

# Scaling unlocks emergent abilities in language models

Jason Wei  
Google Brain



# Outline

- **Emergent abilities** of large language models.
  - Inverse scaling can become U-shaped.
- **Chain-of-thought prompting** elicits reasoning in large language models.
  - Challenging BIG-Bench tasks and whether chain-of-thought can solve them.
  - Language models are multilingual chain-of-thought reasoners.
  - Self-consistency improves chain-of-thought reasoning in language models.
- Feel free to interrupt anytime with questions :)

## 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>

*craffel@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*

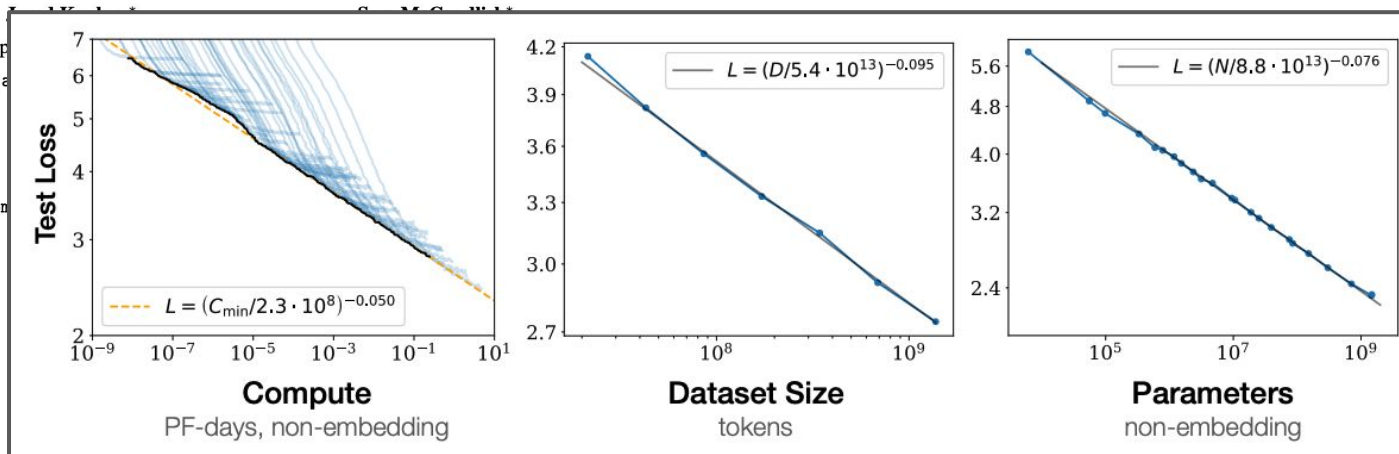
<sup>1</sup>Google Research <sup>2</sup>Stanford University <sup>3</sup>UNC Chapel Hill <sup>4</sup>DeepMind

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  
ja...  
**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:

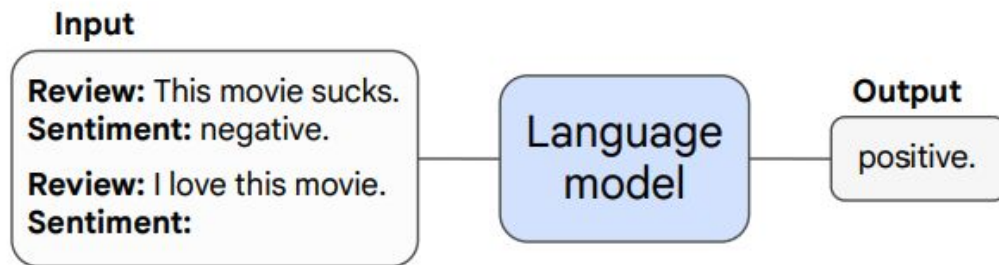
- **Uranium.** With a bit of uranium, nothing special happens; with a large amount of uranium packed densely enough, you get a nuclear reaction.
- **DNA.** Given only small molecules such as calcium, you can't meaningfully encode useful information; given larger molecules such as DNA, you can encode 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

## 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.*

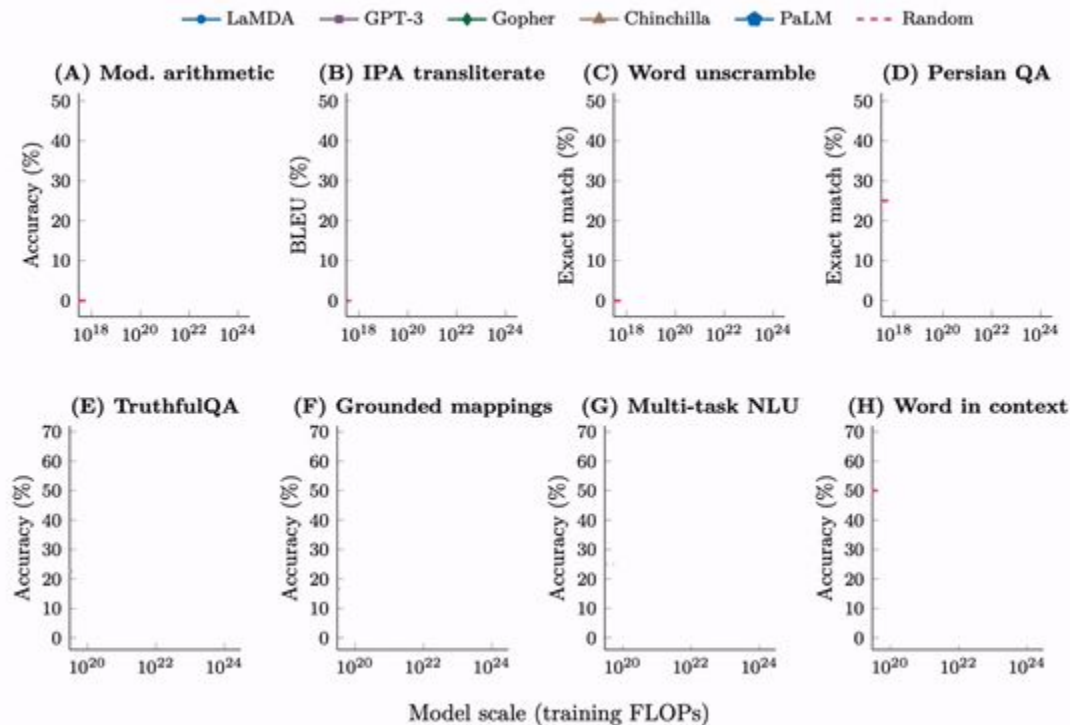
- How to measure the “size” of the model?
  - Training FLOPs
  - Number of model parameters
  - Training dataset size

