Chapter 2 – Chatbots and Prompt Engineering

2.1 ChatGPT vs Bing AI vs Google Bard(Gemini)

Developer & Technology :-

- **ChatGPT (OpenAI)** → Uses GPT-4 / GPT-4o models. Strong at natural conversation, coding help, reasoning, and creative writing.

- **Bing AI (Microsoft)** → Powered by OpenAI's GPT-4, but deeply integrated with Bing search. Provides real-time web answers + citations.

- **Google Bard / Gemini (Google)** → Initially Bard, now rebranded as Gemini. Uses Google's own Gemini 1.5 models, strong in reasoning + multimodal (text, image, code).

2.2 What is Prompt Engineering?

➡ **Prompts** are phrases or sentences that you can write to initiate a conversion with a chatbot.

➡ **Prompt Engineering** is the practice of designing and structuring the prompt given to an AI system like ChatGPT to guide it in generating accurate, useful, and context-appropriate responses.

🔹 **Example**

❌ *Bad Prompt:* "Explain AI."
✅ *Good Prompt:* "Explain Artificial Intelligence in simple terms with 3 real-life examples that a high school student can understand."

➡ Prompt Engineering in various sectors :-

1. Customer support in apps
2. Data analysis
3. Corporate training
4. Healthcare

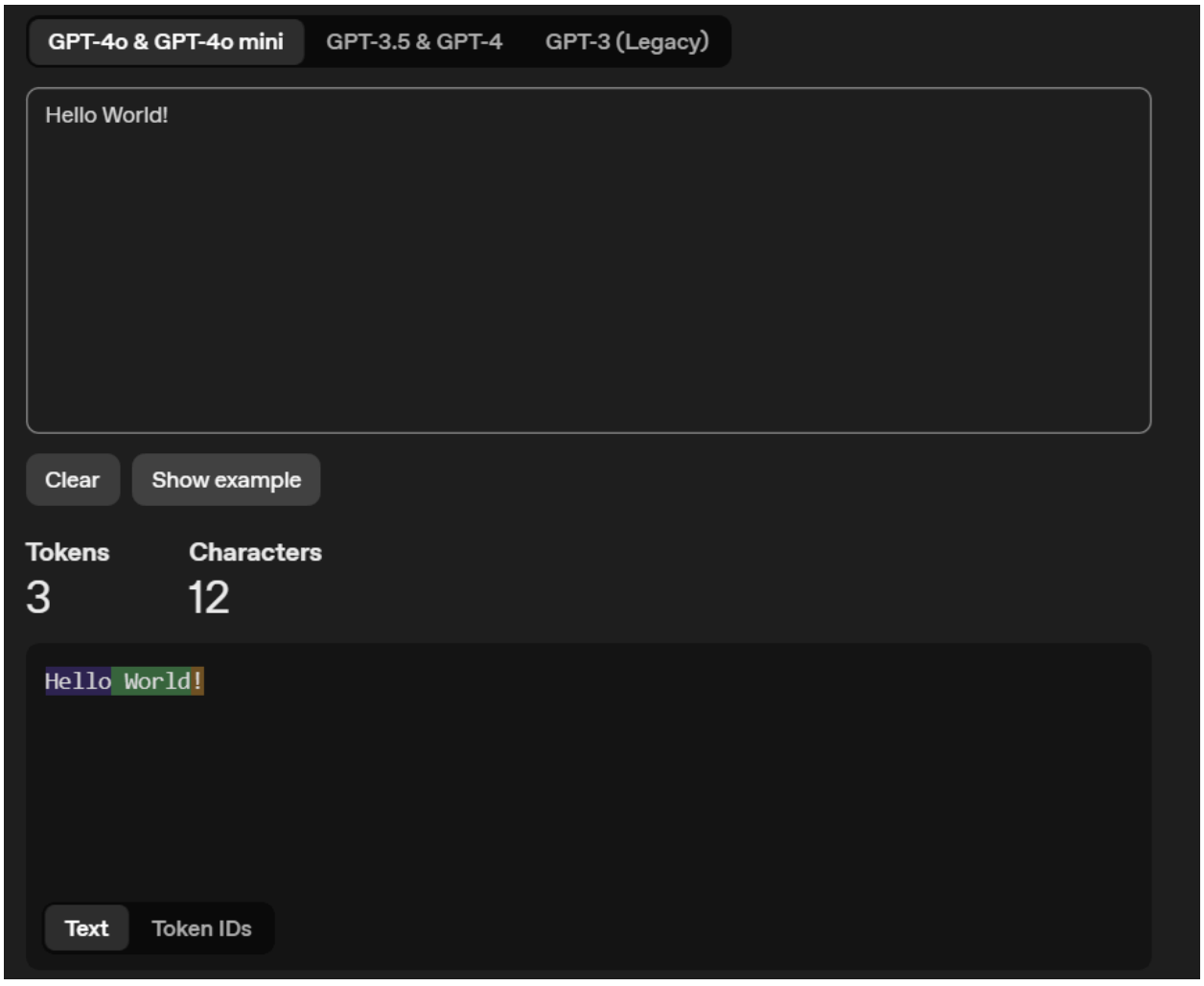
2.3 Understanding Tokens

➡ Till 4.0 version,
According to the original programming of ChatGPT, ChatGPT cannot recognize words. ChatGPT does not know how to recognize words. Because it has been taught to recognize tokens only, not words.

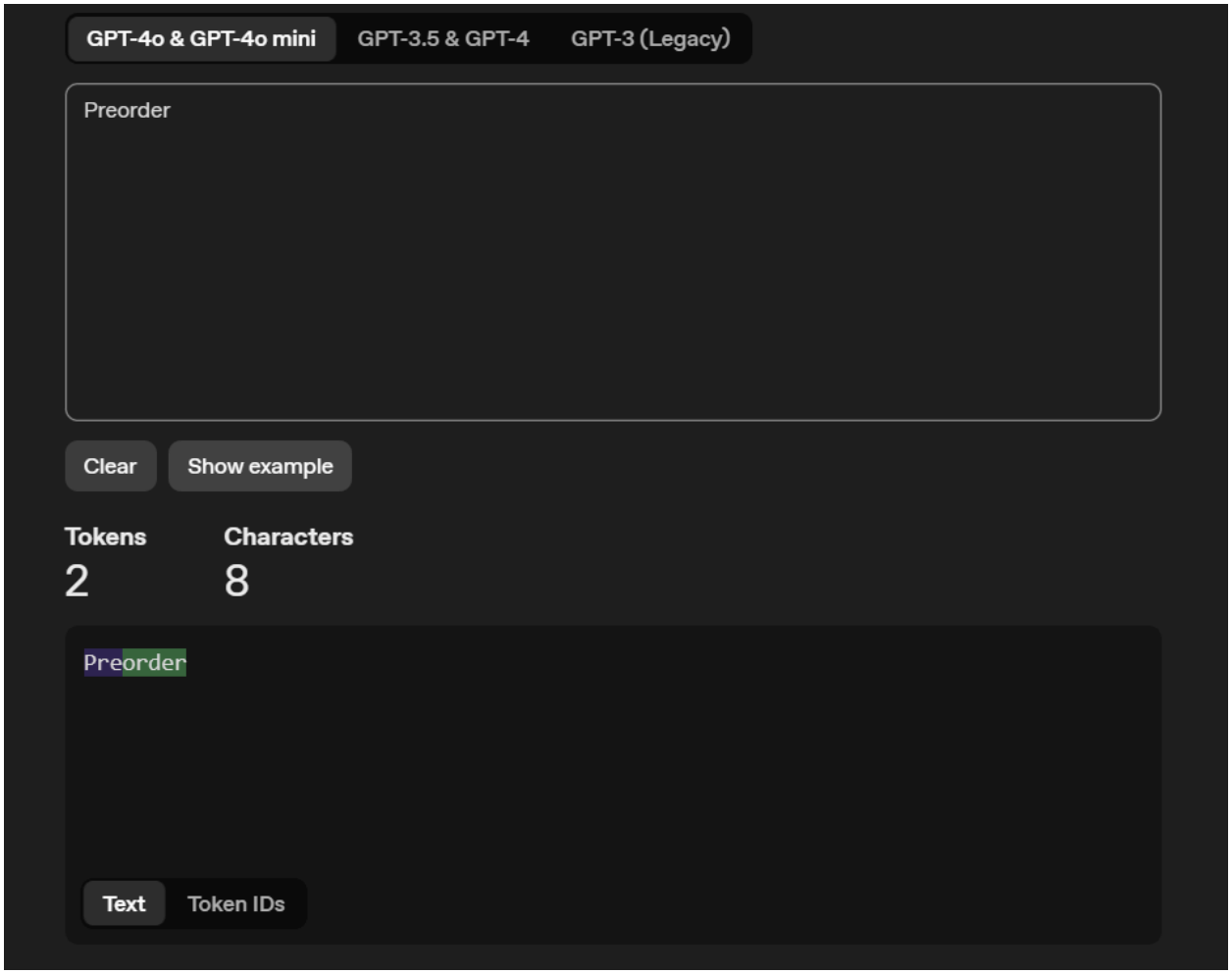
➡ **What are tokens?**
Tokens are individual units of information through which ChatGPT understands the information, takes our input and through which ChatGPT gives its output.
Tokens can be words, only one word can also be one token. But the difference is that sometimes one word can have multiple tokens.

