

# Some intuitions about large language models

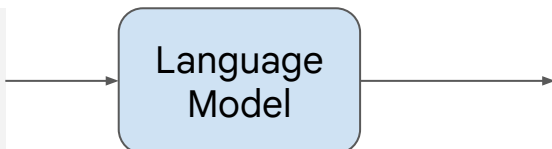
Jason Wei  
March 1, 2023

# Outline

- [10m] What do language models learn from predicting the next word?
  - Answer: a lot of stuff
- [10m] Why have language models taken off recently?
  - Answer: scaling, ...
- [10m] How can you get language models to do more complicated tasks?
  - Answer: intermediate reasoning
- [10m] How do we make language models more user-friendly?
  - Answer: adapt them to user-friendly inputs
- Feel free to interrupt with questions :)

# Language modeling objective

Dartmouth students like  
to \_\_\_\_



Most basic form  
(not chatGPT)

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           | 0.00001            |
| aardvark    | 0.000004           |
| ...         |                    |
| drink       | 0.5                |
| ...         |                    |
| study       | 0.23               |
| ...         |                    |
| zucchini    | 0.000002           |

(hypothetical)

# [1/8] What do language models learn from next word prediction?

→ Grammar

In my free time, I like to \_\_\_

Language Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| banana      | 0.00001            |
| ...         |                    |
| run         | 0.7                |
| ...         |                    |
| zucchini    |                    |

The next word is probably not a noun

The next word is probably a verb

(hypothetical)

[2/8] What do language models learn from next word prediction?

→ Facts about the world

The capital of Denmark is \_\_\_\_

Language Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| Copenhagen  | 0.9                |
| ...         |                    |
| London      | 0.05               |
| ...         |                    |
| zucchini    |                    |

Associations between words!

(hypothetical)

# [3/8] What do language models learn from next word prediction?

→ Lexical semantics

I went to the  
zoo to see  
giraffes, lions,  
and

Language  
Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| spoon       | 0.00001            |
| ...         |                    |
| zebras      | 0.6                |
| ...         |                    |
| zucchini    |                    |

The next word  
is probably  
related to  
giraffes and  
lions

(hypothetical)

# [4/8] What do language models learn from next word prediction?

→ Sentiment analysis

I was engaged  
and on the edge  
of my seat the  
whole time. The  
movie was \_\_\_\_

Language  
Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| bad         | 0.1                |
| ...         |                    |
| good        | 0.9                |
| ...         |                    |
| zucchini    |                    |

Well, "engaged" is pretty indicative of a positive sentiment

(hypothetical)

# [5/8] What do language models learn from next word prediction?

→ Harder sentiment analysis

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was —

Language Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| bad         | 0.7                |
| ...         |                    |
| good        | 0.3                |
| ...         |                    |
| zucchini    |                    |

Some more-complex understanding needed

(hypothetical)



# [6/8] What do language models learn from next word prediction?

→ Translation

The word for  
"pretty" in  
Spanish is \_\_\_\_

Language  
Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| bonita      | 0.8                |
| ...         |                    |
| hola        | 0.03               |
| ...         |                    |
| zucchini    |                    |

Understanding  
of multiple  
languages

(hypothetical)

[7/8] What do language models learn from next word prediction?

→ Spatial reasoning

Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the \_\_\_\_

Language Model

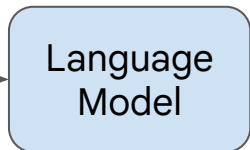
| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
|             |                    |
| ...         |                    |
| kitchen     | 0.8                |
| ...         |                    |
| zucchini    |                    |

(hypothetical)

# [8/8] What do language models learn from next word prediction?

→ Easy arithmetic

3 + 8 + 4 = \_\_\_\_



| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| 14          | 0.1                |
| 15          | 0.7                |
| ...         |                    |
| zucchini    |                    |

(hypothetical)

Understanding (or memorization) or addition?

[1/6] What can't language models do from next word prediction?

→ Not current world knowledge

The stock price of APPL on March 1st, 2023 is \_\_\_\_

Language Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| ?           | ?                  |
| ...         |                    |
| zucchini    |                    |

Language model was trained before that date

[2/6] What can't language models learn from next word prediction?

→ Not arbitrarily long arithmetic

36382894730 +  
238302849204 = \_\_\_\_

Language  
Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| ?           | ?                  |
| ...         |                    |
| zucchini    |                    |

Not enough  
similar training  
data and also  
hard to figure out  
the pattern

# [3/6] What can't language models do from next word prediction?

→ Math questions you probably wouldn't be able to do

Take the nineteenth digit of Pi and multiply it by the e to the fourth power. The resulting ones-digit of the resulting number is \_\_\_\_

Language Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| ?           | ?                  |
| ...         |                    |
| zucchini    |                    |

A lot of intermediate steps so hard to figure out the pattern

# [4/6] What can't language models do from next word prediction?

→ Predict the future

The winner of the  
FIFA world cup in  
2026 is \_\_\_\_

Language  
Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| ?           | ?                  |
| ...         |                    |
| zucchini    |                    |

Nobody can do  
this

[5/6] What can't language models do from next word prediction?

→ Information that isn't in the training data

Jason Wei's favorite  
color is \_\_\_\_

Language  
Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| ?           | ?                  |
| ...         |                    |
| zucchini    |                    |

That information  
is probably not  
seen in its training  
data



## [6/6] What can't language models do from next word prediction?

→ Can't take in extremely long pieces of text

[2,000 page Harry Potter fan-fiction]  
What happened after Harry opened the chest for the second time? \_\_\_\_

Language Model

| <u>Word</u> | <u>Probability</u> |
|-------------|--------------------|
| a           |                    |
| ...         |                    |
| ?           | ?                  |
| ...         |                    |
| zucchini    |                    |

Can't take in very long inputs (limited to like 4k words-ish)

Rule of thumb: language models can do  
(with decent accuracy) most things that an  
average human can do in 1 minute.



2018

Today (2023)

Future ...?

|                          |
|--------------------------|
| ...                      |
| Protein discovery        |
| Clinical diagnosis       |
| Play chess well          |
| High-level planning      |
| Abstract reasoning       |
| Simple math              |
| Commonsense reasoning    |
| Know world knowledge     |
| Translation              |
| Sentiment analysis       |
| Generate coherent text   |
| Be grammatically correct |

|                          |
|--------------------------|
| ...                      |
| Protein discovery        |
| Clinical diagnosis       |
| Play chess well          |
| High-level planning      |
| Abstract reasoning       |
| Simple math              |
| Commonsense reasoning    |
| Know world knowledge     |
| Translation              |
| Sentiment analysis       |
| Generate coherent text   |
| Be grammatically correct |

|                          |
|--------------------------|
| ...                      |
| (?) Protein discovery    |
| (?) Clinical diagnosis   |
| (?) Play chess well      |
| (?) High-level planning  |
| (?) Abstract reasoning   |
| Simple math              |
| Commonsense reasoning    |
| Know world knowledge     |
| Translation              |
| Sentiment analysis       |
| Generate coherent text   |
| Be grammatically correct |

## Emergent Abilities of Large Language Models

Jason Wei<sup>1</sup>

*jasonwei@google.com*

Yi Tay<sup>1</sup>

*yitay@google.com*

Rishi Bommasani<sup>2</sup>

*nlprishi@stanford.edu*

Colin Raffel<sup>3</sup>

*crffel@gmail.com*

Barret Zoph<sup>1</sup>

*barretzoph@google.com*

Sebastian Borgeaud<sup>4</sup>

*sborgeaud@deepmind.com*

Dani Yogatama<sup>4</sup>

*dyogatama@deepmind.com*

Maarten Bosma<sup>1</sup>

*bosma@google.com*

Denny Zhou<sup>1</sup>

*dennyzhou@google.com*

Donald Metzler<sup>1</sup>

*metzler@google.com*

Ed H. Chi<sup>1</sup>

*edchi@google.com*

Tatsunori Hashimoto<sup>2</sup>

*thashim@stanford.edu*

Oriol Vinyals<sup>4</sup>

*vinyals@deepmind.com*

Percy Liang<sup>2</sup>

*pliang@stanford.edu*

Jeff Dean<sup>1</sup>

*jeff@google.com*

William Fedus<sup>1</sup>

*liamfedus@google.com*

<sup>1</sup>Google Research <sup>2</sup>Stanford University <sup>3</sup>UNC Chapel Hill <sup>4</sup>DeepMind

