

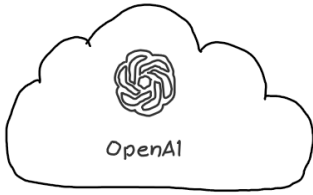


Azure OpenAI

Workshop & Buzz-Word Bingo

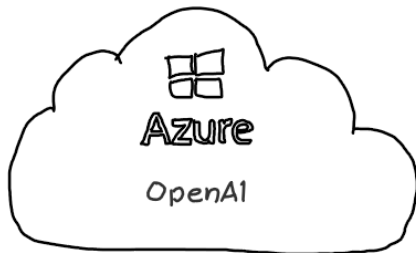
Robert Eichenseer
Senior Services Engineer
Azure Engineering

OpenAI / Azure OpenAI



OpenAI

- “Startup” which runs on Azure and partners with MS
- Hosts the famous <https://chat.openai.com>
- Alpha Models
- Early access
- Prototyping early features



Azure OpenAI

- Enterprise Grade (security, availability, networking, regional availability, financially backed SLA ...)
- Same models as OpenAI
- Fine-tuned models

Azure OpenAI co-develops the APIs with **OpenAI**, ensuring compatibility and a smooth transition from one to the other.

Begriffsdefinition

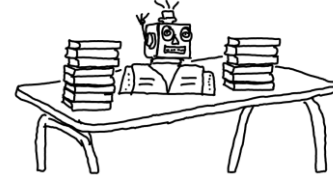


Künstliche Intelligenz (AI)

Die Intelligenz, die von Maschinen demonstriert wird, ähnlich der Intelligenz von Menschen.

Benützen
von Modellen

https://en.wikipedia.org/wiki/Artificial_intelligence



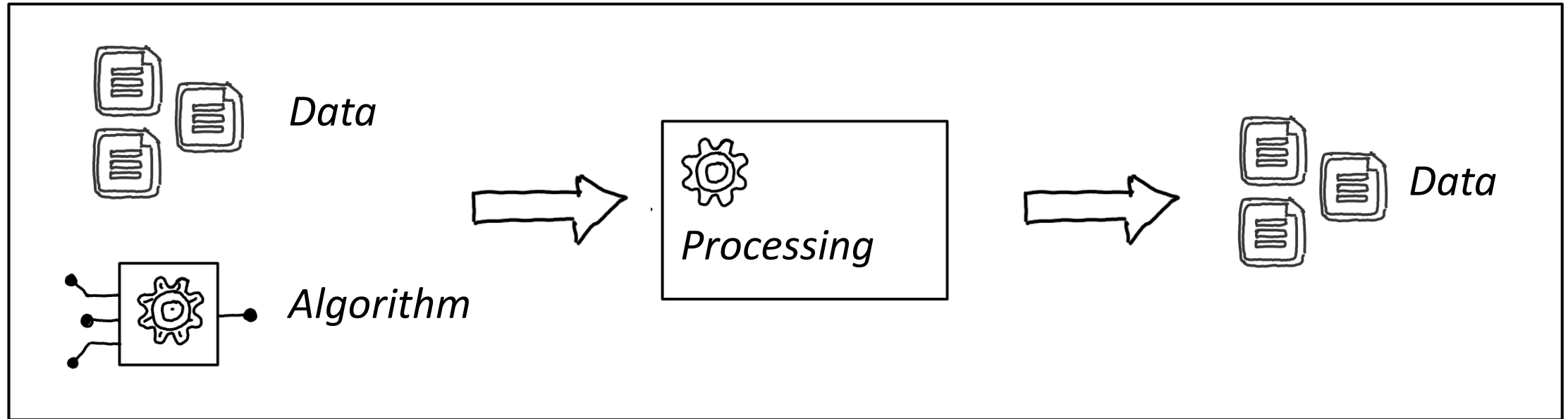
Maschinelles Lernen (ML)

Das Lernen eines künstlichen Systems aus Beispielen und die Fähigkeit der Verallgemeinerung dieser nach Beendigung einer Lernphase.

