ARTIFICIAL INTELLIGENCE AND LLMs

Your First AI App using ChatGPT & LaMDA/PaLM

MAXIMILIANO FIRTMAN

Build Stuff

MAXIMILIANO FIRTMAN

MOBILE+WEB DEVELOPER

HTML since 1996

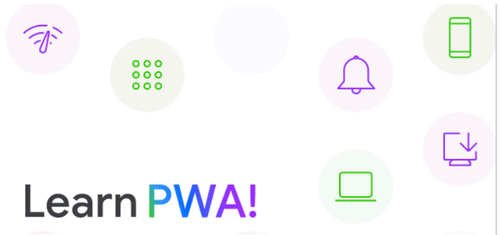
JavaScript since 1998

AUTHOR

Authored 13 books and +70 courses

Published +150 webapps

@FIRT · FIRT.DEV



Search

web.dev/learn/pwa

000

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

11:20 - 12:05 Part 1

12:05 - 12:20 ☕ BREAK

12:20 - 13:05 Part 2

Pre-requisites

Visual Studio Code

Node.js

OpenAI account

Project and Slides:

GitHub.com/firtman/buildstuff

Questions?

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...

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

:-)

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 Enginering

Creating formatted data

IMPORTANT

Same concepts can be used today with many models and providers

LLMs and [Chat]GPT

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

GPTs

plugins

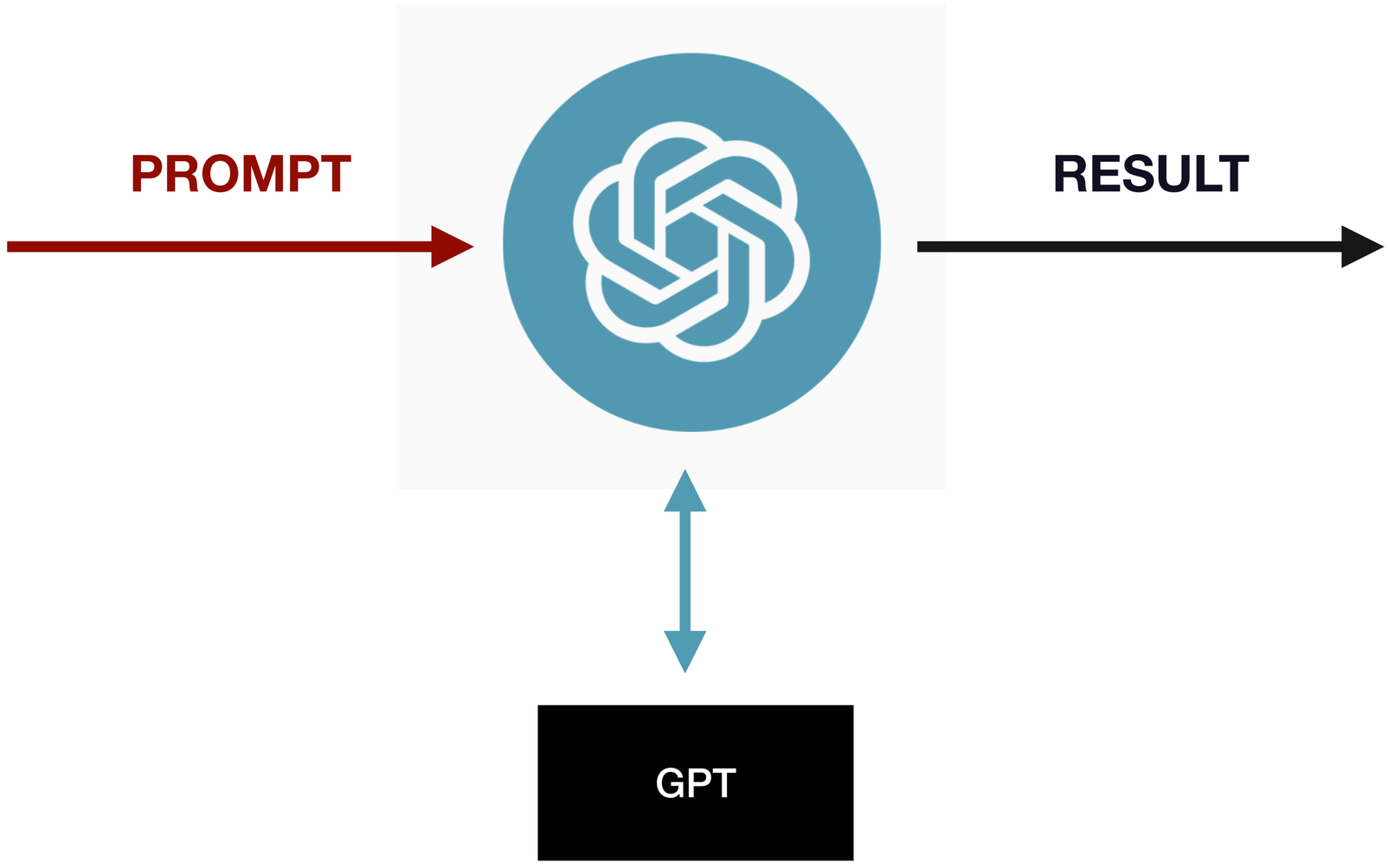
APIs

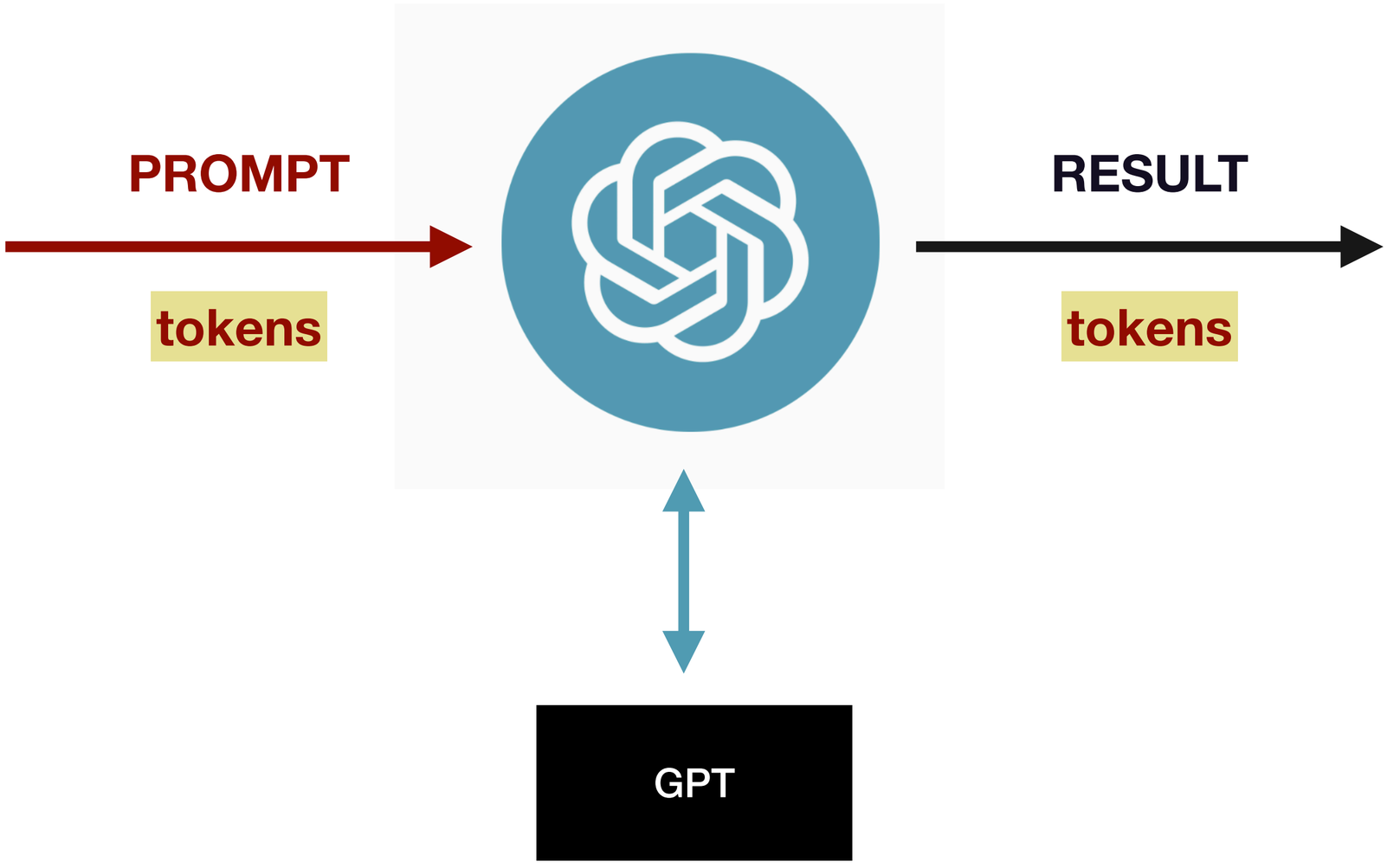
Microsoft

OpenAI

IMPORTANT

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


PROMPT
RESULT
GPT


PROMPT
tokens
RESULT
tokens
GPT

DEFINITION

Token

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

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

- 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
  Better for code and complex tasks
- GPT 4 Turbo image input mode
- Image output models (DALL-E)
- Audio models (input and output)
- Assistants API

