

Prompting Guide

Table of Contents

- [Introduction](#)
 - [Techniques](#)
 - [Good Prompting](#)
 - [Models and Risks](#)
 - [License](#)
-

Introduction

- What is Prompt Engineering?
-

Techniques

Zero-Shot prompting: Zero-shot prompting is a method in artificial intelligence where a language model is given a task or question without any specific training examples. Instead, it uses its pre-existing knowledge and the provided instructions to generate a response or perform the task. It's a way to make AI models more versatile and adaptable to different tasks without extensive training.

Zero-shot prompting is like asking a very smart friend a question without giving them any background information or examples. They rely on their general knowledge and the way you phrase the question to give you an answer. It's a way for AI to answer questions or do tasks it hasn't been specifically trained for. For example, you might ask, "What are the main ingredients in pizza?" and the AI can answer without having seen a pizza recipe before.

Few-Shot Prompting:

Few-shot prompting is a technique used with large-language models to improve their performance on complex tasks. While these models can do well without any prior examples (zero-shot), they struggle with more complicated tasks. Few-shot prompting involves giving the model a few examples (demonstrations) in the prompt to help it learn and generate better responses. This method becomes effective when the models are sufficiently large.

For example, imagine we want the model to use a new word in a sentence, like "whatpu" or "farduddle." By providing a single example sentence using these words in the prompt, the model learns how to use them correctly. This process is called "1-shot" prompting. If the task is harder, we can try with more examples (e.g., 3-shot, 5-shot, 10-shot) to further improve the model's performance.

Prompt:

```
A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses
```

the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.
To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

Output:

When we won the game, we all started to farduddle in celebration.

Chain-of-Thought Prompting

Chain-of-thought (CoT) prompting, introduced in Wei et al. (2022), allows for complex reasoning by breaking tasks into intermediate steps. When combined with few-shot prompting, it improves results on intricate tasks requiring reasoning.

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.

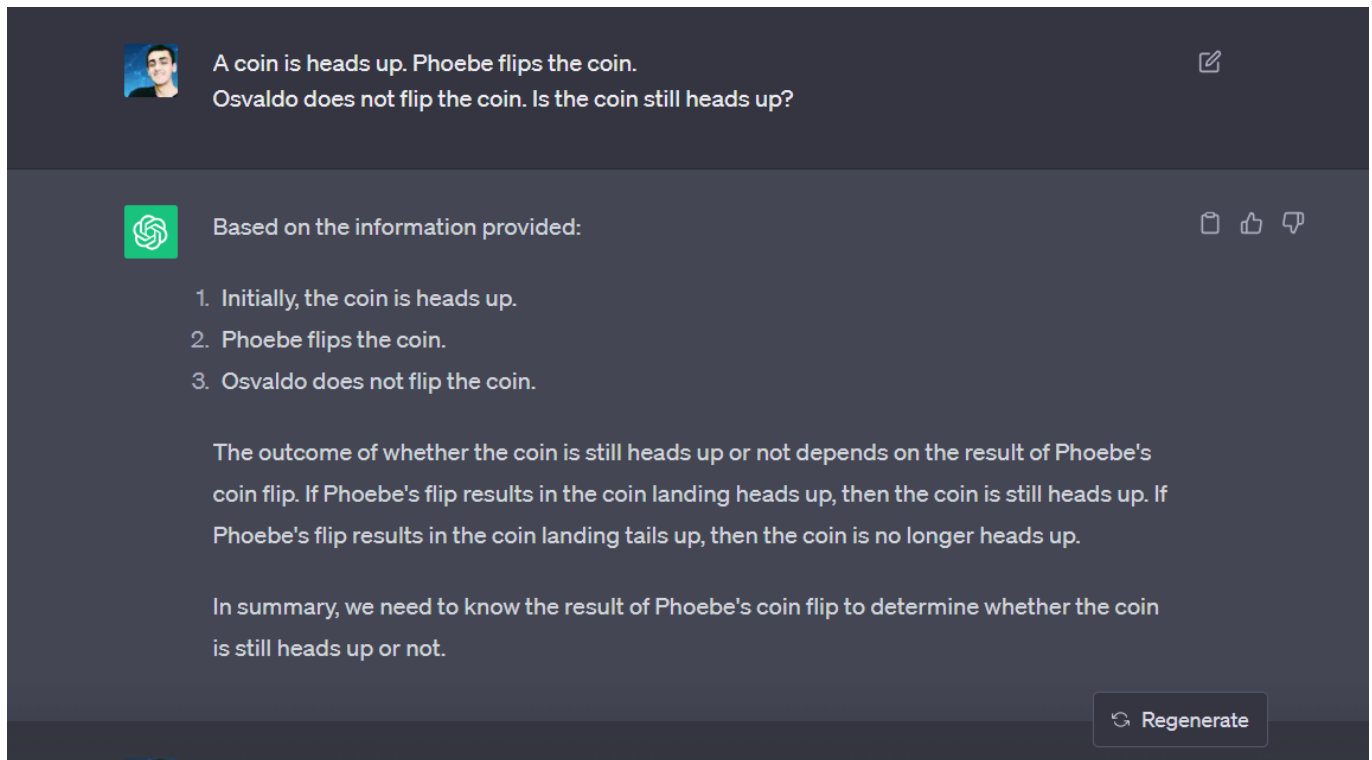
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

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. ✅

In the provided example, the prompt presents groups of numbers and apples. By demonstrating the reasoning process in the answer (A), the model effectively solves the task. It involves breaking down complex reasoning tasks into a series of intermediate steps or thought processes. This helps the model understand and reason through the problem step by step. It's important to note that this capability emerges in large language models, as stated by the authors.