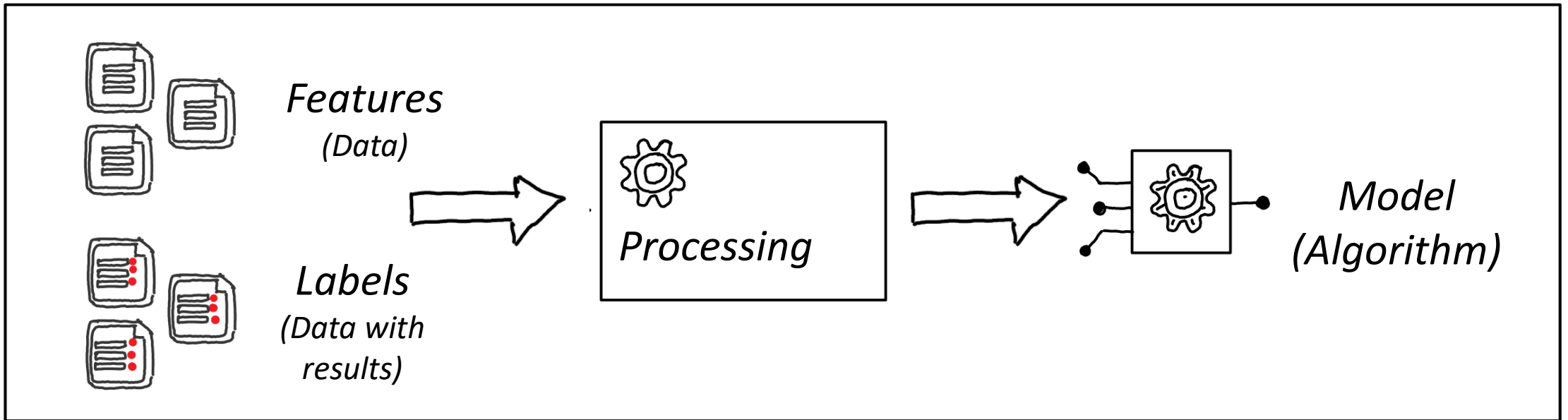
Erstellen
von Modellen

https://en.wikipedia.org/wiki/Machine_learning

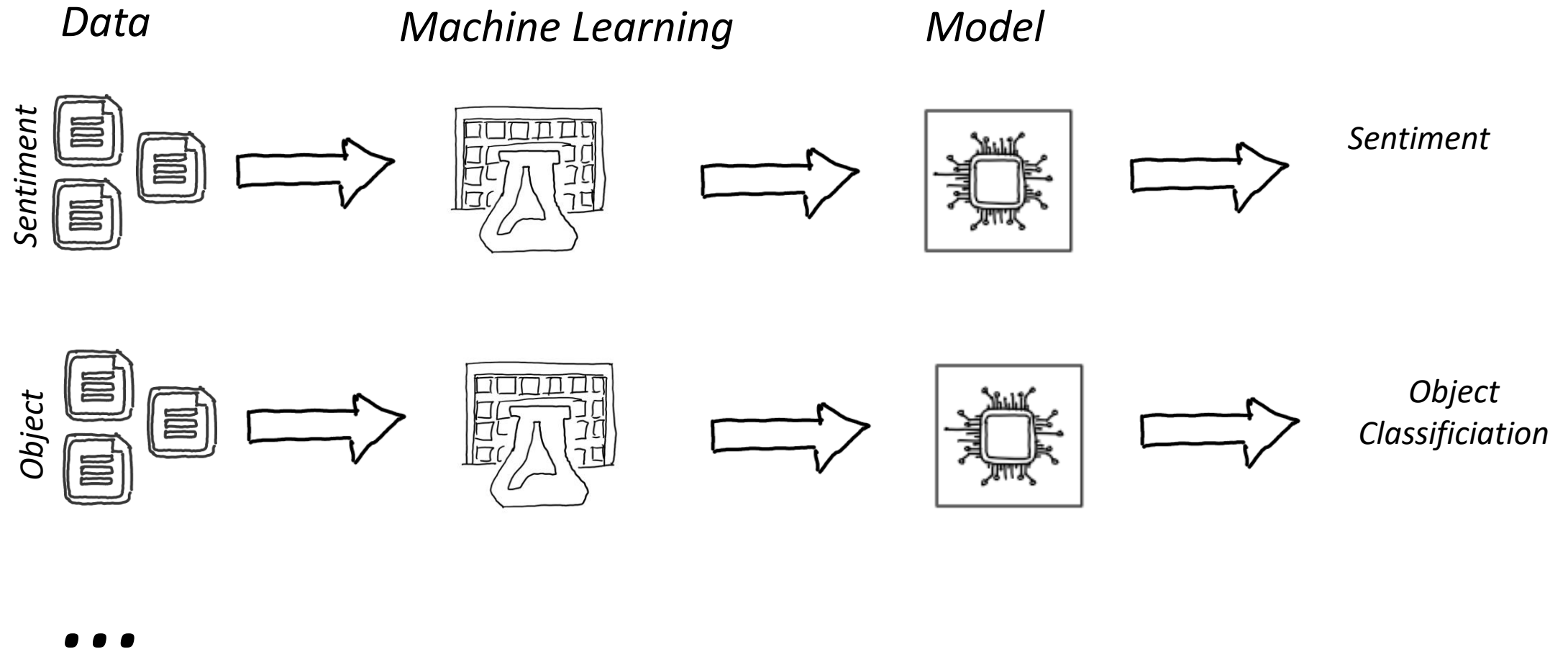
Traditional Software Engineering



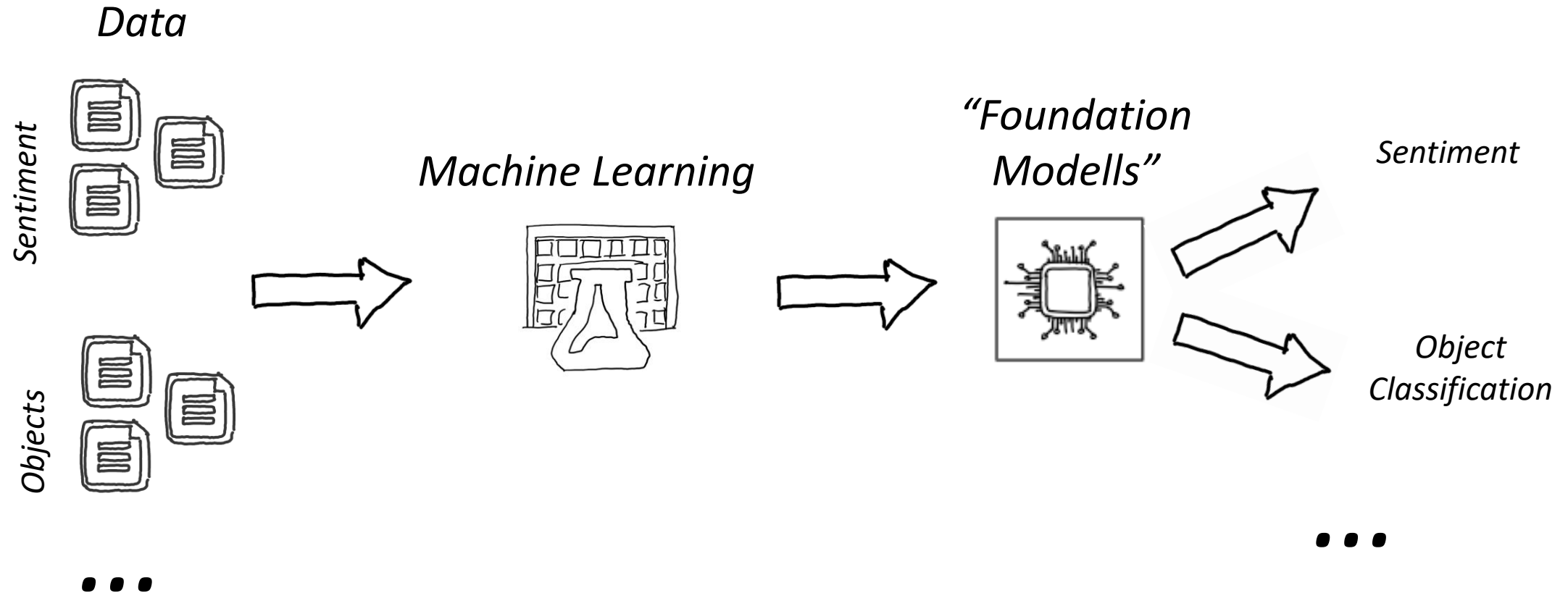
Training Models



Classic Approach



Large Language Models



Azure AI

Speech



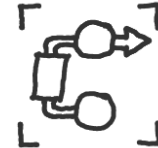
Vision



OpenAI



Content Safety



OpenAI LLM Models



GPT family

Verständnis von Text, Code und Bildern - Erzeugung von Text und Code



DALL-E

Erzeugung und Bearbeitung von Bildern mittels Text



Embeddings

Umwandlung von Text in eine numerische Representation (Vektor) unter Beibehaltung der semantischen Bedeutung



Whisper

Umwandlung von Sprache nach Text

Few Shot Learning vs. Model Fine Tuning

Few Shot Learning

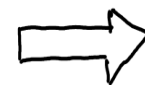
Question



Grounding



Sample
Q&A



LLM



Model Fine Tuning

LLM



Grounding



Fine tuned LLM



Sample
Q&A



Question



Tokens

Text In



Question



Grounding



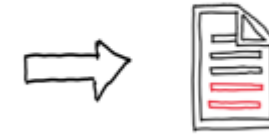
Inference

LLM



Text Out

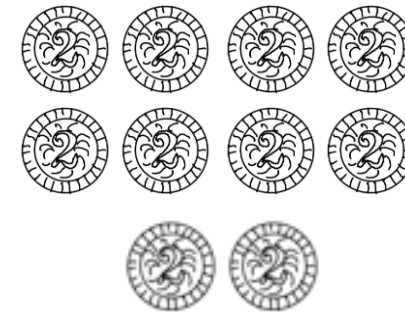
Completion



Tokens



Tokens



Tokens: The fundamental unit of text that the model operates on. It can be a word, sub-word or even a character.
Different models take different amount of tokens (e.g. GPT-4 32k: 24k words/48 pages – GPT-3.5: 3k words/6 pages)

Azure AI

Speech



Vision



OpenAI



Content Safety



Content Safety



Text & Bild Analyse

Überprüfung auf sexuelle Inhalte, Gewalt, Hass und Selbstverletzung



Jailbreak Risiko

Analysiert Text auf Jailbreak Angriffe



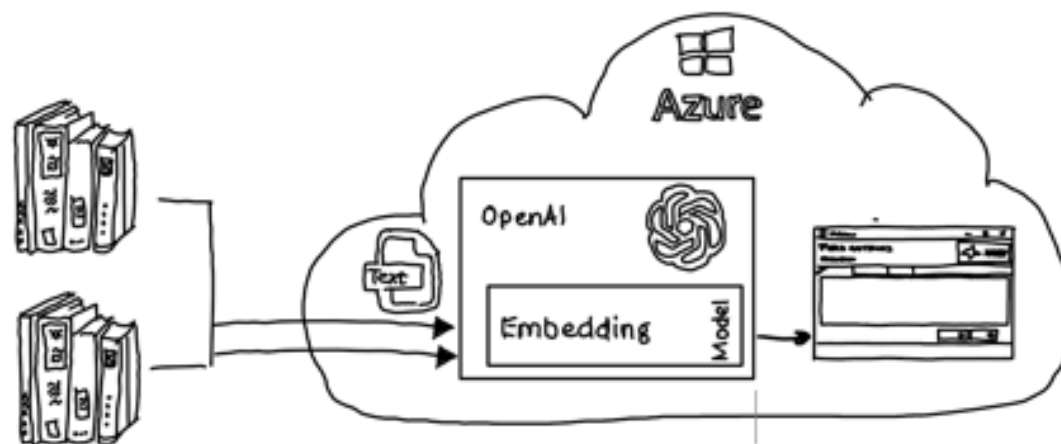
Copyright Verletzung

Scannt KI erzeugten Text auf bekannte und evtl. geschützte Inhalte (Lieder, Artikel ...)

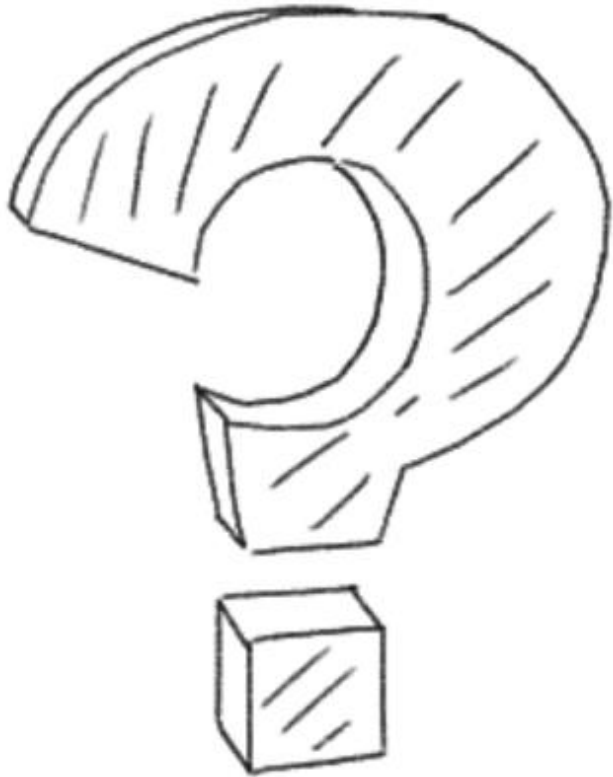
Exploring Crawl

Chat with your own Data

(GPT on your Data)