# 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

## 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: A

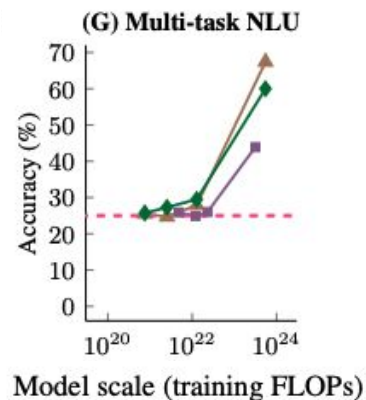
If 4 daps = 7 yaps, and 5 yaps = 3 baps, how many daps equal 42 baps?

(A) 28 (B) 21 (C) 40 (D) 30

Answer: C

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

LaMDA GPT-3 Gopher  
Chinchilla PaLM Random



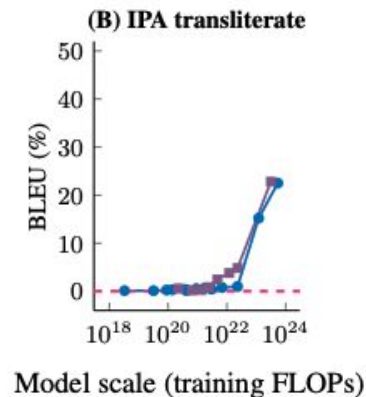
# Emergence in few-shot prompting

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

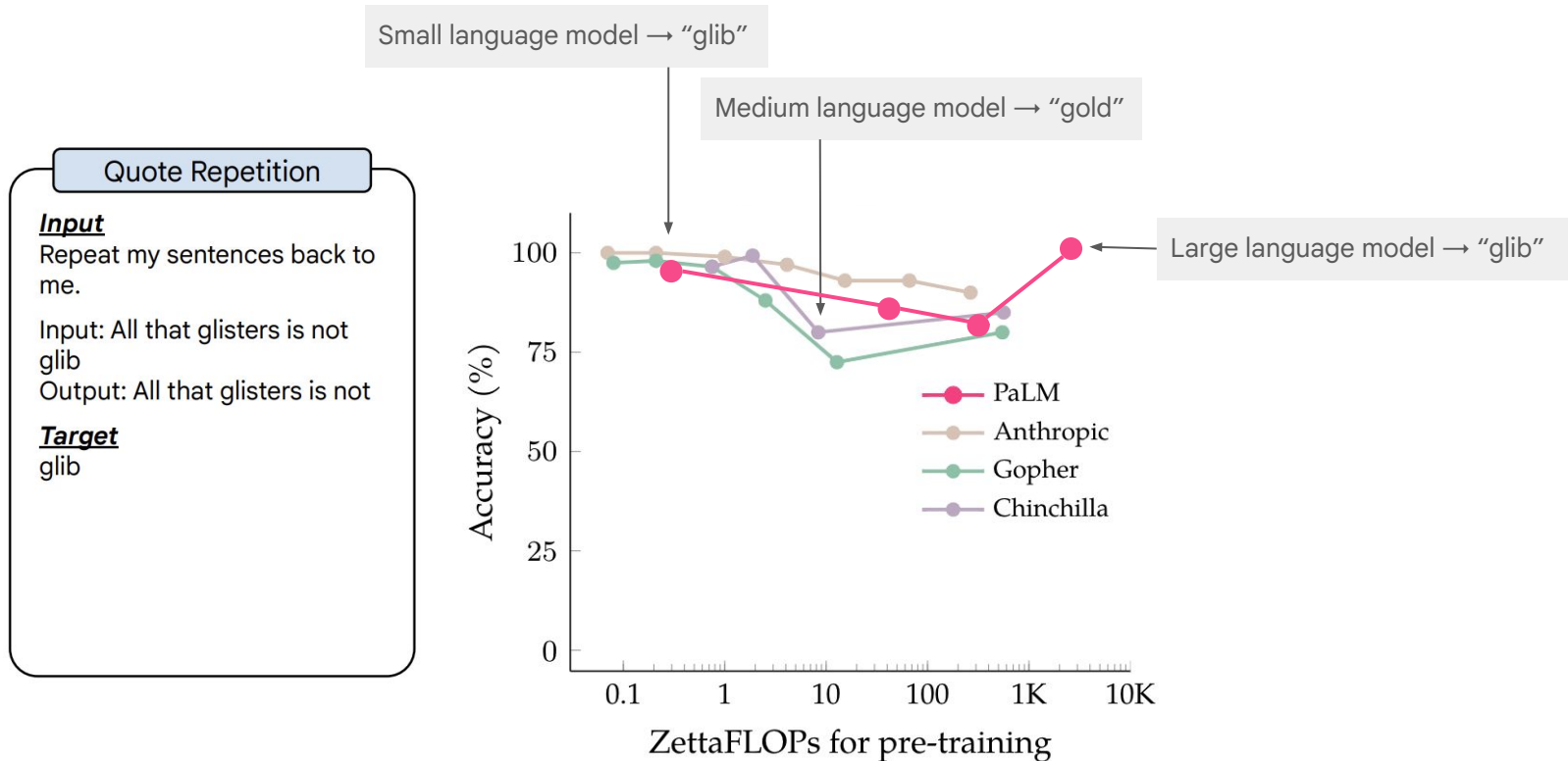
Target (IPA): ðə 1931 'meɪleɪ  
'sɛnsəs wəz ən ə'larm bɛl.

