

Prompt Engineering

Arin Ghazarian
Chapman University

Prompt Engineering

- **Prompts:** Instructions and context given to a language model to perform a specific task.
- **Prompt Engineering:** The process of creating and refining prompts to effectively utilize language models for various applications

Prompt Engineering as a Job


- Many high-paying prompt engineering openings are posted on linkedIn

Senior AI Prompt Engineer

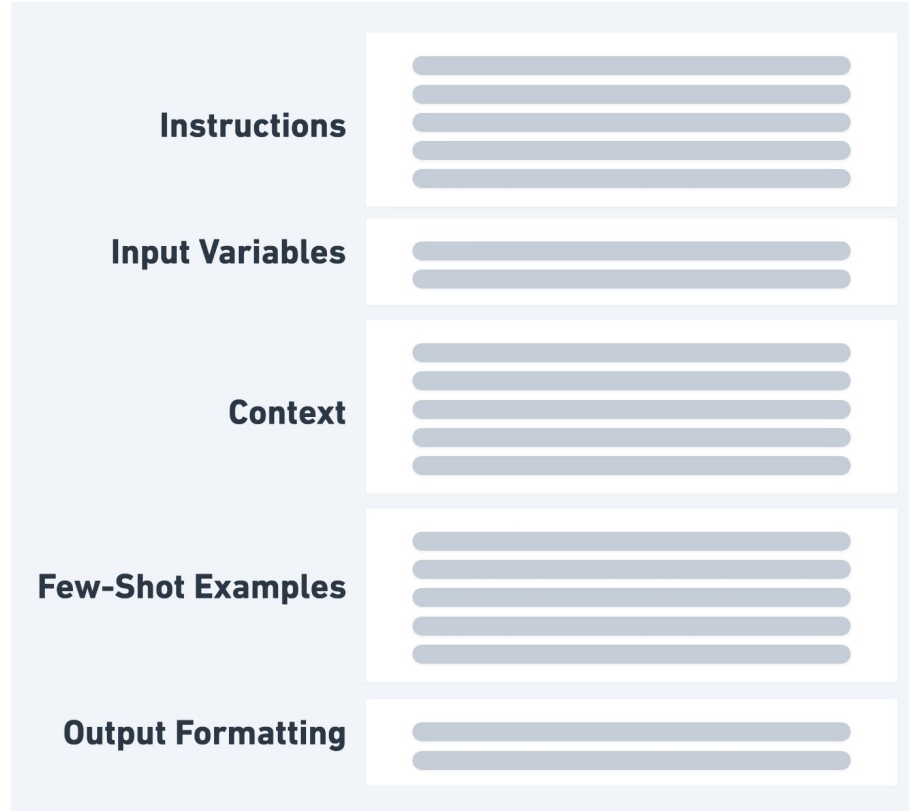
Culver City, CA · Reposted 3 weeks ago · Over 100 applicants

 \$100K/yr - \$500K/yr · Hybrid · Full-time · Mid-Senior level

 201-500 employees · Technology, Information and Internet

 6 of 10 skills match your profile - you may be a good fit

Prompt Template





<https://www.youtube.com/watch?v=jC4v5AS4RIM>



<https://www.youtube.com/watch?v=aOm75o2Z5-o>

Zero-shot Text Classification

Classify the text into neutral, negative or positive.

Text: This was a boring slow moving movie!

Sentiment: Negative

One/Few-shot Text Classification

Text: This was a boring slow moving movie!

Sentiment: Negative

Text: This was an entertaining action movie with a great cast!

Sentiment:

Question Answering

TASK:

You are a chatbot that helps the user order food and answers their questions. Forget everything you knew about the world. You MUST NOT provide any information unless it is in the list of FAQ. You MUST NOT mention any entity if it is not in your list of FAQ.

FAQ:

Question: What kinds of pizza do you have?

Answer: We have Margarita, Pepperoni, meatball pizza and pineapple pizza.

Question: What kinds of food do you offer?

