Natural Language Understanding, Generation, and Machine Translation

Lecture 24: Generation, In-Context Learning, and Reasoning with LLMs

Pasquale Minervini
p.minervini@ed.ac.uk
March 15th, 2024

Natural Language Generation with LLMs

Most LLMs are **auto-regressive text generation models**: at each time step t, the model gets a sequence of tokens $\{y_{<t}\}$ as input, and outputs a new token $\hat{y}_t$

Given a LM $f(\cdot)$ and vocab V, we get scores $\mathbf{s} = f\left(\{y_{<t}\}\right) \in \mathbb{R}^{|V|}$:

Natural Language Generation with LLMs

Most LLMs are **auto-regressive text generation models**: at each time step t, the model gets a sequence of tokens $\{y_{<t}\}$ as input, and outputs a new token $\hat{y}_t$

Given a LM $f(\cdot)$ and vocab V, we get scores $\mathbf{s} = f\left(\{y_{<t}\}\right) \in \mathbb{R}^{|V|}$:

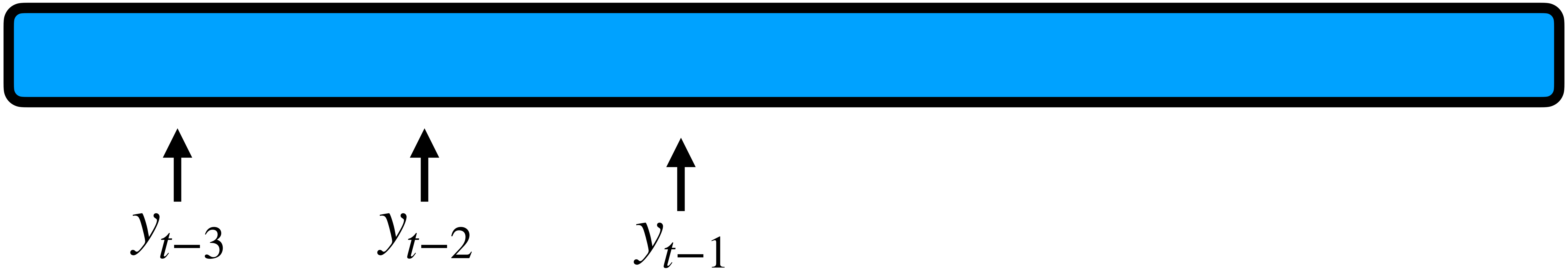
$$P\left(y_t \mid \{y_{<t}\}\right) = \frac{\exp\left(\mathbf{s}_w\right)}{\sum_{w'} \exp\left(\mathbf{s}_{w'}\right)}$$

Natural Language Generation with LLMs

Most LLMs are **auto-regressive text generation models**: at each time step t, the model gets a sequence of tokens $\{y_{<t}\}$ as input, and outputs a new token $\hat{y}_t$

Given a LM $f(\cdot)$ and vocab V, we get scores $\mathbf{s} = f\left(\{y_{<t}\}\right) \in \mathbb{R}^{|V|}$:

$$P\left(y_t \mid \{y_{<t}\}\right) = \frac{\exp\left(\mathbf{s}_w\right)}{\sum_{w'} \exp\left(\mathbf{s}_{w'}\right)}$$

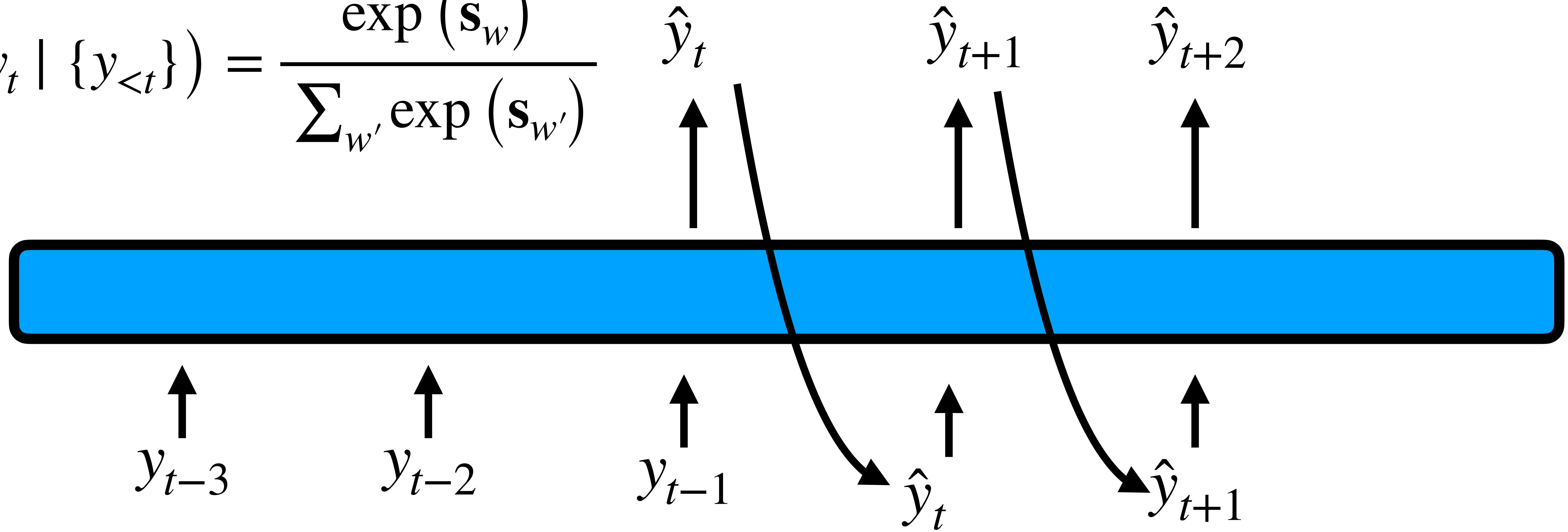


Natural Language Generation with LLMs

Most LLMs are **auto-regressive text generation models**: at each time step t, the model gets a sequence of tokens $\{y_{<t}\}$ as input, and outputs a new token $\hat{y}_t$

Given a LM $f(\cdot)$ and vocab V, we get scores $\mathbf{s} = f\left(\{y_{<t}\}\right) \in \mathbb{R}^{|V|}$:

$$P\left(y_t \mid \{y_{<t}\}\right) = \frac{\exp\left(\mathbf{s}_w\right)}{\sum_{w'} \exp\left(\mathbf{s}_{w'}\right)}$$



Natural Language Generation with LLMs

During inference, the decoding algorithm defines a function to select a token from the distribution over next tokens:

$$\hat{y}_t = \boxed{g}\left(P\left(y_t \mid \{y_{<t}\}\right)\right)$$

$g(\cdot)$ is the decoding algorithm

Natural Language Generation with LLMs

During inference, the decoding algorithm defines a function to select a token from the distribution over next tokens:

$$\hat{y}_t = \boxed{g}\left(P\left(y_t \mid \{y_{<t}\}\right)\right)$$

$g(\cdot)$ is the decoding algorithm

Greedy decoding: at each step, just select the highest-probability next token according to the model:

$$\hat{y}_t = \boxed{\arg\max_{w \in V}} P\left(y_t = w \mid \{y_{<t}\}\right)$$

This already works but — what else is there?

Decoding — Finding the Most Likely String

Greedy decoding: select the highest probability token in $P\left(y_t \mid \{y_{<t}\}\right)$:

$$\hat{y}_t = \boxed{\arg\max_{w \in V}} P\left(y_t = w \mid \{y_{<t}\}\right)$$

Beam Search: wider exploration of candidates using beam search; you saw it in the Machine Translation (MT) lectures 🙂

Decoding — Finding the Most Likely String

Greedy decoding: select the highest probability token in $P\left(y_t \mid \{y_{<t}\}\right)$:

$$\hat{y}_t = \boxed{\arg\max_{w \in V}} P\left(y_t = w \mid \{y_{<t}\}\right)$$

Beam Search: wider exploration of candidates using beam search; you saw it in the Machine Translation (MT) lectures 🙂

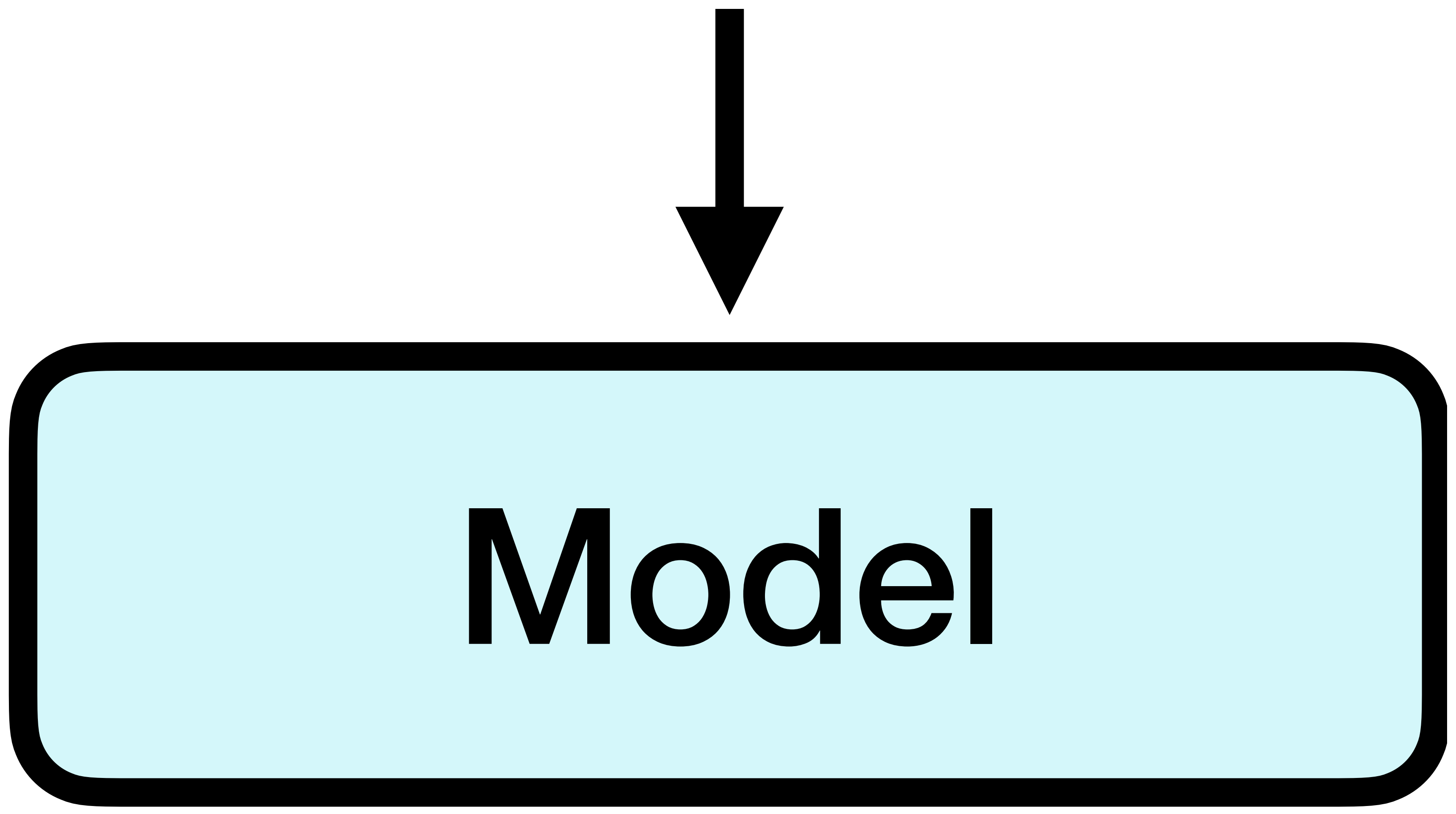
Heuristic: maximum probability decoding is good for low-entropy tasks like MT and summarisation, where the target outputs tend to be more predictable