Other Companies and Models

Meta

Llama

Google

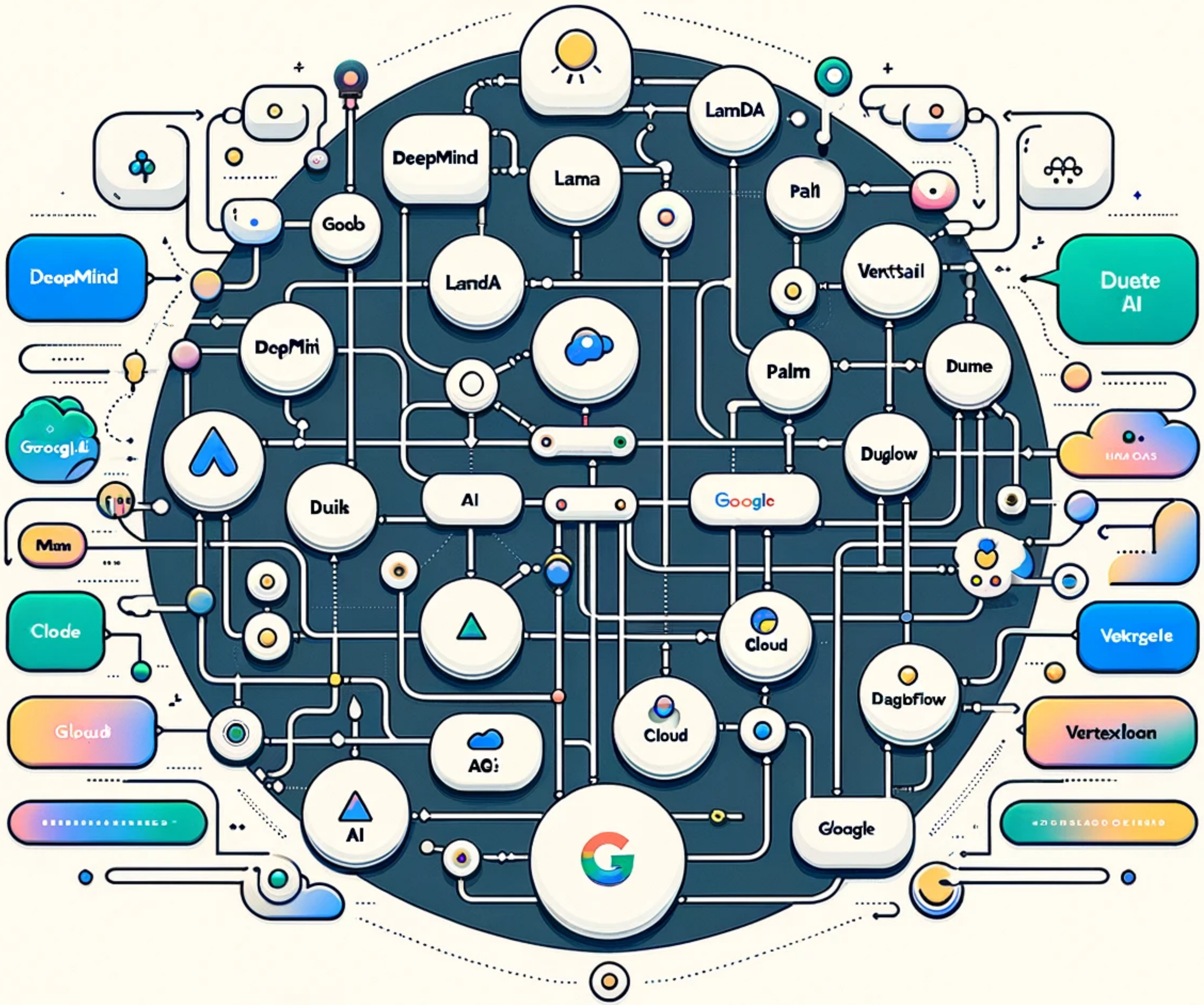
LaMDA / PaLM

Claude

Mistral AI

X? Amazon?


Google AI


LamDA
DeepMind
Lama
Pali
Goob
LandA
Vertsail
DeepMind
Duete AI
DepMiri
Palm
Dume
Duglow
Duik
AI
Google
Mum
Clode
Cloud
Vekrgele
Dagbflow
Gloud
AGi
Cloud
Vertexloan
AI
Gloagle

Google AI LLMs

- Several Models
  - LaMDA
  - **PaLM (available to the public)**
- Ways for accessing them
  - Duet AI
  - DialogFlow
  - Vertex AI
- Families: Imagen, Codey, Chirp
- Google Bard and integration with Google services available

Cloud Services for Generative AI

OpenAI	Microsoft Azure	AWS Bedrock	Google Cloud
GPT	GPT / OpenAI	Llama	PaLM
DALL-E	Llama	Claude	
Other	Hugging Face	Mistral	

Cloud Services for Generative AI

OpenAI	Microsoft Azure	AWS Bedrock	Google Cloud
GPT	GPT / OpenAI	Llama	PaLM
DALL-E	Llama	Claude	
Other	Hugging Face	Mistral	

We can use cloud-based APIs for using LLMs

RESTFul APIs

High level APIs

Multi-vendor libraries

Vercel AI SDK

Langchain

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

GPT 3.5 and 4 models' training data cuts off in Jan-2022 or Apr-2023, so they may not have knowledge of current events.

IMPORTANT

We can fine tune GPT 3.5 and 4 (preview). In this case we pay for training and then we pay more per token to use them

Google Cloud Vertex PaLM API

- Text and Chat
  - **US$0.50 / 1 million tokens**
  - Half the price of GPT 3.5 Turbo
  - Quality is not so good yet in comparison

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.