Recap - LLMs in 15 Seconds

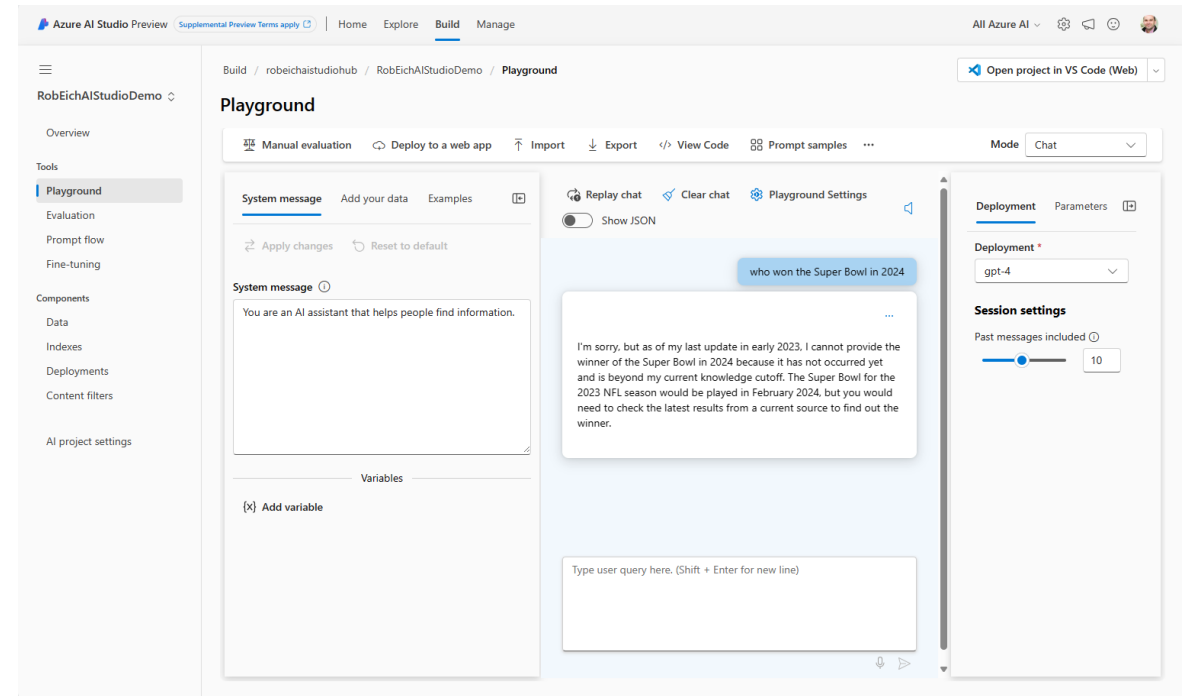


Large language models (LLMs) take a limited number of so-called tokens (words or subwords) as input and they predict the next most likely word or token and add it to the input text.

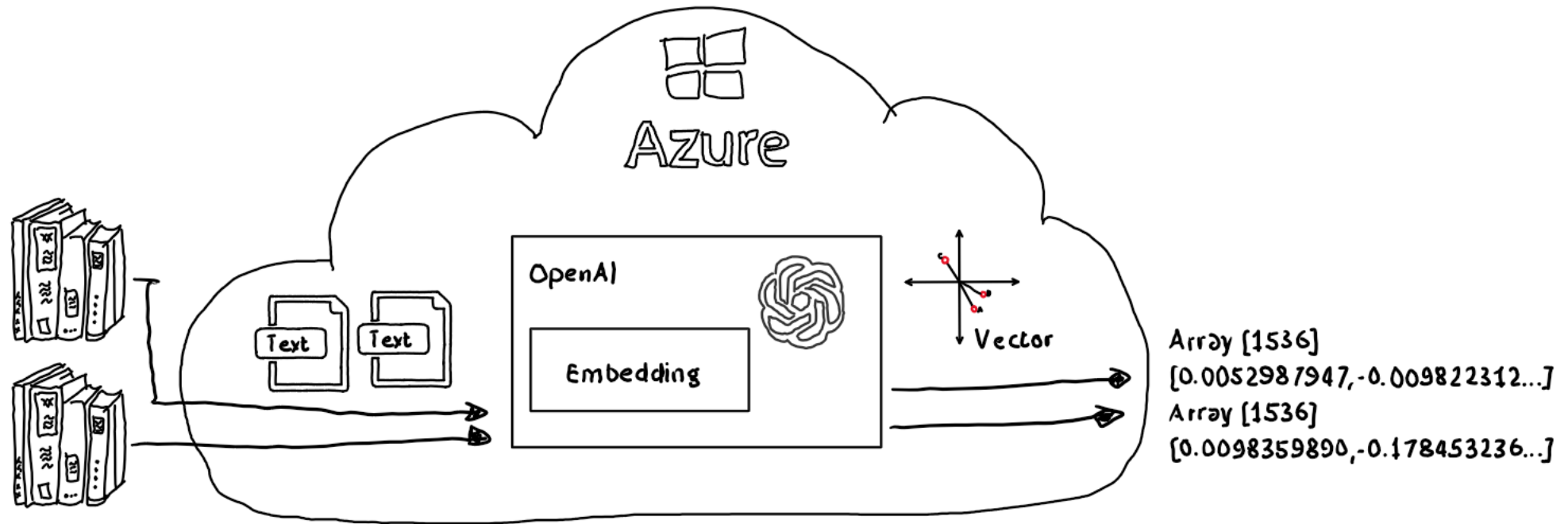
Therefore, the input text or prompt is also of great importance.

