu

Jeff Dean¹

jeff@google.com

William Fedus¹

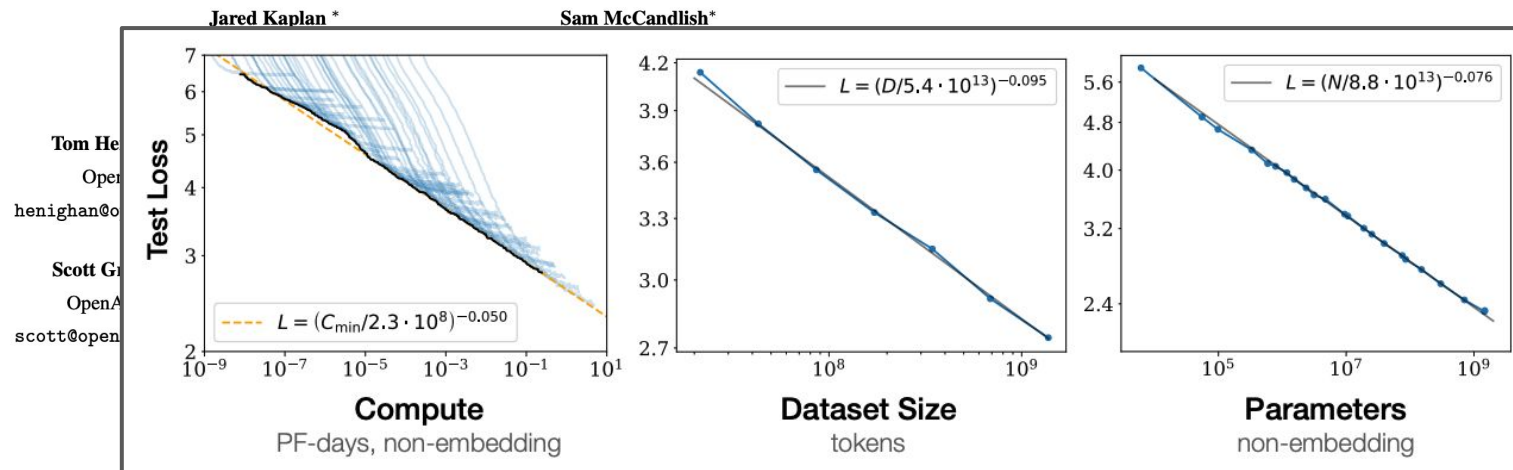
liamfedus@google.com

¹Google Research ²Stanford University ³UNC Chapel Hill ⁴DeepMind

Reviewed on OpenReview: <https://openreview.net/forum?id=yzkSU5zdwD>

Predictable gains as a result of scaling

Scaling Laws for Neural Language Models



Emergence in science

- Emergence:

“a qualitative change that arises from quantitative changes”

<https://bounded-regret.ghost.io/future-ml-systems-will-be-qualitatively-different/>

Bounded Regret | Home

Future ML Systems Will Be Qualitatively Different

JAN 11, 2022 · 7 MIN READ

In 1972, the Nobel prize-winning physicist Philip Anderson wrote the essay "[More is Different](#)". In it, he argues that quantitative changes can lead to qualitatively different and unexpected phenomena. While he focused on physics, one can find many examples of More is Different in other domains as well, including biology, economics, and computer science. Some examples of More is Different include:

- **Uranium.** With a bit of uranium, nothing special happens; with a large amount of uranium packed densely enough, you get a nuclear reaction.
- **DNA.** Given only small molecules such as calcium, you can't meaningfully encode useful information; given larger molecules such as DNA, you can encode a genome.
- **Water.** Individual water molecules aren't wet. Wetness only occurs due to the interaction forces between many water molecules interspersed throughout a fabric (or other material).
- **Traffic.** A few cars on the road are fine, but with too many you get a traffic jam. It could be that 10,000 cars could traverse a highway easily in 15 minutes, but 20,000 on the road at once could take over an hour.
- **Specialization.** Historically, in small populations, virtually everyone needed to farm or hunt to survive; in contrast, in larger and denser communities, enough food is produced for large fractions of the population to specialize in non-agricultural work.

While some of the examples, like uranium, correspond to a sharp transition, others like specialization are more continuous. I'll use **emergence** to refer to qualitative changes that arise from quantitative increases in scale, and **phase transitions** for cases where the change is sharp.

In this post, I'll argue that emergence often occurs in the field of AI, and that this should significantly affect our intuitions about the long-term development and deployment of AI systems. We should expect weird and surprising phenomena to emerge as we scale up systems. This presents opportunities, but also poses important risks.

E

o have to the
e rest of sci-
e of society.
thesis breaks
with the twin
plexity. The
mplex aggre-
cles, it turns
ood in terms
of the prop-
s. Instead, at
entirely new
e understand-
s requires re-
s fundamental
er. That is, it
ay array the
n a hierarchy,
he elementary
y the laws of

