Microsoft ZUYD

Generative AI
Kennis sessie

Dec 2023

Achieving more
together

Agenda

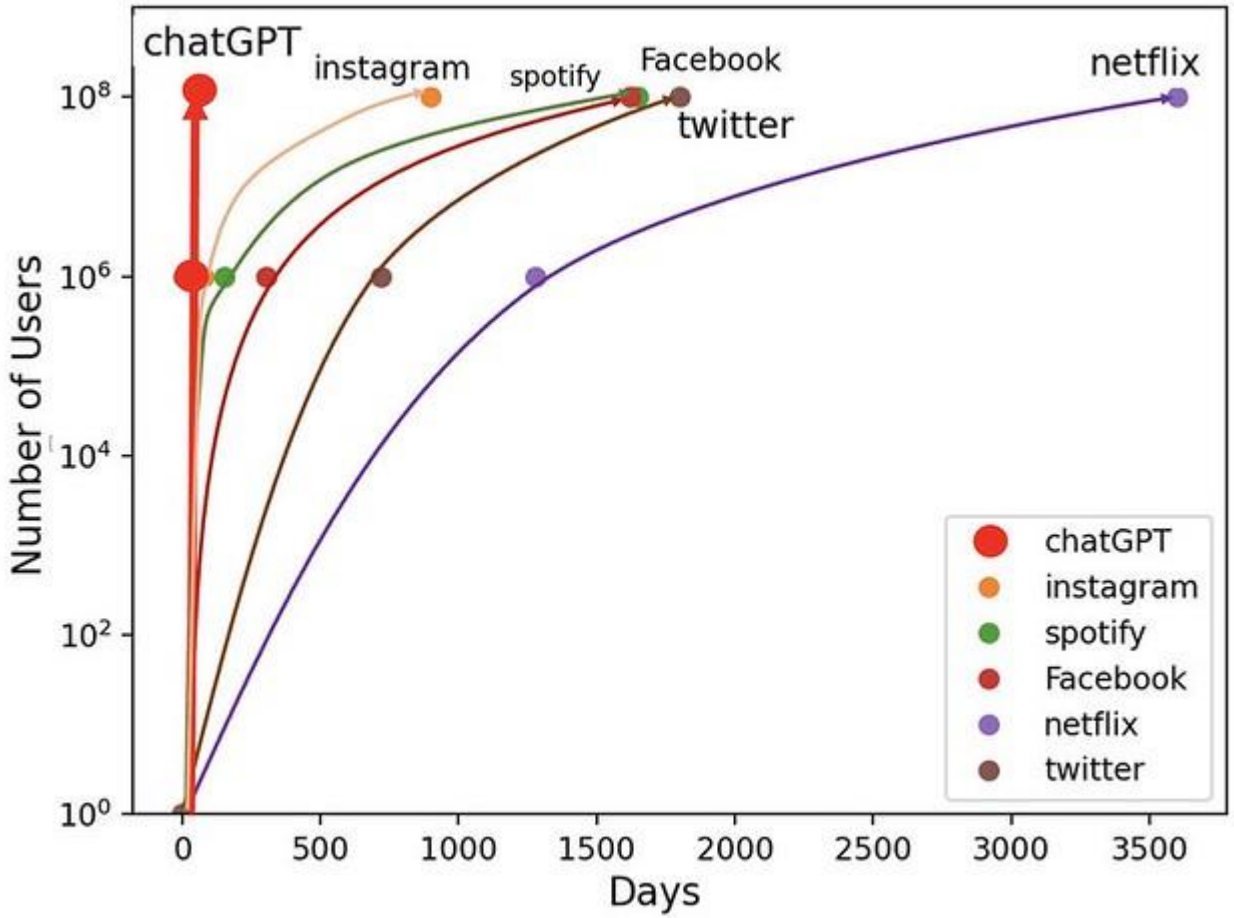
- **Intro**
- **Prompts**
- **Embeddings**
- **Retrieval Augmented Generation (RAG)**

