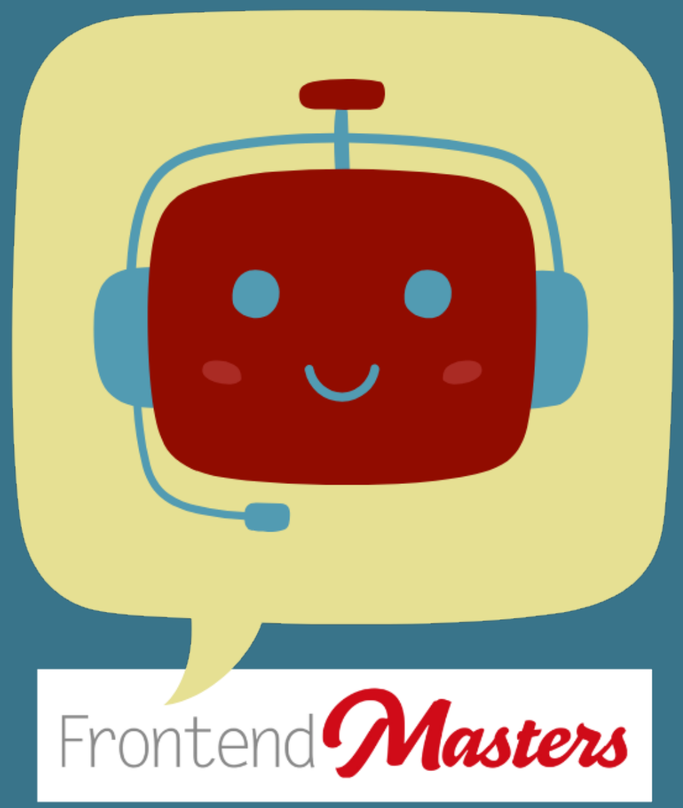


ARTIFICIAL INTELLIGENCE AND LLMs

CHATGPT FOR WEB DEVELOPERS

MAXIMILIANO FIRTMAN

MAXIMILIANO FIRTMAN

MOBILE+WEB DEVELOPER

HTML since 1996

JavaScript since 1998

AUTHOR

Authored 13 books and +70 courses

Published +150 webapps

@FIRT · FIRT.DEV

Model

Default (GPT-3.5)

ChatGPT PLUS

Send a message.

DEFINITION

ChatGPT

ChatGPT is a large language model designed to generate human-like responses to natural language input, using the GPT architecture.

DEFINITION

ChatGPT

ChatGPT is a computer program that can understand and respond to written text like a human would.

DEFINITION

ChatGPT

is like a talking robot that can talk with us by reading what we write and answering us in a way that sounds like a person talking

define ChatGPT in one sentence

ChatGPT is a large language model designed to generate human-like responses to natural language input, using the GPT-3.5 architecture.

define ChatGPT in one sentence for someone that has never heard anything on the AI topic

ChatGPT is a computer program that can understand and respond to written text like a human would.

now as if we are a 5-years old

ChatGPT is like a talking robot that can talk with us by reading what we write and answering us in a way that sounds like a person talking.

IMPORTANT

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

:-)

IMPORTANT

So, what we going to talk about?

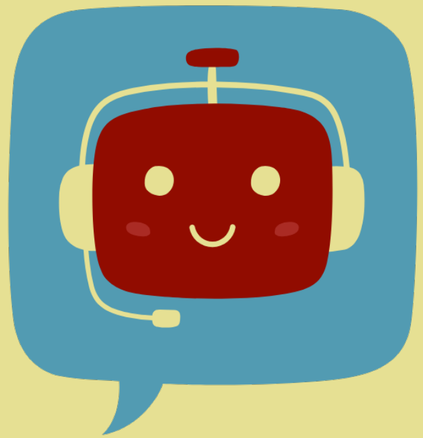
What we'll cover

OpenAI and ChatGPT

IA and Web Development

Prompt Enginering

Embeddings & Fine Tuning



Open IA and ChatGPT

Artificial Inteligence **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 Inteligence **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 Inteligence **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
plugins