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

LaMDA GPT-3 Gopher  
Chinchilla PaLM Random



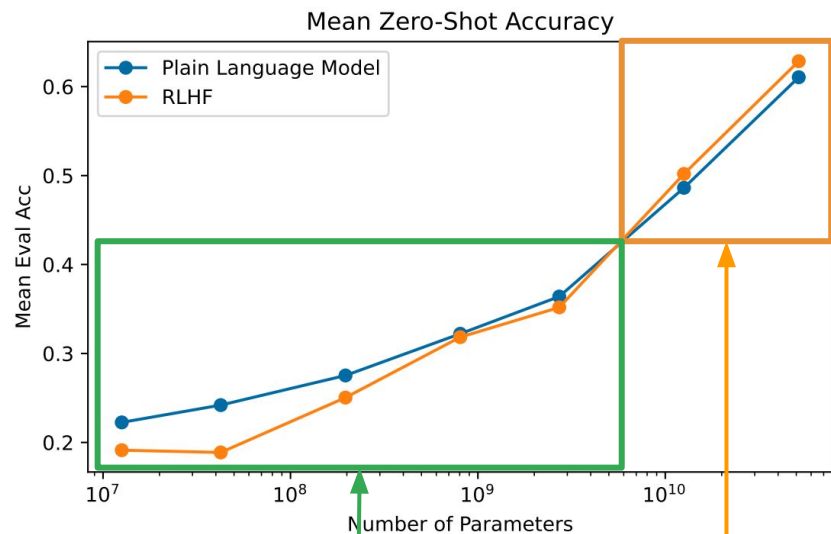
# Inverse scaling can become U-shaped



# Emergent prompting techniques

A prompting technique is emergent if it hurts performance (compared to baseline) for small models, and improves baseline for large models

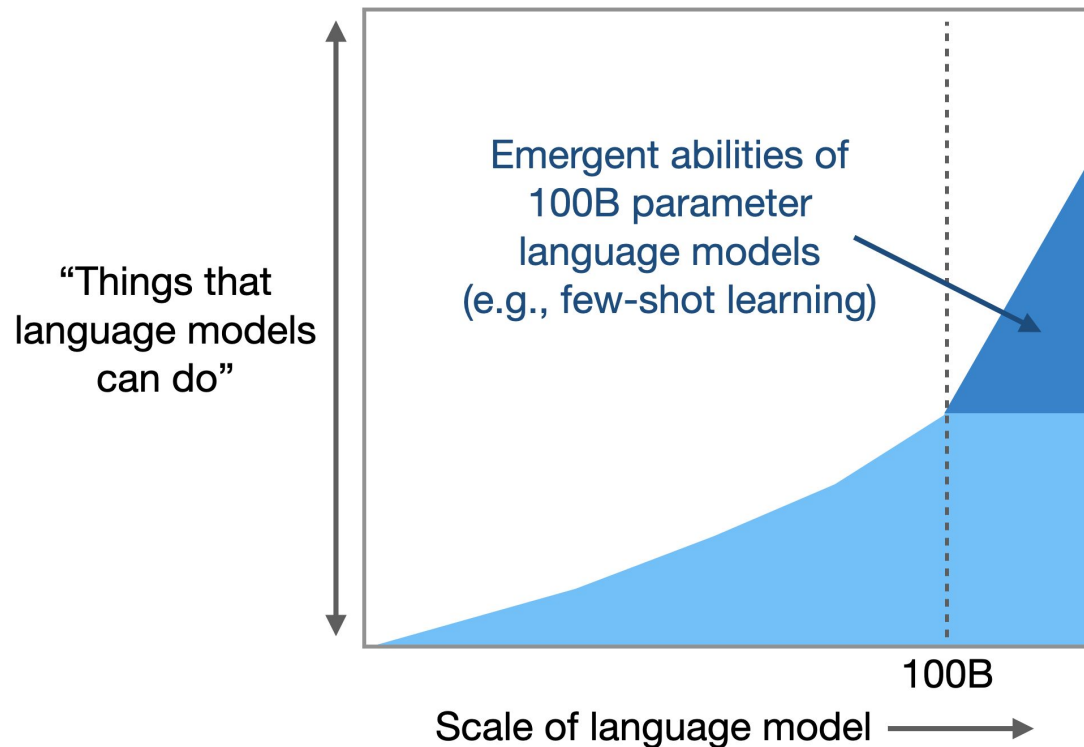
> later: chain-of-thought prompting as an emergent prompting technique



[Bai et al., 2022.](#)

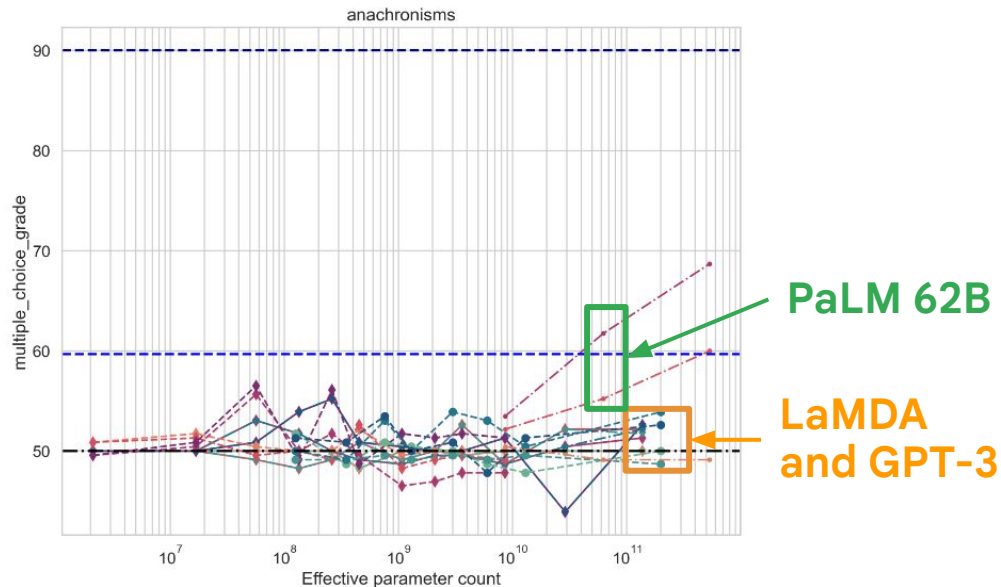
RLHF hurts  
performance

RLHF helps  
performance

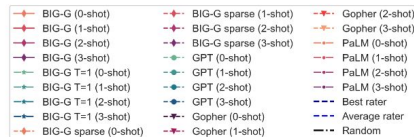


# Emergence: better data

Smaller models with better data can also lead to emergence, even when larger models trained on worse data don't demonstrate worse behavior