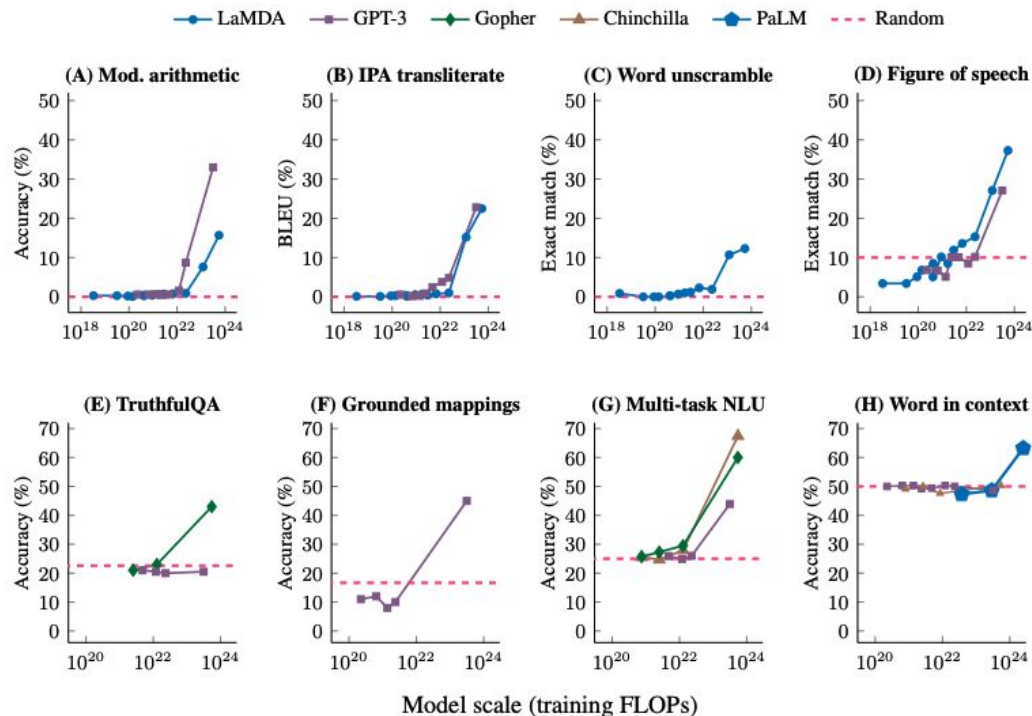
Y
nentary particle
physics
y-body physics
mistry
ecular biology
:
:
:
siology
chology

Emergent abilities in large language models definition

An ability is emergent if it is not present in smaller models but is present in larger models.

- How to measure the size of the model?
 - Training FLOPs
 - Number of model parameters
 - Training dataset size

Emergence in few-shot prompting



Few Shot Prompt and Predicted Answer

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?
(A) 75 (B) 76 (C) 22 (D) 23

Answer: B

Compute $i + i^2 + i^3 + \dots + i^{258} + i^{259}$.

(A) -1 (B) 1 (C) i (D) -i

Answer: B

Input (English): The 1931
Malay census was an alarm
bell.
If 4 d...
how n...
(A) 28
Answer

Target (IPA): ðə 1931
'meɪləɪ 'sensəs wəz ən
ə'lɑ:m bɛl.

Emergent prompting strategies

- Prompting techniques beyond few-shot prompting that improve the ability of language models
 - Only work at sufficient model scale

Finetune on many tasks (“instruction-tuning”)

Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.
How would you accomplish this goal?

OPTIONS:
-Keep stack of pillow cases in fridge.
-Keep stack of pillow cases in oven.

Target
keep stack of pillow cases in fridge

Input (Translation)

Translate this sentence to Spanish:
The new office building was built in less than three months.

Target
El nuevo edificio de oficinas se construyó en tres meses.

Sentiment analysis tasks

Coreference resolution tasks

...

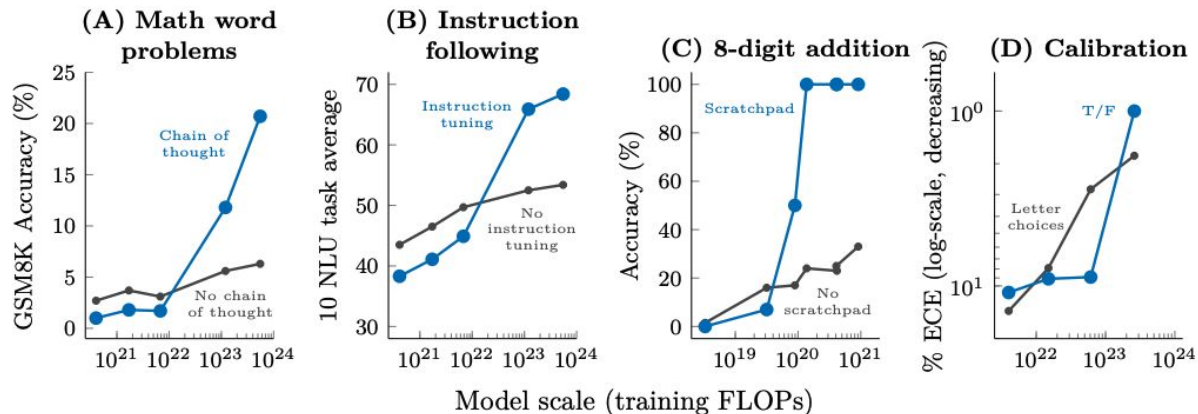
Inference on unseen task type

Input (Natural Language Inference)

Premise: At my age you will probably have learnt one lesson.
Hypothesis: It's not certain how many lessons you'll learn by your thirties.
Does the premise entail the hypothesis?

OPTIONS:
-yes -it is not possible to tell -no

FLAN Response
It is not possible to tell



Emergent abilities of large language model (TMLR '22).

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, & W. Fedus.

Implications

It is possible that if we continue scaling up, language models in the future could have even more abilities.

Emergent abilities of large language model (TMLR '22).

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, & W. Fedus.

Any questions?