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

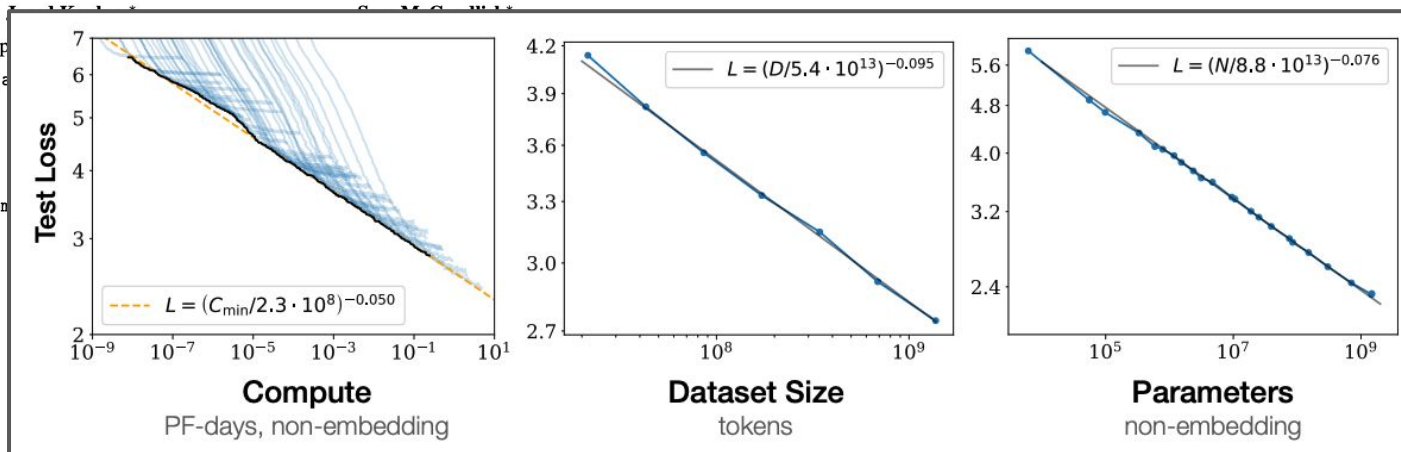
# Predictable gains as a result of scaling

## Scaling Laws for Neural Language Models

Johns Hopkins University  
jeff@cs.jhu.edu

**Tom Henighan**  
OpenAI  
henighan@openai.com

**Scott Gray**  
OpenAI  
scott@openai.com



# Emergence in science

- Emergence: *“a qualitative change that arises from quantitative changes”*

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

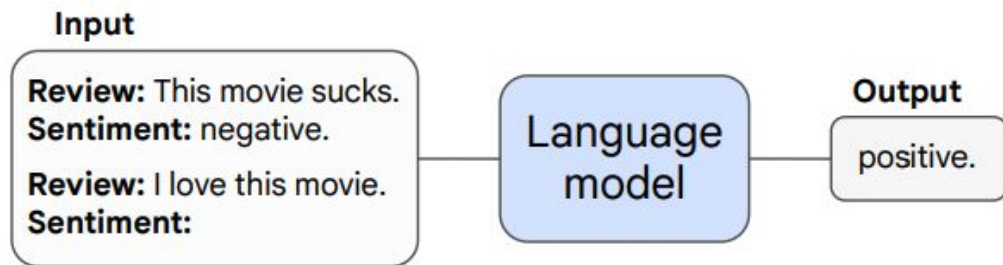
[Jacob Steinhardt, 2022.](#)

# Definition: *emergent abilities* in large language models

*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

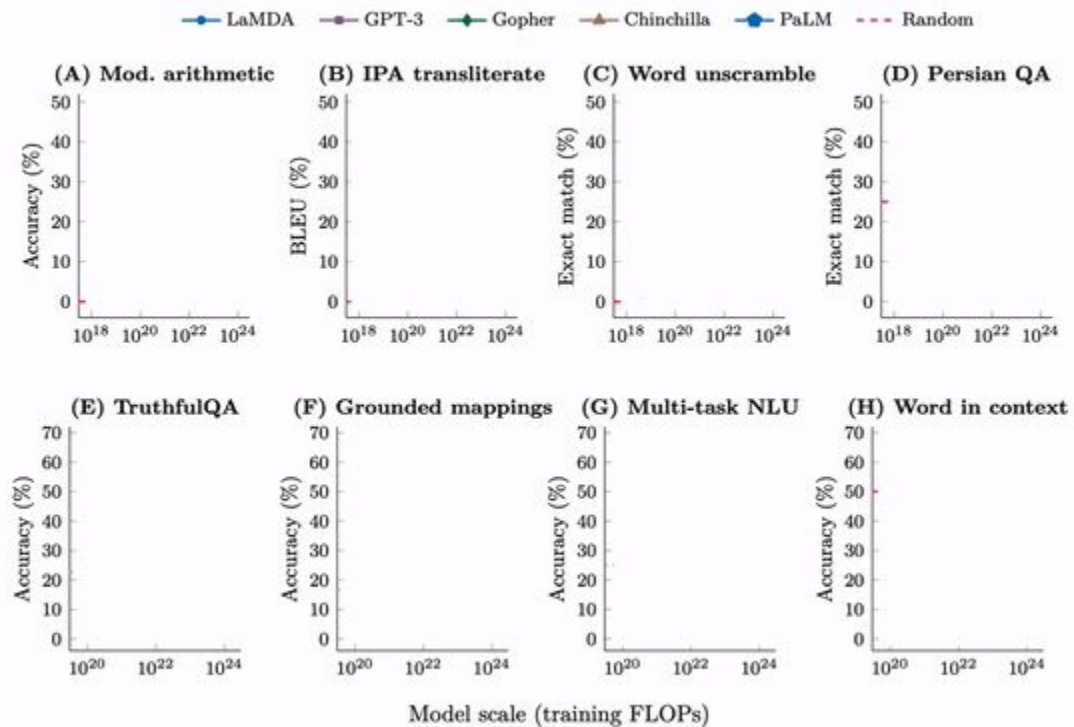
# Emergence in few-shot prompting



*> A few-shot prompted task is emergent if it achieves random accuracy for small models and above-random accuracy for large models.*



# Emergence in few-shot prompting

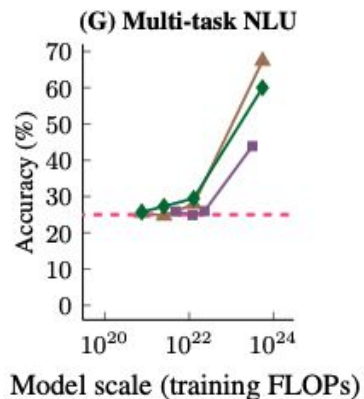


# Emergence in few-shot prompting

|                       |  |   |
|-----------------------|--|---|
| <b>Microeconomics</b> | One of the reasons that the government discourages and regulates monopolies is that      |   |
|                       | (A) producer surplus is lost and consumer surplus is gained.                             | ✗ |
|                       | (B) monopoly prices ensure productive efficiency but cost society allocative efficiency. | ✗ |
|                       | (C) monopoly firms do not engage in significant research and development.                | ✗ |
|                       | (D) consumer surplus is lost with higher prices and lower levels of output.              | ✓ |

|                            |   |   |
|----------------------------|---|---|
| <b>College Mathematics</b> | In the complex $z$ -plane, the set of points satisfying the equation $z^2 =  z ^2$ is a |   |
|                            | (A) pair of points  | ✗ |
|                            | (B) circle  | ✗ |
|                            | (C) half-line   | ✗ |
|                            | (D) line  | ✓ |

[Hendryks et al., 2020.](#)

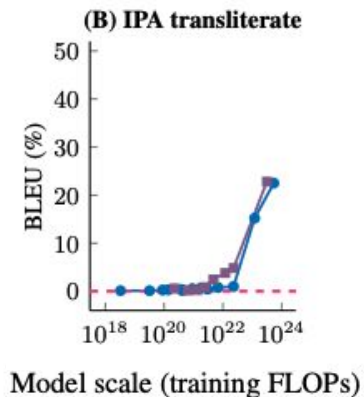


# Emergence in few-shot prompting

Input (English): The 1931  
Malay census was an alarm  
bell.

