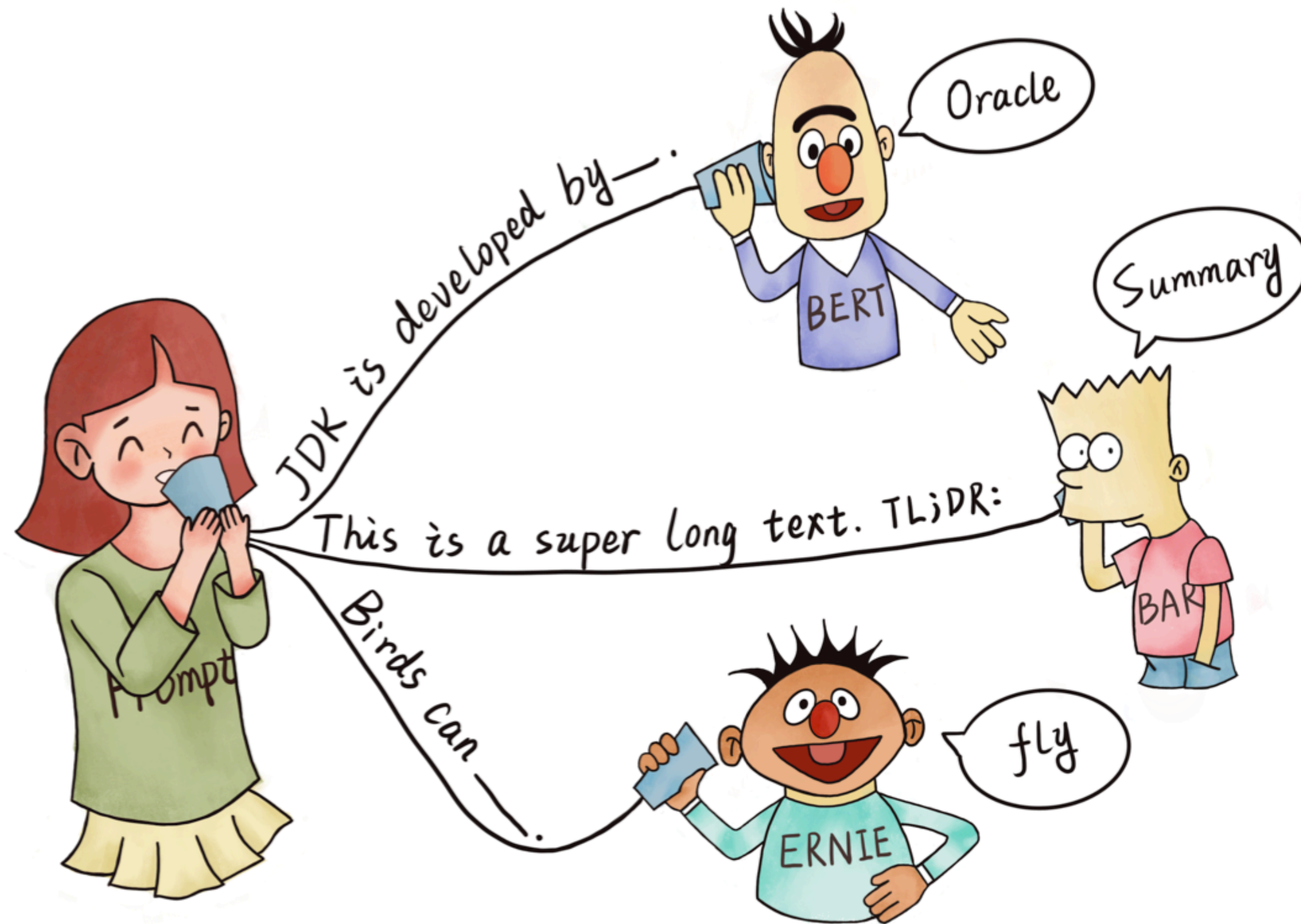


Section 5: Prompt Engineering

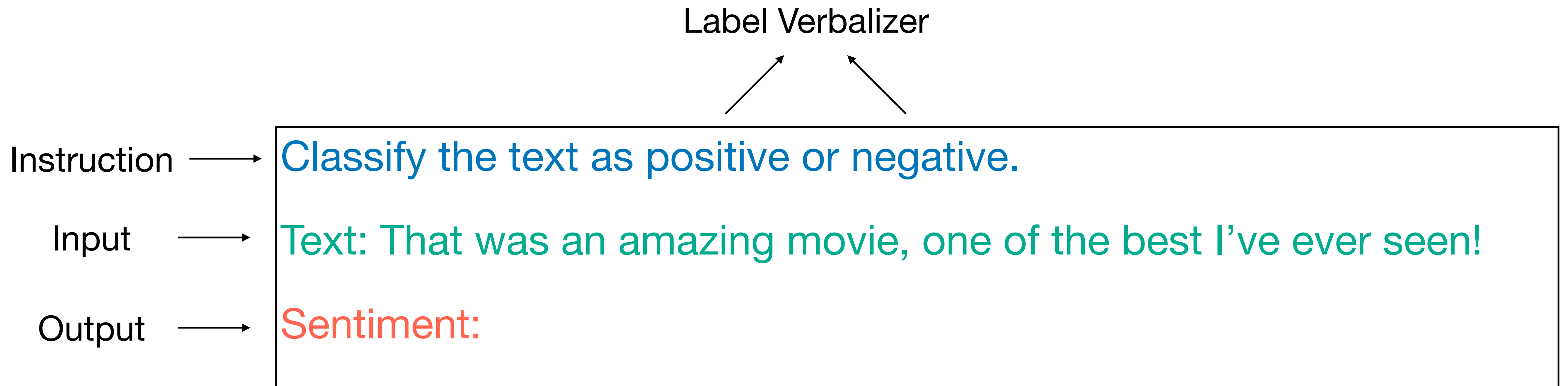
Nitish Joshi, 4th April 2023

What is prompt engineering?



Design prompts to efficiently use Language Models (LMs) for diverse applications

Elements of a prompt



This simple design can be used for lots of different task!!

Task: Summarization

Input

Scarsdale, N.Y., a village about 20 miles north of New York City known for Tudor-style architecture and large, lavish estates, may seem like an unusual setting for an aging, five-story parking garage that neighbors have described as “an eye sore,” “decrepit,” “unsafe” and “seedy.”

But for over 40 years the site has survived multiple attempts to raze and redevelop it. The latest push, in which the village is considering plans to build hundreds of apartments there, including some that would have been affordable to