

PROMPT ENGINEERING

Topics

- Introduction to Prompt Engineering
- Some Advanced Techniques for Prompt Engineering
- Conclusion & Future Directions

What are prompts?

- **Prompts** involve instructions and context passed to a language model to achieve a desired task
- **Prompt engineering** is the practice of developing and optimizing prompts to efficiently use language models (LMs) for a variety of applications
- Why?
 - Helps to test and evaluate the uses and limitations of LLMs
 - Enables all kinds of innovative applications on top of LLMs

Prompt Engineering (stated another way)

What is prompt engineering?

Prompt engineering is a process of creating a set of prompts, or questions, that are used to guide the user toward a desired outcome. It is an effective tool for designers to create user experiences that are easy to use and intuitive. This method is often used in interactive design and software development, as it allows users to easily understand how to interact with a system or product..

<https://github.com/dair-ai/Prompt-Engineering-Guide/blob/main/lecture/Prompt-Engineering-Lecture-Elvis.pdf>

Why Prompt Engineering?

ANTHROPIC

Prompt Engineer and Librarian

APPLY FOR THIS JOB

SAN FRANCISCO, CA / PRODUCT / FULL-TIME / HYBRID

Anthropic's mission is to create reliable, interpretable, and steerable AI systems. We want AI to be safe for our customers and for society as a whole.

Anthropic's AI technology is amongst the most capable and safe in the world. However, large language models are a new type of intelligence, and the art of instructing them in a way that delivers the best results is still in its infancy – it's a hybrid between programming, instructing, and teaching. You will figure out the best methods of prompting our AI to accomplish a wide range of tasks, then document these methods to build up a library of tools and a set of tutorials that allow others to learn prompt engineering or simply find prompts that would be ideal for them.

Compensation and Benefits*

Anthropic's compensation package consists of three elements: salary, equity, and benefits. We are committed to pay fairly and aim for these three elements collectively to be highly competitive with market rates.

Salary - The expected salary range for this position is \$250k - \$335k.

Equity - Equity will be a major component of the total compensation for this position. We aim to offer higher-than-average equity compensation for a company of our size, and communicate equity amounts at the time of offer issuance.

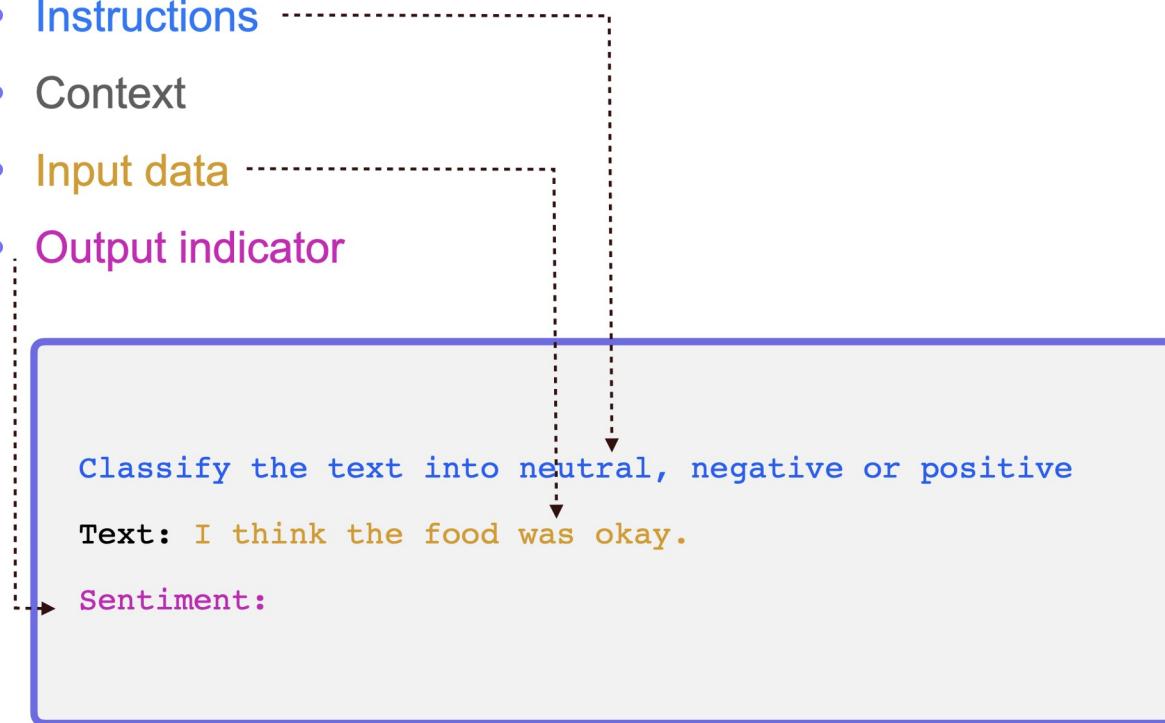
<https://github.com/dair-ai/Prompt-Engineering-Guide/blob/main/lecture/Prompt-Engineering-Lecture-Elvis.pdf>

