Building intelligent and magic-like solutions using Azure OpenAI Services

JUSSI ROINE – aka.ms/jussi

CEO at NOT BAD SECURITY

Novice broccoli enthusiast

CTRL+ALT+AZURE PODCAST

MVP Microsoft® Most Valuable Professional

Thanks to our sponsors!

collabdays BELGIUM

Platinum

AvePoint

The Flow
Rencore

Gold

AMEXIO
ECM & CCM COMPANY

inetum realdolmen
Positive digital flow

ORDINA

spikes

ventigrate

Xylos

Silver

advantive
innovative people valuable solutions

BC/NAV TECH DAYS
mibuso.com

CTG

Docubird™

ESPC

SharePint

Kianda

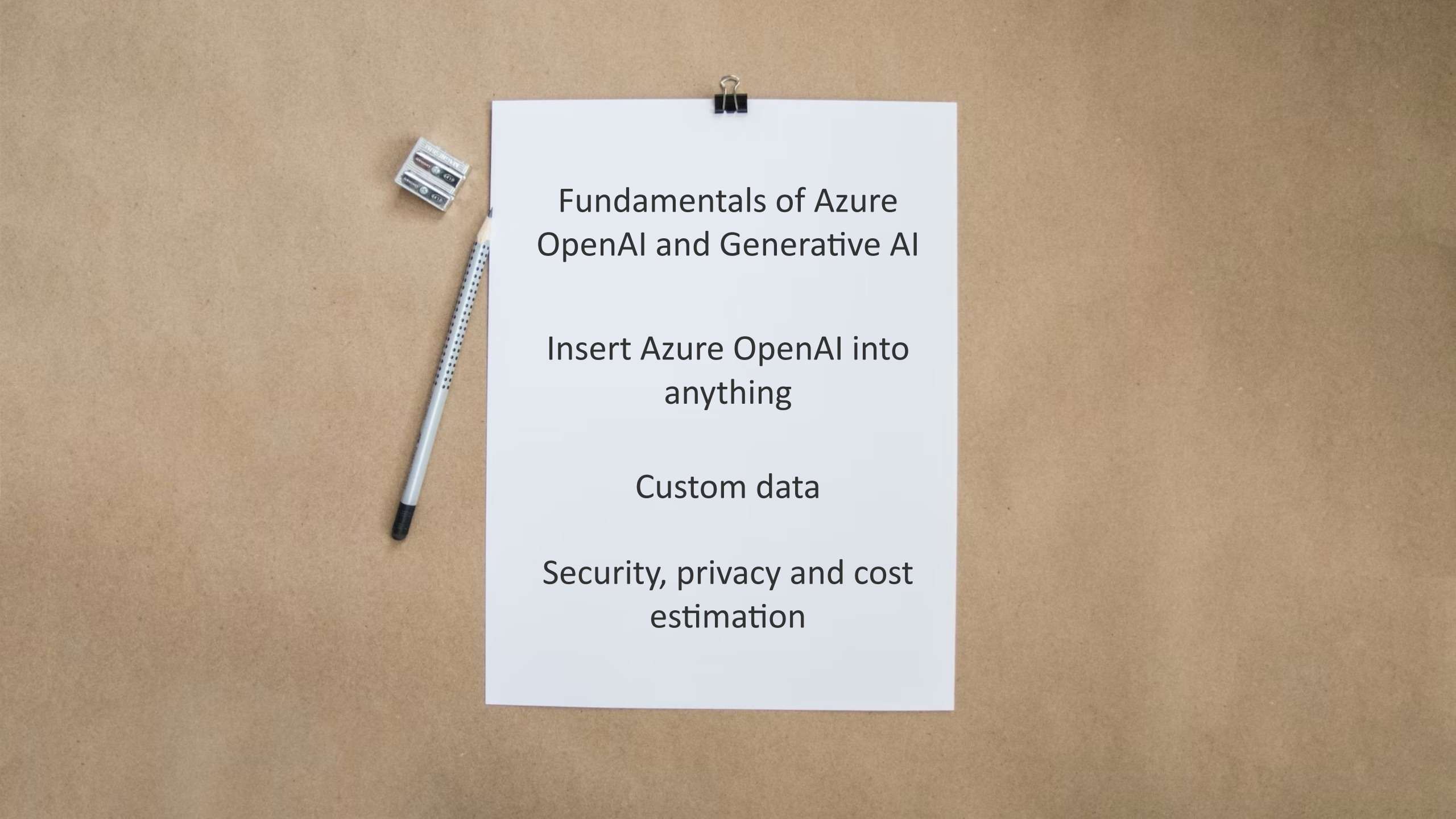
Community

EUROPEAN COLLABORATION SUMMIT 2023

EUROPEAN CLOUD SUMMIT 2023

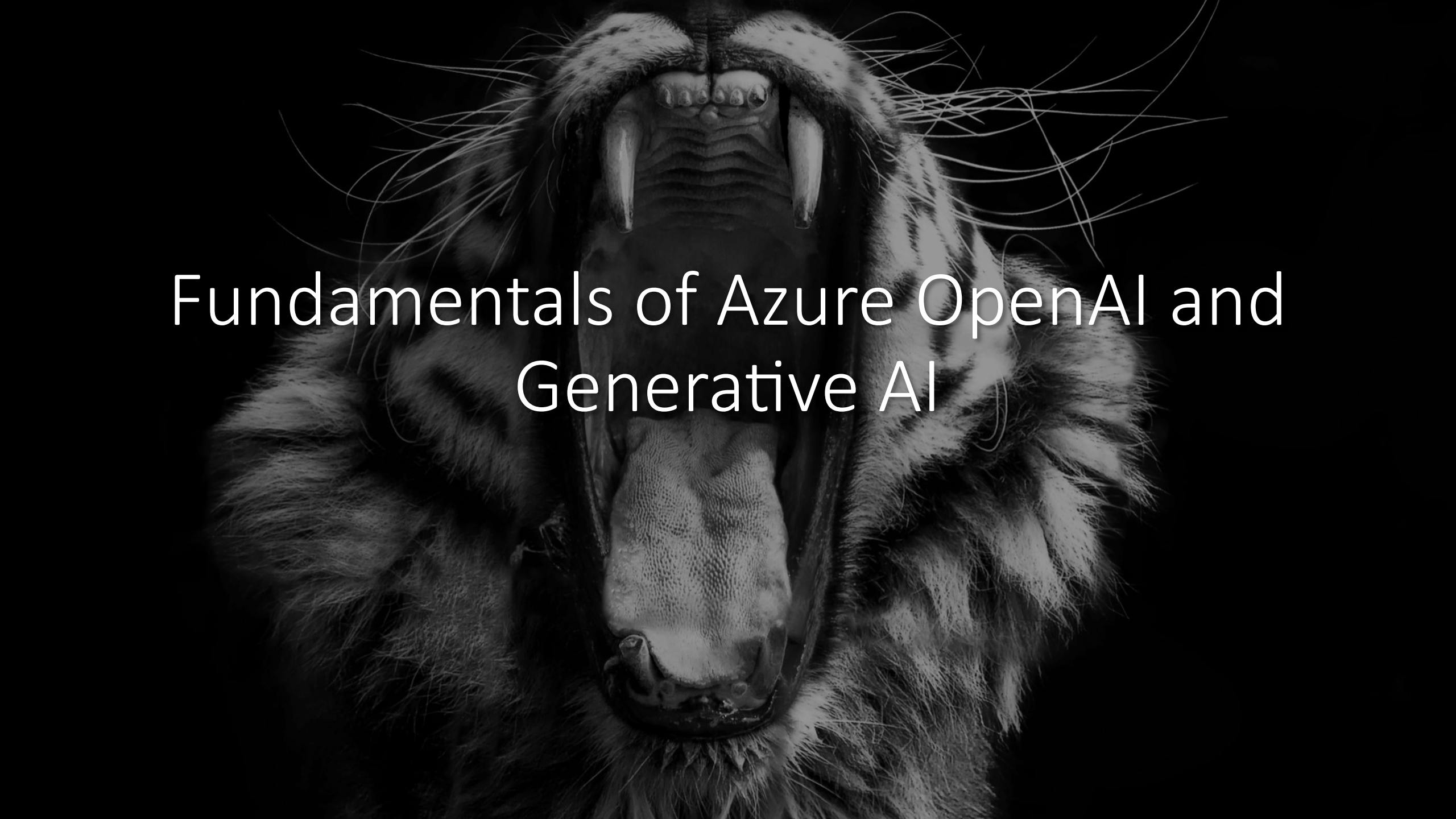
Organized by

BIWUG


Fundamentals of Azure OpenAI and Generative AI
Insert Azure OpenAI into anything
Custom data
Security, privacy and cost estimation

WHAT FINLAND'S FLAG STANDS FOR

VERY
FINNISH
PROBLEMS



Fundamentals of Azure OpenAI and Generative AI

What is OpenAI?

Ensure that artificial general intelligence (AGI) benefits humanity.

OpenAI is an AI research laboratory

- They've introduced models such as DALL-E, GPT-3, GPT-3.5, GPT 4 and Codex

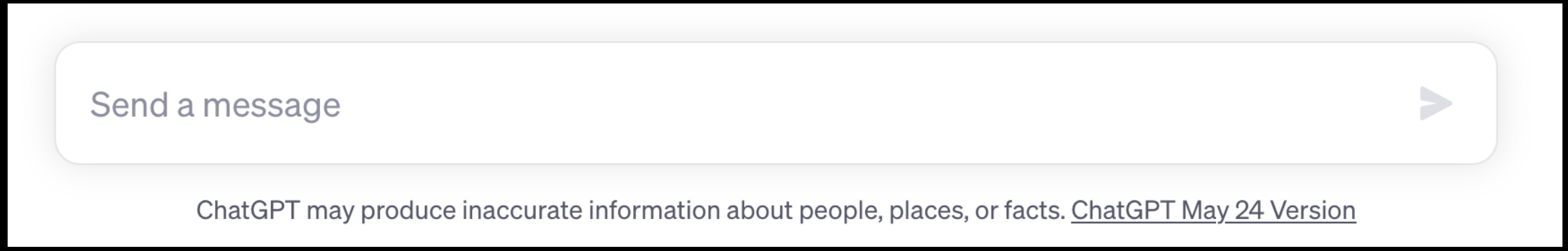
In late 2022, they launched ChatGPT – a new generative AI service

- Based on GPT-3 and then GPT-3.5 Turbo

GPT | Codex | DALL-E | ChatGPT

And what is ChatGPT?

It’s the most exciting service launched in years – a textbox!



Magic-like AI chatbot capability, yet still a black box essentially. How does it work? Nobody knows.

Free and paid subscription available - available at https://chat.openai.com

Why do I love ChatGPT?

How to increase number of open files in Azure ARM template for container instance

```
    --memory 1 \

    --cpu 1 \

