An Introduction to Prompt Engineering

What is Prompt Engineering?

Simply put, It is the process of designing effective prompts to achieve better performance from LLMs.

There are many different techniques for prompt engineering some of which we are going to discuss about.

Prompt Engineering Techniques

- Zero-shot Prompting
- Few-shot Prompting
- Chain-of-thought Prompting
- Chain-of-thought Prompting with Self-Consistency
- Auto CoT
- Tree-of-thoughts

Zero-shot Prompting

It is the simplest type of prompting in which the user asks the question without any exemplars.

Example Prompt:

Classify the text into neutral, negative or positive.

Text: I think the vacation is okay.

Sentiment:

Few-shot Prompting

In few-shot prompting, the user provides the LLM with several exemplars of desired answers.

Example:

```
Classify the text into neutral, negative or positive.
Text: The answer he gave was horrible.
Sentiment: Negative
Text: I think the holiday will be fun.
Sentiment: Positive
Text: I think the vacation is okay.
Sentiment:
```

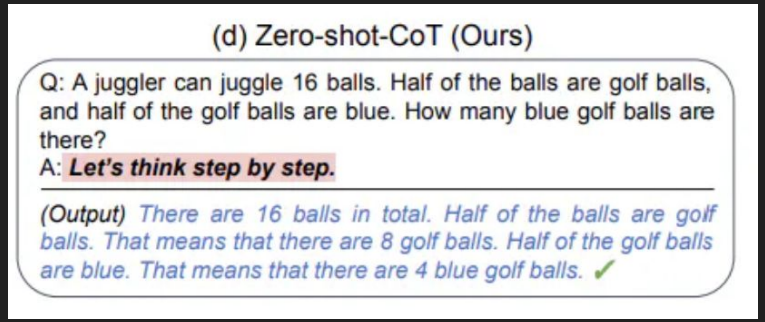
Chain-of-thought Prompting

Chain of thought prompting has two major paradigms. The first one is zero-shot CoT prompting, and the second one is manual CoT prompting.

CoT prompting can be summarized as forcing LLMs to output the reasoning with the answer instead of simply outputting the answer.

Zero-shot CoT

This is the simpler approach in which the LLM is forced to ‘reason’ about the answer. Zero-shot CoT is realized by simply adding ‘Let’s think step by step’ sentence to the end of the prompt.



Manual CoT

In manual CoT, the LLM is provided with exemplars of thought chains, reasonings. Compared to Zero-shot CoT, this approach yields better performance at the expense of manual effort for designing prompts.

Take a look at the next slide for an example.

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

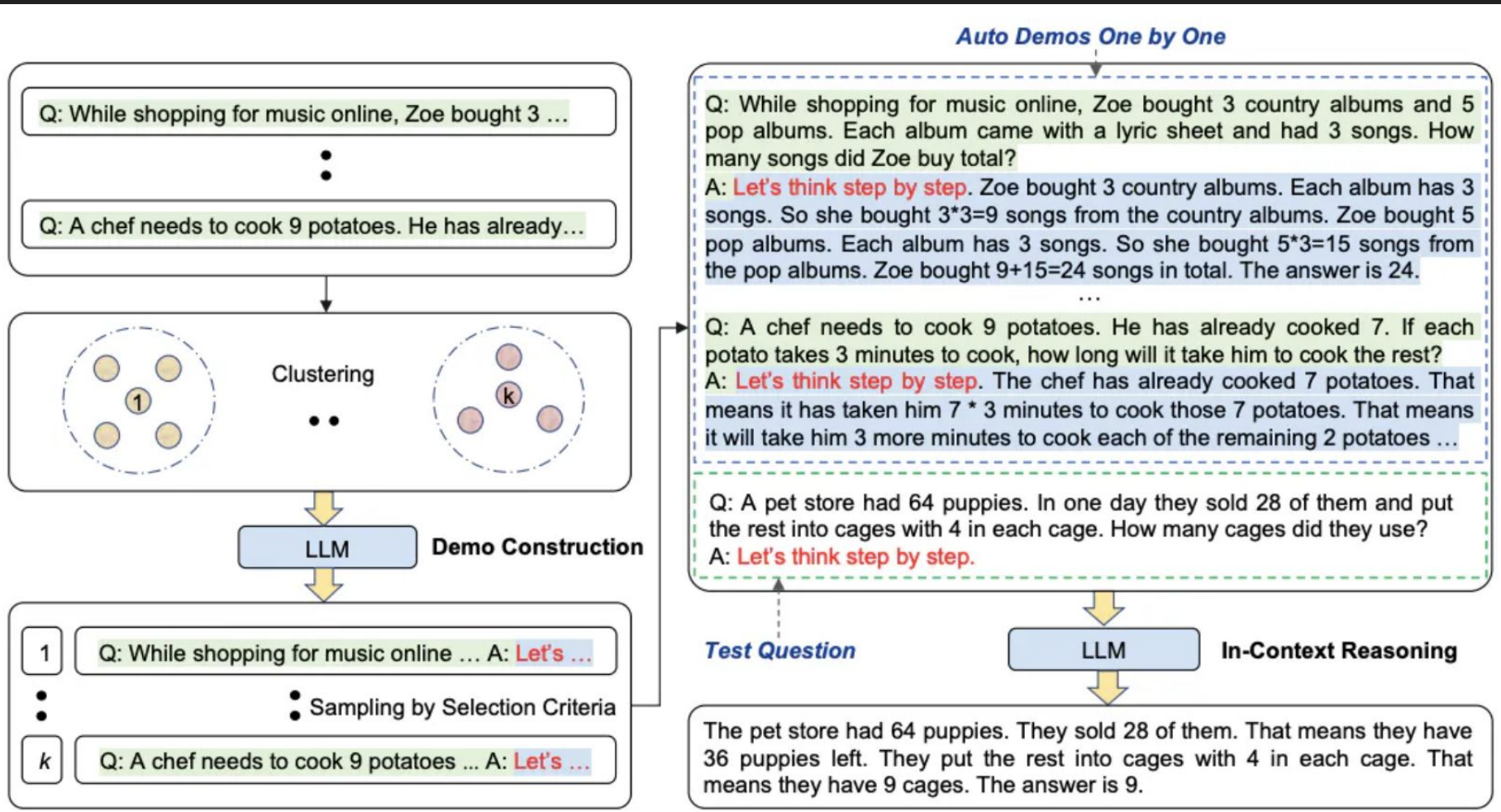
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

