

Some intuitions about large language models

Language modeling objective

Dartmouth students like
to __



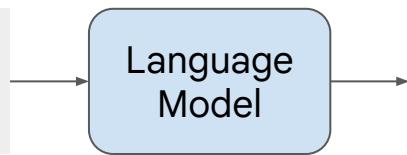
Most basic form
(not chatGPT)

<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 __



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

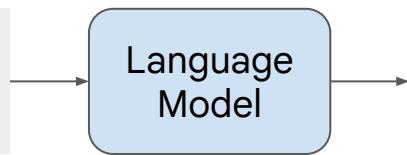
The next word
is probably not
a noun

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 __



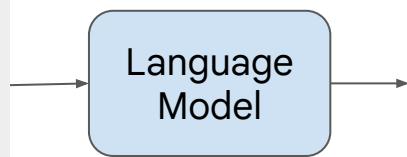
<u>Word</u>	<u>Probability</u>
a	
...	
Copenhagen	0.9
...	
London	0.05
...	
zucchini	

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



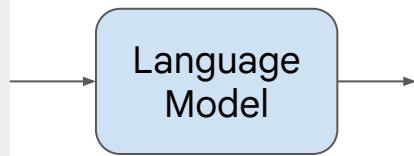
<u>Word</u>	<u>Probability</u>
a	
...	
spoon	0.00001
...	
zebras	0.6
...	
zucchini	

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 __



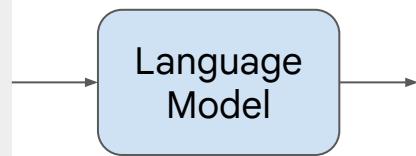
<u>Word</u>	<u>Probability</u>
a	
...	
bad	0.1
...	
good	0.9
...	
zucchini	

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



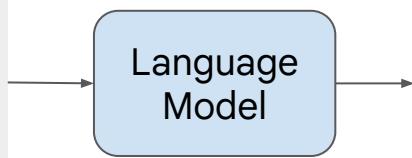
<u>Word</u>	<u>Probability</u>
a	
...	
bad	0.7
...	
good	0.3
...	
zucchini	

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 __



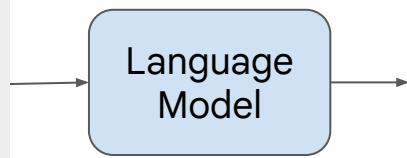
<u>Word</u>	<u>Probability</u>
a	
...	
bonita	0.8
...	
hola	0.03
...	
zucchini	

Understanding
of multiple
languages

(hypothetical)

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

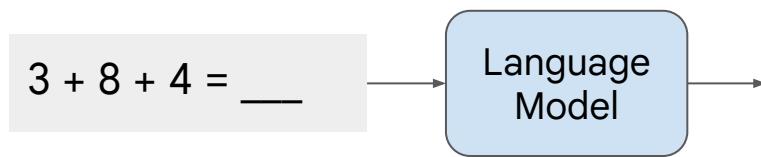
Iroh went into
the kitchen to
make some tea.
Standing next to
Iroh, Zuko
pondered his
destiny. Zuko
left the __



<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

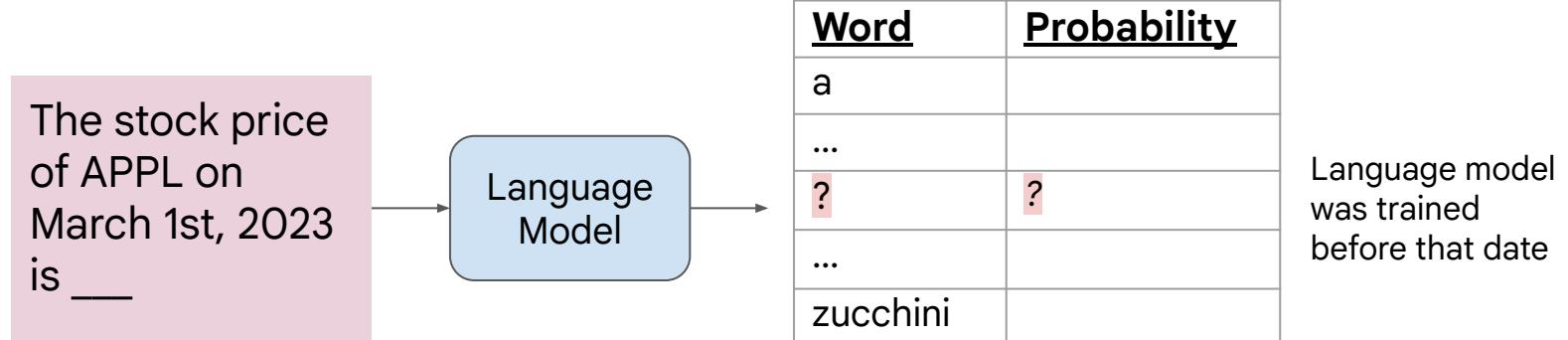


<u>Word</u>	<u>Probability</u>
a	
...	
14	0.1
15	0.7
...	
zucchini	

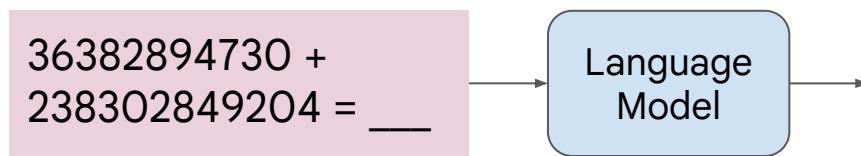
Understanding (or
memorization) or
addition?