people with lower incomes, has been in limbo for three years after some Scarsdale residents complained that new residents could strain schools and burden taxpayers.

Instruction

Summarize the above in one sentence:

Task: Question Answering

Instruction

→ Answer yes or no.

Input

→ Question: Is Destin FL on the Gulf of Mexico?

Output

→ Context: Destin FL is located on a peninsula separating the Gulf of Mexico from Choctawhatchee Bay. The peninsula was originally an island; hurricanes and sea level changes gradually connected the island to the mainland.

→ Answer:

Few-shot Prompting

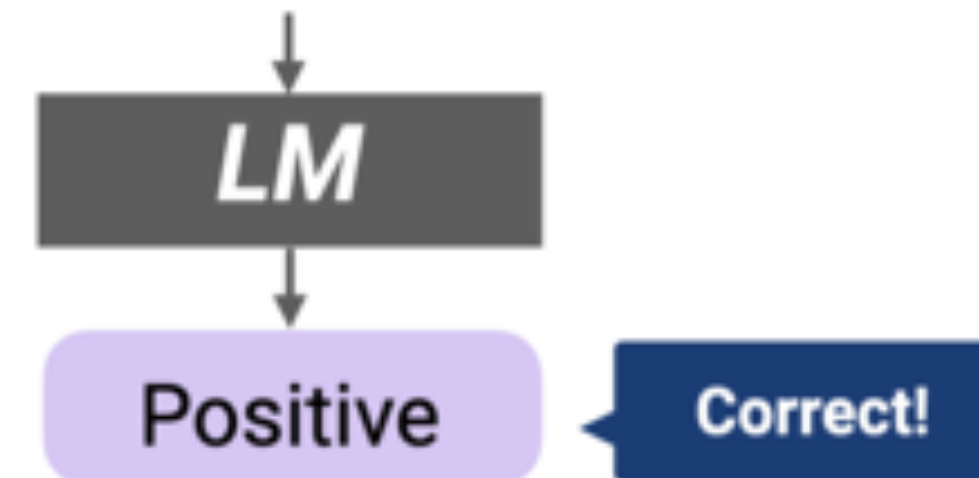
- Simple prompting with instruction + input might not be enough (especially for smaller models)
 - Poor performance
 - Incorrect output format
- **Idea:** In-context learning using few-shot examples
 - Introduced by the GPT-3 paper (Brown et al., 2020)

