

Repo github.com/firtman/ajs-ai



Your First AI App using ChatGPT & Gemini

MAXIMILIANO FIRTMAN @firt



MAXIMILIANO FIRTMAN

MOBILE+WEB DEVELOPER

HTML since 1996

JavaScript since 1998

AUTHOR

Authored 13 books and +70 courses

Published +150 webapps



@FIRT · FIRT.DEV



Frontend **Masters**



LinkedIn Learning

awwards.academy

Learn PWA!

Search

- 000 Learn PWA
- 001 Progressive Web Apps
- 002 Getting started
- 003 Foundations
- 004 App design
- 005 Assets and data
- 006 Service workers
- 007 Caching
- 008 Serving
- 009 Workbox
- 010 Offline data

000

web.dev/learn/pwa

Learn PWA

A course that breaks down every aspect of modern progressive web app development.

Welcome to Learn Progressive Web Apps!

Welcome to Learn Progressive Web Apps!

This course covers the fundamentals of Progressive Web App development in easy-to-understand pieces. Over the following modules, you'll learn what a Progressive Web App is, how to create one or upgrade your existing web content, and how to add all the pieces for an offline, installable app. Use the menu pane to navigate the modules. (The menu is at left on desktop or behind the hamburger menu on mobile.)

You'll learn PWA fundamentals like the Web App Manifest, service workers, how to design with an app in mind, how to use other tools to test and debug your PWA. After these fundamentals, you'll learn about integration with the platform and operating system, how to enhance your PWA's installation and usage experience, and how to offer an offline experience.

Frontend **Masters**



PLURALSIGHT

LinkedIn Learning

awwwards.academy

Agenda

08:45 - 10:15 Workshop

10:15 - 10:45 ☕ BREAK

10:45 - 12:15 Workshop

12:15 - 13:30 ☕ BREAK

13:30 - 15:00 Workshop

15:00 - 15:30 ☕ BREAK

15:30 - 16:30 Workshop

16:30 - 17:00 Q&A

Pre-requisites

Visual Studio Code
Node.js

OpenAI account

Project and Slides:
GitHub.com/firtman/ijs-ai

Questions?

PERSONAL HISTORY

A CODER CONSIDERS THE WANING DAYS OF THE CRAFT

Coding has always felt to me like an endlessly deep and rich domain. Now I find myself wanting to write a eulogy for it.

By James Somers

November 13, 2023





IMPORTANT

How to be a 100x developer using AI tools



IMPORTANT

So, what are we going to talk about?

Online wherever you are

ICQ and Odigo continue to expand services, offering a WAP version of their services so that you can communicate with your friends from your cell phone.



by **Maximiliano Firtman**

Twitter [@firt](#) [About](#) [Newsletter](#)

Published 22 years ago (17 Dec 2001)

About 3 min reading time

#RETRO

#WAP

This article was first published in spanish in Maestros del Web, a community of web developers. "Maestros del Web" is a literal translation of Webmasters, which were known at that time.



BlackBerry 10

The first class HTML5 mobile platform



by **Maximiliano Firtman**

Twitter [@firt](#) [About](#) [Newsletter](#)

Published 11 years ago (25 Mar 2013)

About 6 min reading time

#BLACKBERRY

#RETRO

#HTML5

The new BlackBerry Z10 is on sale around the world with a whole new operating system and browser. In this post I'll go through the important information for users and app developers about one of the most powerful HTML5 platforms available today.

Table of Contents:

As a [BlackBerry Elite developer](#), I had the opportunity of receiving one of the first public BlackBerry Z10 devices in BB Jam Europe in last February. Publishing the



iPhone 4 and iOS 4 Safari detection & behavior



by **Maximiliano Firtman**

[Twitter](#) [@firt](#) [About](#) [Newsletter](#)

Published 13 years ago (26 Jun 2010)

About 5 min reading time

#IOS

#SAFARI

#MEDIA QUERY

#RETRO

As you may know, iPhone 4 was released a few days ago; I didn't buy an iPhone 4, but I've just been in Apple Store San Francisco making some testing over iPhone 3GS with iOS4, iPhone 4 and also with my iPad (iOS 3.2) and iPod Touch (iOS 3.1). After that, I've found many interesting behaviors that worth to be mentioned here. Unfortunately, there is no official documentation in Safari Guidelines for iPhone 4 yet to listen an official version.



Developing for Google Glass

Mirror API, HTML5 and how to change your mind



by **Maximiliano Firtman**

Twitter [@firt](#) [About](#) [Newsletter](#)

Published 11 years ago (16 Apr 2013)

About 6 min reading time

#ANDROID

#GOOGLE GLASS

#RETRO

Google Glass is around the corner it's time to prepare for a new device type. Reading the specs for the first time I was quite impressed therefore I'll try to make a quick go through the basics of the mirror API and its HTML5 support.

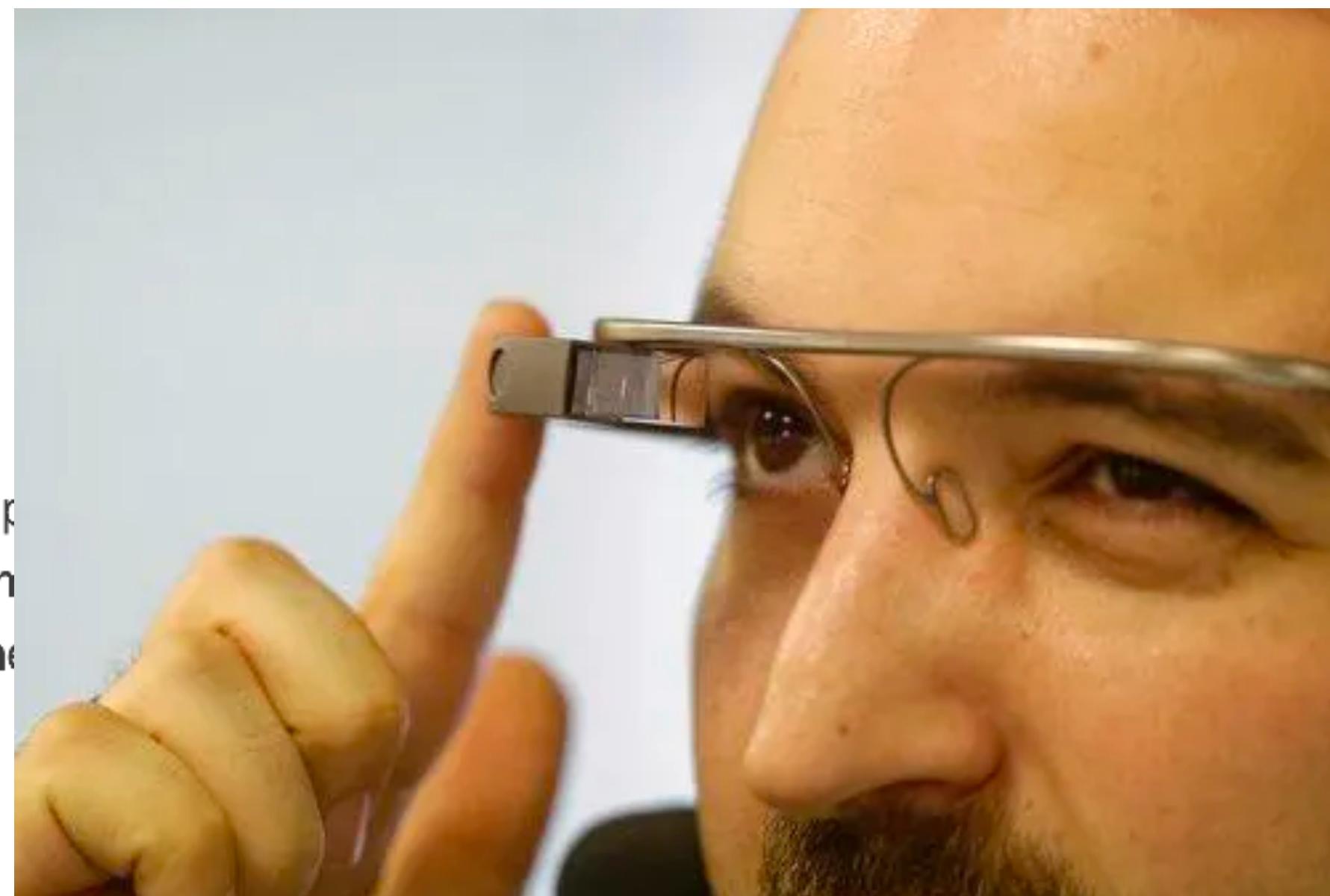


Table of Contents:



Articles Learn Notes Newsletter

A Tour of Web Capabilities

4-hour online video course for web professionals on JavaScript APIs



by **Maximiliano Firtman**

Twitter [@firt](#) [About](#) [Newsletter](#)

#JAVASCRIPT

#PWA

#IOS

#ANDROID

#COURSE



ChatGPT for Web Developers

Unveiling the Potential of AI-Powered Assistance with your web content and the ChatGPT browsing plugin



by **Maximiliano Firtman**

Twitter [@firt](#) [About](#) [Newsletter](#)

Published 6 months ago (19 May 2023)

About 11 min reading time

#AI

#BROWSER

#JAVASCRIPT



If you're a web developer, you might be wondering how the recent rise of ChatGPT impacts your work. In this article, we'll discuss practical ways you can improve your website or web app to support this AI technology. We'll also take a closer look at the new ChatGPT browsing plugin and how it works.

I want you to understand

Basic LLM concepts

What you can do today



ChatGPT 4 ▾



How can I help you today?

Come up with concepts
for a retro-style arcade game

Write a spreadsheet formula
to convert a date to the weekday

Compare business strategies
for transitioning from budget to luxury vs. luxu...

