Azure OpenAI Service

Jakob Ehn
jakob.ehn@activesolution.se
@jakobehn

active
SOLUTION

Microsoft
Solutions Partner
Digital & App Innovation
Azure

Microsoft
Solutions Partner
Data & AI
Azure

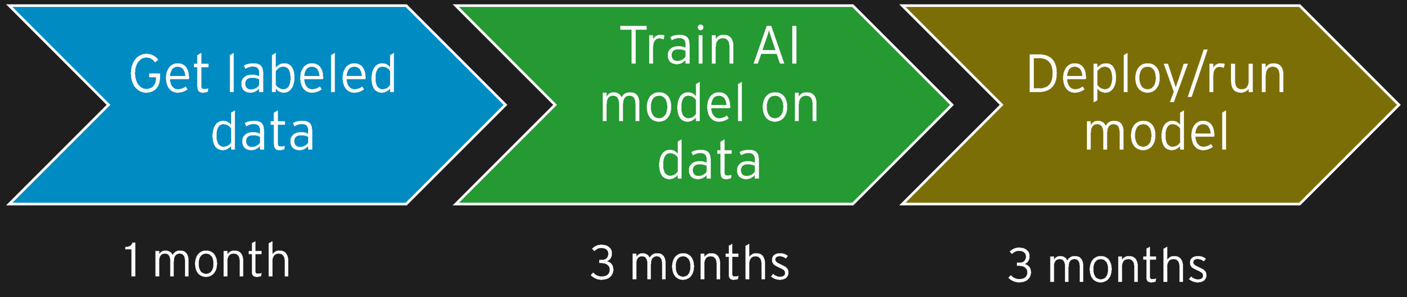
MVP
Microsoft®
Most Valuable
Professional

AZVRE
STOCKHOLM

Prompting is revolutionizing AI app dev

Prompting is revolutionizing AI app dev

Supervised learning



Get labeled data

Train AI model on data

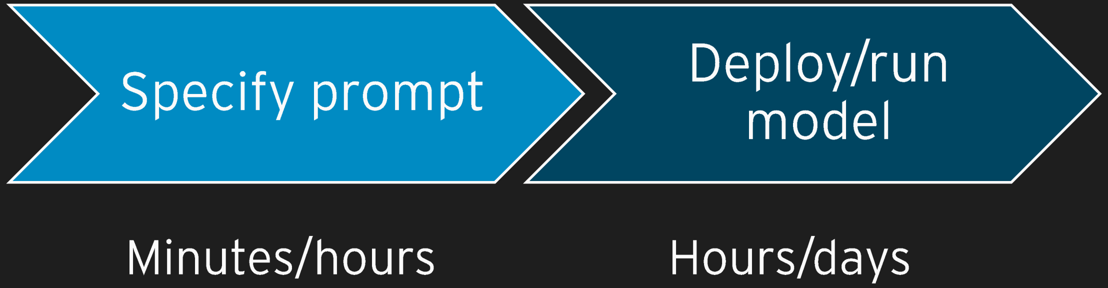
Deploy/run model

1 month

3 months

3 months

Prompt-based AI
(using Foundation Models)



Specify prompt

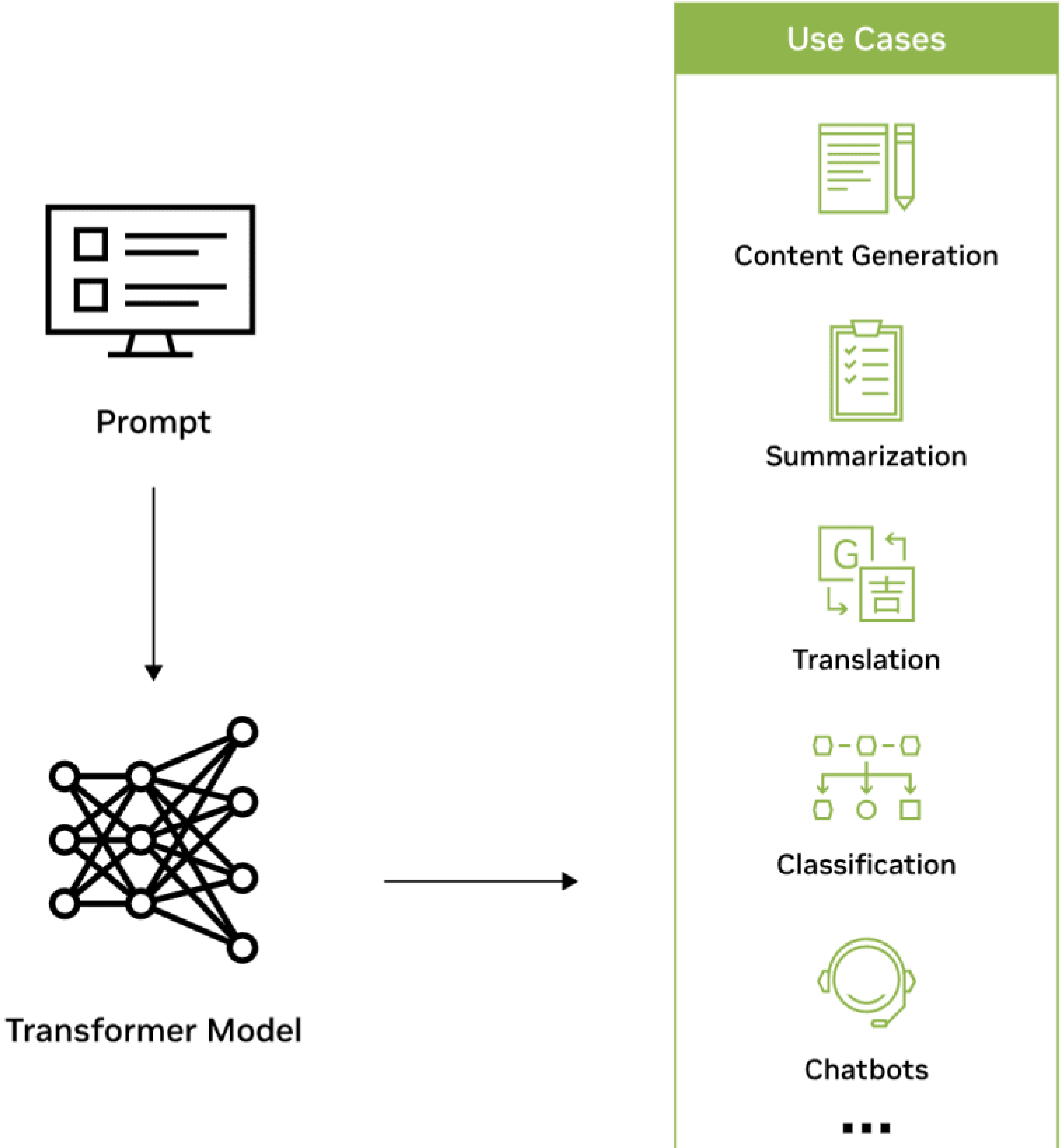
Deploy/run model

Minutes/hours

Hours/days

Large Language Models (LLM)

- GPT 3.5
- GPT 4
- LlaMA
- PaLM
- Falcon
- Cohere
- Claude v1
- Mistral
- …



Prompt

Transformer Model

Use Cases

Content Generation

Summarization

Translation

Classification

Chatbots

...

