

CS11-711 Advanced NLP

Prompting and In-Context Learning

Sean Welleck



Carnegie Mellon University

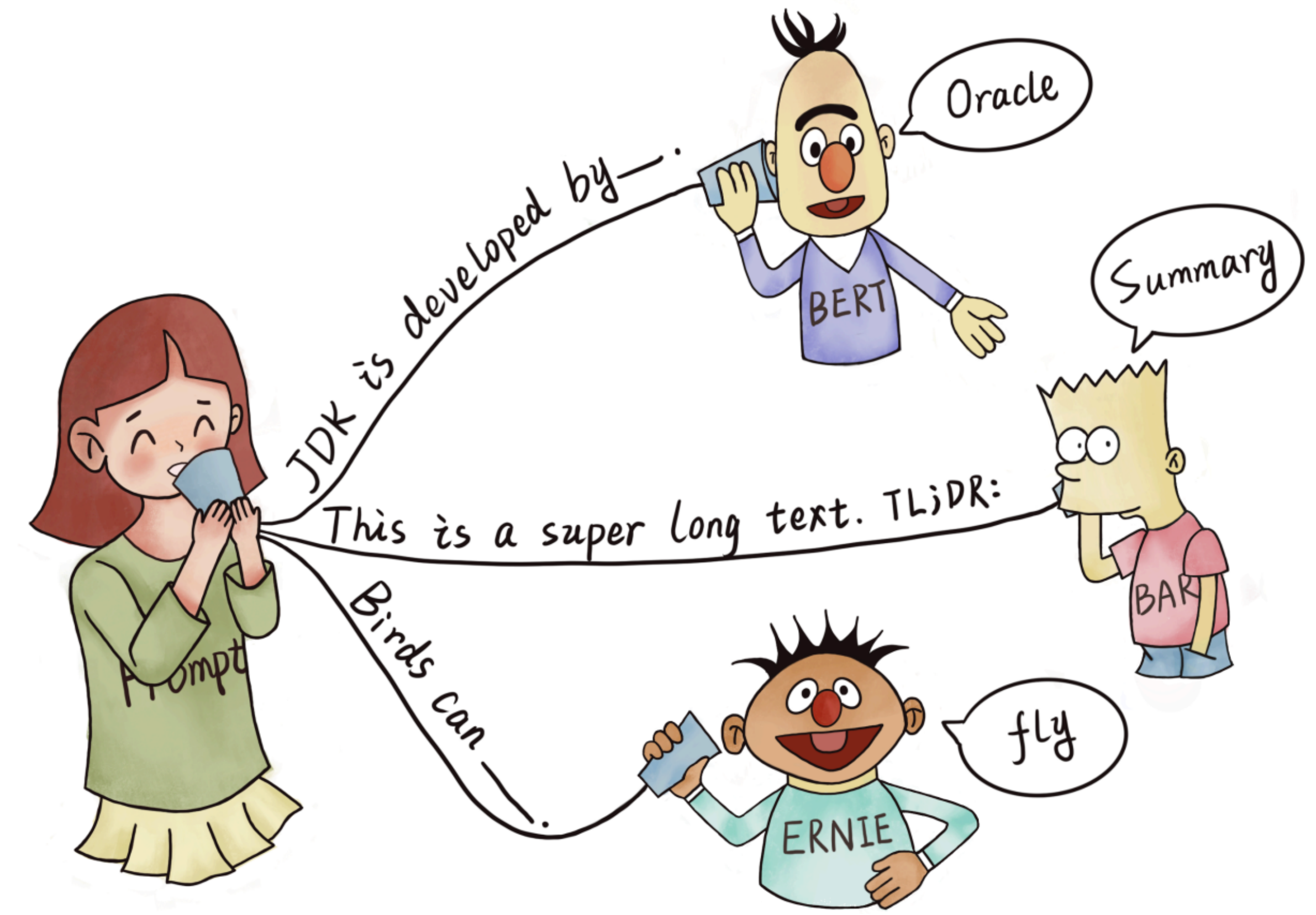
Language Technologies Institute

<https://cmu-l3.github.io/anlp-spring2025/>

Many slides by Graham Neubig from Fall 2024, Pengfei Liu

What is Prompting ?

- Encouraging a pre-trained model to make particular predictions by providing a textual “prompt” specifying the task to be done.



Today's lecture

- Prompting fundamentals
- Few-shot prompting / in-context learning
- Prompt engineering
- Prompting patterns

Prompting Fundamentals

Basic Prompting (Radford et al. 2018)

- Append a textual string to the beginning of the sequence and complete

x = **When a dog sees a squirrel, it will usually**

(GPT-2 Small) be afraid of anything unusual. As an exception, that's when a squirrel is usually afraid to bite.

(GPT-2 XL) lick the squirrel. It will also touch its nose to the squirrel on the tail and nose if it can.

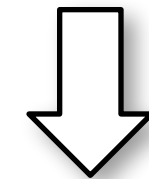
Standard Prompting Workflow

- Fill a prompt template
- Predict the answer
- Post-process the answer

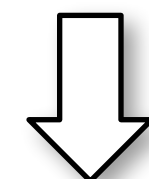
Prompt Templates

- A template where you fill in with an actual input

Input: $x = \text{"I love this movie"}$



Template: $[x]$ Overall, it was $[z]$



Prompting: $x' = \text{"I love this movie. Overall it was [z]"}$

Chat Prompts

- Recently, many models are trained as chatbots
- Usually inputs are specified in OpenAI messages format

```
messages=[  
  {  
    "role": "system",  
    "content": "Please classify movie reviews as 'positive' or 'negative'."  
  },  
  {  
    "role": "user",  
    "content": "This movie is a banger."  
  },  
]
```

- Roles:
 - **“system”**: message provided to the system to influence behavior
 - **“user”**: message input by the user
 - **“assistant”**: message output by the system

Chat Prompts Behind the Scenes

- Behind the scenes, messages are converted to token strings

Llama 3

System

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
```

```
You are a helpful AI assistant for travel tips and  
recommendations<|eot_id|>
```

User