    --ulimit nofile=65535:65535
```

"Hello all, I wanted to clarify this issue. Proposed solutions are based on GPT search results that seem invalid."

What is Azure OpenAI, then?

Azure OpenAI, in contrast to OpenAI, is a REST API providing access to the same OpenAI models

Secured with Entra ID and all the goodness of Azure security.

Available in most regions now – West Europe, Sweden Central, UK South, France Central, East US, etc.

~~To get access, you have to apply for Azure OpenAI, and then apply for the more advanced models: https://aka.ms/oai/access & https://aka.ms/oai/get-gpt4~~

Differences with OpenAI: private networking, privacy, encryption, regional availability, PaaS-like configuration.

- GPT
- Codex
- DALL-E
- ChatGPT

What are Large Language Models?

Large Language Models consist of artificial neural networks

- And these include tens of millions, or billions of parameters

Trained on text – such reddit.com, GitHub, Wikipedia, and so on.

Pre-trained, so inherently and by design, on a large corpus

- To put it simply, predicting the next word in a sentence.

Input and output is numbers – so words have to be tokenized in a tokenizer map

- 1 token maps to ~4 characters (~0.75 words), at least in English
  - Hi from Finland: laskuvarjojääkärikokelas

Let's talk a bit about the models

GPT-3 and GPT-4

Davinci: Most capable
Curie: Very capable, but faster
Babbage: Very straightforward and very fast
Ada: Simple tasks, fastest
GPT-4: Greater accuracy, supports more tokens

Codex

Davinci-codex: Best for applications needing deep understanding of code

Cushman-codex: Powerful and fast, better for generating code

ChatGPT

ChatGPT: Designed for conversational interfaces – conversation in, message out.