Why Prompt Engineering?

You get better results / answers from the LLM

Elements of a Prompt

- A prompt is composed with the following components:
 - Instructions
 - Context
 - Input data
 - Output indicator



<https://github.com/dair-ai/Prompt-Engineering-Guide/blob/main/lecture/Prompt-Engineering-Lecture-Elvis.pdf>

Designing Prompts for Different Tasks

- Examples of common tasks using different prompts
- Tasks covered:
 - Text Summarization
 - Question Answering
 - Text Classification
 - Role Playing
 - Code Generation
 - Reasoning

Text Summarization

Context

Antibiotics are a type of medication used to treat bacterial infections. They work by either killing the bacteria or preventing them from reproducing, allowing the body's immune system to fight off the infection.

----> Antibiotics are usually taken orally in the form of pills, capsules, or liquid solutions, or sometimes administered intravenously. They are not effective against viral infections, and using them inappropriately can lead to antibiotic resistance.

----> Explain the above in one sentence:

Antibiotics are medications used to treat bacterial infections by either killing the bacteria or stopping them from reproducing, but they are not effective against viruses and overuse can lead to antibiotic resistance.

Question & Answering

Answer the question based on the context below. Keep the answer short and concise. Respond "Unsure about answer" if not sure about the answer.

Context: Teplizumab traces its roots to a New Jersey drug company called Ortho Pharmaceutical. There, scientists generated an early version of the antibody, dubbed OKT3. Originally sourced from mice, the molecule was able to bind to the surface of T cells and limit their cell-killing potential. In 1986, it was approved to help prevent organ rejection after kidney transplants, making it the first therapeutic antibody allowed for human use.

Question: What was OKT3 originally sourced from?

Answer: Mice.

Text Classification

Classify the text into neutral, negative or positive.

Text: I think the food was okay.

Sentiment: Neutral

Role Playing

The following is a conversation with an AI research assistant. The assistant tone is technical and scientific.

Human: Hello, who are you?

AI: Greeting! I am an AI research assistant. How can I help you today?

Human: Can you tell me about the creation of blackholes?

AI: Sure! Black holes are regions of spacetime where the gravitational force is so strong that nothing, not even light, can escape from it. They are created when a very massive star dies and its core collapses in on itself, forming a singularity of infinite density. The intense gravity of the singularity pulls in all the matter and radiation around it, creating the black hole.

Code Generation

```
"""
Table departments, columns = [DepartmentId,
DepartmentName]
Table students, columns = [DepartmentId, StudentId,
StudentName]
Create a MySQL query for all students in the Computer
Science Department
"""

SELECT StudentId, StudentName
FROM students
WHERE DepartmentId IN (SELECT DepartmentId FROM
departments WHERE DepartmentName = 'Computer Science');
```

Reasoning

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

Solve by breaking the problem into steps. First, identify the odd numbers, add them, and indicate whether the result is odd or even.

Odd numbers: 15, 5, 13, 7, 1

Sum: 41

41 is an odd number.