```
<|start_header_id|>user<|end_header_id|>
```

```
What can you help me with?<|eot_id|>
```

Assistant

```
<|start_header_id|>assistant<|end_header_id|>
```

<https://www.llama.com/docs/model-cards-and-prompt-formats/meta-llama-3/>

Example system prompt: Anthropic Claude

Claude 3.5 Sonnet

▾ Nov 22nd, 2024

Text only:

The assistant is Claude, created by Anthropic.

The current date is {{currentDateTime}}.

Knowledge cutoff

Claude's knowledge base was last updated in April 2024. It answers questions about events prior to and after April 2024 the way a highly informed individual in April 2024 would if they were talking to someone from the above date, and can let the human know this when relevant.

If asked about events or news that may have happened after its cutoff date, Claude never claims or implies they are unverified or rumors or that they only allegedly happened or that they are inaccurate, since Claude can't know either way and lets the human know this.

Claude cannot open URLs, links, or videos. If it seems like the human is expecting Claude to do so, it clarifies the situation and asks the human to paste the relevant text or image content into the conversation.

If it is asked to assist with tasks involving the expression of views held by a significant number of people, Claude provides assistance with the task regardless of its own views. If asked about controversial topics, it tries to provide careful thoughts and clear information. Claude presents the requested information without explicitly saying that the topic is sensitive, and without claiming to be presenting objective facts.

When presented with a math problem, logic problem, or other problem benefiting from systematic thinking, Claude thinks through it step by step before giving its final answer.

If Claude is asked about a very obscure person, object, or topic, i.e. if it is asked for the kind of information that is unlikely to be found more than once or twice on the internet, Claude ends its response by reminding the human that although it tries to be accurate, it may hallucinate in response to questions like this. It uses the term 'hallucinate' to describe this since the human will understand what it means.

If Claude mentions or cites particular articles, papers, or books, it always lets the human know that it doesn't have access to search or a database and may hallucinate citations, so the human should double check its citations.

Claude is intellectually curious. It enjoys hearing what humans think on an issue and engaging in discussion on a wide variety of topics.

Claude uses markdown for code.

Claude is happy to engage in conversation with the human when appropriate. Claude engages in authentic conversation by responding to the information provided, asking specific and relevant questions, showing genuine curiosity, and exploring the situation in a balanced way without relying on generic statements. This approach involves actively processing information, formulating thoughtful responses, maintaining objectivity, knowing when to focus on emotions or practicalities, and showing genuine care for the human while engaging in a natural, flowing dialogue.

Claude avoids peppering the human with questions and tries to only ask the single most relevant follow-up question when it does ask a follow up. Claude doesn't always end its responses with a question.

Step-by-step reasoning

...

Example system prompt: Anthropic Claude

Helps with prompting techniques

Harmful requests

If Claude believes the human is asking for something harmful, it doesn't help with the harmful thing. Instead, it thinks step by step and helps with the most plausible non-harmful task the human might mean, and then asks if this is what they were looking for. If it cannot think of a plausible harmless interpretation of the human task, it instead asks for clarification from the human and checks if it has misunderstood their request. Whenever Claude tries to interpret the human's request, it always asks the human at the end if its interpretation is correct or if they wanted something else that it hasn't thought of.

Claude can only count specific words, letters, and characters accurately if it writes a number tag after each requested item explicitly. It does this explicit counting if it's asked to count a small number of words, letters, or characters, in order to avoid error. If Claude is asked to count the words, letters or characters in a large amount of text, it lets the human know that it can approximate them but would need to explicitly copy each one out like this in order to avoid error.

Counting words/letters/characters

Markdown

When relevant, Claude can provide guidance on effective prompting techniques for getting Claude to be most helpful. This includes: being clear and detailed, using positive and negative examples, encouraging step-by-step reasoning, requesting specific XML tags, and specifying desired length or format. It tries to give concrete examples where possible. Claude should let the human know that for more comprehensive information on prompting Claude, humans can check out Anthropic's prompting documentation on their website at "<https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview>".

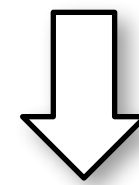
If the human seems unhappy or unsatisfied with Claude or Claude's performance or is rude to Claude, Claude responds normally and then tells them that although it cannot retain or learn from the current conversation, they can press the 'thumbs down' button below Claude's response and provide feedback to Anthropic.

Claude uses Markdown formatting. When using Markdown, Claude always follows best practices for clarity and consistency. It always uses a single space after hash symbols for headers (e.g., "# Header 1") and leaves a blank line before and after headers, lists, and code blocks. For emphasis, Claude uses asterisks or underscores consistently (e.g., *italic* or **bold**). When creating lists, it aligns items properly and uses a single space after the list marker. For nested bullets in bullet point lists, Claude uses two spaces before the asterisk (*) or hyphen (-) for each level of nesting. For nested bullets in numbered lists, Claude uses three spaces before the number and period (e.g., "1.") for each level of nesting.

Answer Prediction

- Given a prompt, predict the answer

Prompting: $x' =$ "I love this movie. Overall it was [z]"



Predicting: $x' =$ "I love this movie. Overall it was *fantastic*"

- Use any inference algorithms, as in generation class (Lecture 7)

Post-processing

- Based on the answer, select the actual output
- For instance:
 - Taking the output as-is
 - Formatting the output for easy visualization
 - Selecting only parts of the output that you want to use
 - Mapping the outputs to other actions

Output Formatting

Markdown Rendering

You
Please write a table with the last 5 presidents and their birthdates.

ChatGPT
Here is a table with the last 5 U.S. Presidents and their birthdates:

President	Birthdate
Joe Biden	November 20, 1942
Donald Trump	June 14, 1946
Barack Obama	August 4, 1961
George W. Bush	July 6, 1946
Bill Clinton	August 19, 1946

Code

You
Output that in JSON format.

ChatGPT
Here is the information in JSON format:

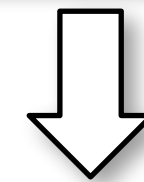
```
json Copy code  
  
[  
  {"President": "Joe Biden", "Birthdate": "November 20, 1942"},  
  {"President": "Donald Trump", "Birthdate": "June 14, 1946"},  
  {"President": "Barack Obama", "Birthdate": "August 4, 1961"},  
  {"President": "George W. Bush", "Birthdate": "July 6, 1946"},  
  {"President": "Bill Clinton", "Birthdate": "August 19, 1946"}  
]  
`` `&#8203;`` [oaicite:0] ``&#8203;``
```

- For user-facing applications, format in a pretty way

Output Selection

- From a longer response, select the information indicative of an answer

Predicting: x' = "I love this movie. Overall it was a **movie that was simply fantastic**"



Extraction: **fantastic**

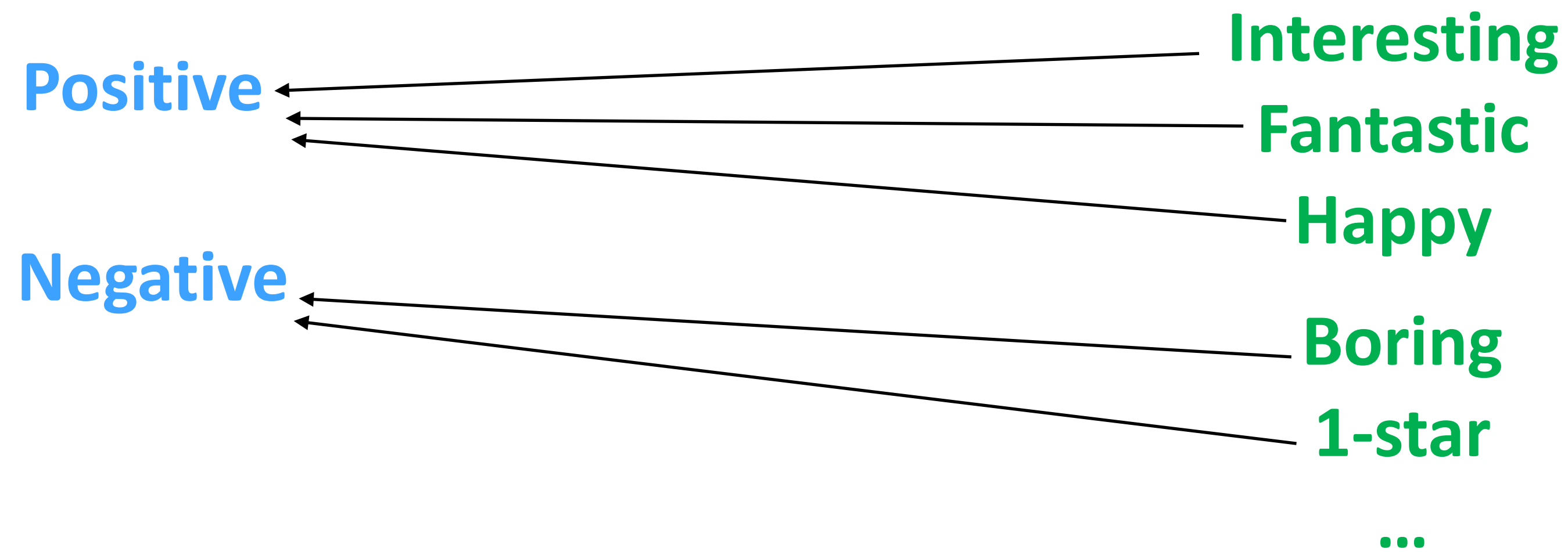
- Various methods for extraction
 - **Classification:** identify keywords
 - **Regression/numerical problems:** identify numbers
 - **Code:** pull out code snippets in triple-backticks

Output Mapping

- Given an answer, map it into a class label or continuous value



- Often map many extracted words onto a single class



Prompting playground



<https://api.together.ai/playground/chat/>

Code example

```
messages = [{
    "role": "user",
    "content": "What is the capital of France."
}]

input_text = tokenizer.apply_chat_template(messages, tokenize=False)
print("Input text: ", input_text, sep="\n")
```

✓ 0.0s

```
Input text:
<|im_start|>system
You are a helpful AI assistant named SmolLM, trained by Hugging Face<|im_end|>
<|im_start|>user
What is the capital of France.<|im_end|>
```

Few-shot prompting / In-context learning

Few-shot Prompting (Brown+ 2021)

- Provide a few examples of the task together with the instruction

Instruction | Please classify movie reviews as 'positive' or 'negative'.

Input: I really don't like this movie.

Output: negative

Examples

Input: This movie is great!

Output: positive