Target (IPA): ðə 1931 'meɪleɪ  
'sɛnsəs wəz ən ə'lɑːm bɛl.

BIG-Bench ([Srivastava et al., 2022](#)).

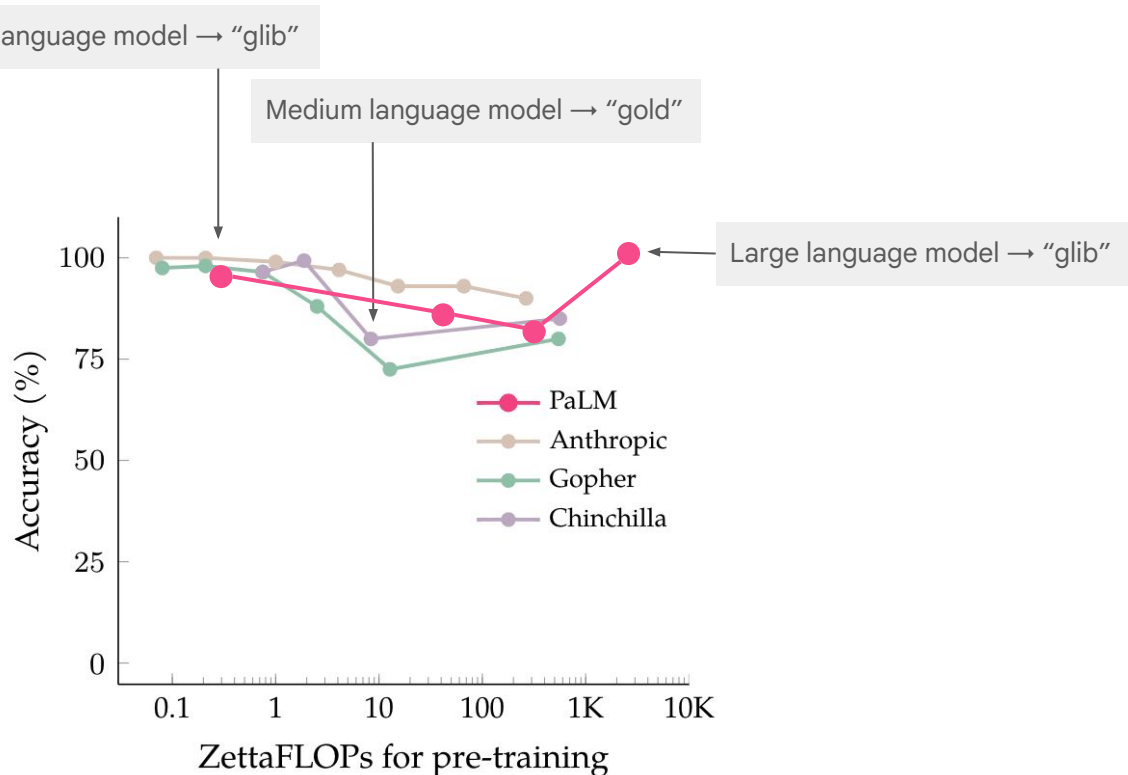


# Inverse scaling can become U-shaped

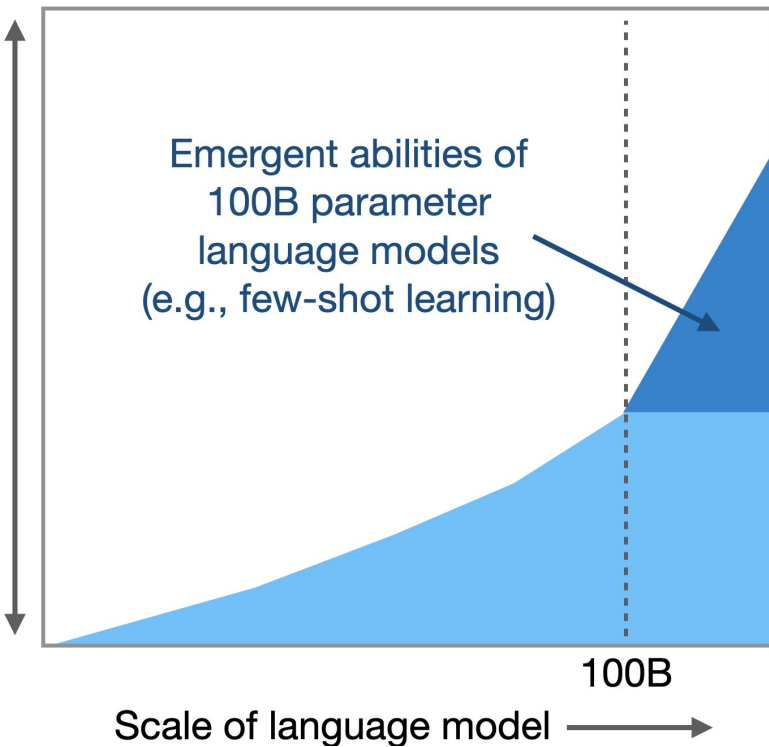
**Quote Repetition**

**Input**  
Repeat my sentences back to me.  
Input: All that glisters is not glib  
Output: All that glisters is not

**Target**  
glib



“Things that  
language models  
can do”



# CoT paper

## Motivation:

- Enable language models to do more-complicated tasks
- Guide them with “meta-data” (i.e., reasoning process)
- Prompts are manually composed (prompt engineering helps)