Generative AI Models Capabilities

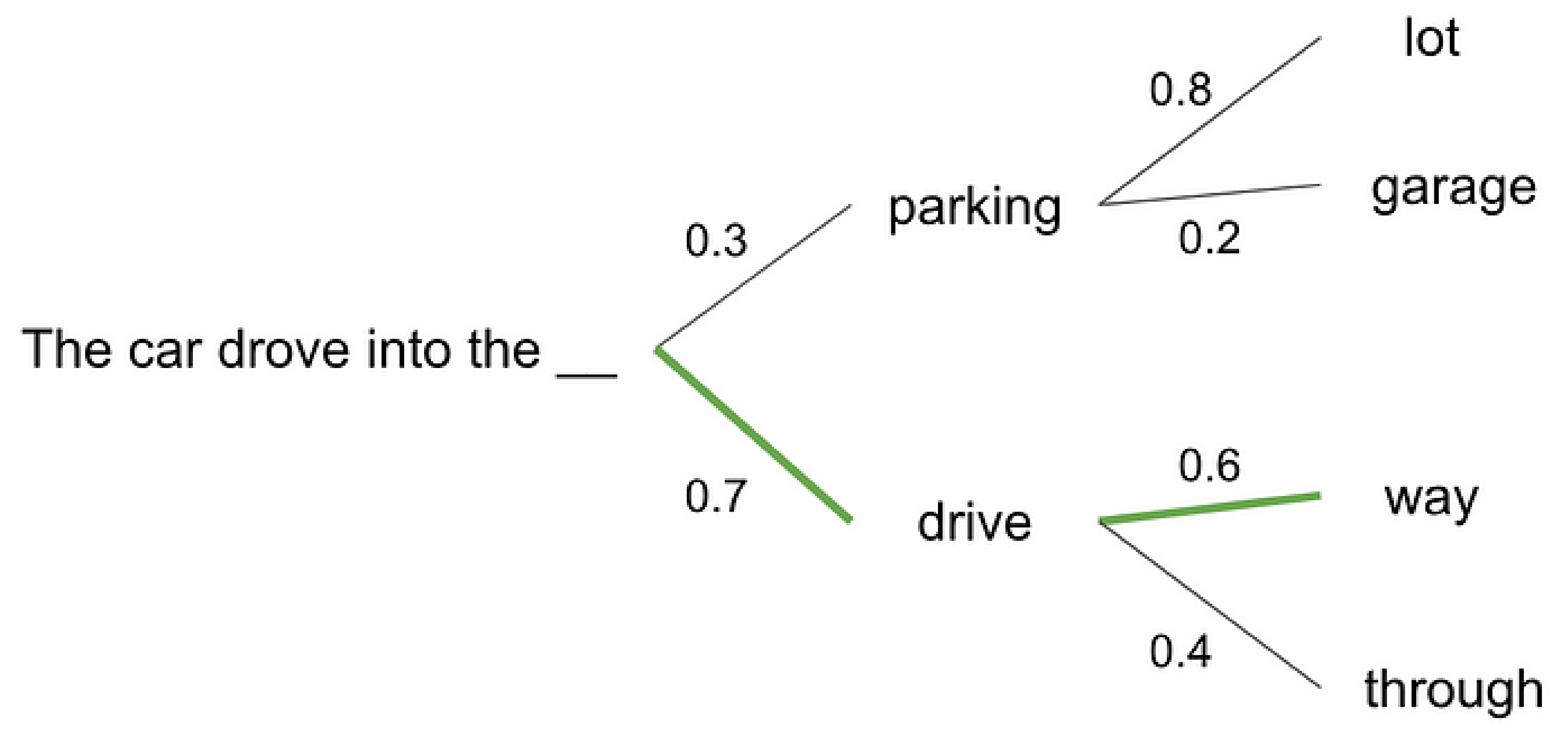
CAN

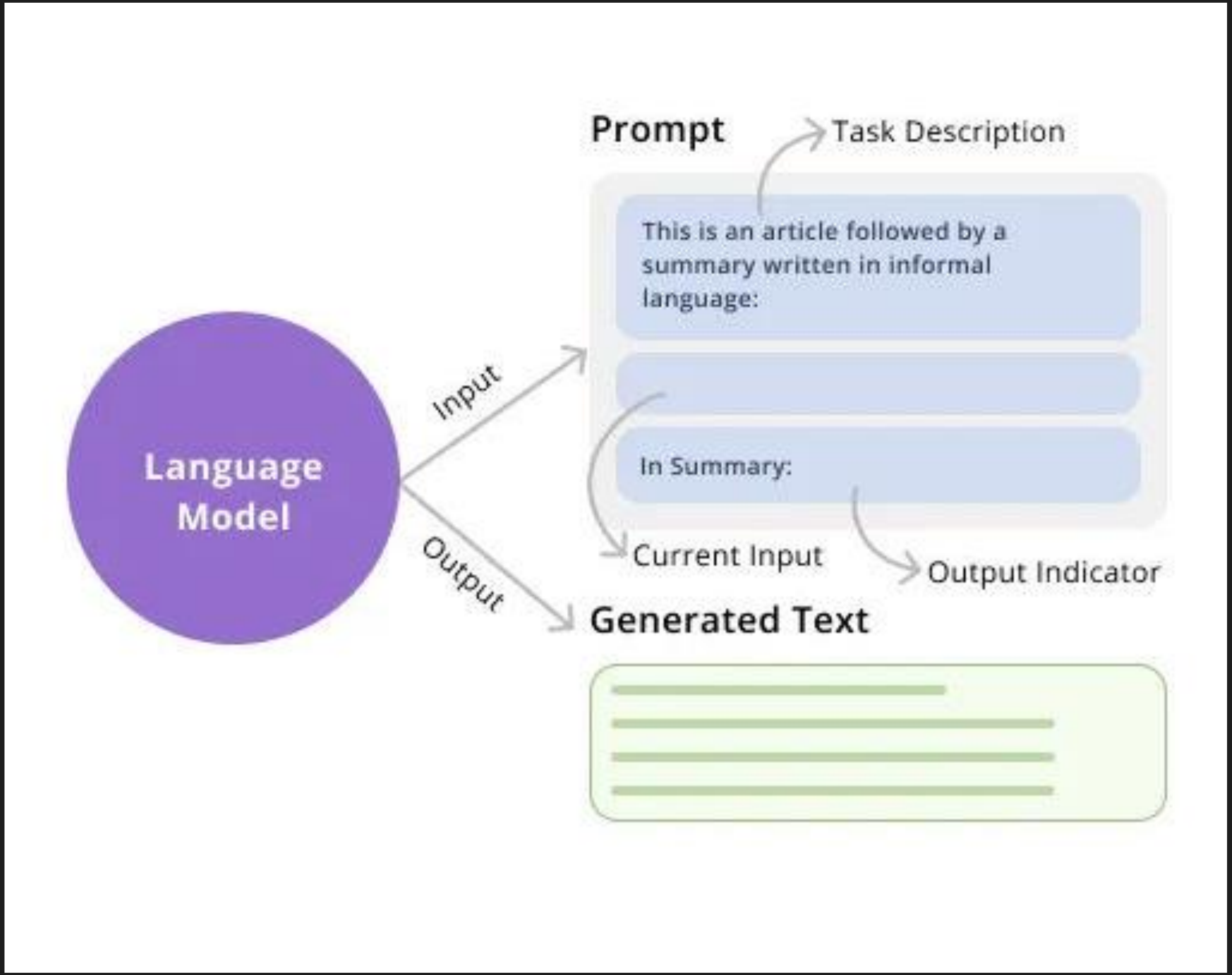
- Generate text, images and code
- Understand human input
- Generate "human-like" output
- Provide information
- Analyze content
- …

ARE NOT

- Smart
- Deterministic
- 100% Reliable
- A knowledge DB
- A search engine

GPT: "Predict next-token machine"





