AN ADVANCED PERSPECTIVE

PROMPT ENGINEERING SERIES

online

INNOVATION LAB

INNOVATION LAB

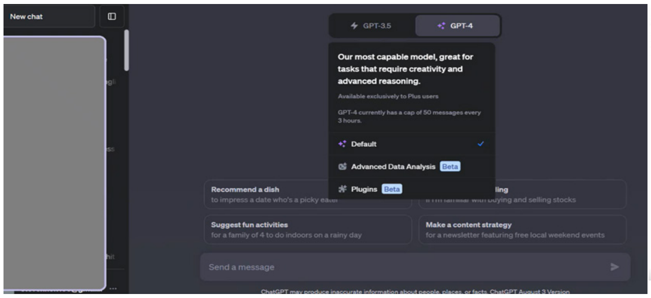
UNDER THE HOOD: ADVANCED PROMPTS.

AI being so new it is important and exciting to see interest in learning how to be utilize this tool. Maybe this has even prompted you to explore the topic more and better understand the technology.

Continue reading to see how your organization can utilize advanced prompt engineering!

online

CHAT GPT VS. OPENAI PLAYGROUND



[redacted]

Our most capable model, great for tasks that require creativity and advanced reasoning.

Available exclusively to Plus users

GPT-4 currently has a cap of 50 messages every 3 hours.

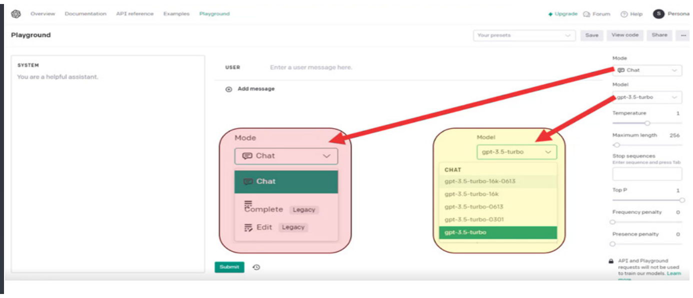
Recommend a dish to impress a date who's a picky eater

…ling …if I'm familiar with buying and selling stocks

Suggest fun activities for a family of 4 to do indoors on a rainy day

Make a content strategy for a newsletter featuring free local weekend events

ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT August 3 Version



Playground

SYSTEM You are a helpful assistant.

USER Enter a user message here.

Mode: Chat

Model: gpt-3.5-turbo

Temperature 1

Maximum length 256

Stop sequences: Enter sequence and press Tab

Top P 1

Frequency penalty 0

Presence penalty 0

API and Playground requests will not be used to train our models. Learn more

Chat-GPT is designed specifically for chat-based conversational AI. It allows users to create dynamic conversations by providing a series of messages as input, each with a user or system role and corresponding content. This format enables context preservation and makes it easier to have back-and-forth interactions with the model. Chat-GPT's primary focus is on generating rich, multi-turn conversations and implementing interactive AI systems.

On the other hand, the OpenAI Playground offers a broader experience for developers and users to test and explore GPT-3's capabilities. It provides a simple interface where users can input a prompt, and GPT-3 responds with a generated completion. The Playground also includes features like "temperature" control to adjust the randomness of the model's output and "max tokens" to limit response length. Moreover, the Playground allows users to fine-tune the model's behavior by providing certain instructions in the prompt.

While Chat-GPT is tailored for conversations and follows a structured input format, the Playground allows greater flexibility as a general-purpose interface for any types of prompts and completions. It caters to a wider range of use cases including text generation, question-answering, code generation, and more.

ADJUSTING VERBOSITY

LLMs such as ChatGPT often tend to be needlessly verbose.