APIs

Microsoft

OpenAI

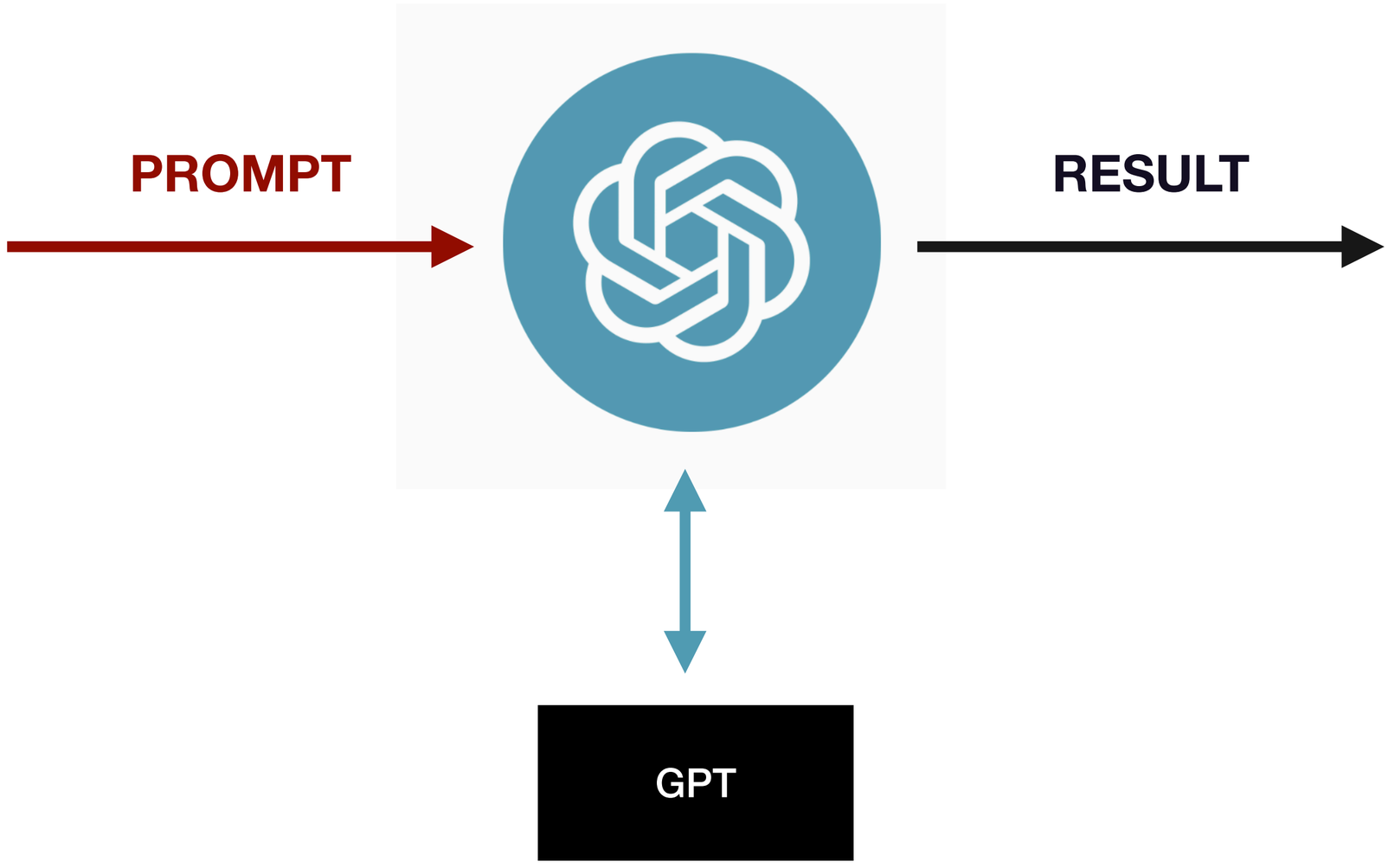
IMPORTANT

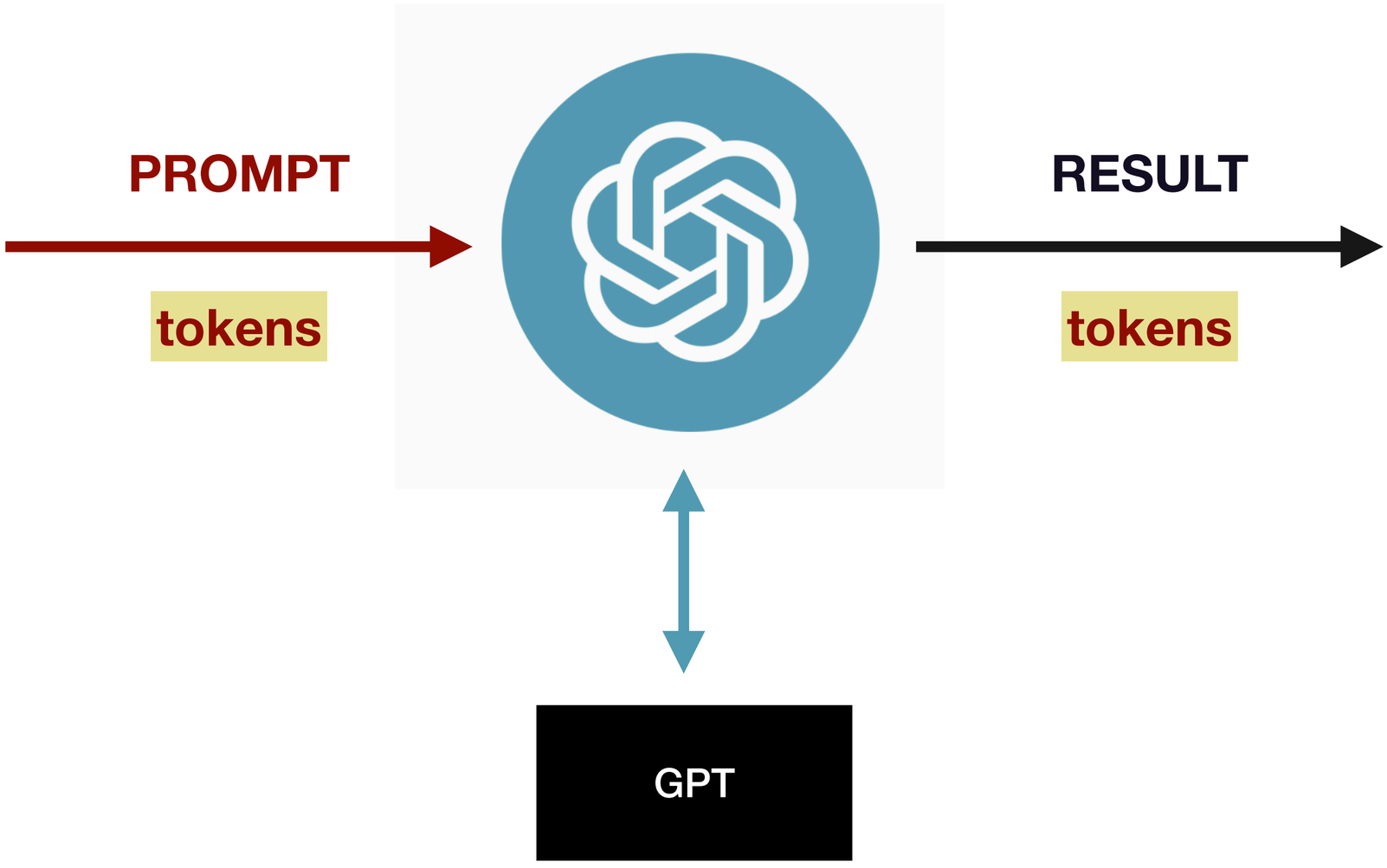
Large Language Models can be used for several tasks without changing or training models

DEFINITION

Token

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


PROMPT
RESULT
GPT


PROMPT
tokens
RESULT
tokens
GPT

Tokens and OpenAI

- OpenAI will charge our credit card based on the amount of tokens we've used
- It 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