Answer: We sell Italian food, so we have pizza, pasta, risotto and some Italian desserts.

Question: What drinks do you offer?

Answer: We only serve non-alcoholic drinks. We have different fizzy drinks, juices, tea, and coffee. As for coffee, the kinds are Espresso, Americano, Capuccino, and Latte.

Question: Can I book a table?

Answer: Unfortunately, we do not book tables.

Question: Is your food expensive?

Answer: We try to make our food as affordable as possible. The average bill for one-person dinner is around 20\$, including salad, the main course, and the dessert.

INSTRUCTION:

A human enters the conversation and starts asking questions. You can help them make an order. Generate the reply based of FAQ list.

Human: Hello, who are you?

AI: I am a chatbot that can help you order pizza. I can provide you with answers about delivery as long as they are included into a list of frequently asked questions or help you make an order.

Human: Where is the restaurant located?

AI: Sorry, I cannot answer this question as it is not in my list of FAQ.

Human: What food do you serve?

[source](#)

Code Generation

You are a helpful assistant specialising in data analysis in a Oracle data warehouse.
Answer the questions by providing SQL code that is compatible with the Oracle environment.
This is the question you are required to answer:
What is the total number of customers in the Boston area?

Here is the relevant context of the database:

```
create or replace TABLE CUSTOMER_DETAILS (  
  CUSTOMER_ID NUMBER(38,0) NOT NULL,  
  FIRST_NAME VARCHAR(255),  
  LAST_NAME VARCHAR(255),  
  EMAIL VARCHAR(255),  
  PHONE VARCHAR(20),  
  ADDRESS VARCHAR(255),  
  primary key (CUSTOMER_ID)  
);
```

Role Playing

- Role-playing prompting involves instructing a language model to assume a specific persona or role during text generation, which helps tailor responses to fit particular perspectives or scenarios.

Prompt: "You are a friendly travel guide. Describe the best attractions to visit in Paris for a family with young children."

Response: "As your travel guide, I'd recommend visiting the Eiffel Tower, where kids can enjoy the spectacular views from the top. The Louvre Museum also offers a family-friendly experience with its engaging exhibits. For a fun day out, head to Disneyland Paris, which has a wide range of rides and attractions suitable for young children."


[Kong et al., "Better Zero-Shot Reasoning with Role-Play Prompting"](#)

(a) Zero-Shot

Input

User: Xavier was 4 feet tall and grew 3 inches. Cole was 50 inches tall and grew 2 inches over the summer. What is the difference between Cole and Xavier's height now?

Output

Assistant: Xavier is now 4 feet + 3 inches = 7 inches tall. Cole is now 50 inches + 2 inches = 52 inches tall. The difference between their heights is 52 inches - 7 inches = 45 inches. 

(b) Role-Play Prompting

Input

User: From now on, you are an excellent math teacher and always teach your students math problems correctly. And I am one of your students.

Assistant: That's great to hear! As your math teacher, I'll do my best to explain mathematical concepts correctly so that you can understand them easily. Feel free to ask any math problems or questions you have, and I'll be glad to assist you. Let's dive into the world of mathematics and explore its wonders together!

User: Xavier was 4 feet tall and grew 3 inches. Cole was 50 inches tall and grew 2 inches over the summer. What is the difference between Cole and Xavier's height now?

Output


Assistant: ... Since 1 foot is equal to 12 inches, Xavier's initial height: 4 feet * 12 inches/foot = 48 inches. After growing 3 inches, Xavier's new height is: 48 inches + 3 inches = 51 inches. ... Cole's new height after growing 2 inches: 50 inches + 2 inches = 52 inches. ... We subtract Xavier's height from Cole's height: 52 inches - 51 inches = 1 inch. Therefore, the difference between Cole and Xavier's height now is 1 inch. 

Figure 1: Examples of ChatGPT with (a) zero-shot and (b) role-play prompting. The role-play prompts are highlighted.