Decoding — Generation via Sampling

We can sample a token from the next token distribution:

$$\hat{y}_t \sim P\left(y_t = w \mid \{y_{<t}\}\right)$$

I ate the pizza while it was still

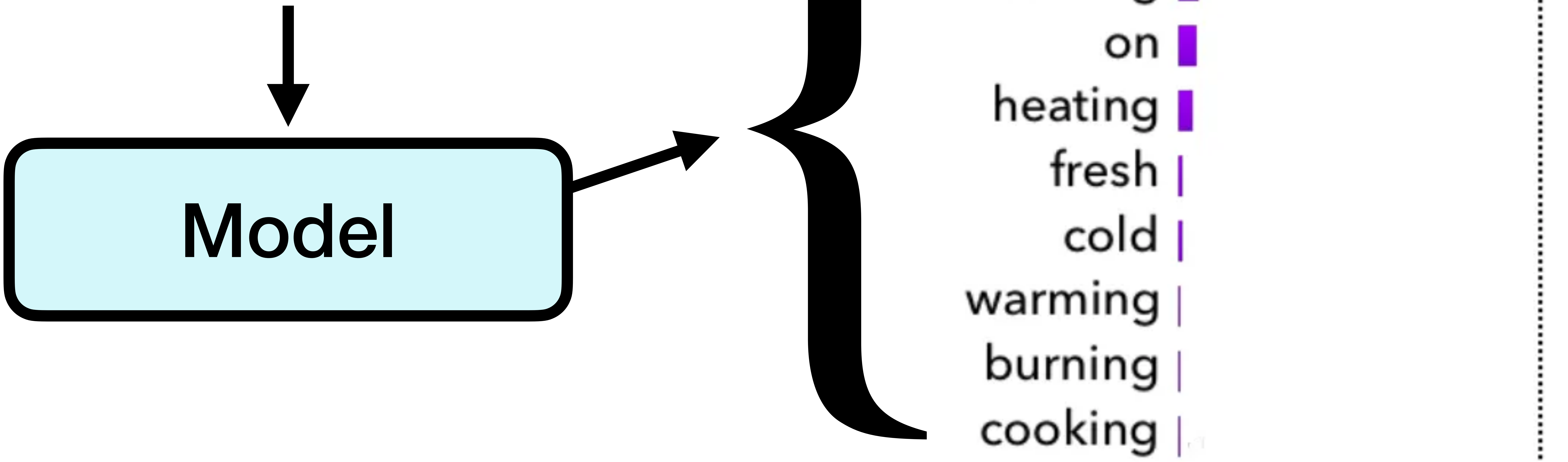


Decoding — Generation via Sampling

We can sample a token from the next token distribution:

$$\hat{y}_t \sim P\left(y_t = w \mid \{y_{<t}\}\right)$$

I ate the pizza while it was still



Decoding — Top-k Sampling

Problem with "naive" sampling: often $P\left(y_t \mid \{y_{<t}\}\right)$ is "heavy-tailed" — the tail of the distribution can be very long and, in aggregate, have considerable mass; however, some tokens are **really wrong!**

Decoding — Top-k Sampling

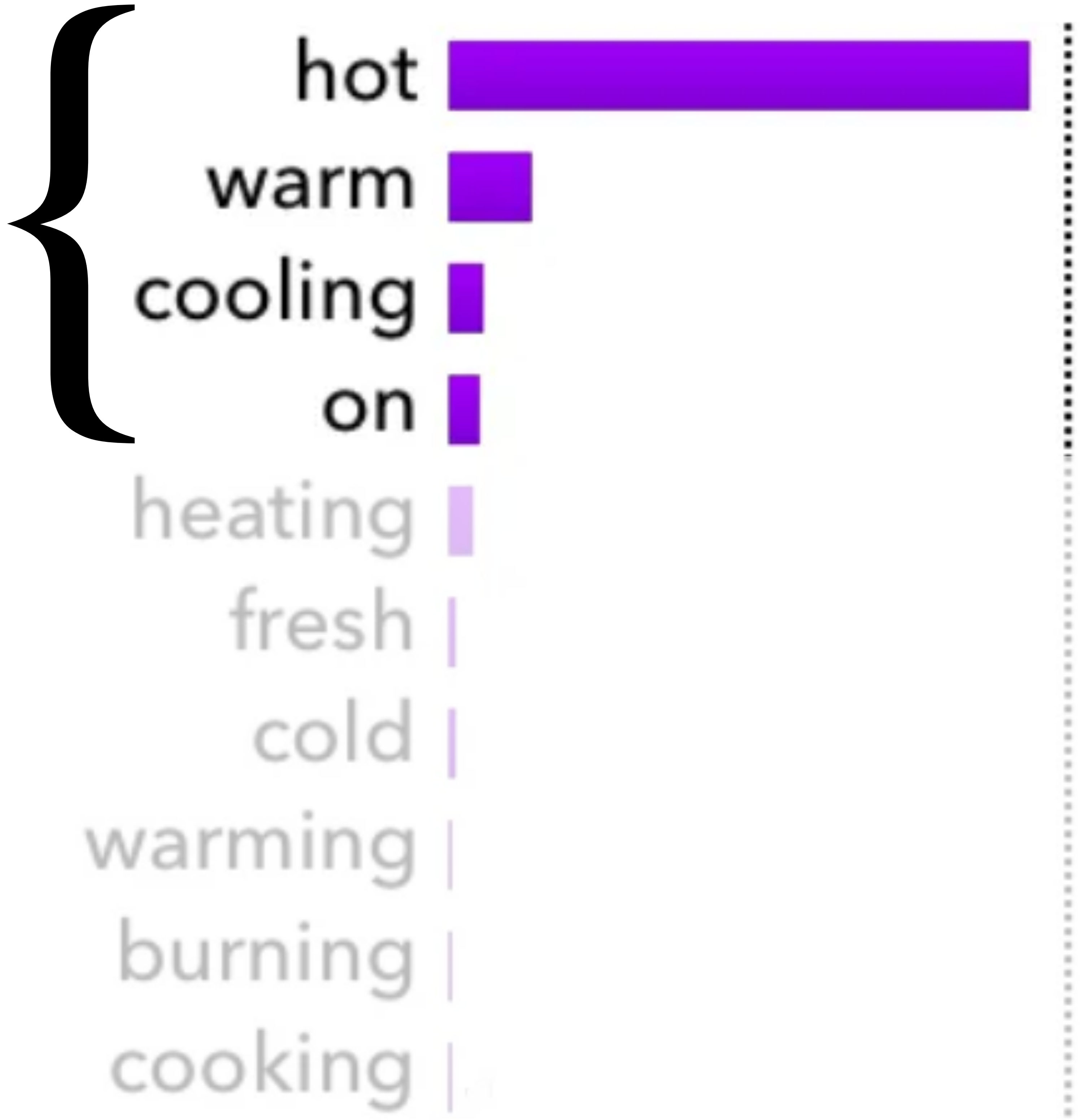
Problem with "naive" sampling: often $P\left(y_t \mid \{y_{<t}\}\right)$ is "heavy-tailed" — the tail of the distribution can be very long and, in aggregate, have considerable mass; however, some tokens are **really wrong!**

Solution: Top-k sampling — we only sample from the top k tokens!

$k = 4$

Model

I ate the pizza while it was still



Decoding — Top-k Sampling

Problem with Top-k sampling: the cut-off can be **too quick/slow** —

- When P is flatter, a small k can remove too many viable options
- When P is sharper, a high k can allow for too many options to have a chance of being selected

Decoding — Top-p Sampling

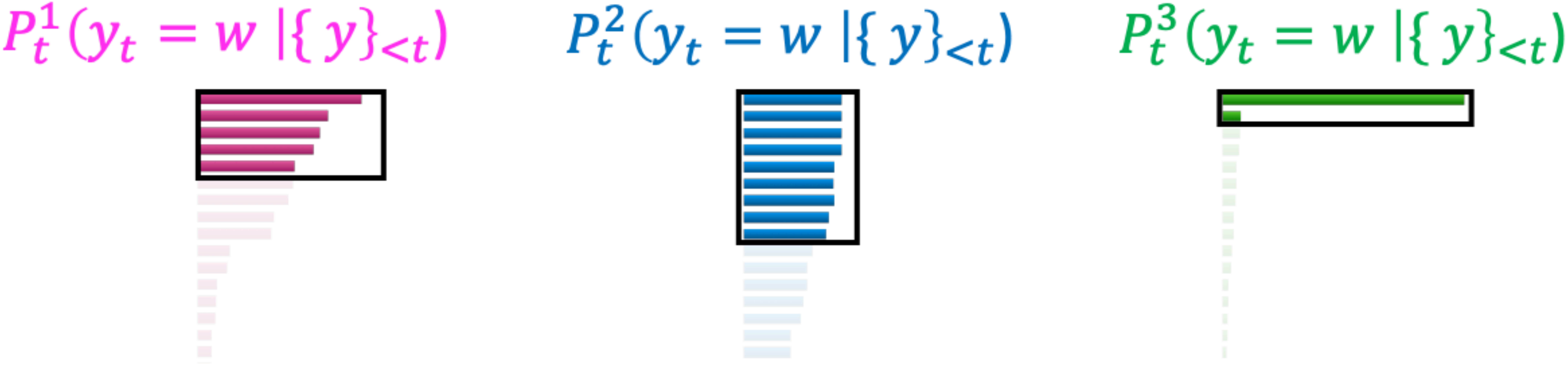
Problem with Top-k sampling: the cut-off can be **too quick/slow** –

- When P is flatter, a small k can remove too many viable options
- When P is sharper, a high k can allow for too many options to have a chance of being selected

Solution: Top-p Sampling (*Nucleus sampling* [Holtzman et al., 2020]) – sample from all tokens in the top-p cumulative probability mass

Decoding — Top-p Sampling

Solution: Top-p Sampling (*Nucleus sampling* [Holtzman et al., 2020]) — sample from all tokens in the top-p cumulative probability mass



Decoding — Temperature

Recap: at timestep t, the model computes a distribution $P\left(y_t \mid \{y_{<t}\}\right)$ over possible next tokens

$$P\left(y_t \mid \{y_{<t}\}\right) = \frac{\exp\left(\mathbf{s}_w\right)}{\sum_{w'} \exp\left(\mathbf{s}_{w'}\right)} \text{ with } \mathbf{s} \in \mathbb{R}^{|V|}$$

Decoding — Temperature

Recap: at timestep t, the model computes a distribution $P\left(y_t \mid \{y_{<t}\}\right)$ over possible next tokens

$$P\left(y_t \mid \{y_{<t}\}\right) = \frac{\exp\left(\mathbf{s}_w\right)}{\sum_{w'} \exp\left(\mathbf{s}_{w'}\right)} \text{ with } \mathbf{s} \in \mathbb{R}^{|V|}$$

We can apply a **temperature hyper-parameter** τ to re-balance P:

$$P\left(y_t \mid \{y_{<t}\}\right) = \frac{\exp\left(\mathbf{s}_w/\tau\right)}{\sum_{w'} \exp\left(\mathbf{s}_{w'}/\tau\right)} \text{ with } \mathbf{s} \in \mathbb{R}^{|V|}$$

When $\tau > 1$, then P becomes **more uniform**

When $\tau < 1$, then P becomes **more concentrated/spiky**

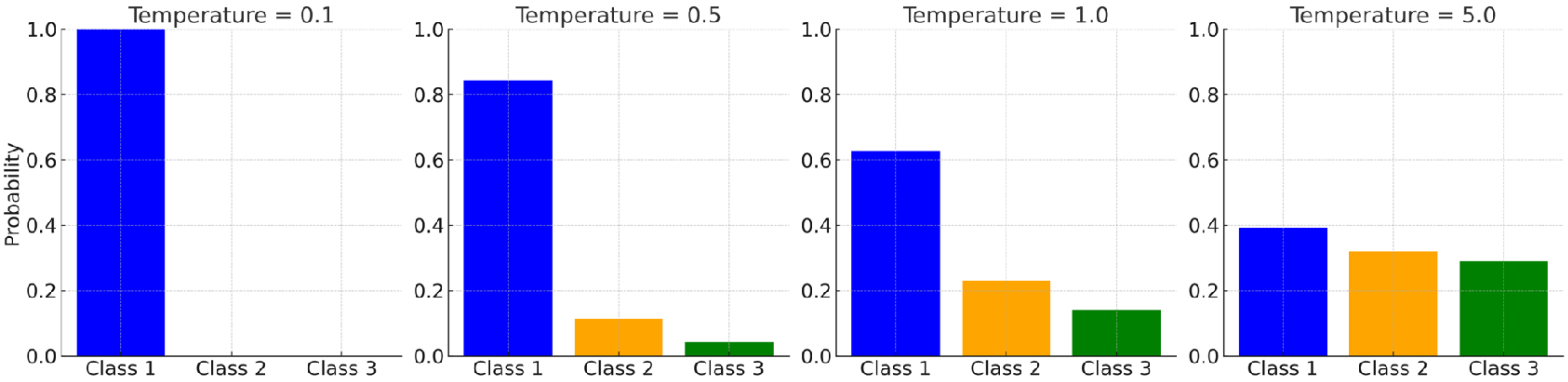
Decoding — Temperature

We can apply a **temperature hyper-parameter** τ to re-balance P:

$$P\left(y_t \mid \{y_{<t}\}\right) = \frac{\exp\left(\mathbf{s}_w/\tau\right)}{\sum_{w'} \exp\left(\mathbf{s}_{w'}/\tau\right)} \quad \text{with } \mathbf{s} \in \mathbb{R}^{|V|}$$

