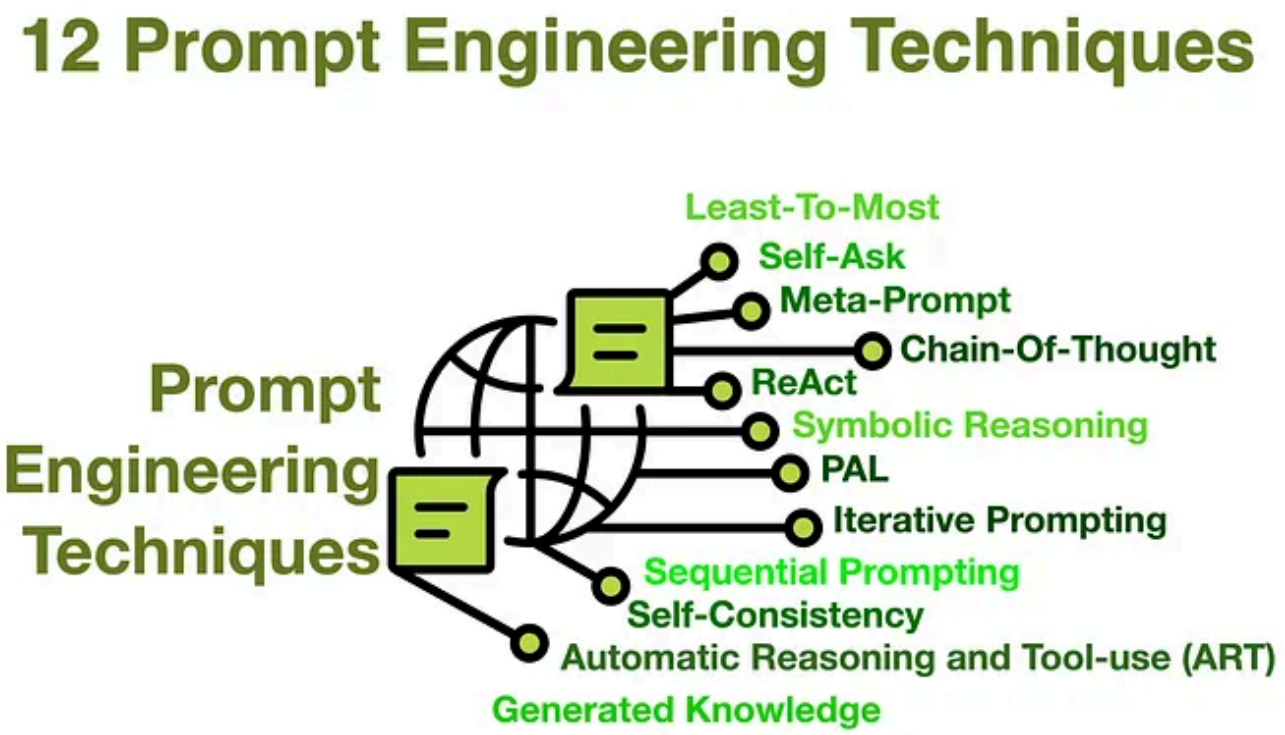
12 Prompt Engineering Techniques

Prompt Engineering can be described as an art form, creating input requests for Large Language Models (LLMs) that will lead to an envisaged output. Here are twelve different techniques in crafting a single or a sequence of prompts.

Author: Cobus Greyling | 7 min read | Aug 2, 2023
URL: https://cobusgreyling.medium.com/12-prompt-engineering-techniques-644481c857aa

---



Least-To-Most Prompting

The process of ***inference*** is reaching a conclusion based on evidence and reasoning. And in turn reasoning can be engendered with LLMs by providing the LLM with a few examples on how to reason and use evidence.

Hence a novel prompting strategy was developed, named *least-to-most prompting.* This method is underpinned by the following strategy:

1. Decompose a complex problem into a series of simpler sub-problems.
2. And subsequently solving for each of these sub-questions.

Solving each subproblem is facilitated by the answers to previously solved subproblems.