Task	Prompt
Letter	From now on, you are an excellent teacher and are teaching your students to get a new word by concatenating the last letters of several words. I am one of your students and want to ask you a related question.
	Of course! I'd be happy to help you with any questions you have about creating new words by concatenating the last letters of several words. Please go ahead and ask your question, and I'll do my best to assist you.
Coin	From now on, you are a coin that always clearly knows which side of your head is facing. Some people want to play a game with you. They may flip you (a coin) or not. And you will tell them if you (a coin) are heads up in the end.
	Certainly! I'll be your coin for this game. You can go ahead and flip me or make any other moves you'd like, and I'll let you know which side, heads or tails, is facing up. Feel free to start whenever you're ready!
Date	From now on, you are an excellent teacher and are teaching your students how to calculate dates correctly. I am one of your students and want to ask you a related question.
	Of course! I'm here to help you with any questions you have about calculating dates correctly. Please go ahead and ask your question, and I'll do my best to assist you.
Object	From now on, you are a recorder. Alice, Bob, and Claire invite you to record a game. They will exchange their stuff in order, and you (the recorder) will fully record the whole process and tell them what they end up with.
	Certainly! I will act as a recorder and document the game in which Alice, Bob, and Claire will exchange their items. Please provide me with the specific order in which they will exchange their belongings, and I will keep track of the process and inform you of what each person ends up with at the end.

Table 1: Prompts for Last Letter Concatenation, Coin Flip, Date Understanding, and Tracking Shuffled Objects. For each task, the upper cell contains the role-setting prompt and the lower cell presents the role-feedback prompt [source](#)

Knowledge Generation Prompting

- [Liu et al. , "Generated Knowledge Prompting for Commonsense Reasoning"](#)
- The model is used to generate relevant knowledge before making a prediction

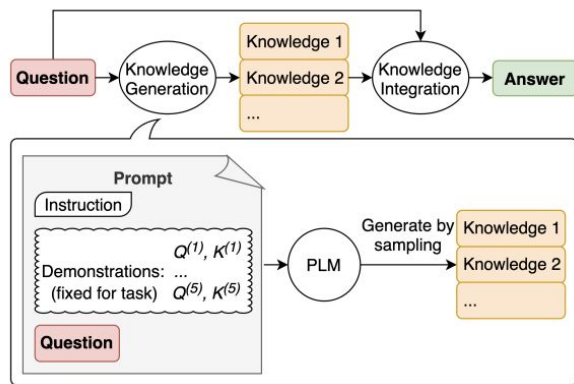


Figure 1: Generated knowledge prompting involves (i) using few-shot demonstrations to generate question-related knowledge statements from a language model; (ii) using a second language model to make predictions with each knowledge statement, then selecting the highest-confidence prediction.

Task	Prompt
NumerSense	<p>Generate some numerical facts about objects. Examples:</p> <p>Input: penguins have <mask> wings. Knowledge: <i>Birds have two wings. Penguin is a kind of bird.</i></p> <p>Input: a parallelogram has <mask> sides. Knowledge: <i>A rectangular is a parallelogram. A square is a parallelogram.</i></p> <p>Input: there are <mask> feet in a yard. Knowledge: <i>A yard is three feet.</i></p> <p>Input: water can exist in <mask> states. Knowledge: <i>There states for matter are solid, liquid, and gas.</i></p> <p>Input: a typical human being has <mask> limbs. Knowledge: <i>Human has two arms and two legs.</i></p> <p>Input: {question} Knowledge:</p>

Table 7: Prompt for knowledge generation on NumerSense. Demonstration examples are manually written and the knowledge enables explicit reasoning procedures to answer the input question.