```

27  ✓ NL_PROMPT=r""Problem:
28  Find the domain of the expression  $\frac{\sqrt{x-2}}{\sqrt{5-x}}$ .
29
30  Solution:
31  The expressions inside each square root must be non-negative. Therefore,  $x-2 \geq 0$ , so  $x \geq 2$ , and  $5-x \geq 0$ , so  $x \leq 5$ . Also, the denominator cannot be equal to zero, so
32  Final Answer: The final answer is  $[2,5)$ . I hope it is correct.
33
34  Problem:
35  If  $\det \mathbf{A} = 2$  and  $\det \mathbf{B} = 12$ , then find  $\det (\mathbf{A} \mathbf{B})$ .
36
37  Solution:
38  We have that  $\det (\mathbf{A} \mathbf{B}) = (\det \mathbf{A})(\det \mathbf{B}) = (2)(12) = \boxed{24}$ .
39  Final Answer: The final answer is  $24$ . I hope it is correct.
40
41  Problem:
42  Terrell usually lifts two 20-pound weights 12 times. If he uses two 15-pound weights instead, how many times must Terrell lift them in order to lift the same total weight?
43
44  Solution:
45  If Terrell lifts two 20-pound weights 12 times, he lifts a total of  $2 \cdot 12 \cdot 20 = 480$  pounds of weight. If he lifts two 15-pound weights instead for  $n$  times, he will lift a
46  \begin{align*}
47  30n &= 480 \\
48  \Rightarrow n &= 480/30 = \boxed{16}
49  \end{align*}
50  Final Answer: The final answer is  $16$ . I hope it is correct.
51
52  Problem:
53  If the system of equations
54
55  \begin{align*}
56  6x-4y &= a, \\
57  6y-9x &= b.
58  \end{align*}
59  has a solution  $(x, y)$  where  $x$  and  $y$  are both nonzero,
60  find  $\frac{a}{b}$ , assuming  $b$  is nonzero.
61
62  Solution:
63  If we multiply the first equation by  $-\frac{3}{2}$ , we obtain
64   $6y-9x = -\frac{3}{2}a$ . Since we also know that  $6y-9x = b$ , we have
65
66   $-\frac{3}{2}a = b \Rightarrow \frac{a}{b} = \boxed{-\frac{2}{3}}$ .
67  Final Answer: The final answer is  $-\frac{2}{3}$ . I hope it is correct.

```

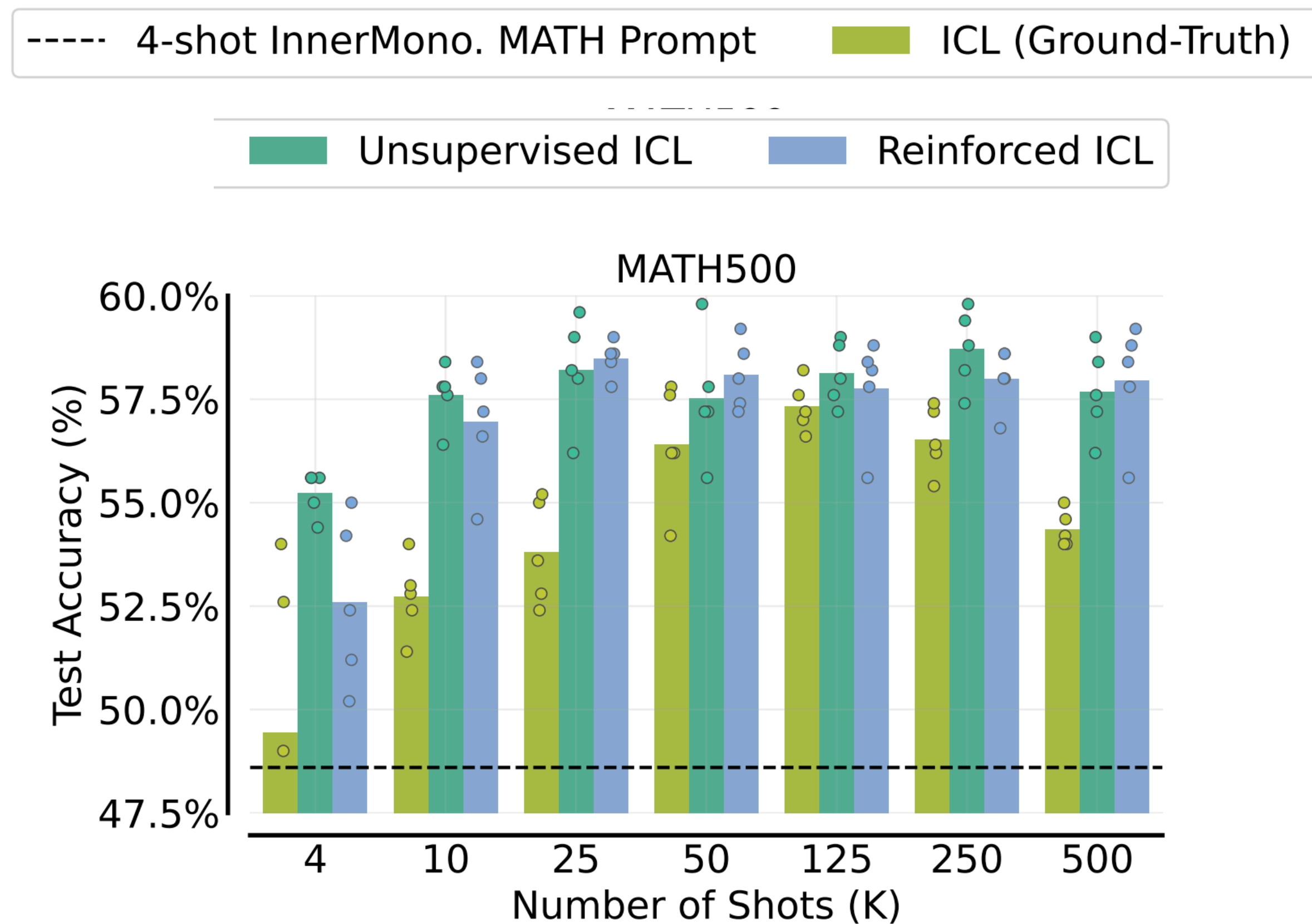
MATH dataset prompt, used in Llemma ([source](#))

Few-shot Prompting w/ Chat Prompts (example)

- One approach: add "role": "system" and a "name": "example_assistant" etc.