Next: chain-of-thought prompting

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

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

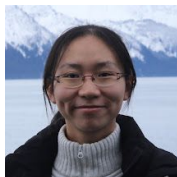
Abstract

We explore how generating a *chain of thought*—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning. In particular, we show how such reasoning abilities emerge naturally in sufficiently large language models via a simple method called *chain-of-thought prompting*, where a few chain of thought demonstrations are provided as exemplars in prompting.

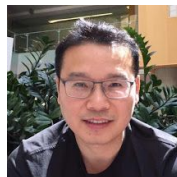
Experiments on three large language models show that chain-of-thought prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. For instance, prompting a PaLM 540B with just eight chain-of-thought exemplars achieves state-of-the-art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier.



Jason Wei



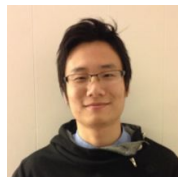
Xuezhi Wang



Denny Zhou



Nathaneal
Schärli



Le Hou



Nathan
Scales



Maarten
Bosma



Aakanksha
Chowdhery



Sharan
Narang



Brian Ichter



Fei Xia



Dale
Schuurmans



Olivier
Bousquet



Quoc Le



Ed Chi

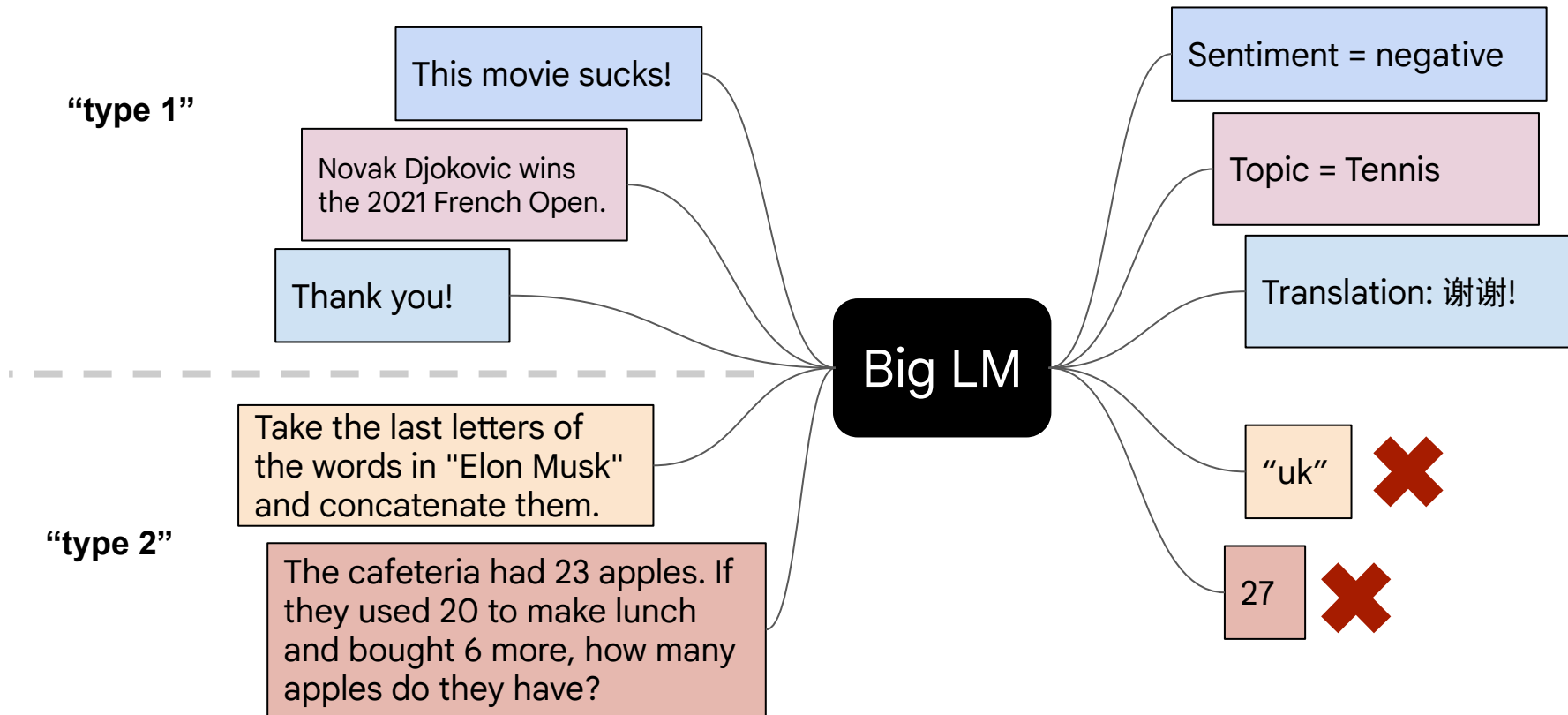


Google I/O 2022

Video:

<https://twitter.com/Google/status/1525188695875366912>

Motivation



Example of chain of thought

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?

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

Standard Prompting

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: 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 answer is 27. ❌

Chain of Thought Prompting

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?