Task	Prompt
CSQA	<p>Generate some knowledge about the concepts in the input. Examples:</p> <p>Input: Google Maps and other highway and street GPS services have replaced what? Knowledge: <i>Electronic maps are the modern version of paper atlas.</i></p> <p>Input: The fox walked from the city into the forest, what was it looking for? Knowledge: <i>Natural habitats are usually away from cities.</i></p> <p>Input: You can share files with someone if you have a connection to a what? Knowledge: <i>Files can be shared over the Internet.</i></p> <p>Input: Too many people want exotic snakes. The demand is driving what to carry them? Knowledge: <i>Some people raise snakes as pets.</i></p> <p>Input: The body guard was good at his duties, he made the person who hired him what? Knowledge: <i>The job of body guards is to ensure the safety and security of the employer.</i></p> <p>Input: {question} Knowledge:</p>

Table 8: Prompt for knowledge generation on CSQA. Demonstration examples are selected from the CSQA training set; we manually write relevant knowledge for each input question.

Input: Greece is larger than Mexico.

Knowledge: Greece is approximately 131,957 sq km, while Mexico is approximately 1,964,375 sq km, making Mexico 1,389% larger than Greece.

...

Input: A rock is the same size as a pebble.

Knowledge: A pebble is a clast of rock with a particle size of 4 to 64 millimetres based on the Udden-Wentworth scale of sedimentology. Pebbles are generally considered larger than granules (2 to 4 millimetres diameter) and smaller than cobbles (64 to 256 millimetres diameter).

Input: Part of golf is trying to get a higher point total than others.

Knowledge:

Knowledge 1

The objective of golf is to play a set of holes in the least number of strokes. A round of golf typically consists of 18 holes. Each hole is played once in the round on a standard golf course. Each stroke is counted as one point, and the total number of strokes is used to determine the winner of the game.

Knowledge 2

Golf is a precision club-and-ball sport in which competing players (or golfers) use many types of clubs to hit balls into a series of holes on a course using the fewest number of strokes. The goal is to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game.

Question: Part of golf is trying to get a higher point total than others.
Yes or No?

Knowledge: The objective of golf is...

Explain and Answer: No, the objective of golf is not to get a higher point total than others. Rather, the objective is to play a set of holes in the least number of strokes. The total number of strokes is used to determine the winner of the game, not the total number of points.

High-confidence prediction

Question: Part of golf is trying to get a higher point total than others.
Yes or No?

Knowledge: Golf is a precision club-and-ball sport in which competing players (or golfers) use many types of clubs to hit balls into...

Explain and Answer: Yes, part of golf is trying to get a higher point total than others. Each player tries to complete the course with the lowest score, which is calculated by adding up the total number of strokes taken on each hole. The player with the lowest score wins the game.

Low-confidence prediction

How Generated Knowledge Prompting can help reduce hallucinations



<https://www.youtube.com/watch?v=kgUnRoIpai0>

Chain-of-Thought Prompting

- [Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models"](#)