ChatGPT answers the question with purely Chain of Thought



Good Prompting

Use Positive Instructions, Not Negative Ones

This guide emphasizes the importance of providing positive instructions ("do") rather than negative ones ("don't") in your project. Positive instructions tend to be more specific and effective. Instead of telling the model what not to do, it's usually better to specify exactly what we want it to do, assuming we know what that is.

For example, consider the situation where we want to ensure that the Language Model (LM) doesn't generate overly lengthy titles, which it tends to do. Instead of saying

Don't make the titles too long

It's more effective to be specific:

Each title should have a length between two and five words.

By using positive instructions and specifying what you want to achieve, you can provide clearer guidance to the model and improve the quality of your project.

When working with your project's models or AI, remember to:

- Clearly define the desired behavior.
- Provide explicit instructions on what the model should do.
- Avoid vague or negative commands that can lead to unintended results.

Following these guidelines can help you achieve better results and ensure your project's success.

Structure your prompt in a meaningful way

When working with Language Models (LLMs), it's essential to structure your prompts in a meaningful way. Just as elements like quotation marks, bullet points, and line breaks make it easier for humans to understand text, they also benefit LLMs. Let's apply this insight to our previous example and provide a well-structured prompt:

Prompt:

Generate a list of ten titles for my autobiography. The book is about my journey as an adventurer who has lived an unconventional life, meeting many different personalities and finally finding peace in gardening. Each title should be between two and five words long. Examples of well-structured titles:

- "Long walk to freedom"
- "Wishful drinking"
- "I know why the caged bird sings"

By structuring your prompts clearly, you can enhance the model's understanding and improve the quality of generated responses.

When designing prompts for your project:

- Use clear and concise language.
- Provide all necessary context.
- Consider formatting elements that aid comprehension.
- Test and iterate to optimize prompt effectiveness.

Following these guidelines will help you achieve better results when working with Language Models.

Models and Risks				
Model	Release Date	Size (B)	Checkpoints	Description
Falcon LLM	May 2023	7, 40	Falcon-7B , Falcon-40B	Falcon LLM is a foundational large language model (LLM) with 40 billion parameters trained on one trillion tokens. TII has now released Falcon LLM – a 40B model.
PaLM 2	May 2023	-	-	A Language Model that has better multilingual and reasoning capabilities and is more compute-efficient than its predecessor PaLM.
Med-PaLM 2	May 2023	-	-	Towards Expert-Level Medical Question Answering with Large Language Models
Gorilla	May 2023	7	Gorilla	Gorilla: Large Language Model Connected with Massive APIs
RedPajama-INCITE	May 2023	3, 7	RedPajama-INCITE	A family of models including base, instruction-tuned & chat models.

Model	Release Date	Size (B)	Checkpoints	Description
LIMA	May 2023	65	-	A 65B parameter LLaMa language model fine-tuned with the standard supervised loss on only 1,000 carefully curated prompts and responses, without any reinforcement learning or human preference modeling.
Replit Code	May 2023	3	Replit Code	replit-code-v1-3b model is a 2.7B LLM trained on 20 languages from the Stack Dedup v1.2 dataset.
h2oGPT	May 2023	12	h2oGPT	h2oGPT is a large language model (LLM) fine-tuning framework and chatbot UI with document(s) question-answer capabilities.

Risks:

This section emphasizes the importance of considering the potential dangers and misuses of language models (LLMs), especially when applied in real-world applications. It discusses risks related to prompt injections and harmful behaviors. Additionally, it mentions the need to address issues such as generalizability, calibration, biases (including social biases), and factuality when working with LLMs.

Factuality:

When utilizing Language Models (LLMs), it's common for them to generate responses that appear coherent but can sometimes be entirely fictitious. Improving their accuracy and dependability hinges on the effective optimization of prompts. Here, we discuss some technical strategies to address this issue:

Ground Truth Integration

One way to bolster the reliability of LLM-generated responses is to incorporate verified information, such as excerpts from pertinent articles or Wikipedia entries, into the contextual information provided. This practice significantly diminishes the chances of the model generating fabricated or inaccurate responses.

Generate a response that incorporates verified information, such as an excerpt from a relevant article or Wikipedia entry.

Response

What are the symptoms of COVID-19?

The most common symptoms of COVID-19 are fever, cough, shortness of breath, and fatigue. Other symptoms may include muscle aches, headache, sore throat, congestion or runny nose, nausea or vomiting, and diarrhea.

According to the Centers for Disease Control and Prevention (CDC), the following are the most common symptoms of COVID-19:

- Fever or chills
- Cough
- Shortness of breath or difficulty breathing

- Fatigue
- Muscle aches
- Headache
- Sore throat
- Congestion or runny nose
- Nausea or vomiting
- Diarrhea
- Loss of taste or smell

Explanation

This response incorporates verified information by including an excerpt from the CDC website. This excerpt provides a comprehensive list of the most common symptoms of COVID-19, and it is also a credible source of information.

Ground truth integration is an important technique for improving the reliability of NLP models. By incorporating verified information into responses, NLP models can reduce the risk of fabricating information and provide more accurate and informative responses.

Response Calibration

Response calibration involves fine-tuning the LLM by adjusting its parameters to minimize response variability. Furthermore, you can instruct the model to explicitly acknowledge uncertainty by generating responses like "I don't know" when it lacks a precise answer. This enhances the model's ability to provide accurate and meaningful information.

Mixed Prompt Approach

By designing prompts that consist of a combination of questions and potential answers, encompassing both within and beyond the model's knowledge domain, you help the LLM better discern the boundaries of its expertise. This mixed prompt approach contributes to more contextually appropriate and factual responses.

License

This project is licensed under the MIT License

Copyright (c) 2023 Interact-Brands