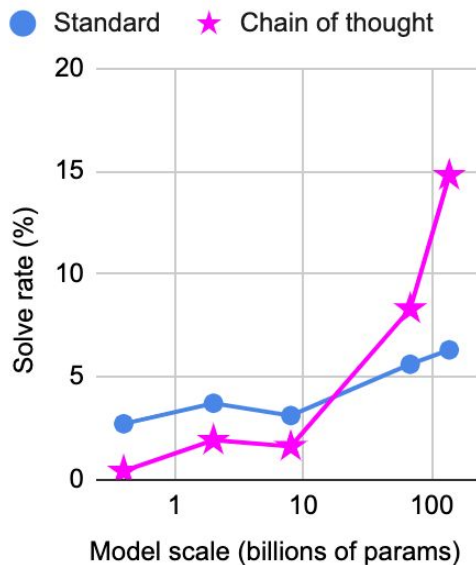
Model Output

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. ✅

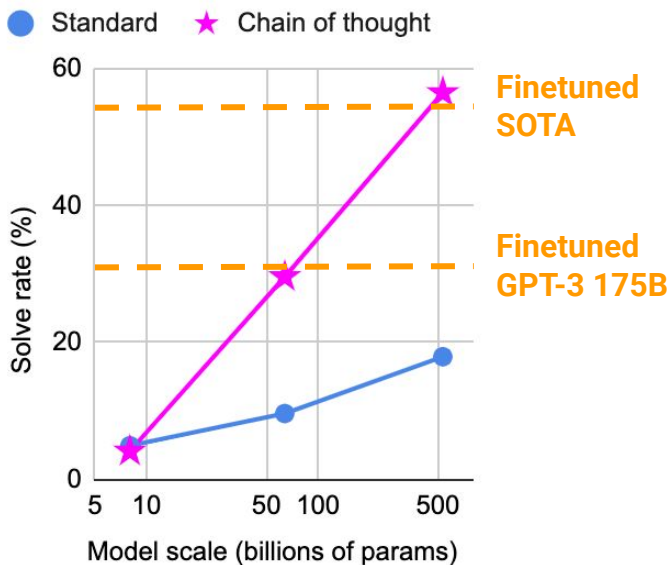
GPT-3 demo

Results on grade-school math problems

LaMDA (GSM8K)



PaLM

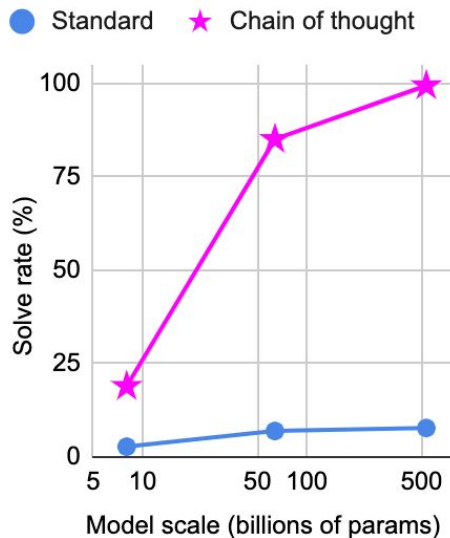


Last Letter Concatenation

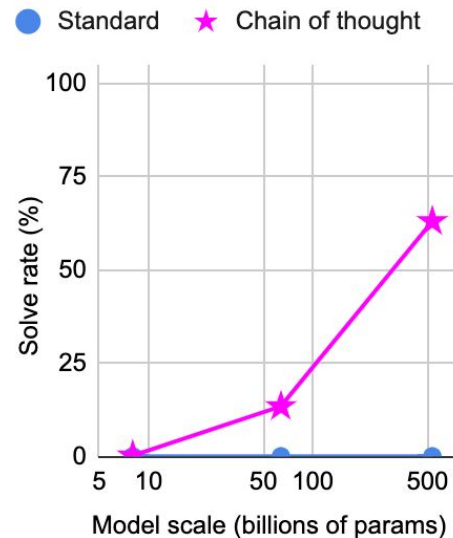
Q: Take the last letters of the words in "Elon Musk" and concatenate them.

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

"In domain" (2 letters)



OOD length generalization (4 letters)



StrategyQA

Real model output #1

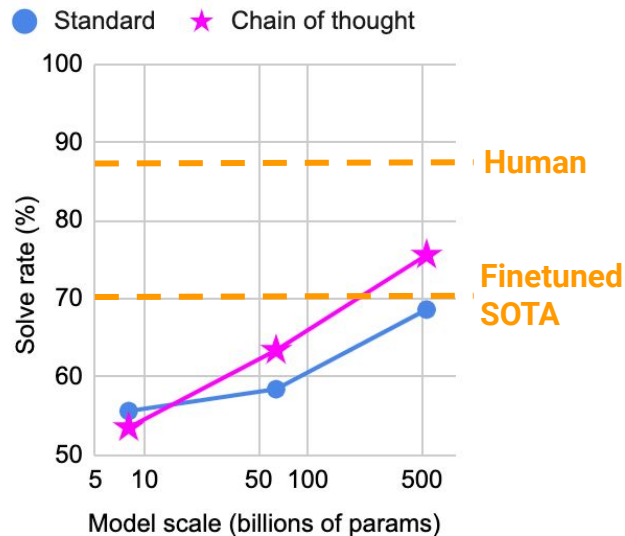
Q: Can you hide a basketball in a sand cat's ear?

A: A basketball is about 30 cm in diameter. A sand cat's ear is about 2 cm in diameter. Thus, a basketball would not fit in a sand cat's ear. So the answer is no.

Real model output #2

Q: Can oysters be used in guitar manufacturing?

A: Oysters can be used to make pearls. Pearls can be used to make guitar inlays. Thus, oysters can be used in guitar manufacturing. So the answer is yes.



Why chain of thought?