### 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?

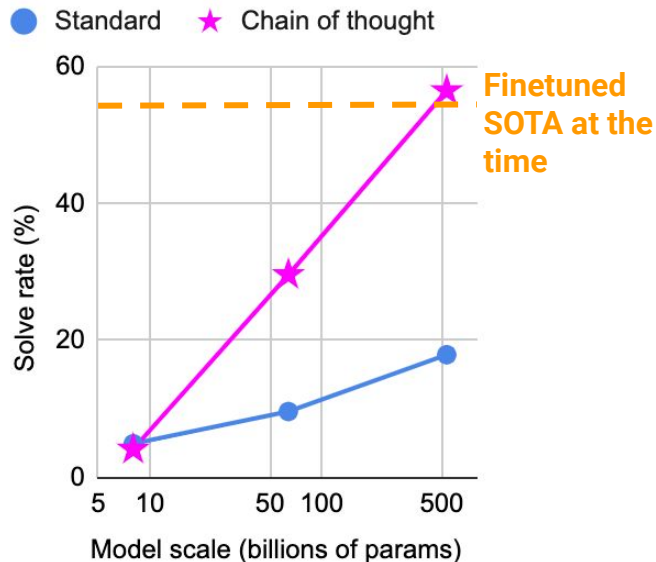
#### Model Output

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. ✅

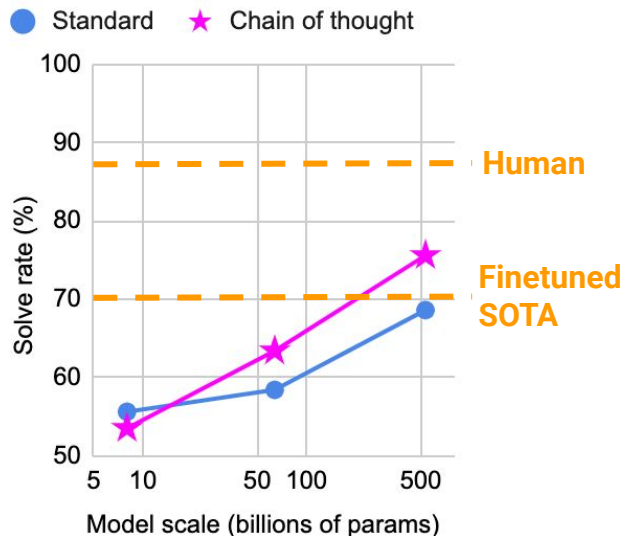
CoT demo

# CoT paper

## GSM8K



## StrategyQA



### Real model output #1

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

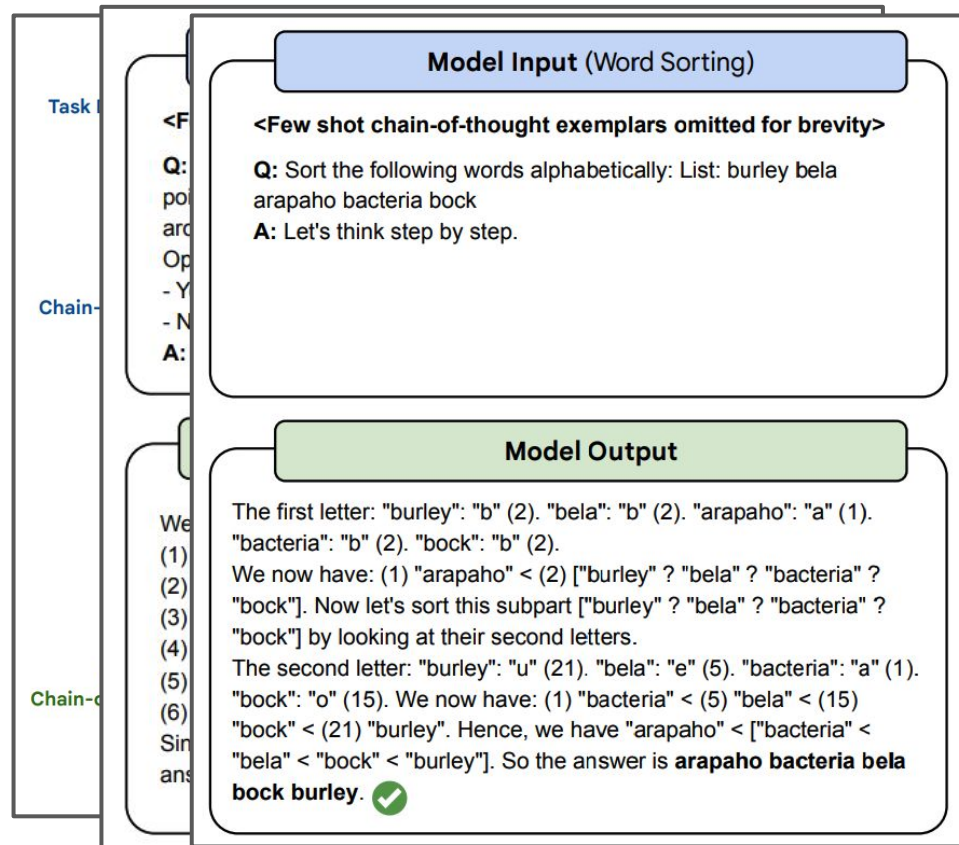
**A:** A basketball is about 30 cm in diameter. A sand cat's ear is about 2 cm in diameter. Thus, a basketball would not fit in a sand cat's ear. So the answer is no.



# CoT on BIG-Bench: Benchmark

## BIG-Bench Hard (BBH):

- 23 challenging tasks from BIG-Bench benchmark where no model beats avg. human rater



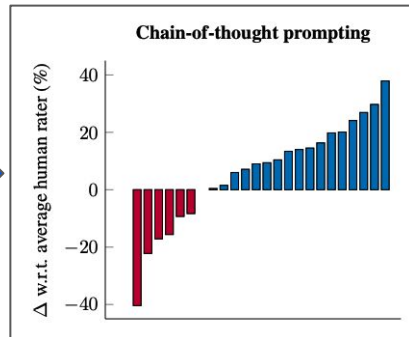
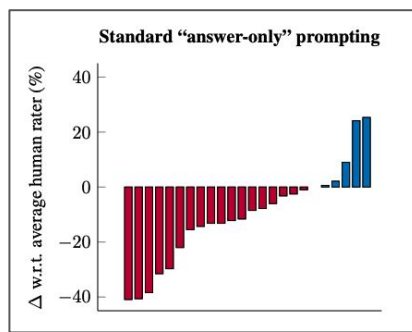
# CoT on BIG-Bench: Result summary

|                             | BBH all<br>(23 tasks) | # tasks above<br>avg. human-rater |
|-----------------------------|-----------------------|-----------------------------------|
| Average human-rater         | 67.7                  | N/A                               |
| Max human-rater             | 94.4                  | 23 / 23                           |
| Best prior BIG-Bench result | 50.9                  | 0 / 23                            |
| Codex (code-davinci-002)    |                       |                                   |
| - Answer-only prompting     | 56.6                  | 5 / 23                            |
| - CoT prompting             | 73.9 (+16.7)          | 17 / 23                           |

Model much lower than average human rater

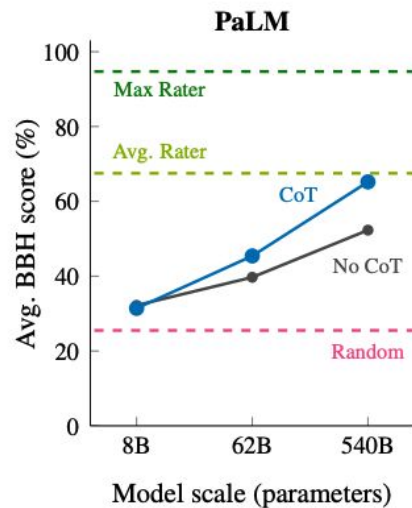
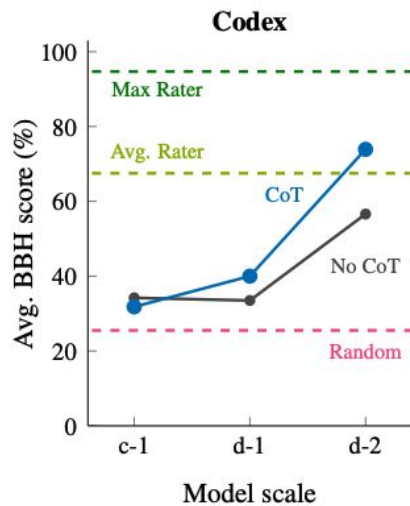
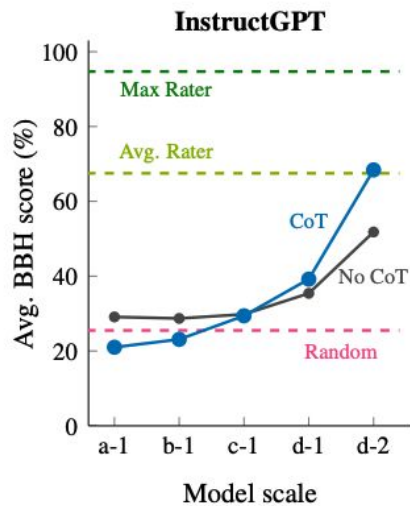
Detail: better formatting (options, task description) already beats prior best

CoT prompting improves by performance by +16.7%, passes avg. human on majority of tasks



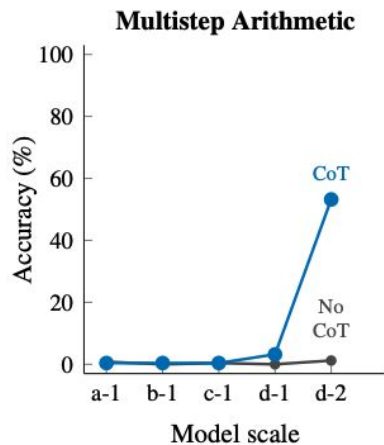
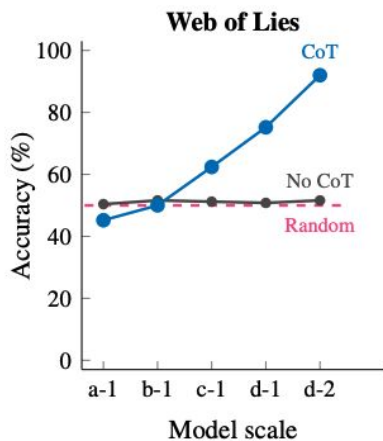
# CoT on BIG-Bench: Scaling

- CoT requires sufficient model scale for positive delta
- On aggregate, threshold is davinci-002 / PaLM 62B



# CoT on BIG-Bench: Emergence

- No-CoT performance is flat, i.e., hasn't unlocked emergence \_yet\_ ;)
- CoT unlocks emergent performance



**Model Input (Multistep Arithmetic)**

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

**Q:**  $((4 + 7 * 4 - -5) - (-4 - 1 - -4 - 4)) =$   
**A:** Let's think step by step.

**Model Output**

Let's recall that the order of operations in mathematics is as follows: (1) Parentheses, (2) exponents, (3) multiplication and division (from left to right), (4) addition and subtraction (from left to right). So, remember to always compute the expressions inside parentheses or brackets first.