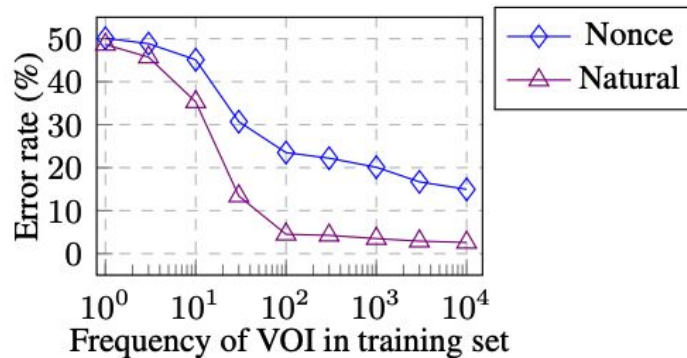


BIG-Bench ([Srivastava et al., 2022](#)).



# Emergence: better data

Better (in-domain) data makes a big difference when compute, model parameters, and dataset size are fixed



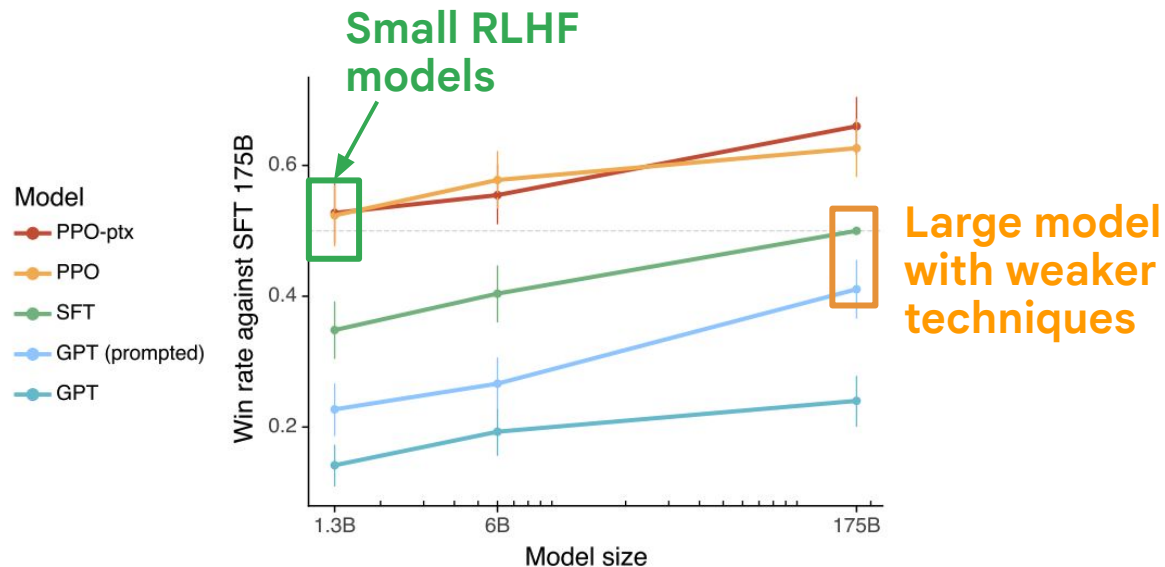
(Setup: small BERT models pre-trained from scratch, task is subject-verb agreement)

Figure 3: Effect of absolute frequency of a verb of interest (VOI) when the ratio between singular and plural forms is held constant at 1:1. The error rate for sixty VOI is shown for BERT models that have seen the sixty VOI at different frequencies in the pre-training dataset.

[Wei et al. 2021.](#)

# Emergence: finetuning for desired behaviors

Desired behaviors can be induced in smaller models via finetuning and RLHF



[Ouyang et al., 2022.](#)



# Emergence: measure of model “scale”

What's the right x-axis for emergence?