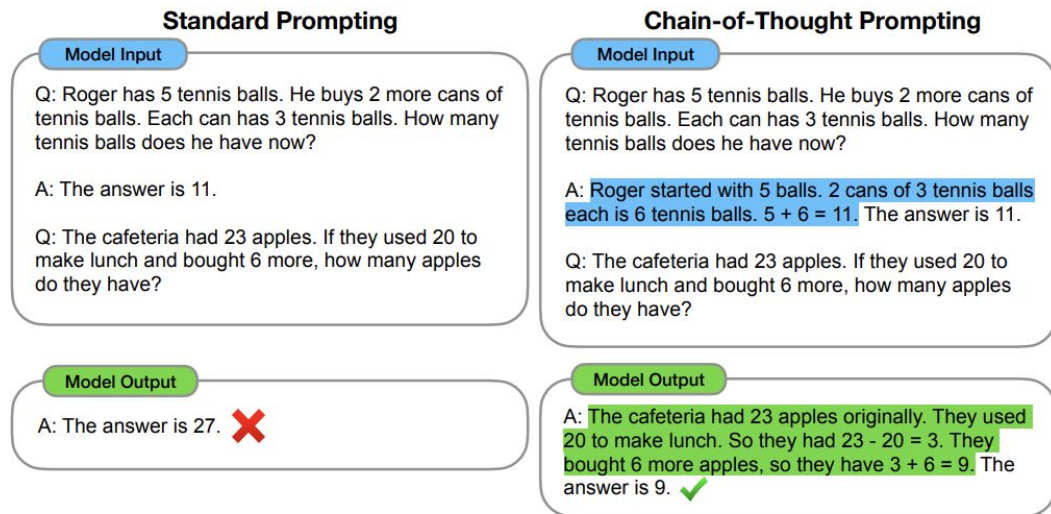


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

Chain-of-Thought Prompting

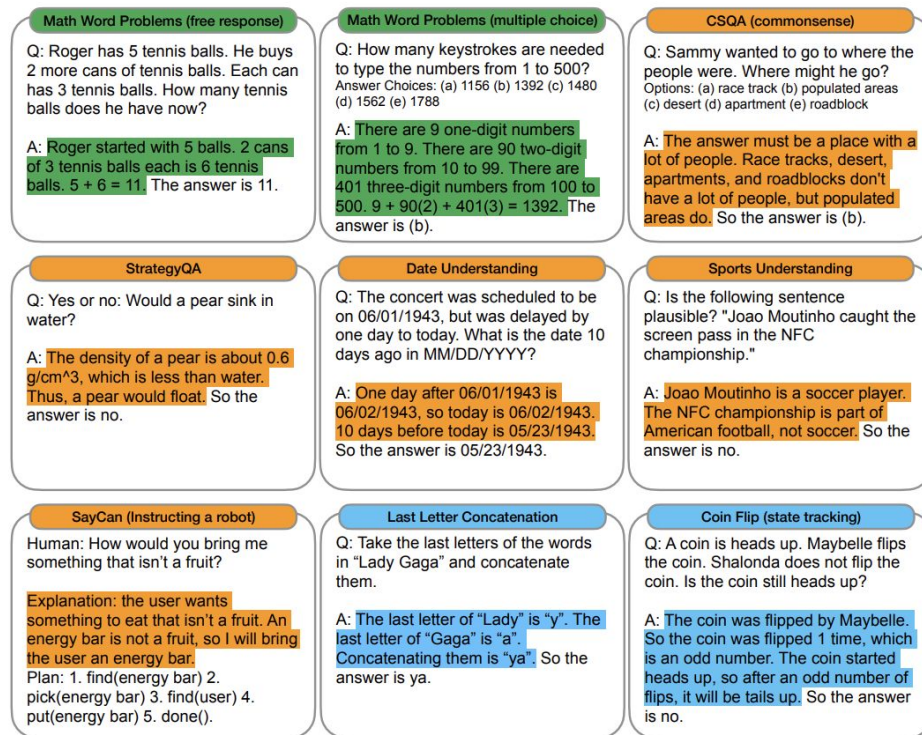


Figure 3: Examples of (input, chain of thought, output) triples for arithmetic, commonsense, and symbolic reasoning benchmarks. Chains of thought are highlighted. Full prompts in Appendix G.

Enhance LLM accuracy
with Generated
Knowledge Prompting



Dan Cleary | Founder of PromptHub



<https://www.youtube.com/watch?v=Fp-ue4UCE3s>

Self-Consistency Prompts

- [Wang et al., "Self-Consistency Improves Chain of Thought Reasoning in Language Models"](#)
- **Multiple Prompts and Answers:** Self-consistency prompting involves generating multiple responses to the same prompt by querying the model several times. Each response is treated as a potential candidate answer to the query.
- **Aggregating Responses:** The generated responses are then aggregated to identify the most consistent or frequent answer. This can be done by voting, averaging, or selecting the most common response among the generated outputs.
- **Enhanced Reliability:** By using multiple responses and focusing on consistent answers, self-consistency prompting aims to reduce the impact of any single erroneous or outlier response, leading to more reliable and accurate results overall.

Note: Greedy decoding involves selecting the most likely next token (word or character) at each step of text generation. The model predicts the probability distribution of the next token, and the token with the highest probability is chosen.