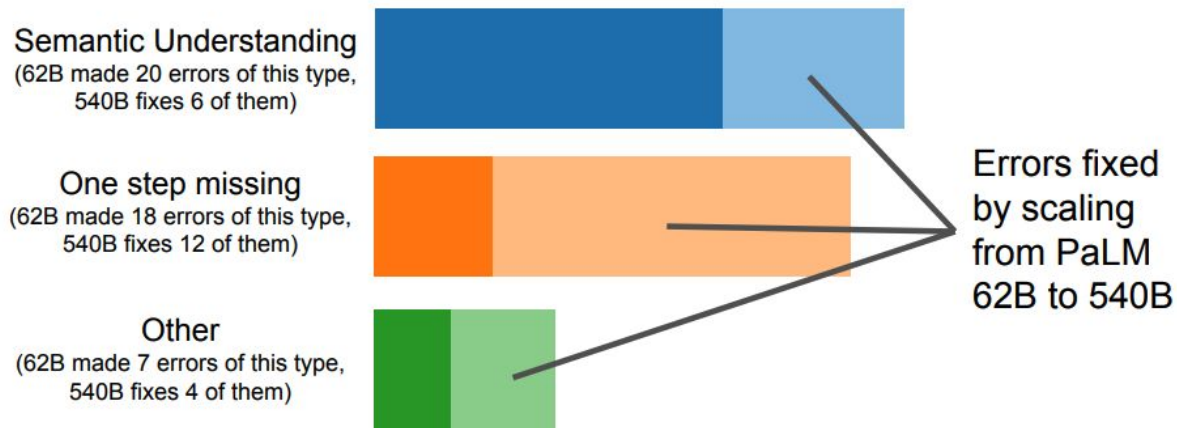
This equation can be written as "A - B", where A =  $(4 + 7 * 4 - -5)$  and B =  $(-4 - 1 - -4 - 4)$ .

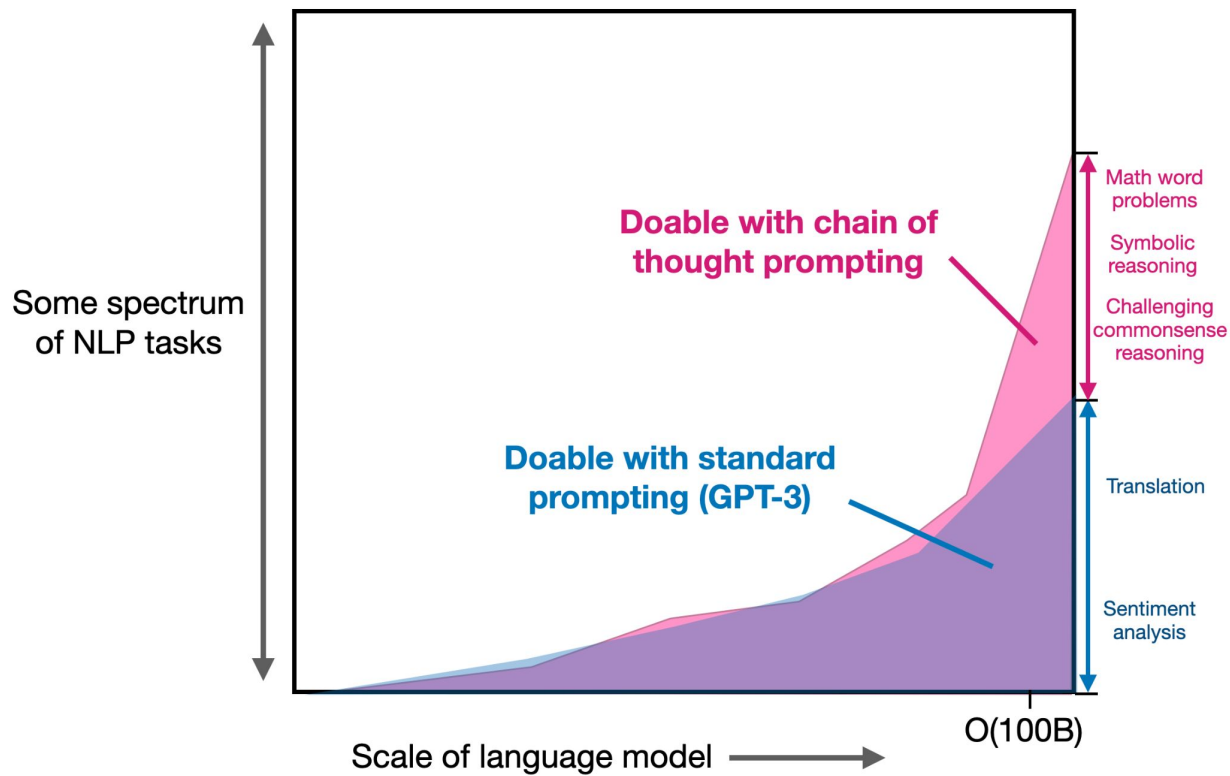
Let's calculate A =  $(4 + 7 * 4 - -5) = (4 + (7 * 4) - -5) = (4 + 28) - -5) = (4 + 28 - -5) = (4 + 28 + 5) = 37$ .

Let's calculate B =  $(-4 - 1 - -4 - 4) = ((-4 - 1) - -4 - 4) = ((-5) - -4 - 4) = ((-5 - -4) - 4) = ((-5 + 4) - 4) = (-1 - 4) = -5$ .

Then, the final equation is A - B =  $37 - -5 = 37 + 5 = 42$ . So the answer is **42**. ✓

# Why does scaling up improve chain-of-thought?





# 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.**

# 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

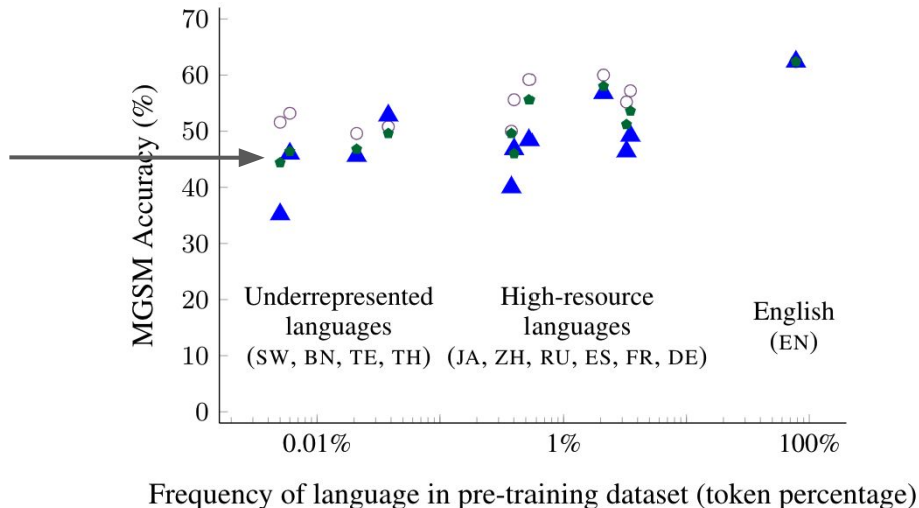
ধাপে ধাপে উত্তর: প্রতিদিন 16টি ডিম পাড়ে। প্রতিদিন তিনটি ডিম খান এবং চারটি ডিম দিয়ে মাফিন তৈরি করেন। তাই প্রতিদিন  $16 - 3 - 4 = 9$ টি ডিম অবশিষ্ট থাকে। প্রতিটি ডিমের মূল্য \$2 হলে প্রতিদিন  $9 * 2 = 18$  ডলার উপার্জন করেন। উত্তর হল 18। ✓