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

Repo

github.com/firtman/buildstuff

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

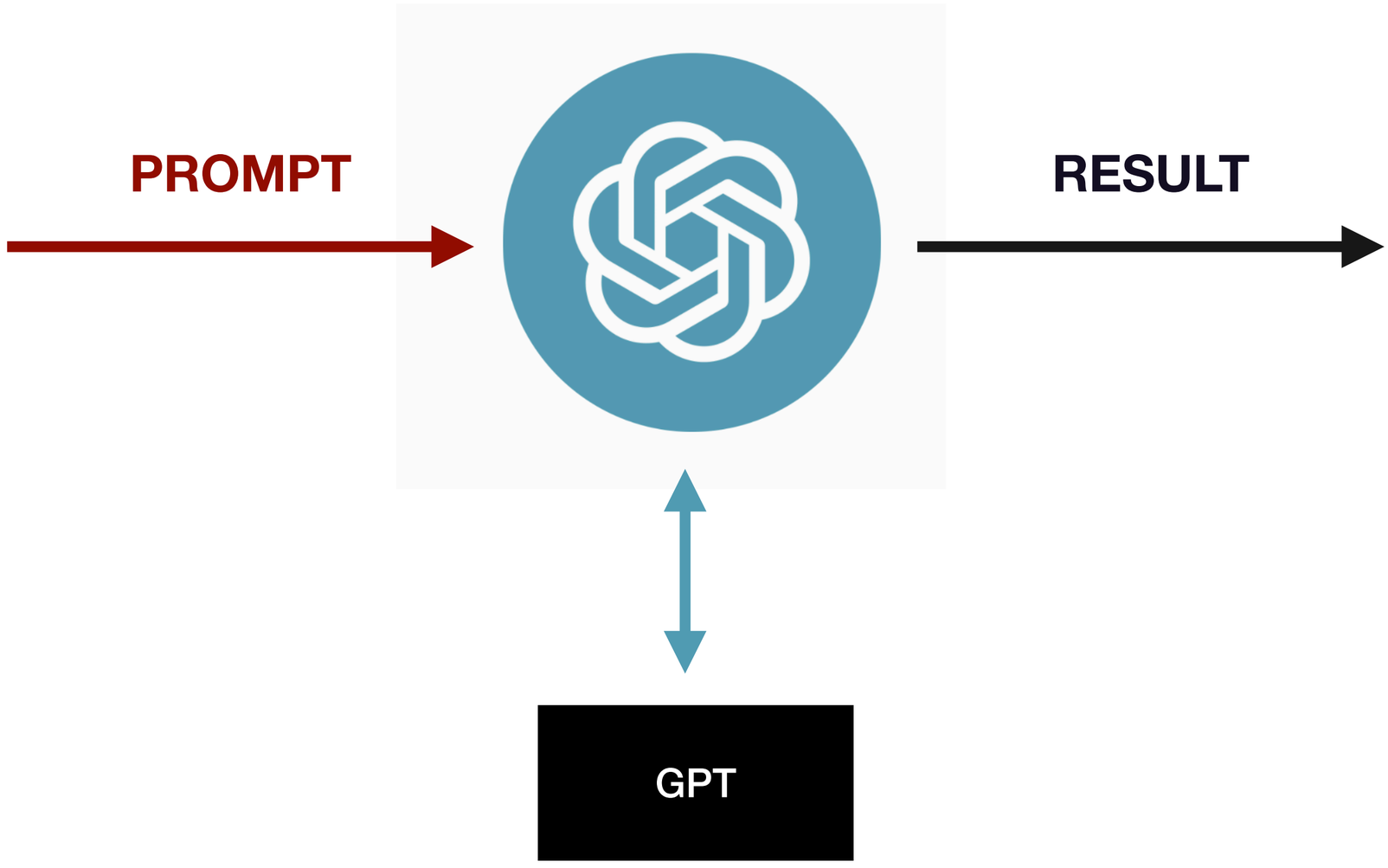
Process of designing and refining prompts or inputs for language models like GPT to generate desired outputs or responses.

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 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 %

https://twitter.com/predict_addict/status/1724381491898060816

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

What's Next

What's Next

- Fine Tuning
- Assistants - LangChain Actions
- Vector-based DBs
- Embeddings

What we've covered

OpenAI and Google APIs

AI and Web Development

Prompt Enginering

Creating formatted data

Thanks! Dėkoju! 🤗

Your First AI App using ChatGPT & LaMDA/PaLM

MAXIMILIANO FIRTMAN

Build Stuff