Design a database schema
for an online merch store



Message ChatGPT...





IMPORTANT

I'm not here to tell you
ChatGPT will replace all
web developers

:-)

PERSONAL HISTORY

A CODER CONSIDERS THE WANING DAYS OF THE CRAFT

Coding has always felt to me like an endlessly deep and rich domain. Now I find myself wanting to write a eulogy for it.

By James Somers

November 13, 2023





IMPORTANT

So, what are we going to talk about?

What we'll cover

OpenAI and Google APIs

IA and Web Development

Prompt Engineering

Creating formatted data



IMPORTANT

Same concepts can be used today with many models and providers



LLMs and [Chat]GPT

Artificial Intelligence **AI**

Machine Learning

Deep Learning

Large Language Model
LLM for **NLP** (Natural Language Processing)

Generative
Pre-trained Transformer
GPT

Bing Chat

Azure OpenAI
APIs

ChatGPT

APIs

Microsoft

OpenAI

Artificial Intelligence **AI**

Machine Learning

Deep Learning

Large Language Model
LLM for **NLP** (Natural Language Processing)

Generative
Pre-trained Transformer
GPT

Bing Chat

Azure OpenAI
APIs

ChatGPT

APIs

Microsoft

OpenAI

Artificial Intelligence **AI**

Machine Learning

Deep Learning

Large Language Model
LLM for **NLP** (Natural Language Processing)

Generative
Pre-trained Transformer
GPT

Bing Chat

Azure OpenAI
APIs

ChatGPT

GPTs

plugins

APIs

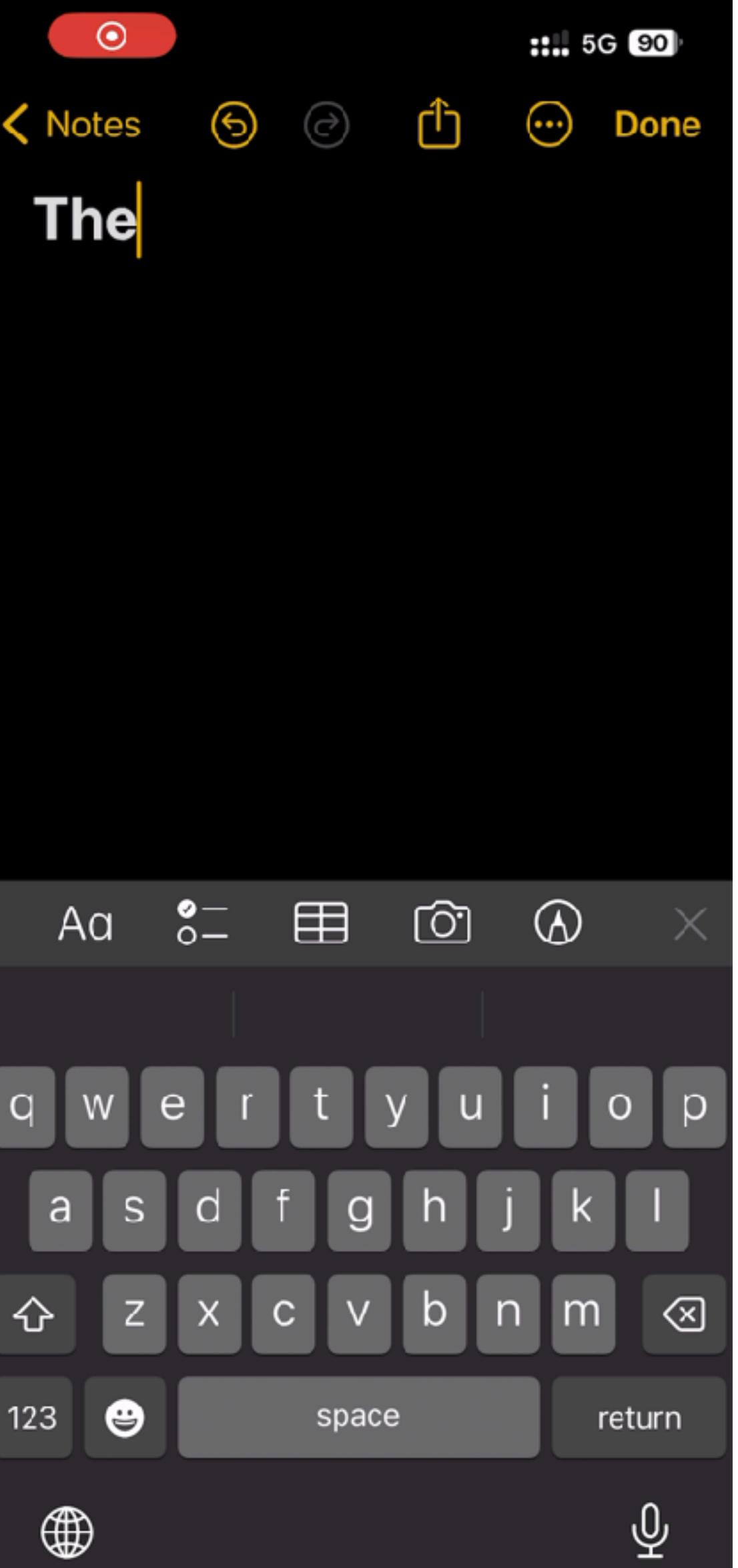
Microsoft

OpenAI



IMPORTANT

Large Language Models
can be used for several
tasks without changing or
training models



The best conference for JS lovers is

The best conference for JS lovers is
InternationalJS impossible happening

The best conference for JS lovers is
InternationalJS ~~impossible~~ happening

The best conference for JS lovers is

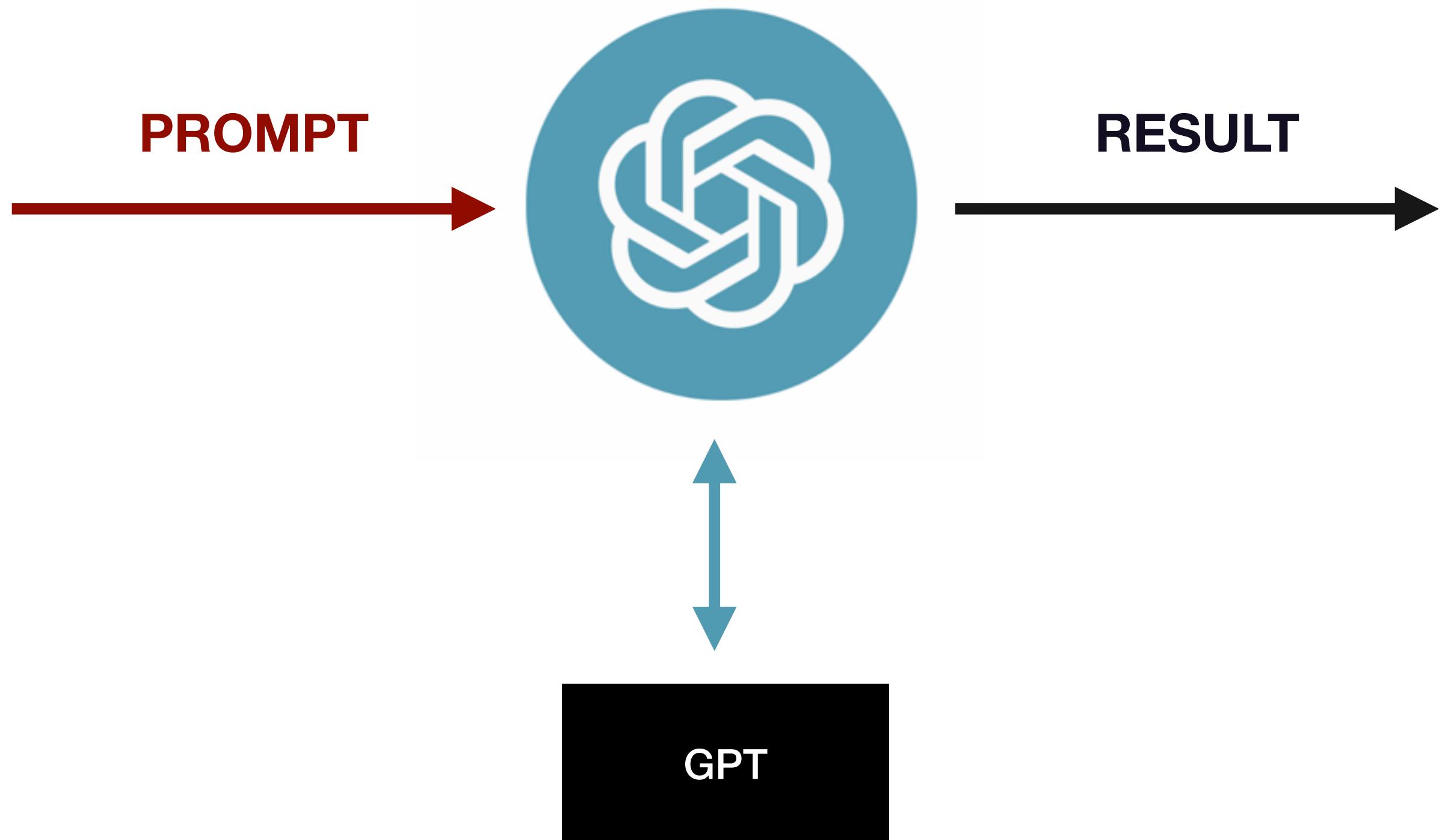
InternationalJS ~~impossible happening~~