➡ It means the RLHF method uses tokens not words.

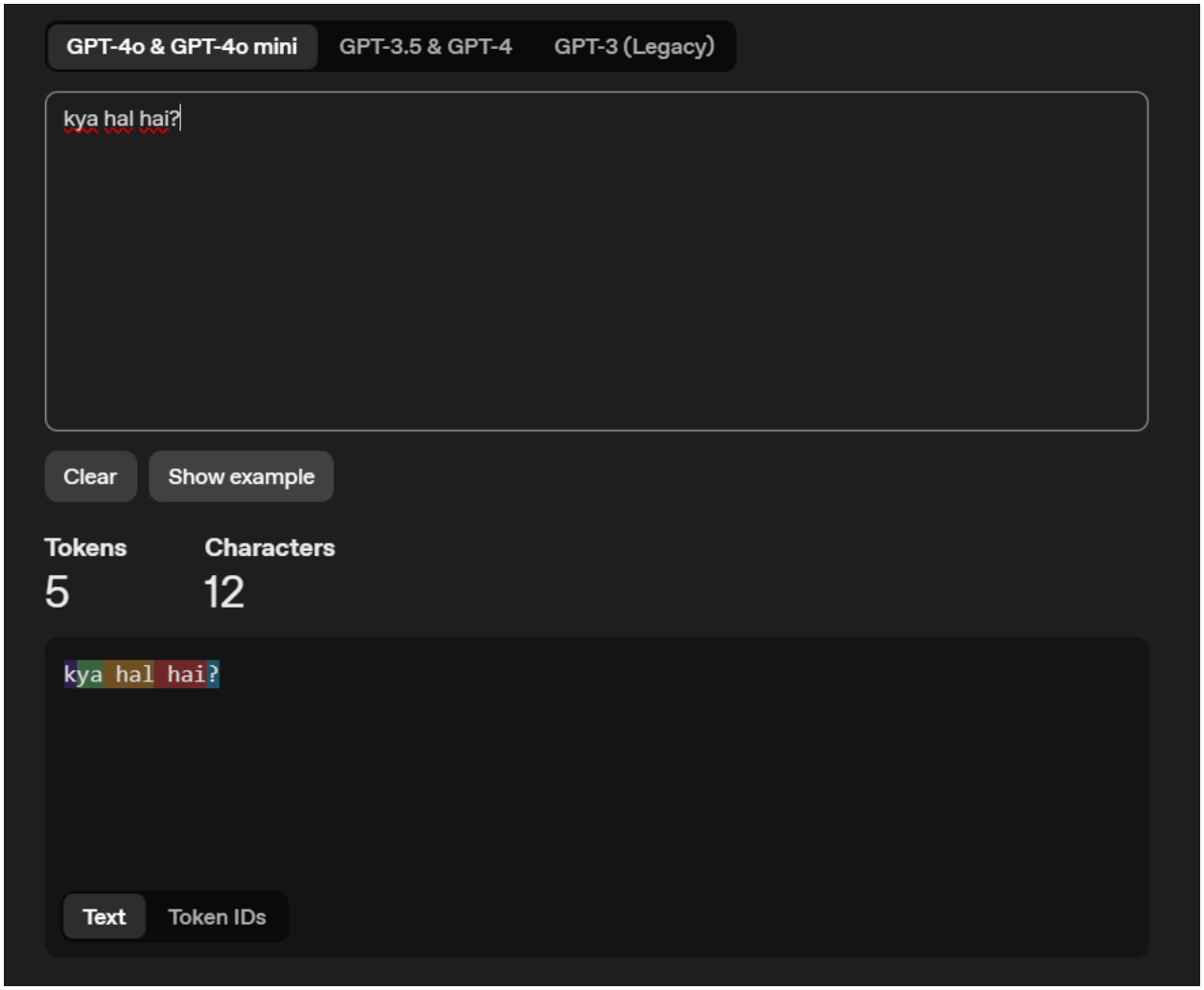
Tokenizer tool :- https://platform.openai.com/tokenizer