This can lead to leaving you the user, with excessive text that you need to read and/or edit when using free or fixed-cost interfaces; or worse, drastically increasing the overall cost of utilizing these models though API calls, which tend to bill by the "Token" of I/O *(...think approx. 4 English characters).*

On the Input side of the I/O equation we can be diligent in our Prompt Engineering, being efficient when crafting input or instruction prompts, and limiting the total amount of data that needs to be provided. When utilizing a UX like ChatGPT, this is more aout staying within the overall Token limit for each Request-Response pair. When utilizing these tools via billable APIs, this is about being thoughtfully cost-effcient with our inputs and instruction prompts.

On the Output side of the I/O equation, our options will depend on which tool and/or interface we are using. This could be as simple as asking ChatGPT to "Provide your answer in a brief point form list of all of the most relevant information."

CONTROLLING RANDOMNESS

Another way to think about "controlling randomness" when using LLMs like ChatGPT, is to think about it as choosing the balance between Coherence Vs. Creativity. This opens a whole new realm of possibilities when it comes to trying to fine-tune your prompts to accomplish specific tasks.

It *is not* currently possible to control the randomness of the outputs from the models when using the standard ChatGPT interface.

It *IS* currently possible to control the randomness of the outputs from the models when using the OpenAI Playground.

CHAINING PROMPTS

Chaining Prompts is an extremely useful technique to learn. It is especially useful when trying to build out large and complex outputs from these models.
The exact prompts you use will vary greatly depending on your specific use case. However, the general idea is to craft a series of prompts that will utilize the outputs from previous prompts as part of the inputs for the next prompts.This process can be linked in as long a chain as is necessary to complete the task.

CHAIN-OF-THOUGHT

In Prompt Engineering, Chain-of-Thought is a prompting Technique which is used to "get a peek under the hood", so to speak, when it comes to understanding HOW an LLM has come to the conclusion that it has to some question. This can be particularly useful when it comes to trying to evaluate the outputs of these models for both Validity and Correctness.

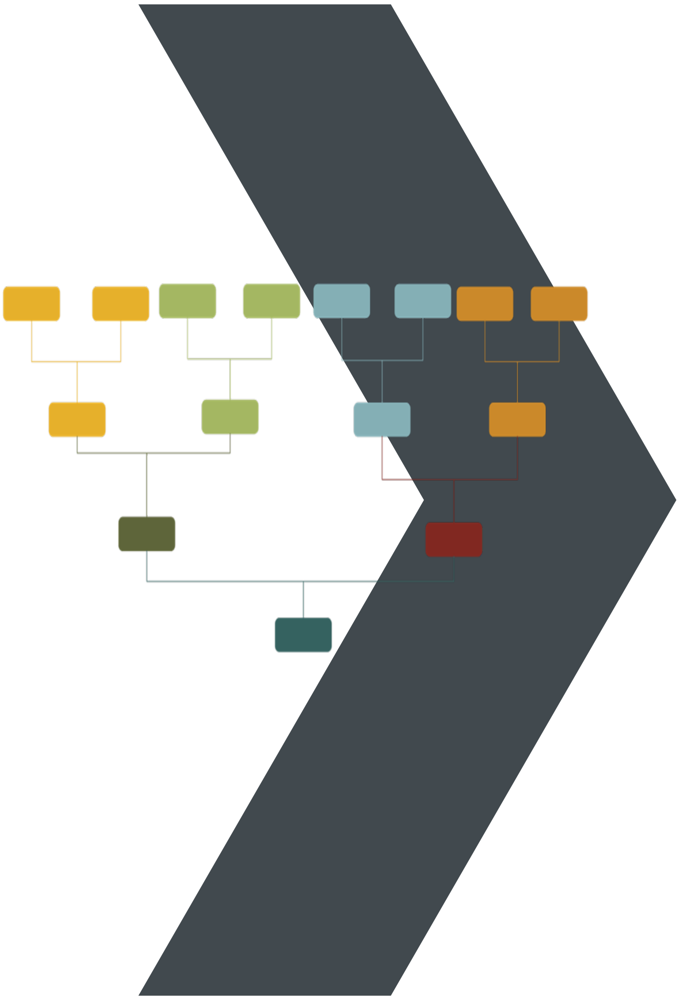
This is especially useful when trying to implement a "Process Reenforced Learning" pattern, verses a standard "Outcome Reenforced Learning" pattern. Or more simply put, if it's not ONLY important to you that the model comes to the correct conclusion; but rather it is ALSO important to you that the model uses valid rationale, logic and reasoning when coming to it's conclusion, then Chain-of-Thought is the technique for you!