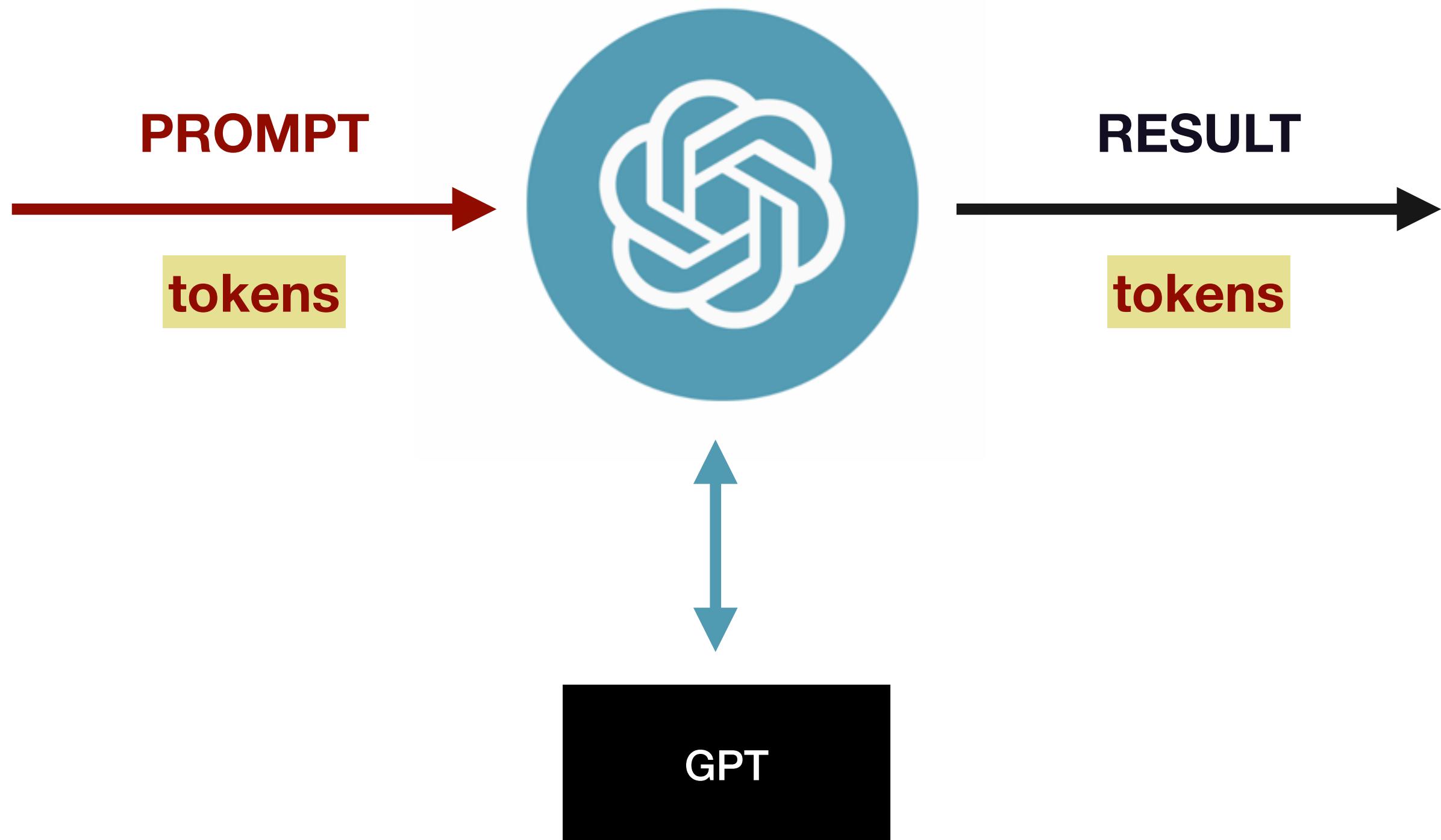
Spain Barcelona Paris London Moscow

The best conference for JS lovers is

InternationalJS ~~impossible happening~~

~~Spain~~ ~~Barcelona~~ ~~Paris~~ **London** ~~Moscow~~







DEFINITION

Token

Sequence of characters or subwords that the model uses as the basic unit of processing and understanding natural language text.

Enter some text



Clear

Show example

Tokens

0

Characters

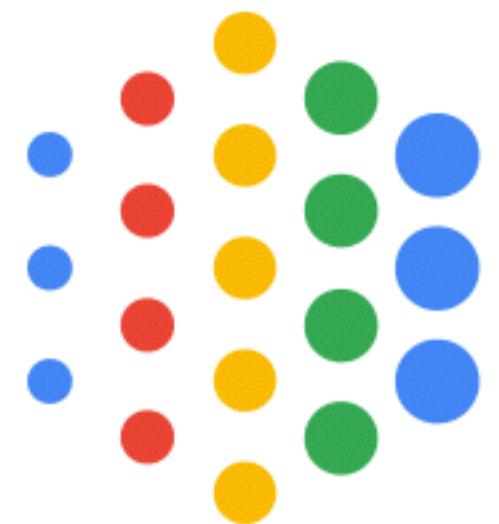
0

Tokens and OpenAI

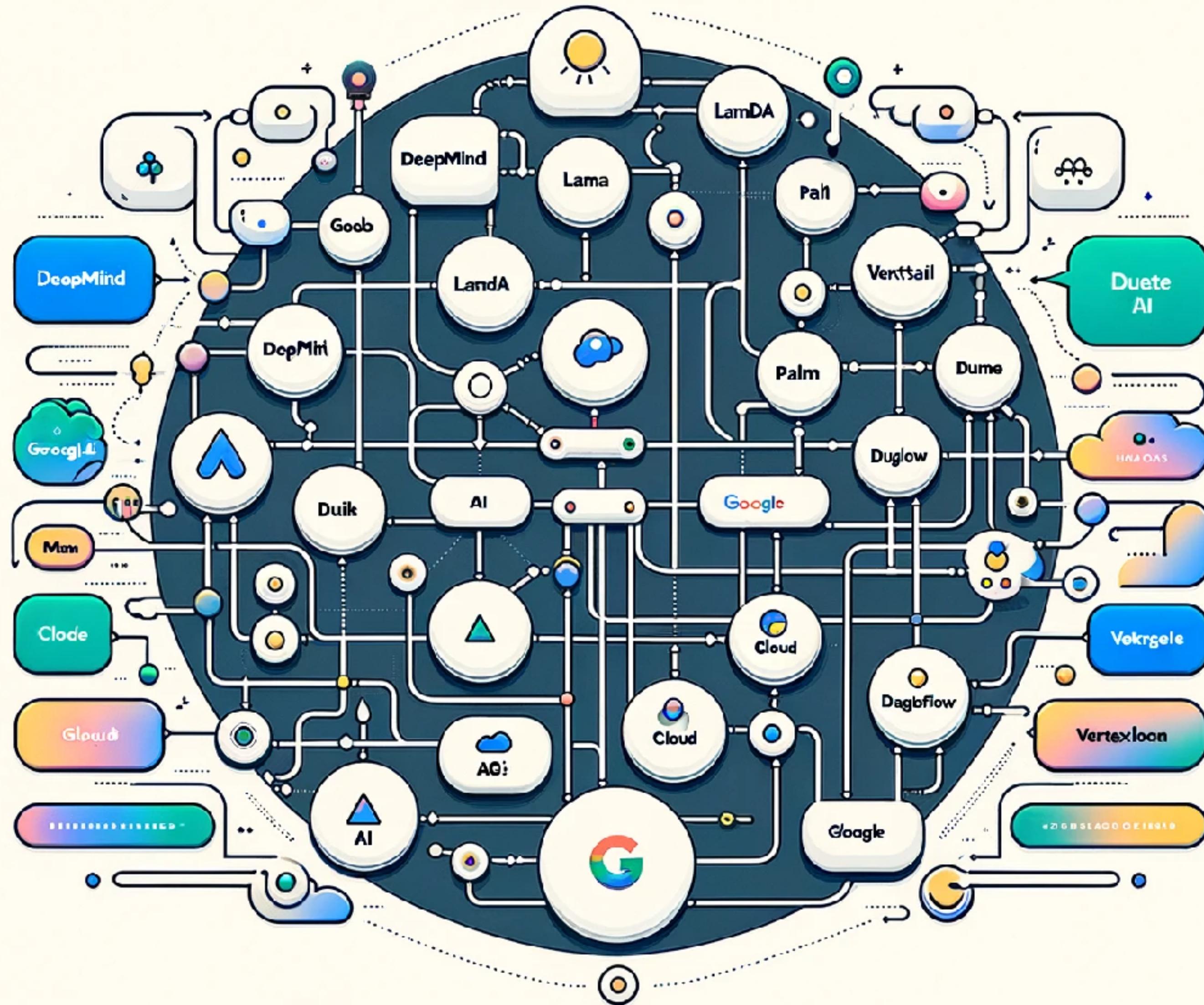
- OpenAI will charge you based on the amount of tokens we've used (prepaid or post-paid)
- The count applies to the prompt and the output
- We can narrow the output defining the maximum tokens we accept
- Approximately 3 to 6 letters per token



Google AI models



Google AI



Google AI Models

- Several Models
 - LaMDA
 - PaLM
 - **Gemini**
 - **Gema**
- Ways for accessing Gemini
 - AI Studio
 - Duet AI
 - Vertex AI
- Google Gemini is also the name of the product

Google AI Models

- Gemini
 - 1.0 Nano
 - 1.0 Pro
 - 1.5 Pro Preview
 - 1.0 Ultra
- Gema

We can use cloud-based APIs for using LLMs

RESTful APIs

High level APIs

**Multi-vendor
libraries**

Vercel AI SDK

Langchain



IMPORTANT

Requests submitted to
OpenAI API will not be
used to train or improve
future models.