Prompt Engineering Techniques

- Many prompting techniques have been designed to improve performance on complex tasks:
 - Zero-shot prompt
 - Few-shot prompts
 - Chain-of-thought (CoT) prompting
 - Self-Consistency
 - ReAct

Prompt Engineering: “Query Crafting”



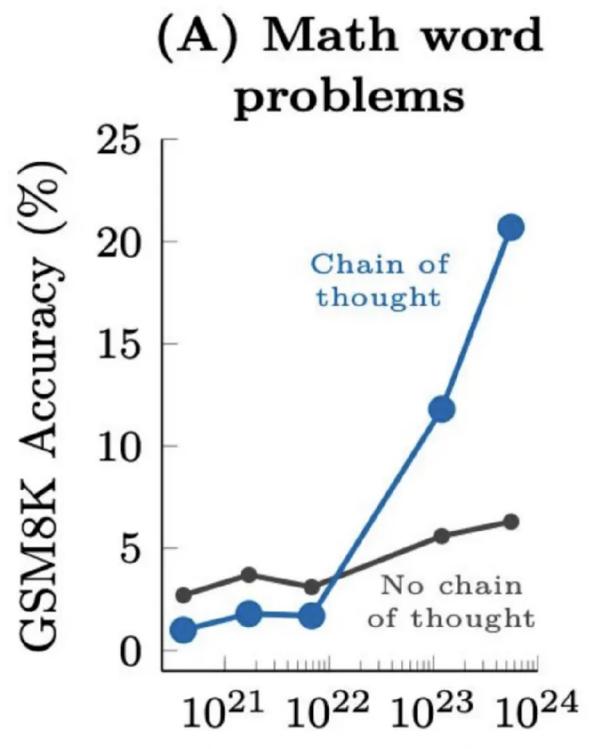
Improving the output with actions like **phrasing queries, specifying styles, providing context, or assigning roles** (e.g., 'Act as a mathematics teacher') (Wikipedia, 2023).

Some hints can be found in OpenAI’s “GPT best practices” (OpenAi, 2023).

Chain-of-thought: popular technique consisting in “guiding [LLMs] to produce a sequence of intermediate steps before giving the final answer” (Wei et al., 2022).

Sources:

- Wei, J. et al., 2022. Emergent abilities of large language models, arxiv.org/abs/2206.07682
- OpenAI, 2023, platform.openai.com/docs/guides/gpt-best-practices/six-strategies-for-getting-better-results
- Wikipedia, 2023, , Prompt Engineering, en.wikipedia.org/wiki/Prompt_engineering



(graph from Wei et al., 2022)

About GSM8K benchmark: arxiv.org/abs/2110.14168

Zero-Shot Prompting

Zero-shot prompting provides no examples in prompts. The LLM is steered towards better performance via instructions

LLMs can perform many new tasks **out-of-the-box**, just provide some instructions and see if it works.

See example on the right.

User	LLM
Translate from German to English but use only words that start with an "f". German: Die Katze schläft gerne in der Box. English:	Feline friend finds fluffy fortress.

<https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f>



Prompt Engineering: In-Context Learning (ICL)

In-Context Learning (ICL) consists in “a few input-output examples in the model’s context (input) as a preamble before asking the model to perform the task for an unseen inference-time example” (Wei et al., 2022).

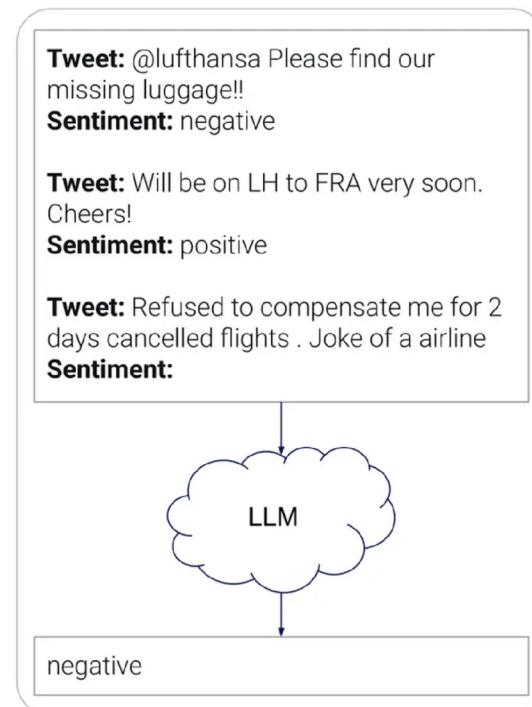
It is a kind of “**ephemeral supervised learning**.”

- **Zero-shot prompting** or Zero-shot learning: no example given (for largest LLMs, smaller ones may struggle).
- **One-shot prompting**: one example provided.
- **Few-shot prompting**: a few examples (typically 3~6).