What are prompts, and prompt engineering?

You interact with the models almost solely through prompts

- Prompt → model → completion

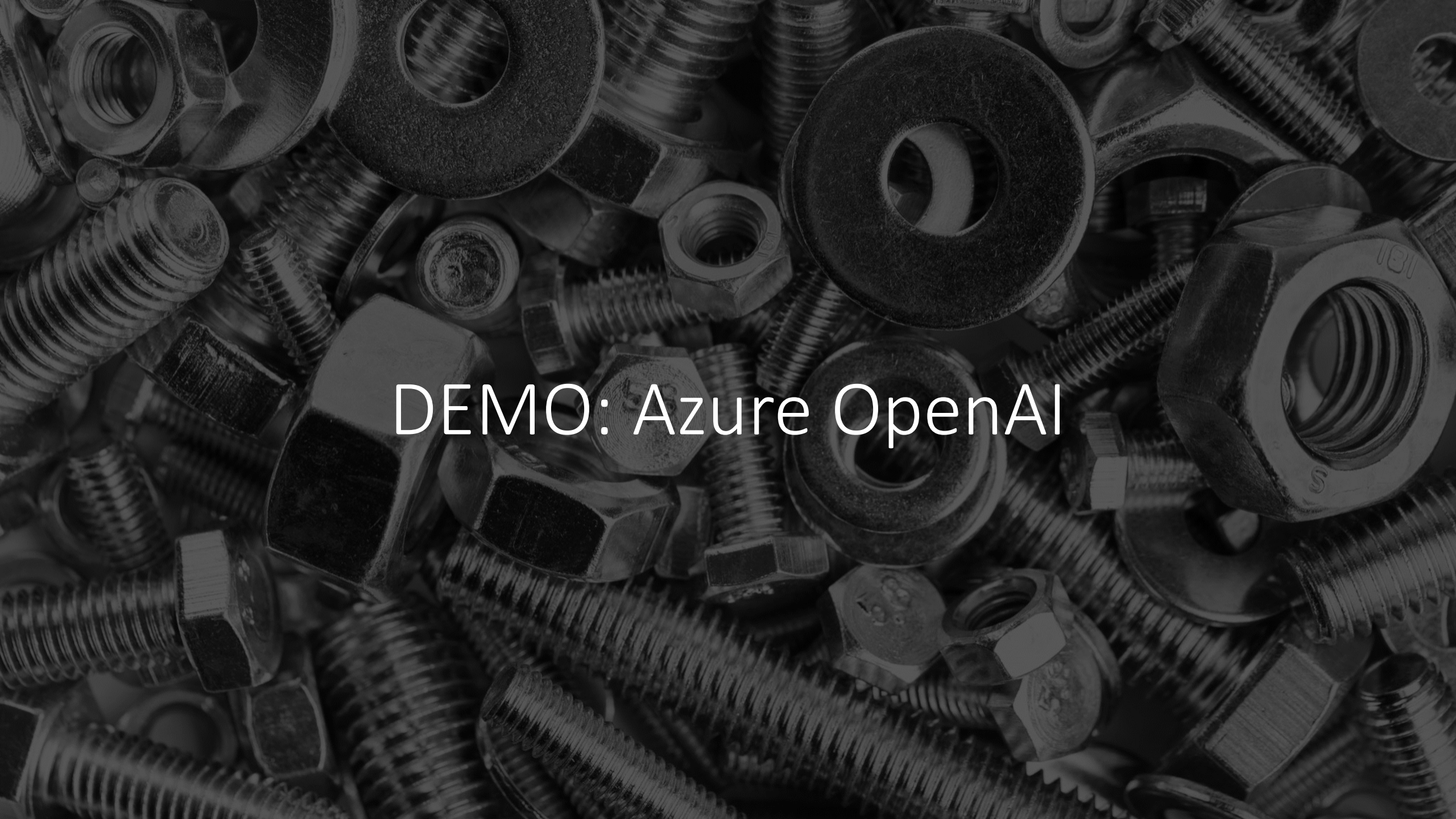
The first prompt I used back in the day:

- "What does WTF stand for?"

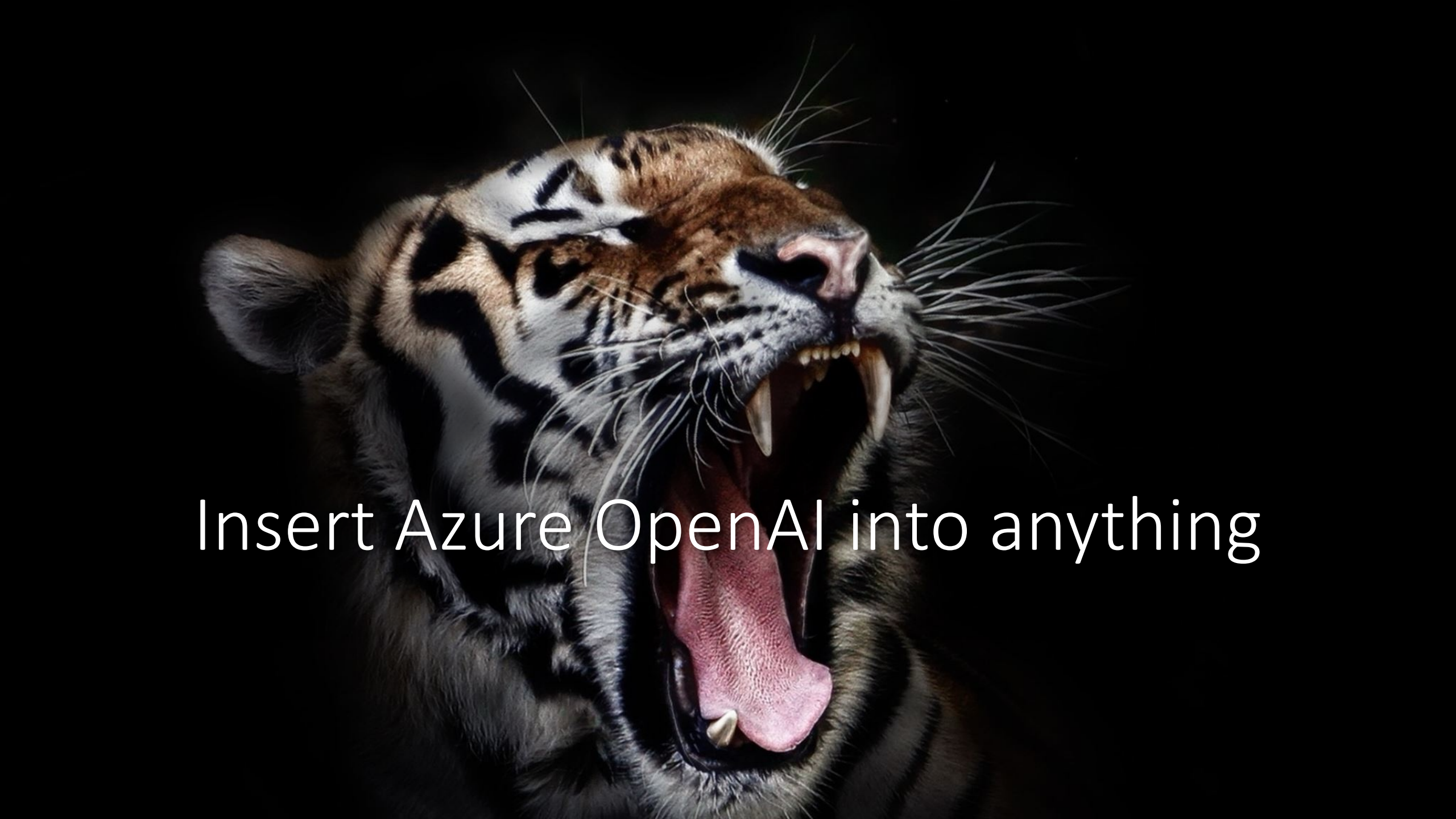
Prompt engineering is more art than strict science