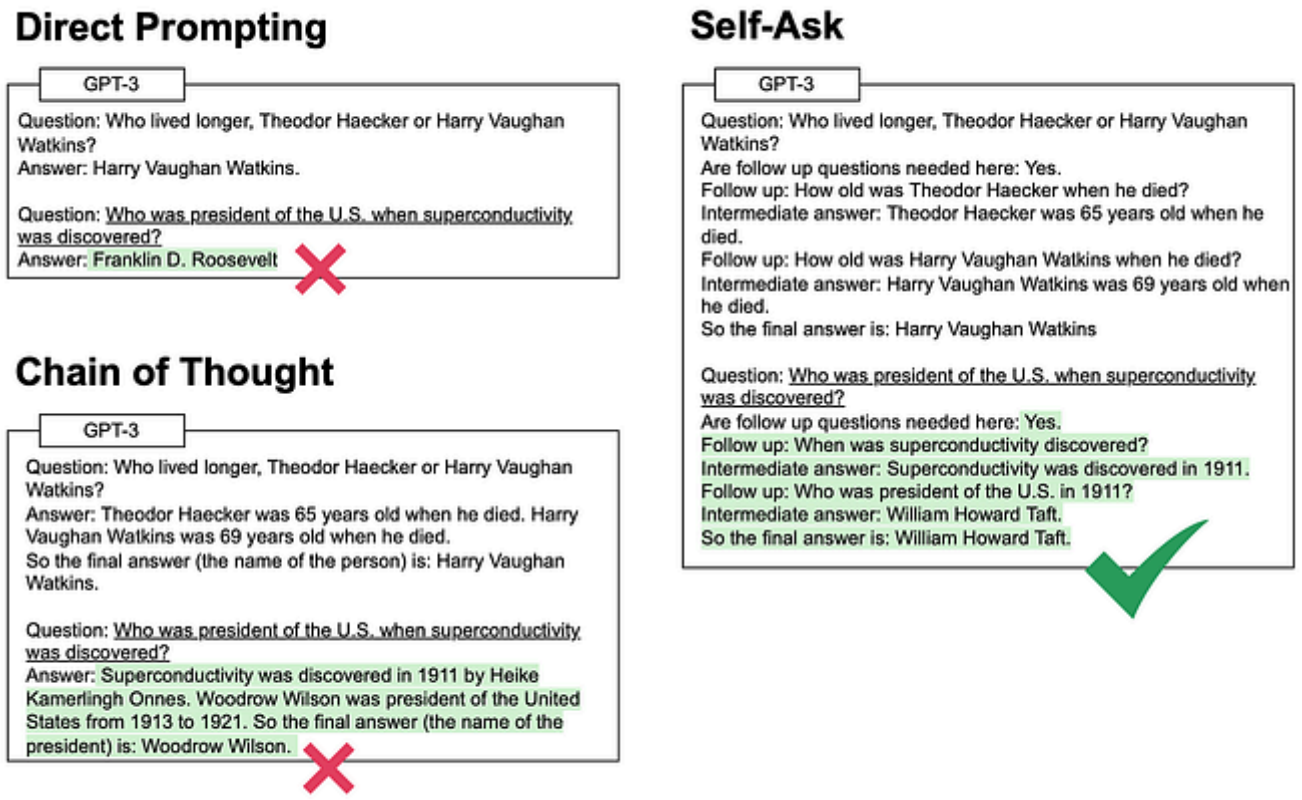
Hence least to most prompting is a technique of using a progressive sequence of prompts to reach a final conclusion.

Self-Ask Prompting

Considering the image below, it is evident that Self-Ask Prompting is a progression from Direct and Chain-Of-Thought prompting.

The interesting thing about self-ask prompting is that the LLM reasoning is shown explicitly and the LLM also decomposes the question into smaller follow-up questions.

The LLM knows when the final answer is reached and can move from follow up intermediate answers to a final answer.



Source

Meta-Prompting

The key principle underpinning Meta-Prompting is to cause the agent to reflect on its own performance and amend its own instructions accordingly.

While simultaneously using one overarching meta-prompt.

Chain-Of-Thought Prompting

Intuitively we as humans break a larger task or problem into sub-tasks, and then we chain these sub-tasks together. Using the output of one sub-task as the input for the next sub-task.