Can be viewed through training FLOPs, model parameters, Wiki-text103 perplexity

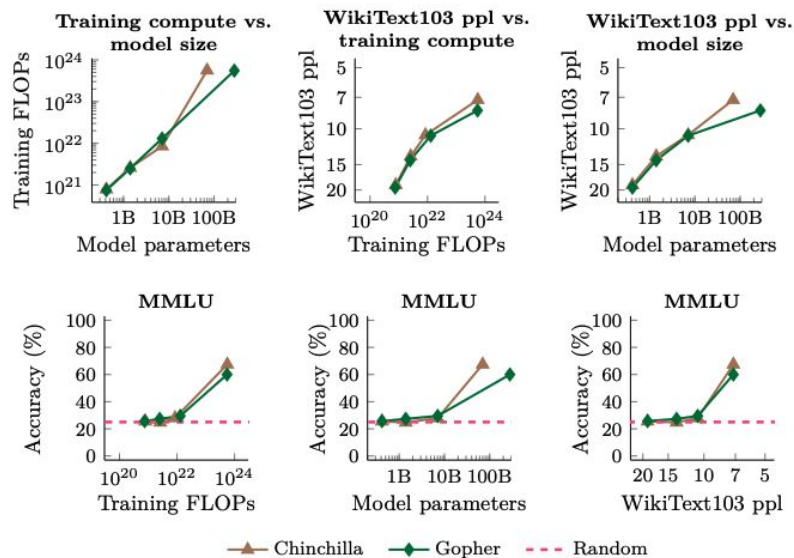
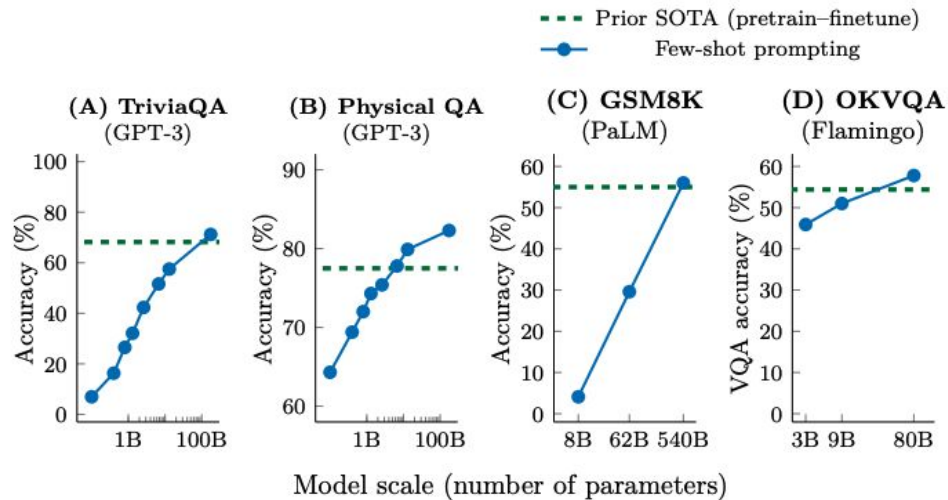


Figure 4: Top row: the relationships between training FLOPs, model parameters, and perplexity (ppl) on WikiText103 (Merity et al., 2016) for Chinchilla and Gopher. Bottom row: Overall performance on the massively multi-task language understanding benchmark (MMLU; Hendrycks et al., 2021a) as a function of training FLOPs, model parameters, and WikiText103 perplexity.

# Emergence: surpassing finetuning

Sociological change in the AI community: finetuned task-specific models are outperformed by few-shot prompted large model



## Summary of emergence:

- Emergent abilities can only be observed in large models
  - Their emergence cannot be predicted by scaling plots with small models only

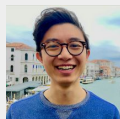
## Reflection:

- Framing for viewing these abilities, which are not intentionally built in
  - Subtext: “why we should keep scaling; these abilities are hard to find otherwise,” context around this
- Tension between emergence (task-general; bigger models) and many production tasks (task-specific; compute constraints; in-domain data)
- Haven’t seen a lot of work on predicting future emergence
  - Why? Too hard, only task-specific answers? Maybe Anthropic is working on it

# 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

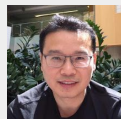
Google Research, Brain Team  
{jasonwei, dennyzhou}@google.com



Jason Wei



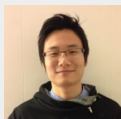
Xuezhi Wang



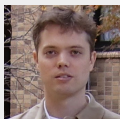
Denny Zhou



Nathaneal  
Schärli



Le Hou



Nathan  
Scales



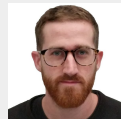
Maarten  
Bosma



Aakanksha  
Chowdhery



Sharan  
Narang



Brian Ichter



Fei Xia



Dale  
Schuurmans



Olivier  
Bousquet



Quoc Le



Ed Chi

...s of intermediate reasoning  
language models to perform  
reasoning abilities emerge  
ple method called *chain-of-*  
onstrations are provided as

chain-of-thought prompting  
nonsense, and symbolic  
For instance, prompting a  
rs achieves state-of-the-art  
problems, surpassing even



Google I/O 2022

Video:

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

# 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?

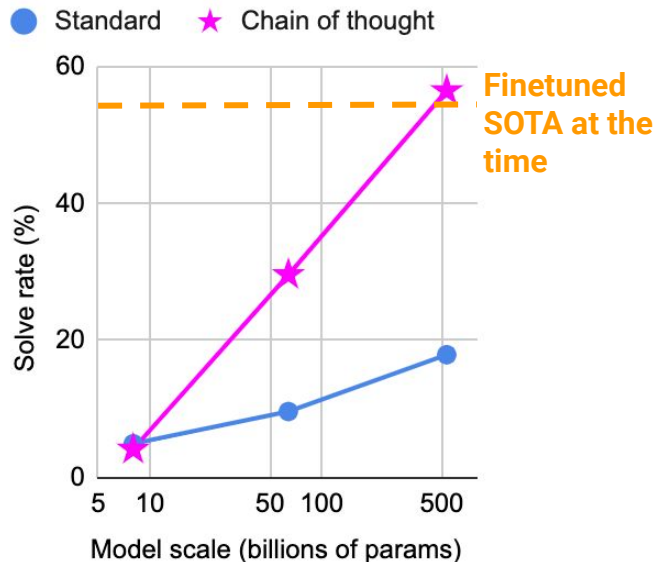
#### Model Output

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. ✅

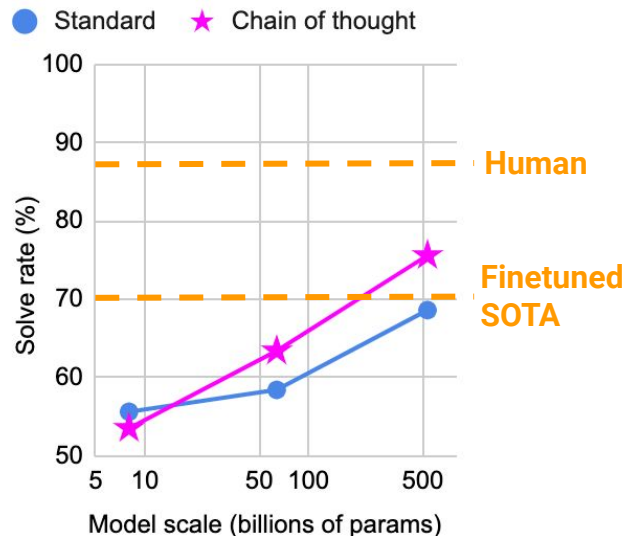
## CoT demo

# CoT paper

## GSM8K



## StrategyQA



### Real model output #1

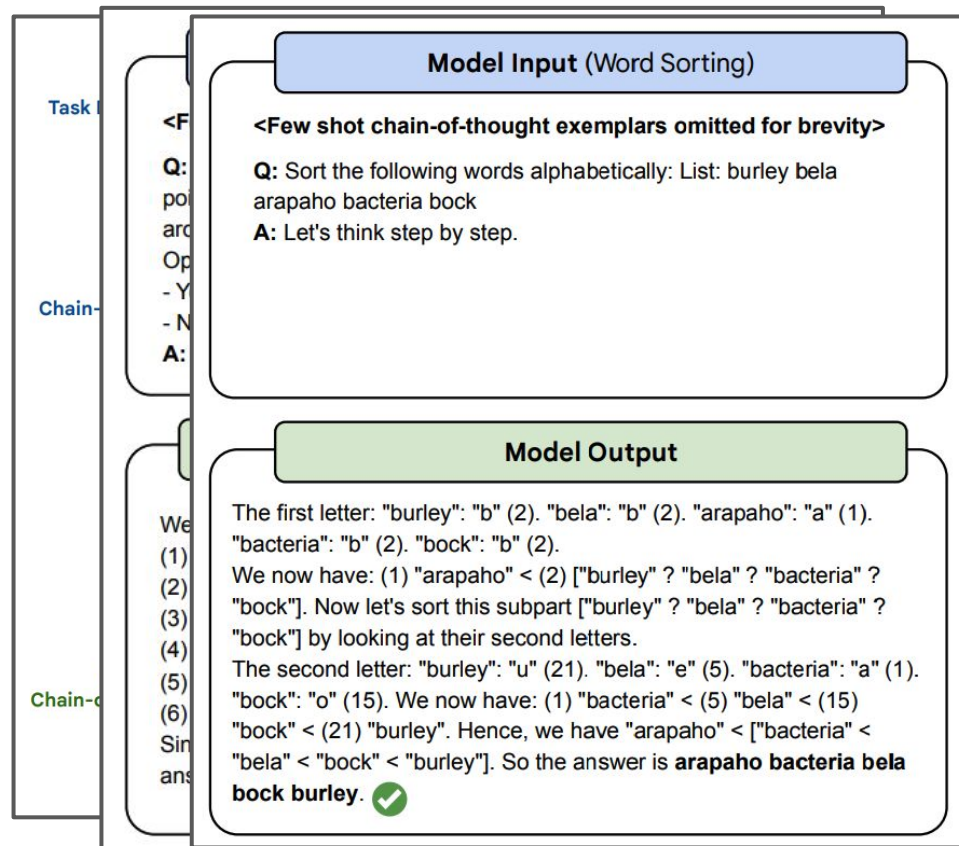
**Q:** Can you hide a basketball in a sand cat's ear?

**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





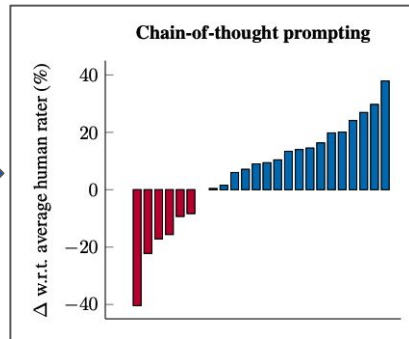
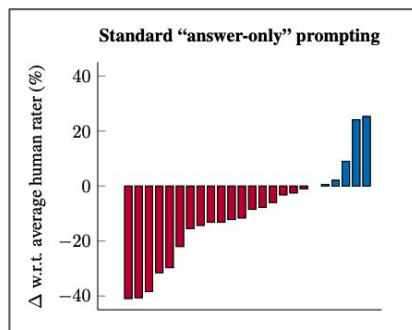
# CoT on BIG-Bench: Result summary

	BBH all (23 tasks)	# tasks above 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

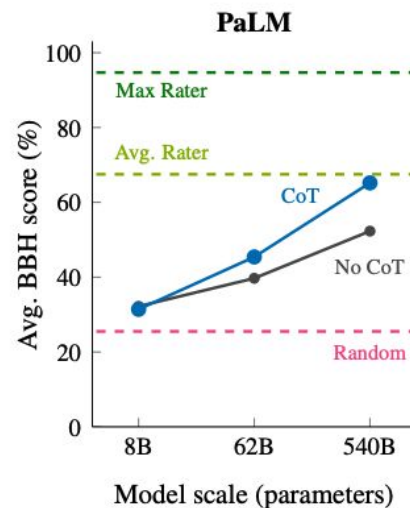
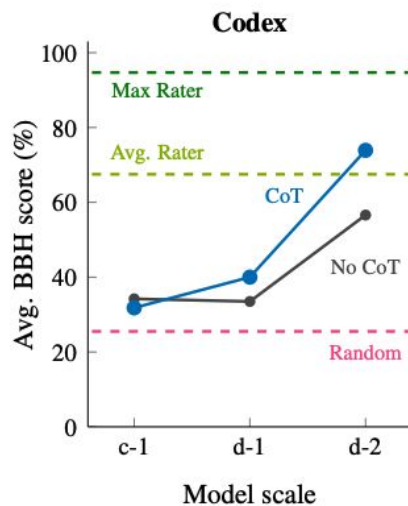
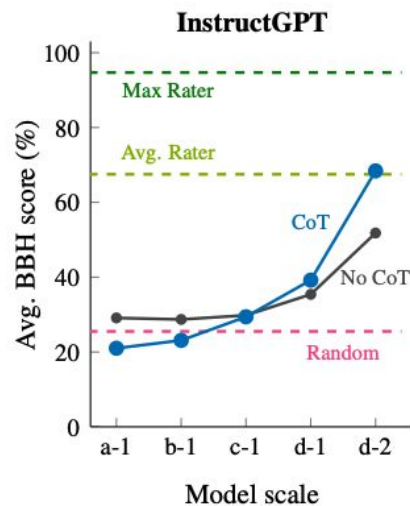
Detail: better formatting (options, task description) already beats prior best