- 😊 Variable computation.
- 😊 Human-readable reasoning.
- 😊 Natural language is broadly applicable.
- 😊 Works with just prompting! →

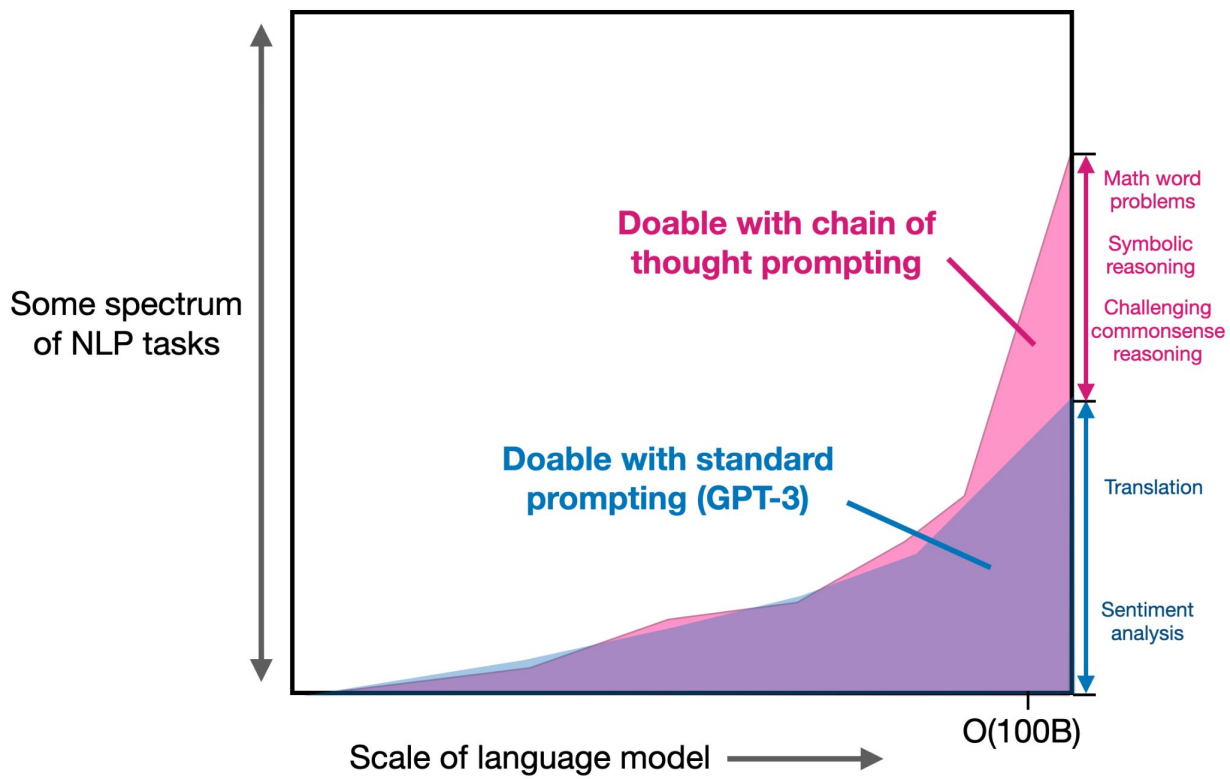
Before (finetuning)

- 😞 Large dataset with intermediate step annotations
- 😞 Many task-specific models

Now (prompting)

- 😊 Chain of thought annotations, but few-shot only
- 😊 Single model for all tasks

- Program induction by rationale generation: learning to solve and explain algebraic word problems ([ACL '17](#)).
- Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge ([NeurIPS '20](#)).
- Training verifiers to solve math word problems ([OpenAI '21](#)).
- More...



Any questions?

Next (if time): extensions to chain of thought prompting

SELF-CONSISTENCY IMPROVES CHAIN OF THOUGHT REASONING IN LANGUAGE MODELS

Xuezhi Wang^{†‡}, Jason Wei[†], Dale Schuurmans[†], Quoc Le[†], Ed H. Chi[†],
Sharan Narang[†], Aakanksha Chowdhery[†], Denny Zhou^{†§}

[†]Google Research, Brain Team

[‡]xuezhiw@google.com, [§]dennyzhou@google.com

ABSTRACT

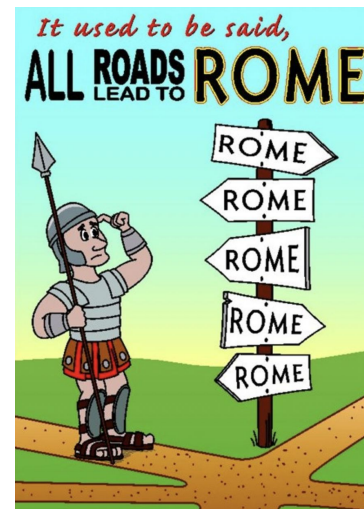
Chain-of-thought prompting combined with pre-trained large language models has achieved encouraging results on complex reasoning tasks. In this paper, we propose a new decoding strategy, *self-consistency*, to replace the naive greedy decoding used in chain-of-thought prompting. It first samples a diverse set of reasoning paths instead of only taking the greedy one, and then selects the most consistent answer by marginalizing out the sampled reasoning paths. Self-consistency leverages the intuition that a complex reasoning problem typically admits multiple different ways of thinking leading to its unique correct answer. Our extensive empirical evaluation shows that self-consistency boosts the performance of chain-of-thought prompting with a striking margin on a range of popular arithmetic and commonsense reasoning benchmarks, including GSM8K (+17.9%), SVAMP (+11.0%), AQuA (+12.2%), StrategyQA (+6.4%) and ARC-challenge (+3.9%).