By using chain-of-thought prompting within the OpenAI Playground, a method wherein specific examples of chain of thought are provided as guidance, it is possible to showcase how large language models can develop sophisticated reasoning capabilities.

Research has shown that sufficiently large language models can enable the emergence of reasoning abilities when prompted in this way.

ReAct

With humans the tight synergy between *reasoning* & *acting* allows for humans to learn new tasks quickly and perform robust reasoning and decision making. We can perform this even when unforeseen circumstances, information or uncertainties are faced.

LLMs have demonstrated impressive results in chain-of-thought reasoning(CoT) and prompting, and acting (generation of action plans).

The idea of ReAct is to combine reasoning and taking action.

Reasoning enables the model to induce, track and update action plans, while actions allow for gathering additional information from external sources.

Combining these to ideas are named ReAct, and it was applied to a diverse set of language and decision making tasks to demonstrate its effectiveness over state-of-the-art baselines in addition to improved human interpretability and trustworthiness.

Symbolic Reasoning & PAL

LLMs should not only be able to perform mathematical reasoning, but also symbolic reasoning which involves reasoning pertaining to colours and object types.

Consider the following question:

```
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?
```

The LLM should convert the input into a dictionary with entities and values according to their quantities, while filtering out non-vegetable entities.

Get Cobus Greyling's stories in your inbox

Finally, the answer is the sum of the dictionary values, below the PAL output from the LLM:

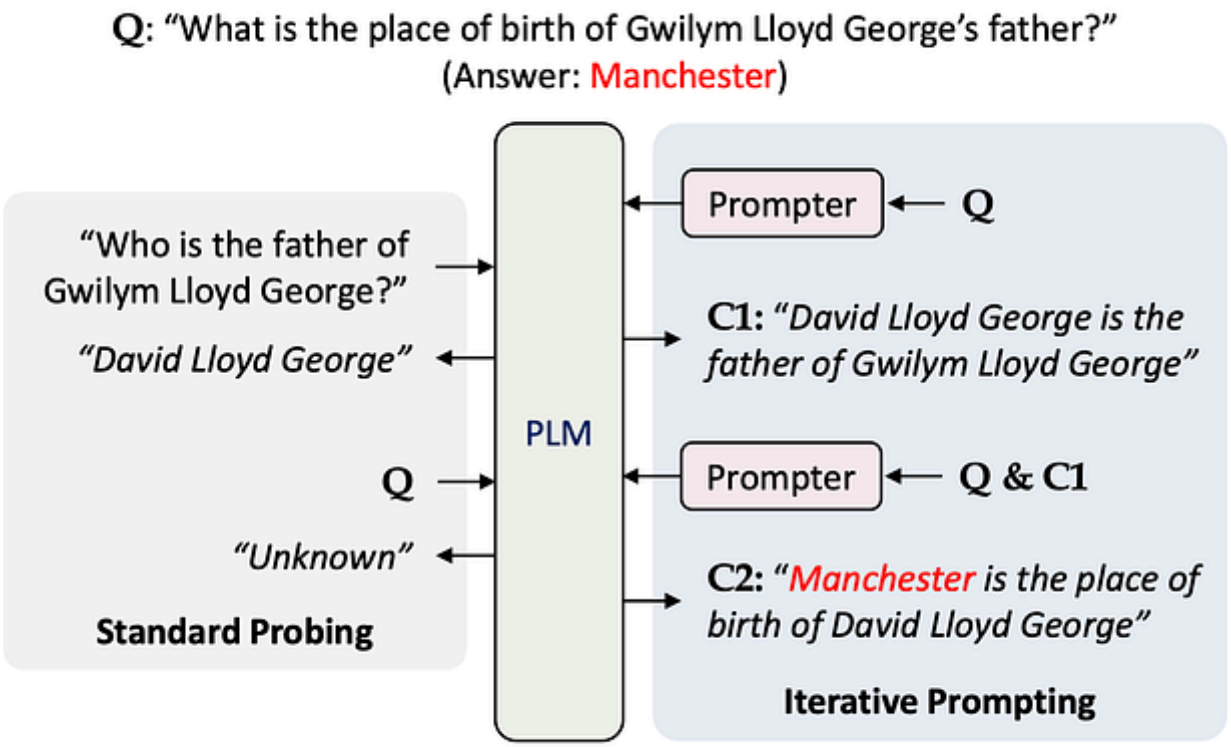
```
# note: I'm not counting the chair, tables, or fridges
  vegetables_to_count = {
      'potato': 2,
      'cauliflower': 1,
      'lettuce head': 1,
      'cabbage': 1,
      'onion': 2
  }
  answer = sum(vegetables_to_count.values())
```

Iterative Prompting

Of late the focus has shifted from LLM fine-tuning to enhanced prompt engineering. Ensuring that prompts are contextual, contains few-shot training examples and conversation history.