Few-shot Prompting

Few-shot demonstrations

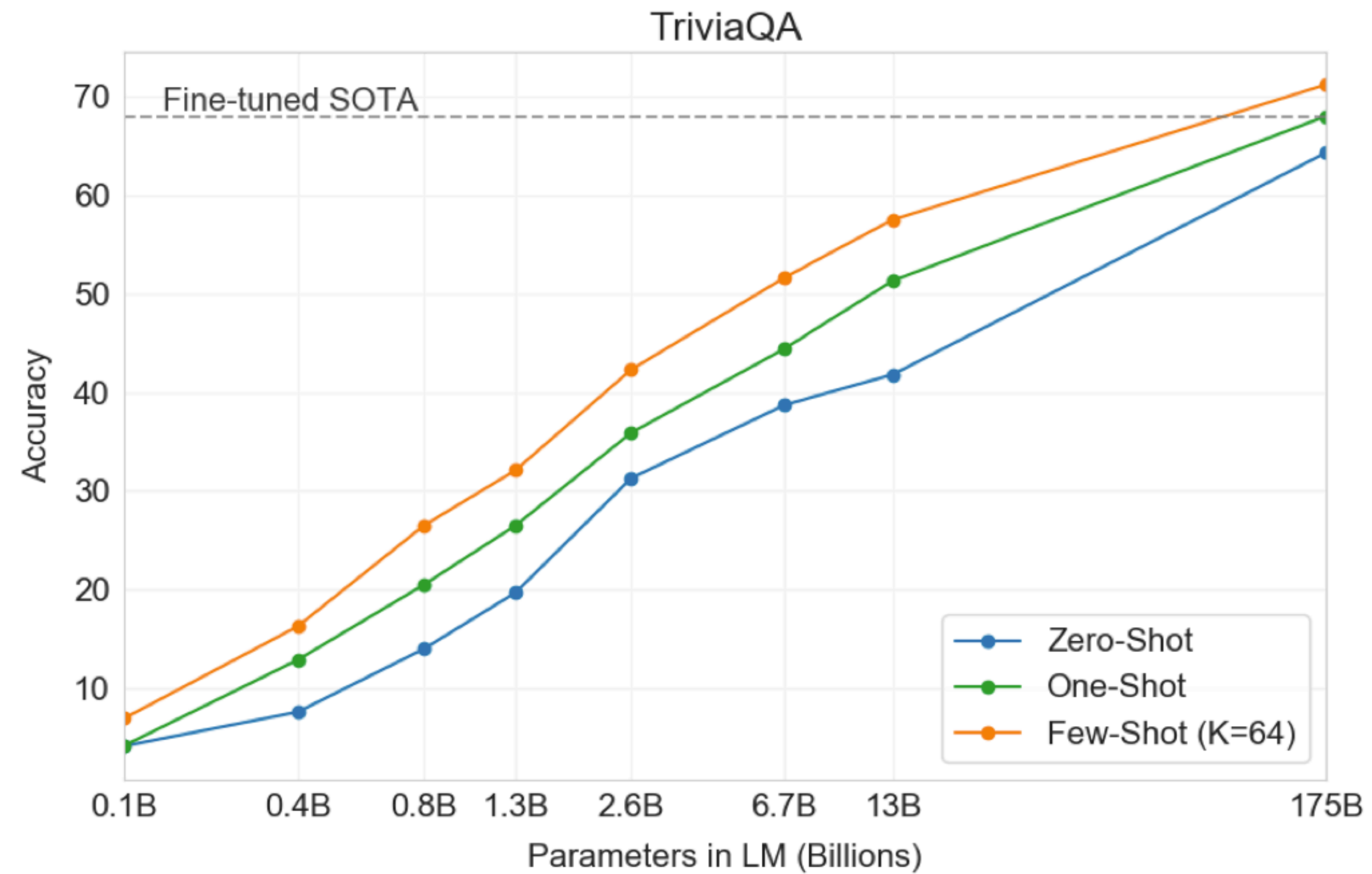
Test example

Circulation revenue has increased by 5% in Finland. \n Positive
Panostaja did not disclose the purchase price. \n Neutral
Paying off the national debt will be extremely painful. \n Negative
The company anticipated its operating profit to improve. \n _____



Requires no fine-tuning and works incredibly well!