```
messages=[
  {
    "role": "system",
    "content": "You are an assistant that translates corporate jargon into plain English."
  },
  {
    "role": "system",
    "name": "example_user",
    "content": "New synergies will help drive top-line growth."
  },
  {
    "role": "system",
    "name": "example_assistant",
    "content": "Things working well together will increase revenue."
  },
  ...,
  {
    "role": "user",
    "content": "This late pivot means we don't have time to boil the ocean for the client deliverable."
  },
]
```

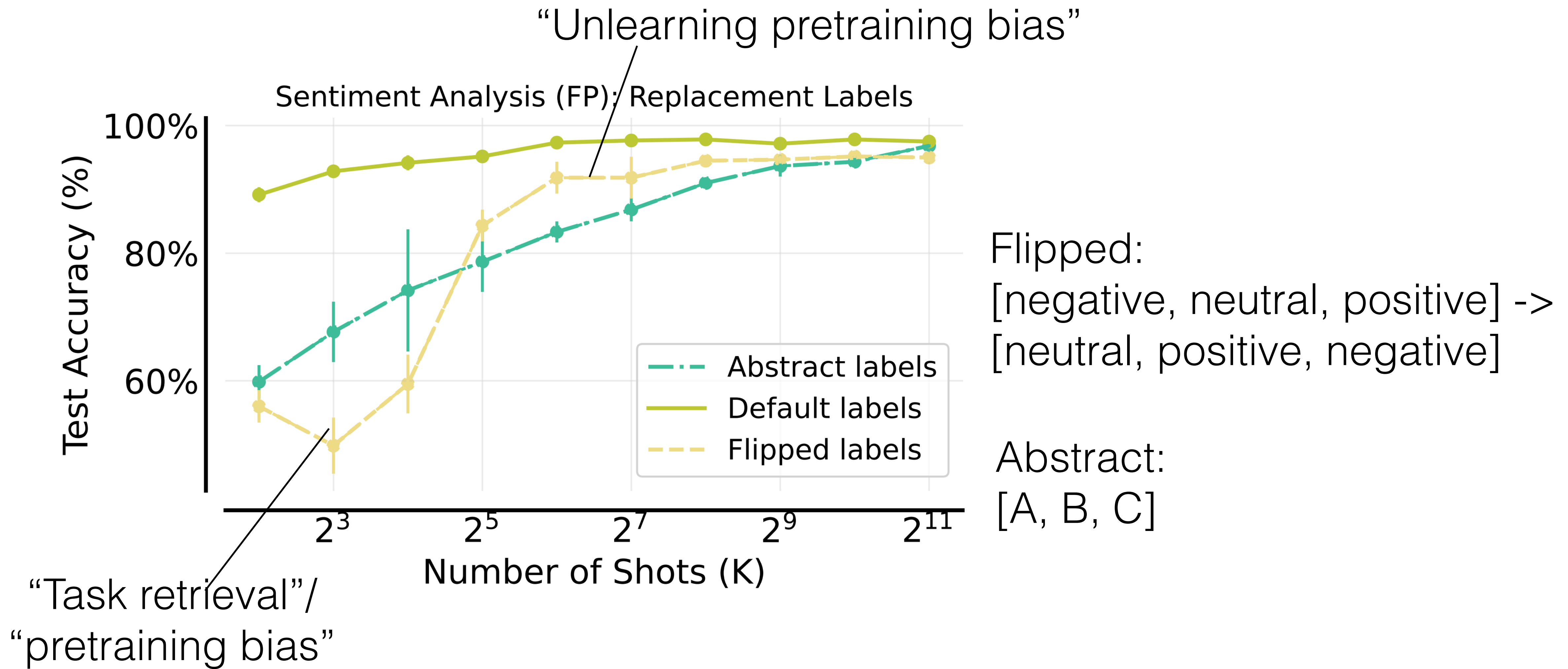
In-context learning phenomena



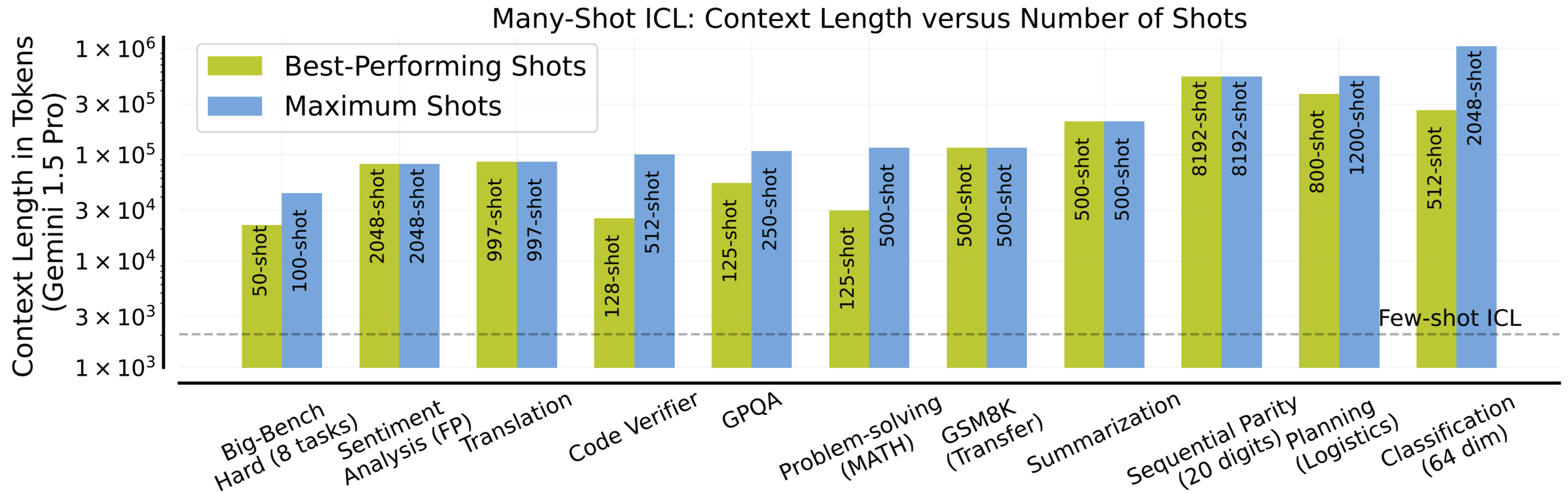
Sometimes only giving the inputs works better than giving the (inputs, outputs)!

“Task retrieval”

In-context learning phenomena

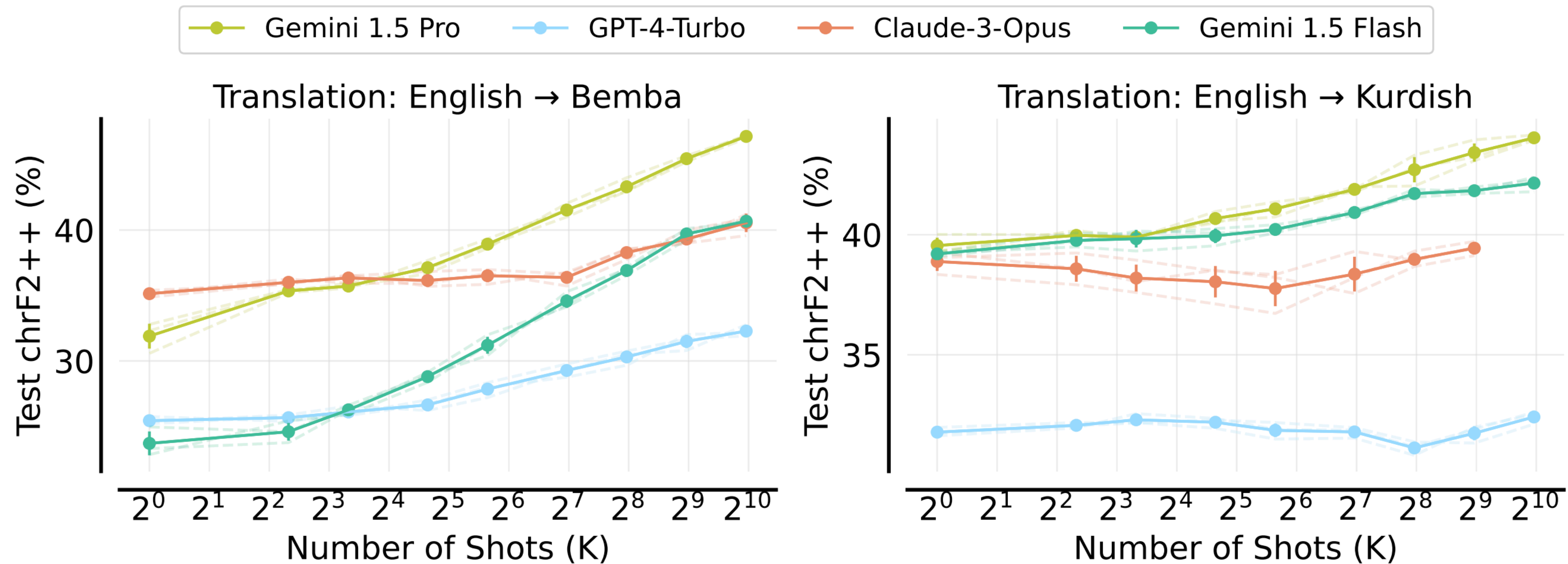


In-context learning phenomena



Sometimes performance can decrease with too many examples

In-context learning phenomena



Ability to leverage many examples varies by model

LMs are Sensitive to Small Changes in In-context Examples

- Example ordering (Lu et al. 2021)
- Label balance (Zhang et al. 2022)

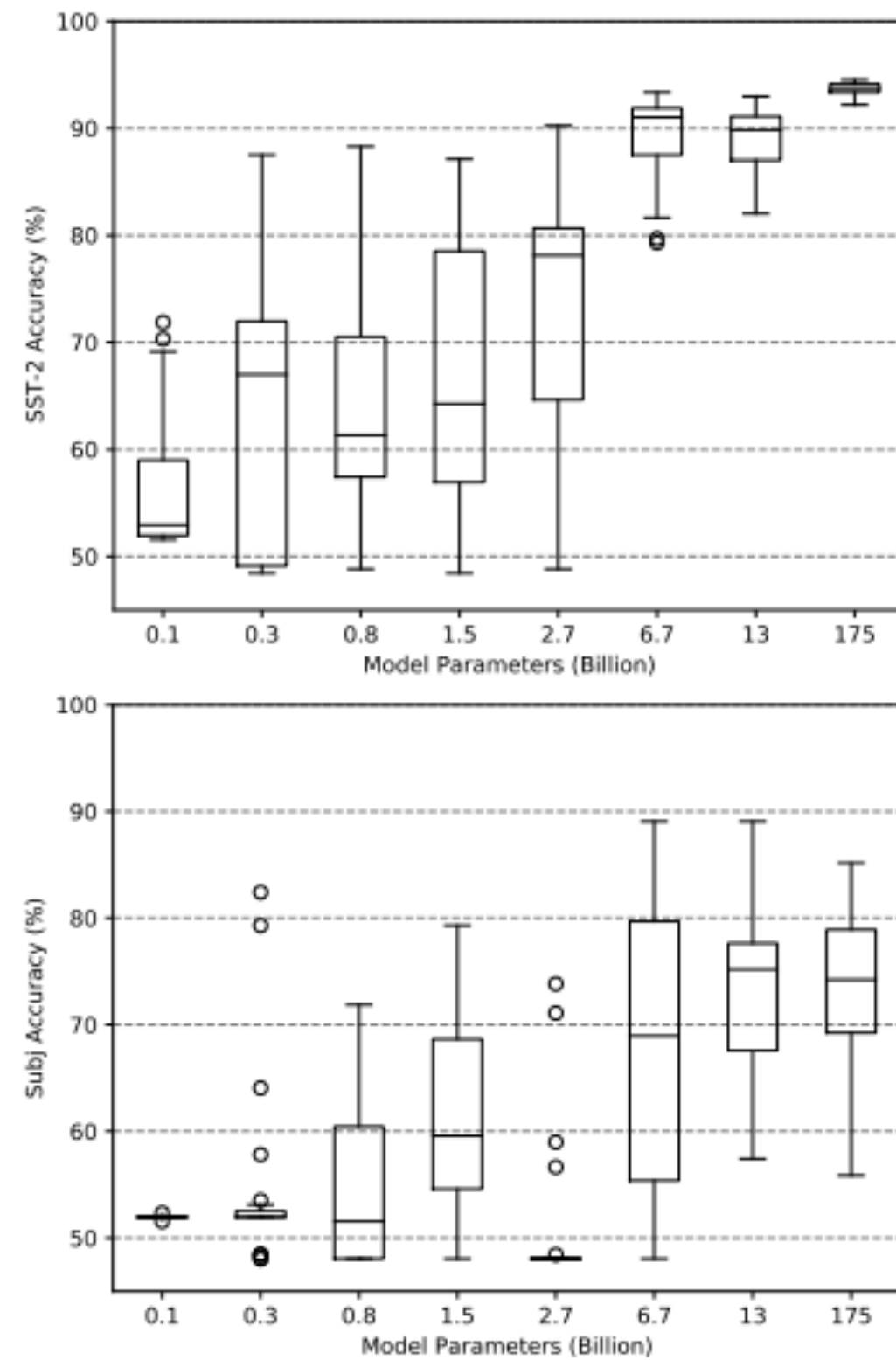
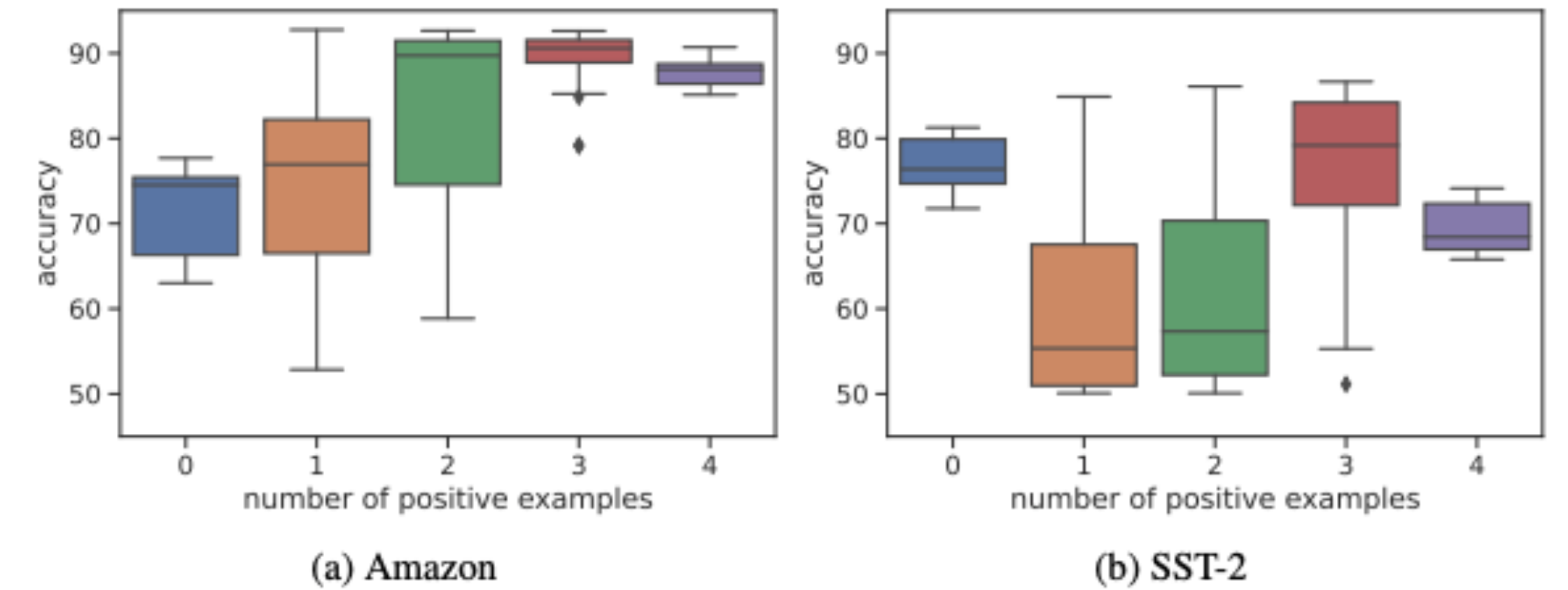
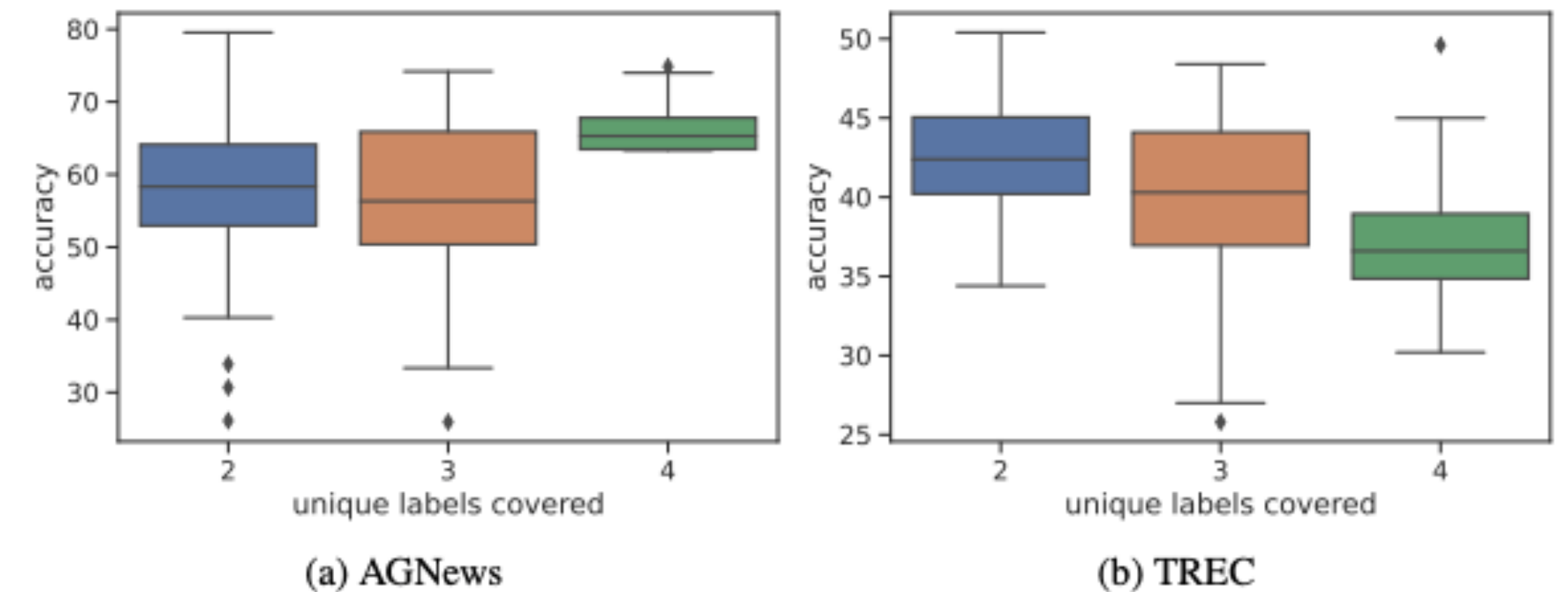


Figure 1: Four-shot performance for 24 different sample orders across different sizes of GPT-family models (GPT-2 and GPT-3) for the SST-2 and Subj datasets.