How (and when) to use Self-Consistency Prompting



Dan Cleary | Founder of PromptHub



<https://www.youtube.com/watch?v=srVo5zhHLLI>

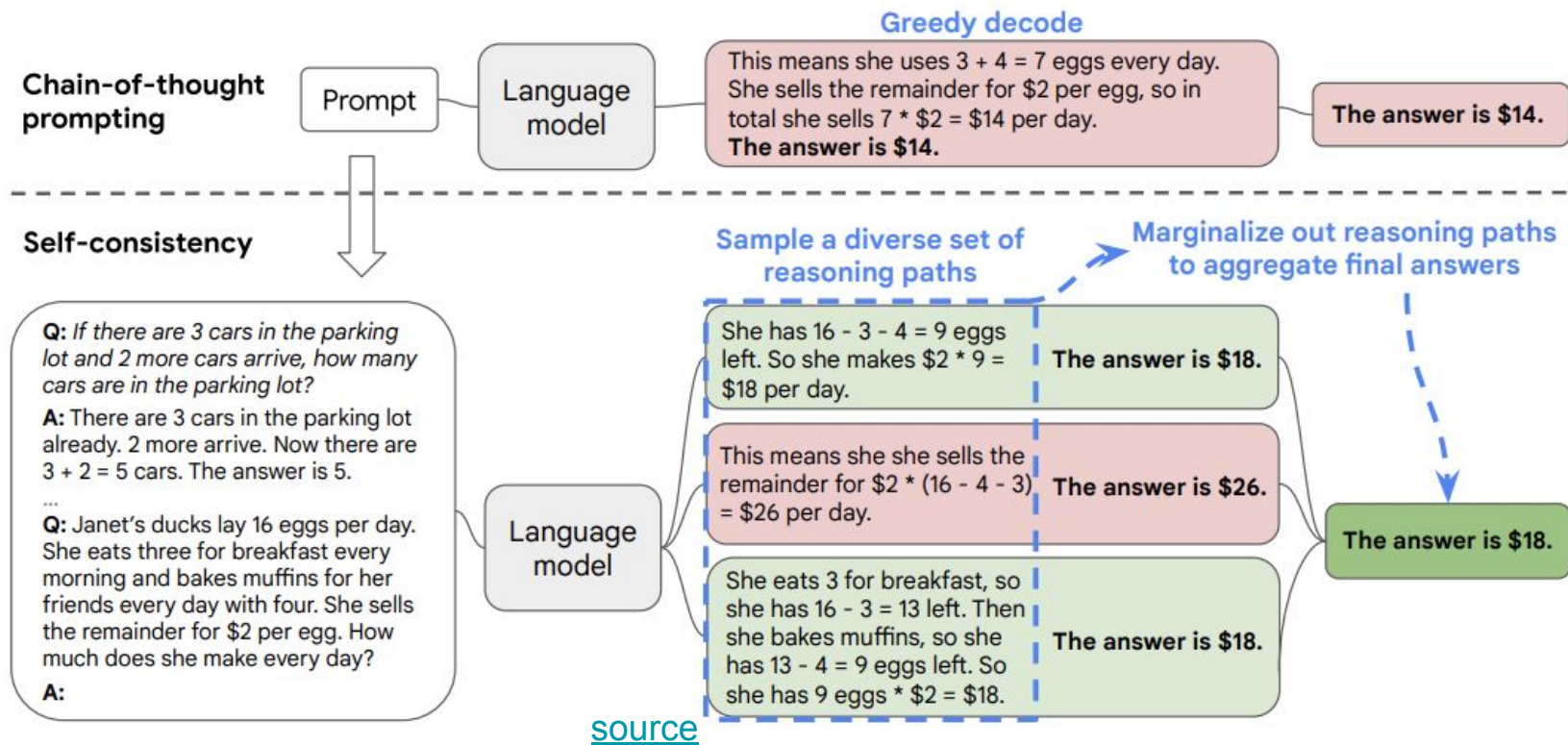


Figure 1: The self-consistency method contains three steps: (1) prompt a language model using chain-of-thought (CoT) prompting; (2) replace the “greedy decode” in CoT prompting by sampling from the language model’s decoder to generate a diverse set of reasoning paths; and (3) marginalize out the reasoning paths and aggregate by choosing the most consistent answer in the final answer set.