When $\tau > 1$, then P becomes **more uniform**

When $\tau < 1$, then P becomes **more concentrated/spiky**



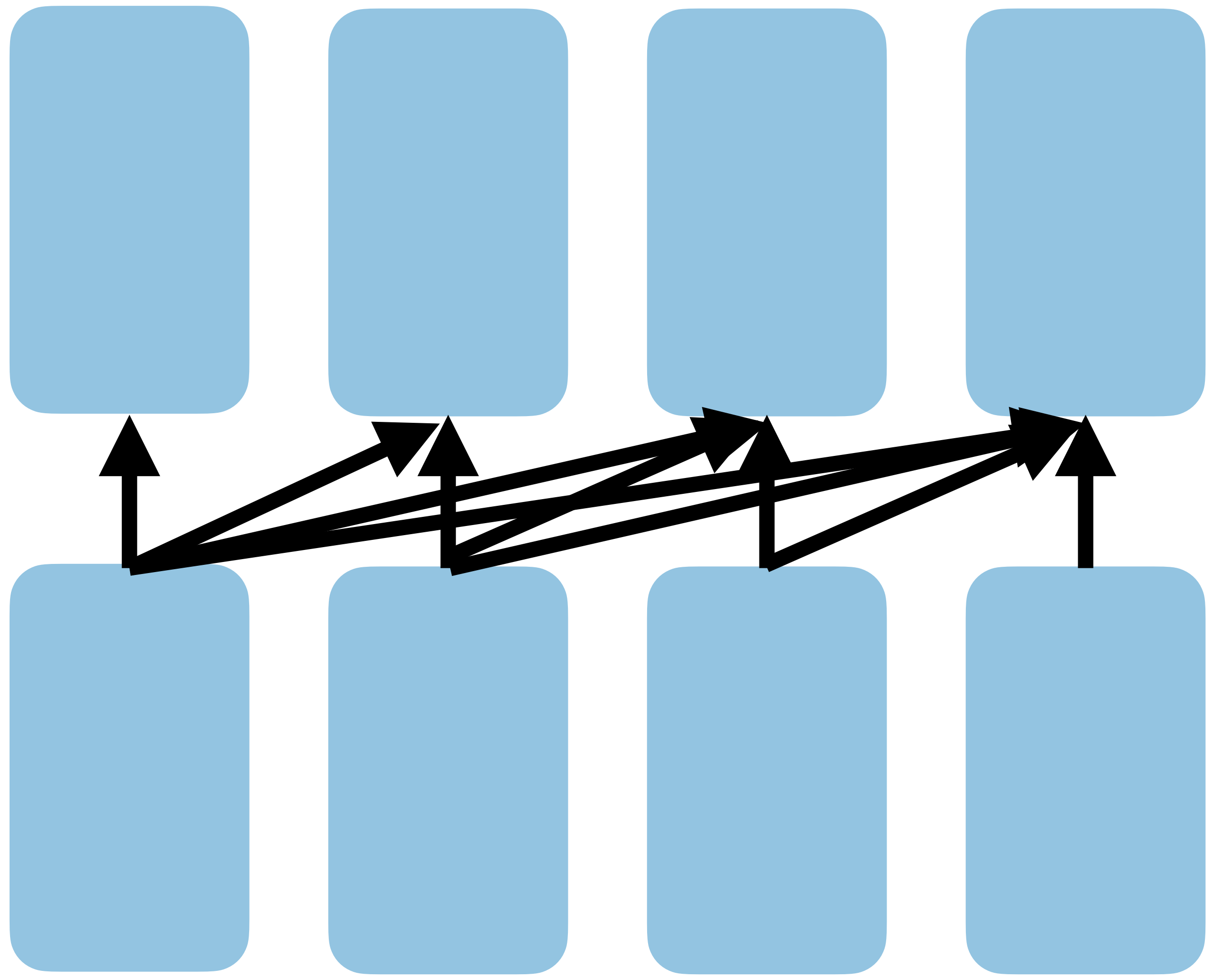
Back to In-Context Learning!

Emerging Abilities of LLMs: GPT-3 (2020)

GPT-3 [Brown et al., 2020]:

- Parameter increase: 1.5B → **175B**
- Trained or more data: (40GB → **>600GB**)

Recap



Language Models are Few-Shot Learners

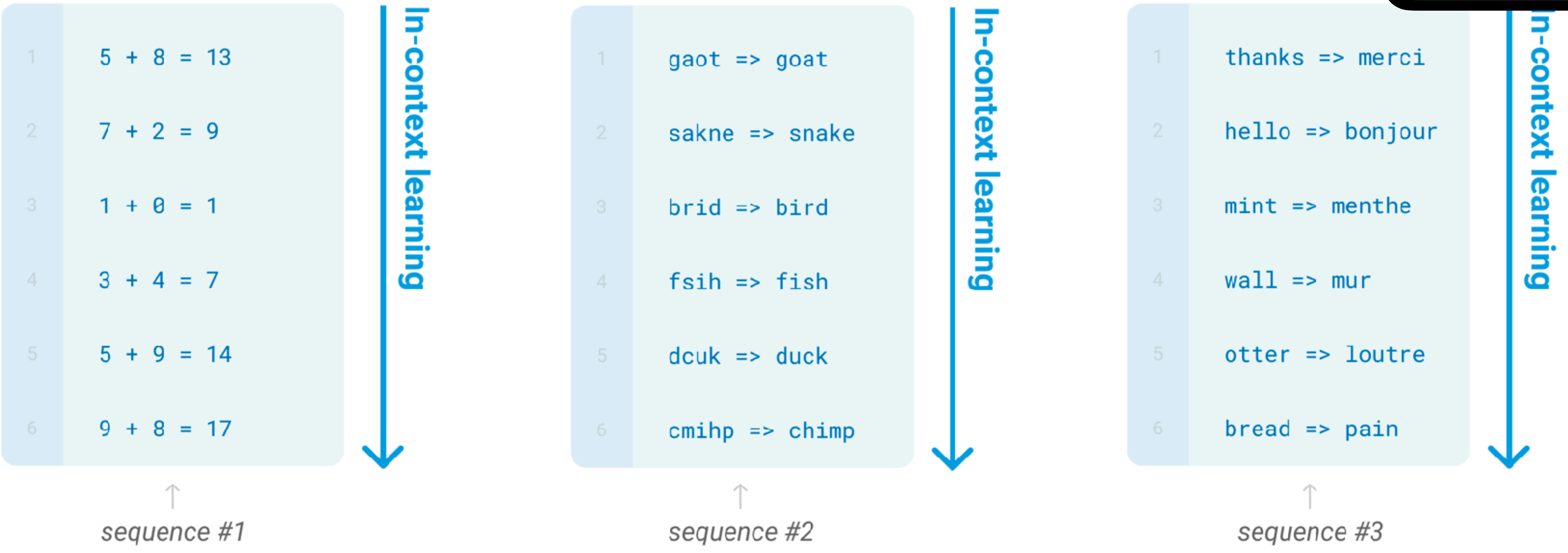
Tom B. Brown* **Benjamin Mann*** **Nick Ryder*** **Melanie Subbiah***

Emergent Few-Shot Learning Abilities

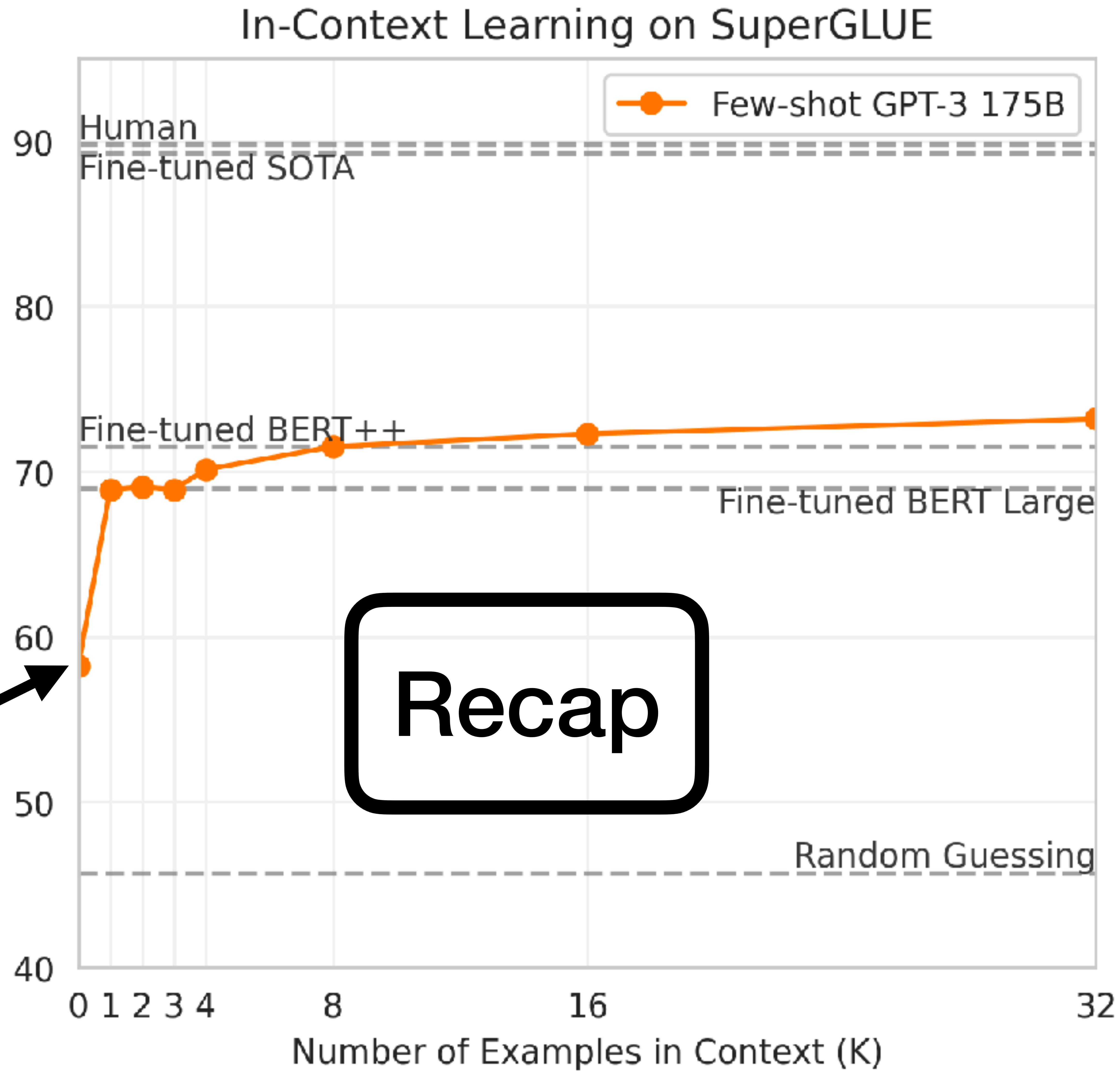
Specify a task by pre-pending examples of the task before your input

Referred to as **In-Context Learning** — we can teach the model a new task *without performing any gradient updates*

