Open AI Cheat Sheet 2.0

GPT-3 and ChatGPT APIs, Updated 3/7/23

THE AI EXCHANGE

Getting started

Visit platform.openai.com/playground and create an account if you don't already have one

You are charged by token usage, managed in Billing.

Modes

Completion

Traditionally called "GPT-3" and used to complete text that it's given.

Chat

Access the ChatGPT interface with Systems, Users and Assistants

Insert

Tell the model to generate text anywhere you have the tag [insert]

Edit

Give the model input, then a set of instructions, and it will edit based on your instructions.

Presets

Browse Open AI's prebuilt "Presets"

Create and save your own.

Even share a link with your team, and they can save your preset.

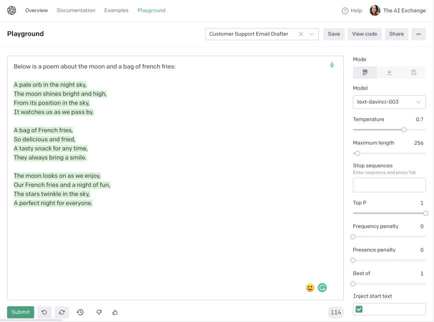
Perfect for Prompt Sharing!

Prompts

GPT-3 Prompts

Write your prompt in the big white text box. Click Submit. GPT-3 writes back in green.

Click ♻️ to regenerate



Playground

Customer Support Email Drafter

Below is a poem about the moon and a bag of french fries:

A pale orb in the night sky,
The moon shines bright and high,
From its position in the sky,
It watches us as we pass by.

A bag of French fries,
So delicious and fried,
A tasty snack for any time,
They always bring a smile.

The moon looks on as we enjoy,
Our French fries and a night of fun,
The stars twinkle in the sky,
A perfect night for everyone.

Mode

Model: text-davinci-003

Temperature: 0.7

Maximum length: 256

Stop sequences: Enter sequence and press Tab

Top P: 1

Frequency penalty: 0

Presence penalty: 0

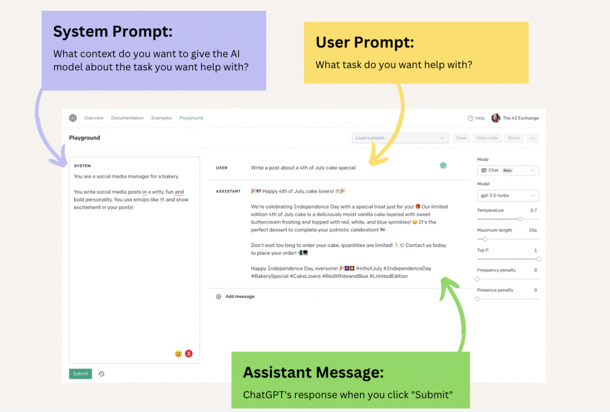
Best of: 1

Inject start text: ☑

114

ChatGPT Prompts

Set a System Message as "context" for the chat, and User Message as the task at hand. ChatGPT will respond as an Assistant.



System Prompt: What context do you want to give the AI model about the task you want help with?

User Prompt: What task do you want help with?

Playground

SYSTEM
You are a social media manager for a bakery.

You write social media posts in a witty, fun and bold personality. You use emojis like 🍰 and show excitement in your posts!

USER
Write a post about a 4th of July cake special.

ASSISTANT
🎉🇺🇸 Happy 4th of July, cake lovers! 🍰🎉

We're celebrating Independence Day with a special treat just for you! 🎁 Our limited edition 4th of July cake is a deliciously moist vanilla cake layered with sweet buttercream frosting and topped with red, white, and blue sprinkles! 😋 It's the perfect dessert to complete your patriotic celebration! 🇺🇸

Don't wait too long to order your cake, quantities are limited! 🏃💨 Contact us today to place your order! 📲💻

Happy Independence Day, everyone! 🎉🎆🎇 #4thofJuly #IndependenceDay #BakerySpecial #CakeLovers #RedWhiteandBlue #LimitedEdition

Mode: Chat Beta

Model: gpt-3.5-turbo

Temperature: 0.7

Maximum length: 256

Top P: 1

Frequency penalty: 0

Presence penalty: 0

Assistant Message: ChatGPT's response when you click "Submit"

GPT-3 vs ChatGPT

Price

GPT-3 (text-davinci-003) is $0.02 / 1000 tokens

ChatGPT (gpt-3.5-turbo) is $0.002 / 1000 tokens

Verbosity

ChatGPT is more verbose. Can be an issue with use cases such as generating code, where the model insists on explaining the code.

Hallucinations

ChatGPT is slightly better, including better at math. But still an issue.

Learning new tasks

GPT-3 is better at "few shot" learning like classification where examples are given.

ChatGPT is better at "zero shot" learning like classification where no examples are given.

Role playing

GPT-3 will role play with no issues.

ChatGPT sometimes reminds you that it's a language model.

Moderation and restraints

GPT-3 has little to no moderation. Developers are responsible for handling moderation, and can use Open AI's moderation endpoint.

ChatGPT cannot talk about inappropriate content or unethical behavior.