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



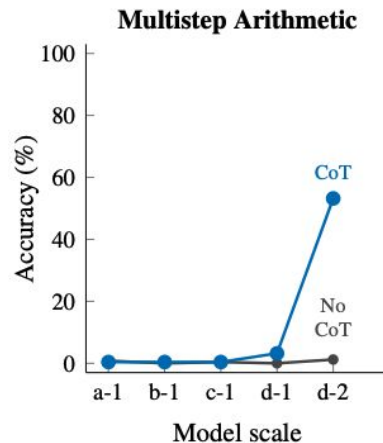
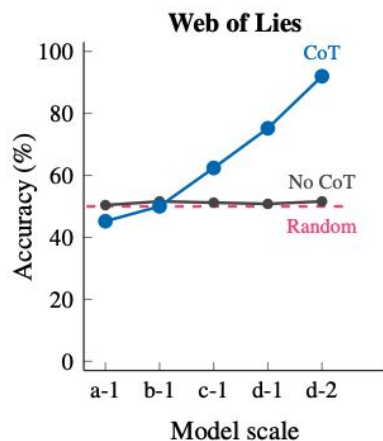
# CoT on BIG-Bench: Scaling

- 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 =  $(-4 - 1 - -4 - 4) = ((-4 - 1) - -4 - 4) = ((-5) - -4 - 4) = ((-5 + 4) - 4) = (-1 - 4) = -5$ .

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

# Multilingual chain-of-thought prompting

- Manually translated version of 250 examples from GSM8K into 10 languages
- Prompt the model with Bengali math problems and Bengali reasoning
- This input is highly improbable (Bengali is 0.01% of pre-training data)

## Model Input

প্রশ্ন: রজারের ৫টি টেনিস বল আছে। সে আরও ২ ক্যান টেনিস বল কিনেছে। প্রতিটি ক্যানে ৩টি করে টেনিস বল আছে। তার কাছে এখন কতগুলি টেনিস বল আছে?

ধাপে ধাপে উত্তর: রজারের প্রথমে ৫টি বল ছিল। ২টি ক্যানের প্রতিটিতে ৩টি টেনিস বল মানে ৬টি টেনিস বল।  $5 + 6 = 11$ । উত্তর হল ১১।

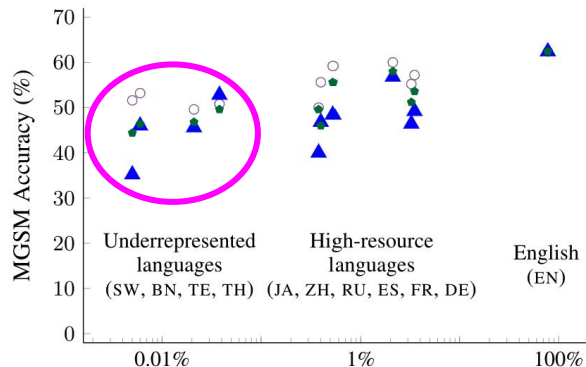
প্রশ্ন: জেনেটের হাঁসগুলি প্রতিদিন ১৬টি করে ডিম পাড়ে। তিনি প্রতিদিন প্রাতরাশে তিনটি করে ডিম খান এবং বন্ধুদের জন্য প্রতিদিন চারটি ডিম দিয়ে মাফিন তৈরি করেন। অবশিষ্ট হাঁসের ডিমগুলি তিনি প্রতিদিন কৃষকদের বাজারে প্রতি ডিম \$২ দরে বিক্রয় করেন। তিনি কৃষকদের বাজারে প্রতিদিন কত ডলার উপার্জন করেন?

## Model Output

ধাপে ধাপে উত্তর: প্রতিদিন ১৬টি 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 \$২ হলে প্রতিদিন  $9 * 2 = 18$  ডলার উপার্জন করেন। উত্তর হল ১৮। ✓

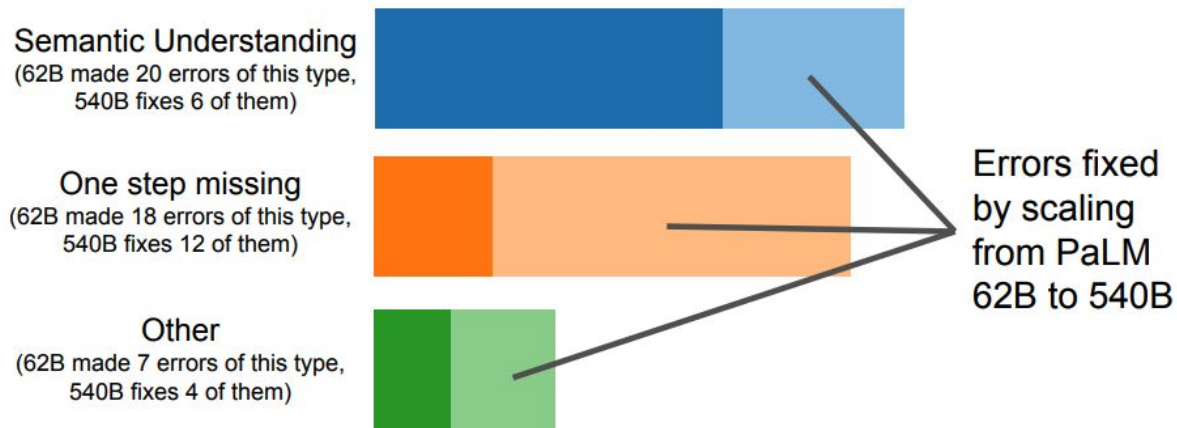
- Expected correlation between language frequency and performance
- Underrepresented languages did surprisingly well
- Implication: nice demonstration of compositionality of the model