Prefix :-


GPT-4o & GPT-4o mini
GPT-3.5 & GPT-4
GPT-3 (Legacy)
Preorder
Clear
Show example
Tokens
2
Characters
8
Preorder
Text
Token IDs

Hindi words :-



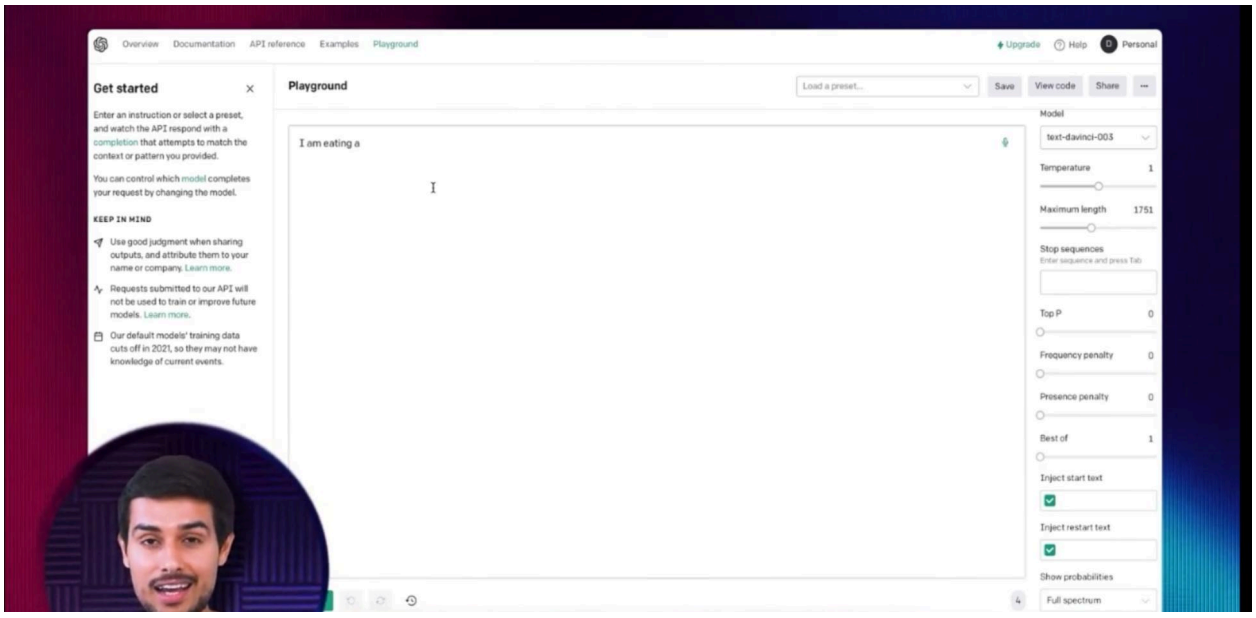
➡ The relation between words and tokens in the English language is something like this, that approximately ¾th of a word is a token.
100 tokens ~ 75 words
2048 tokens ~ 1500 words
➡ If we talk in different languages, then the token can vary a lot.
➡ ChatGPT 3.5 version,
Limitation :- 4096 tokens per prompt including question & answer.
➡ ChatGPT 4-32k version,
Limitation :- 32,768 tokens per prompt including question & answer.
(Approximately 24,000 english words)