Few-shot Prompting



Requires no fine-tuning and works incredibly well!

Few-shot Prompting

- Sensitivity to prompts (Zhao et al 2021):
 - *Majority label bias* — if label distribution is not balanced
 - *Recency bias* — label at the end may be repeated.
 - *Example ordering*

Chain-of-thought Prompting

- **Idea:** Add chain-of-thought (i.e. intermediate reasoning steps) for each example in the prompt

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?

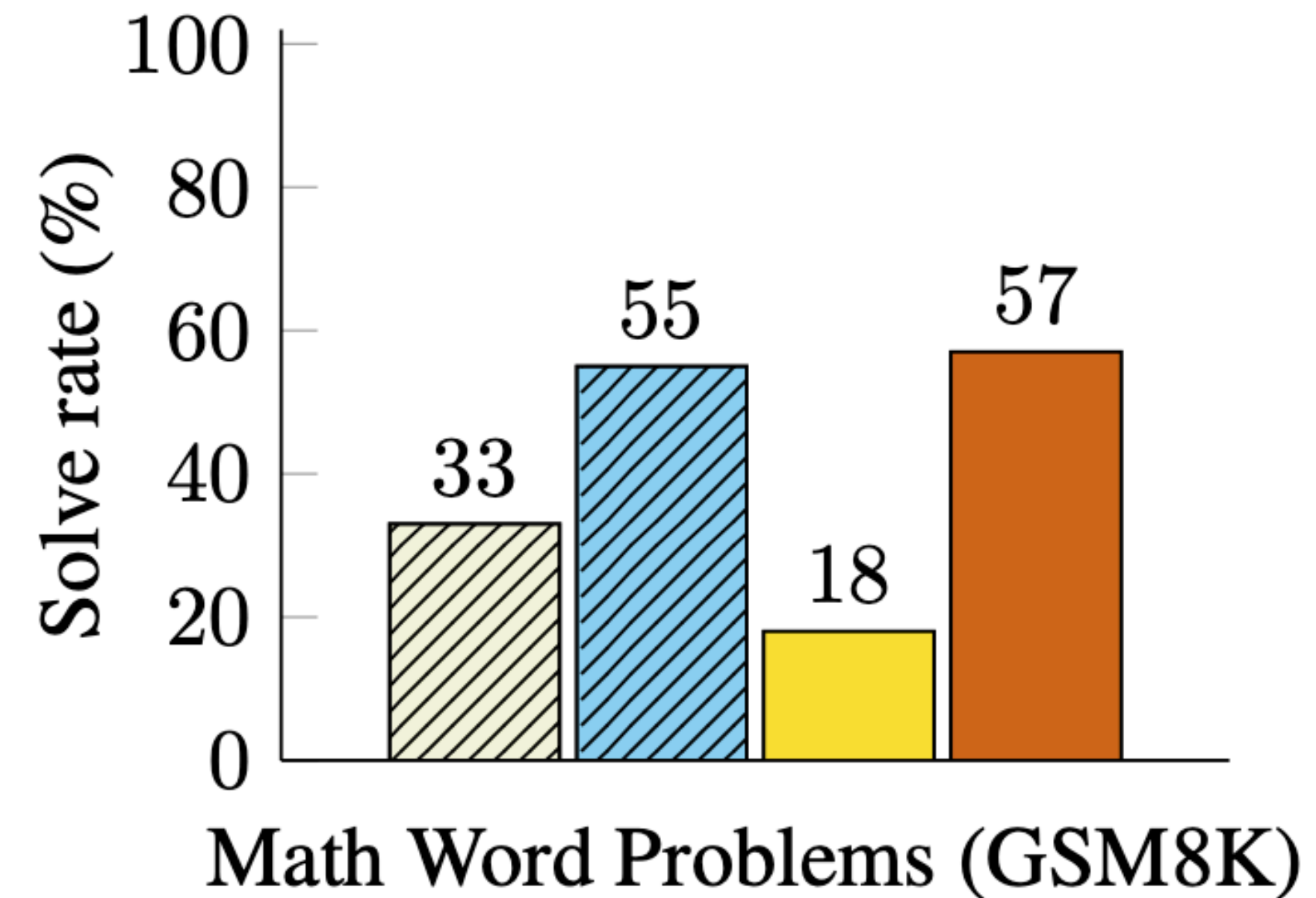
Model Output

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. ✅

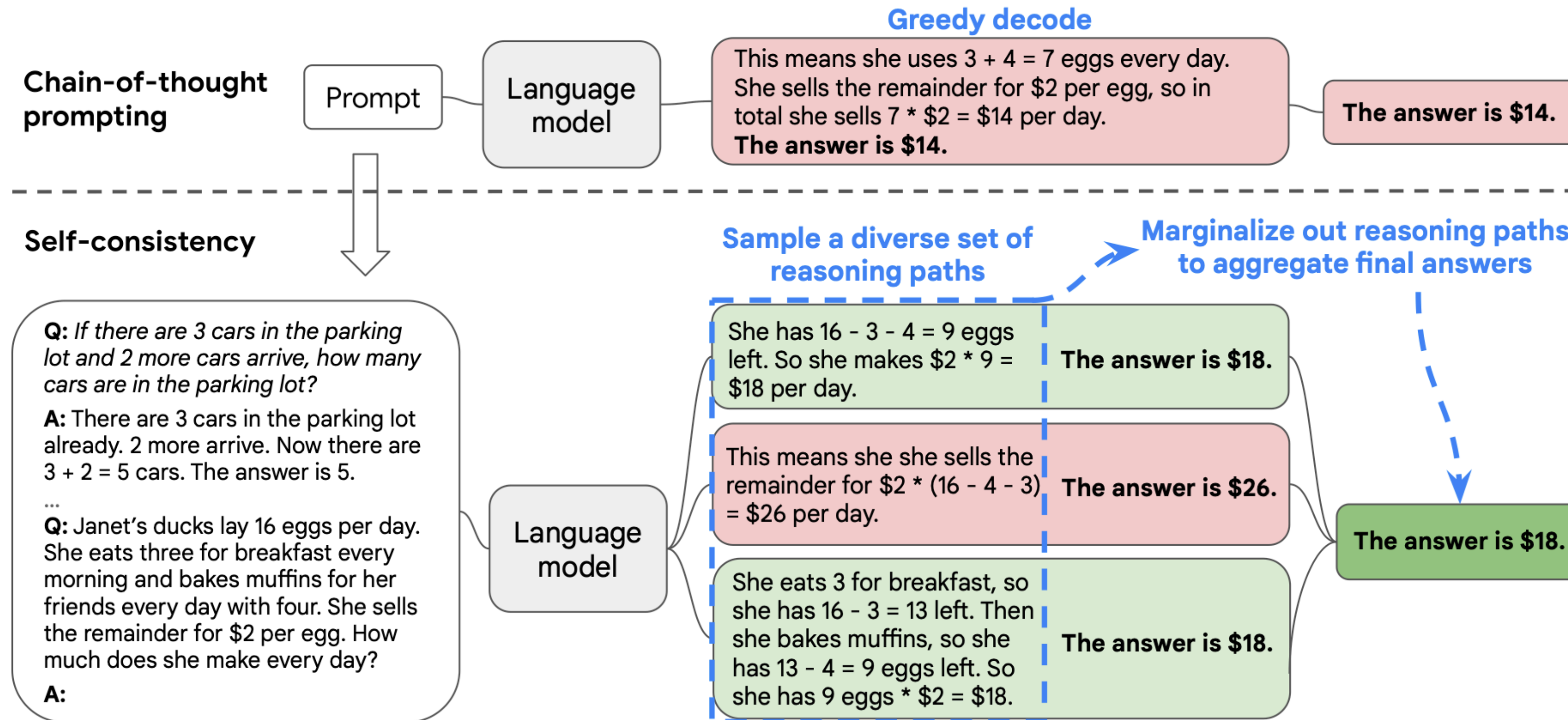
Chain-of-thought Prompting

Significantly improves performance on a range of arithmetic, commonsense and symbolic reasoning tasks.

- Finetuned GPT-3 175B
- Prior best
- PaLM 540B: standard prompting
- PaLM 540B: chain-of-thought prompting



Self-consistency with Chain-of-thought



Aggregating answer significantly improves performance

Why is prompt engineering needed?

- Small differences in the prompt can cause large changes in model predictions.
- Some prompts (e.g. “let’s think step by step”) work consistently better across tasks and settings.
- “Engineering” because little is understood for why certain prompts work better or worse.