Motivation

Given a question, do people always solve it in the same way?

Key ideas:

1. It is actually optional to generate the chain-of-thought, only the final answer is used for evaluation.
2. Ask the LM to generate a diverse set of possible solutions.
3. If some of them lead to the same answer, then it's more likely to be the correct answer.



Chain of thought with greedy decode

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: 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:

Prompt example with chain of thought

Greedy decode

Language model

This means she uses $3 + 4 = 7$ eggs every day. She sells the remainder for \$2 per egg, so in total she sells $7 * \$2 = \14 per day.
The answer is \$14.

- Wrong answer due to incorrect reasoning path
- Is there a way to self-check this solution?

Self-consistency: sample decode

Prompt with example chain of thought

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: 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 $\$2 * 9 =$

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 \text{ eggs} * \$2 = \18 .

The answer is \$18.

Note the reasoning paths are optional, so you don't have to look at them after getting the final answer

Self-consistency: majority vote

Prompt with example chain of thought

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: 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 $\$2 * 9 =$

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 \text{ eggs} * \$2 = \18 .

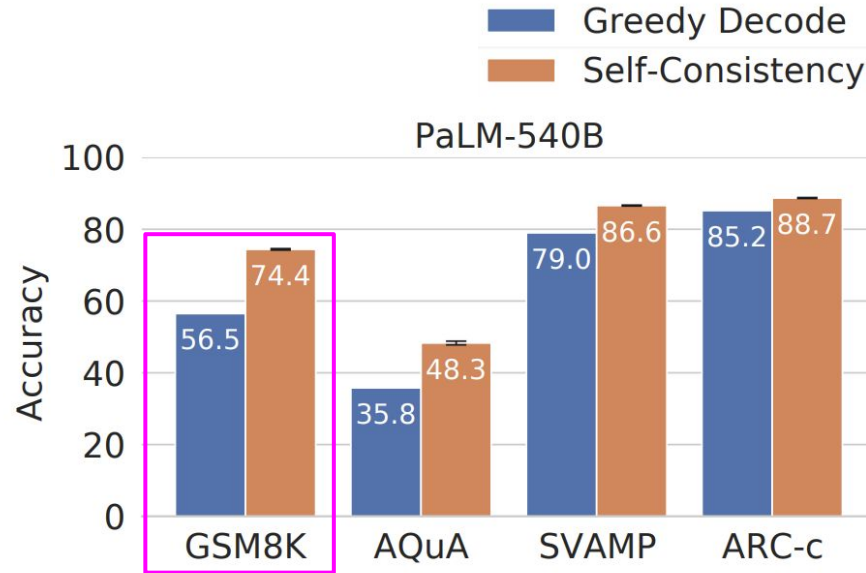
The answer is \$18.

Majority vote on the answers

The answer is \$18.

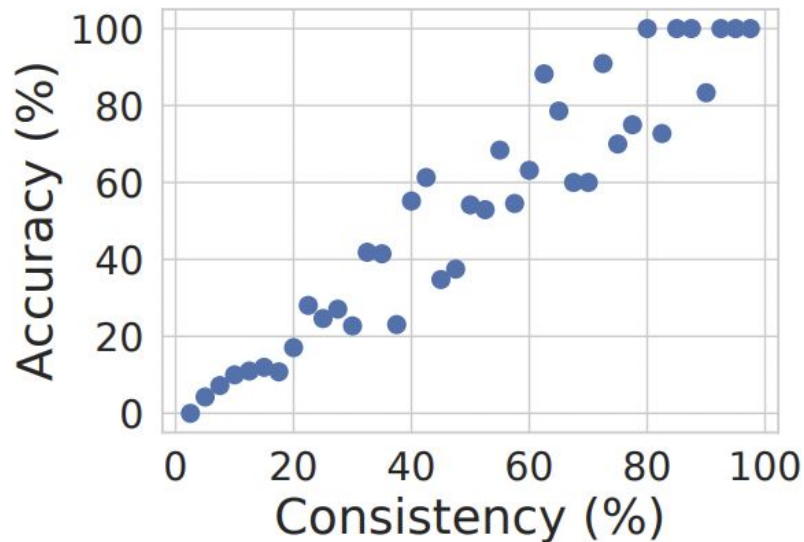
Self-consistency demo

Results



- When combined with PaLM-540B: beats SoTA on 8 out of 10 reasoning tasks
- Previous SoTA are from various papers with task-specific training, while self-consistency is a generic approach over the same frozen LM