- Prompts vary from simple to complex, and they can be vastly comprehensive and detailed – see https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/prompt-engineering

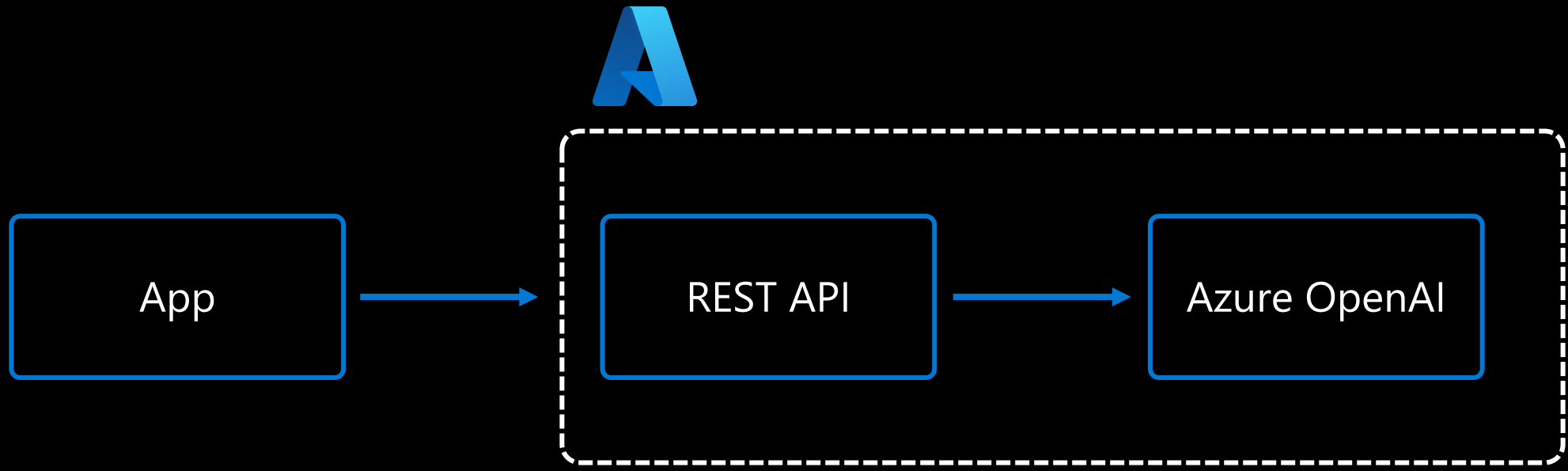
Prompts count for the number of tokens consumed.



DEMO: Azure OpenAI


Insert Azure OpenAI into anything

A word about the architecture



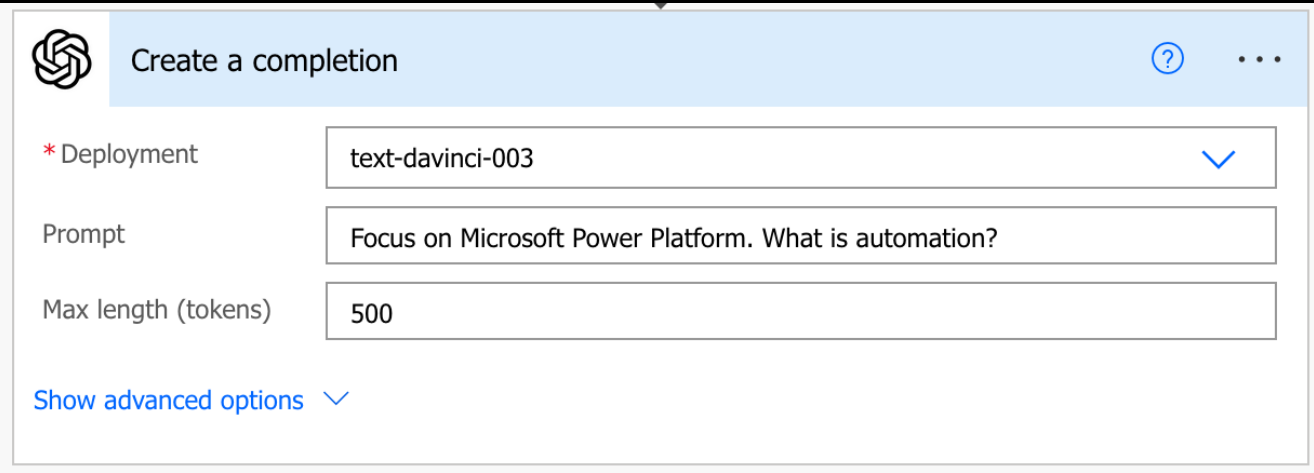
Connecting to Azure OpenAI – basic approach

A traditional REST API – requires an API Key, API version, and optional parameters (such as max_tokens).