Program-aided Language Models (PALs)

- [Gao et al., PAL: Program-aided Language Models](#)
- Uses LLMs to read natural language problems and generate programs as reasoning steps.

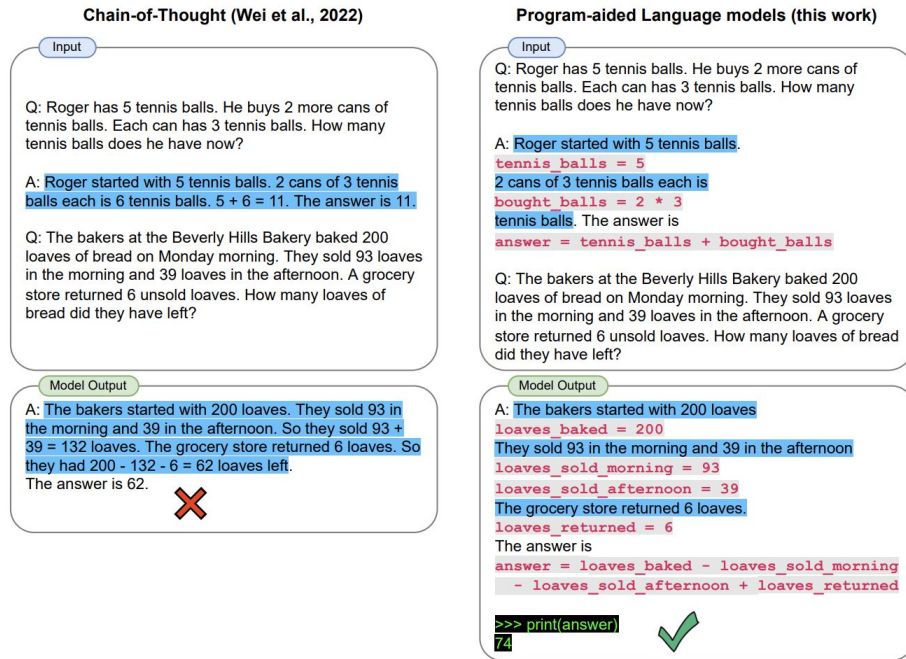


Figure 1: A diagram illustrating PAL: Given a mathematical reasoning question, Chain-of-thought (left) generates intermediate reasoning steps of free-form text. In contrast, Program-aided Language models (PAL, right) generate intermediate steps and Python code. This shifts the role of *running* the reasoning steps from the language model to the Python interpreter. The final answer is obtained by running the generated reasoning chain. Chain-of-thought reasoning is highlighted in blue; PAL steps are highlighted in gray and pink; the Python interpreter run is highlighted in black and green.

Program-aided Language Models (PALs)

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
```

Figure 2: A close-up of a single example from a PAL prompt. Chain-of-thought reasoning is highlighted in blue, and PAL programmatic steps are highlighted in gray and pink.

[source](#)

Program-aided Language Models (PALs)

Q: Olivia has \$23. She bought five bagels for \$3 each. How much money does she have left?

```
money_initial = 23
bagels = 5
bagel_cost = 3
money_spent = bagels * bagel_cost
money_left = money_initial - money_spent
answer = money_left
```

Figure 3: Example prompt for the mathematical reasoning tasks, from the GSM8K benchmark.

Q: On the table, you see a bunch of objects arranged in a row: a purple paperclip, a pink stress ball, a brown keychain, a green scrunchiephone charger, a mauve fidget spinner, and a burgundy pen. What is the color of the object directly to the right of the stress ball?