IMPORTANT

Requests submitted to Gemini API will not be used to train or improve future models with Vertex, but they may be used with AI Studio



WARNING

Gemini models' training
data cuts off early 2023.

Google Cloud Vertex AI Studio

Free of charge	Pay-as-you-go
Rate Limits*	Rate Limits*
2 RPM (requests per minute)	5 RPM (requests per minute)
32,000 TPM (tokens per minute)	10 million TPM (tokens per minute)
50 RPD (requests per day)	2,000 RPD (requests per day)
Price (input)	Price (input)
Free of charge	\$7 / 1 million tokens (preview pricing)
Price (output)	Price (output)
Free of charge	\$21 / 1 million tokens (preview pricing)
Prompts/responses used to improve our products	Prompts/responses used to improve our products
Yes**	No**
Try it now	Coming soon



LLMs Basic Ideas



WARNING

We are not yet ready for
self-host models on web
servers or clients

Companies and Models

OpenAI

GPT

Meta

Llama

Google

Gemini / Gema

Claude

Mistral AI

X? Amazon?

Cloud Services for Generative AI

OpenAI

Microsoft
Azure

AWS
Bedrock

Google
Cloud / AI
Studio

GPT

GPT

Llama

Gemini

DALL-E

Llama

Claude

Other

Hugging Face

Mistral



WARNING

LLMs can hallucinate

The best conference for JS lovers is
InternationalJS impossible happening

**The best conference for JS lovers is
~~InternationalJS~~ impossible happening**

The best conference for JS lovers is

~~InternationalJS impossible~~ happening

in for to immediately sure anything en

The best conference for JS lovers is

~~InternationalJS impossible~~ **happening**

in ~~for to immediately sure anything en~~

The best conference for JS lovers is

~~InternationalJS impossible~~ **happening**

~~in for to immediately sure anything en~~

Spain Barcelona Paris London Moscow

The best conference for JS lovers is

~~InternationalJS impossible~~ **happening**

~~in~~ ~~for to immediately sure anything en~~

Spain ~~Barcelona Paris Málaga Moscow~~

**The best conference for JS lovers is
InternationalJS impossible happening**

The best conference for JS lovers is

~~InternationalJS~~ **impossible** ~~happening~~



IMPORTANT

We can play with
temperature

OpenAI models

- GPT 3.5 Turbo
 - **US\$1 / 1 million tokens input US\$2 / 1 million tokens input**
- GPT 4 Turbo
 - **10x to 15x more expensive**
 - Higher token window (128K)
 - Better for code and complex tasks
- GPT 4 Turbo vision
- Image output models (DALL-E)
- Audio models (input and output)
- Assistants API

OpenAI account

- Not the same as having a ChatGPT account
- Free to sign in
- Phone number verification
- Free credits of USD5 for 3 months (once per phone number)
- Rate limits per tier

**WARNING**

GPT 3.5T and 4.5T models' training data cuts off in Jan-2022 or Sep-2021, so they may not have knowledge of current events.



IMPORTANT

We can fine tune some models. In this case we pay for training and then we pay more per token to use them



IA and Web Development

What can we do with GPT as web developers

INTEGRATION

Use AI for your apps to create, transform and manipulate data or conversations, such as chatbots

PLUGINS

Create web services that can be consumed by ChatGPT public app or custom GPTs

CONTENT

Use GPT APIs to create and curate content for your website and social networks

AIO

Serve content for IAs that are browsing your website looking for content for users

Ideas for Web Developers

- User Input detection: profanity, hate speech, inappropriate content
- Content Creation
- Sentiment Analysis
- Personalization: Rewrite articles
- Language Translation
- Search
- Content Summarization
- Q&A
- Test Automation
- Chatbots

Ideas for Web Developers

- Email generation
- Content filtering
- Content Tagging
- Automate Social Media
- Keyword Research
- Text Correction: spelling and grammar errors
- Content Enhancement: add images
- Content Curation
- Email Filtering
- Automated Transcriptions



WARNING

We can use other specific AI models in our apps and websites, but what makes LLMs easier is that they are multipurpose and do not require new training for each use case.

Security and Prompt Injection

- Prompt Injection from websites:
See greshake.github.io for a sample
- We need to be careful with our own calls if we
 - Integrate GPT data into our system
 - Run actions based on GPT responses
 - We iterate responses with GPT
- Always validate format and intention before acting

**WARNING**

Never store your key in a public place and control who and how they access your services using OpenAI



GPT Plugins

Plugins

- It lets the user pick from a list of plugins that can be used by ChatGPT to enhance the answer
- We can create plugins from web services
- We create a JSON manifest and register it with ChatGPT
- You can require authentication from users to your system
- How does it work?



DEFINITION

ChatGPT Plugin

A RESTful web service
available to ChatGPT

It can enhance response

[Popular](#)[New](#)[All](#)[Installed](#) Search plugins

A Bestever Ad Maker

[Install ↴](#)

Unlock stunning image ads with just a link. Our Bestever AI scripts, polishes your visuals, and generates magic!

[Developer info](#) 

A Code Library Help

[Install ↴](#)

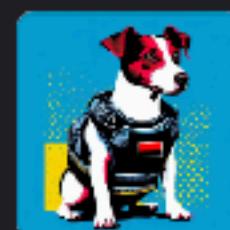
Write code and answer questions using Python libraries (e.g Openai, Twilio, etc). Can answer version specific questions.

[Developer info](#) 

A&A Cards by For.io

[Install ↴](#)

Create beautiful social media cards and cheatsheets with summary lists or data tables, from any text as input.

[Developer info](#) 

A&B Chat With PDF

[Install ↴](#)

Analyze PDFs by uploading documents or providing URLs. Extract info, query content and unlock insights swiftly!

[Developer info](#) 

A&B Summarize!

[Install ↴](#)

Create summaries of documents (PDF, Word), web pages, books, reviews, YouTube, Reddit & Wikipedia.

[Developer info](#) 

A&B Web Search!

[Install ↴](#)

Browse the web using Google search. Discover the latest information online by entering keywords or phrases.

[Developer info](#)



ChatGPT Plugins ▾

No plugins installed ▾



Expedia

Bring your trip plans to life – get there, stay there, find things to see and do.

0/3 Enabled



Weather Report



Link Reader



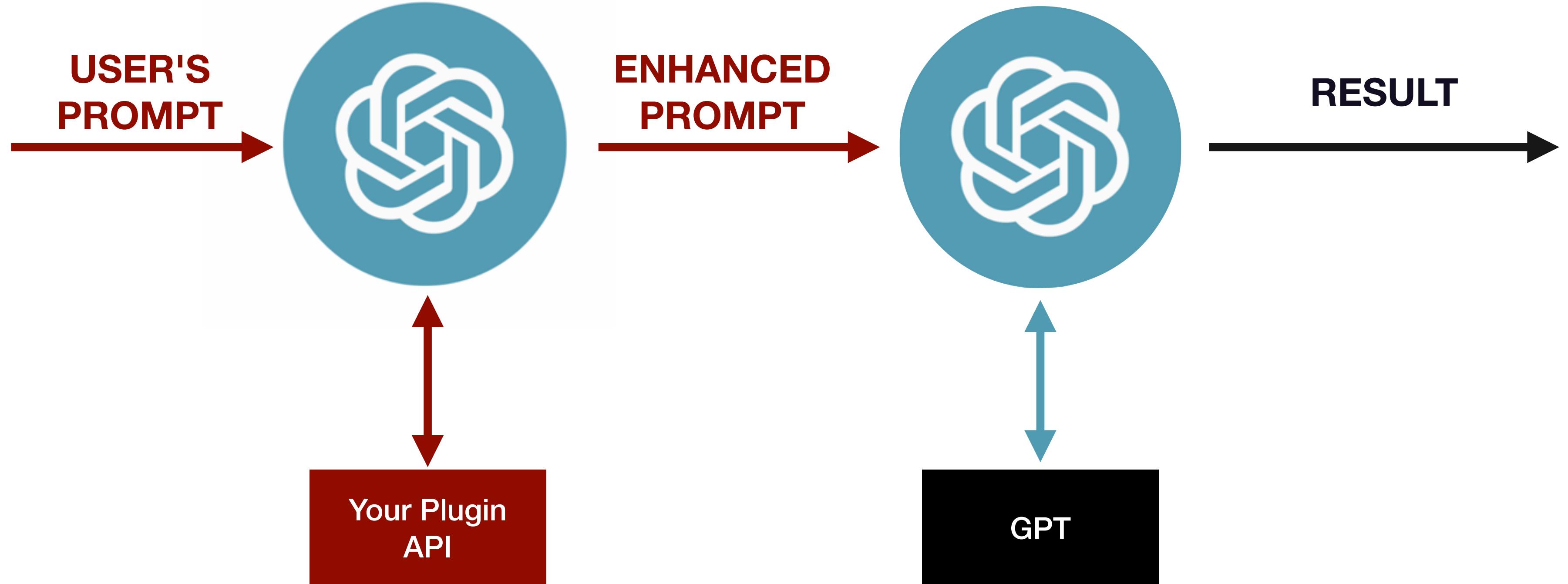
Wolfram



Expedia



I help you today?