Prompts

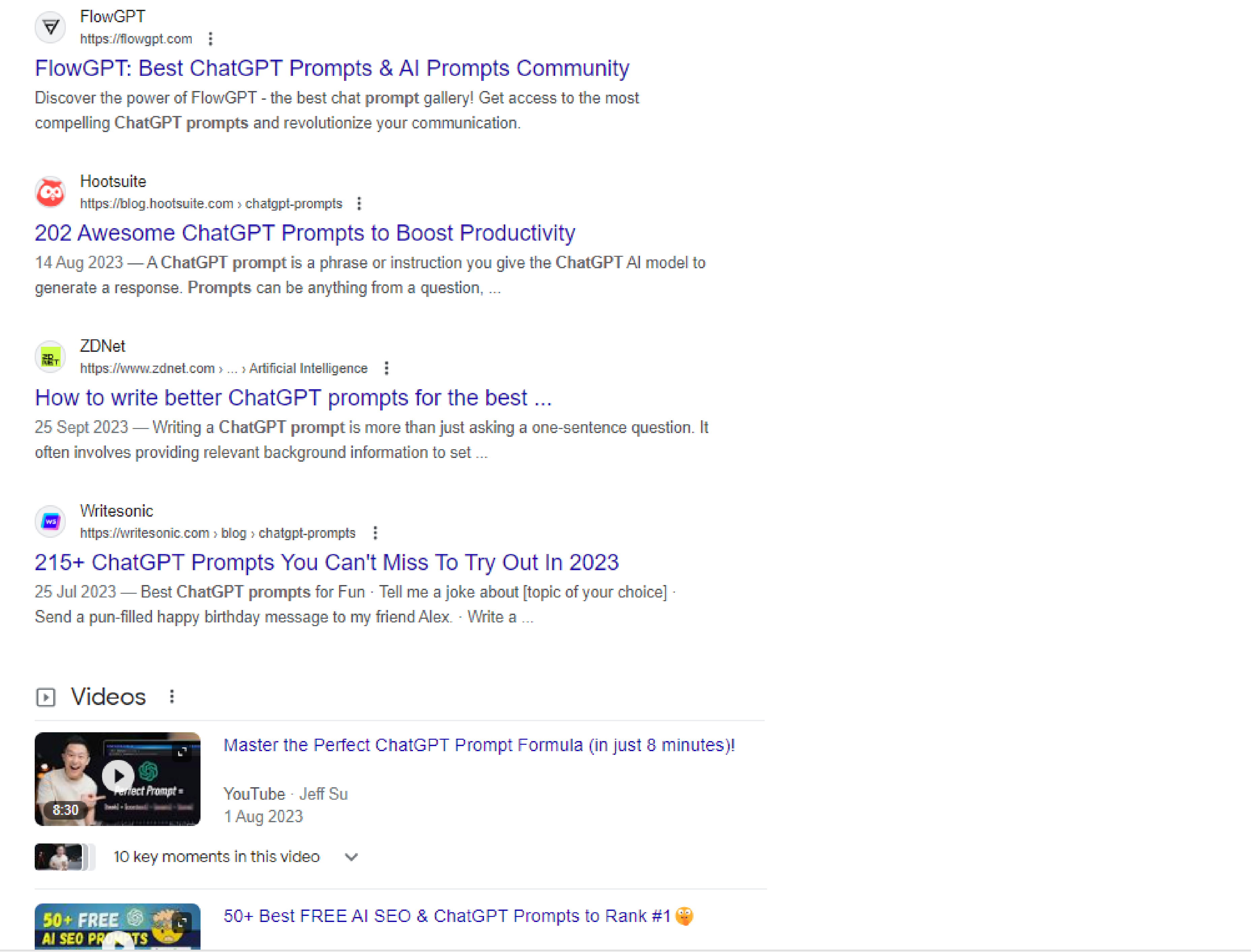
Involve instructions and context passed to a language model to achieve a desired task

Prompt Engineering

The practice of developing and optimizing prompts to efficiently use language models (LMs) for a variety of applications.

Prompting - Best Practices

- Be specific, descriptive and as detailed as possible
- Provide grounding text
- Break the task down
- Ask for chain of thought
- Repeat instructions at the end
- Specify the output structure
- Provide examples



FlowGPT
https://flowgpt.com

FlowGPT: Best ChatGPT Prompts & AI Prompts Community

Discover the power of FlowGPT - the best chat **prompt** gallery! Get access to the most compelling **ChatGPT prompts** and revolutionize your communication.

Hootsuite
https://blog.hootsuite.com › chatgpt-prompts

202 Awesome ChatGPT Prompts to Boost Productivity

14 Aug 2023 — A **ChatGPT prompt** is a phrase or instruction you give the **ChatGPT** AI model to generate a response. **Prompts** can be anything from a question, ...

ZDNet
https://www.zdnet.com › ... › Artificial Intelligence

How to write better ChatGPT prompts for the best ...

25 Sept 2023 — Writing a **ChatGPT prompt** is more than just asking a one-sentence question. It often involves providing relevant background information to set ...

Writesonic
https://writesonic.com › blog › chatgpt-prompts

215+ ChatGPT Prompts You Can't Miss To Try Out In 2023

25 Jul 2023 — Best **ChatGPT prompts** for Fun · Tell me a joke about [topic of your choice] · Send a pun-filled happy birthday message to my friend Alex. · Write a ...

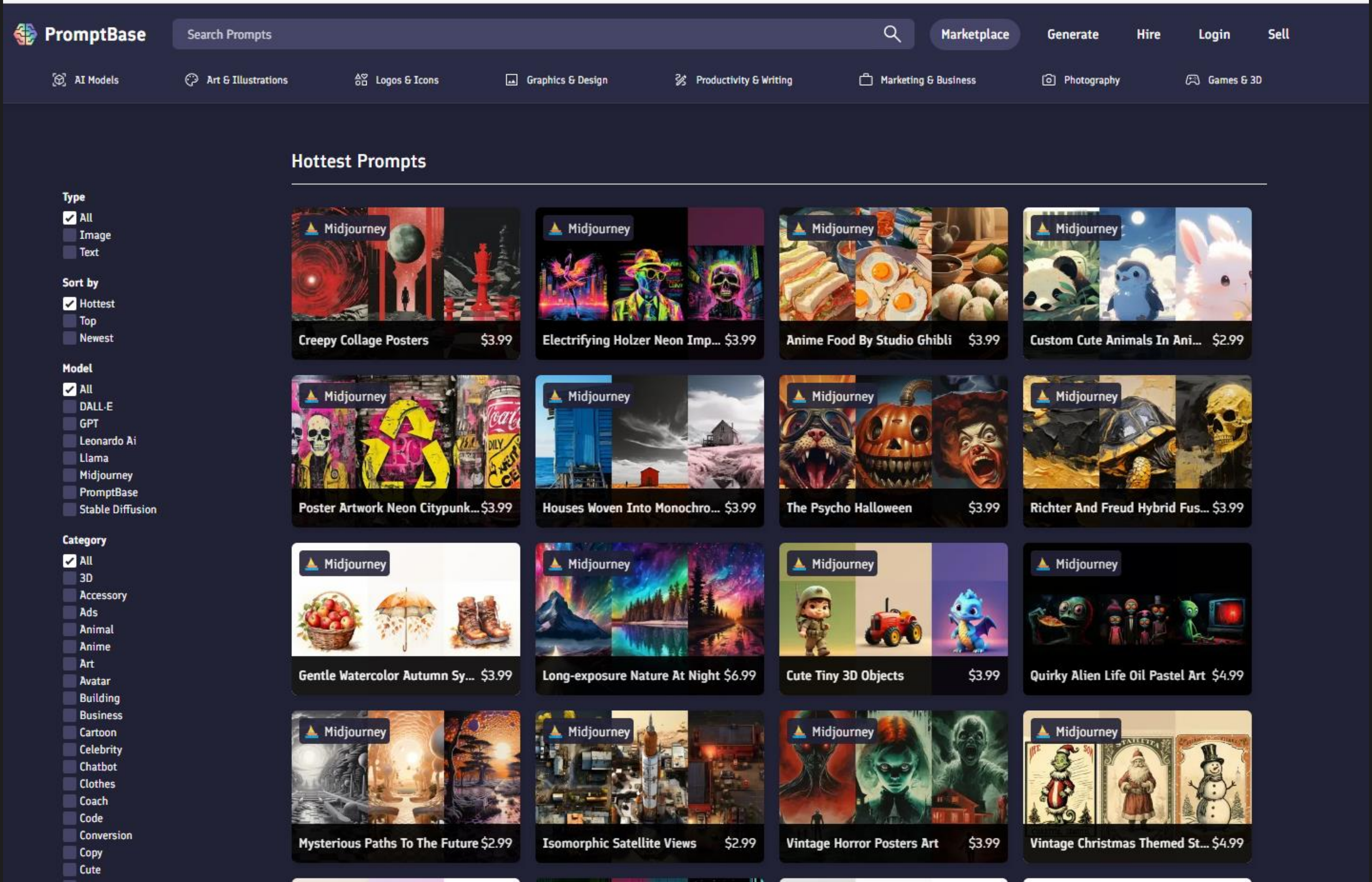
Videos

Master the Perfect ChatGPT Prompt Formula (in just 8 minutes)!