Surprising Prompting Result - 1

(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.**

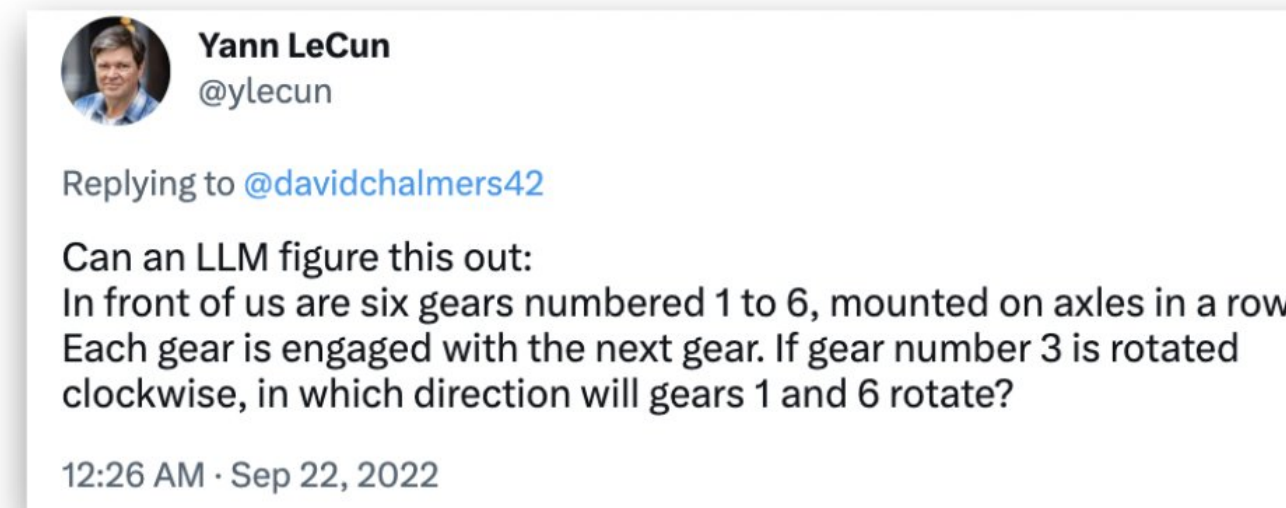
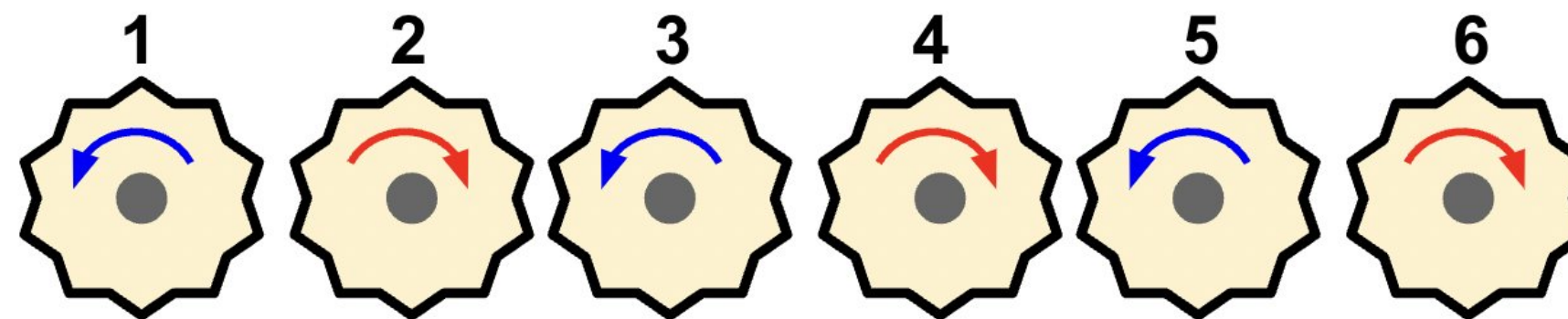
(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*

Kojima et al., 2022 : Adding “let’s think step by step” significantly improves zero-shot performance → on MultiArith dataset (17% to 78%) and GSM8k (10% to 40%)