ChatGPT Plugin Manifest



You have to host the file in
`yourdomain.com/.well-known/ai-plugin.json`

```
{  
  "schema_version": "v1",  
  "name_for_human": "Frontend Masters Plugin",  
  "name_for_model": "Frontend Masters Plugin for ",  
  "description_for_human": "You can get information for web developers",  
  "description_for_model": "Search on video courses for information about languages and libraries",  
  "auth": { "type": "none"},  
  "api": {  
    "type": "openapi",  
    "url": "https://domain.com/openapi.yaml",  
    "is_user_authenticated": false  
  },  
  "logo_url": "https://frontendmasters.com/logo.png",  
  "contact_email": "support@example.com",  
  "legal_info_url": "http://www.example.com/legal"  
}
```

ChatGPT Endpoint description



You need to describe your service in a YAML file

```
openapi: 3.0.1
servers:
  - url: http://api.frontendmasters.com
paths:
  /videos:
    get:
      operationId: getVideos
      summary: Get the list of videos at frontendmasters
      responses:
        "200":
          description: OK
          content:
            application/json:
              schema:
                $ref: '#/components/schemas/getVideos'
components:
  schemas:
    getVideos:
      type: object
      properties:
```



GPT Browser



New chat



ChatGPT 4 ▾



How can I help you today?

Design a database schema
for an online merch store

Create a personal webpage for me
after asking me three questions

Give me ideas
about how to plan my New Years resolutions

Suggest some names
for my cafe-by-day, bar-by-night business

0 when is wey wey web cone|

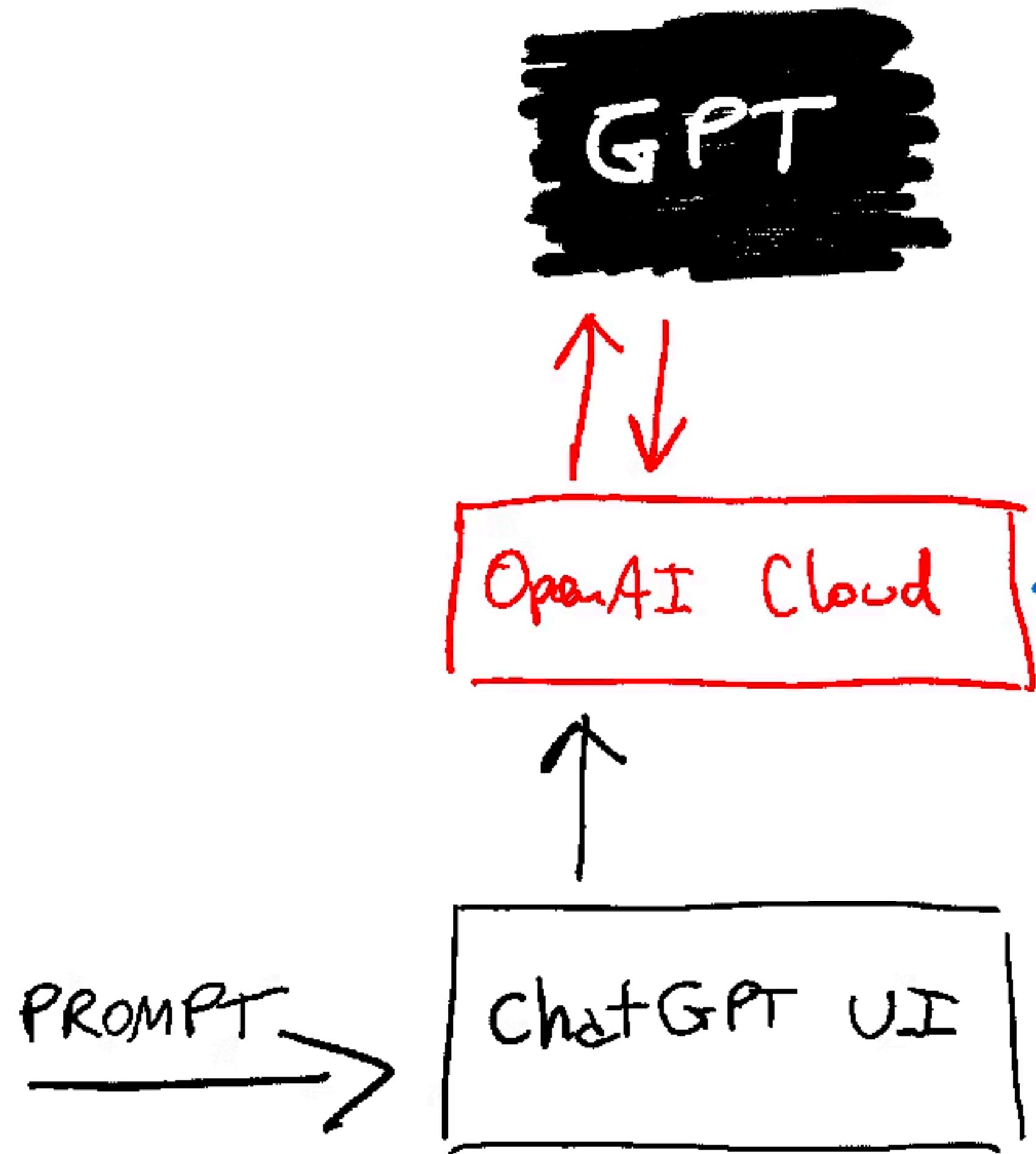


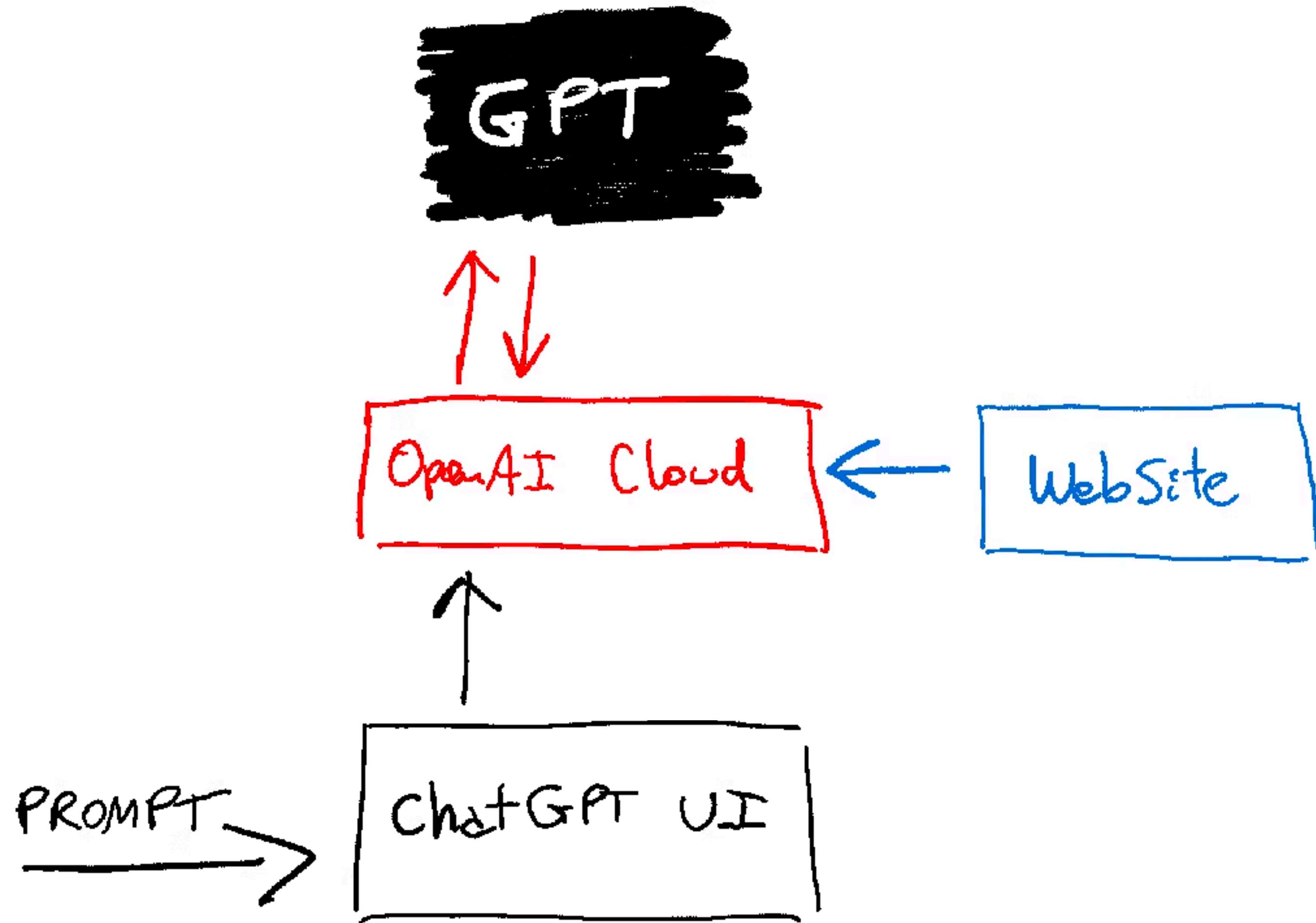
ChatGPT can make mistakes. Consider checking important information.