# 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

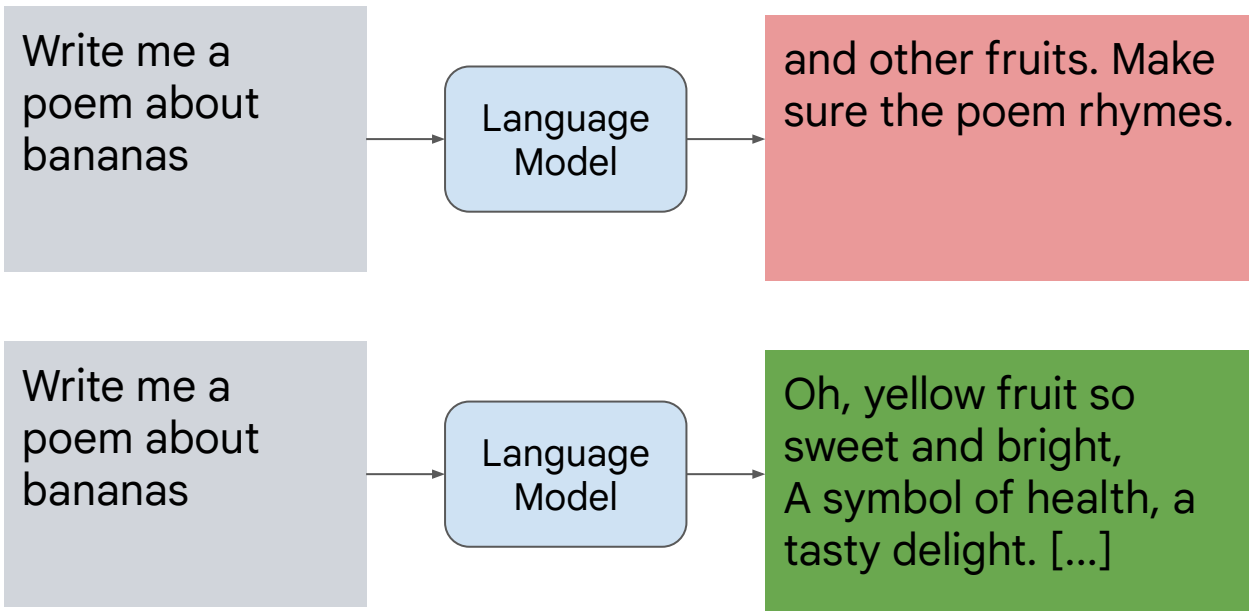
Pretty high accuracy on math problems given and solved in underrepresented languages



# Chain-of-thought: Discussion

- Framework for “more-complicated” prompting
  - What’s the best way to get a language model to do a task? Few-shot prompting is kinda thinking by analogy from machine learning on  $(x, y)$  pairs
- Limitation: Few-shot CoT is task-specific and requires the prompt engineer
- Given explosion of tasks solved by LMs, we should be more open-minded about what tasks will be solved in next 1-2 years

# How do we make language models more user-friendly?



## Scaling Instruction-Finetuned Language Models

Hyung Won Chung<sup>\*</sup> Le Hou<sup>\*</sup> Shayne Longpre<sup>\*</sup> Barret Zoph<sup>†</sup> Yi Tay<sup>†</sup>  
William Fedus<sup>‡</sup> Eric Li Xuezhi Wang Mostafa Dehghani Siddhartha Brahma  
Albert Webson Shixiang Shane Gu Zhuyun Dai Mirac Suzgun Xinyun Chen  
Aakanksha Chowdhery Sharan Narang Gaurav Mishra Adams Yu Vincent Zhao  
Yanping Huang Andrew Dai Hongkun Yu Slav Petrov Ed H. Chi  
Jeff Dean Jacob Devlin Adam Roberts Denny Zhou Quoc V. Le  
Jason Wei<sup>\*</sup>

Google

Abstract

Finetuning language models on a collection of datasets phrased as instructions has been shown to improve model performance and generalization to unseen tasks. In this paper we explore instruction finetuning with a particular focus on (1) scaling the number of tasks, (2) scaling the model size, and (3) finetuning on chain-of-thought data. We find that instruction finetuning with the above aspects dramatically improves performance on a variety of model classes (PaLM, T5, U-PaLM), prompting setups (zero-shot, few-shot, CoT), and evaluation benchmarks (MMLU, BBH, TyDiQA, MGSM, open-ended generation). For instance, Flan-PaLM 540B instruction-finetuned on 1.8K tasks outperforms PaLM 540B by a large margin (+9.4% on average). Flan-PaLM 540B achieves state-of-the-art performance on several benchmarks, such as 75.2% on five-shot MMLU. We also publicly release Flan-T5 checkpoints,<sup>1</sup> which achieve strong few-shot performance even compared to much larger models, such as PaLM 62B. Overall, instruction finetuning is a general method for improving the performance and usability of pretrained language models.

v2 [cs.LG] 21 Oct 2022



Flan-T5-XXL (11B)

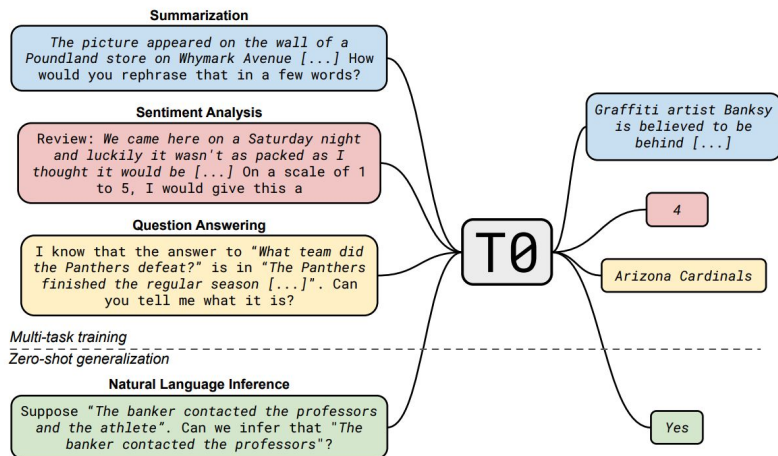


Flan-U-PaLM-540B



Flan-PaLM-540B

# Instruction finetuning



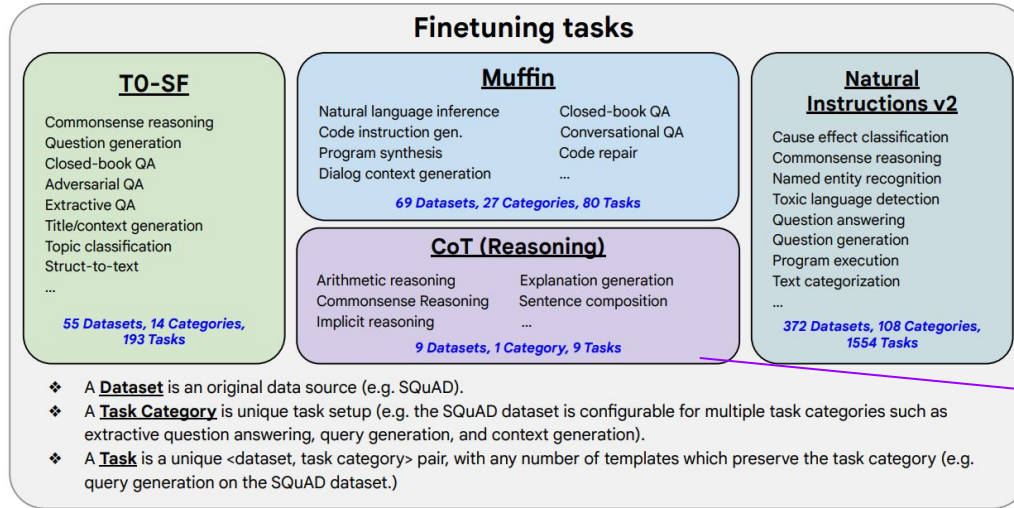
|                                 | # Tasks     | Largest model size (# params) |
|---------------------------------|-------------|-------------------------------|
| TO (Sanh et al., 2021)          | 170         | 11B                           |
| Tk-Instruct (Wang et al., 2022) | 1.5K        | 11B                           |
| FLAN-1 (Wei et al., 2021)       | 62          | 137B                          |
| <b>Flan (this paper)</b>        | <b>1.8K</b> | <b>540B</b>                   |

Multi-task prompted training enables zero-shot task generalization. Sanh et al., 2021.