(hypothetical)

[1/6] What can't language models do from next word prediction?
→ Not current world knowledge



[2/6] What can't language models learn from next word prediction?
→ Not arbitrarily long arithmetic

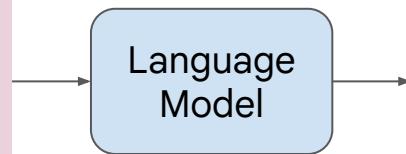


<u>Word</u>	<u>Probability</u>
a	
...	
?	?
...	
zucchini	

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 __

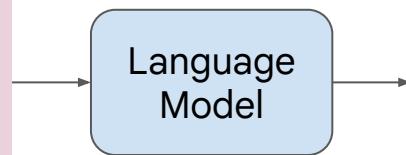


<u>Word</u>	<u>Probability</u>
a	
...	
?	?
...	
zucchini	

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

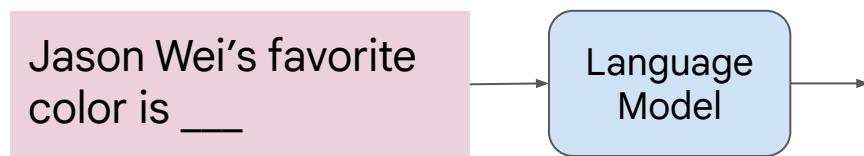
The winner of the
FIFA world cup in
2026 is __



<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

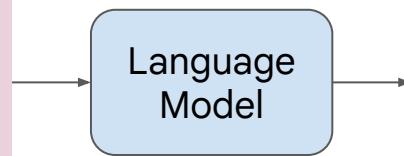


<u>Word</u>	<u>Probability</u>
a	
...	
?	?
...	
zucchini	

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? __



<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

...
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

Today

...
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

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

Emergent Abilities of Large Language Models

Jason Wei¹

jasonwei@google.com

Yi Tay¹

yitay@google.com

Rishi Bommasani²

nlprishi@stanford.edu

Colin Raffel³

craffel@gmail.com

Barret Zoph¹

barrettzoph@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

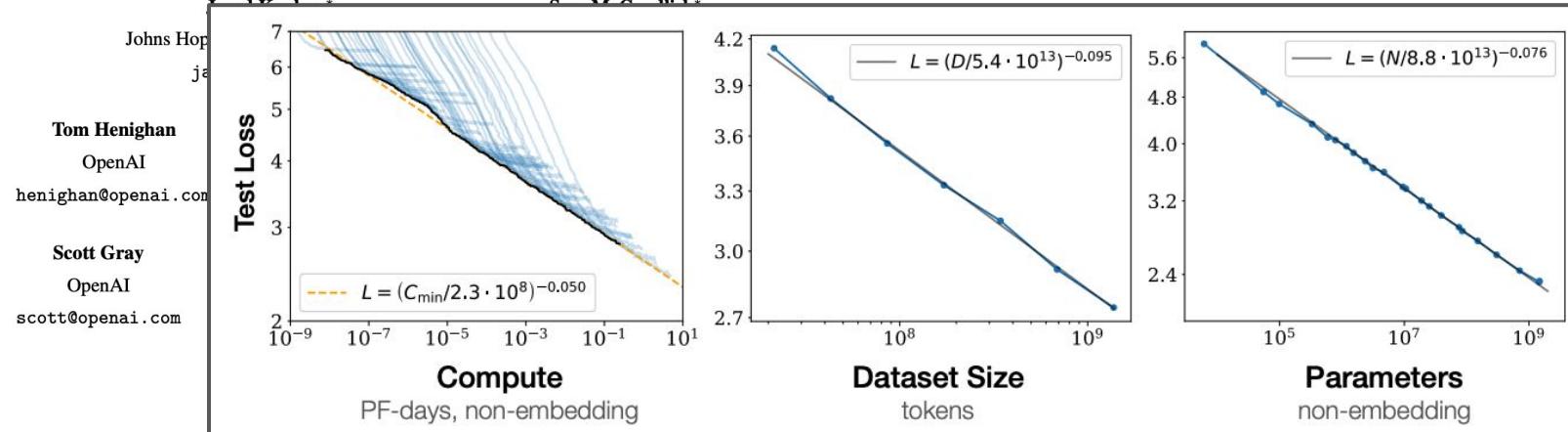
William Fedus¹

liamfedus@google.com

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

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*”