Don’t forget what Azure can offer: API Management and Azure Functions to abstract the actual logic.

Test directly via Azure AI Studio, or with any command line tool: curl is a fantastic option!

Connecting to Azure OpenAI with Logic Apps and Power Automate



You can utilize the Custom Connector – by Daniel Laskewitz, Andrew Coates and Robin Rosengrün https://github.com/microsoft/PowerPlatformConnectors/tree/dev/custom-connectors/AzureOpenAIService

Use cases for Azure OpenAI

Chatbots	Injecting intelligence	Accelerating business
Power Virtual Agents	Canvas & model-driven apps	Multiple LLMs as part of a solution
Embedded chat agents	Augmenting workflows	Custom Copilots
Custom web apps	Enriching content & data	Industry—specific tools


crypto.com
PERONI
0.0%
aramco
JCB
ASTON MARTIN
Cognizant
FORMULA ONE TEAM
SentinelOne

DEMO: Connecting to Azure OpenAI

Custom data

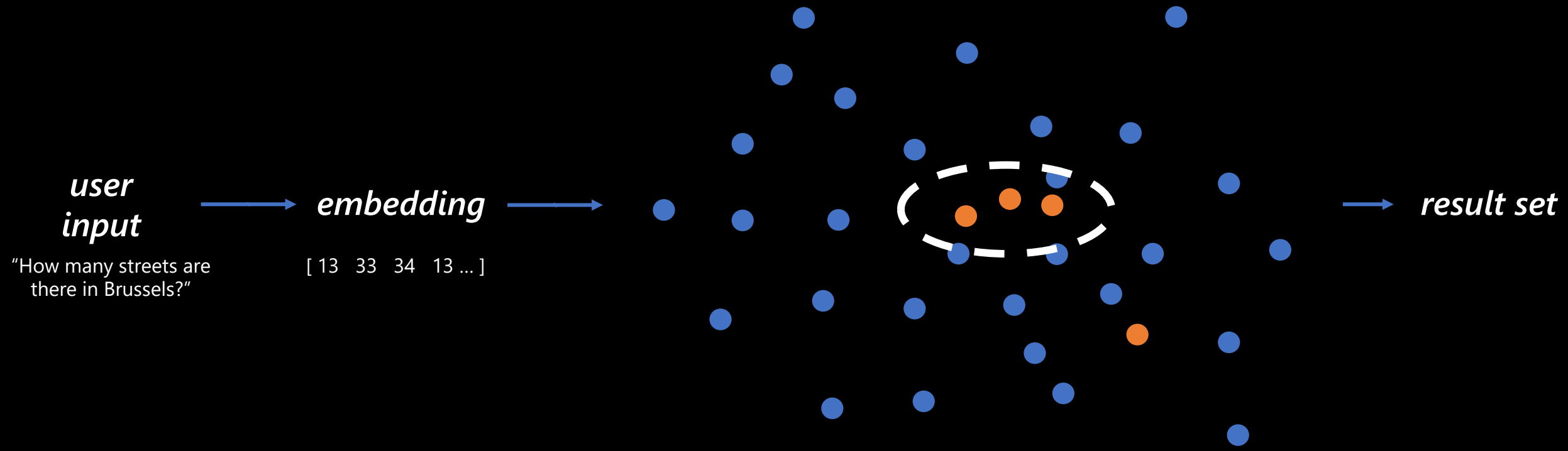
Why custom data? Doesn't GPT know everything?

GPT models know too much – and then they start to hallucinate.

Custom data aims to guardrail and contain the information.

The idea is to add custom data, and control the level of hallucination at the same time.

About embeddings and vectors

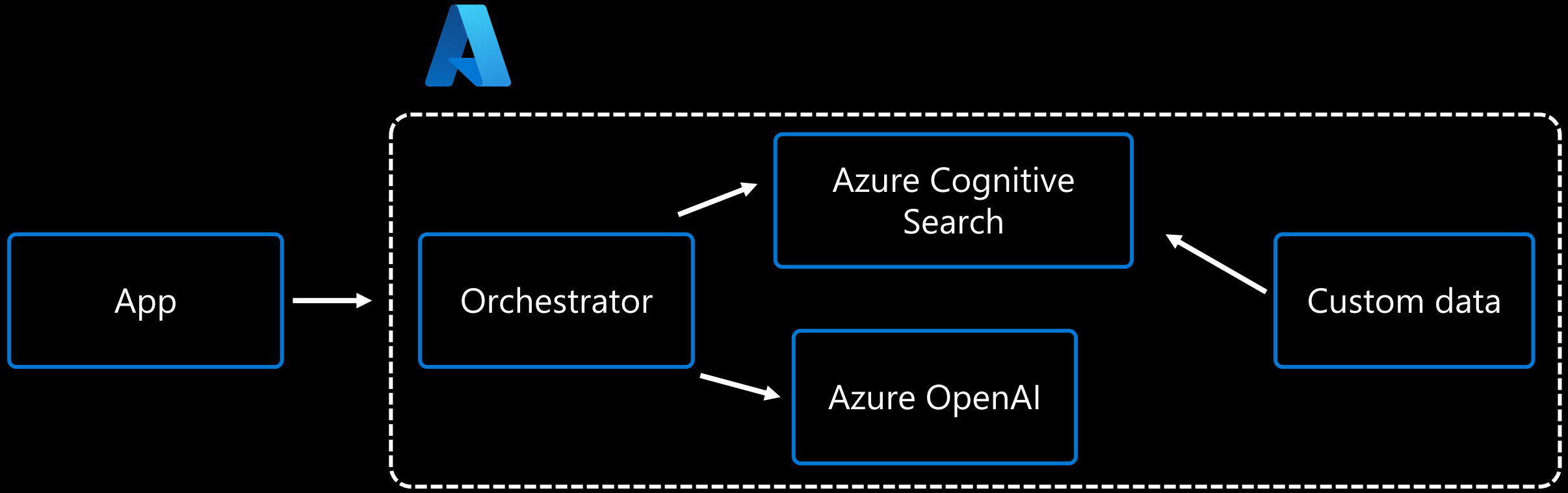


Representation of text in semantic meaning

Vectors (arrays of numbers) – think of them as points on a line, and similar text has shorter distance with each other.