Recap



Emergent Few-Shot Learning Abilities



Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:      ← task description
2  cheese =>                         ← prompt
```

Recap

Emergent Few-Shot Learning Abilities

One-shot

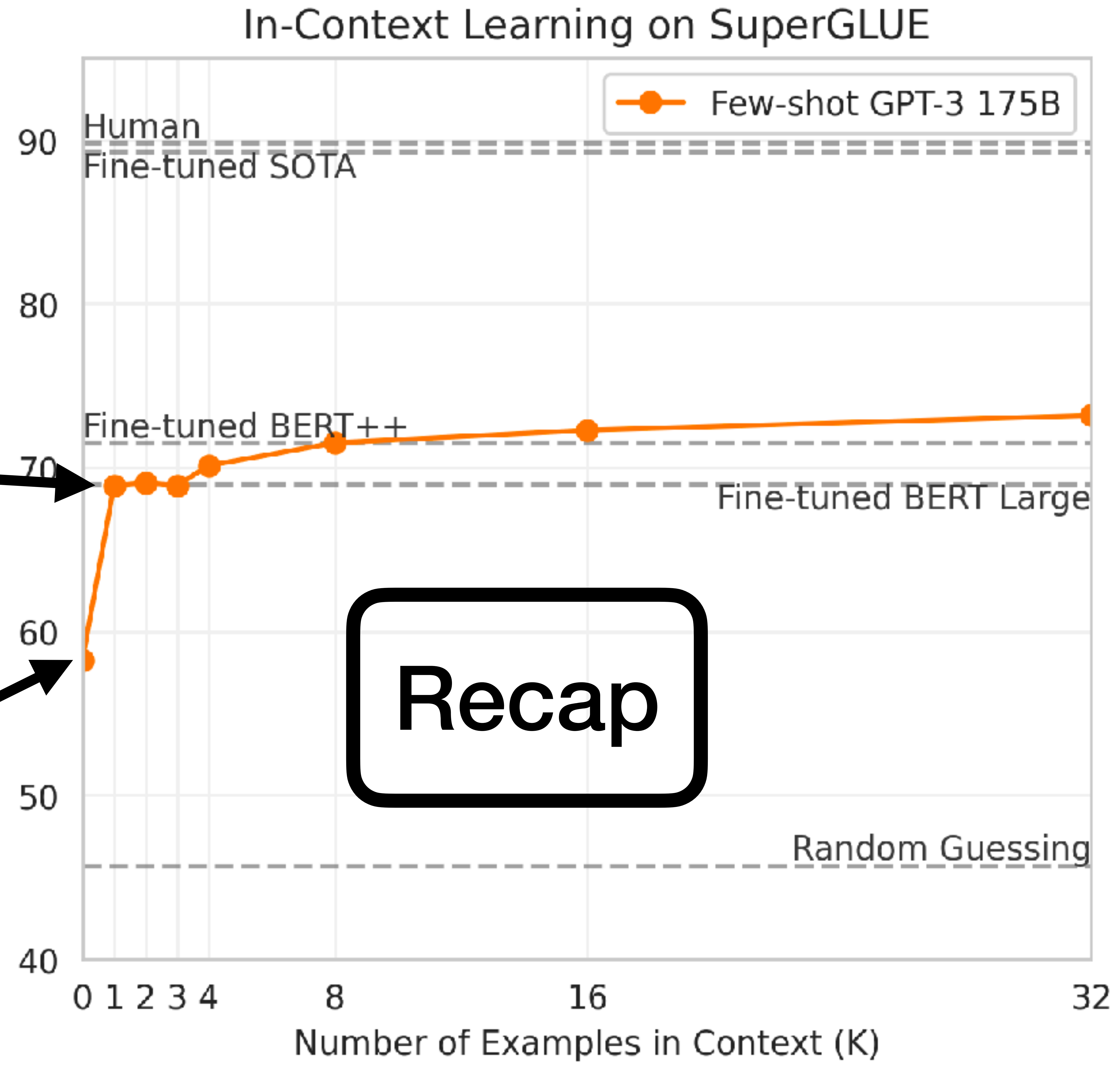
In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:      ← task description
2  sea otter => loutre de mer        ← example
3  cheese =>                         ← prompt
```

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:      ← task description
2  cheese =>                         ← prompt
```



Emergent Few-Shot Learning Abilities

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← examples
peppermint => menthe poivrée        ←
plush girafe => girafe peluche      ←
cheese => ..................        ← prompt
```

One-shot

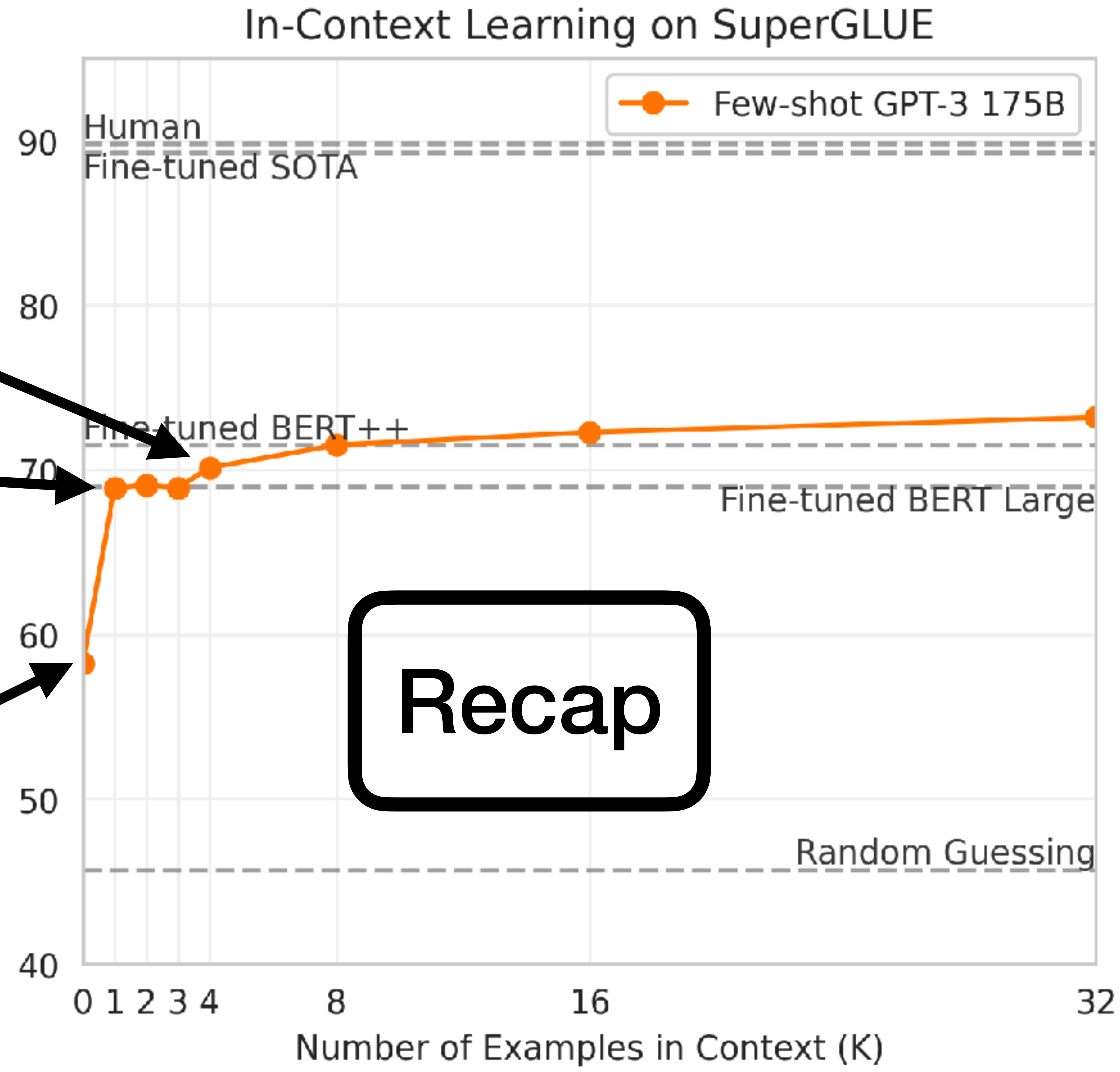
In addition to th
example of the

```
Translat
sea otte
cheese =>  ..................       ← prompt
```