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

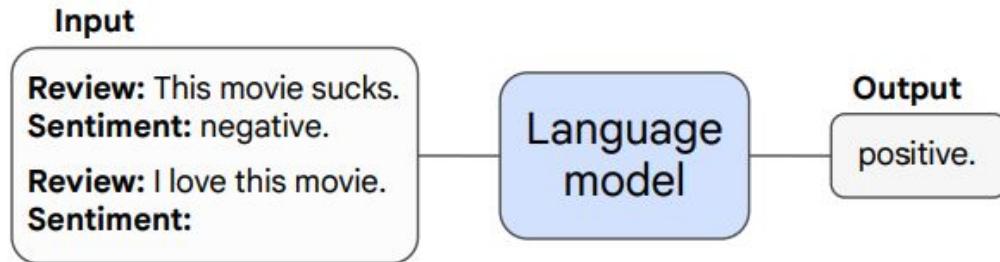
be an obvious candidate of reduction. — Jacob Steinhardt, 2022

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.*

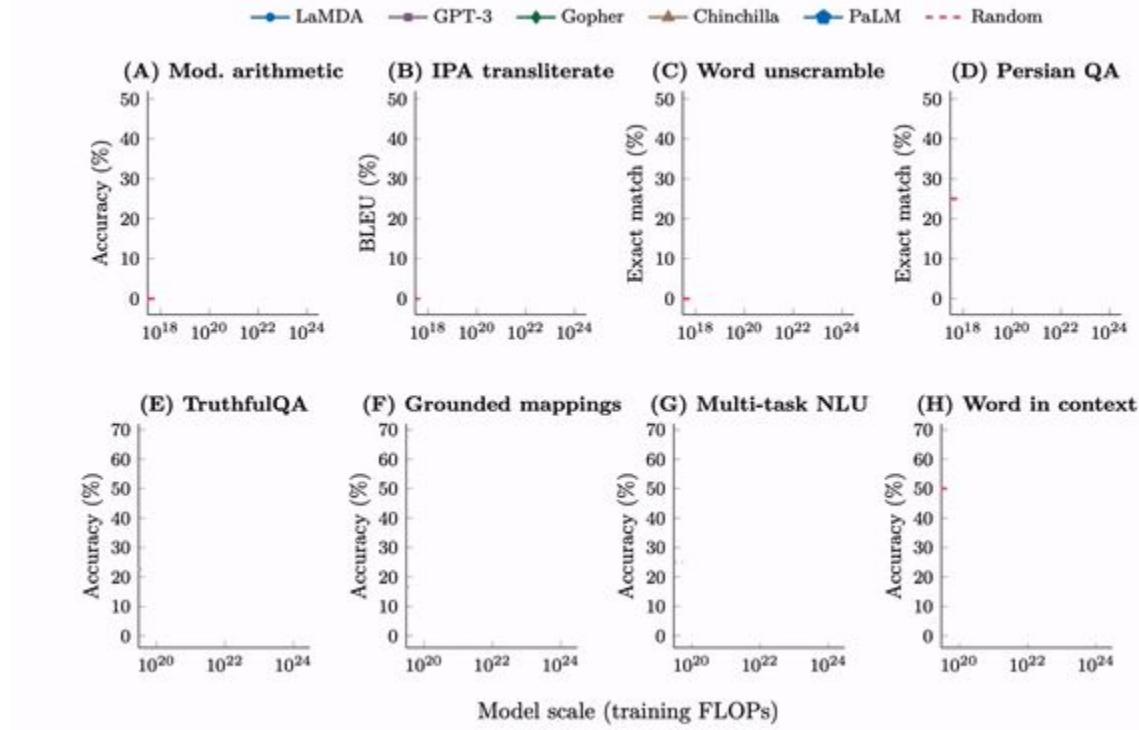
- How to measure the “size” of the model?
 - Training FLOPs
 - Number of model parameters
 - Training dataset size

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

Microeconomics

- 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.

✗
✗
✗
✓

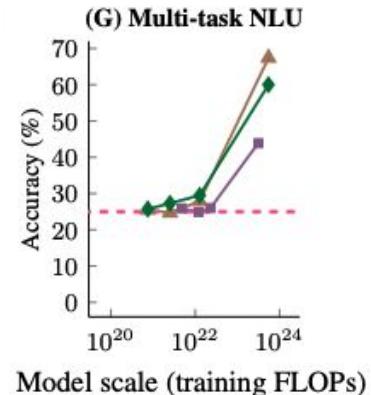
College Mathematics

- 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.](#)