About RAG and grounding

Retrieval Augmented Generation is a fancy phrase for "just do what I tell you, and don't forget that!". Generative AI does not have context.



Custom data and cracking documents

Embedding your data requires that text and content is tokenized, and stored as vectors

Vectors for embedding data can be generated with the help of Azure OpenAI, or additional tools

Vectors have to be stored in a database, such as Pinecone or Redis with RediSearch. Azure SQL works also.

Fine tuning

Instead of embedding data, you can also fine-tune the existing LLM models.

It’s a time consuming and often costly effort, but might yield best results in specific scenarios.

Data must be formatted as JSONL – JSON with Line breaks.

```
{"prompt": "How to optimize making filtered coffee at home with a
  Moccamaster", "completion": "Water the filter paper"}

{"prompt": "How to make best coffee on a Moccamaster", "completion": "Use
  cold and clean water"}
```

Fine tuning process

Status: Training succeeded

Finished training on: 3/17/2023 1:51 PM

Training file: coffee-azure-openai.jsonl

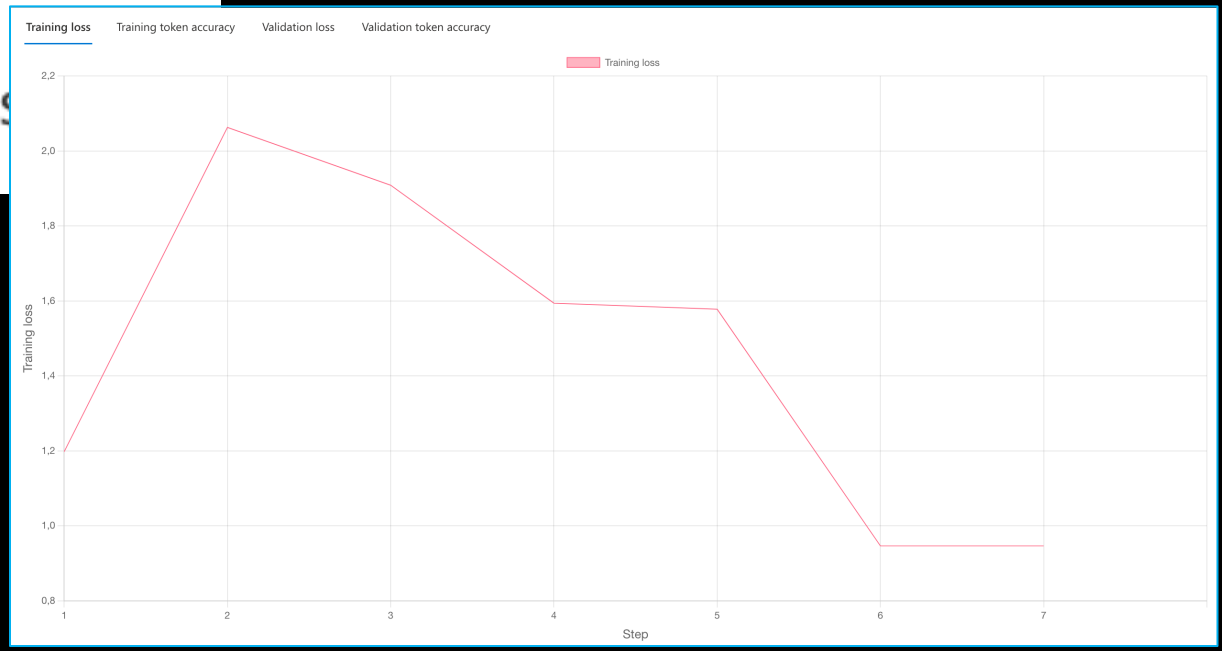
Base model: curie

Total training time: 35 minutes, 46 seconds

Statistics:

Total tokens: 159

Total examples: 7



DEMO: Custom data


Security, privacy and
cost estimation

What happens with my data?

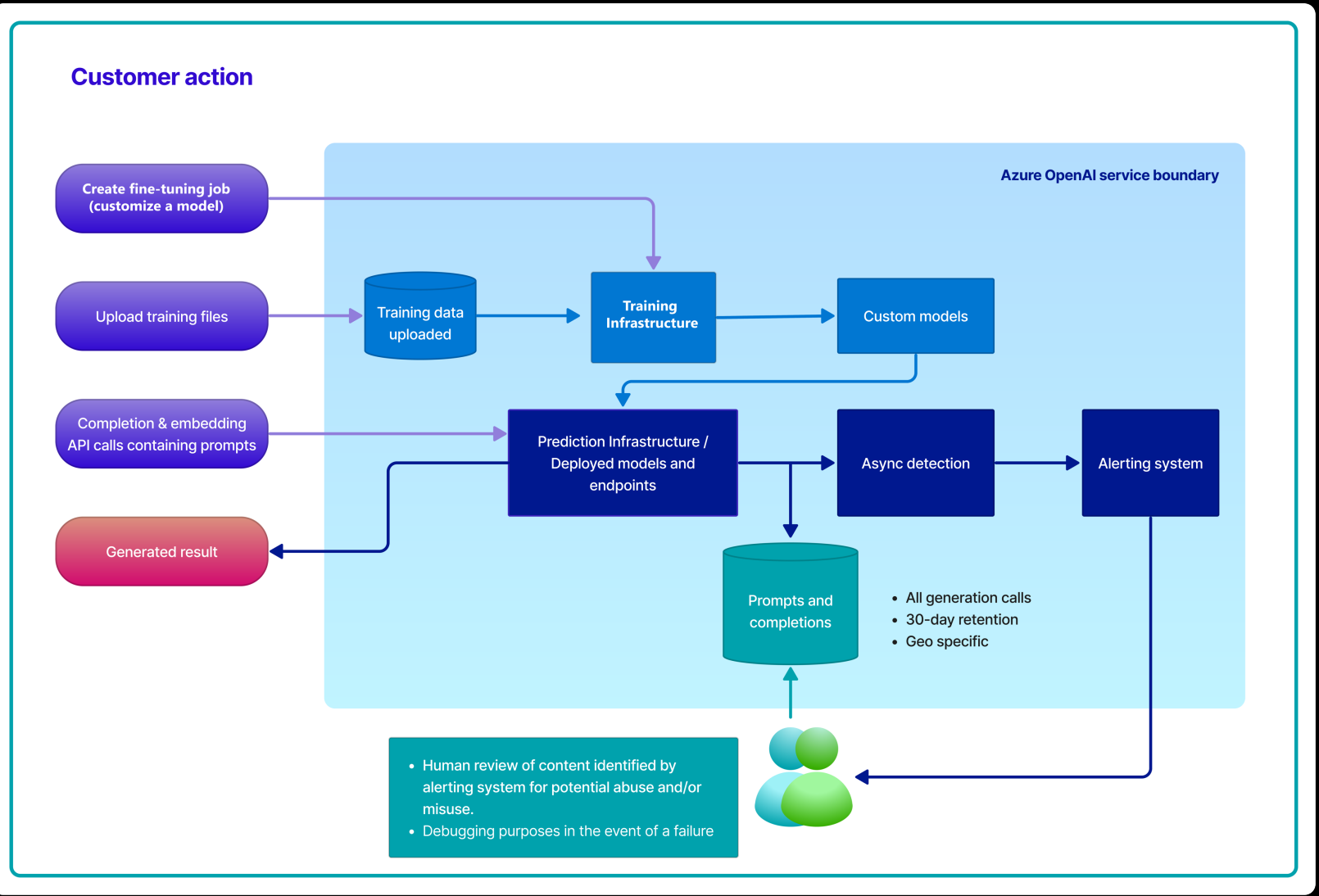
Azure OpenAI processes prompts, completions, training & validation data, and results from trainings