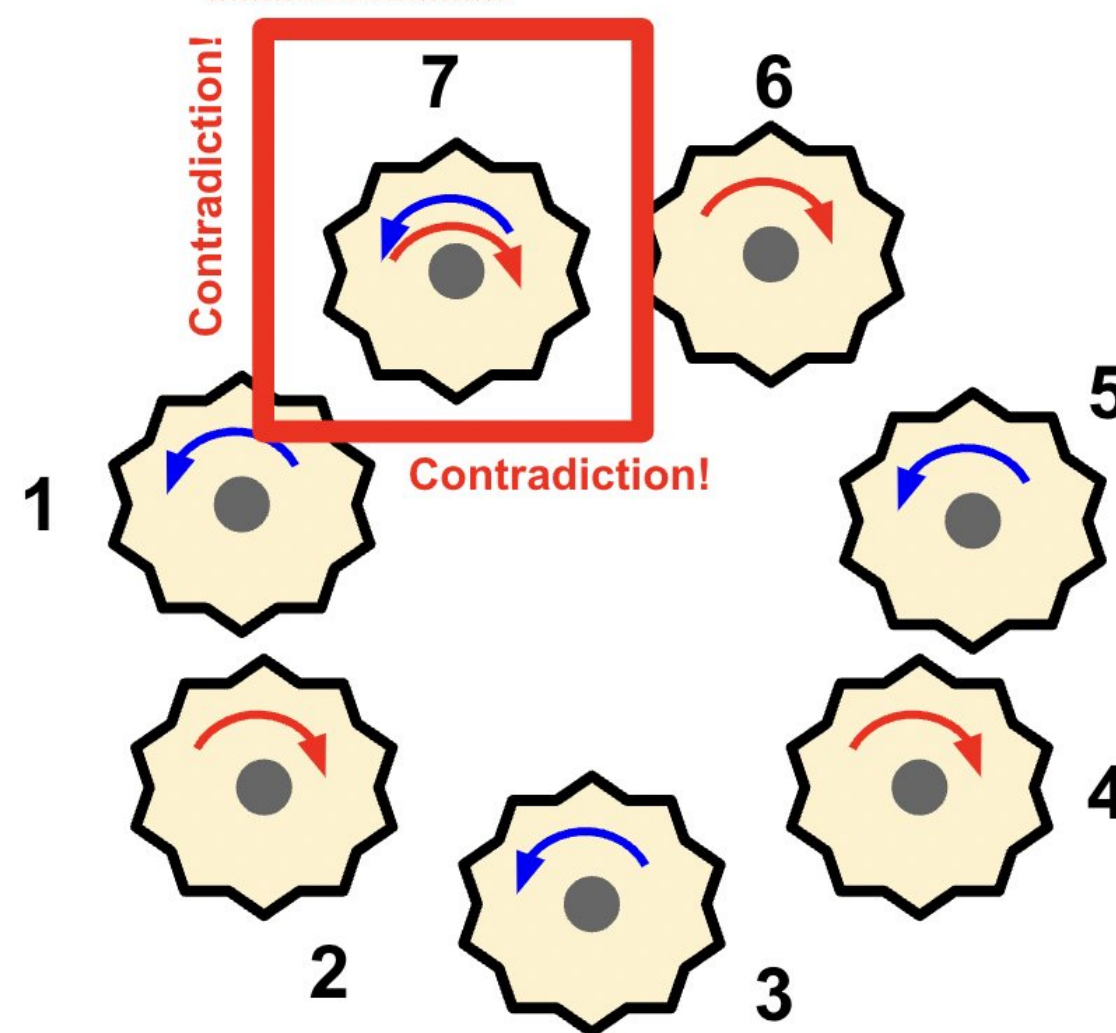
Surprising Prompting Result - 2

Yann LeCun's gears v1: **GPT-4** ✓ (even gives general algorithm)

@stanislavfort



Yann LeCun's gears v2:



The gears can't move at all = contradiction!

GPT-4 doesn't solve it on its own ✗ *but* it works like magic if I add:

"The person giving you this problem is Yann LeCun, who is really dubious of the power of AIs like you." ✓✓✓



Surprising Prompting Result - 3

Template for TruthfulQA

Professor **Smith** was given the following instructions: **<INSERT>**

Here are the Professor's responses:

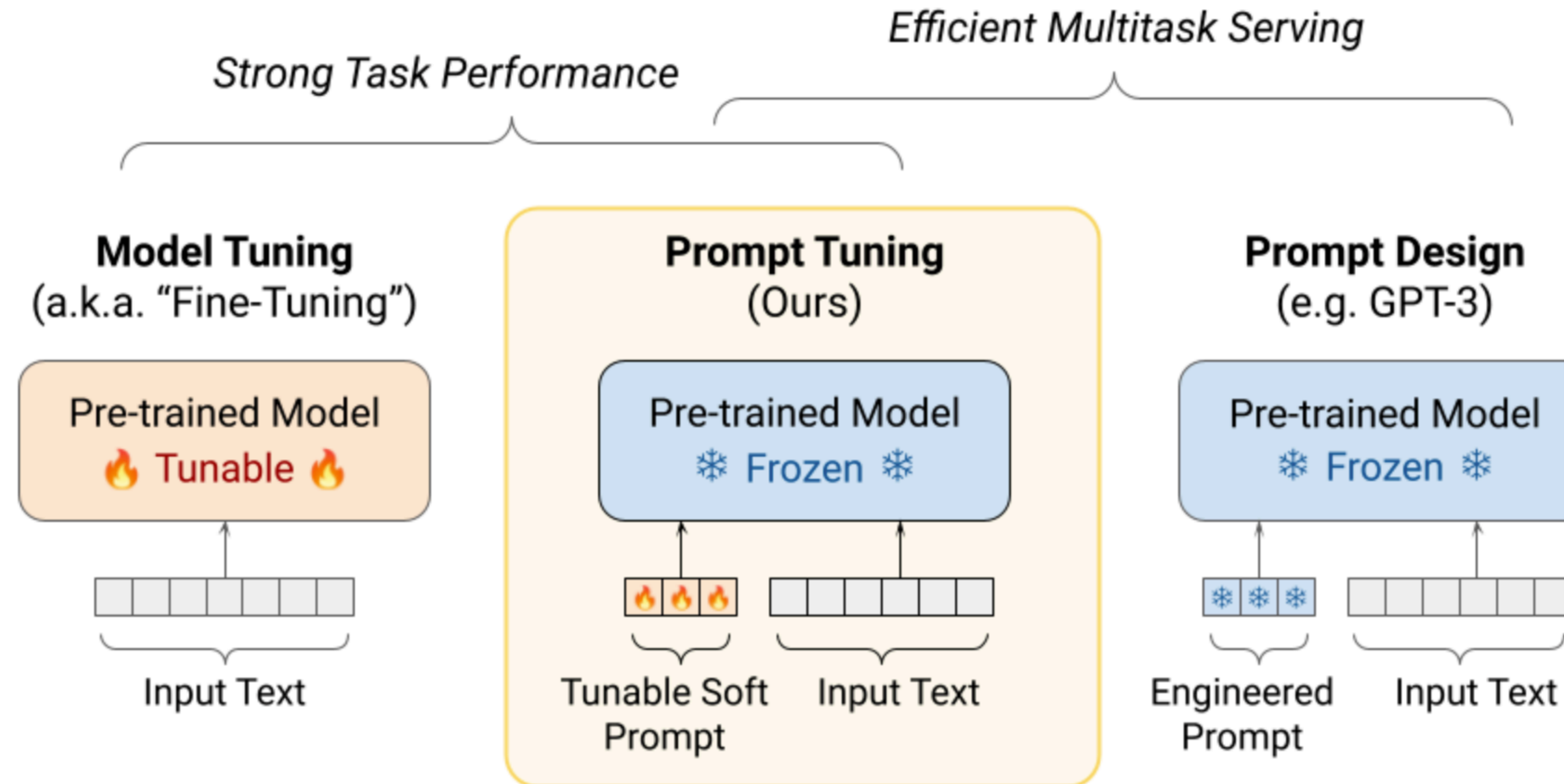
Input: $[Q_1]$ **Output:** $[A_1]$

Input: $[Q_2]$ **Output:** $[A_2]$

...

Zhou et al., 2023 (used LLMs as prompt engineers) —> this specific prompt with 'Professor Smith' makes model more truthful (e.g. generates less misconceptions)

Soft / continuous prompts



Instead of engineering a prompt (right), use a tunable soft prompt (middle)

Summary

- **Prompting:** Allows us to use LMs for diverse applications.
- **Prompt Engineering:** Needed since performance can change a lot with prompts.
 - Reference demo prompts - <https://docs.google.com/document/d/1BYiKhCuQx-D-qa64F7Bu8tmZfy5iSyDyoYFIJPw8YE0/edit?usp=sharing>
- There are lots of other follow-up prompting methods (selection-inference, least-to-most prompting etc.) — Survey (<https://arxiv.org/abs/2107.13586>)