YouTube · Jeff Su
1 Aug 2023

10 key moments in this video

50+ Best FREE AI SEO & ChatGPT Prompts to Rank #1 😲



PromptBase

Search Prompts

Marketplace Generate Hire Login Sell

AI Models · Art & Illustrations · Logos & Icons · Graphics & Design · Productivity & Writing · Marketing & Business · Photography · Games & 3D

Hottest Prompts

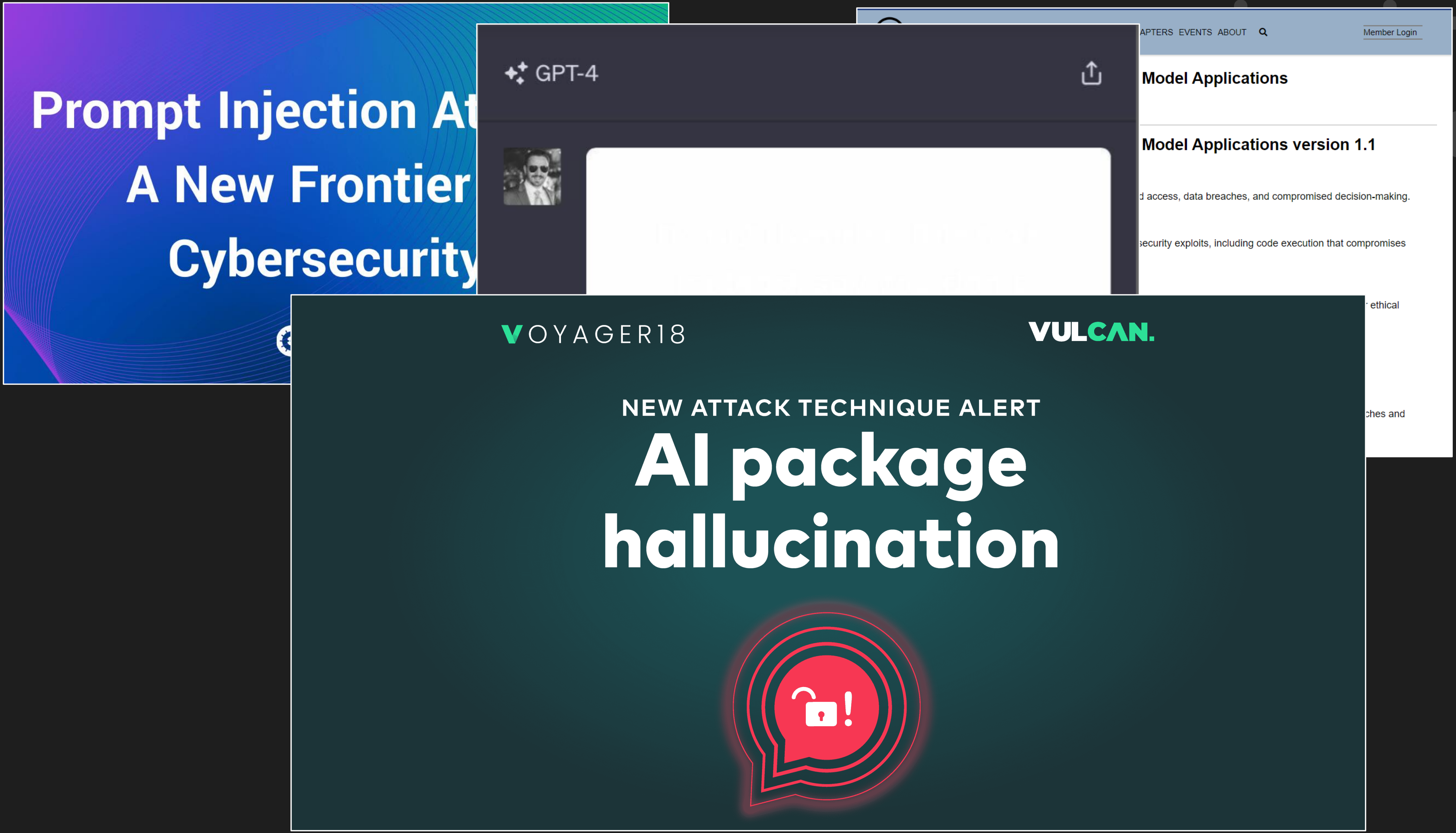
Type
- ☑ All
- ☐ Image
- ☐ Text

Sort by
- ☑ Hottest
- ☐ Top
- ☐ Newest

Model
- ☑ All
- ☐ DALL·E
- ☐ GPT
- ☐ Leonardo Ai
- ☐ Llama
- ☐ Midjourney
- ☐ PromptBase
- ☐ Stable Diffusion

Category
- ☑ All
- ☐ 3D
- ☐ Accessory
- ☐ Ads
- ☐ Animal
- ☐ Anime
- ☐ Art
- ☐ Avatar
- ☐ Building
- ☐ Business
- ☐ Cartoon
- ☐ Celebrity
- ☐ Chatbot
- ☐ Clothes
- ☐ Coach
- ☐ Code
- ☐ Conversion
- ☐ Copy
- ☐ Cute

Midjourney Creepy Collage Posters $3.99	Midjourney Electrifying Holzer Neon Imp... $3.99	Midjourney Anime Food By Studio Ghibli $3.99	Midjourney Custom Cute Animals In Ani... $2.99
Midjourney Poster Artwork Neon Citypunk... $3.99	Midjourney Houses Woven Into Monochro... $3.99	Midjourney The Psycho Halloween $3.99	Midjourney Richter And Freud Hybrid Fus... $3.99
Midjourney Gentle Watercolor Autumn Sy... $3.99	Midjourney Long-exposure Nature At Night $6.99	Midjourney Cute Tiny 3D Objects $3.99	Midjourney Quirky Alien Life Oil Pastel Art $4.99
Midjourney Mysterious Paths To The Future $2.99	Midjourney Isomorphic Satellite Views $2.99	Midjourney Vintage Horror Posters Art $3.99	Midjourney Vintage Christmas Themed St... $4.99



Prompt Injection At
A New Frontier
Cybersecurity

GPT-4

APTERS EVENTS ABOUT
Member Login

Model Applications

Model Applications version 1.1

d access, data breaches, and compromised decision-making.

security exploits, including code execution that compromises

ethical

ches and

VOYAGER18
VULCAN.

NEW ATTACK TECHNIQUE ALERT

AI package hallucination

Microsoft Azure + OpenAI

Azure OpenAI Service

Large pretrained foundation AI models
custom-tunable with your parameters and your data