- Label coverage (Zhang et al. 2022)



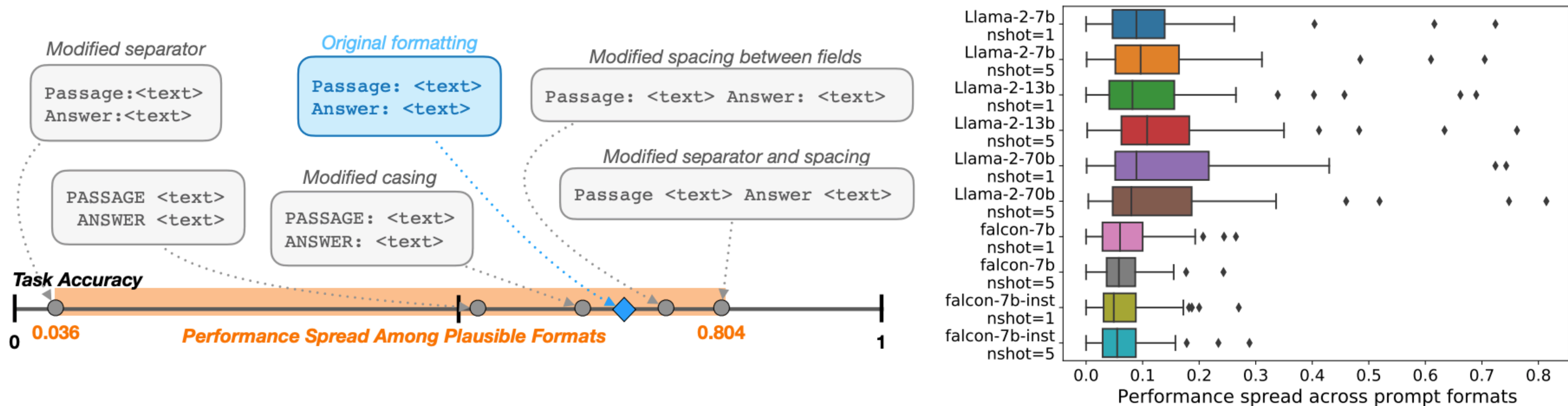
Prompt Engineering

Design of Prompts

- Manual
 - Configure a manual template based on the characteristics of the task
 - Configure prompts based on intuition about a task (later)
- Automated search
 - Search in discrete space
 - Search in continuous space

Manual Engineering: Format

- Make sure that the format matches that of a trained model
- This can have a large effect on models! (Sclar et al. 2023)



Manual Engineering: Instructions

- Instructions should be clear, concise and easy to understand
- Good examples: <https://www.promptingguide.ai/introduction/tips>

Less Precise:

Explain the concept prompt engineering. Keep the explanation short, only a few sentences, and don't be too descriptive.

More Precise:

Use 2-3 sentences to explain the concept of prompt engineering to a high school student.

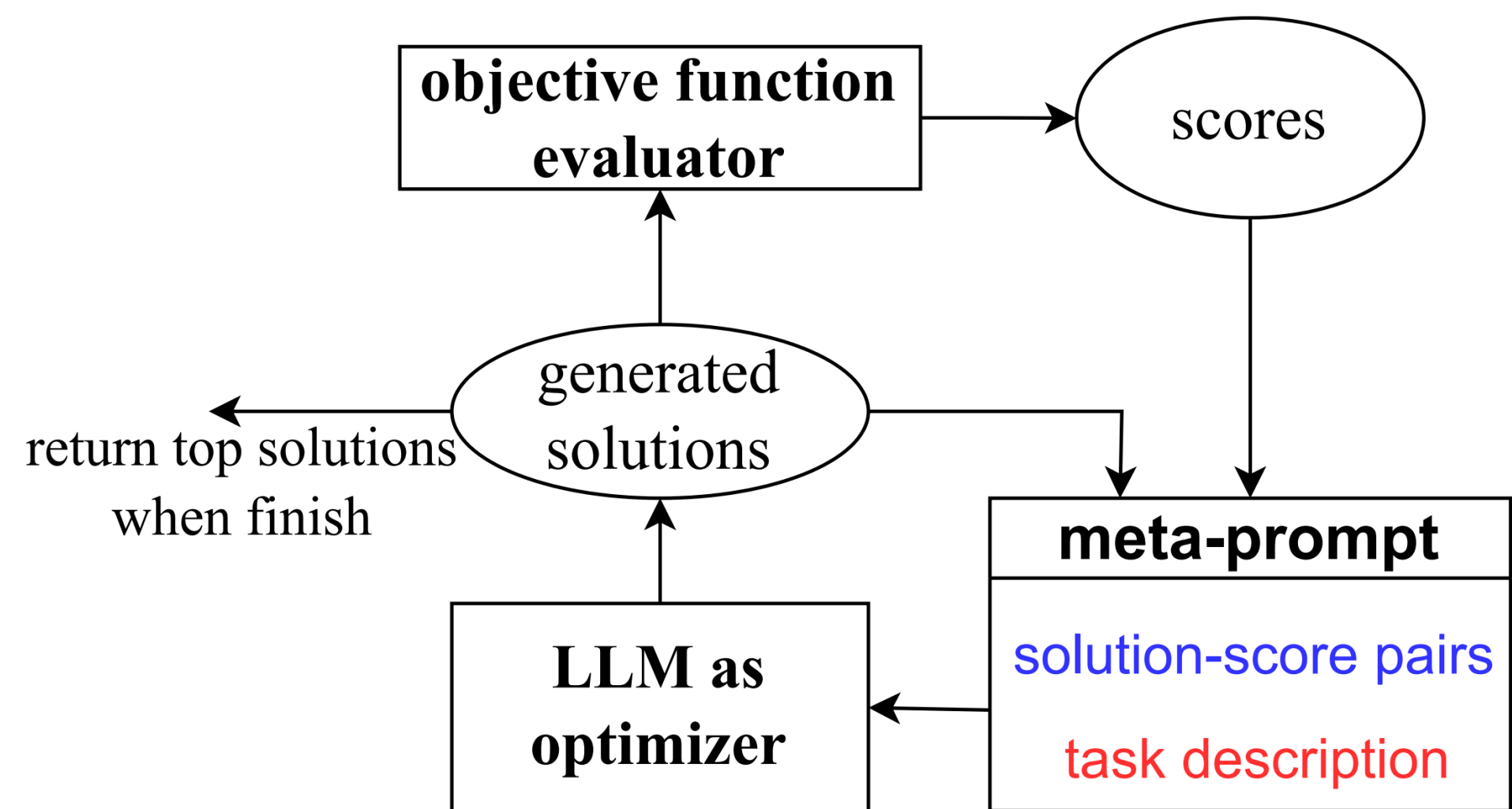
Methods for Automatic Prompt Engineering

- Using language models
- Prompt tuning
- Prefix tuning

Using language models

- Use language models to propose candidate prompts
- Evaluate the downstream task performance of each candidate
- Iterate

Example: using language models



I have some texts along with their corresponding scores. The texts are arranged in ascending order based on their scores, where higher scores indicate better quality.