> Ensure the prompt holds contextual information via an iterative process.

Iterative prompting should establish a contextual chain-of-thought, which negates the generation of irrelevant facts and hallucination. Interactive context-aware & contextual prompting.



Source

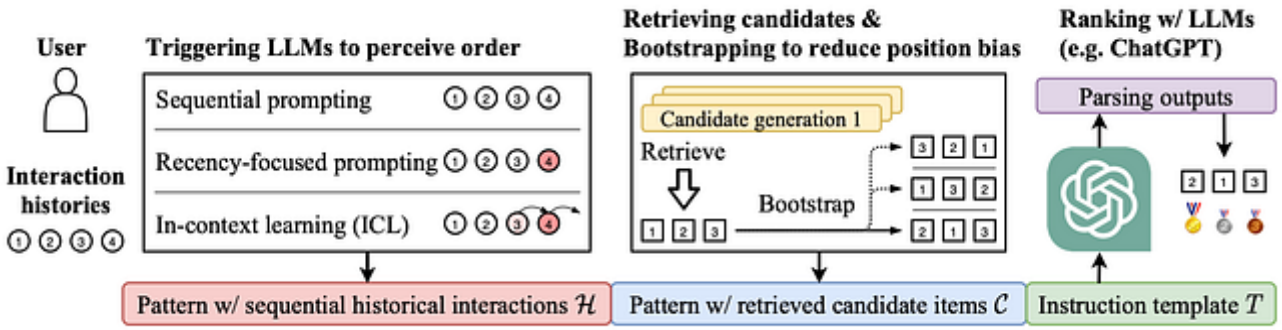
Considering the image above, at C1 and C2 knowledge is important for accurately answering the question. The approach of iterative prompting contains strong elements of chain-of-thought prompting and pipelines.

Sequential Prompting

Sequential prompting considers the possibility of building a capable recommender with LLMs. Usually recommender systems are developed in a pipeline architecture, consisting of multi-stage candidate generation (*retrieving more relevant items*) and ranking (*ranking relevant items at a higher position*) procedures.

Sequential Prompting focuses on the ranking stage of recommender systems, since LLMs are more expensive to run on a large-scale candidate set.