OpenAI models

- ChatGPT or GPT 3.5
  US$2 / 1 million tokens
- GPT 4
  30x to 60x more expensive
  Higher token window
  Better for code and complex tasks
- GPT 4 image and prompt (future)
- InstructGPT models
- Image models
- Audio models

OpenAI account

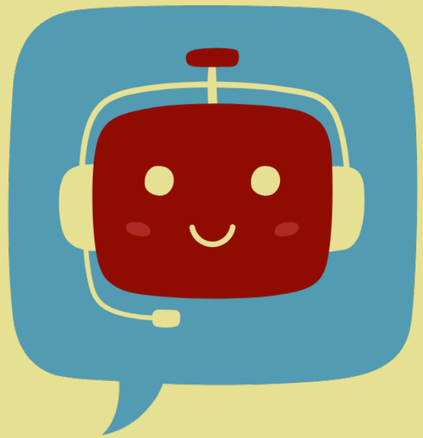
- Not the same as having a ChatGPT account
- Free to sign in
- Phone number verification
- Free credits of USD5 for 3 months (once per phone number)
- Rate limits
  - GPT 3.5, 3500 RPM
  - GPT 4, 200 RPM

IMPORTANT

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

WARNING

OpenAI GPT 3.5 and 4 models' training data cuts off in 2021, so they may not have knowledge of current events.



IA and Web Development

What can we do with GPT as web developers

INTEGRATION	PLUGINS	CONTENT	AIO
Use AI for your apps to create, transform and manipulate data or conversations, such as chatbots	Create web services that can be consumed by ChatGPT public app	Use GPT APIs to create and curate content for your website and social networks	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.

Plugins

- A new feature to ChatGPT
- 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?

USER'S PROMPT

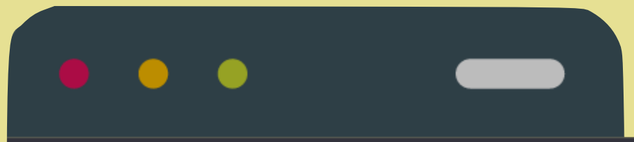
ENHANCED PROMPT

RESULT

Your Plugin API

GPT

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

```yaml
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:
```

Browser Plugin

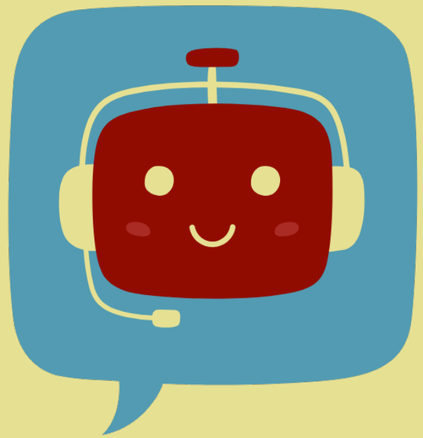
- ChatGPT can browse the web, similar to how Bing Chat works
- It honors robots.txt
- The keyword for the User Agent is ChatGPT-User
- SEO for AI, AIO?: You don't need to do anything in special
-