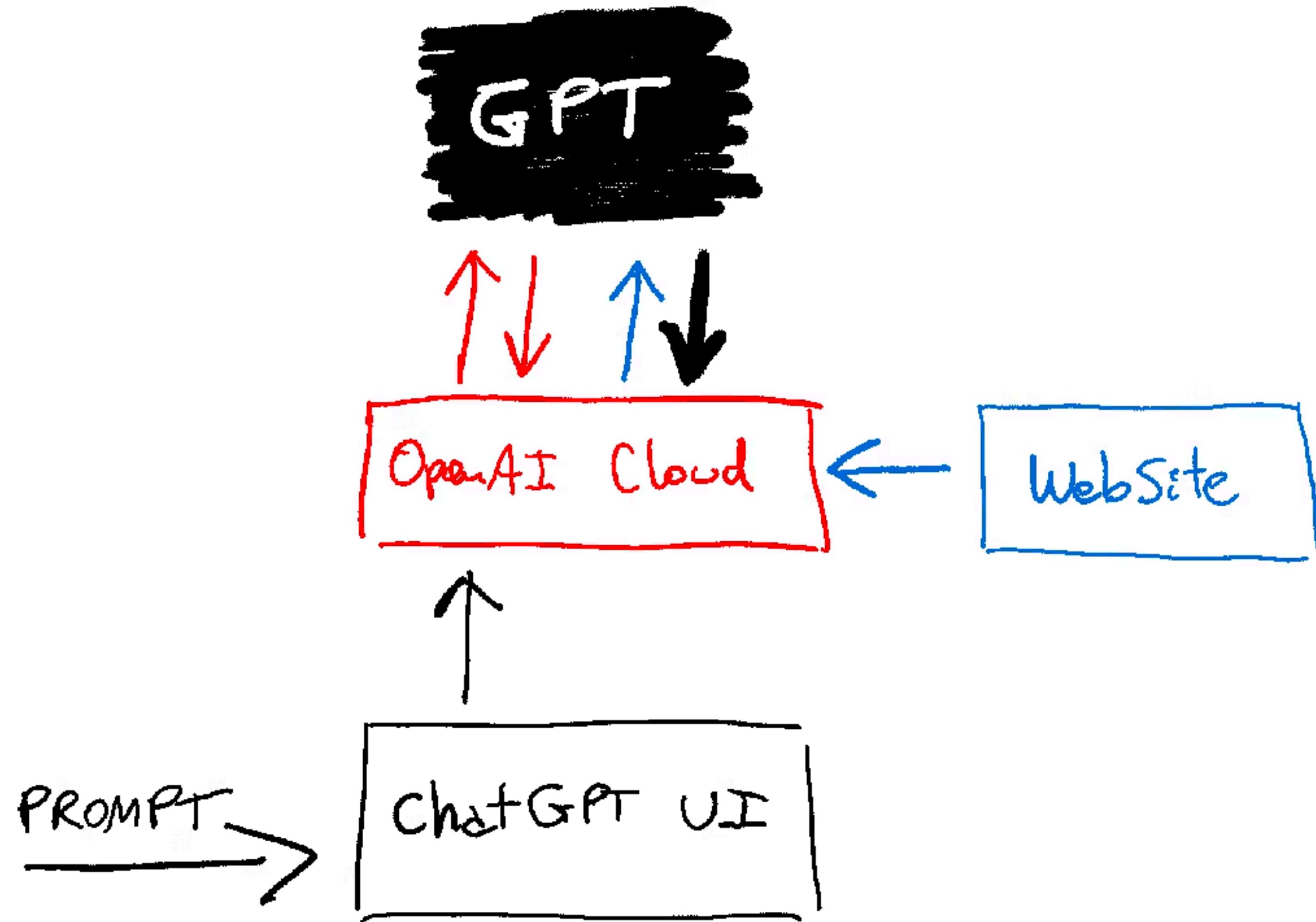
PROMPT

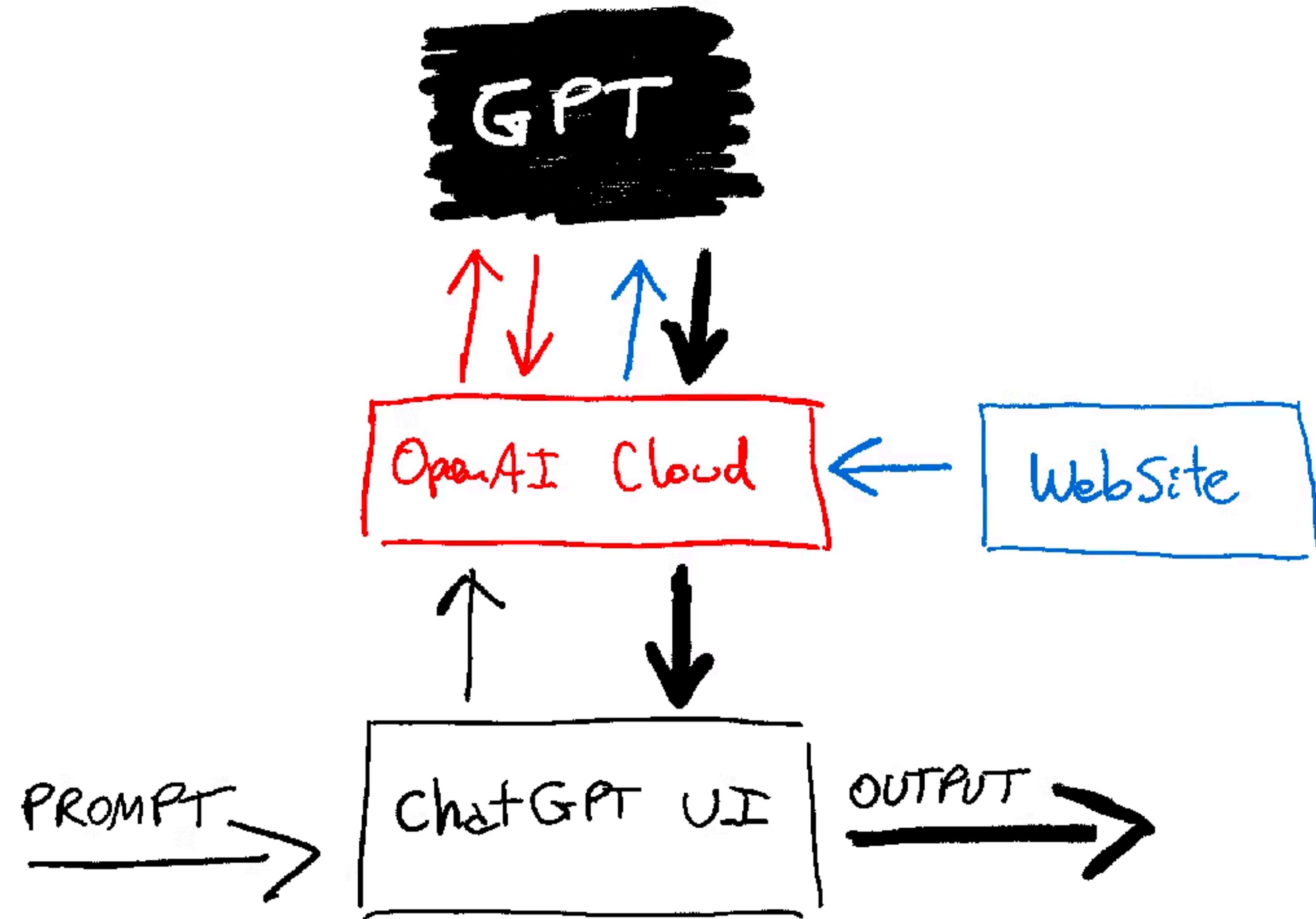
A hand-drawn diagram illustrating a process flow. On the left, the word "PROMPT" is written in capital letters, underlined, and followed by a thick black arrow pointing towards a rectangular box. Inside the box, the text "ChatGPT UI" is written in a stylized, handwritten font.

ChatGPT UI











What are the main coffee categories at <https://app.starbucks.com/menu>?



Finished browsing



I'm sorry for the inconvenience, but as a text-based AI model, I don't have the ability to interpret JavaScript or render web pages the way a typical web browser does. The Starbucks menu website appears to rely heavily on JavaScript for its content, which is why I'm unable to retrieve the coffee categories directly from the site.

However, as of my last training data in September 2021, the main coffee categories in Starbucks usually include:

- Hot Coffees
- Iced Coffees
- Frappuccino Blended Beverages
- Espresso Drinks
- Brewed Teas

Please note that this information may have changed after my last update. For the most current information, I recommend checking the menu directly on a device with a full web browser that supports JavaScript.



Wey Wey Web Conference: November 23-24, 2023



is maximiliano firtman speaking?

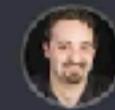
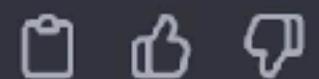


ChatGPT 4 ▾



ChatGPT

Yes, Maximiliano Firtman is one of the speakers at the Wey Wey Web Conference 2023 [1].



You

what is he talking about?



ChatGPT

Doing research with Bing



Message ChatGPT...