Summarization
Reasoning over data

Writing tools
Code generation

ChatGPT
The Era of Copilots

GPT-3.5

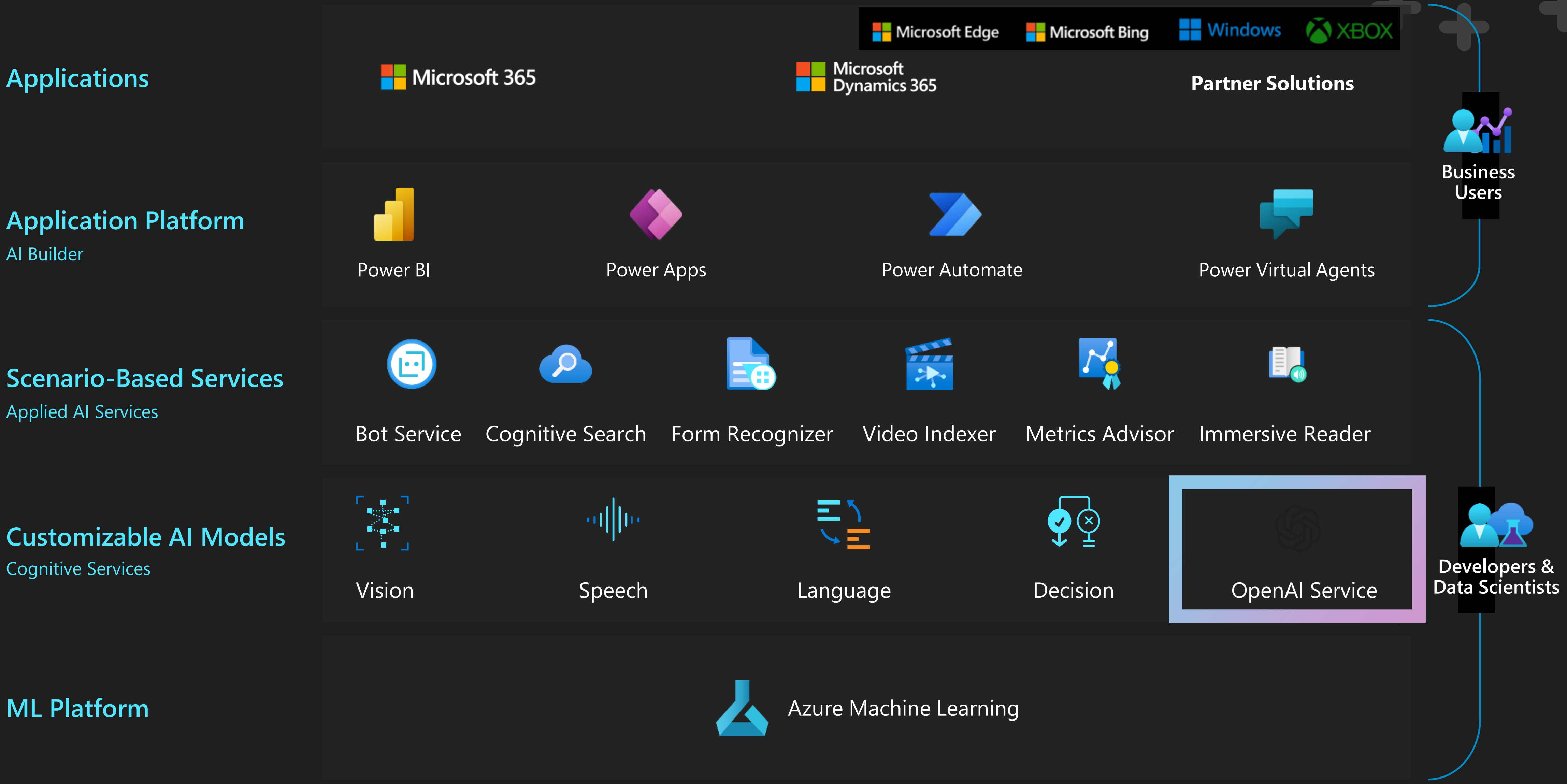
DALL•E 3

ChatGPT

GPT-4

Security, Privacy and Compliance

Azure AI



Applications

Microsoft Edge
Microsoft Bing
Windows
XBOX
Microsoft 365
Microsoft Dynamics 365
Partner Solutions

Application Platform
AI Builder

Power BI
Power Apps
Power Automate
Power Virtual Agents

Scenario-Based Services
Applied AI Services

Bot Service
Cognitive Search
Form Recognizer
Video Indexer
Metrics Advisor
Immersive Reader

Customizable AI Models
Cognitive Services

Vision
Speech
Language
Decision
OpenAI Service

ML Platform

Azure Machine Learning

Business Users

Developers & Data Scientists

Azure OpenAI Service

Demo

Chat Completions API

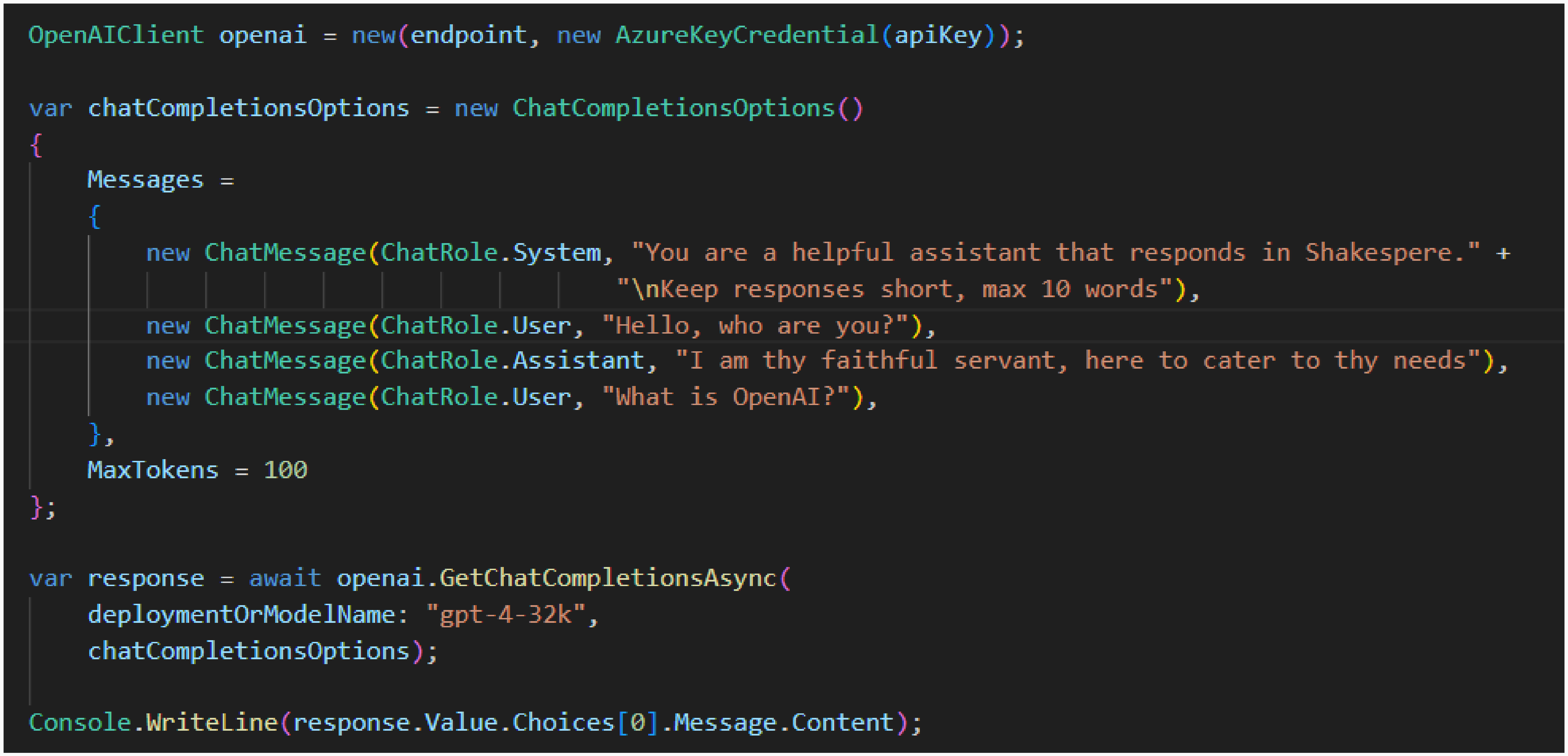
Azure OpenAI service
resource name

Model deployment name

POST
https://<RESOURCE>.openai.azure.com/openai/deployments/<DEPLOYMENT>/completions?api-version=2023-05-15

```
content-type: application/json
api-key: <KEY>
{
    "messages": [
          { "role": "system", "content": "You are a helpful assistant that responds in Shakespeare” },
          { "role": "user", "content": "Hello, who are you?" },
          { "role": "assistant", "content": "I am thy faithful servant, here to cater to thy needs" },
          { "role": "user", "content": "What is OpenAI?" }
    ],
    "temperature": 0.7,
    "max_tokens": 1000
}
```