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



Prompt Engineering

DEFINITION

Prompt Engineering

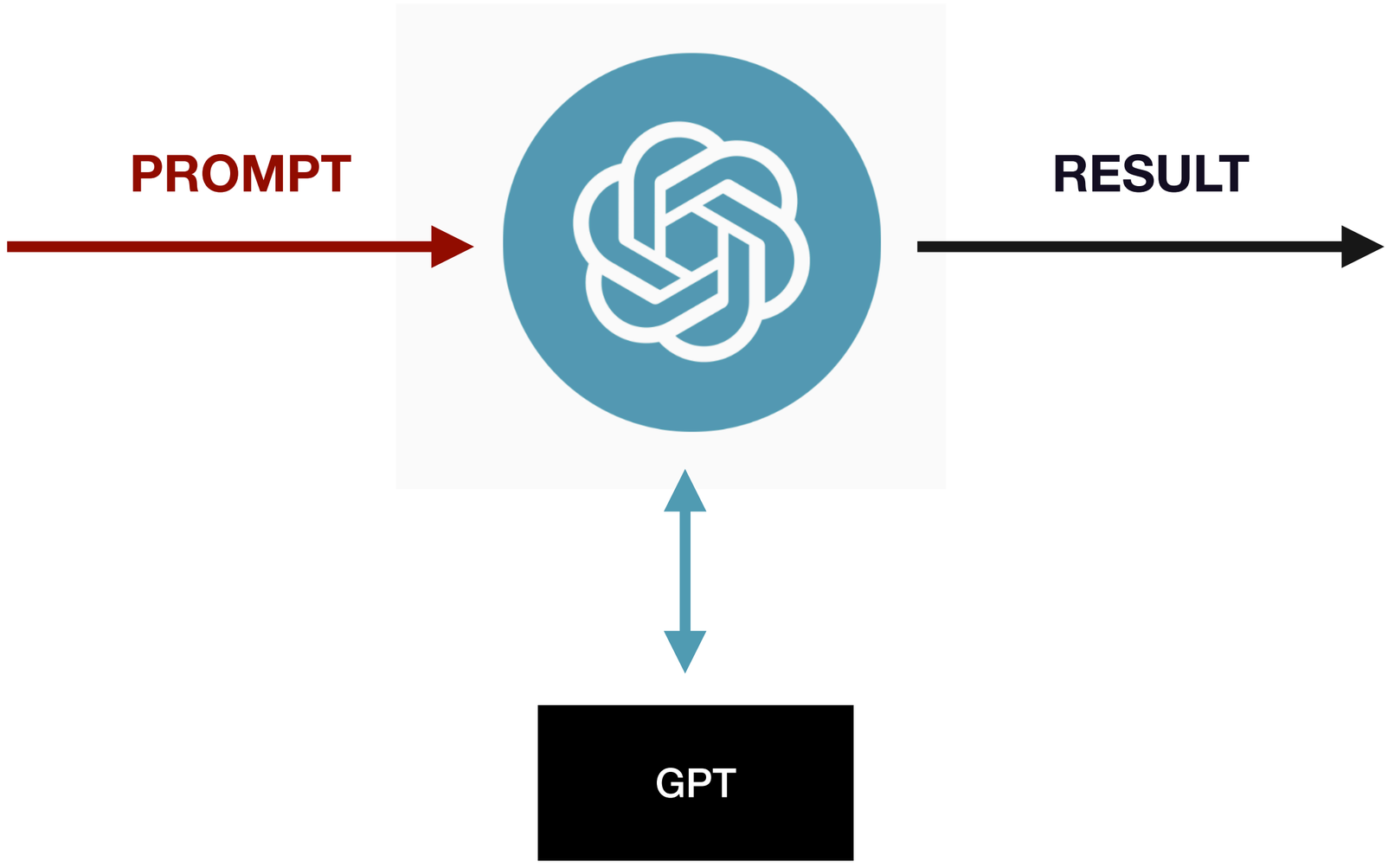
Process of designing and refining prompts or inputs for language models like GPT to generate desired outputs or responses.

WARNING

Engineering or Hacking?

IMPORTANT

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


PROMPT
RESULT
GPT

Prompt Engineering for Developers

- We want consistent and deterministics 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 that user generated content that goes into the prompt is valid
- We want to stop prompt injection

WARNING

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

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

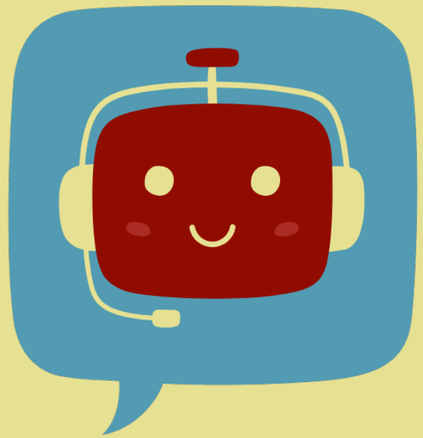
Specific and Clear Instructions

- Ask for data in a structured format you want (JSON, HTML, CSV, or any string format you need)
- 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 "Answer with 'false' in case..."