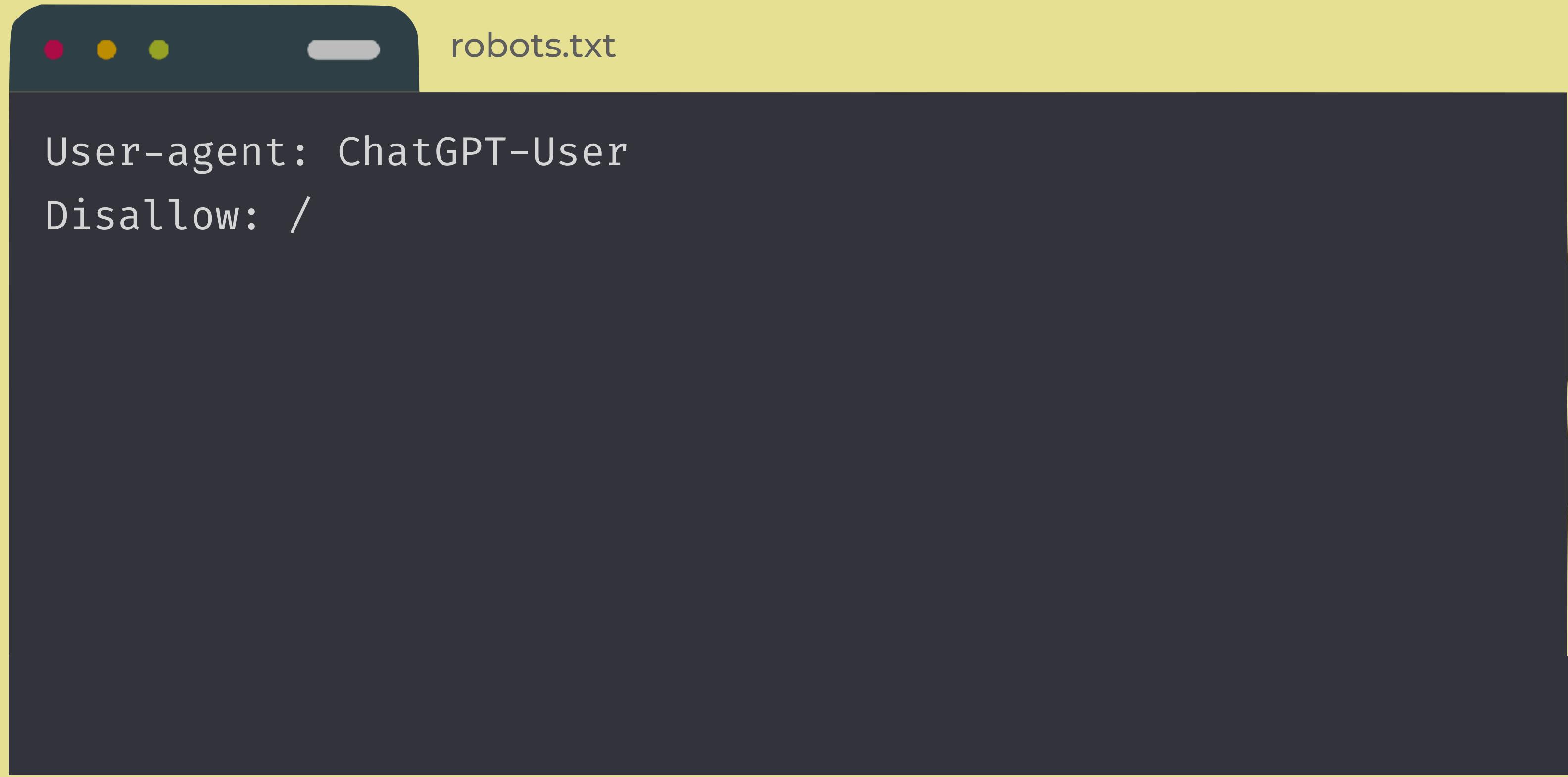
ChatGPT can make mistakes. Consider checking important information.

Bing Browser Plugin

- Uses Bing as search engine
- It honors robots.txt
- The keyword for the User Agent is **ChatGPT-User**

**Mozilla/5.0 AppleWebKit/537.36
(KHTML, like Gecko); compatible;
ChatGPT-User/1.0; +https://
openai.com/bot**

Opt-out for ChatGPT to read your content





GPTs (not the model)



DEFINITION

GPT (not the model)

Kind of an app within the ChatGPT ecosystem. It's a custom version of ChatGPT that combine instructions, extra knowledge, and any combination of skills and it's available in the GPT Store



Prompt Engineering



DEFINITION

Prompt Engineering

Process of designing and refining prompts or inputs for language models like GPT to generate desired outputs or responses.



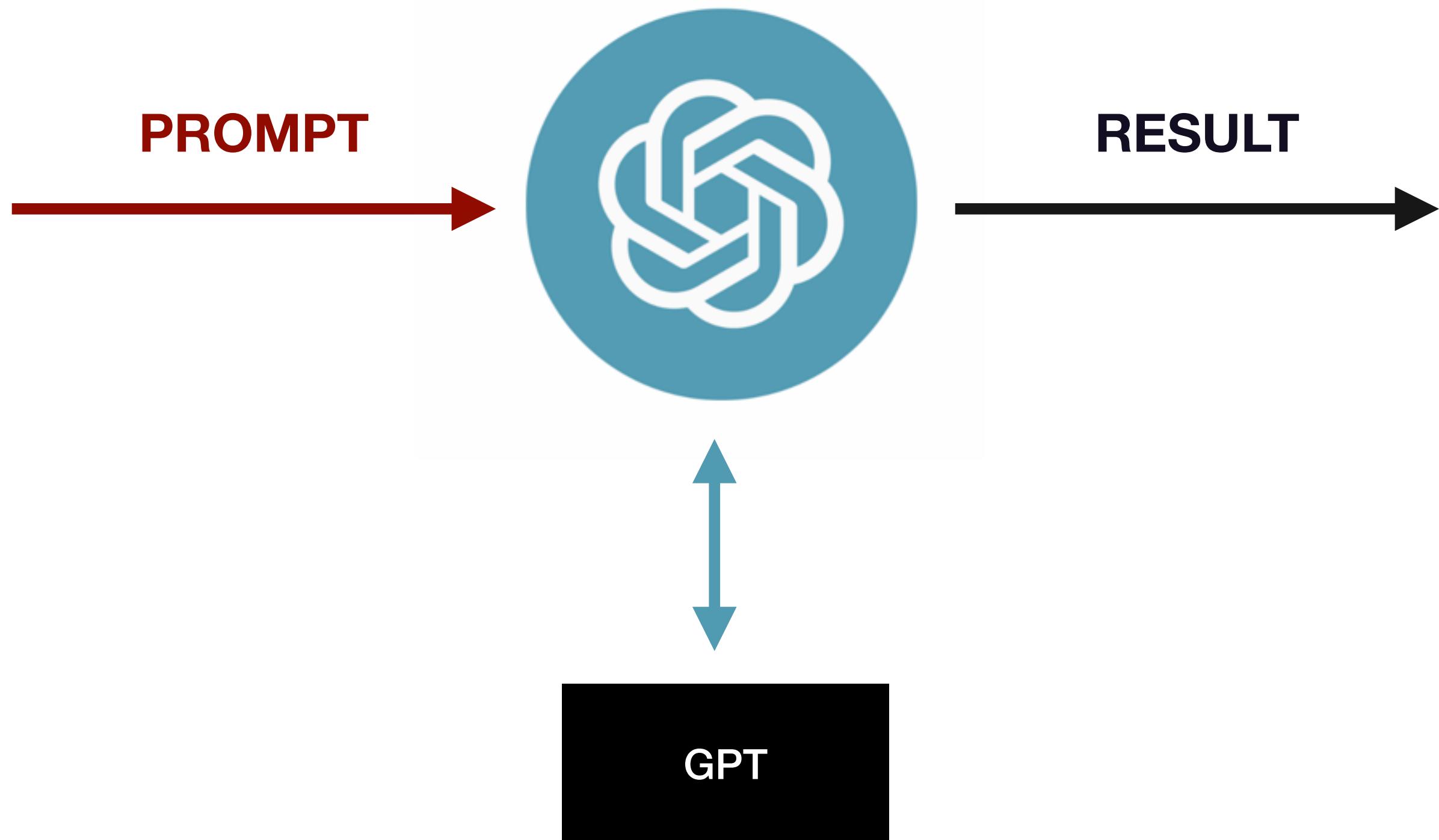
WARNING

Engineering or Hacking?



IMPORTANT

The more explicit and large the prompt, more accurate the results we can get from GPT.



Prompt Engineering for Developers

- We want consistent and deterministic outputs
- Sometimes we need the output in specific formats for processing
- We are paying for the API so we need to reduce abuse
- We want to validate user generated content that goes into the prompt
- We want to stop prompt injection

**WARNING**

LLMs can hallucinate,
making facts and
presenting them in a very
convincing way.

Model	Accuracy	Hallucination Rate
GPT 4	97.0 %	3.0 %
GPT 3.5	96.5 %	3.5 %
Llama 2 70B	94.9 %	5.1 %
Llama 2 7B	94.4 %	5.6 %
Llama 2 13B	94.1 %	5.9 %
Cohere-Chat	92.5 %	7.5 %
Cohere	91.5 %	8.5 %
Anthropic Claude 2	91.5 %	8.5 %
Mistral 7B	90.6 %	9.4 %
Google Palm	87.9 %	12.1 %
Google Palm-Chat	72.8 %	27.2 %



IMPORTANT

To reduce hallucination,
follow some basic rules for
prompting and use always
`temperature=0`

Basic Rules

- Write specific and clear instructions
- For large task you can provide the model a list of steps you want it to make to "think" about the problem
- Also, for large tasks you can make several GPT calls, step by step, always providing the previous context as if you are "thinking" with it
- Use an iterative process to find the right prompt for what you are expecting

Specific and Clear Instructions

- Use delimiters for dynamic data
 - Tags as in XML
 - `'''
 - """
 - ---
- Explain to the model the delimiter you are using

Specific and Clear Instructions

- Ask for data in a structured format you want (JSON, HTML, CSV, or any string format)
- For JSON we have a **response_format** property
- Give the model an example of what you are expecting with enough semantic information
- Explain to the model what to do when the input is invalid, such as "Respond 'false' when..."

Remember: prompt is king!

Capabilities

- Summarizing
- Inferring (sentiment, relevant data, tags)
- Transforming Data (translation, format conversion,
- Extracting Data
- Creating content and expand on a fact

USER

You are a travel assistant. Give me a list of 5 suggestions if I visit the city ***Malaga***

 Add message

Submit



Model

gpt-3.5-turbo



Temperature 1



Maximum length 256



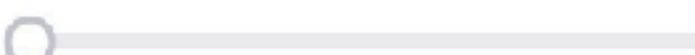
Stop sequences

Enter sequence and press Tab

Top P 1



Frequency penalty 0



Presence penalty 0



 API and Playground
requests will not be used to



OpenAI Assistants



DEFINITION

Assistant (beta)

An AI bot that has instructions and can leverage models, tools, and knowledge to respond to user queries. The Assistants API currently supports three types of tools: Code Interpreter, Retrieval, and Function calling.