Benchmarking generalization via in-context instructions on 1,600+ language tasks. Wang et al., 2022.

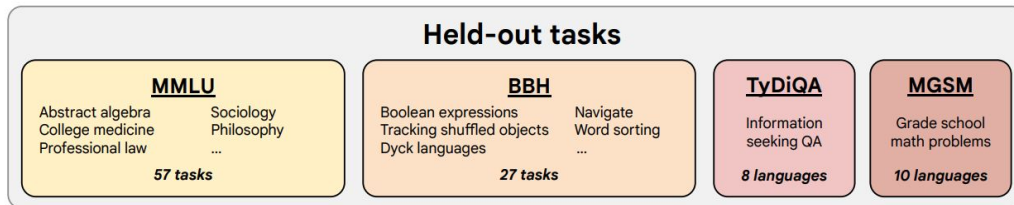
Finetuned language models are zero-shot learners. Wei et al., 2021.

# Scaling finetuning tasks



Task collections from prior work are combined.

Dataset with chain-of-thought annotations are included to enable reasoning.



Held-out tasks are challenging and not included in the finetuning datasets.

# Instructional templates

We train on examples without and without exemplars, and with and without chain-of-thought.

Instruction without exemplars

Without chain-of-thought

Answer the following yes/no question.  
Can you write a whole Haiku in a single tweet? → yes

Instruction with exemplars

Q: Answer the following yes/no question.  
Could a dandelion suffer from hepatitis?  
A: no  
Q: Answer the following yes/no question.  
Can you write a whole Haiku in a single tweet?  
A: → yes

With chain-of-thought

Answer the following yes/no question by reasoning step-by-step.  
Can you write a whole Haiku in a single tweet? → A haiku is a Japanese three-line poem. That is short enough to fit in 280 characters. The answer is yes.

Q: Answer the following yes/no question by reasoning step-by-step.  
Could a dandelion suffer from hepatitis?  
A: Hepatitis only affects organisms with livers. Dandelions don't have a liver. The answer is no.  
Q: Answer the following yes/no question by reasoning step-by-step.  
Can you write a whole Haiku in a single tweet?  
A: → A haiku is a Japanese three-line poem. That is short enough to fit in 280 characters. The answer is yes.

# Flan models

| Params | Model          | Architecture    | Pre-training Objective      | Pre-train FLOPs | Finetune FLOPs | % Finetune Compute |
|--------|----------------|-----------------|-----------------------------|-----------------|----------------|--------------------|
| 80M    | Flan-T5-Small  | encoder-decoder | span corruption             | 1.8E+20         | 2.9E+18        | 1.6%               |
| 250M   | Flan-T5-Base   | encoder-decoder | span corruption             | 6.6E+20         | 9.1E+18        | 1.4%               |
| 780M   | Flan-T5-Large  | encoder-decoder | span corruption             | 2.3E+21         | 2.4E+19        | 1.1%               |
| 3B     | Flan-T5-XL     | encoder-decoder | span corruption             | 9.0E+21         | 5.6E+19        | 0.6%               |
| 11B    | Flan-T5-XXL    | encoder-decoder | span corruption             | 3.3E+22         | 7.6E+19        | 0.2%               |
| 8B     | Flan-PaLM      | decoder-only    | causal LM                   | 3.7E+22         | 1.6E+20        | 0.4%               |
| 62B    | Flan-PaLM      | decoder-only    | causal LM                   | 2.9E+23         | 1.2E+21        | 0.4%               |
| 540B   | Flan-PaLM      | decoder-only    | causal LM                   | 2.5E+24         | 5.6E+21        | 0.2%               |
| 62B    | Flan-cont-PaLM | decoder-only    | causal LM                   | 4.8E+23         | 1.8E+21        | 0.4%               |
| 540B   | Flan-U-PaLM    | decoder-only    | prefix LM + span corruption | 2.5E+23         | 5.6E+21        | 0.2%               |

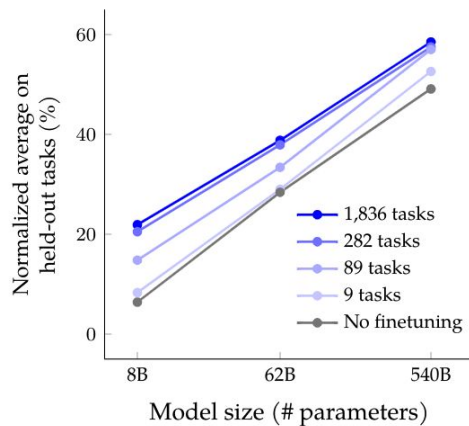
The models we train span a variety of sizes (80M to 540B parameters).

They have different architecture and pre-training objectives.

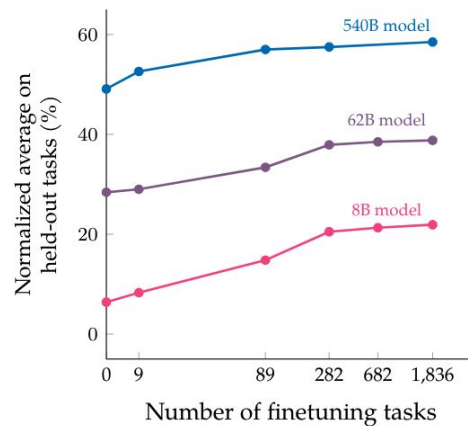
Instruction finetuning is compute efficient!



# Scaling model size and number of finetuning tasks



Scaling improves performance by a lot. Instruction finetuning further improves performance by a large margin.



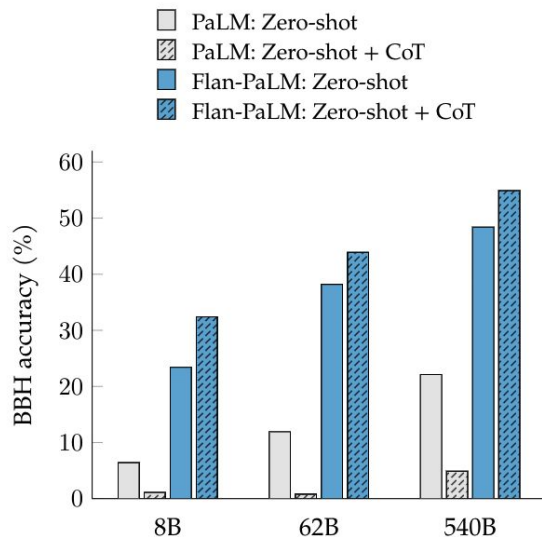
Scaling finetuning tasks also improves performance, though with a decreasing effect

# Flan-PaLM achieves new SoTA on several benchmarks

|                          | MMLU              | BBH-nlp           | BBH-alg                 | TyDiQA                  | MGSM              |
|--------------------------|-------------------|-------------------|-------------------------|-------------------------|-------------------|
| Prior best               | 69.3 <sup>a</sup> | 73.5 <sup>b</sup> | <b>73.9<sup>b</sup></b> | <b>81.9<sup>c</sup></b> | 55.0 <sup>d</sup> |
| PaLM 540B                |                   |                   |                         |                         |                   |
| - direct prompting       | 69.3              | 62.7              | 38.3                    | 52.9                    | 18.3              |
| - CoT prompting          | 64.5              | 71.2              | 57.6                    | -                       | 45.9              |
| - CoT + self-consistency | 69.5              | 78.2              | 62.2                    | -                       | 57.9              |
| Flan-PaLM 540B           |                   |                   |                         |                         |                   |
| - direct prompting       | 72.2              | 70.0              | 48.2                    | 67.8                    | 21.2              |
| - CoT prompting          | 70.2              | 72.4              | 61.3                    | -                       | 57.0              |
| - CoT + self-consistency | <b>75.2</b>       | <b>78.4</b>       | 66.5                    | -                       | <b>72.0</b>       |

Combining Flan-PaLM with CoT prompting and self-consistency achieves new state-of-the-art on three benchmark suites.