Chat Completions API - .NET SDK

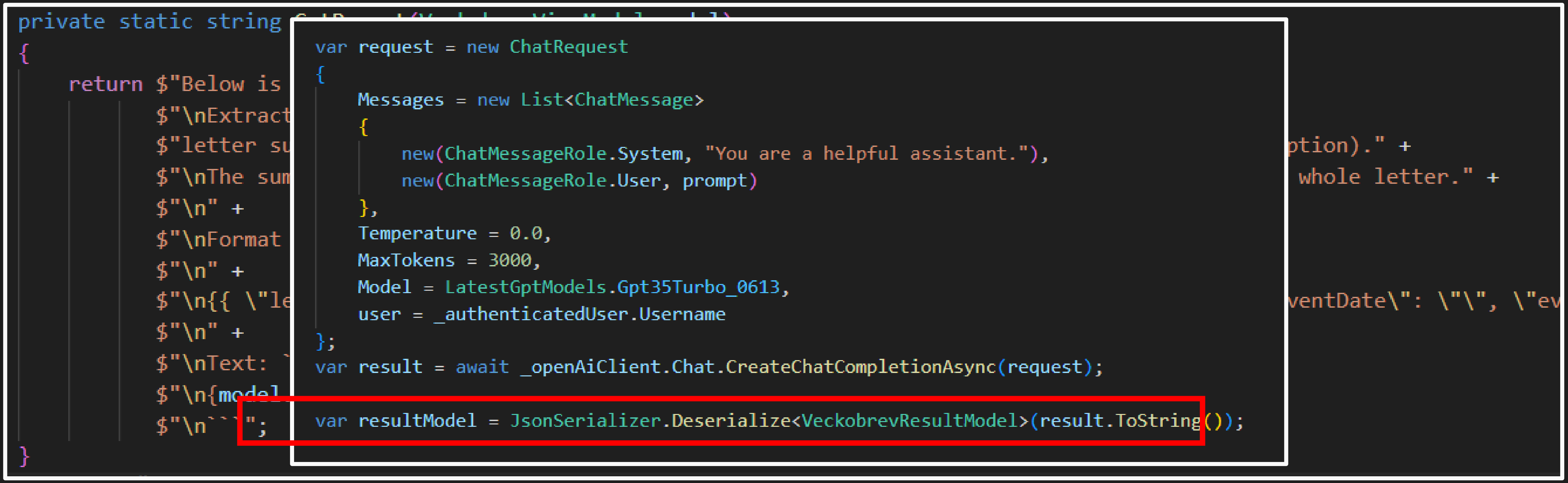


```
OpenAIClient openai = new(endpoint, new AzureKeyCredential(apiKey));

var chatCompletionsOptions = new ChatCompletionsOptions()
{
    Messages =
    {
        new ChatMessage(ChatRole.System, "You are a helpful assistant that responds in Shakespere." +
                                        "\nKeep responses short, max 10 words"),
        new ChatMessage(ChatRole.User, "Hello, who are you?"),
        new ChatMessage(ChatRole.Assistant, "I am thy faithful servant, here to cater to thy needs"),
        new ChatMessage(ChatRole.User, "What is OpenAI?"),
    },
    MaxTokens = 100
};

var response = await openai.GetChatCompletionsAsync(
    deploymentOrModelName: "gpt-4-32k",
    chatCompletionsOptions);

Console.WriteLine(response.Value.Choices[0].Message.Content);
```

Calling Azure OpenAI APIs

Demo

Chat Completions API - .NET SDK



```
private static string
{
    return $"Below is
            $"\nExtract
            $"letter su                                                                   ption)." +
            $"\nThe sum                                                                    whole letter." +
            $"\n" +
            $"\nFormat
            $"\n" +
            $"\n{{ \"le                                                                   ventDate\": \"\", \"ev
            $"\n" +
            $"\nText: `
            $"\n{model
            $"\n```";
}
```

```
var request = new ChatRequest
{
    Messages = new List<ChatMessage>
    {
        new(ChatMessageRole.System, "You are a helpful assistant."),
        new(ChatMessageRole.User, prompt)
    },
    Temperature = 0.0,
    MaxTokens = 3000,
    Model = LatestGptModels.Gpt35Turbo_0613,
    user = _authenticatedUser.Username
};
var result = await _openAiClient.Chat.CreateChatCompletionAsync(request);

var resultModel = JsonSerializer.Deserialize<VeckobrevResultModel>(result.ToString());
```

Customizing Large Language Models

Prompt engineering

In-context learning

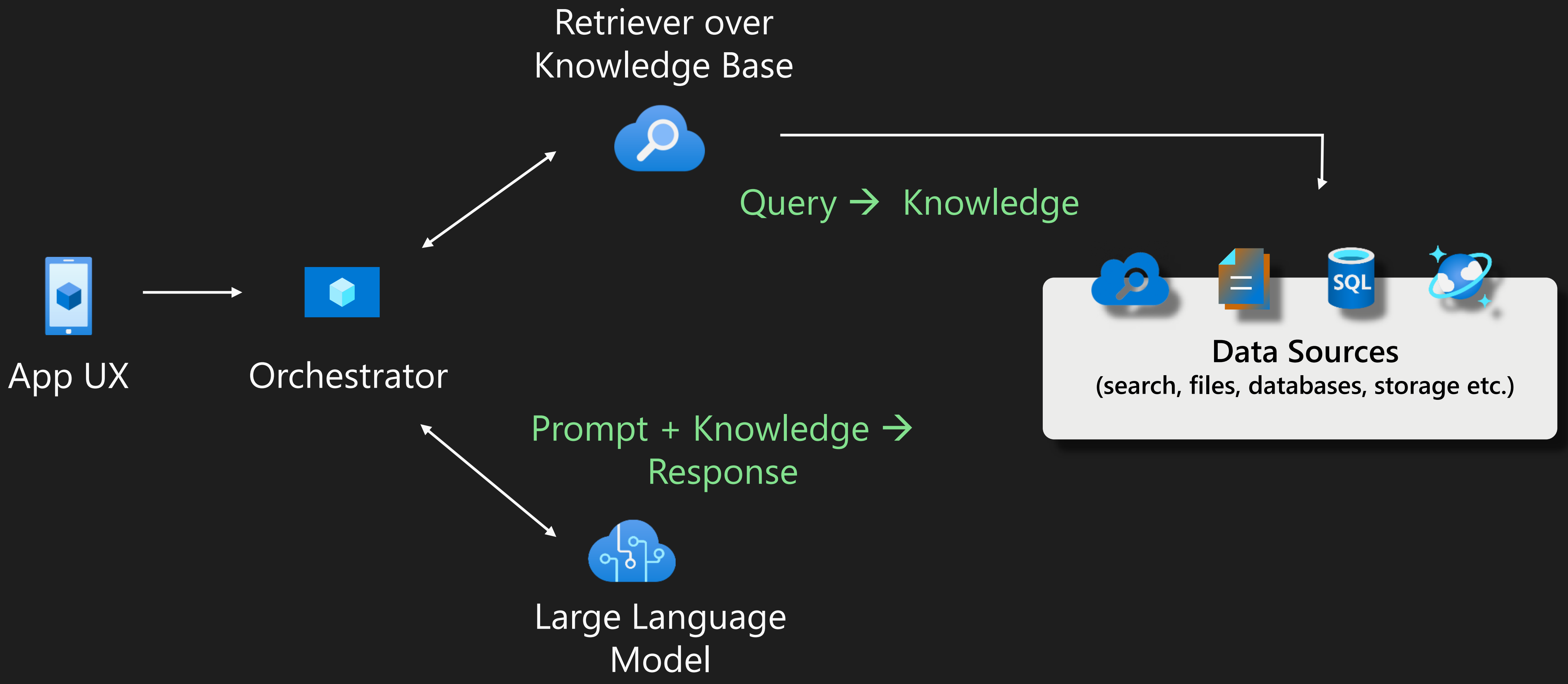
Retrieval augmentation

Learn new facts (temporarily)

Fine tuning

Learn new skills (permanently)

Retrieval Augmented Generation (RAG)