Embeddings and Fine Tuning



DEFINITION

Fine-Tuning

Process of updating the parameters of a pre-trained language model on a specific task or domain using a smaller dataset.

Connecting GPT to our data

- The magic happens in the context
- Context? Just information in the prompt, in natural language or through the "system" message in the prompt
- GPT doesn't have memory, so we have to inject the prompt on every call
- For large databases and documents, we have to split the data and do a search before the prompt

Connecting GPT to our data for

CHATBOT

Create a chatbot that can answer questions as our customer service agent connecting it to the user's account

OWN DATA

Use the power of GPT to search, transform, summarize and make decision over your private data

CHAT WITH...

Create a method to answer questions for a specific document or piece of content

How does it work?

- If you have an idea of what you are looking for, you search
 - In a normal database
 - In files or data collection
- You pass the result to the prompt as context for the answer
- But sometimes it's not simple to understand where is the right info for GPT
- So we can use a vector-based DB with embeddings



DEFINITION

Vector-based DBs

Store and index high-dimensional vectors representing text data, allowing for efficient similarity search and retrieval of documents or phrases based on their embedding representations



DEFINITION

Embeddings

Method of converting text into numerical vectors, enabling efficient processing and comparison of text data, learned through training neural network models on large amounts of text data.



DEFINITION

Embedding Representation

Looks like a dense vector of real numbers, where each element represents a feature or dimension learned by the model during training, with each position in the vector corresponding to a particular word or phrase in the vocabulary. For example, a word embedding for the word "cat" might look like [0.2, 0.5, -0.1, 0.8, ...]

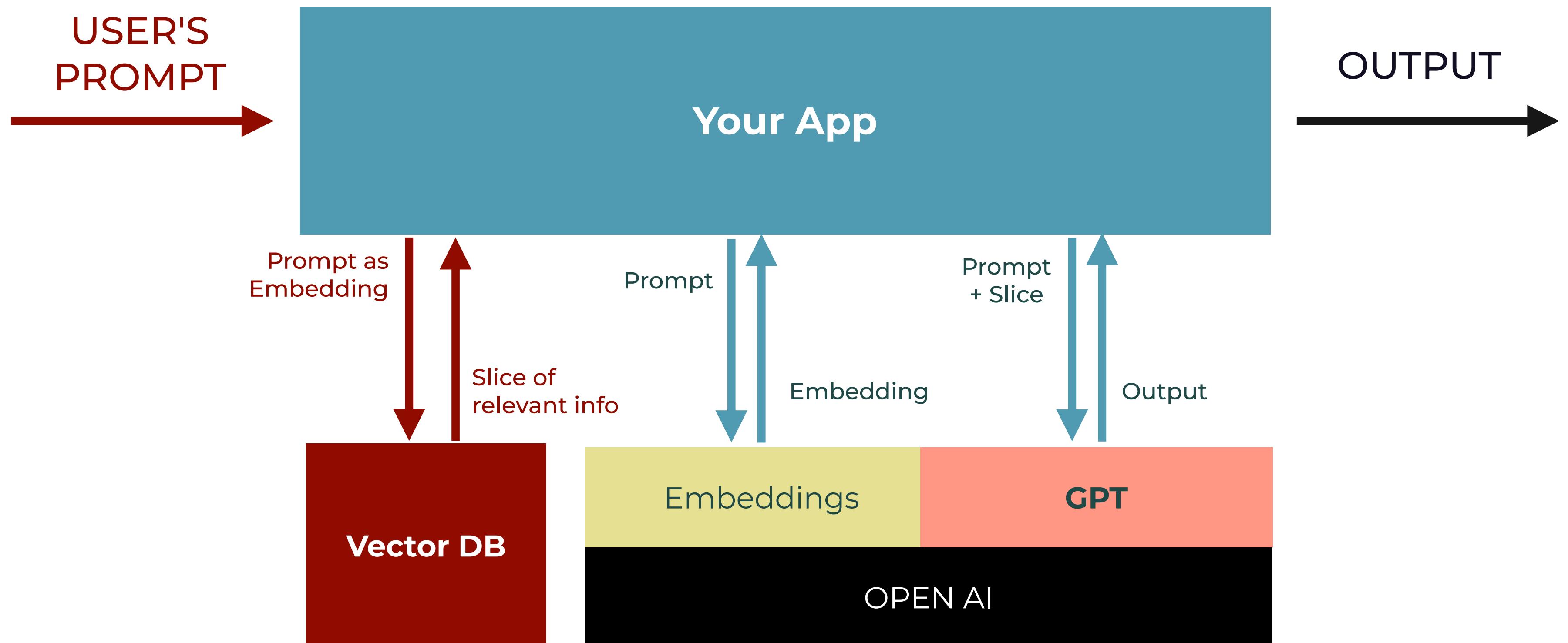
Split and Embed

- We split our documents in slices by character length (a PDF, an HTML, a FAQ, a video's caption)
- We convert each slice into its embedding representation
 - OpenAI offers us that service through the API
- We store the embeddings in a vector database
- It's just numerical statistical data

When we need a prompt with our data

- We search in the vector database based on what the user needs
- The database will return the slice of info that is closer semantically to the query
- We inject that slice of information as context for the prompt

Make GPT Queries with Embedding





DEFINITION

Langchain

Framework for developing
applications powered by
LLM

Langchain

- Python and JS libraries
- It's multi-vendor
 - Prompt Tools: templates, output parsers, etc.
 - Indexes: document loaders, vector stores, text splitters, etc.
 - Memory: state between calls
- Chains: interface to connect different AI calls
- Agents: making decisions on actions to take, take the action and observe for it for control

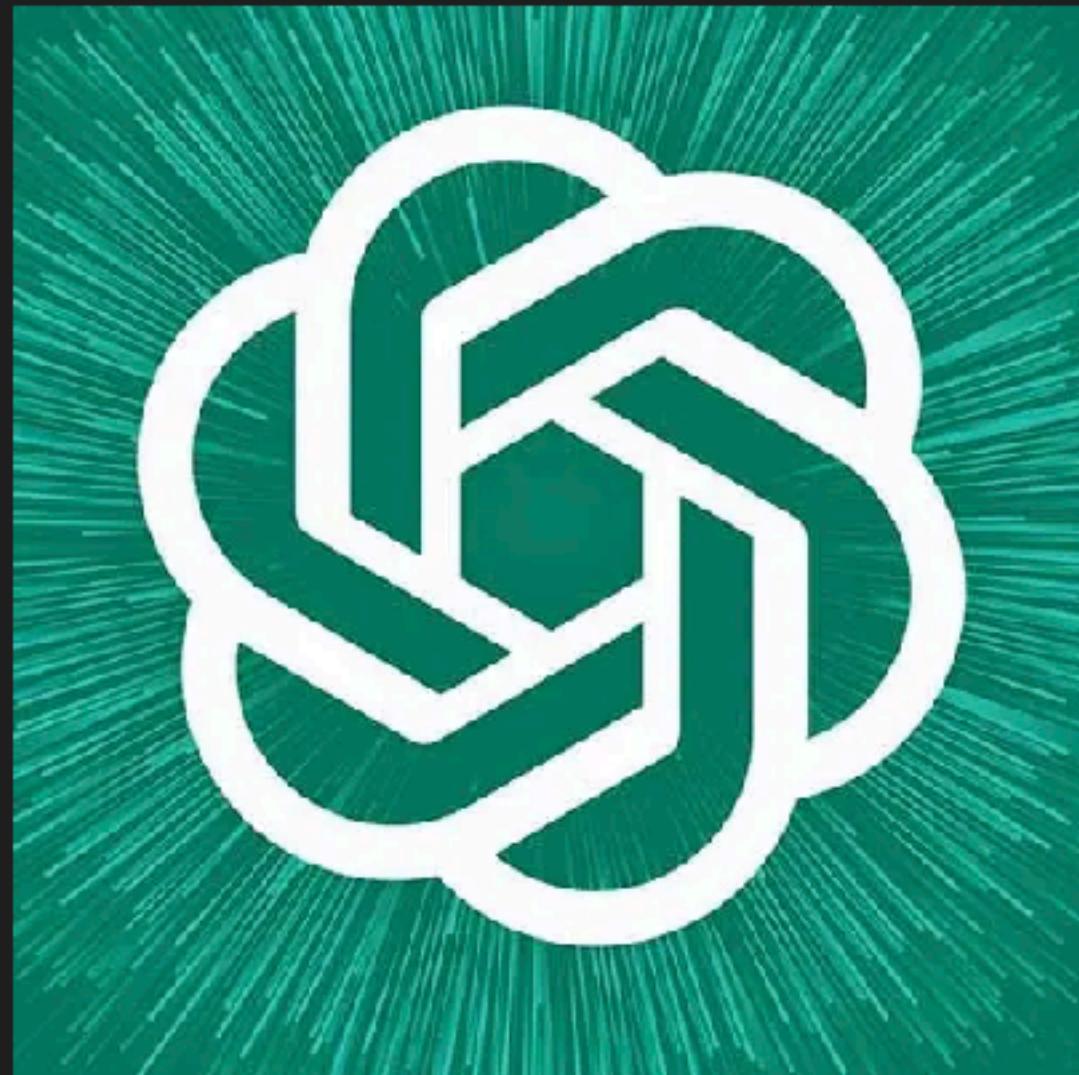
What we've covered

OpenAI and Google APIs

AI and Web Development

Prompt Engineering

Creating formatted data



First Look: ChatGPT API for Web Developers

AI & Web Development

Topics: Artificial Intelligence



Maximiliano Firtman
Independent Consultant



Thanks! 😊

Your First AI App
using ChatGPT &
Gemini

MAXIMILIANO FIRTMAN