Created by The AI Exchange

Get our free newsletter at theaiexchange.beehiiv.com

Open AI Cheat Sheet 2.0

GPT-3 and ChatGPT APIs, Updated 3/7/23

THE AI EXCHANGE

ChatGPT system prompt

Below is the rough prompt used for chat.openai.com

> *You are ChatGPT, a large language model trained by OpenAI. Answer as concisely as possible. Knowledge cutoff: {knowledge_cutoff} Current date: {current_date}*

🚨 Open AI states that the model sometimes forgets System Messages. Add important information in the User Message

Customizing GPT

1. Write a great prompt
2. Give GPT a few example inputs and outputs

> *input => "What is a good dog name?"*
> *output => "Princess because humans spoil their pets."*
>
> *input => "What is a good cat name?"*
> *output =>*

3. Create a fine tuned model
4. Add context dynamically to your prompt via an outside data source using embeddings

Using via API

1. Go to https://platform.openai.com/account/api-keys.
2. Create a new secret key. Save this. You can't access again.
3. Go back to the playground. Click "View Code"
4. Use these parameters in an API wrapper. Or use the code directly in your application.

Important settings

Completion Models

GPT-3 models can understand and generate language. Today's most powerful is text-davinci-003.

Codex models can understand and generate code.

Fine-tuned models are created by your organization.

Temperature

Controls randomness in the way responses are generated. Higher values are more random.

Max length

How many tokens can be used in the response.

Token = word chunk, about 750 words per 1000 tokens

GPT-3 and ChatGPT have a 4096 token limit, including the prompt.

Stop sequences

Used to stop GPT-3/ChatGPT from continuing on.

Popular stop sequences:

- \n *for new lines*
- ## *for headings*
- Question: *for question and answer*

Other settings

Mode*

- Complete - generates text to complete the input prompt
- Insert - generates text where [insert] tag is used
- Edit - takes existing text and rewrites based on instructions

Show Probabilities*

View the probability that a token was generated. Helpful for understanding and debugging generations.

Start Text*

Injects starting text after completion is finished. Useful for chat bot conversations.

Top P

An alternative to temperature for controlling randomness. Top P = 0.25 is less random than TopP = 1

Frequency Penalty

Reduces redundancy in answer by making it less likely to generate the same token based on how frequent is shows up so far.

Presence Penalty

Another method to reduce redundancy that penalizes based on presence of the token, not frequency.

Best Of*

Generates multiple versions of the completion on Open AI's server and sends you the best one.

* Not currently available in API

Created by The AI Exchange

Get our free newsletter at theaiexchange.beehiiv.com

Open AI Models Cheat Sheet

GPT-3, ChatGPT (3.5), & GPT-4, Updated 3/15/23

THE AI EXCHANGE

	GPT-3 (text-davinci-003)	ChatGPT (gpt-3.5-turbo)	GPT-4 (gpt-4 & gpt-4-32k)
Format	Send input, get output	*Send chat messages, get output*	*Send chat or regular messages, get output*
Useful settings	*Temperature, Stop sequences, # of Completions*	*Temperature, Stop sequences, # of Completions*	*Not confirmed yet, but most likely the same*
Max tokens	*4,096 tokens*	*4,096 tokens*	*8,192 tokens & 32,768 tokens*
Verbosity	*Direct and specific*	*More verbose, because it was fine tuned for chat*	*Less verbose than ChatGPT, but more verbose than GPT-3*
Access to internet	*None*	*None*	*None*
Accuracy and hallucinations	*The lowest accuracy rating*	*Better than GPT-3 at math, but still not perfect*	*Significantly better at complex reasoning situations, math, and following instructions*
Classification	*Outperforms ChatGPT on few-shot classification [source]*	*Outperforms GPT-3 on zero-shot classification [source]*	*Outperforms ChatGPT on zero-shot and few-shot classification [source]*
Ability to code	*Yes, and will generate ready-to-use code*	*Yes, but will also explain the code which means the code needs to be parsed out*	*Yes, but will also explain the code which means the code needs to be parsed out*
Role playing	*Will role play without issue*	*Sometimes remembers that it's a large language model*	*Will adhere more closely to user instructions and system message (higher steerability)*
Moderation and restraints	*Very little moderation or restraints*	*Cannot talk about inappropriate content or unethical behavior.*	*Improved response to disallowed behavior by 83% from GPT-3.5 and improved responses to sensitive requests by 29%*
Cost	*$0.02/1k tokens*	*$0.002/1k tokens* *1/10th the cost of GPT-3*	*8k context: $0.03/1k prompt tokens, $0.06/1k completion tokens* *32k context: $0.06/1k prompt tokens, $0.12/1k completion tokens*
Speed	*Davinci takes an average of 1 minute to return response with 2048 tokens [source]*	*"10x faster" according to the internet so far*	*As of 3/15, GPT-4 has rate limits and delayed response time due to current capacity constraints*
Modalities	*Text inputs > text outputs*	*Text inputs > text outputs*	*Text & image inputs > text outputs*
Where to access	*Accessible via the Playground and API*	*Accessible via the Playground, API, and ChatGPT Free/Plus*	*As of 3/15, available via ChatGPT Plus and via API from waitlist*
Knowledge cutoff	*June 2021*	*September 2021*	*September 2021*

Created by The AI Exchange

Get our free newsletter at theaiexchange.beehiiv.com