Auto-CoT

Auto-CoT is an approach which tries to utilize both CoT paradigms to achieve better performance.

Here is the outline of the process:

1. The questions are gathered.
2. Questions are clustered according to their similarity.
3. A representative question is selected from each cluster.
4. Zero-shot CoT prompting is applied to each representative question.
5. The answers of the representative questions are gathered. At this point, the question and the answer pairs form pseudo manual CoT prompts.
6. The pseudo manual CoT prompts are used as if they were manual CoT prompts.

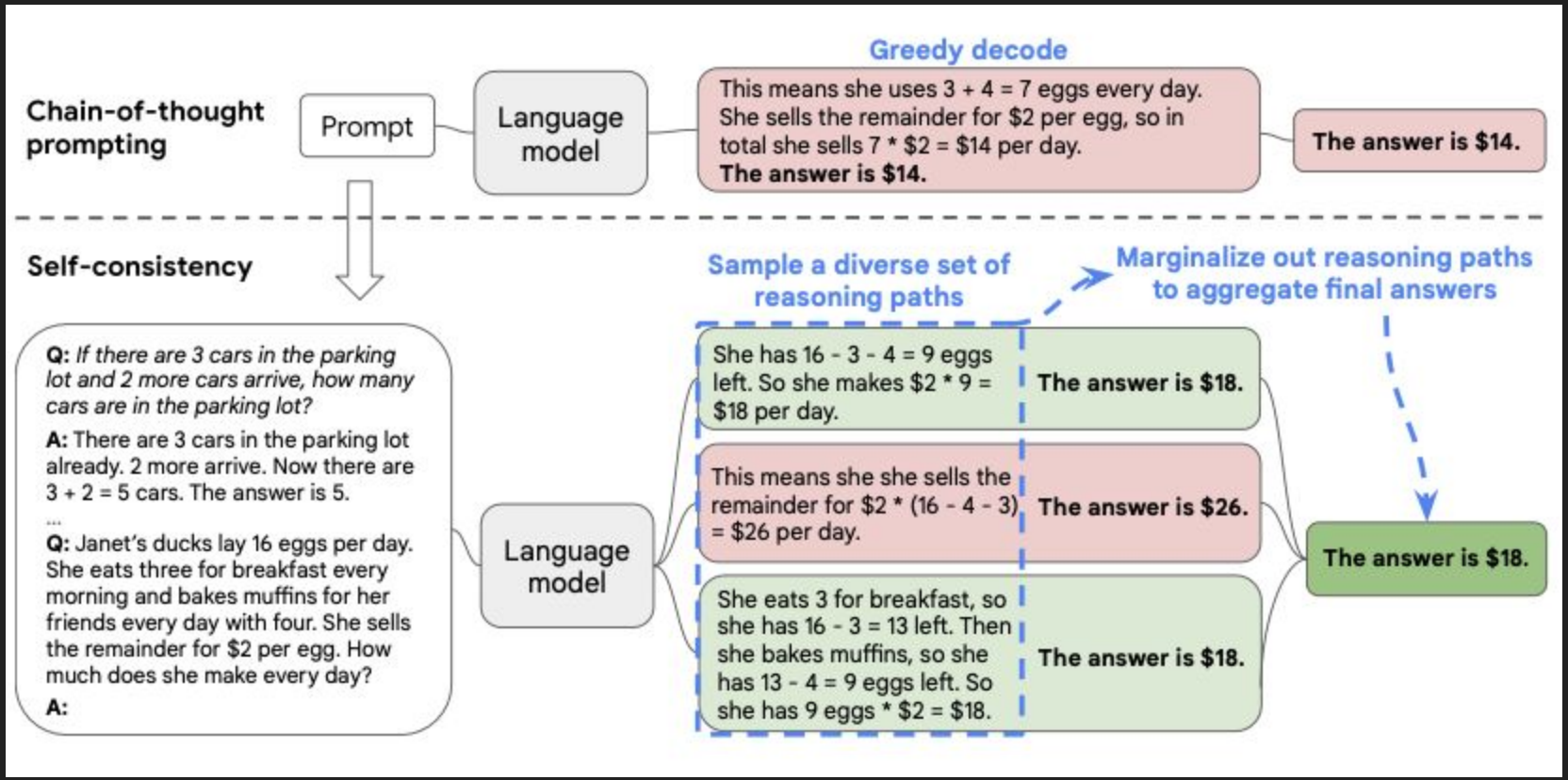

Q: While shopping for music online, Zoe bought 3 …
Q: A chef needs to cook 9 potatoes. He has already…
Clustering
1
k
LLM
Demo Construction
1
Q: While shopping for music online … A: Let's …
Sampling by Selection Criteria
k
Q: A chef needs to cook 9 potatoes ... A: Let's …
Auto Demos One by One
Q: While shopping for music online, Zoe bought 3 country albums and 5 pop albums. Each album came with a lyric sheet and had 3 songs. How many songs did Zoe buy total?
A: Let's think step by step. Zoe bought 3 country albums. Each album has 3 songs. So she bought 3*3=9 songs from the country albums. Zoe bought 5 pop albums. Each album has 3 songs. So she bought 5*3=15 songs from the pop albums. Zoe bought 9+15=24 songs in total. The answer is 24.
…
Q: A chef needs to cook 9 potatoes. He has already cooked 7. If each potato takes 3 minutes to cook, how long will it take him to cook the rest?
A: Let's think step by step. The chef has already cooked 7 potatoes. That means it has taken him 7 * 3 minutes to cook those 7 potatoes. That means it will take him 3 more minutes to cook each of the remaining 2 potatoes …
Q: A pet store had 64 puppies. In one day they sold 28 of them and put the rest into cages with 4 in each cage. How many cages did they use?
A: Let's think step by step.
Test Question
LLM
In-Context Reasoning
The pet store had 64 puppies. They sold 28 of them. That means they have 36 puppies left. They put the rest into cages with 4 in each cage. That means they have 9 cages. The answer is 9.

Chain-of-Thought Prompting with Self Consistency

CoT with self consistency is a self-ensemble technique which uses multiple outputs to arrive at the correct answer.

Here is the process:

1. Use CoT prompting, either manual or zero-shot.
2. Sample multiple outputs from the same LLM, hence the term self-ensemble.
3. Marginalize out the thought chain of each output and find the most consistent answer, i.e. the most frequent answer.