Demo

Grounding Private Data



Embeddings

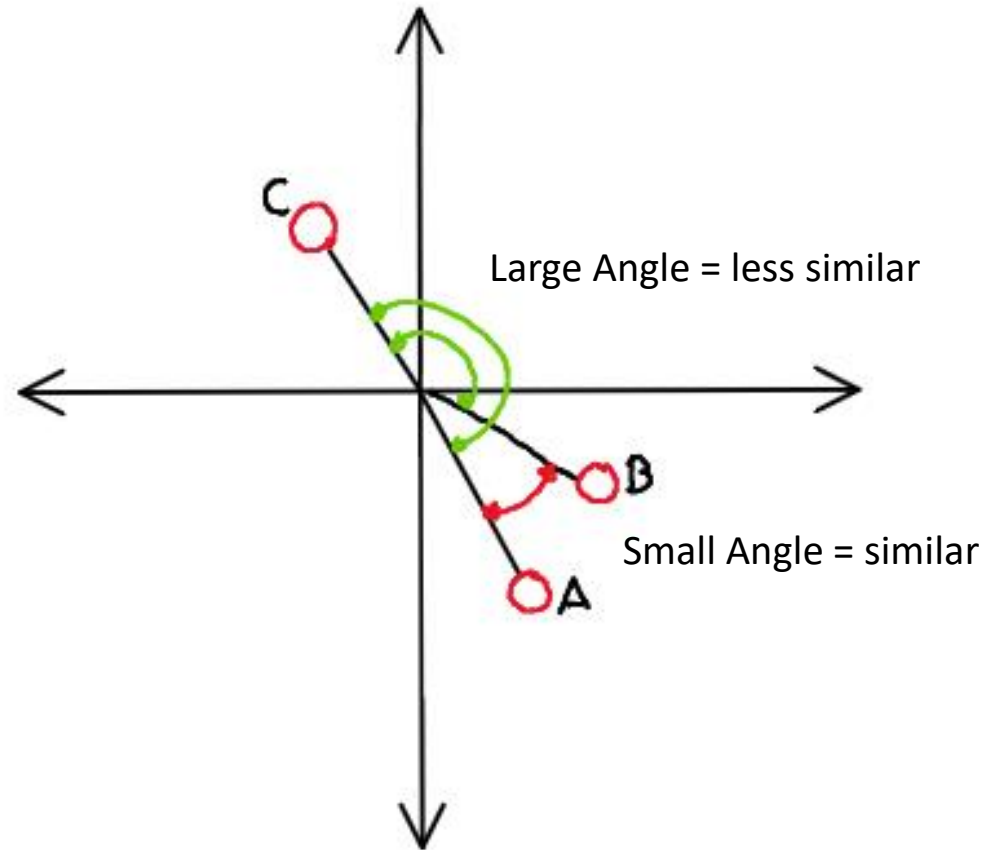


Embeddings

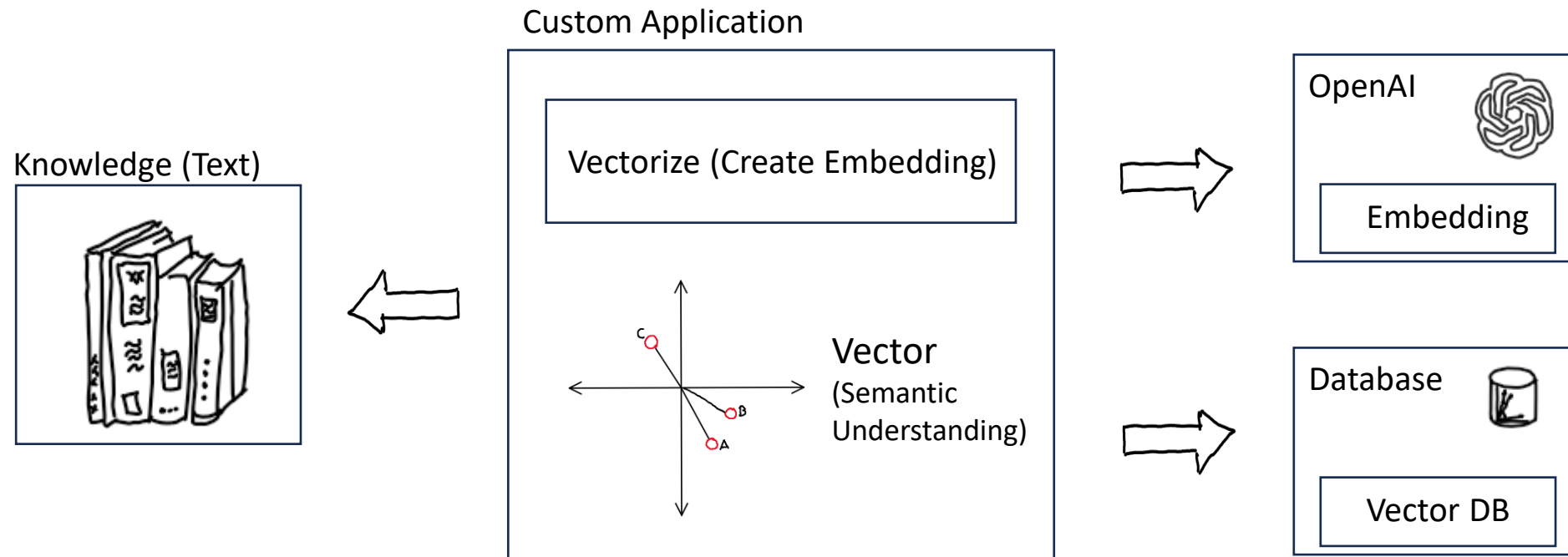
A = Boeing 747

B = Airbus A 380

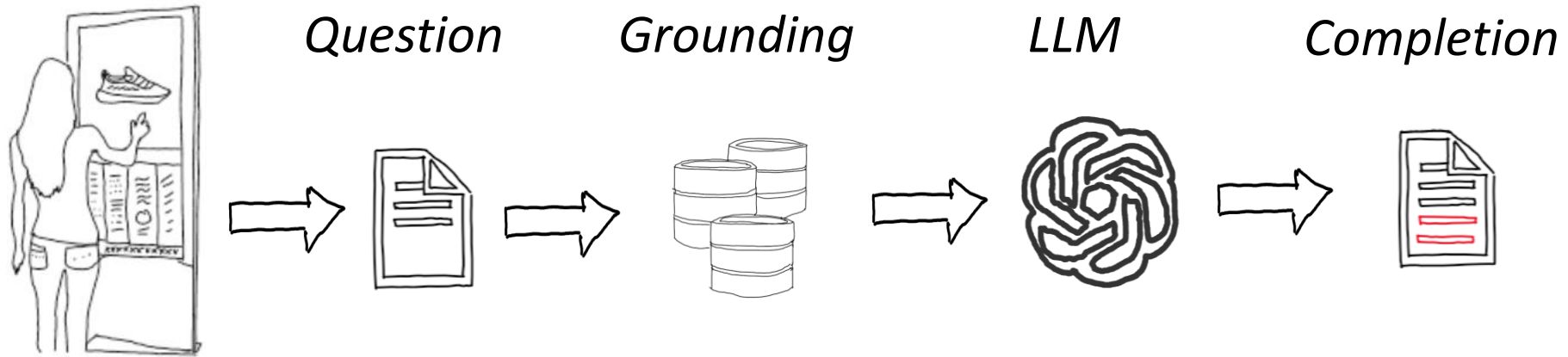
C = Cat



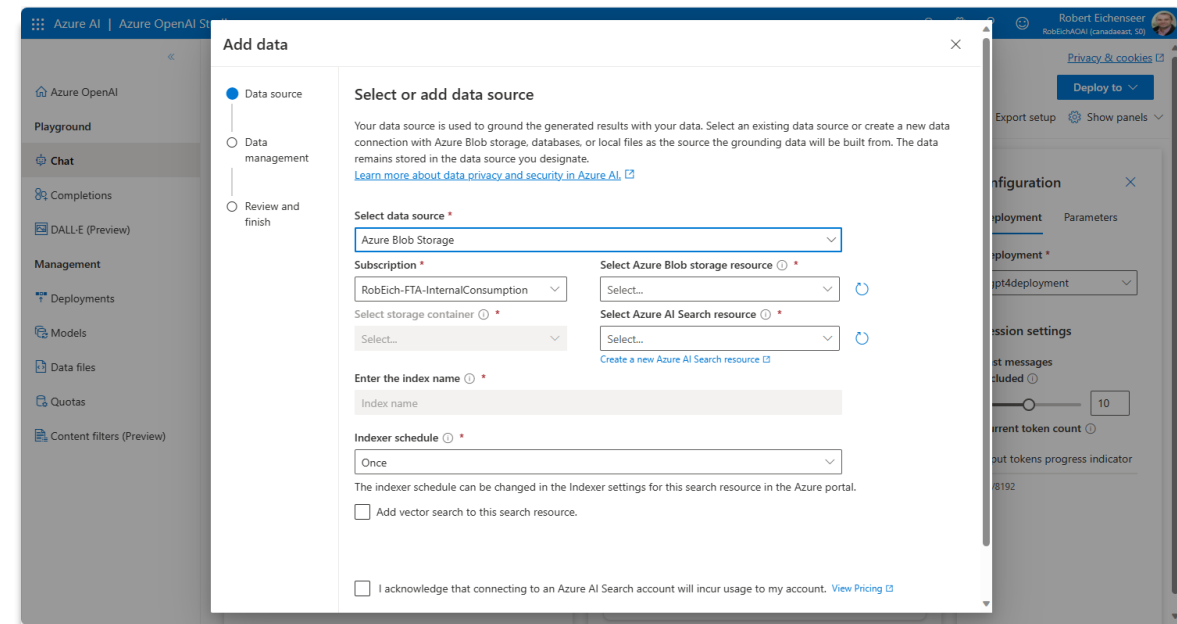
Embeddings / Vector DB



Grounding



Demo: GPT on your data



The screenshot shows the 'Add data' dialog box in the Azure AI Studio interface. The dialog is titled 'Add data' and has a close button (X) in the top right corner. It features a progress indicator on the left with three steps: 'Data source' (selected), 'Data management', and 'Review and finish'. The main content area is titled 'Select or add data source' and includes a descriptive paragraph about data sources. Below this, there are several configuration options: 'Select data source' (a dropdown menu showing 'Azure Blob Storage'), 'Subscription' (a dropdown menu showing 'RobEich-FTA-InternalConsumption'), 'Select storage container' (a dropdown menu showing 'Select...'), 'Select Azure Blob storage resource' (a dropdown menu showing 'Select...'), 'Select Azure AI Search resource' (a dropdown menu showing 'Select...'), 'Enter the index name' (a text input field showing 'Index name'), 'Indexer schedule' (a dropdown menu showing 'Once'), and a checkbox for 'Add vector search to this search resource.' At the bottom, there is a checkbox for 'I acknowledge that connecting to an Azure AI Search account will incur usage to my account.' and a link to 'View Pricing'.

Add data

Select or add data source

Your data source is used to ground the generated results with your data. Select an existing data source or create a new data connection with Azure Blob storage, databases, or local files as the source the grounding data will be built from. The data remains stored in the data source you designate. [Learn more about data privacy and security in Azure AI](#)

Select data source *

Azure Blob Storage

Subscription *

RobEich-FTA-InternalConsumption

Select storage container *

Select...

Select Azure Blob storage resource *

Select...

Select Azure AI Search resource *

Select...

[Create a new Azure AI Search resource](#)

Enter the index name *

Index name

Indexer schedule *

Once

The indexer schedule can be changed in the Indexer settings for this search resource in the Azure portal.

☐ Add vector search to this search resource.

☐ I acknowledge that connecting to an Azure AI Search account will incur usage to my account. [View Pricing](#)

Open project in VS Code (Web)

robeichazaistudiopr...

Overview

Tools

Playground

Evaluation

Prompt flow

Fine-tuning

Components

Data

Indexes

Deployments

Content filters

Project settings

Build / robeichazaistudiohub / robeichazaistudioproject / Playground

Playground

Manual evaluation | Deploy to a web app | Import | Export | View Code | Prompt samples | ...

Mode Chat

System message | Add your data | Examples

Replay chat | Clear chat | Playground settings

Show JSON

Gain insights into your own data source. Your data is stored securely in your Azure subscription. [Learn more about how your data is protected.](#)

Index:
[placid-mango-vgsl41jdjv](#)

Search type:

- Hybrid (vector + keyword)
- Semantic
- Keyword
- Hybrid + semantic
- ✓ Hybrid (vector + keyword)
- Vector



Start chatting

Test your assistant by sending queries below. Then adjust your assistant setup to improve the assistant's responses.

