



How to Use GPT-3.5 and GPT-4 with Python




Presenter: Qiao Jin, M.D. (@DrQiaoJin)

Yale



National Library of Medicine
National Center for Biotechnology Information

Three ways to use GPT-3.5 and GPT-4

		Pros	Cons
ChatGPT		Easy-to-use	Less flexible
OpenAI API		Flexible & Up-to-date	Not compliant to HIPAA
Azure OpenAI API		Compliant to HIPAA	Features not up-to-date

Model Overview

GPT-4o New

Our fastest and most affordable flagship model

- ✧ Text and image input, text output
- 📄 128k context length
- 📄 Input: \$5 | Output: \$15*

GPT-4 Turbo

Our previous high-intelligence model

- ✧ Text and image input, text output
- 📄 128k context length
- 📄 Input: \$10 | Output: \$30*

GPT-3.5 Turbo

Our fast, inexpensive model for simple tasks

- ✧ Text input, text output
- 📄 16k context length
- 📄 Input: \$0.50 | Output: \$1.50*

** prices per 1 million tokens*

ChatGPT Plus / month = \$20 = 2M tokens for GPT-4o API

Tokens

Language models read and write text in chunks called **tokens**

GPT-4o (coming soon) **GPT-3.5 & GPT-4**

The p53 protein, often referred to as the "guardian of the genome," is a crucial tumor suppressor involved in the regulation of the cell cycle and apoptosis. It plays a key role in preventing cancer formation by maintaining genomic stability.

Clear Show example

Tokens 57 Characters 281

The p53 protein, often referred to as the "guardian of the genome," is a crucial tumor suppressor involved in the regulation of the cell cycle and apoptosis. It plays a key role in preventing cancer formation by maintaining genomic stability. Here are the primary functions of p53:

Text Token IDs

1 token = 0.75 word (very rough estimation)

Chat Completions API Usage

python ▾



```
1 from openai import OpenAI
2 client = OpenAI()
3
4 response = client.chat.completions.create(
5     model="gpt-4-turbo",
6     messages=[
7         {"role": "system", "content": "You are a helpful assistant for answering biomedical questions."},
8         {"role": "user", "content": "What is the function of p53?"},
9     ]
10 )
```

Chat Completions Result

```
1 {
2   "choices": [
3     {
4       "finish_reason": "stop",
5       "index": 0,
6       "message": {
7         "content": "The p53 protein ...",
8         "role": "assistant"
9       },
10      "logprobs": null
11    }
12  ],
13  "created": 1677664795,
14  "id": "chatcmpl-7QyqpwdfhqwajicIEznoc6Q47XAYW",
15  "model": "gpt-3.5-turbo-0613",
16  "object": "chat.completion",
17  "usage": {
18    "completion_tokens": 17,
19    "prompt_tokens": 57,
20    "total_tokens": 74
21  }
22 }
```

The Anatomy of Messages

Role	Content	Example
System (input / prompt)	High-level task instruction	“You are ChatGPT, a helpful assistant for ...”
User (input / prompt)	Specific user requests	“What’s the function of p53?”
Assistant (output / completion)	LLM generation	“The function of protein p53 includes ...”

Coding & QA

https://colab.research.google.com/drive/1mBcFuELGuL18NyGDv3RCHuoT_cVSOVK7?usp=sharing

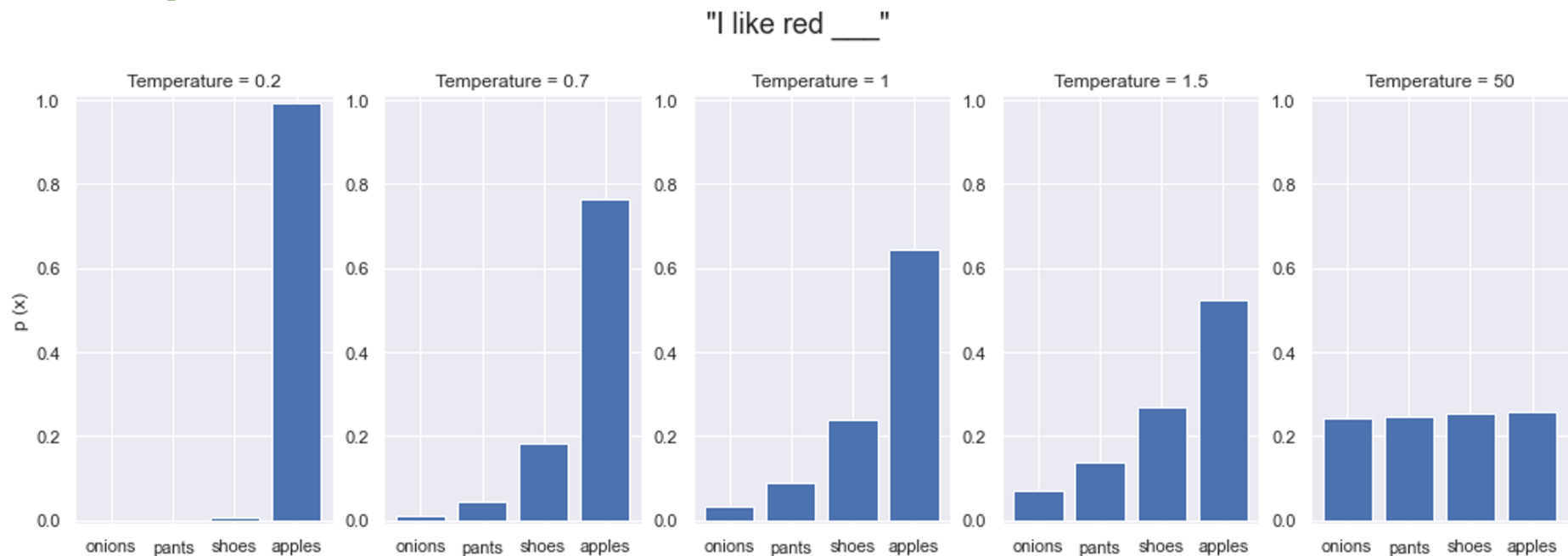
Parameters in OpenAI API

Only `messages` and `model` are the required parameters for GPT API

Other useful optional parameters include

Property	Parameter
Number of outputs	<code>n</code>
Randomness	<code>temperature</code> , <code>top_p</code> , <code>seed</code>
Repeatedness	<code>frequency_penalty</code> , <code>presence_penalty</code>
Termination	<code>max_tokens</code> , <code>stop</code>
Tools	<code>tools</code> , <code>tool_choice</code> , <code>parallel_tool_calls</code>

Temperature



Setting temperature=0 can almost get you deterministic results

Coding & QA

https://colab.research.google.com/drive/1mBcFuELGuL18NyGDv3RCHuoT_cVSOVK7?usp=sharing

OpenAI documentation

Start with the basics

Quickstart tutorial

Make your first Chat Completions API request

Prompt examples

Explore what OpenAI models can do with prompts

Many other interesting technical content not covered in this tutorial!

Thank you!

How to Use Open-source LLMs with Python