LaMDA GPT-3 Gopher
Chinchilla PaLM Random



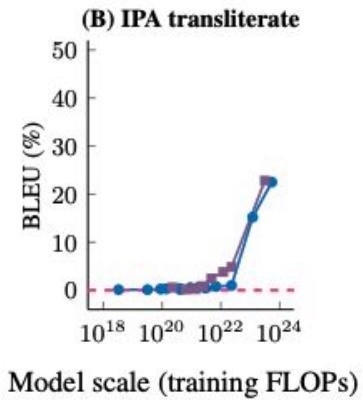
Emergence in few-shot prompting

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

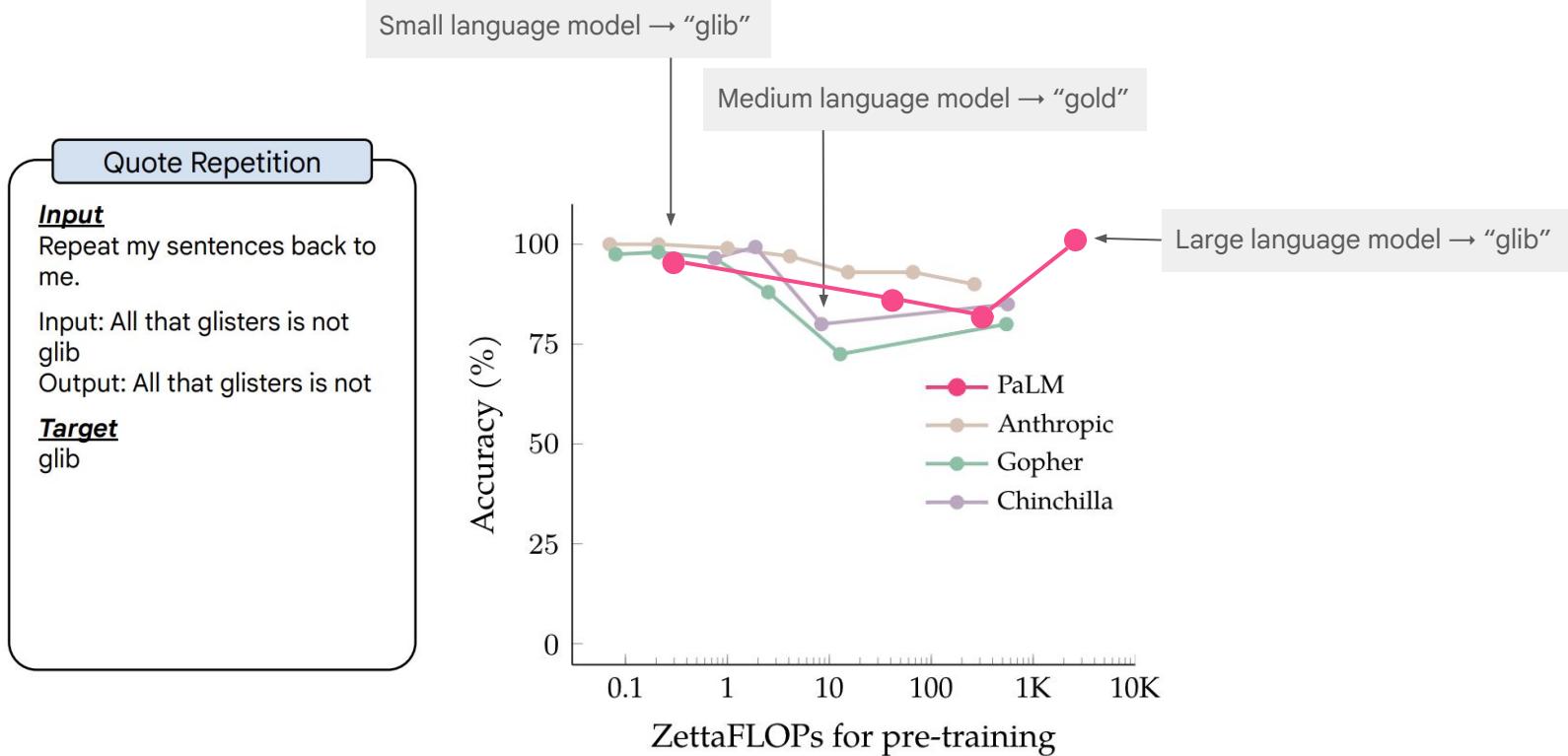
Target (IPA): ðə 1931 'meɪləɪ 'sɛnsəs waz ən ə'larm bel.

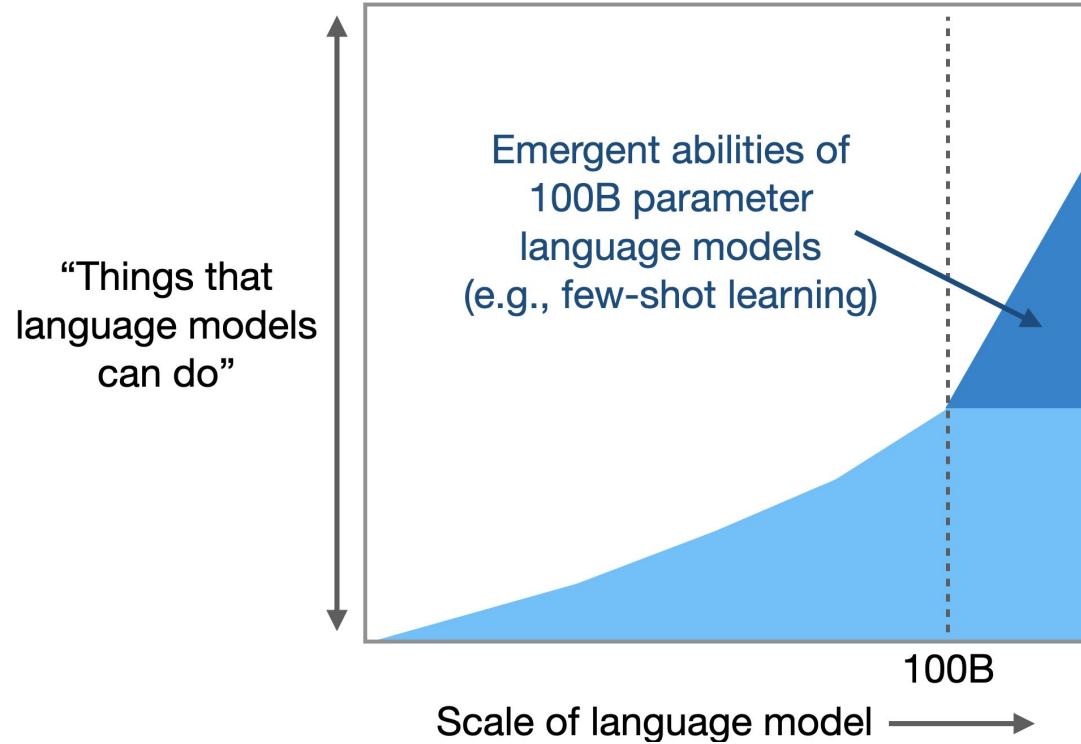
BIG-Bench ([Srivastava et al., 2022](#)).