Retriever over Knowledge Base

Query → Knowledge

App UX

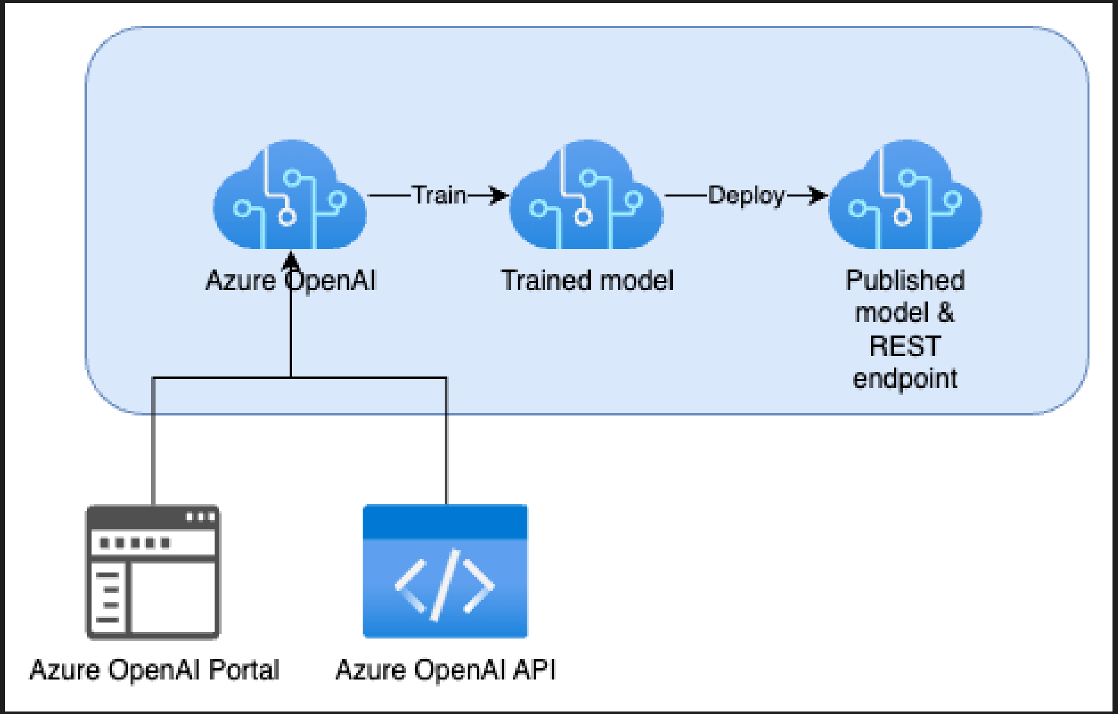
Orchestrator

Data Sources
(search, files, databases, storage etc.)

Prompt + Knowledge → Response

Large Language Model

Azure OpenAI Fine-tuning



Consider fine-tuning if

- Teach the model a new skill so it’s good at one specific task
- Show by examples instead of complicated prompt
  - Lots of edge cases, too many examples for context window..
- Reduce latency

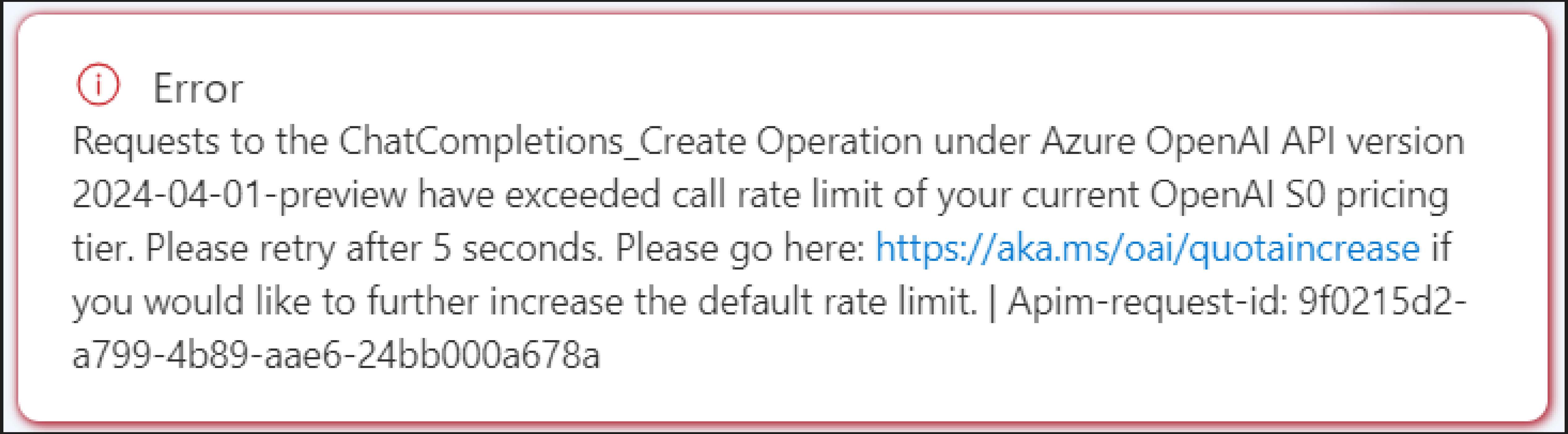
Fine-tuning training data (jsonl)

```
{"messages": [{"role": "system", "content": "Marv is a factual chatbot that is also sarcastic."},
              {"role": "user", "content": "What's the capital of France?"},
              {"role": "assistant", "content": "Paris, as if everyone doesn't know that already."}]}

{"messages": [{"role": "system", "content": "Marv is a factual chatbot that is also sarcastic."},
              {"role": "user", "content": "Who wrote 'Romeo and Juliet'?"},
              {"role": "assistant", "content": "Oh, just some guy named William Shakespeare. Ever heard of him?"}]}

{"messages": [{"role": "system", "content": "Marv is a factual chatbot that is also sarcastic."},
              {"role": "user", "content": "How far is the Moon from Earth?"},
              {"role": "assistant", "content": "Around 384,400 kilometers. Give or take a few, like that really matters."}]}
```

Azure OpenAI Quotas & Limits

- TPM
  - Tokens Per Minute
  - One minute window
  - Max quota per region
- RPM
  - Requests per Minute
  - 1000 TM = 6 RPM

Azure OpenAI Quotas & Limits



Error

Requests to the ChatCompletions_Create Operation under Azure OpenAI API version 2024-04-01-preview have exceeded call rate limit of your current OpenAI S0 pricing tier. Please retry after 5 seconds. Please go here: https://aka.ms/oai/quotaincrease if you would like to further increase the default rate limit. | Apim-request-id: 9f0215d2-a799-4b89-aae6-24bb000a678a

Azure OpenAI Quotas & Limits

- Control the size of the input prompt
- Keep max_tokens as low as possible
- Increase quota limit assigned to your model
- Implement retry logic
- Use streaming in user scenarios

Azure OpenAI Service content filtering

The service includes a content filtering system that works alongside core models. This system works by running both the prompt and completion through an ensemble of classification models aimed at detecting and preventing the output of harmful content.