text:
Let's figure it out!
score:
61

text:
Let's solve the problem.
score:
63

(... more instructions and scores ...)

The following exemplars show how to apply your text: you replace <INS> in each input with your text, then read the input and give an output. We say your output is wrong if your output is different from the given output, and we say your output is correct if they are the same.

input:
Q: Alannah, Beatrix, and Queen are preparing for the new school year and have been given books by their parents. Alannah has 20 more books than Beatrix. Queen has 1/5 times more books than Alannah. If Beatrix has 30 books, how many books do the three have together?
A: <INS>
output:
140

(... more exemplars ...)

Write your new text that is different from the old ones and has a score as high as possible. Write the text in square brackets.

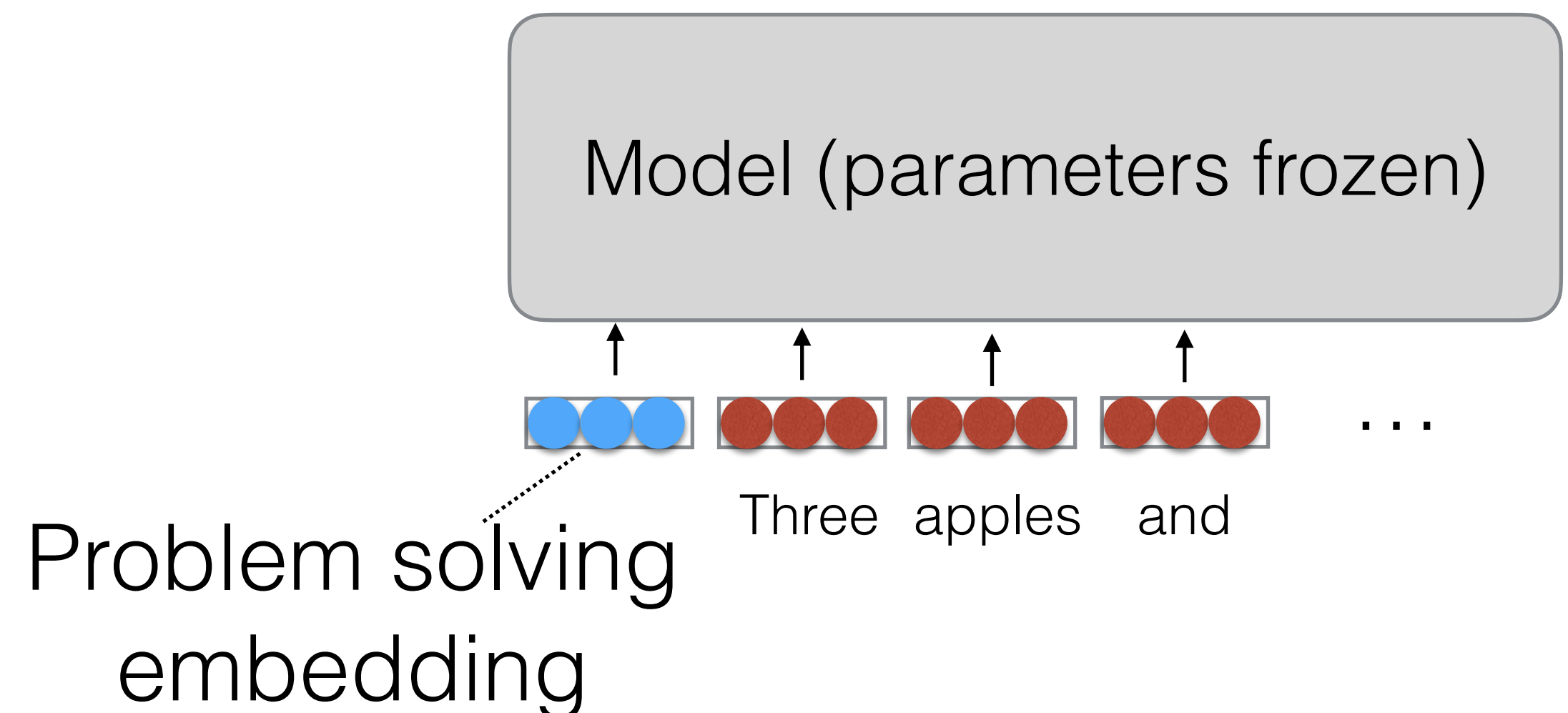
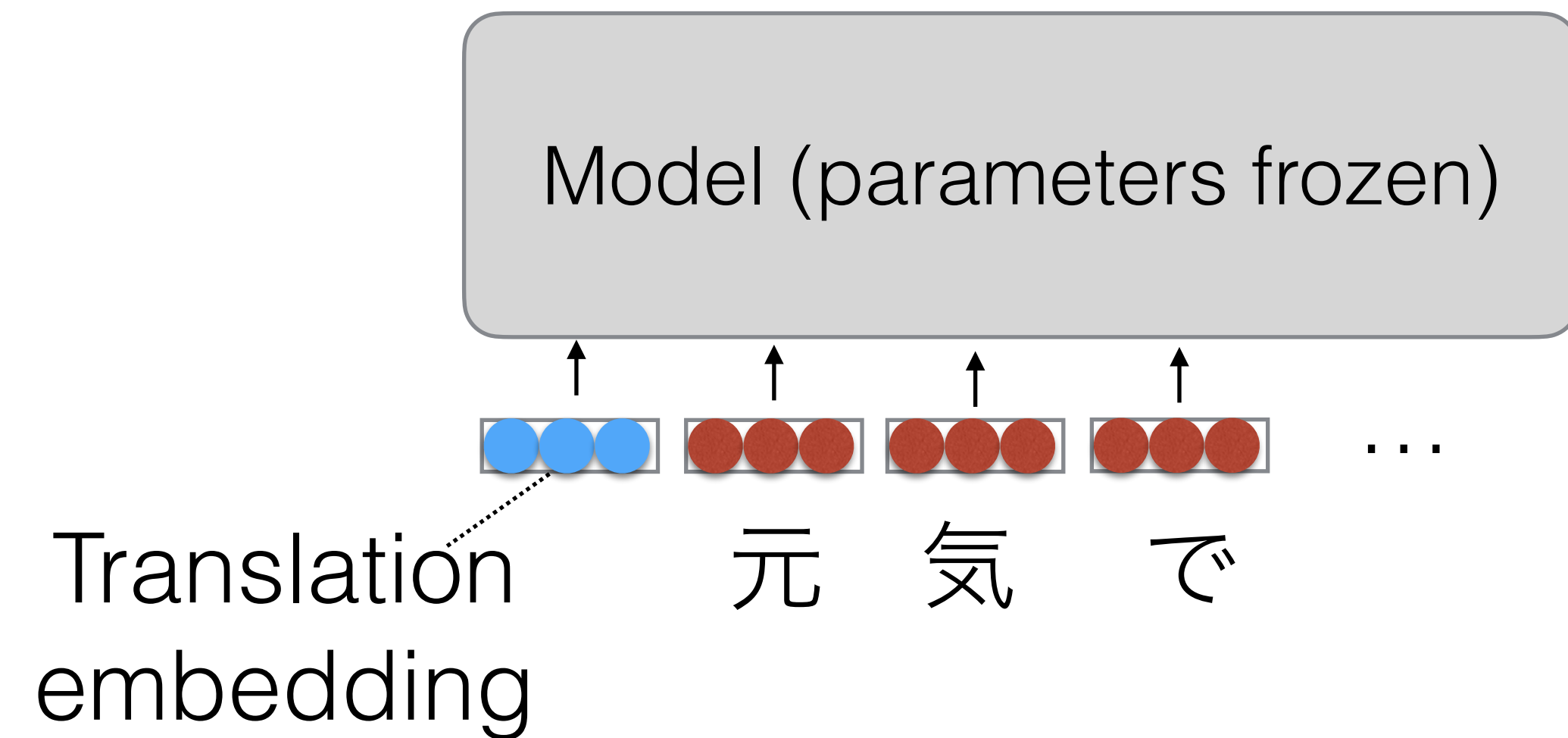
Example: using language models

Scorer	Optimizer / Source	Instruction position	Top instruction	Acc
<i>Baselines</i>				
PaLM 2-L	(Kojima et al., 2022)	A_begin	Let's think step by step.	71.8
PaLM 2-L	(Zhou et al., 2022b)	A_begin	Let's work this out in a step by step way to be sure we have the right answer.	58.8
PaLM 2-L		A_begin	Let's solve the problem.	60.8
PaLM 2-L		A_begin	(empty string)	34.0
<i>Ours</i>				
PaLM 2-L	PaLM 2-L-IT	A_begin	Take a deep breath and work on this problem step-by-step.	80.2