—●— LaMDA —■— GPT-3 —◆— Gopher
—▲— Chinchilla —●— PaLM - - - Random



Inverse scaling can become U-shaped

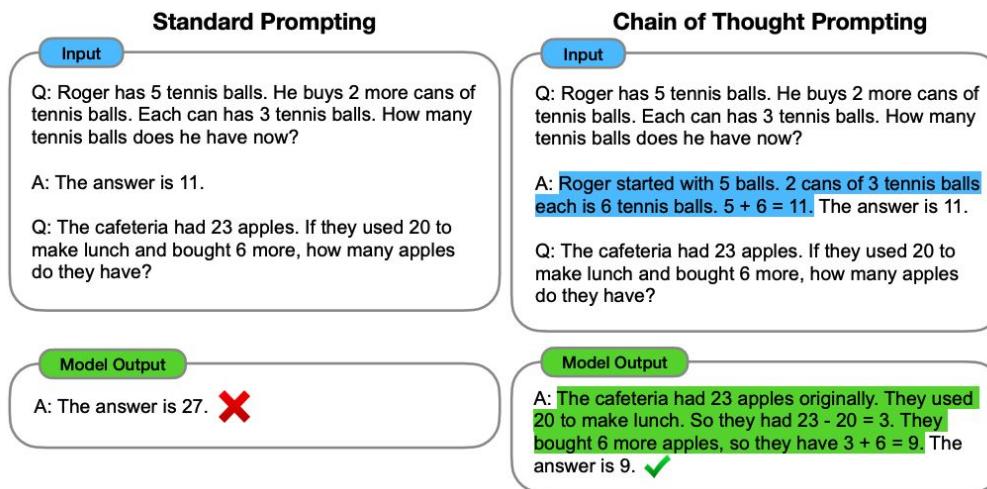




CoT paper

Motivation:

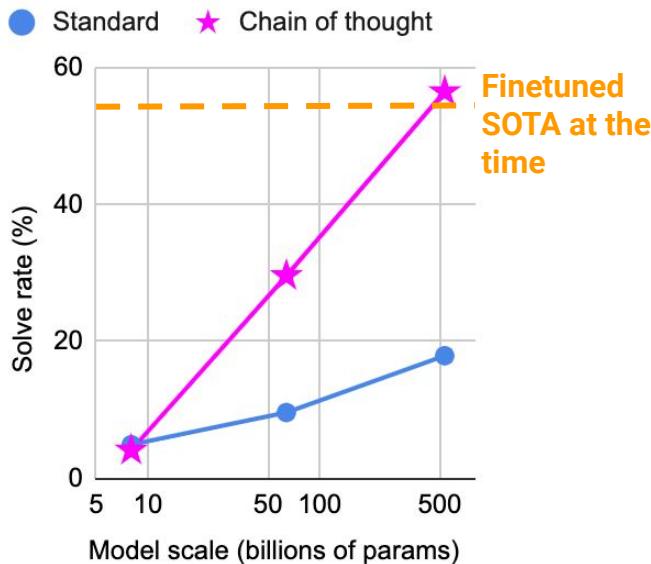
- Enable language models to do more-complicated tasks
- Guide them with “meta-data” (i.e., reasoning process)
- Prompts are manually composed (prompt engineering helps)