Type user query here. (Shift + Enter for new line)

Deployment | Parameters

Deployment *

gpt-35-turbo

Session settings

Past messages included ⓘ

10

Open project in VS Code (Web)

Build / robeichazaistudiohub / robeichazaistudioproject / Flows / robeichazaistudio

robeichazaistudio Chat

View batch runs

Runtime * Select runtime

Clone

Save

Deploy

Evaluate

Chat

Flow

+ LLM + Prompt + Python + More tools Raw file mode Wrap text

Inputs

Name	Type	Value	Chat input	Tag	Action
chat_history	list			Chat history	
query	string	Please input content in chat b			

+ Add input

Outputs

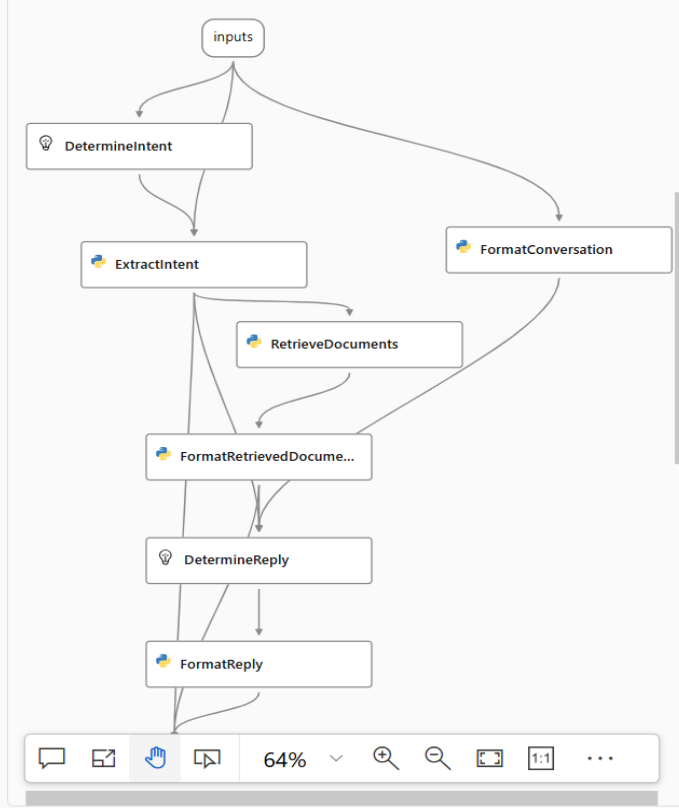
Name	Value	Chat output	Action
reply	\${FormatReply.output}		
search_intents	\${ExtractIntent.output.search_intents}		
fetches_docs	\${FormatRetrievedDocuments.output}		
current_query_i	\${ExtractIntent.output.current_message_intent}		

+ Add output

DetermineIntent Show variants Generate variants

Files

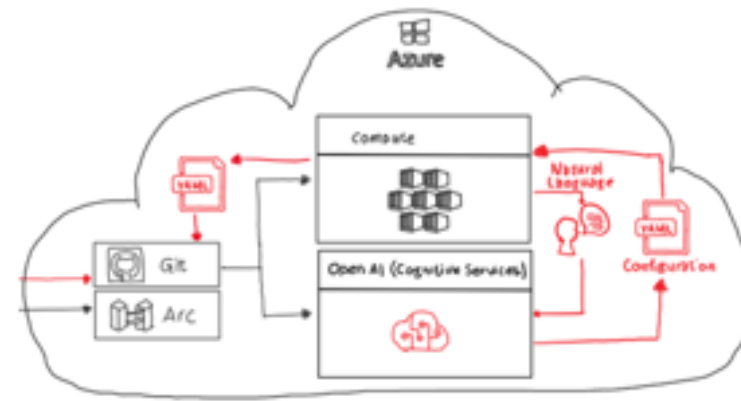
Graph



Scaling Walk

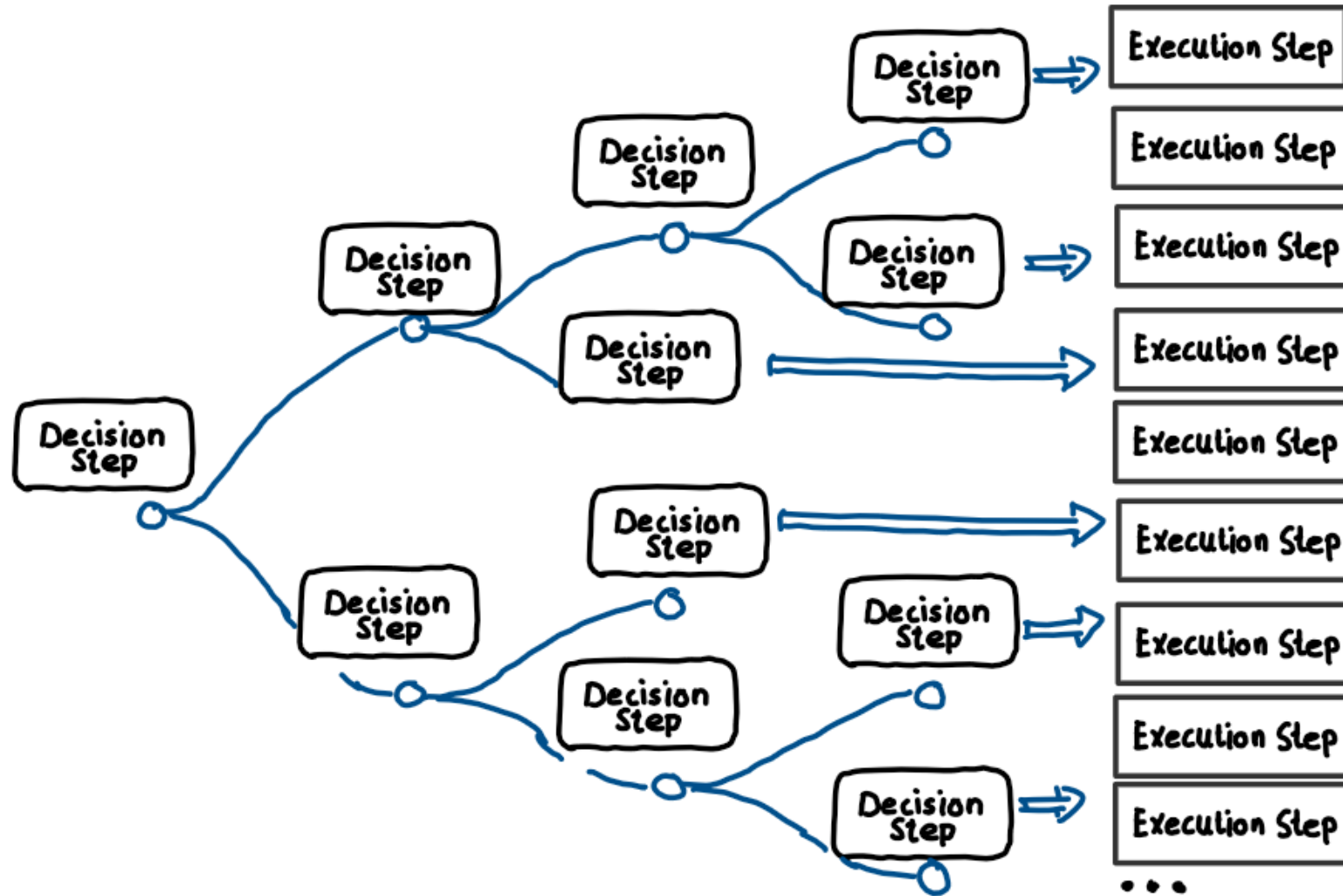
Dynamic Execution

(Make your app agile)