- No data is used to train the models – also no data is sent to OpenAI

Prompts and completion are temporarily stored for up to 30 days – it's encrypted.

All other data, such as custom data and training, is encrypted and isolated within the Azure subscription

How is my data managed?



Source: [Microsoft]

What data should you use with Azure OpenAI?

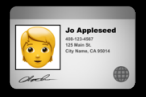
Sensitive data can be used – as opposed to cannot be used with OpenAI (the service)

Encrypt all data, as you would usually anyway when you store it in cloud.

At a certain point, you have to trust Microsoft – customer-managed key will be available (with Key Vault support)
https://aka.ms/cogsvc-cmk

Security capabilities to consider

Bind your Azure OpenAI instances to private endpoints or VNETs you control → closed from the public Internet



Managed Identities should be used when possible

Build monitoring around Azure OpenAI instance use – tokens used, total calls, total errors, blocked calls, etc.

What about costs?

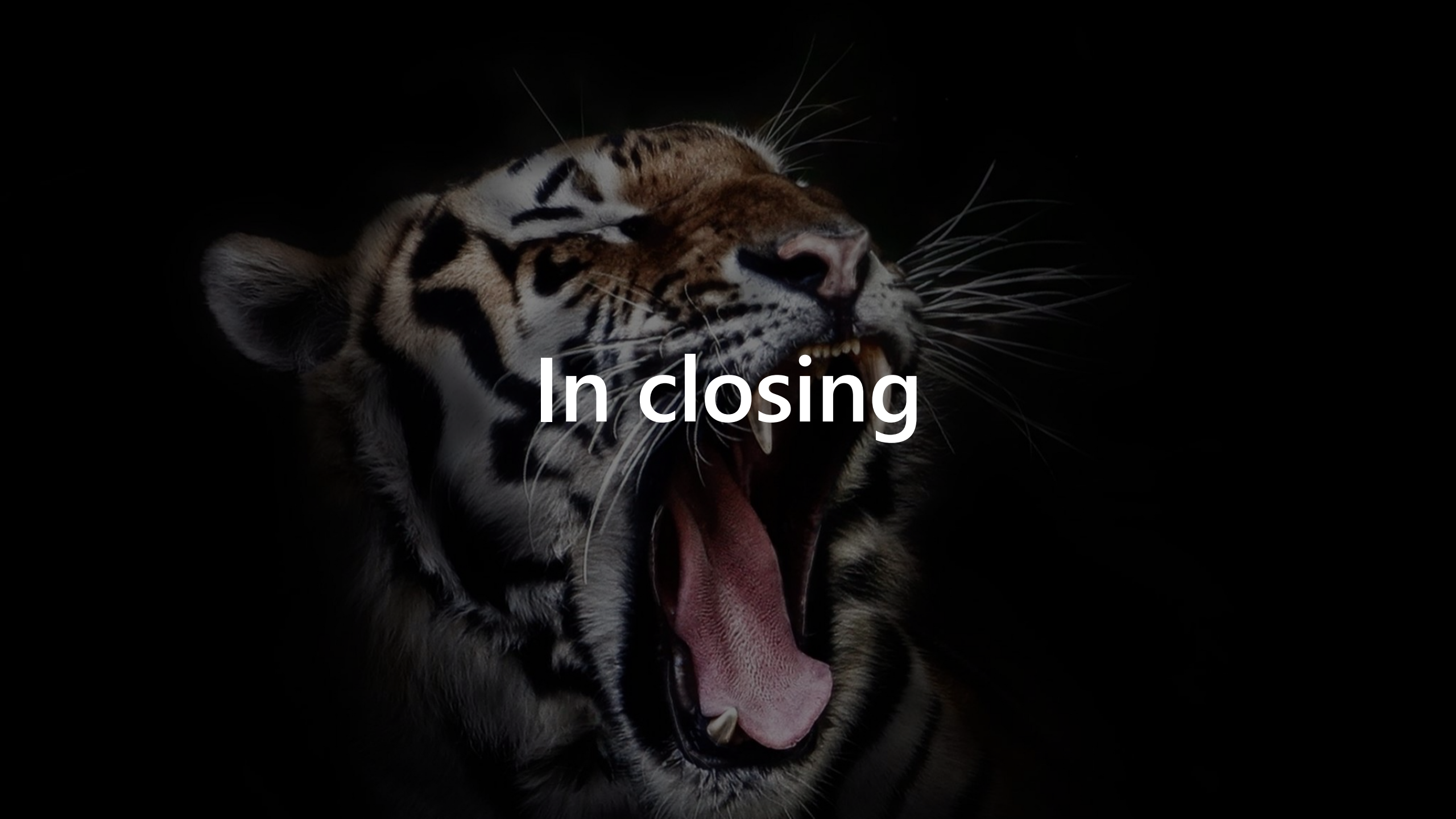
Tokens will cost you – per 1000 tokens, and depending on which LLM models you are using.