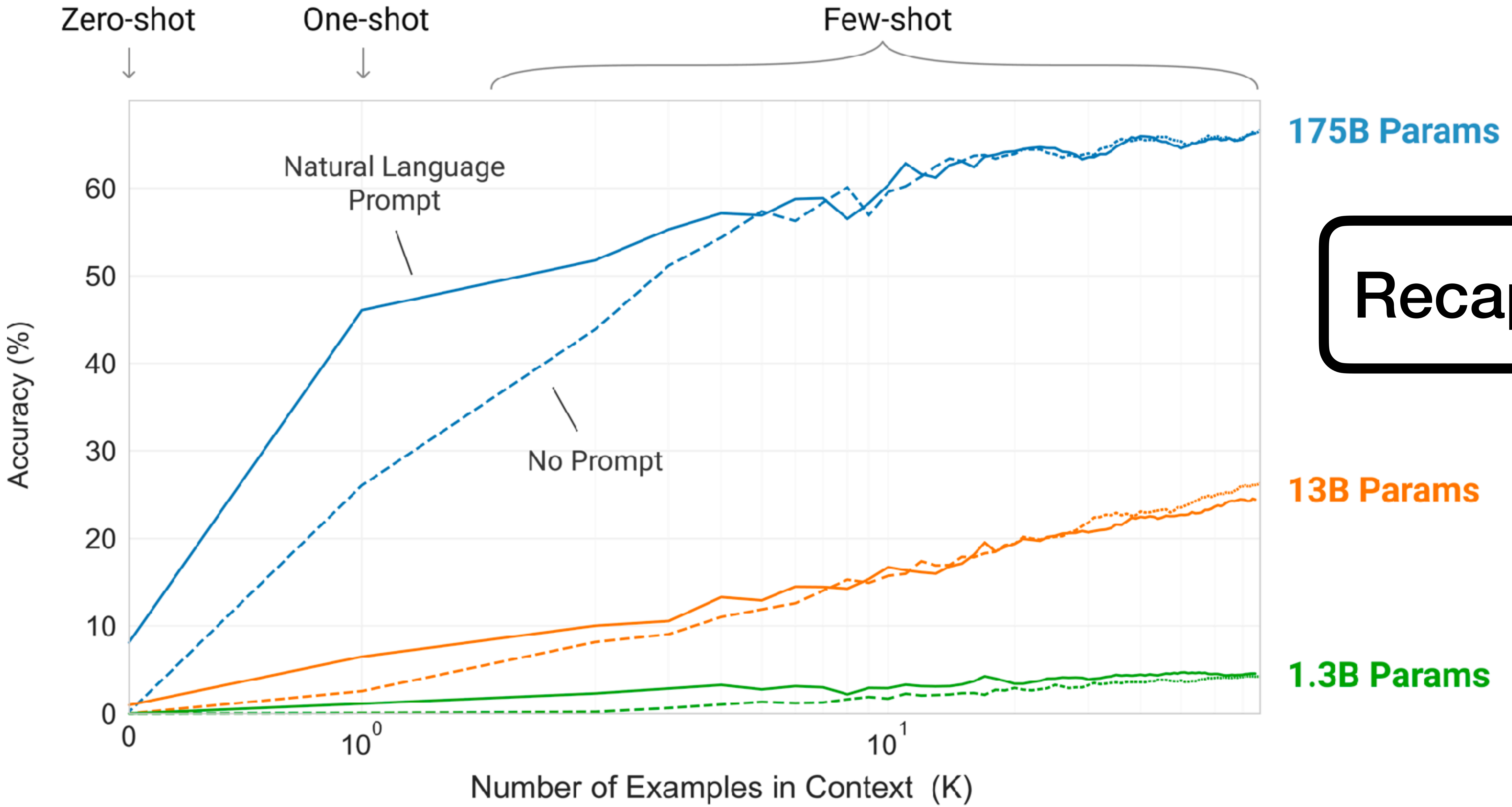
Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
cheese => ..................        ← prompt
```

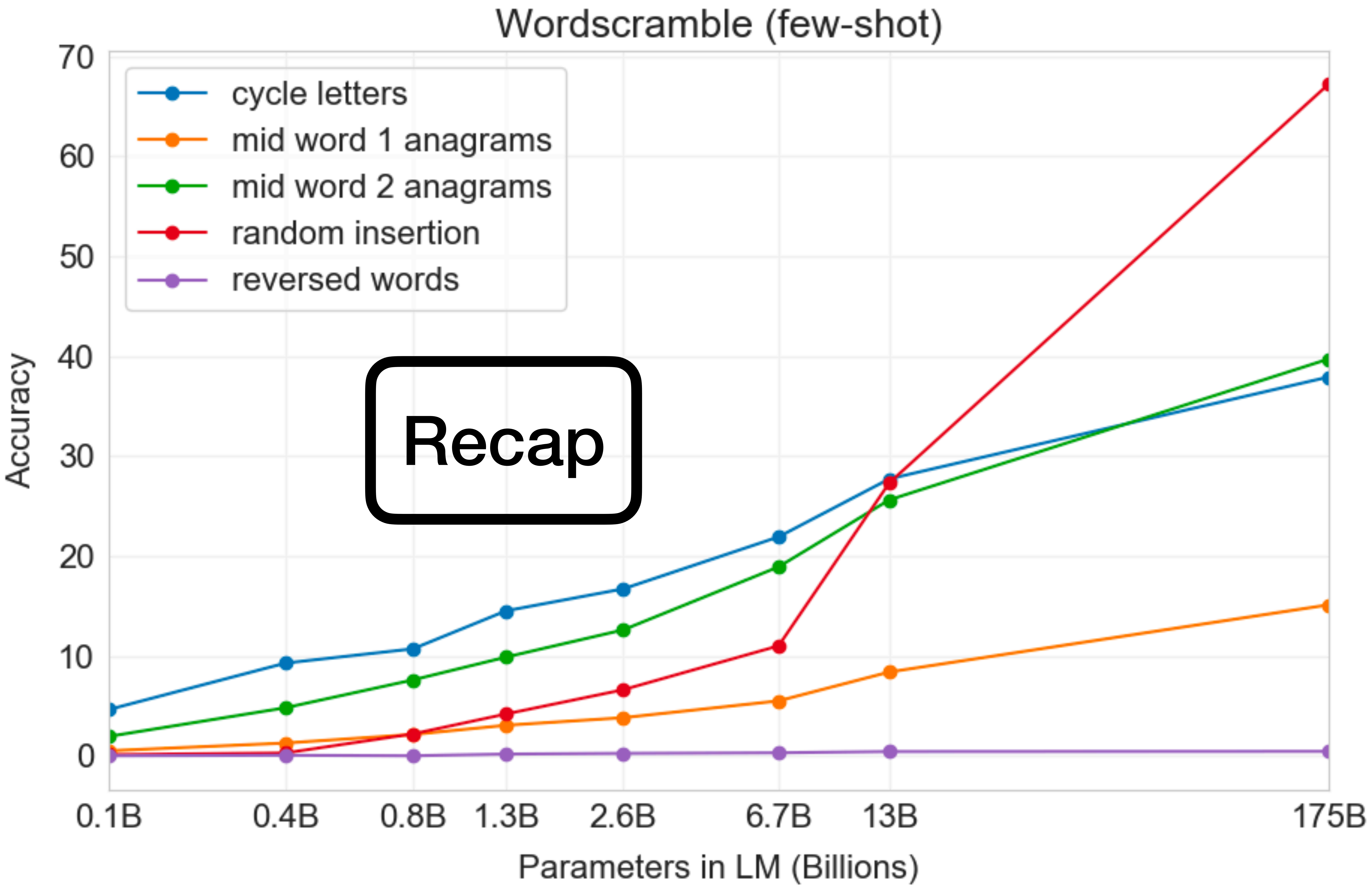


Emergent Few-Shot Learning Abilities

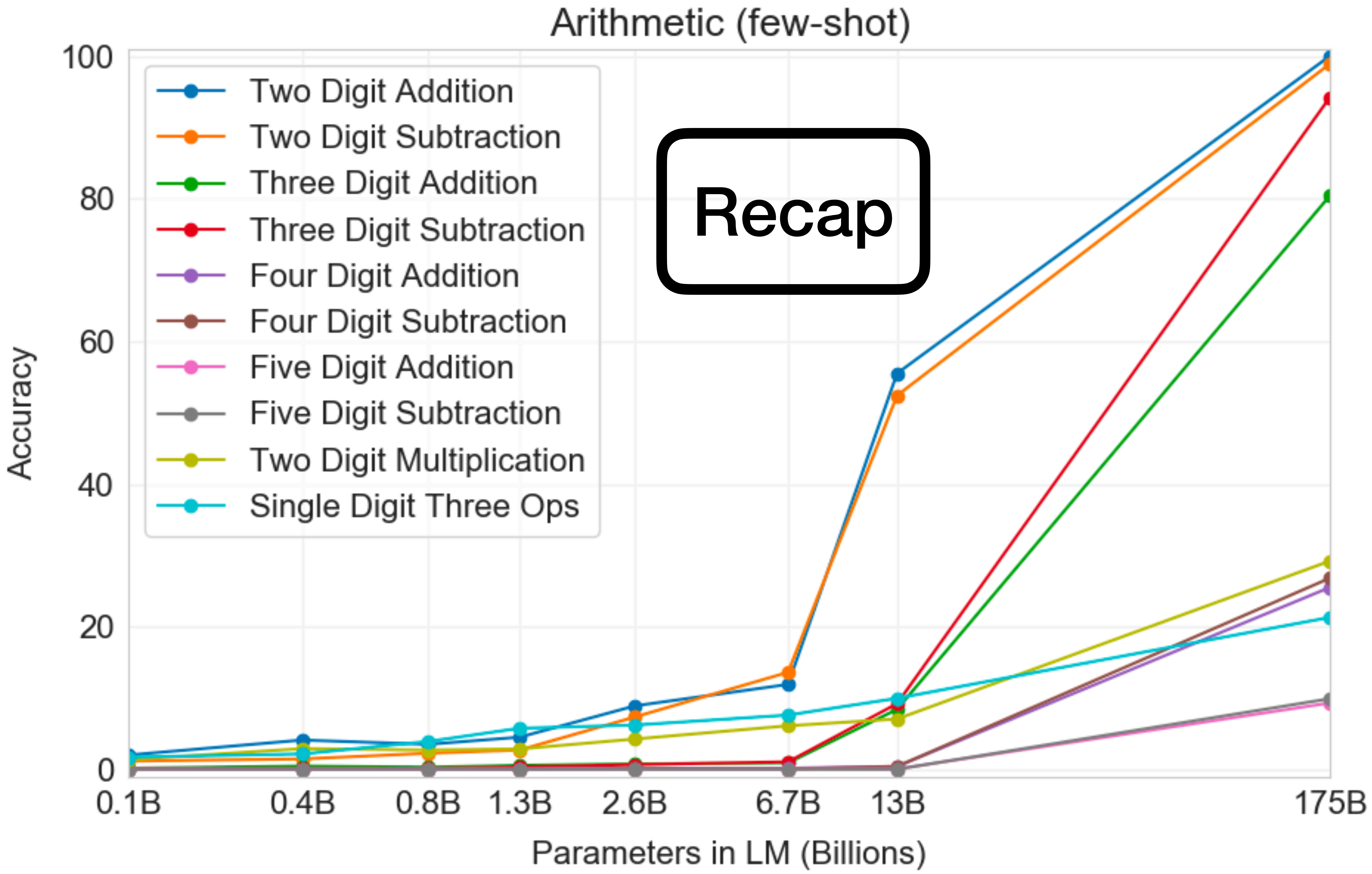


Recap

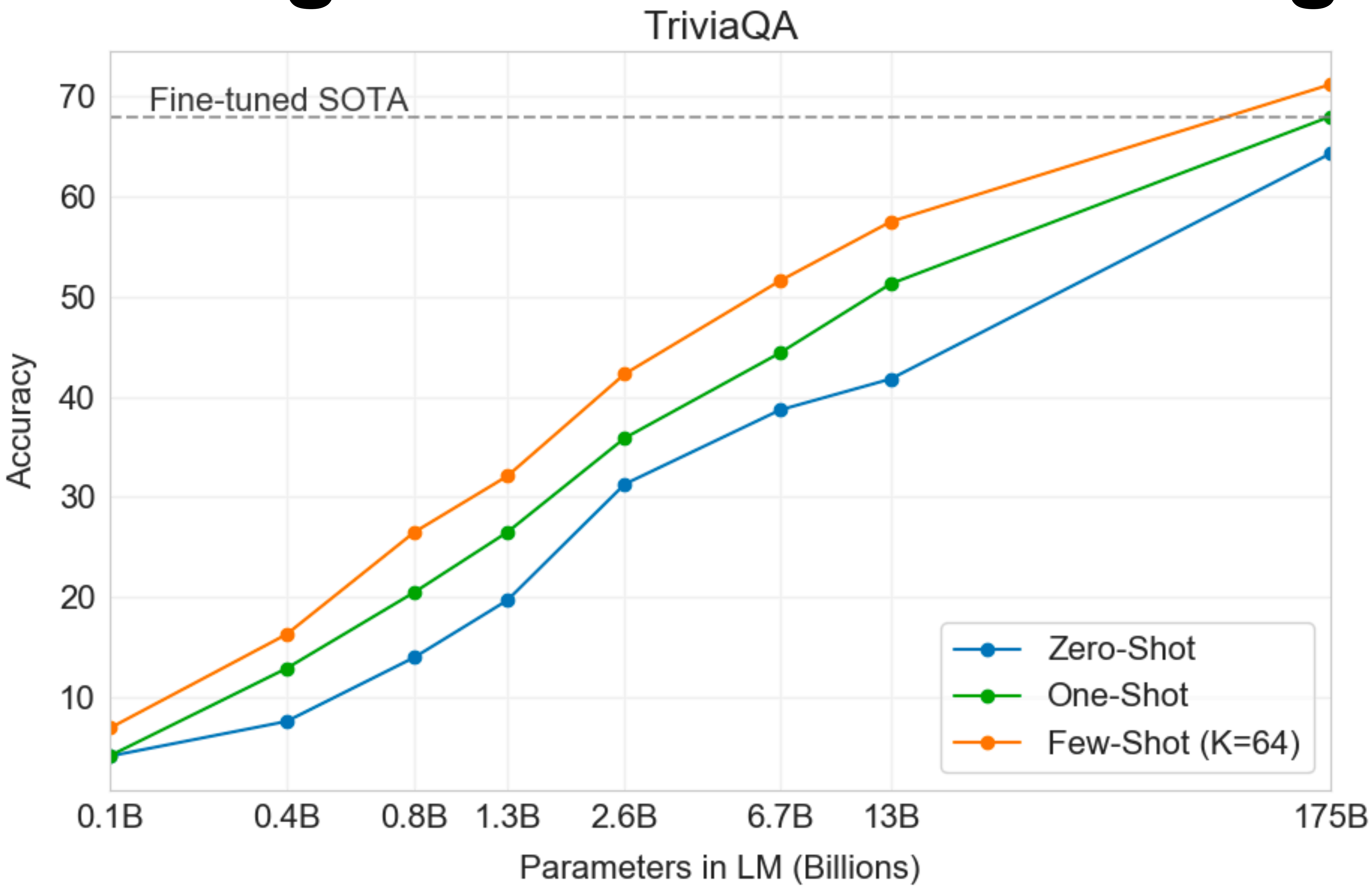
Emergent Few-Shot Learning Abilities



Emergent Few-Shot Learning Abilities

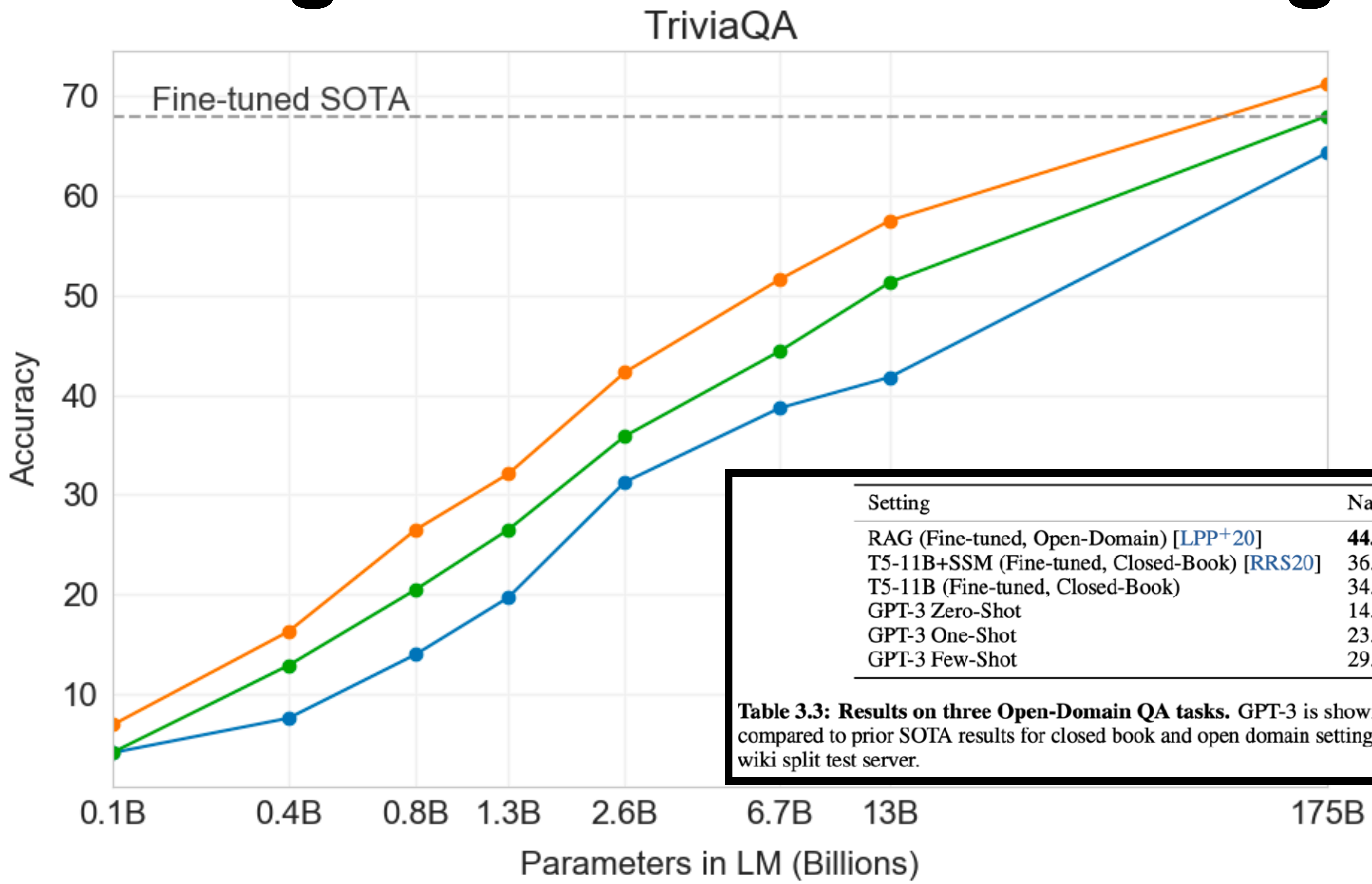


Emergent Few-Shot Learning Abilities



Recap

Emergent Few-Shot Learning Abilities



Recap

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP+20]	**44.5**	**45.5**	**68.0**
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	**68.0**
GPT-3 Few-Shot	29.9	41.5	**71.2**

Table 3.3: Results on three Open-Domain QA tasks. GPT-3 is shown in the few-, one-, and zero-shot settings, as compared to prior SOTA results for closed book and open domain settings. TriviaQA few-shot result is evaluated on the wiki split test server.

In-Context Learning

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // ________

LM

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // ________

LM

Sensitivity to Prompts

Poor template choices can severely degrade model performance

Possible solutions: Template Ensembles [Voronov et al., 2024], Global and Local Entropy of the predictions [Lu et al., 2022]

[Voronov et al., 2024]

ICL Fails on Complex Reasoning Tasks

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: The answer is **5**

Arithmetic Reasoning (AR)
(+ − ×÷…)

ICL Fails on Complex Reasoning Tasks

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: The answer is **5**

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

A: The answer is **nk.**

Arithmetic Reasoning (AR)
$(+ - \times \div \ldots)$

Symbolic Reasoning (SR)

ICL Fails on Complex Reasoning Tasks

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: The answer is **5**

Arithmetic Reasoning (AR)
(+ − ×÷…)

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

A: The answer is **nk.**

Symbolic Reasoning (SR)

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

A: The answer is **(c).**

Commonsense Reasoning (CR)

Arithmetic Problems

Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

72

Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

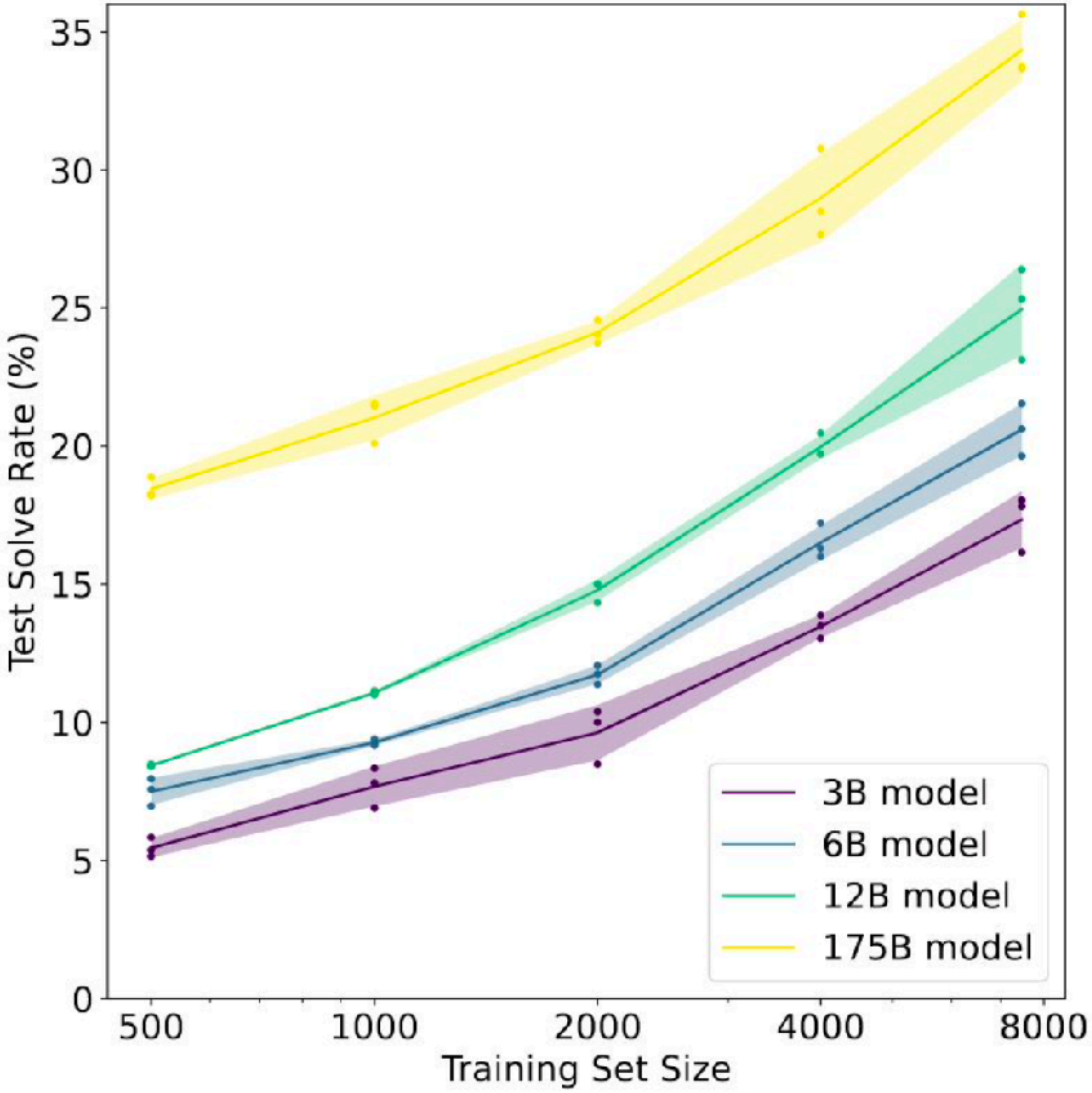
10

James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?

624

Results for GPT-3 fine-tuned on GSM8K [Cobbe et al., 2021]

Arithmetic Problems



Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?

72

Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?

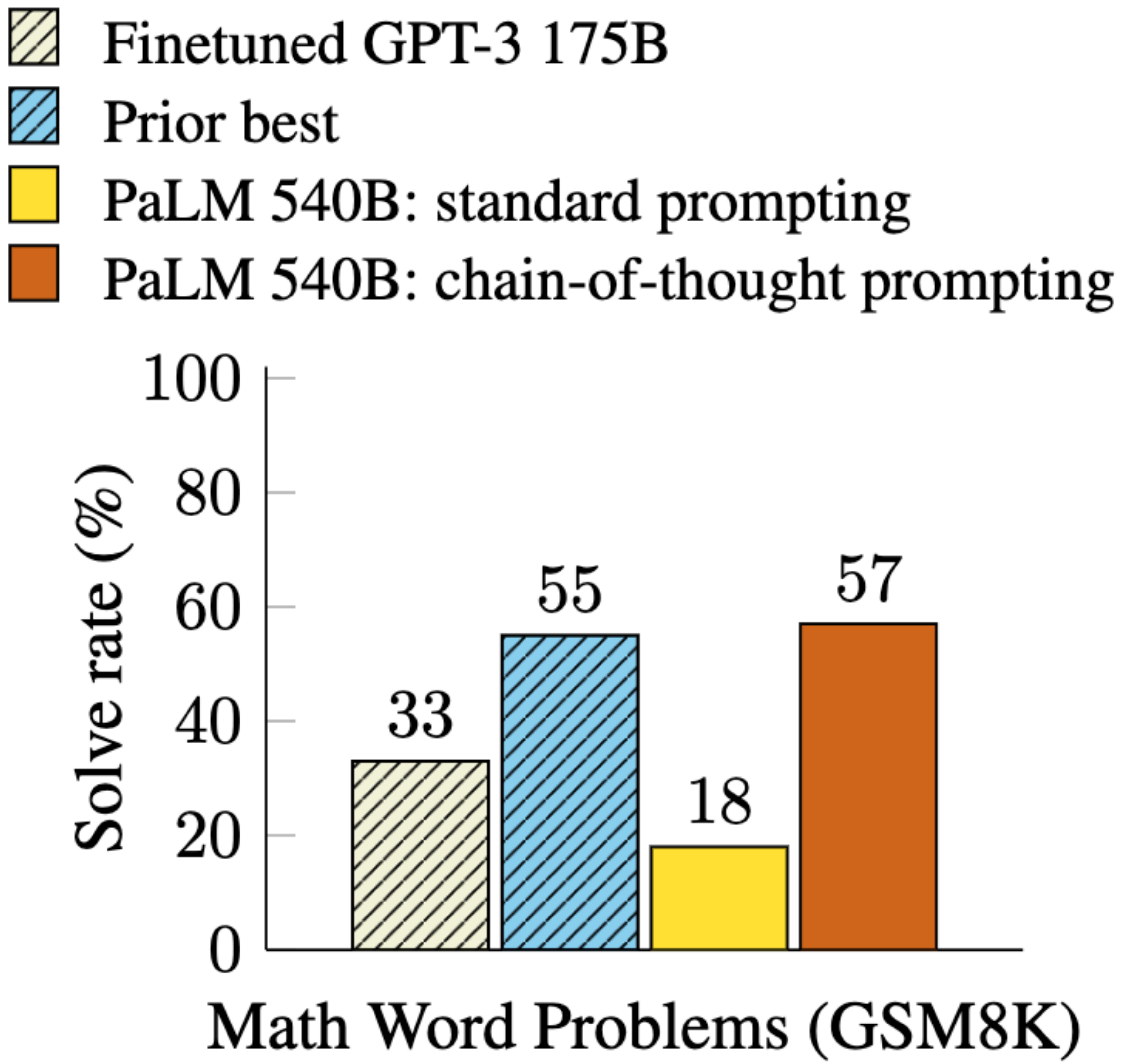
10

James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?

624

Results for GPT-3 fine-tuned on GSM8K [Cobbe et al., 2021]

Arithmetic Problems



Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May?
72

Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn?
10

James writes a 3-page letter to 2 different friends twice a week. How many pages does he write a year?
624

Results for PaLM 540B “trained” via ICL on GSM8K

ICL Fails on Complex Reasoning Tasks

Scaling up LMs does not efficiently achieve accurate results in **Arithmetic Reasoning** (*AR*), **Commonsense Reasoning** (*CR*), and **Symbolic Reasoning** (*SR*) tasks

Proposed solution: **Chain of Thought Prompting**

Chain of Thought Prompting

Few-Shot

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**

Zero-Shot

Large Language Models are Zero-Shot Reasoners

Takeshi Kojima
The University of Tokyo
t.kojima@weblab.t.u-tokyo.ac.jp

Shixiang Shane Gu
Google Research, Brain Team

Machel Reid
The University of Tokyo

Yutaka Matsuo
The University of Tokyo

Yusuke Iwasawa
The University of Tokyo

Chain of Thought Prompting

Definition: a **Chain of Thought** is a **series of intermediate natural language reasoning steps** that lead to the final output

Chain of Thought Prompting

Definition: a **Chain of Thought** is a **series of intermediate natural language reasoning steps** that lead to the final output

In a way, it is similar to the **backward-chaining reasoning algorithm** from logic programming, where complex task are (recursively) decomposed into simpler tasks

Chain of Thought Prompting

Definition: a **Chain of Thought** is a **series of intermediate natural language reasoning steps** that lead to the final output

In a way, it is similar to the **backward-chaining reasoning algorithm** from logic programming, where complex task are (recursively) decomposed into simpler tasks

It can provide several benefits:

- Intermediate problems can be easier to solve for a LLM

Chain of Thought Prompting

Definition: a **Chain of Thought** is a **series of intermediate natural language reasoning steps** that lead to the final output

In a way, it is similar to the **backward-chaining reasoning algorithm** from logic programming, where complex task are (recursively) decomposed into simpler tasks

It can provide several benefits:

- Intermediate problems can be easier to solve for a LLM
- The reasoning step can provide an **explanation** for the prediction

Chain of Thought Prompting

Definition: a **Chain of Thought** is a **series of intermediate natural language reasoning steps** that lead to the final output

In a way, it is similar to the **backward-chaining reasoning algorithm** from logic programming, where complex task are (recursively) decomposed into simpler tasks

It can provide several benefits:

- Intermediate problems can be easier to solve for a LLM
- The reasoning step can provide an **explanation** for the prediction
- Only requires **inference** with a LLM — no fine-tuning!

Chain of Thought Prompting

Few-shot CoT [Wei et al., 2022]

Standard Prompting