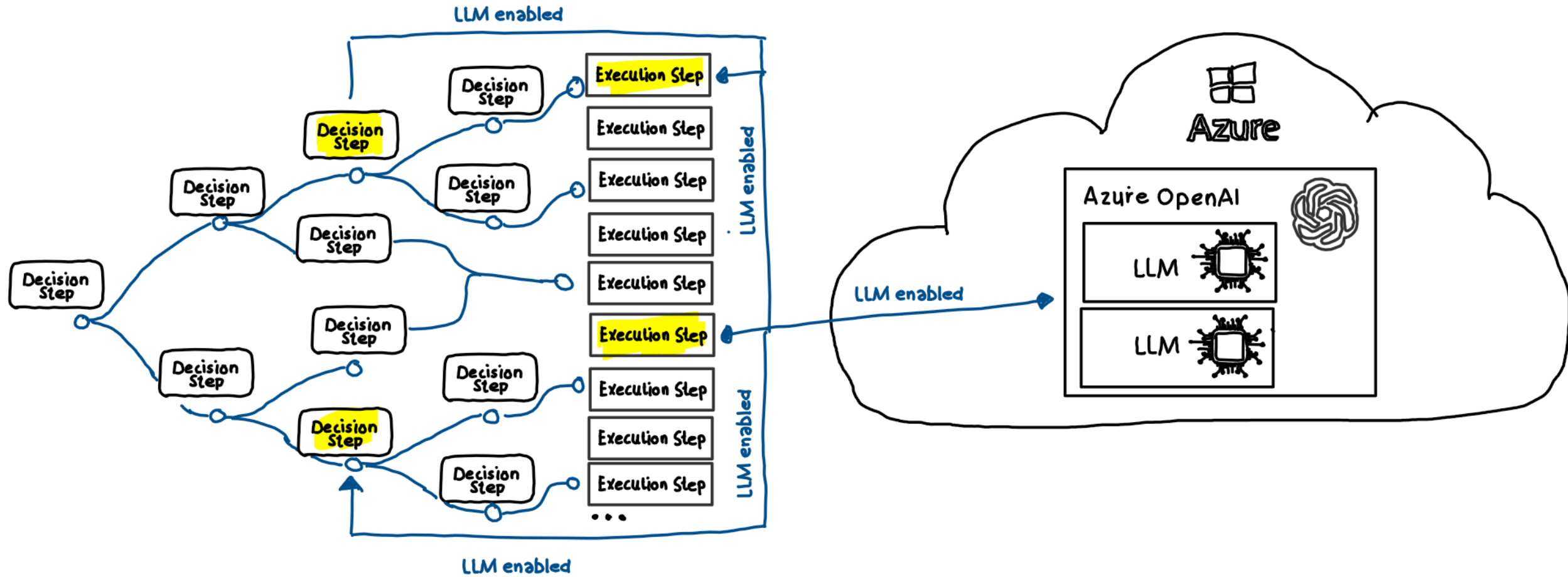


Specific Use case AI supported

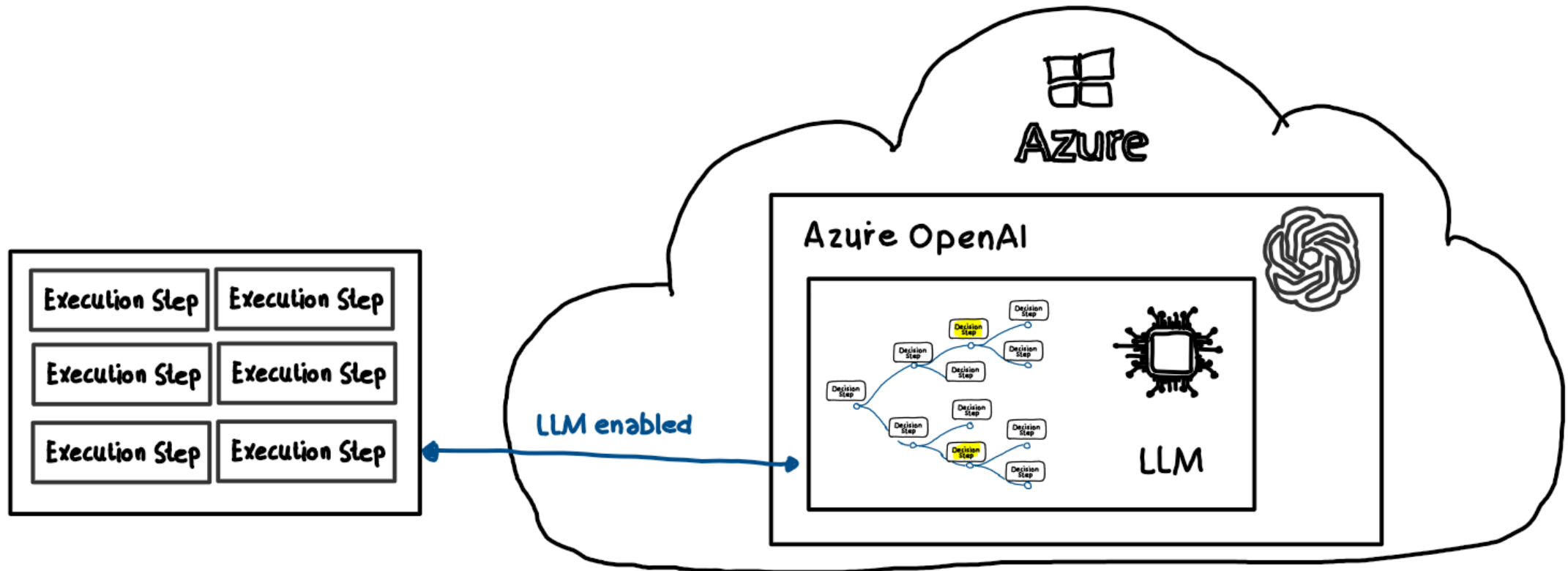
Decision Matrix – Classic Application



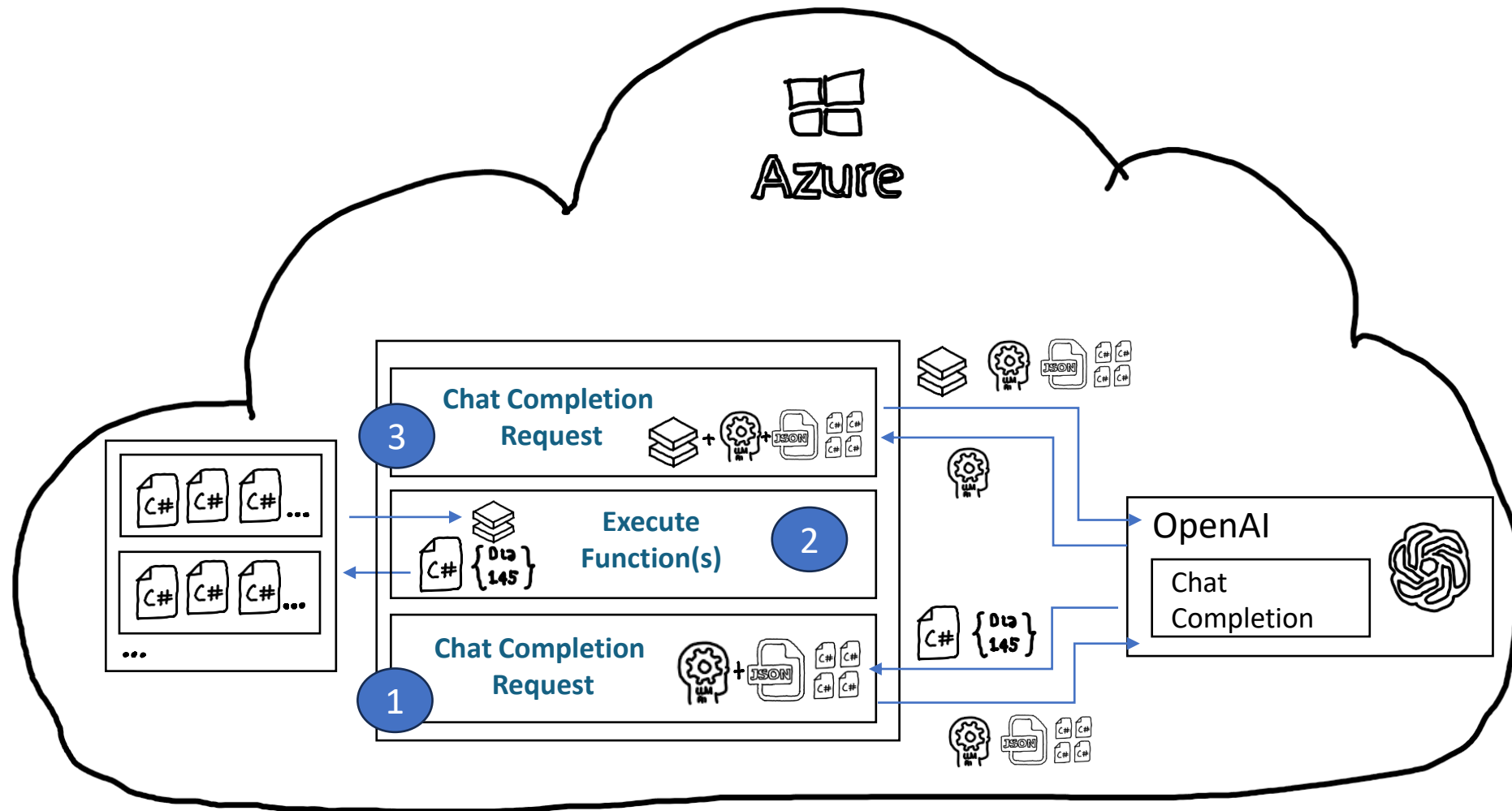
Decision Matrix



Decision Matrix



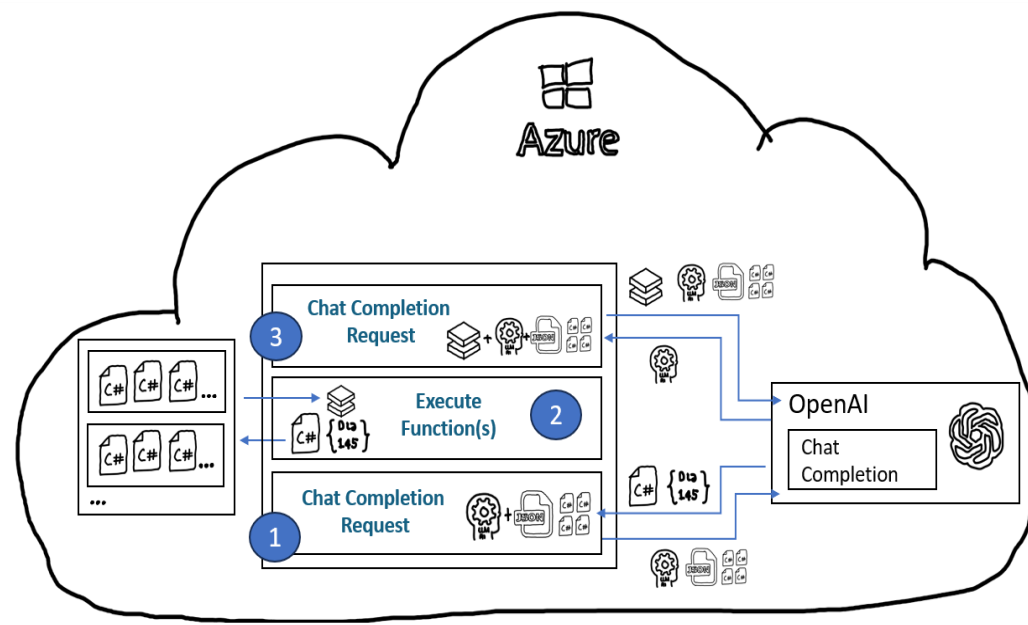
LLM Function Calling / Tools



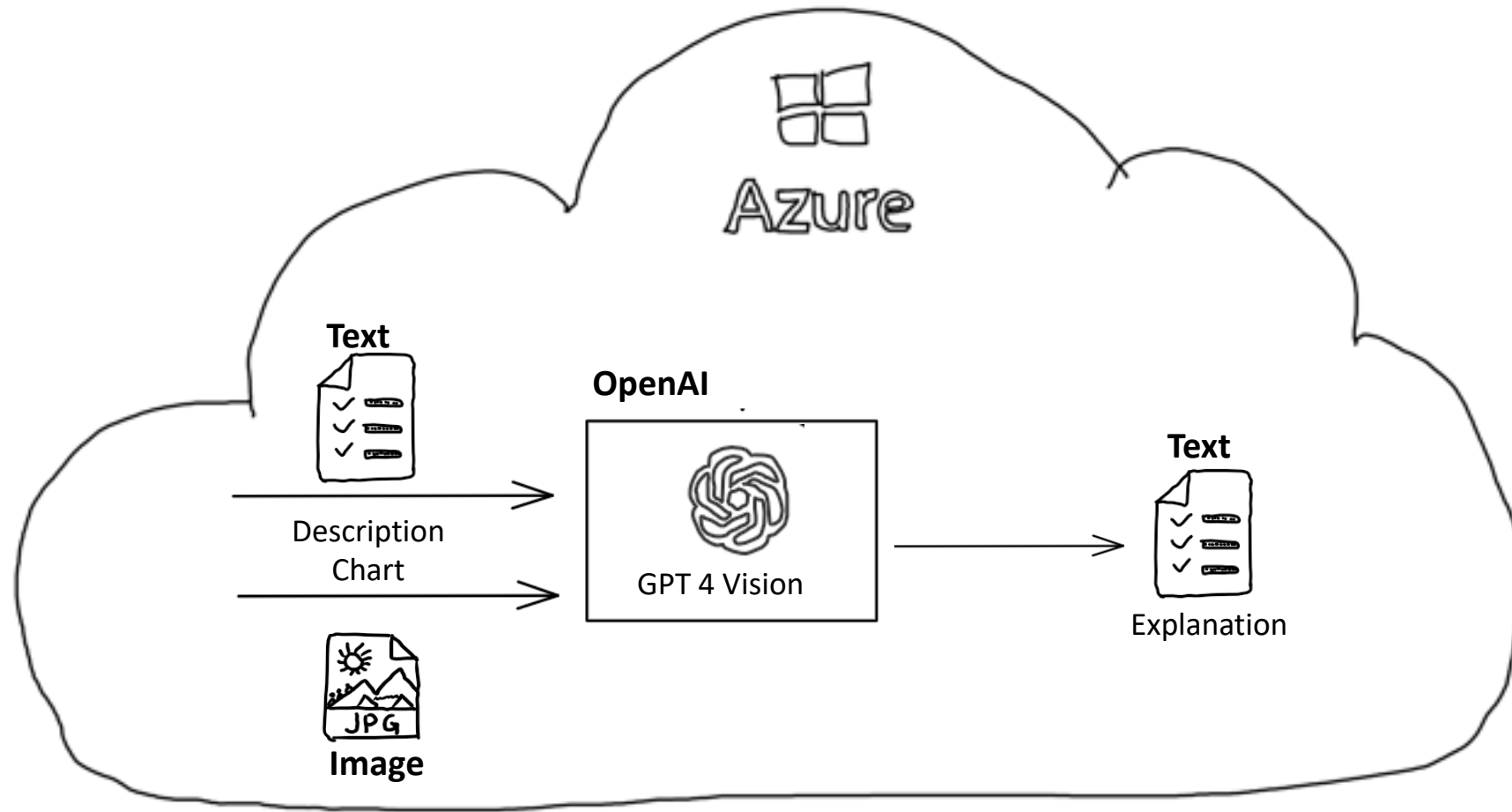
Demo

LLM Function Calling

[aoai-net-starterkit/docs/03_SDK/05_ChatTools.ipynb](#) at main · Azure-Samples/aoai-net-starterkit (github.com)



GPT-4 / Vision



Demo

Analyze Chart

Azure AI | Azure OpenAI Studio

Robert Eichenseer
RobEichGPT4Turbo (Sweden Central, S2)

Privacy & cookies

Deploy to

Chat playground

Import setup Export setup Show panels

Setup

Prompt Add your data

Apply changes

Use a system message template

Select a template

System message

You help users analyzing financial graphs.

Examples

Using examples

Add examples to show the chat what responses you want. It will try to mimic any responses you add here so make sure they match the rules you laid out in the system message.

+ Add

Clear chat Playground settings View code Show JSON

Explain the financial report to me

Allianz 181,24 euros

Revenues (blue) + net income (red) 2009-2016 + estimate 2020

Year	Revenues (blue)	Net Income (red)
2009	~80,000	~2,000
2010	~85,000	~3,000
2011	~90,000	~4,000
2012	~95,000	~5,000
2013	~100,000	~6,000
2014	~105,000	~7,000
2015	~110,000	~8,000
2016	~115,000	~9,000
2017	~120,000	~10,000
2018	~125,000	~11,000
2019	~130,000	~12,000
2020	~135,000	~13,000

valuespectrum.com

Analyst

181,24 euros

This financial graph represents the revenues (in blue) and net income (in red) of Allianz, an insurance company, from the years 2009 to 2016, with an estimate for the year 2020.

From the bar chart (blue bars), we can observe the following about the company's

Type user query here. (Shift + Enter for new line)