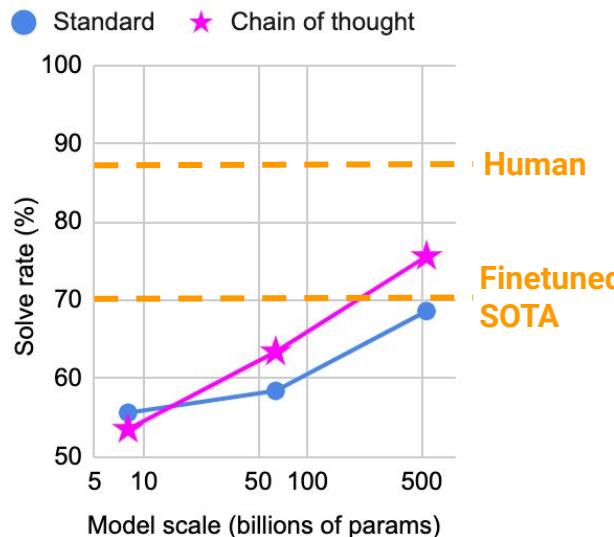
CoT demo

CoT paper

GSM8K



StrategyQA



Real model output #1

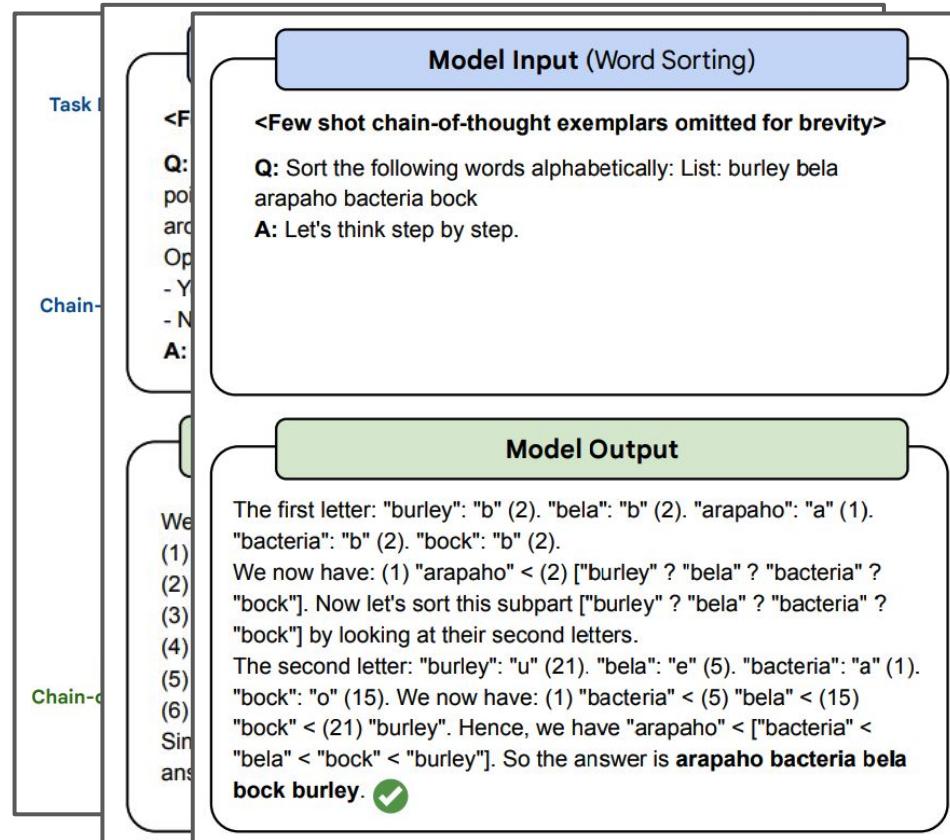
Q: Can you hide a basketball in a sand cat's ear?