Consistency \leftrightarrow Uncertainty

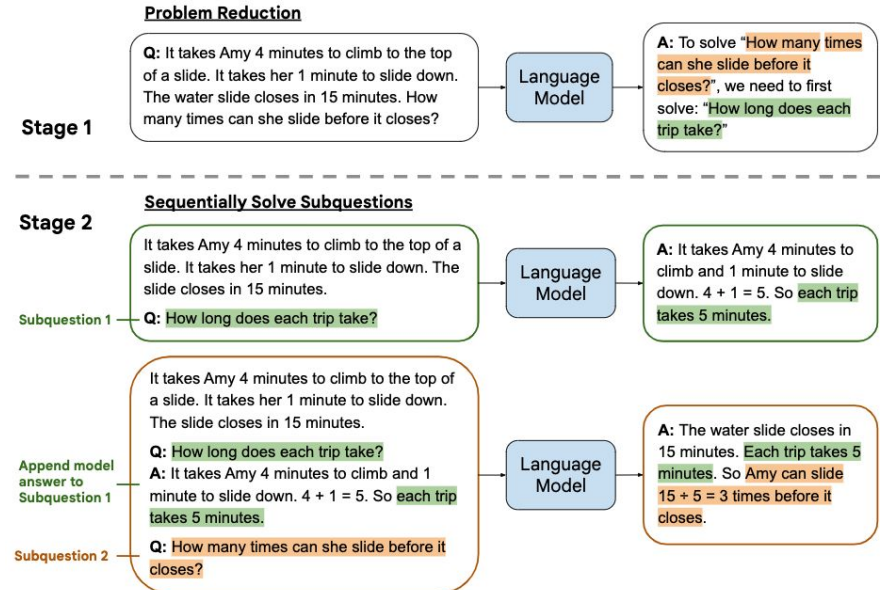


- Consistency: %decodes agreeing with the majority vote
- The consistency is nicely correlated with the accuracy
- May use to provide uncertainty estimates:
 - low “consistency” -> “uncertain”

Least-to-most prompting

Hierarchical prompting method

- Even more degrees of freedom



Least-to-most prompting

Command	Action Sequence
“look thrice after jump”	JUMP LOOK LOOK LOOK
“run left and walk”	TURN_LEFT RUN WALK
“look opposite right”	TURN_RIGHT TURN_RIGHT LOOK

Table 5: Example commands in SCAN and their corresponding action sequences. An agent successfully executes a natural language command by performing its corresponding action sequence.

Least-to-most prompting

Q: “jump right”

A: The output of “jump right” concatenates: the output of “turn right”, the output of “jump”. “turn right” outputs “TURN_RIGHT”. “jump” outputs “JUMP”. So concatenating the output of “turn right” and the output of “JUMP” leads to “TURN_RIGHT” + “JUMP”. So the output of “jump right” is “TURN_RIGHT” + “JUMP”.

Q: “jump right twice”

A: The output of “jump right twice” concatenates: the output of “jump right”, the output of “jump right”. “jump right” outputs “TURN_RIGHT” + “JUMP”. So repeating the output of “jump right” two times leads to (“TURN_RIGHT” + “JUMP”) * 2. So the output of “jump right twice” is (“TURN_RIGHT” + “JUMP”) * 2.

Q: “jump right twice after walk”

A: The output of “jump right twice after walk” concatenates: the output of “walk”, the output of “jump right twice”. “walk” outputs “WALK”. “jump right twice” outputs (“TURN_RIGHT” + “JUMP”) * 2. So concatenating the output of “walk” and the output of “jump right twice” leads to “WALK” + (“TURN_RIGHT” + “JUMP”) * 2. So the output of “jump right twice after walk” is “WALK” + (“TURN_RIGHT” + “JUMP”) * 2.

Table 8: A test case of least-to-most prompting on SCAN. Generated with code-davinci-002 of GPT-3. Note that the outputs of longer commands are built on the outputs of shorter commands.

Least-to-most prompting

Prompting method	<code>code-davinci-002</code>	<code>code-davinci-001</code>	<code>text-davinci-002*</code>
Standard prompting	16.7	0.4	6.0
Chain-of-Thought	16.2	0.0	0.0
Least-to-Most	99.7	60.7	76.0

Table 9: Accuracies (%) of different prompting methods on the test set of SCAN under the length-based split. The results of `text-davinci-002` are based on a random subset of 100 commands.

Multilingual chain-of-thought prompting

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?

Step-by-Step 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: 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 at the farmers' market daily for \$2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?

Step-by-Step Answer: Lays 16 eggs every day. Eat three eggs a day and make muffins with four eggs. So every day $16 - 3 - 4 = 9$ eggs are left. If the price of each egg is \$2, you earn $9 * 2 = \$18$ per day. The answer is 18.

Model Input

প্রশ্ন: রজারের 5টি টেনিস বল আছে। সে আরও 2 ক্যান টেনিস বল কিনেছে। প্রতিটি কানে 3টি করে টেনিস বল আছে। তার কাছে এখন কতগুলি টেনিস বল আছে?

ধাপে ধাপে উত্তর: রজারের প্রথমে 5টি বল ছিল। 2টি ক্যানের প্রতিটিতে 3টে টেনিস বল মানে 6টি টেনিস বল। $5 + 6 = 11$ । উত্তর হল 11।

প্রশ্ন: জেনেটের হাঁসগুলি প্রতিদিন 16টি করে ডিম পাড়ে। তিনি প্রতিদিন প্রাতরাশে তিনটি করে ডিম খান এবং বন্ধুদের জন্য প্রতিদিন চারটি ডিম দিয়ে মাফিন তৈরি করেন। অবশিষ্ট হাঁসের ডিমগুলি তিনি প্রতিদিন কৃষকদের বাজারে প্রতি ডিম \$2 দরে বিক্রয় করেন। তিনি কৃষকদের বাজারে প্রতিদিন কত ডলার উপার্জন করেন?