Model 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

Few-shot CoT [Wei et al., 2022]

Standard Prompting

Model 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

Model 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?

Chain of Thought Prompting

Few-shot CoT [Wei et al., 2022]

Standard Prompting

Model 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

Model 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.

Chain of Thought Prompting

Zero-shot CoT [Kojima et al., 2022]

Chain of Thought Prompting

Zero-shot CoT [Kojima et al., 2022]

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. **X**

Chain of Thought Prompting

Zero-shot CoT [Kojima et al., 2022]

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ***X***

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

Chain of Thought Prompting

Zero-shot CoT [Kojima et al., 2022]

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ***X***

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ***X***

Chain of Thought Prompting

Zero-shot CoT [Kojima et al., 2022]

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. **X**

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 **X**

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Chain of Thought Prompting

Zero-shot CoT [Kojima et al., 2022]

【1st prompt】
Reasoning Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?
A: Let's think step by step.

Chain of Thought Prompting

Zero-shot CoT [Kojima et al., 2022]

【1st prompt】
Reasoning Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?
A: Let's think step by step.

LLM

In one minute, Joe throws 25 punches.
In three minutes, Joe throws 3 * 25 = 75 punches.
In five rounds, Joe throws 5 * 75 = 375 punches.

Chain of Thought Prompting

Zero-shot CoT [Kojima et al., 2022]

【1st prompt】
Reasoning Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?
A: Let's think step by step.

LLM

In one minute, Joe throws 25 punches.
In three minutes, Joe throws 3 * 25 = 75 punches.
In five rounds, Joe throws 5 * 75 = 375 punches.

【2nd prompt】
Answer Extraction

Q: On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 ･･･
A: Let's think step by step.

In one minute, Joe throws 25 punches. ･･･In five rounds, Joe throws 5 * 75 = 375 punches. .
Therefore, the answer (arabic numerals) is

Chain of Thought Prompting

Zero-shot CoT [Kojima et al., 2022]

CoT Prompting — Experiments

Free Response

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: **There are originally 3 cars. 2 more cars arrive. 3 + 2 = 5.** The answer is **5.**

- Eight handcrafted examples
- All with equations with flexible formats
- Benchmarked on:
  - **GSM8K** [Cobbe et al., 2021]
  - **SVAMP** [Patel et al., 2021]
  - **MAWPS** [Koncel-Kedziorski et al., 2016]

Few-shot CoT [Wei et al., 2022]

CoT Prompting — Experiments

Multiple Choice

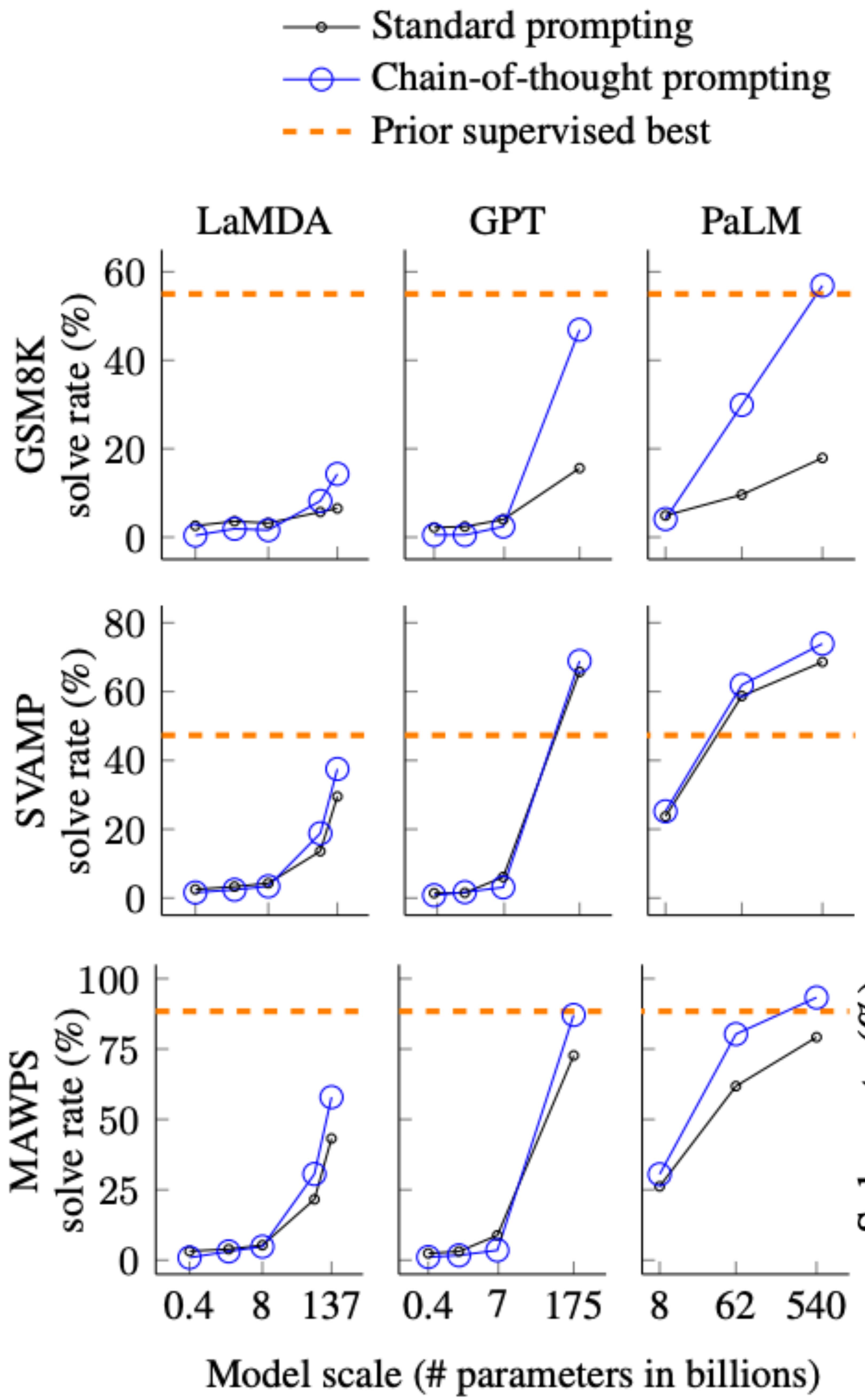