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

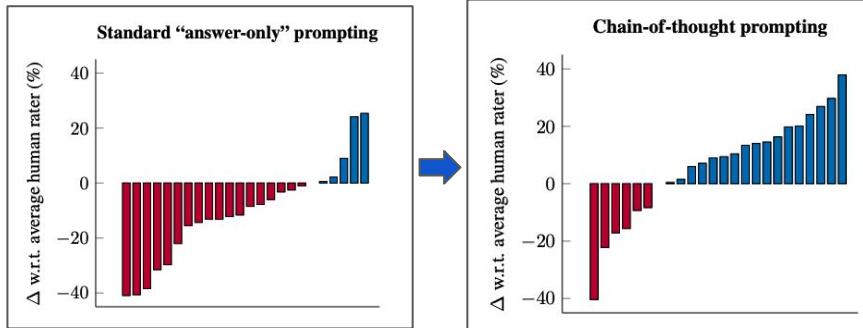


CoT on BIG-Bench: Result summary

	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

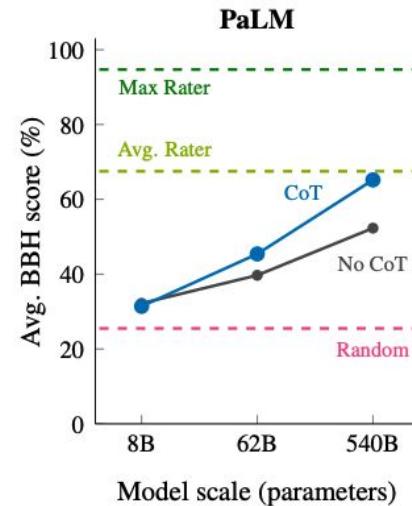
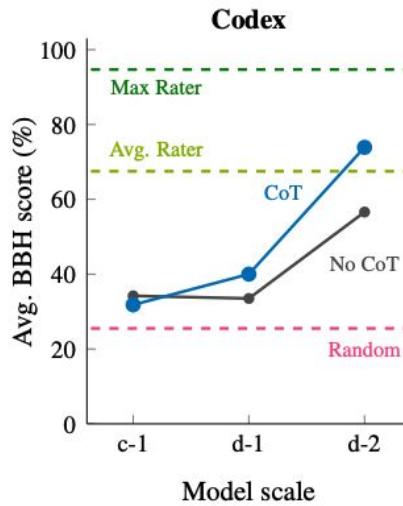
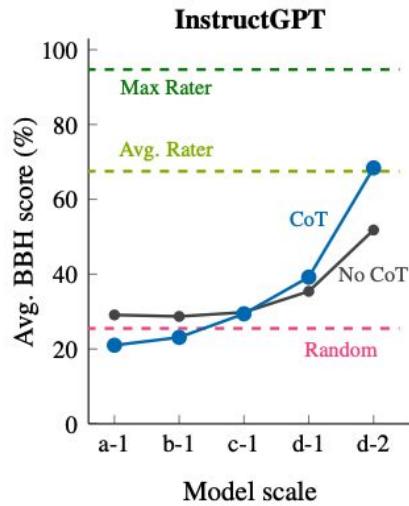
Detail: better formatting (options, task description) already beats prior best



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

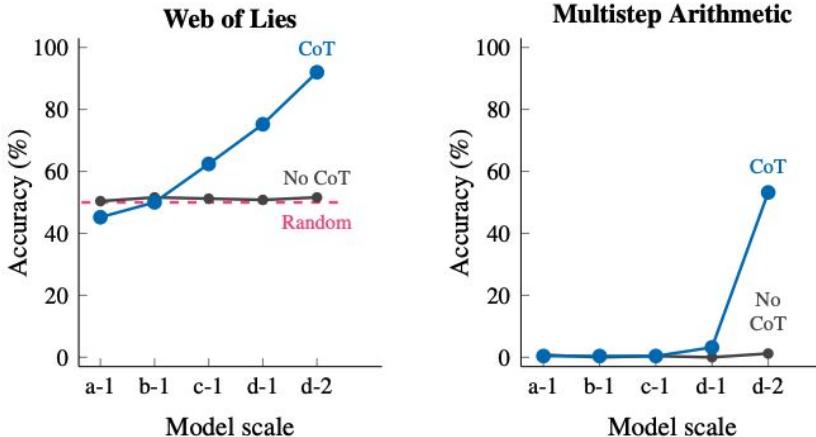
CoT on BIG-Bench: Scaling

- 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 multiplication (from left to right). So, remember to always compute the expressions inside parentheses or brackets first.

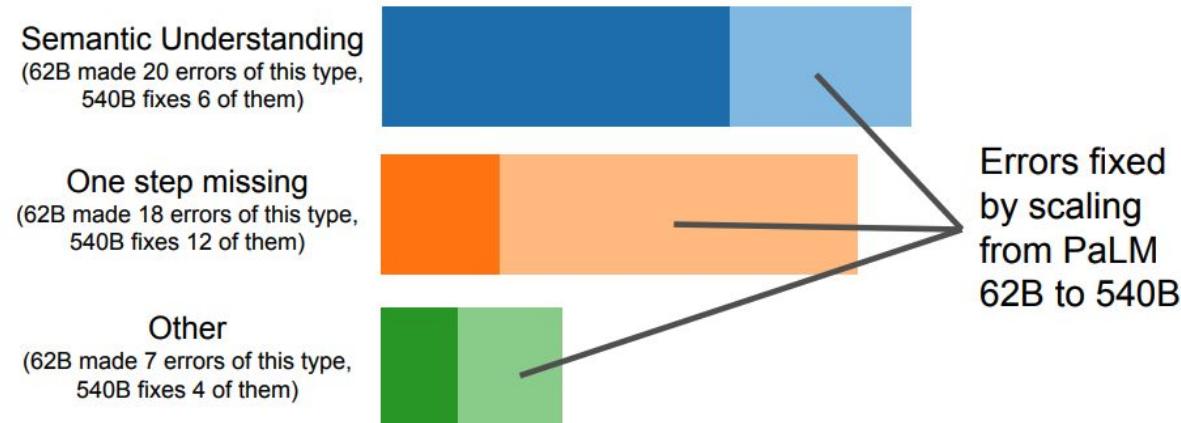
This equation can be written as "A - B", where $A = (4 + 7 * 4 - 5)$ and $B = (4 - 1 - 4 - 4)$.