Implementing Chain-of-Thought is as simple as including something like the following at the end of the existing question prompt:

"Think carefully and logically, explaining your answer step-by-step."

TREE OF THOUGHT

The primary goal of crafting Tree-of-Thought prompts is an attempt to enable these models to “self-correct” for errors in logic, as well as the ever notorious “hallucinations” that the models are known to be occasionally plagued by.



Imagine three different experts are answering this question. All experts will write down one step of their thinking, then share it with the group. Each expert will consider and evaluate all other experts thinking at this step, to evaluate both their own thoughts, as well as the other experts’ thoughts, for correctness and logical validity. Then all experts will go on to the next step, etc. If any expert realizes they’re wrong, at any point, then they will inform the other experts, and leave the discussion. The question is...”

Three experts with exceptional logical thinking skills are collaboratively answering a question using a tree of thoughts method. Each expert will share their thought process in detail, taking into account the previous thoughts of others and admitting any errors. They will iteratively refine and expand upon each other’s ideas, giving credit where it’s due. The process continues until a conclusive answer is found. The question is...”

WHAT IS THE ADVANCED DATA ANALYSIS TOOL?

Formerlly referred to as the Code Interpreter, this plug-in feature is a tool that processes and executes code written in various programming languages to provide insights and perform complex data analysis tasks.

TWO VERY HELPFUL PHRASES WHEN USING CODE INTERPRETER / ADVANCED DATA ANALYSIS [BETA]:

1. "Proceed without further questions."

2. "Output a downloadable file."

SYNTEHTIC DATA

"I need this imaginary data about an imaginary persons' finances. Please produce an annual income/expenses data table for this imaginary person. They should make 25% more than the average working Canadian for their age, which is 44. They work as a Project Manager in IT, and have over 15 years experience in this field. They work and live in Winnipeg, so their expenses and income should both be representative of what would be expected for the city of Winnipeg, Canada. Provide as much detail as possible to make the data seem as realistic as possible. Please proceed without further questions. Output a downloadable File."

UNIVERSAL JAILBREAKS

We, as Prompt Engineers, and users of LLMs like ChatGPT, need to be aware of the security risks when using or implementing these tools. We have previously discussed the need to protect any sensitive client information. This could include any variety of PPI or Private Corporate Data in relation to these clients.

What we ALSO need to be aware of, as these tools become more prolific, and more companies seek to develop customer/user facing tools that utilize these LLMs in the back-end… is that currently, these tools are EXTREMELY difficult to “secure” against malicious use. This includes the models being “coaxed” into behaving in ways that may exist outside the limits being imposed on them by the leading LLM Engineers.

This could be anything from stepping outside the sentiment bounds that may have been placed on them, such as always being courteous and polite with customers utilizing these services, to providing answers regarding subject matter that the models have been explicitly instructed to not provide answers about.

DON'T FORGET ABOUT SECURITY

As you start your prompt engineering journey, it's important to not forget about security. Remember these key points:

- PROTECTING CLIENT DATA IS PARAMOUNT.
- LLMS HAVE ACCESS TO SENSITIVE CLIENT INFORMATION.
- SECURITY BREACHES CAN LEAD TO SEVERE CONSEQUENCES.
- WE MUST PRIORITIZE PRIVACY AND CONFIDENTIALITY.

For more information about Prompt Engineering, contact Online today!

obsglobal.com/innovation-lab

Results. Guaranteed.