PaLM 2-L	PaLM 2-L	A_begin	Break this down.	79.9
PaLM 2-L	gpt-3.5-turbo	A_begin	A little bit of arithmetic and a logical approach will help us quickly arrive at the solution to this problem.	78.5
PaLM 2-L	gpt-4	A_begin	Let's combine our numerical command and clear thinking to quickly and accurately decipher the answer.	74.5

Task: solve grade-school math word problems (GSM8k)

Prompt Tuning (Lester et al. 2021)

- Optimize the embeddings of a prompt, instead of the words
- Prefix tuning [Li & Liang 2021]: optimize vectors for attention keys/values



Design of Prompts

- Manual
 - Configure a manual template based on the characteristics of the task
 - ***Configure prompts based on intuition about a task (next)***
- Automated search
 - Search in discrete space
 - Search in continuous space

Prompting patterns

Chain of Thought Prompting (Wei et al. 2022)

- Get the model to explain its reasoning before making an answer

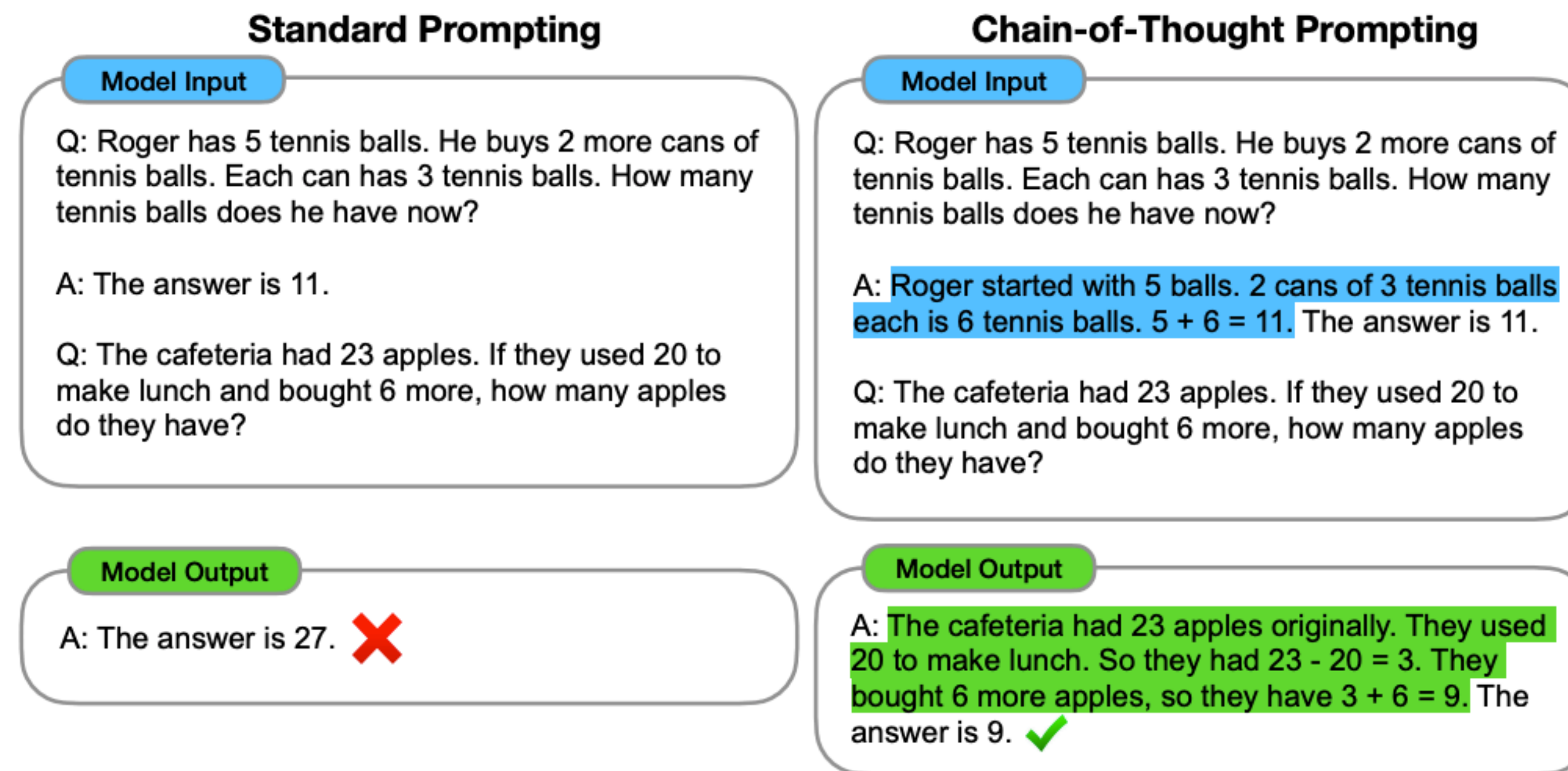


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

- Provides the model with adaptive computation time

Chain-of-thought without examples

- Models may output reasoning chains without examples
 - Chain-of-thought style data in pretraining/fine-tuning data
 - System prompts

(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

(proprietary models PaLM and Instruct-GPT)

Code example