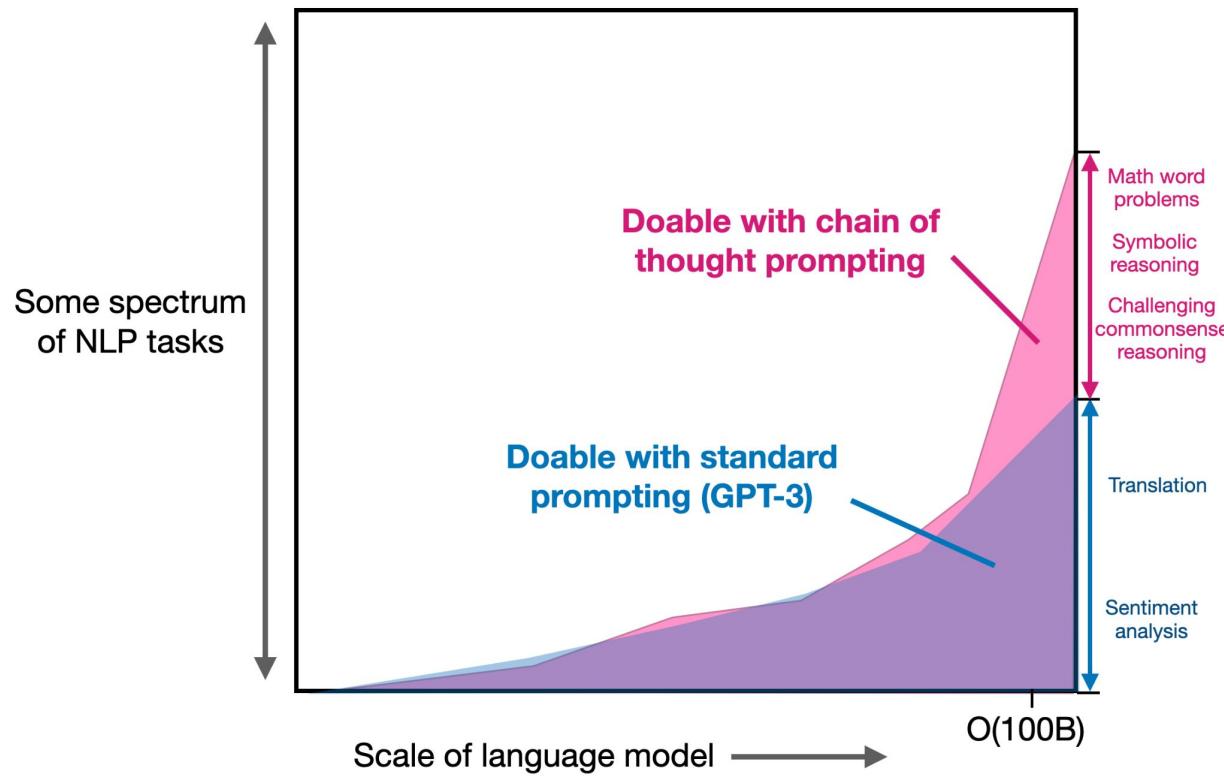
Let's calculate $A = (4 + 7 * 4 - 5) = (4 + (7 * 4) - 5) = (4 + (28) - 5) = (4 + 28 - 5) = (4 + 23) = 27$.

Let's calculate $B = (4 - 1 - 4 - 4) = ((4 - 1) - 4 - 4) = (3 - 4 - 4) = (-1 - 4) = -5$.

Then, the final equation is $A - B = 27 - (-5) = 27 + 5 = 32$. So the answer is **32**. ✓

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.

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 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 11।

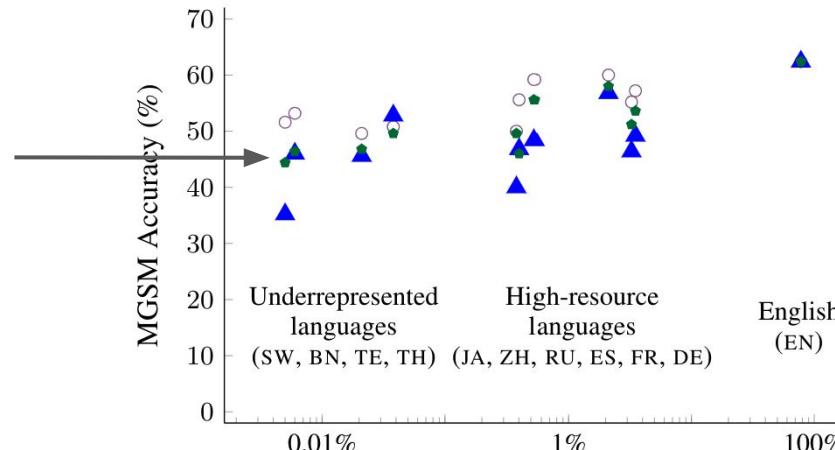
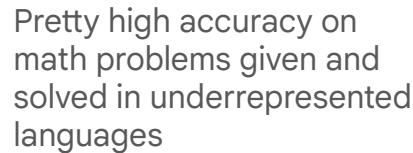
প্রশ্ন: জেনেটের হাঁসগুলি প্রতিদিন 16টি করে ডিম পাড়ে। তিনি প্রতিদিন প্রাতরাশে তিনটি করে ডিম খান এবং বন্ধুদের জন্য প্রতিদিন চারটি ডিম দিয়ে মাফিন তৈরি করেন। অবশিষ্ট হাঁসের ডিমগুলি তিনি প্রতিদিন কৃষকদের বাজারে প্রতি ডিম \$2 দরে বিক্রয় করেন। তিনি কৃষকদের বাজারে প্রতিদিন কত ডলার উপার্জন করেন?

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



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

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 :)