Configuration

Deployment Parameters

Deployment

gpt4vision

Enhancements

Vision Azure AI Services

Session settings

Past messages included

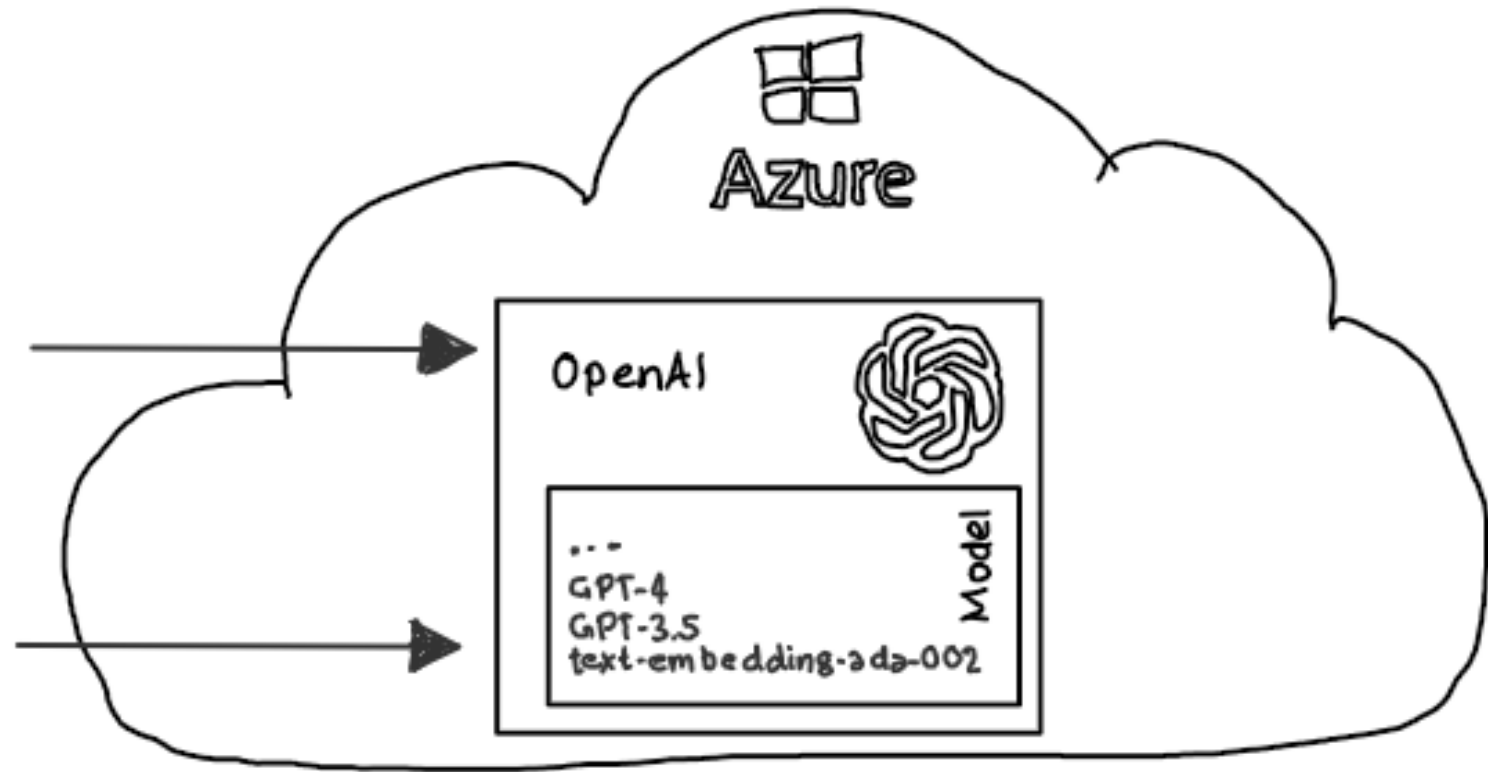
Current token count

10

Input tokens progress indicator

536/128000

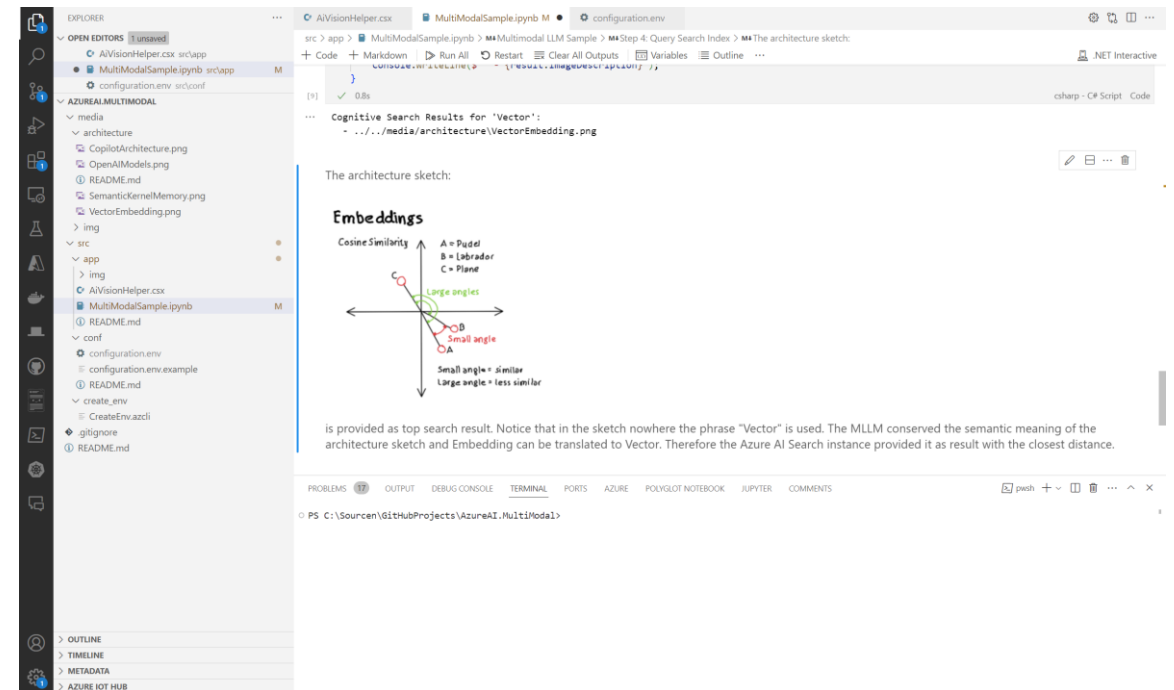
Multi Modal Models



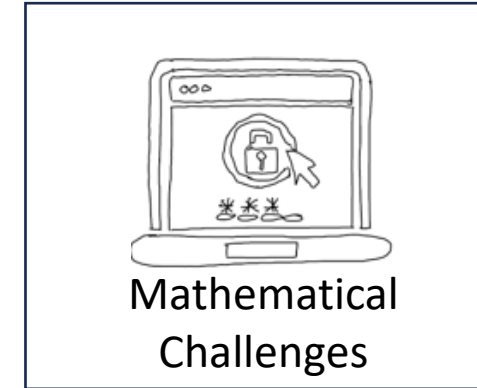
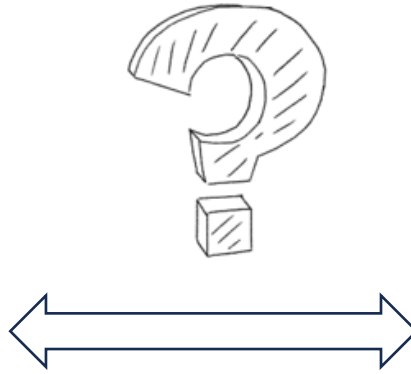
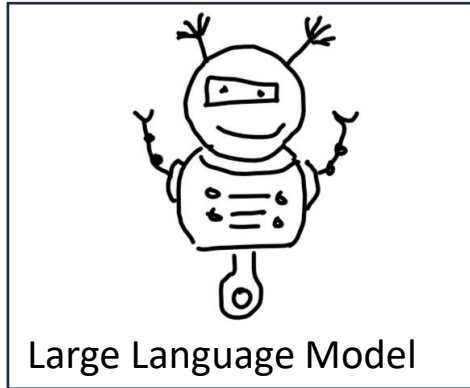
Demo

Multi Modal Embedding

[AzureAI.MultiModal/src/app/MultiModalSample.ipynb](https://github.com/RobertEichenseer/AzureAI.MultiModal/blob/main/src/app/MultiModalSample.ipynb)
at main · RobertEichenseer/AzureAI.MultiModal
(github.com)



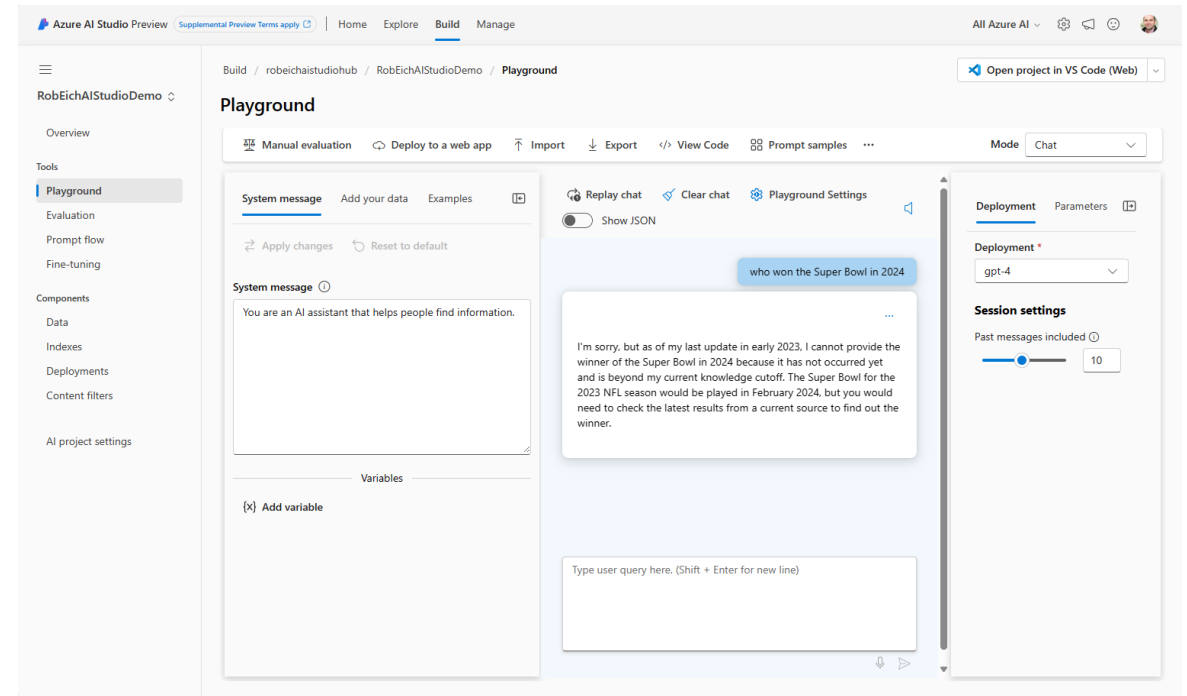
Challenges



Demo

Assistants API

Code Interpreter



Questions