Q: A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance?
Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km

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

- Four exemplars, whose questions, intermediate reasoning, and answers are sampled from the **training set**
- Exemplars have flexible formats
- Benchmarked on:
  - **AQuA-RAT** [Ling et al., 2017]

Few-shot CoT [Wei et al., 2022]

CoT Prompting — Results

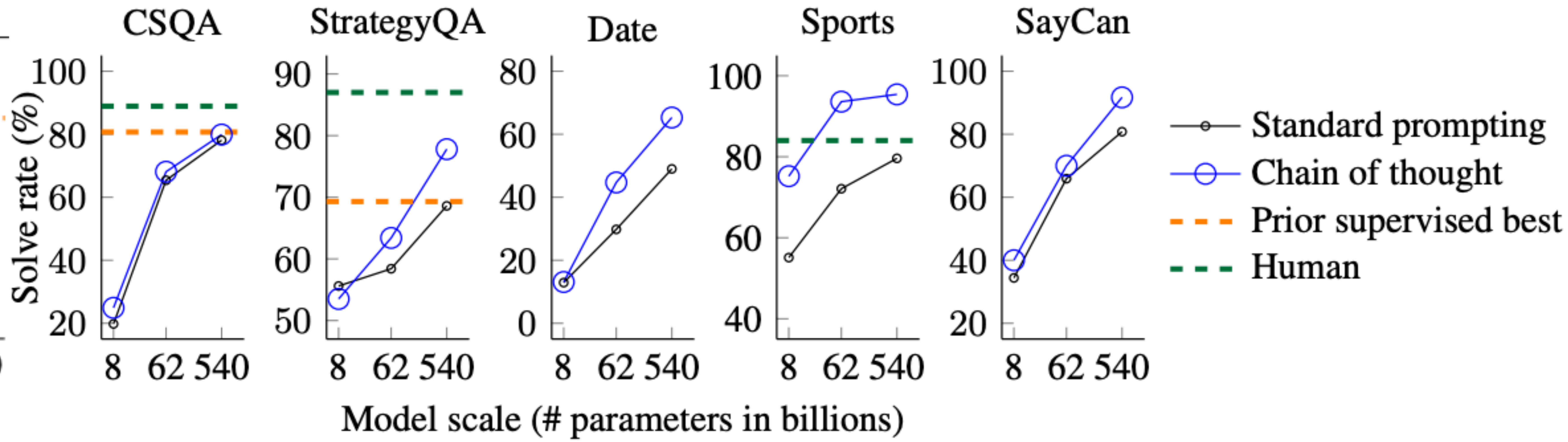


GSM8K
Josh decides to try flipping a house. He buys a house for \$80,000 and then puts in \$50,000 in repairs. This increased the value of the house by 150%. How much profit did he make?

MAWPS - MultiArith
The school cafeteria ordered 42 red apples and 7 green apples for students lunches. But, if only 9 students wanted fruit, how many extra did the cafeteria end up with?

SVAMP
Each pack of dvds costs 76 dollars. If there is a discount of 25 dollars on each pack. How much do you have to pay to buy each pack?

AQuA-RAT
A person is traveling at 20 km/hr and reached his destiny in 2.5 hr then find the distance? Answer Choices: (a) 53 km (b) 55 km (c) 52 km (d) 60 km (e) 50 km

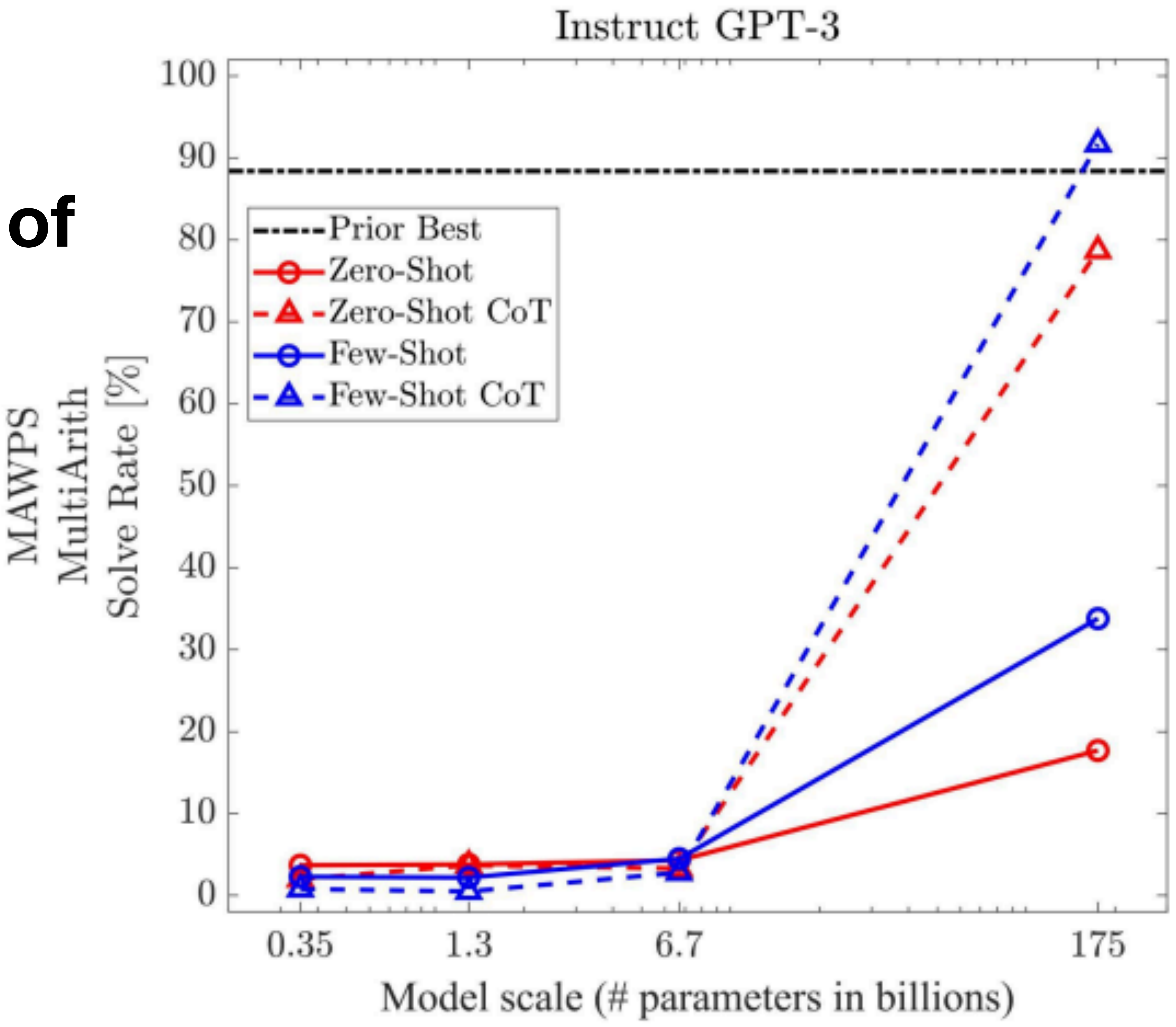


Arithmetic Reasoning — Observations

Both zero-shot and few-shot CoT prompting are **emergent abilities of model scale**

CoT does not help for smaller models, while it yields **significant improvements** for models with **>100B parameters**

Few-shot CoT achieves better accuracy than zero-shot CoT



Symbolic Reasoning — Last Letter Concatenation

Last letter concatenation

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

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

Generate full names by **randomly concatenating** names from the top names from name census data

Four exemplars with **strict format**

Symbolic Reasoning — Coin Flip

Coin Flip

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

A: **The coin was flipped by no one. So the coin was flipped 0 times. The coin started heads up, and it was not flipped, so it is still heads up.** So the answer is **yes**.

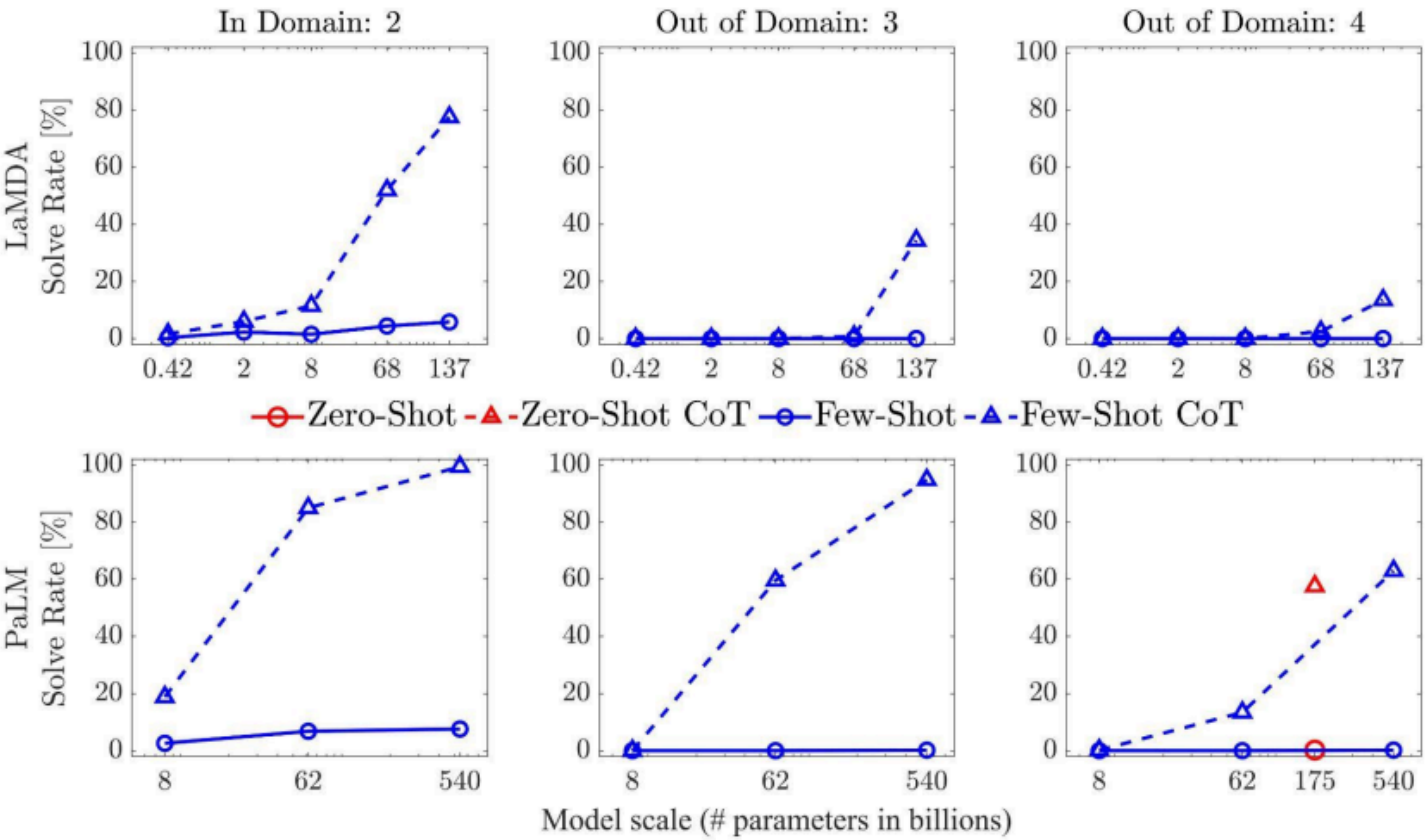
Eight exemplars with **strict format**

In & Out-of-Domain Tests

In-domain Test Set: test examples have the same number of steps as the few-shot training examples

Out-of-Domain (OOD) Test Set: examples have more steps than the few-shot training examples

Symbolic Reasoning — Last Letter Concatenation



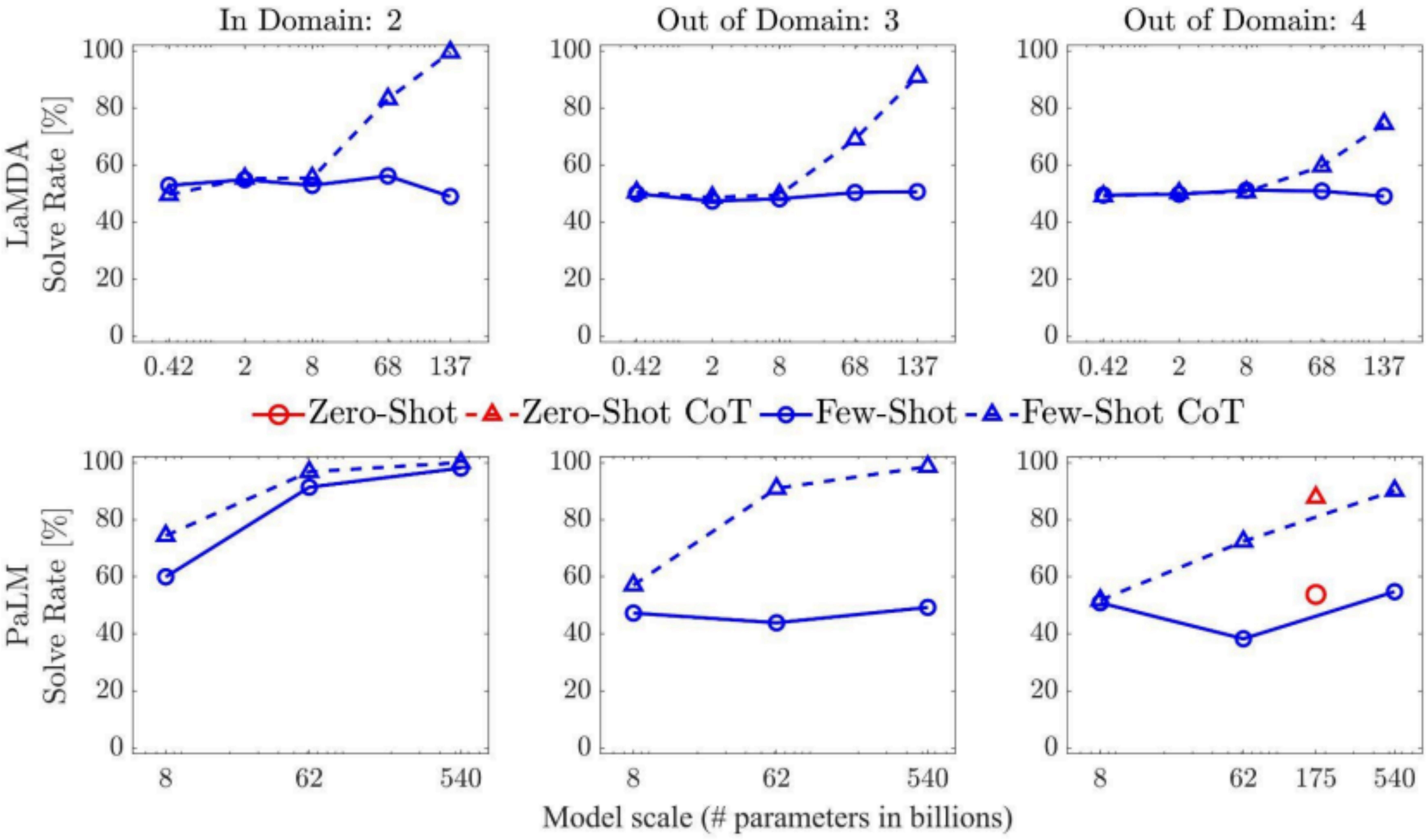
In-Domain

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

Out-of-Domain

Take the last letters of the words in "**Johann Sebastian Bach**" and concatenate them.

Symbolic Reasoning — Coin Flip



In-Domain

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

Out-of-Domain

A coin is heads up. **Tom does not flip the coin**. ***Mike does not flip the coin***. **Jake flips the coin**. Is the coin still heads up?