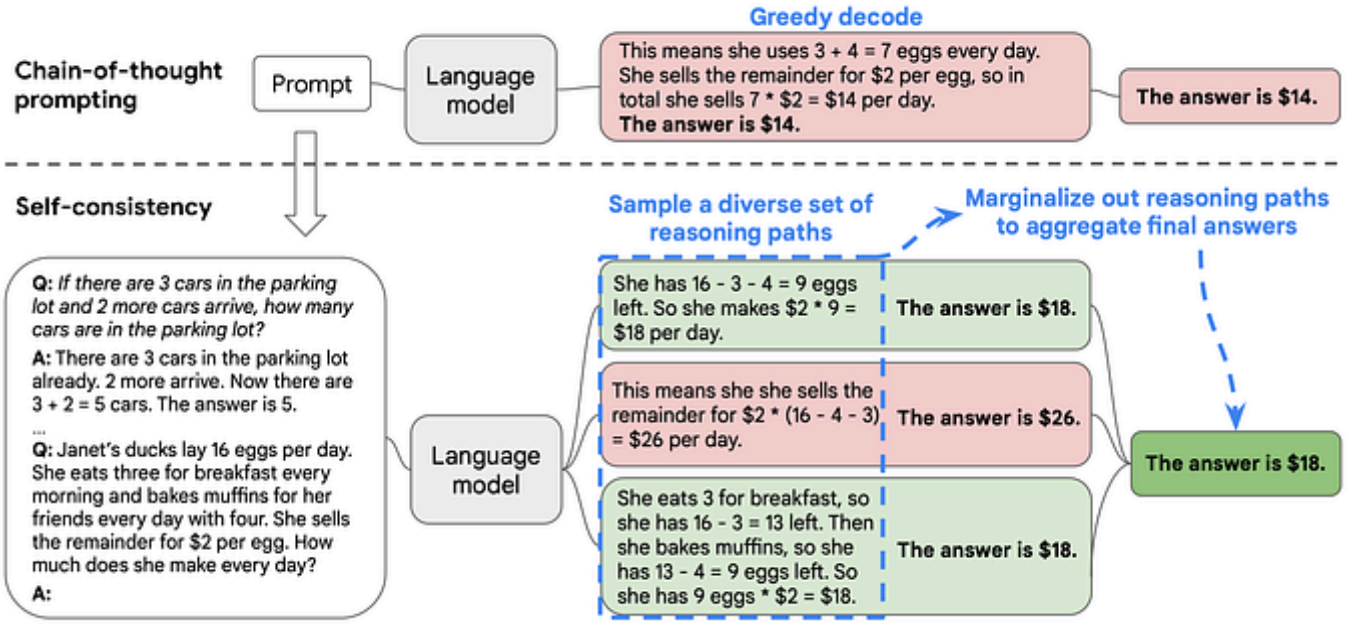
The ranking performance is sensitive to the retrieved top-ranked candidate items, which is more suitable to examine the subtle differences in the recommendation abilities of LLMs.



Source

Self-Consistency

With Chain-Of-Thought reasoning a path of thought is generated which then in turn is followed. And on the contrary, self-consistency leverages the intuition that a complex reasoning problem typically admits multiple different ways of thinking leading to its unique correct answer.



Source

The self-consistency method is constituted by three steps:

1. prompt the LLM to generate a chain of thought (CoT) reasoning part.
2. Generate a diverse set of reasoning paths.
3. Select the most consistent output for the final answer.

The approach followed by the self-consistency method can introduce increased overhead; especially if the steps of each CoT involves the calling of external tools and APIs. The overhead will manifest in the form of additional cost and time to complete the round-trips.

Automatic Reasoning & Tool Use (ART)

It has been illustrated that Chain Of Thought prompting elicits complex and sequential reasoning from LLMs. It has also been proven that for each step external tools can be used to improve the specific node's generated output.

The premise of developing these methodologies is to leverage a *frozen* large language model (LLM). Hence augmenting a previously trained model which is time-stamped.

Automatic Reasoning and Tool-use (ART) is a framework which also leverages frozen models to generate intermediate reasoning steps as a program.

The approach of ART strongly reminds of the principle of Agents, that of decomposing a problem, and making use of tools for each decomposed step.

> With ART, a frozen LLM decomposes instances of a new task into multiple steps using external tools whenever appropriate.

ART is a fine-tuning free approach to automate multi-step reasoning and automatic tool selection and use.

Generated Knowledge

The principle of generated knowledge is that knowledge can be integrated at inference time. Showing that reference knowledge can be used instead of model fine tuning.

Tests were performed across multiple datasets, common-sense reasoning, etc.

The principle of generated knowledge is supported by developments like RAG, pipelines, and more.

Other mentions by Author

- Least To Most Prompting - Least To Most Prompting for Large Language Models (LLMs) enables the LLM to handle complex reasoning.
- Self-Ask Prompting - Self-Ask Prompting is a progression from Chain Of Thought Prompting. Below are a few practical examples and an…
- Meta-Prompt - Meta-Prompt for building self-improving agents with a single universal prompt.
- Chain-Of-Thought Prompting & LLM Reasoning - When we as humans are faced with a complicated reasoning task, such as a multi-step math word problem, we segment our…
- ReAct: Synergy Between Reasoning & Acting In LLMs - An element of human intelligence is the ability to seamlessly combine task-oriented actions with verbal or inner…
- Symbolic Reasoning & PAL: Program-Aided Large Language Models - LLMs should not only be able to perform mathematical reasoning, but also symbolic reasoning which involves reasoning…