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



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.

WARNING

We don't fine-tune LLM models such as ChatGPT!

We always use it as a black box working with the prompt.

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	OWN DATA	CHAT WITH...
Create a chatbot that can answer questions as our customer service agent connecting it to the user's account	Use the power of GPT to search, transform, summarize and make decision over your private data	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 Dos

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, ...]**

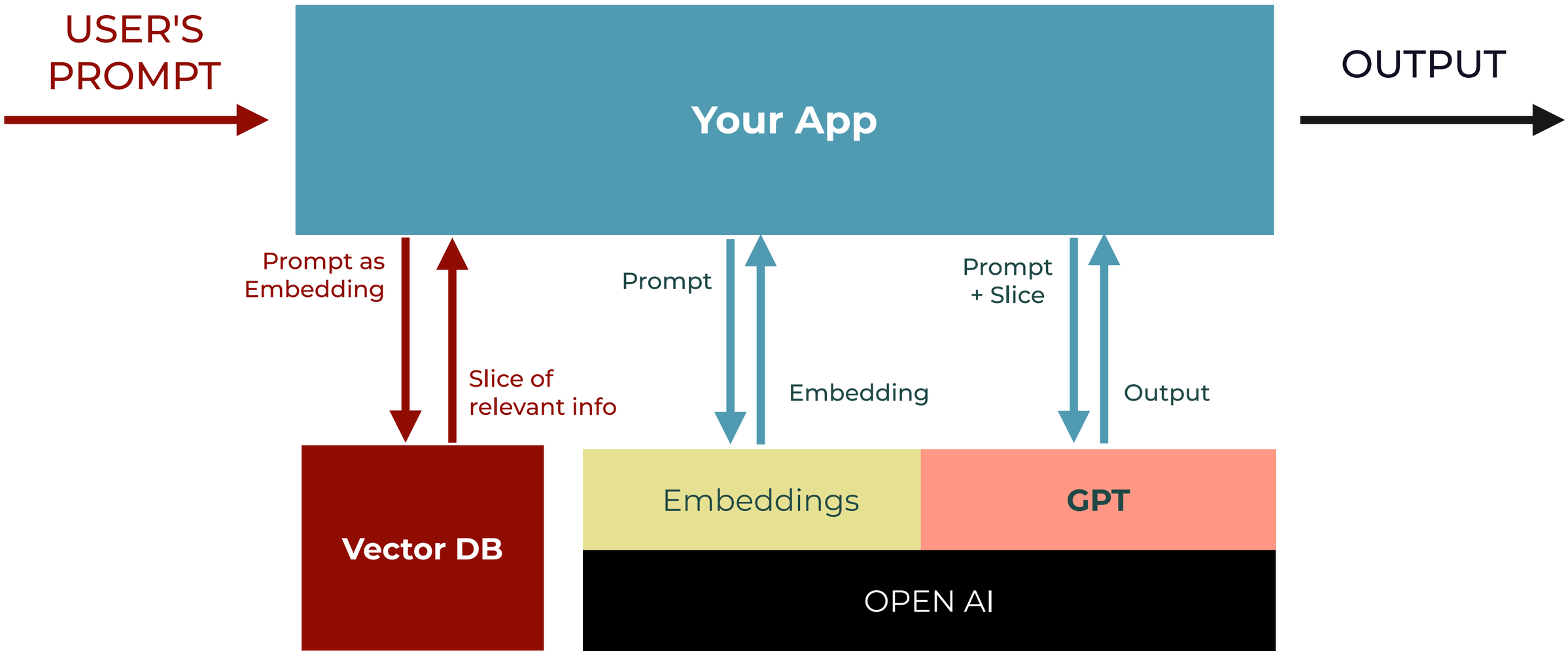
Split and Embed

- 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,e tc.
  - 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 ChatGPT

IA and Web Development

Prompt Enginering

Embeddings & Fine Tuning

THANKS! 🤗

CHATGPT FOR WEB DEVELOPERS

MAXIMILIANO FIRTMAN