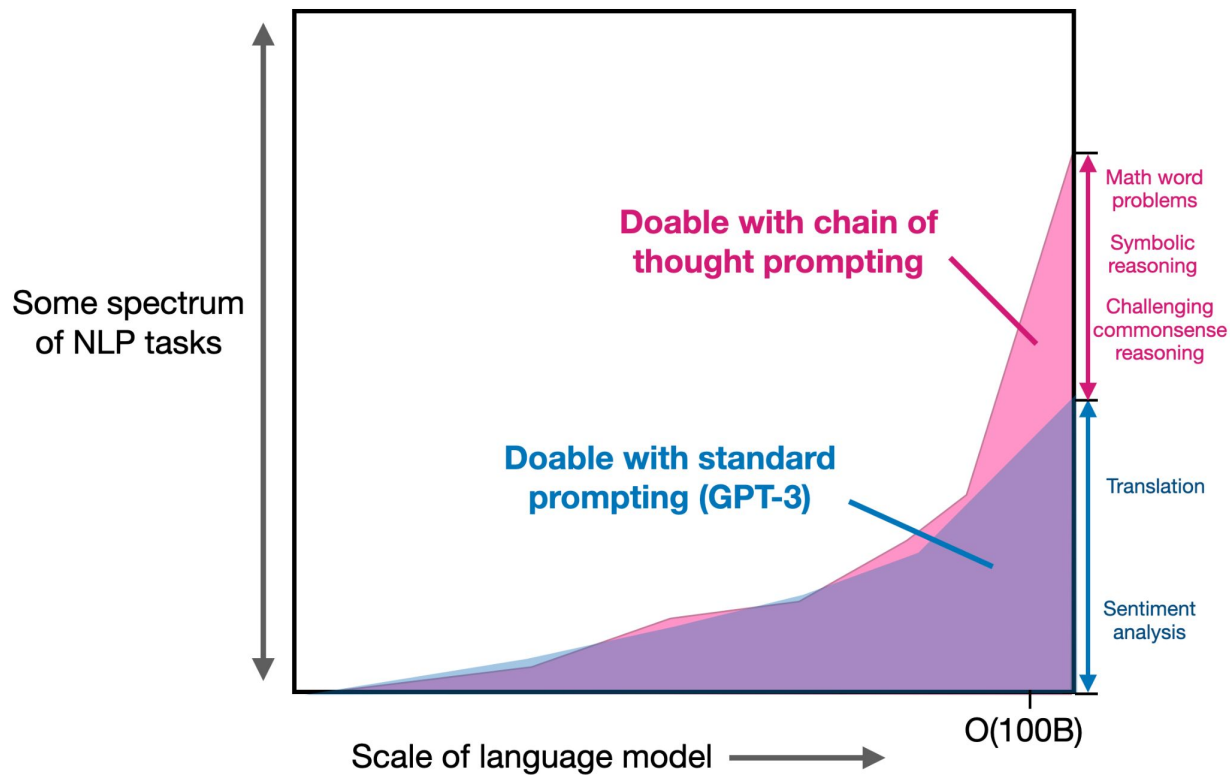
- Translate to English with Google Translate and solve with English intermediate steps
- ▲ Intermediate reasoning steps in the language of the question
- Intermediate reasoning steps in English



Frequency of language in pre-training dataset (token percentage)

# 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  $\$2 * 9 =$   
 $\$18$  per day.

The answer is \$18.

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

The answer is \$14.

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

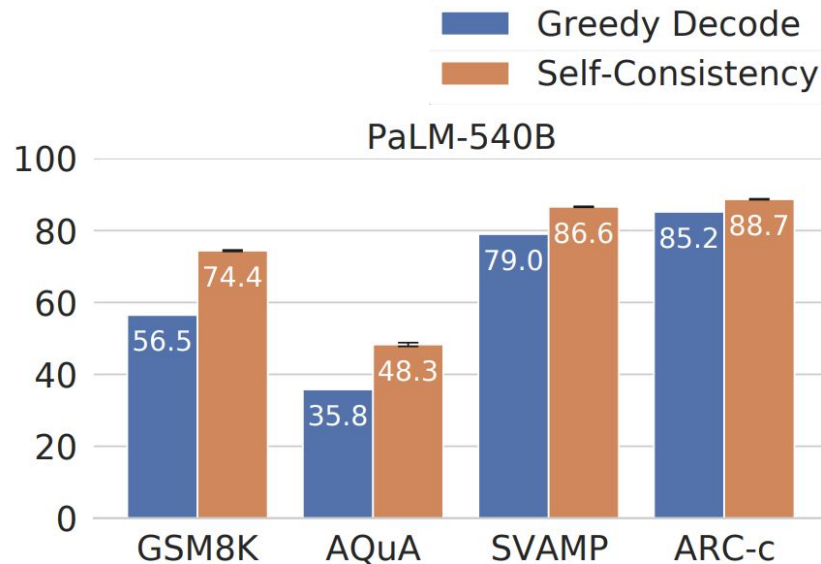
The answer is \$18.

Majority vote  
on the answers

The answer is \$18.

# Self-consistency: results

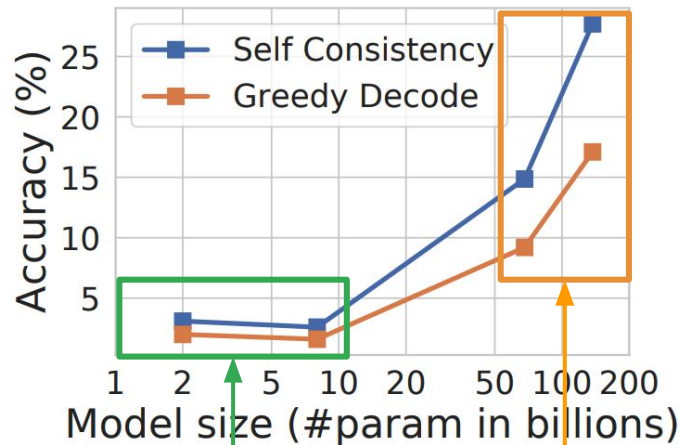
Simple trick but big performance delta





# Self-consistency: emergence

Self-consistency doesn't work for small models, but can help a lot for large models



Increases  
performance  
by a lot

Doesn't work

# 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

# 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)