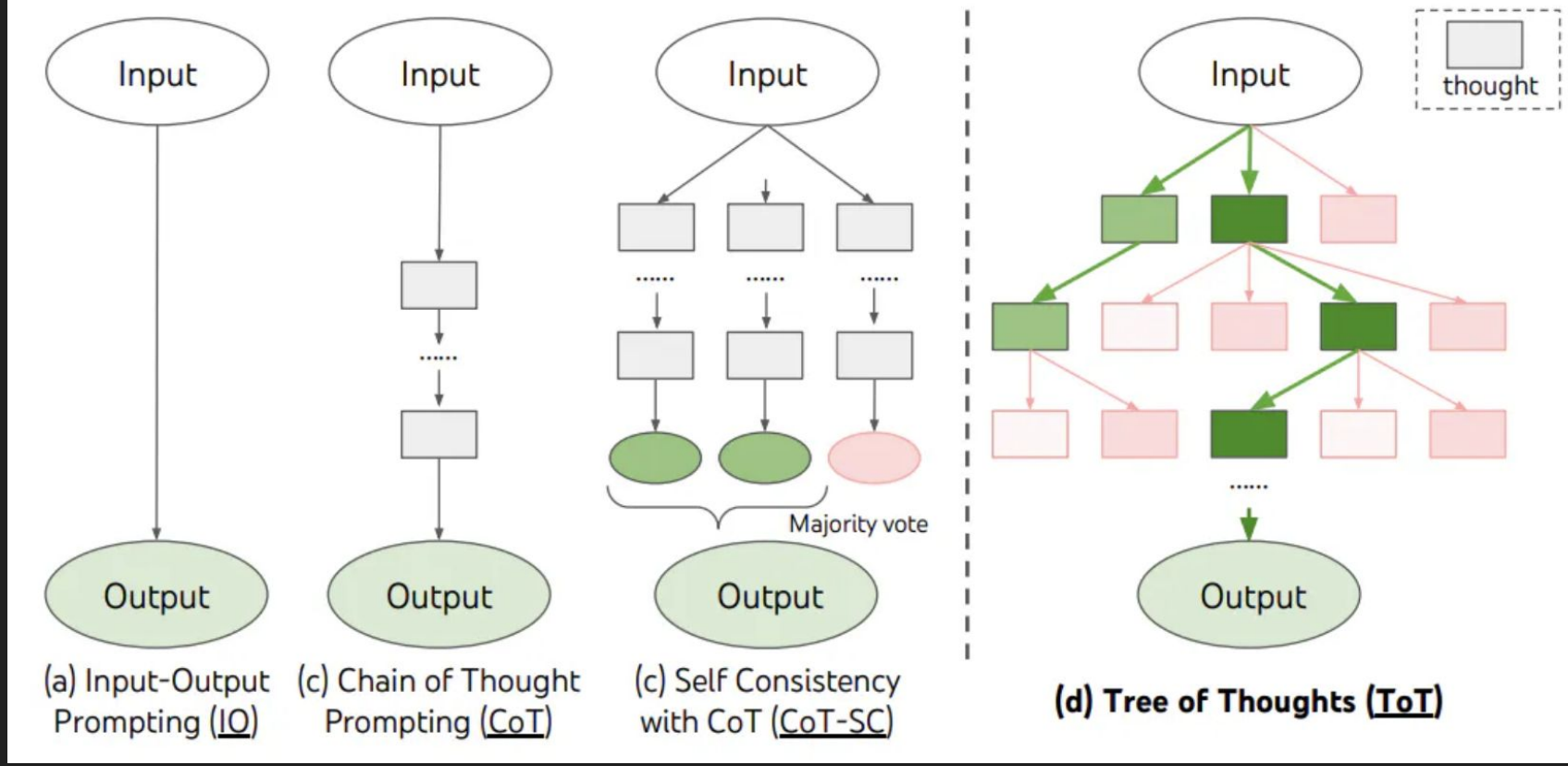

Chain-of-thought prompting
Prompt
Language model
Greedy decode
This means she uses 3 + 4 = 7 eggs every day. She sells the remainder for $2 per egg, so in total she sells 7 * $2 = $14 per day. The answer is $14.
The answer is $14.
Self-consistency
Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?
A: There are 3 cars in the parking lot already. 2 more arrive. Now there are 3 + 2 = 5 cars. The answer is 5.
...
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 a diverse set of reasoning paths
Marginalize out reasoning paths to aggregate final answers
She has 16 - 3 - 4 = 9 eggs left. So she makes $2 * 9 = $18 per day.
The answer is $18.
This means she she sells the remainder for $2 * (16 - 4 - 3) = $26 per day.
The answer is $26.
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 eggs * $2 = $18.
The answer is $18.
The answer is $18.

Tree-of-Thoughts

For complex tasks that require exploration or strategic lookahead, traditional or simple prompting techniques fall short.

How does ToT work?

- Generate Multiple Thoughts: For each sub-problem or step, the LLM is instructed to generate multiple potential "thoughts" (e.g., equations, lines of code, text snippets, etc.).
- Evaluate Thoughts: The LLM then evaluates the quality of each generated thought, often using a self-evaluation process or additional heuristics.
- Build the Tree: The best thoughts are selected, forming branches of a "thought tree". The LLM moves along a branch, exploring different solution paths.
- Search and Refinement: Search strategies (like breadth-first or depth-first) help navigate the tree, and the process can involve backtracking and refining previous thoughts.


Input
Output
(a) Input-Output Prompting (IO)
Input
Output
(c) Chain of Thought Prompting (CoT)
Input
Majority vote
Output
(c) Self Consistency with CoT (CoT-SC)
thought
Input
Output
(d) Tree of Thoughts (ToT)