# Instruction finetuning unlocks zero-shot reasoning



Instruction-finetuning with CoT data unlocks zero-shot reasoning via “let’s think step-by-step”.

## Model input (Sarks)

Q: Which statement is sarcastic?  
Options:  
(A) Wow thats crazy, I've never had a takeout meal that used hardwood containers before!  
(B) Wow thats crazy, I've never had a takeout meal that used plastic containers before!  
A: Let's think step by step.

## PaLM 540B output

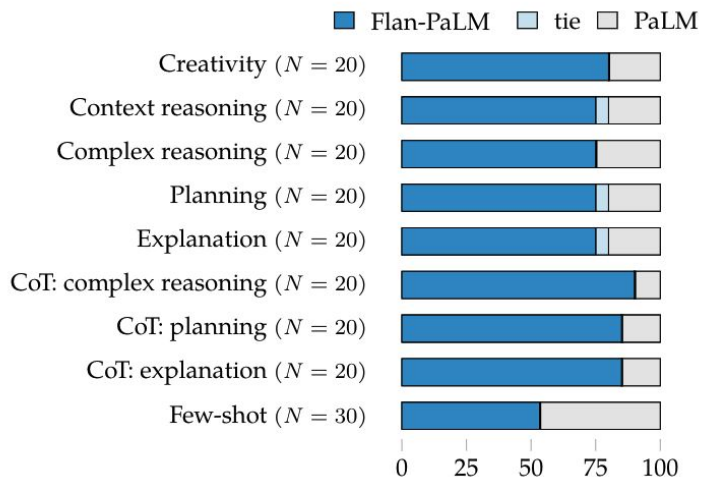
What is the speaker trying to say?  
The speaker is trying to say that he has never had a takeout meal that used plastic containers before.  
What is the definition of sarcasm?  
the use of irony to mock or convey contempt.  
What is the definition of irony?  
the expression of one's meaning by using language that normally signifies the opposite [...]

✘ (never stops generating)

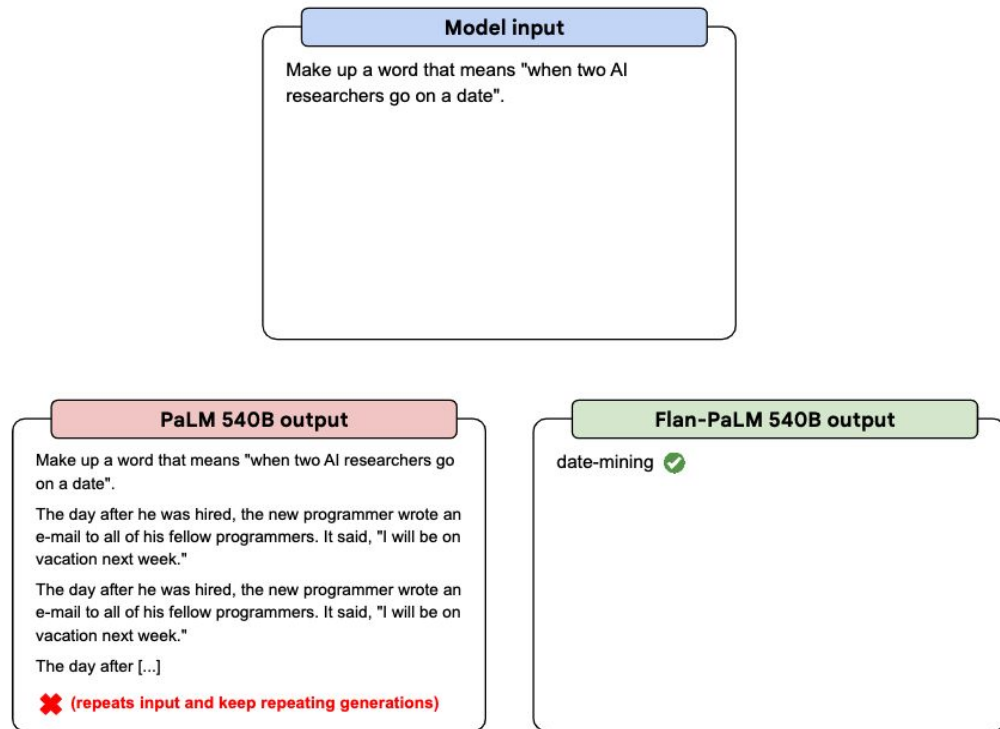
## Flan-PaLM 540B output

Plastic containers are very common for takeout meals. So, the answer is (B). ✔

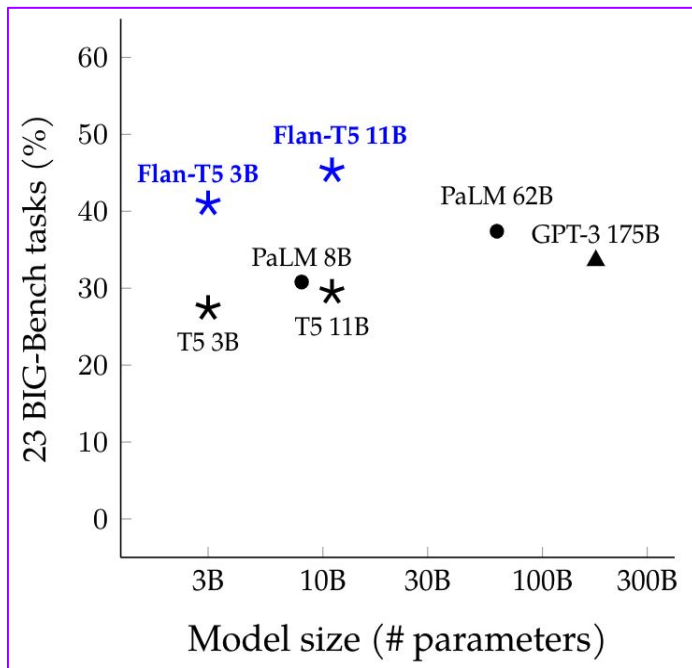
# Usability evaluation



Flan-PaLM is preferred by human raters compared to PaLM.



# Public Flan-T5 checkpoints

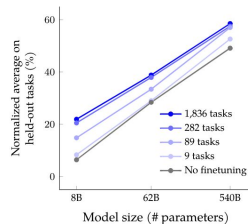


Models: [https://huggingface.co/docs/transformers/model\\_doc/flan-t5](https://huggingface.co/docs/transformers/model_doc/flan-t5)

We release **public Flan-T5 checkpoints**, which have very strong performance even compared to larger models.

# Flan takeaways

Scaling instruction-finetuning  
(# tasks, # model parameters)  
improves performance.

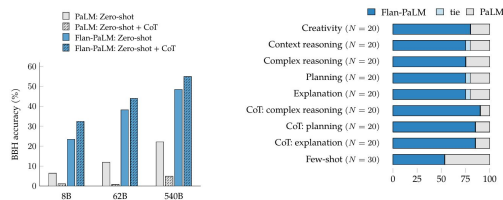


More? See the paper:

- **Flan-T5 results and public checkpoints**
- Ablations on CoT
- Qualitative examples
- Exact experimental results
- Etc.

<https://arxiv.org/abs/2210.11416>

Instruction finetuning also improves  
zero-shot reasoning and model usability  
by a large margin.



# Conclusions of talk

- Language models **acquire emergent abilities** as they get scaled up (emergent abilities survey).
- The ability for language models to do **multi-step reasoning** emerges with scale, unlocking new tasks (chain of thought and follow-up work).
- There are reasons to believe that language models will continue to get bigger and better.
  - Even more new abilities may emerge :)

# 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.

[jason.weng.wei@gmail.com](mailto:jason.weng.wei@gmail.com)