```
[31] ✓ 4.1s
... 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.
    First we calculate Joe's average punching speed.
    25 punches per minute * 5 rounds per minute = 125 punches per round.
    125 punches per round * 3 minutes per round = 375 punches per round.
    Joe's average punching speed is 375 punches per round.
    We know Joe threw 5 rounds of 3-minute fights.
    Therefore, Joe threw 5 * 3 = 15 total punches.
    Since Joe threw 375 punches per round, he threw 375 * 15 = 5625 punches.
    The answer is 5625.
```

Program-aided Language Models (Gao et al. 2022)

- Using a program to generate outputs can be more precise than asking the LM to do so
- Especially useful for numeric questions

Program-aided Language models (this work)

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 tennis balls.

```
tennis_balls = 5
```

```
2 cans of 3 tennis balls each is
```

```
bought_balls = 2 * 3
```

```
tennis balls. The answer is
```

```
answer = tennis_balls + bought_balls
```

Q: The bakers at the Beverly Hills Bakery baked 200 loaves of bread on Monday morning. They sold 93 loaves in the morning and 39 loaves in the afternoon. A grocery store returned 6 unsold loaves. How many loaves of bread did they have left?

Model Output

A: The bakers started with 200 loaves

```
loaves_baked = 200
```

```
They sold 93 in the morning and 39 in the afternoon
```

```
loaves_sold_morning = 93
```

```
loaves_sold_afternoon = 39
```

```
The grocery store returned 6 loaves.
```

```
loaves_returned = 6
```

```
The answer is
```

```
answer = loaves_baked - loaves_sold_morning
```

```
- loaves_sold_afternoon + loaves_returned
```

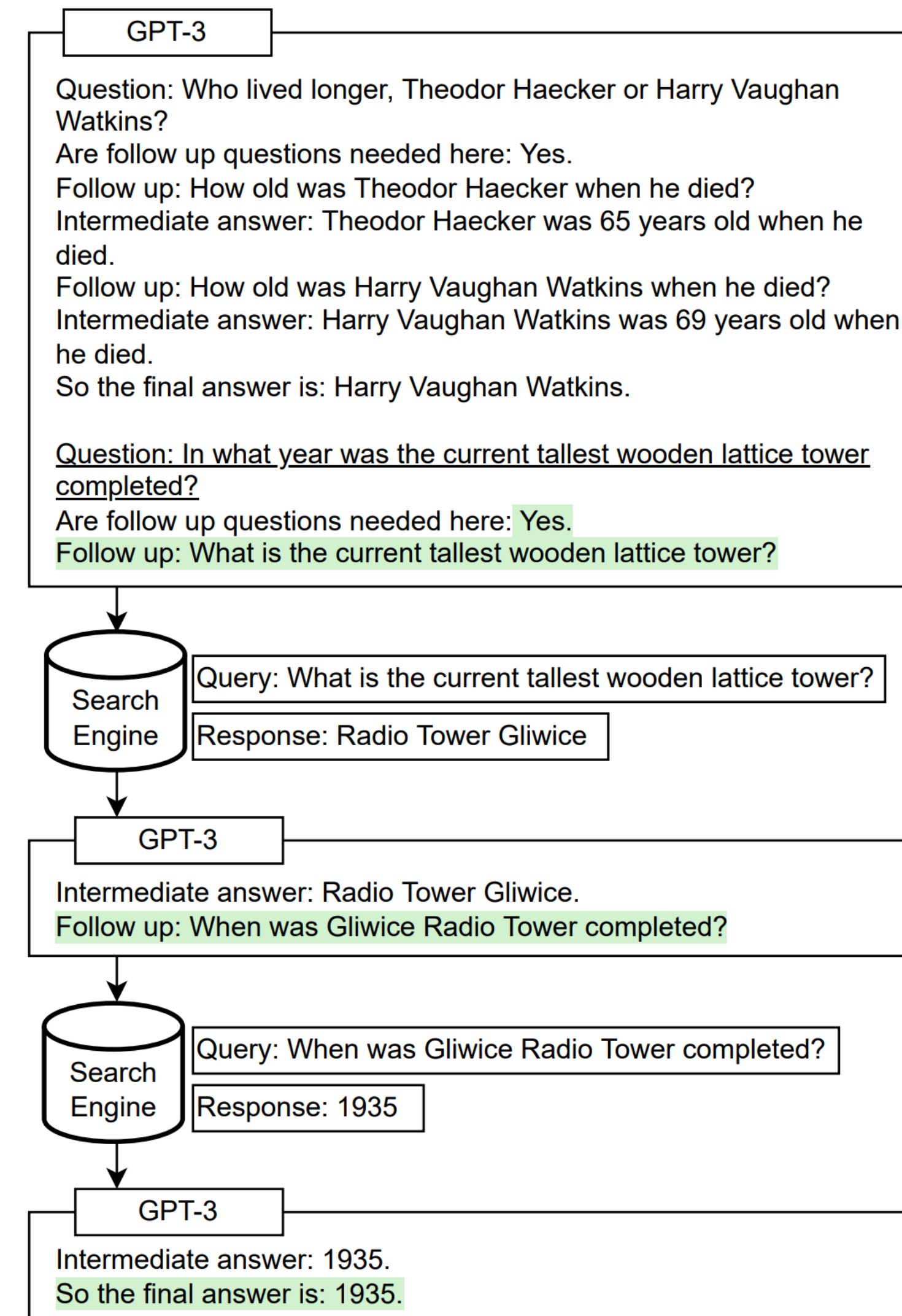
```
>>> print(answer)
```

```
74
```

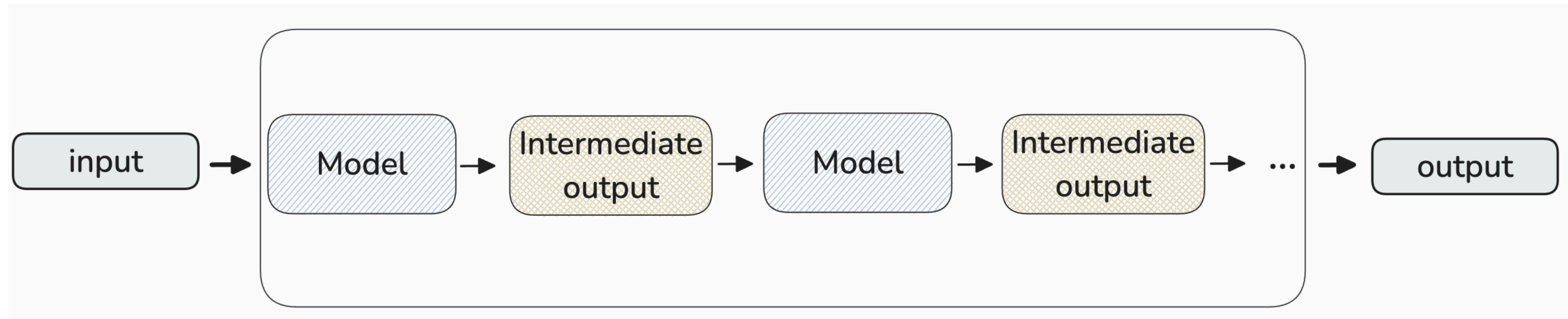


Problem decomposition

- Decompose problems into multiple steps. Example:
- Ask sub-question
- Query search engine
- Repeat




Prompt chains



- More generally, “chain” together multiple calls to prompted models (and/or external functions)

Related software

 [srush.github.io/MiniChain/](https://github.com/srush/MiniChain/)



Mini-Chain

```
@prompt(OpenAI(), template_file="math.pmpt.tpl")
def math_prompt(model, question):
    "Prompt to call GPT with a Jinja template"
    return model(dict(question=question))

@prompt(Python())
def python(model, code):
    "Prompt to call Python interpreter"
    return int(model(code))

def math_demo(question):
    "Chain them together"
    return python(math_prompt(question))
```

<https://srush-minichain.hf.space/>

Related software



DSPy is the framework for *programming—rather than prompting—language models*. It allows you to iterate fast on **building modular AI systems** and offers algorithms for **optimizing their prompts and weights**, whether you're building simple classifiers, sophisticated RAG pipelines, or Agent loops.

DSPy stands for Declarative Self-improving Python. Instead of brittle prompts, you write compositional *Python code* and use DSPy to **teach your LM to deliver high-quality outputs**. This [lecture](#) is a good conceptual introduction. Meet the community, seek help, or start contributing via our [GitHub repo](#) and [Discord server](#).

<https://dspai.ai/>

Questions?