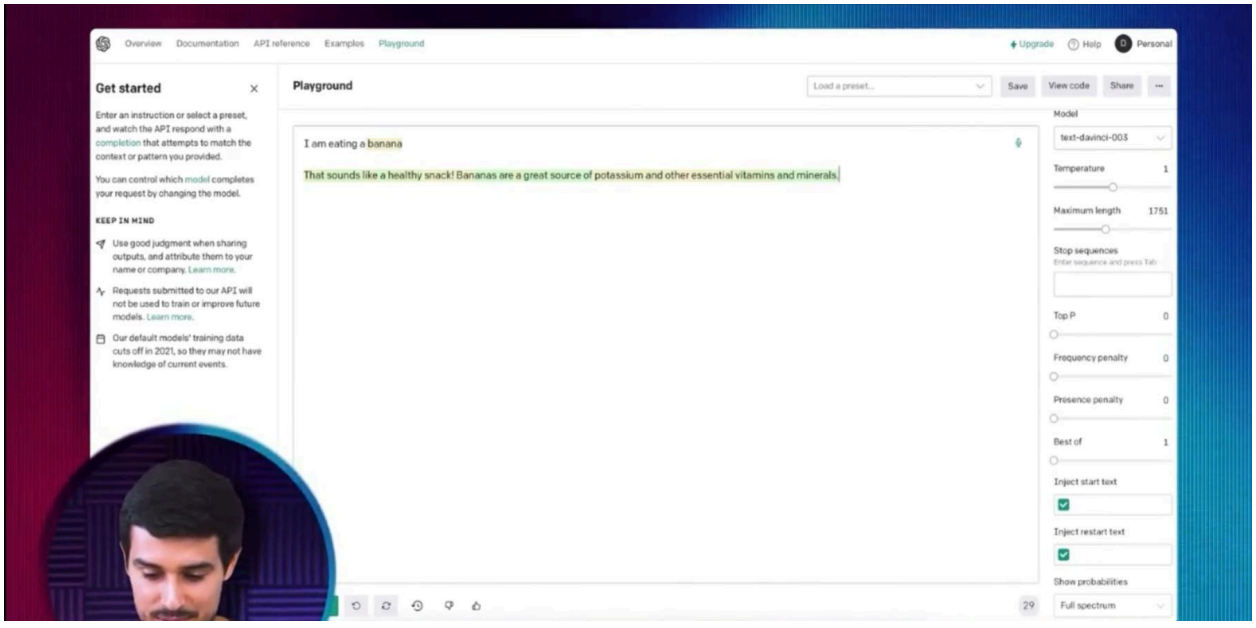
➡ **To know token limits of different AI models :-** https://platform.openai.com/docs/models