GPT-3.5-Turbo – 0.0020 € (completion)

GPT-4 8K – 0.058 € (completion)

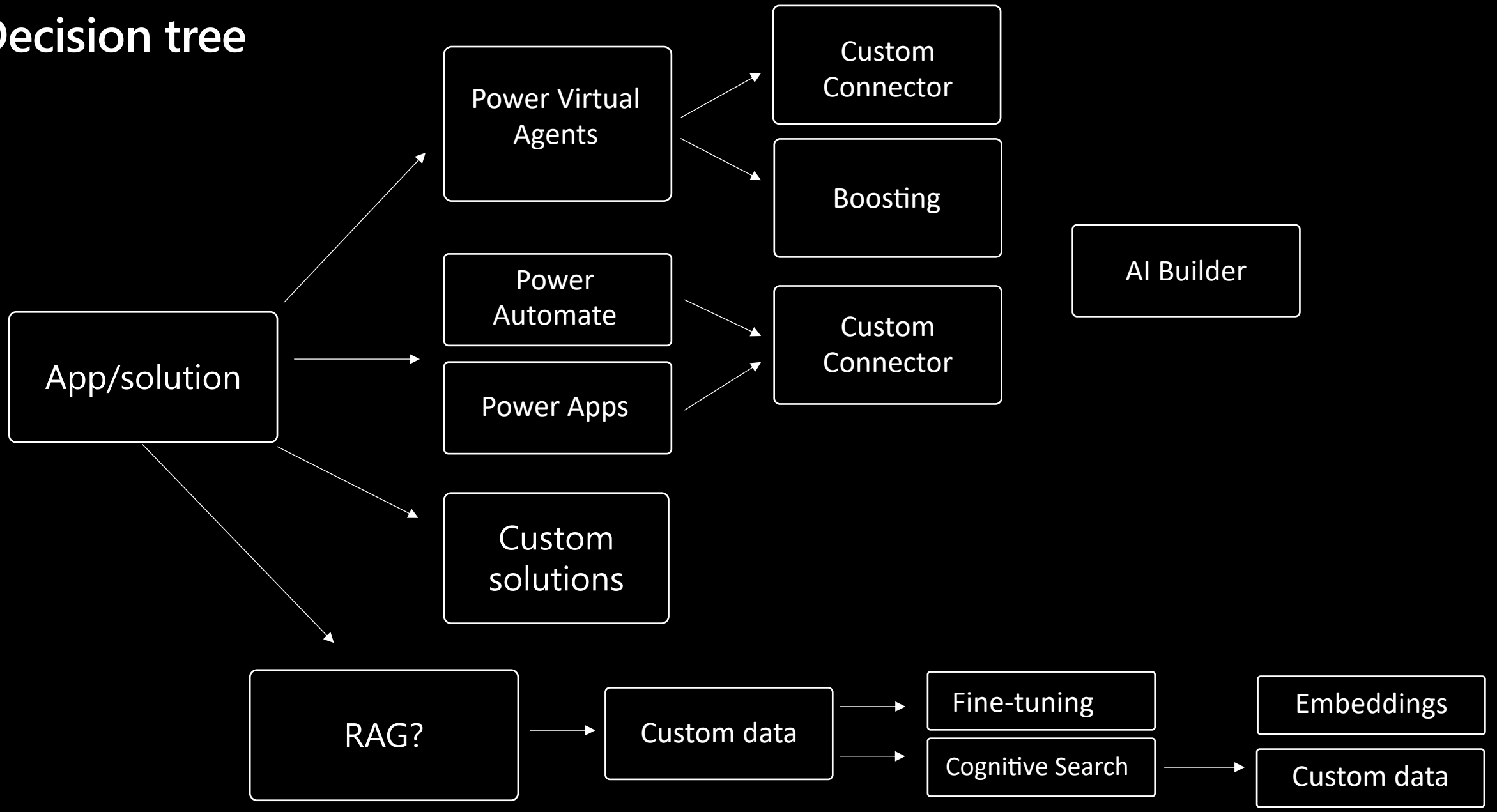
Factor in that tokens are consumed in RAG, and if you build a memory, it adds up *quite* quickly.

Azure Cognitive Search is about 233 €/month, fixed.



In closing

Decision tree



Tooling and projects - recap

Azure AI Studio

Azure Machine Learning Studio

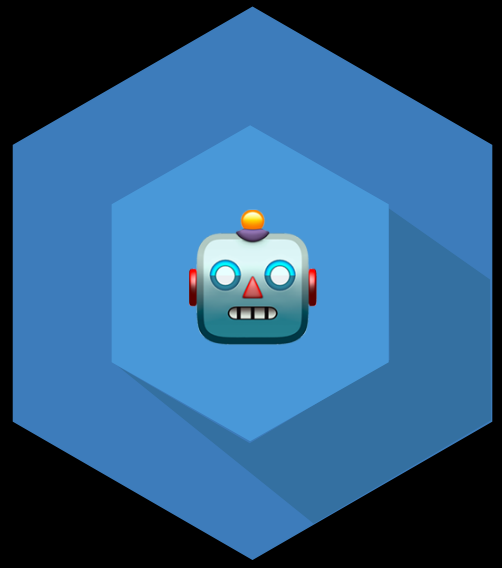
Azure Cognitive Search

Your private ChatGPT in Azure - https://github.com/microsoft/azurechat

Chat with your data - https://github.com/Azure-Samples/chat-with-your-data-solution-accelerator

Enterprise end-to-end demo-- https://github.com/Azure-Samples/azure-search-openai-demo

Now, go and build something!



Start building with Azure OpenAI – today!

Learn to use RAG – it's key for memory, and brings more "intelligence" for your solutions.

For custom data, things will become easier in the future.

Thank you!

aka.ms/jussi