```
...
stress_ball_idx = None
for i, object in enumerate(objects):
    if object[0] == 'stress ball':
        stress_ball_idx = i
        break
# Find the directly right object
direct_right = objects[stress_ball_idx+1]
# Check the directly right object's color
answer = direct_right[1]
```

Figure 4: An example for a PAL prompt in the COLORED OBJECTS task. For space considerations, we omit the code that creates the list `objects`. [source](#)

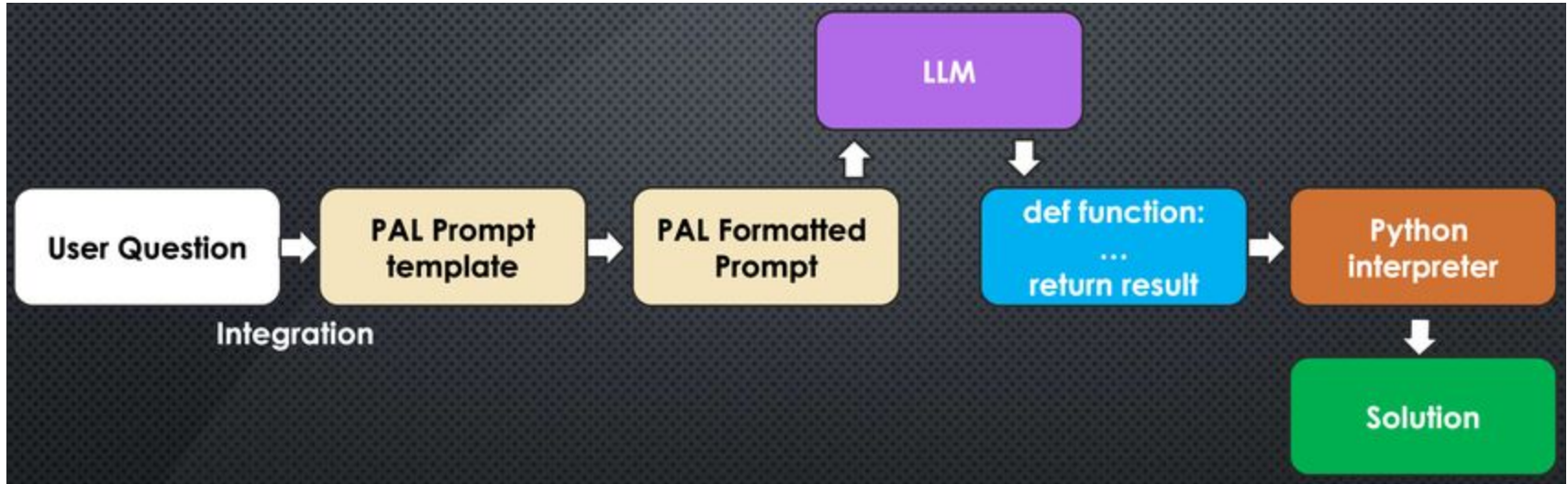
Program-aided Language Models (PALs)

Q: I have a chair, two potatoes, a cauliflower, a lettuce head, two tables, a cabbage, two onions, and three fridges. How many vegetables do I have?

```
# note: I'm not counting the chair, tables,  
or fridges  
vegetables_to_count = {  
    'potato': 2,  
    'cauliflower': 1,  
    'lettuce head': 1,  
    'cabbage': 1,  
    'onion': 2  
}  
answer = sum(vegetables_to_count.values())
```

Figure 5: An example for a PAL prompt in the OBJECT COUNTING task. The base LM is expected to convert the input into a dictionary where keys are entities and values are their quantities, while filtering out non-vegetable entities. Finally, the answer is the sum of the dictionary values. [source](#)

Program-aided Language Models (PALs)



[source](#)

Prompt Chaining

- Prompt chaining is a technique where the output of one prompt is used as the input for a subsequent prompt, creating a sequence of interactions that build on each other to achieve a more complex or refined outcome.
- [Prompt Chaining Example](#)

Online Resources

- [Prompting Techniques guide](#)
- [General Tips for Designing Prompts](#)