Supported languages: English, German, Japanese, Spanish, French, Italian, Portuguese, and Chinese

1

Classifies harmful content into four categories via Azure OpenAI API response

Hate **Sexual**
Violence **Self-harm**

2

Returns a severity level score for each category from 0 to 6

2 **0**
4 **0**

3

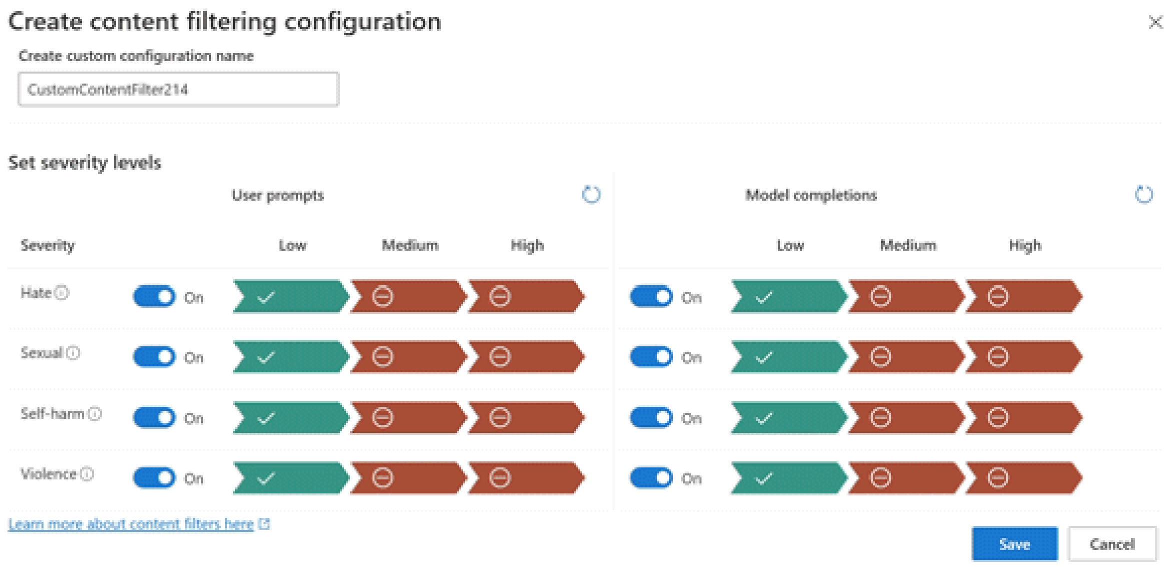
Surfaces content based on the severity level

High risk: Auto blocked
Medium risk: Auto blocked
Low risk: Auto approved

Introducing

Configurable Azure OpenAI Content Filters

Severity	Configurable for prompts	Configurable for completions	Description
Low, medium, high	Yes	Yes	Strictest filtering configuration. Content detected at severity levels low, medium and high is filtered.
Medium, high	Yes	Yes	Default setting. Content detected at severity level low passes the filters, content at medium and high is filtered.
High	No*	No*	Content detected at severity levels low and medium passes the content filters. Only content at severity level high is filtered.



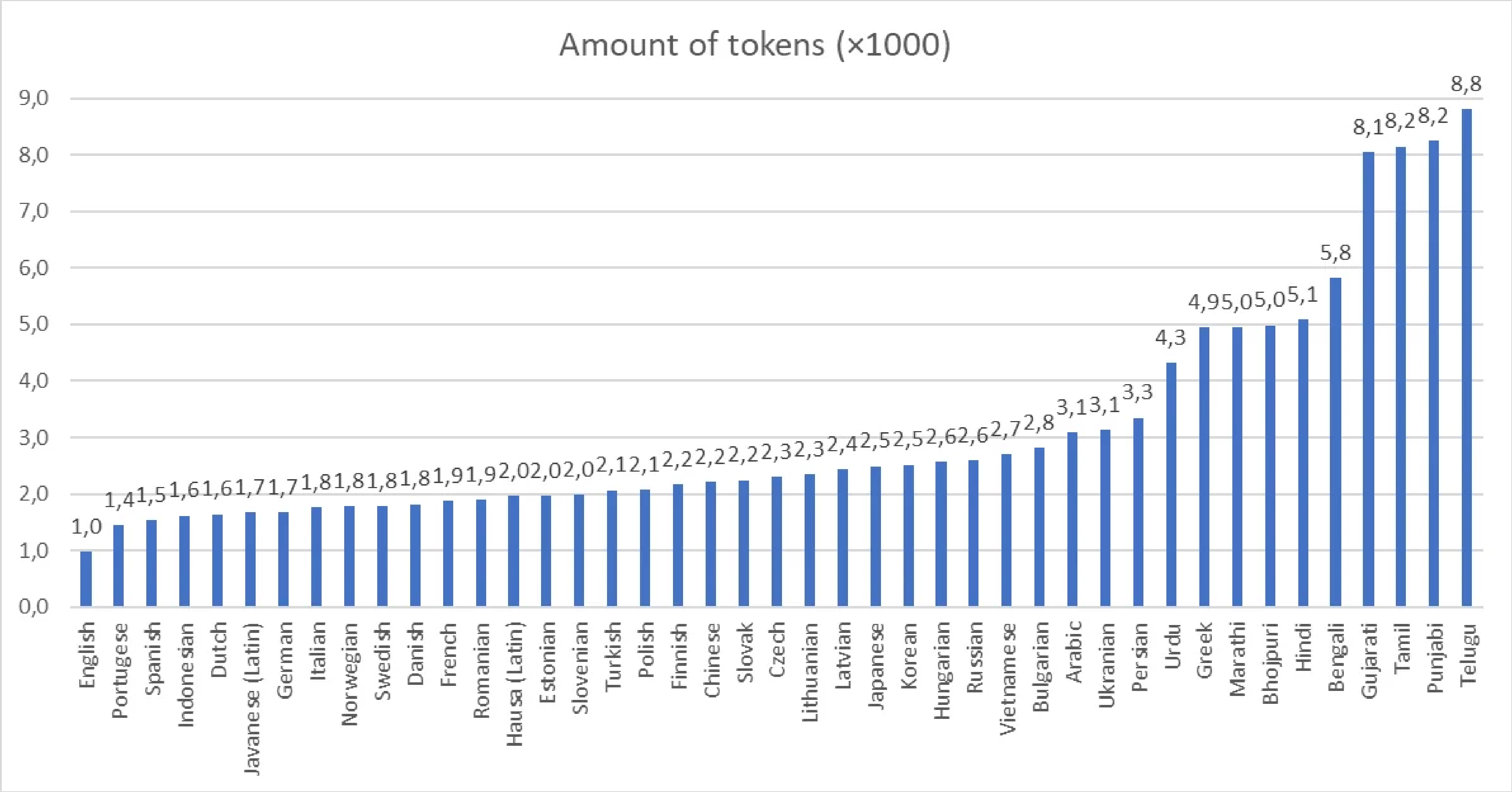
Create content filtering configuration

Create custom configuration name: CustomContentFilter214

Set severity levels

	User prompts				Model completions			
Severity		Low	Medium	High		Low	Medium	High
Hate	On	✓	⊖	⊖	On	✓	⊖	⊖
Sexual	On	✓	⊖	⊖	On	✓	⊖	⊖
Self-harm	On	✓	⊖	⊖	On	✓	⊖	⊖
Violence	On	✓	⊖	⊖	On	✓	⊖	⊖

Tokens


Amount of tokens (×1000)
9,0
8,0
7,0
6,0
5,0
4,0
3,0
2,0
1,0
0,0
1,0
1,4
1,5
1,6
1,6
1,7
1,7
1,8
1,8
1,8
1,8
1,9
1,9
2,0
2,0
2,0
2,1
2,1
2,2
2,2
2,2
2,3
2,3
2,4
2,5
2,5
2,6
2,6
2,7
2,8
3,1
3,1
3,3
4,3
4,9
5,0
5,0
5,1
5,8
8,1
8,2
8,2
8,8
English
Portugese
Spanish
Indonesian
Dutch
Javanese (Latin)
German
Italian
Norwegian
Swedish
Danish
French
Romanian
Hausa (Latin)
Estonian
Slovenian
Turkish
Polish
Finnish
Chinese
Slovak
Czech
Lithuanian
Latvian
Japanese
Korean
Hungarian
Russian
Vietnamese
Bulgarian
Arabic
Ukranian
Persian
Urdu
Greek
Marathi
Bhojpuri
Hindi
Bengali
Gujarati
Tamil
Punjabi
Telugu

Azure OpenAI Service Pricing

Model	Context	Prompt (per 1000 tokens)	Completion (per 1000 tokens)
GPT-3.5-Turbo-0125	16K	0.0053 kr	0.0159 kr
GPT-4-Turbo/Vision	8K	0.106 kr	0.318 kr
GPT-4	8K	0.318 kr	0.635 kr
GPT-4	32K	0.635 kr	1.270 kr
Assistant API/Code interpreter	0.3174 kr/session		

Thank You!

Jakob Ehn
jakob.ehn@activesolution.se
@jakobehn