Context window limits (e.g., 4096 tokens).

Source: Wei, J. et al., 2022. Emergent abilities of large language models, arxiv.org/abs/2206.07682



Example of an input and output for two-shot prompting

Few-Shot Prompting

Few-shot prompting allows us to provide exemplars in prompts to steer the model towards better performance

Providing **examples** helps the LLM understand and follow your task.

This is especially helpful to ensure a specific **output format**.

User
Standardize different currency formats.

Examples:

- Sandwich EUR 8.9 -> €8.90
- 14.50 pounds Burger-> £14.00
- 4\$ for a Milkshake -> \$4.00
- Toppings are 99 cents each -> €0.99
- Steak: 24.99 USD ->

LLM
\$24.99

<https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f>

Few-Shot Prompting

The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.

A: The answer is False.

The odd numbers in this group add up to an even number: 17, 10, 19, 4, 8, 12, 24.

A: The answer is True.

The odd numbers in this group add up to an even number: 16, 11, 14, 4, 8, 13, 24.

A: The answer is True.

The odd numbers in this group add up to an even number: 17, 9, 10, 12, 13, 4, 2.

A: The answer is False.

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

A: The answer is True.

Chain-of-Thought (CoT) Prompting

- Instructing the model to reason about the task when responding:
 - This is very useful for tasks that requiring reasoning
 - You can combine it with few-shot prompting to get better results
 - You can also do zero-shot CoT where exemplars are not available

Ask the model to solve complex tasks
step by step.

Why does this work?

It gives the model a **working memory**, similar to humans.

User
Who won the World Cup in the year before Lionel Messi was born? Think step by step.

LLM
Lionel Messi was born on June 24, 1987. The World Cup that took place before his birth was the 1986 World Cup. The winner of the 1986 FIFA World Cup was Argentina.

<https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f>

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 Elicits Reasoning in Large Language Models

https://rdi.berkeley.edu/understanding_llms/assets/jan19.pdf

Chain-of-Thought (CoT) Prompting

The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.

A: Adding all the odd numbers (9, 15, 1) gives 25. The answer is False.

The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.

A: Adding all the odd numbers (15, 5, 13, 7, 1) gives 41. The answer is False.

Zero-Shot Chain-of-Thought Prompting

Involves adding "Let's think step by step" to the original prompt

I went to the market and bought 10 apples. I gave 2 apples to the neighbor and 2 to the repairman. I then went and bought 5 more apples and ate 1. How many apples did I remain with?

11 apples

I went to the market and bought 10 apples. I gave 2 apples to the neighbor and 2 to the repairman. I then went and bought 5 more apples and ate 1. How many apples did I remain with?

Let's think step by step.

First, you started with 10 apples.
You gave away 2 apples to the neighbor and 2 to the repairman, so you had 6 apples left.
Then you bought 5 more apples, so now you had 11 apples.
Finally, you ate 1 apple, so you would remain with 10 apples.

<https://github.com/dair-ai/Prompt-Engineering-Guide/blob/main/lecture/Prompt-Engineering-Lecture-Elvis.pdf>

Self Consistency

- Self-Consistency aims to improve chain-of-thought prompting
- The idea is to sample multiple, diverse reasoning paths through few-shot CoT, and use the generations to select the most consistent answer.
- This helps to boost the performance of CoT prompting on tasks involving arithmetic and commonsense reasoning

When I was 6 my sister was half my age. Now
I'm 70 how old is my sister?

35

Many examples were generated but the
model kept responding 35 as the answer

Self Consistency

When I was 6 my sister was half my age. Now I'm 70 how old is my sister?

35

Many examples were generated but the model kept responding 35 as the answer

Is this still an issue? On OpenAI 3.5? On OpenAI 4o?

CODE OF PROMPT ENGINEERING WITH OPENAI API

LAB 3C

- Have the bot:
 - Get a user's question
 - Then (the bot) answers the question
 - Then the bot should ask "DO YOU WANT MORE INFO".
- If the user says 'yes',
 - The bot should then provide more information
 - The bot then re-asks "DO YOU WANT MORE INFO"
- If the user says no
 - bot asks the user what question can the bot help with
- Have the bot give answers such that someone who is 10 years old can understand the chatbot's answer.
- How does it do with "what is basefull?"