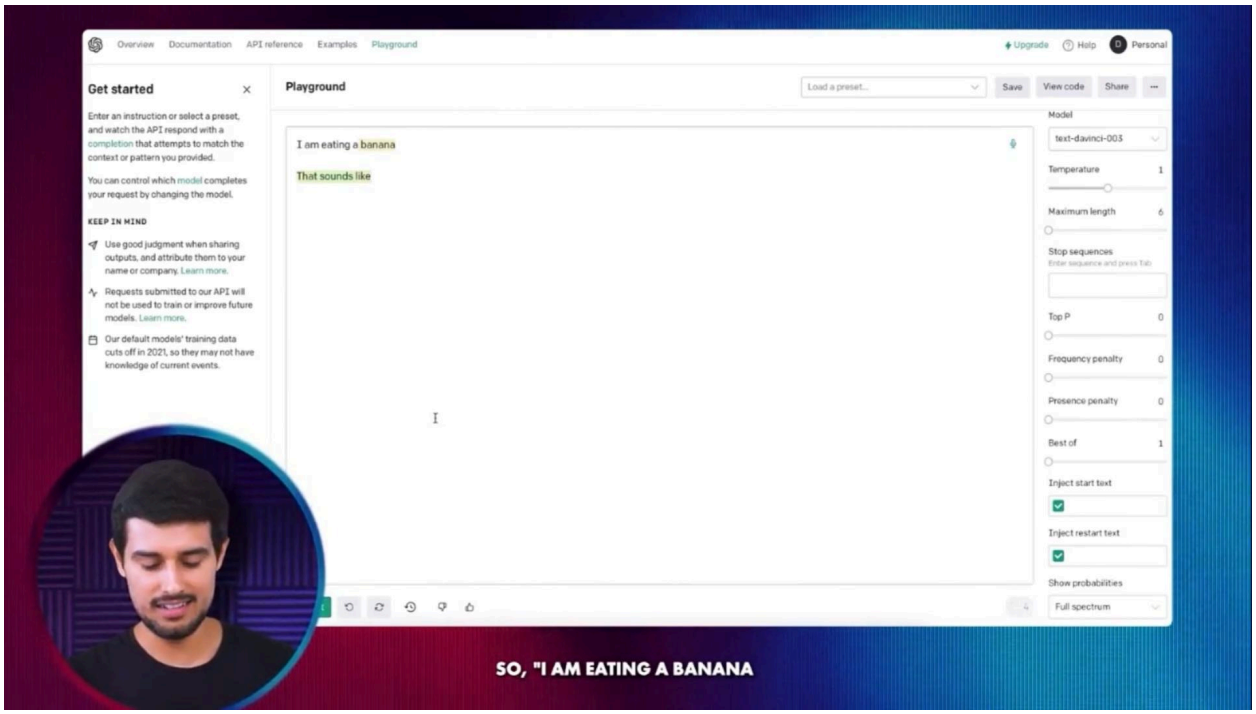
2.4 Temperature and Other Parameters

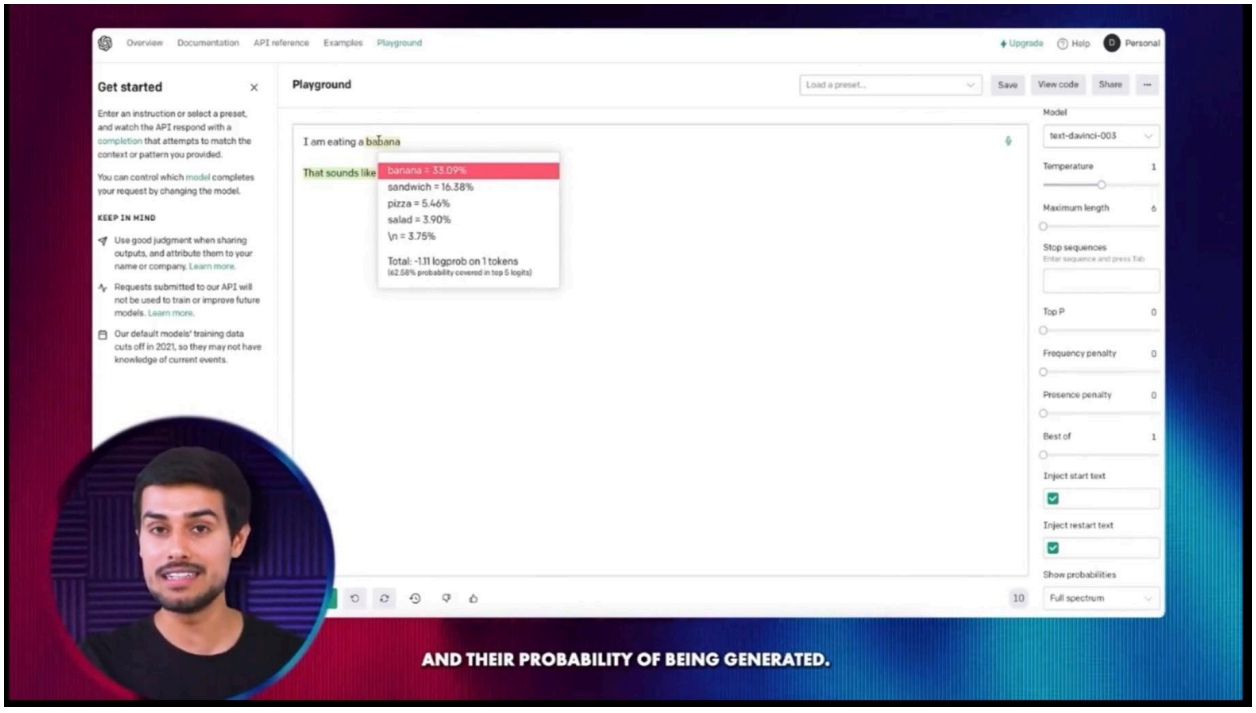
https://platform.openai.com/chat/edit

It will require payment to use.









➡ **Temperature :-**

How randomized your output will be.