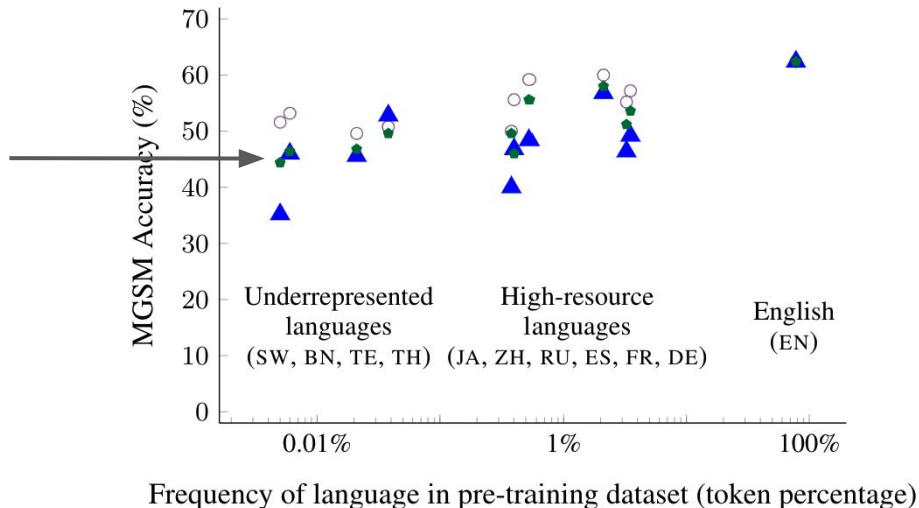
Model Output

ধাপে ধাপে উত্তর: প্রতিদিন 16টি ডিম পাড়ে। প্রতিদিন তিনটি ডিম খান এবং চারটি ডিম দিয়ে মাফিন তৈরি করেন। তাই প্রতিদিন $16 - 3 - 4 = 9$ টি ডিম অবশিষ্ট থাকে। প্রতিটি ডিমের মূল্য \$2 হলে প্রতিদিন $9 * 2 = 18$ ডলার উপার্জন করেন। উত্তর হল 18। ✓

Multilingual chain-of-thought prompting

- Translate to English with Google Translate and solve with English intermediate steps
- ▲ Intermediate reasoning steps in the language of the question
- Intermediate reasoning steps in English

Pretty high accuracy on math problems given and solved in underrepresented languages



Chain-of-thought on BIG-Bench tasks

Model Input (Navigate)

<Few shot chain-of-thought exemplars omitted for brevity>

Q: If you follow these instructions, do you return to the starting point? Turn left. Turn right. Take 5 steps. Take 4 steps. Turn around. Take 9 steps.

Options:

- Yes
- No

A: Let's think step by step.

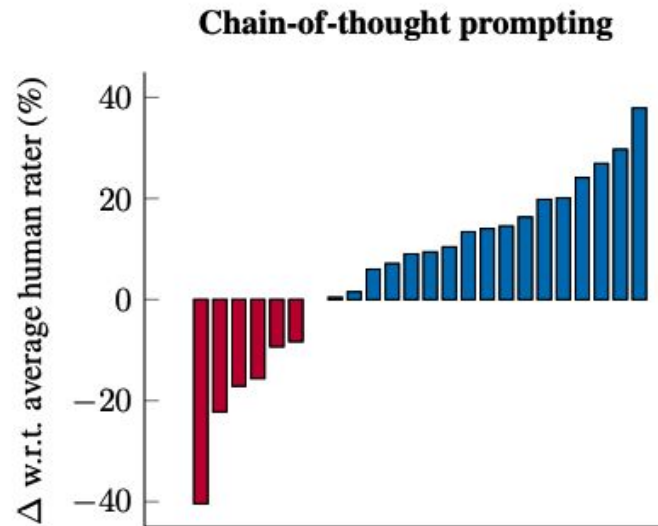
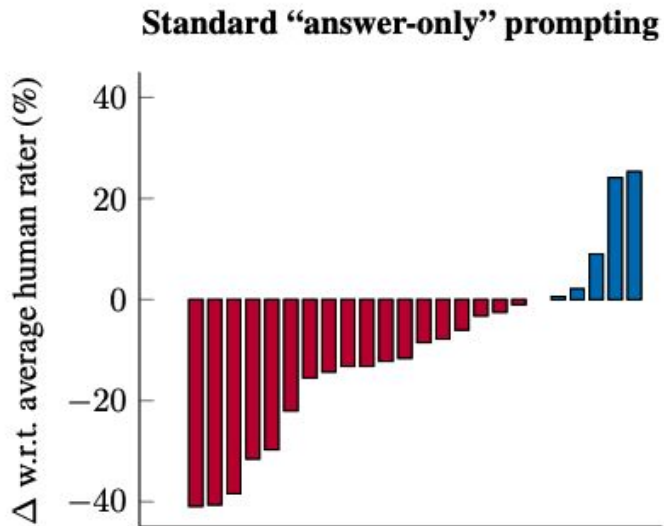
Model Output

We start at the origin (0, 0), facing the positive y-axis.

- (1) Turn left: (0, 0), facing the negative x-axis.
- (2) Turn right: (0, 0), facing the positive y-axis.
- (3) Take 5 steps: (0, 5), facing the positive y-axis.
- (4) Take 4 steps: (0, 9), facing the positive y-axis.
- (5) Turn around: (0, 9), facing the negative y-axis.
- (6) Take 9 steps: (0, 0), facing the negative y-axis.

Since (0, 0) is (0, 0), we are indeed where we started. So the answer is **Yes**. ✓

Chain-of-thought on BIG-Bench tasks



Looking forward (just my personal interests)

- Scaling
- Better prompting and characterization of language model abilities
- Applied work (therapy, creative writing, science)
- Benchmarks
- Compute-efficient methods for better language models

Thanks.

jasonwei@google.com

We're hiring interns!