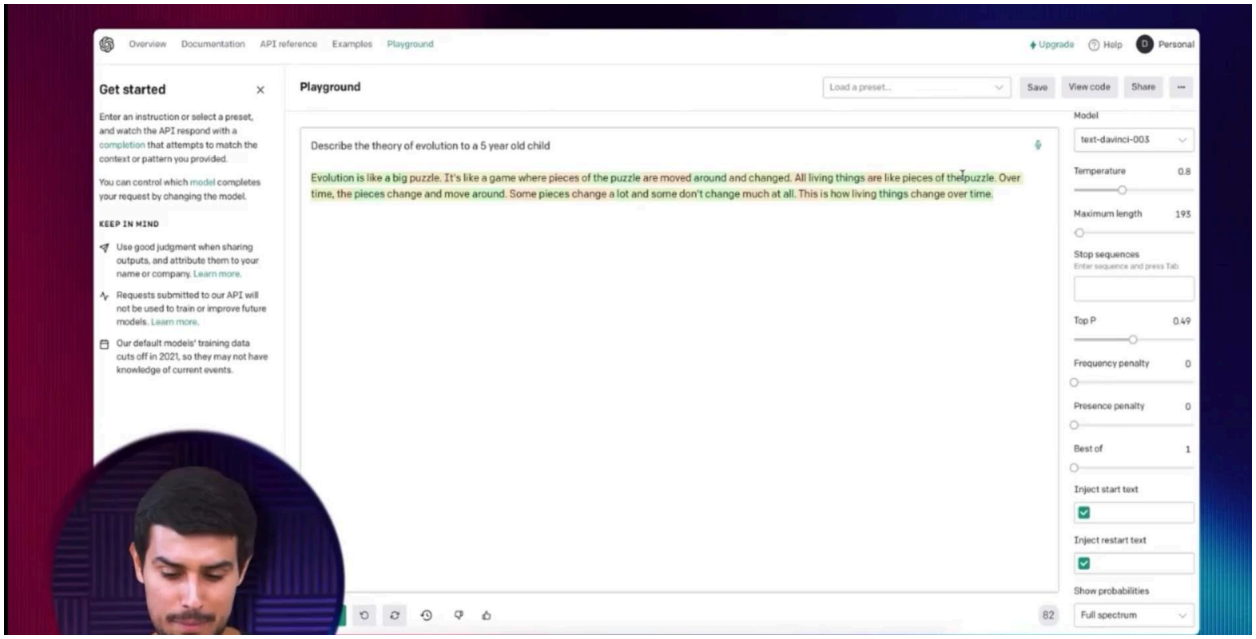
If hover over the temperature, it says that it “controls randomness : Lowering results in less random completions. As the temperature approaches zero, the model will become deterministic and repetitive.”
Temperature is a metaphorical parameter.
👉 Controls **how random or deterministic** the choice is.

➡ **Maximum length :-**
To control the length of the token that is going to be the output.

➡ **Top P :-**
P stands for Probability.
Top P also controls the randomness in a way but from a different perspective.
If I increase P to the maximum then random results will be generated once again.
But if it is reduced completely, then a definite,direct result will be produced, which has the highest probability.
Keep the smallest set of words that together reach probability p.
Example: with **top-p = 0.9**, the model only samples from words that make up the top 90% likelihood.

👉 Controls **how many word options** are considered.

➡ **Frequency penalty & Presence penalty :-**



Frequency Penalty

- Reduces chance of repeating words proportionally to how often they were used.
- Example :
    - No penalty : I like pizza. I like pizza. I like pizza.
    - With penalty : I like pizza. I enjoy pasta. I prefer burgers.

Presence Penalty

- Reduces chance of repeating words once they have already appeared at all.
- Example :
    - No penalty : Cats are cute. Cats are fluffy. Cats like milk.
    - With penalty : Cats are cute. Dogs are loyal. Birds can sing.

2.5 Toolkit for Best Results

For complex calculations :- https://www.wolframalpha.com

Hallucination :-
It means imagining things.

Toolkit of Prompt Engineering basic set of rules :-

1. Make ChatGPT pretend to be an expert.
   Ex. Pretend you are a teacher/doctor, etc.
2. Define your objective as specifically as possible.
3. Ask your answers in a certain format.
4. Prompt by example.
5. Ask ChatGPT to ask you questions.
6. Refine your answers and counter questions.