Intro
chatGPT

Chapter slide subtitle

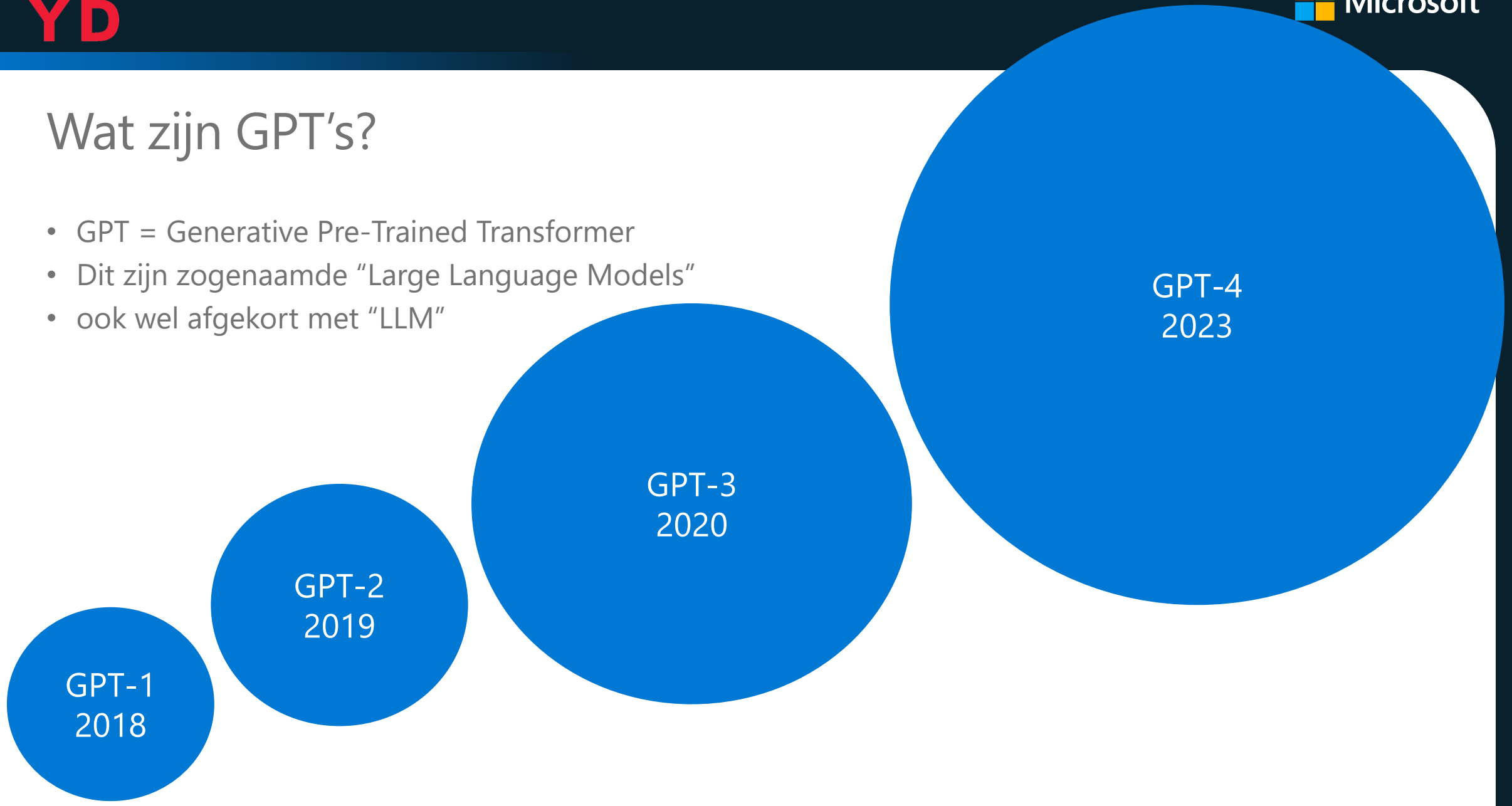
De ChatGPT adoptie explosie

- **100 miljoen gebruikers** in een paar dagen
- Ter vergelijking:
  Instagram heeft daar **een paar jaar** over gedaan



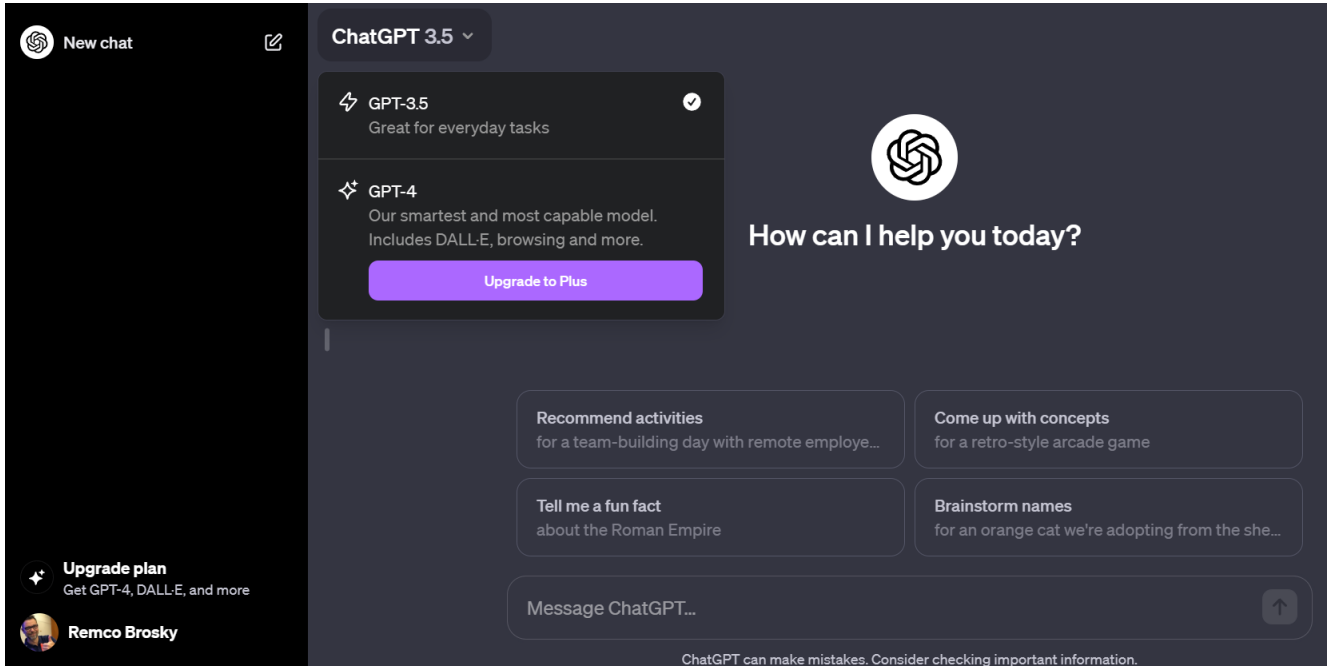
Wat zijn GPT's?

- GPT = Generative Pre-Trained Transformer
- Dit zijn zogenaamde "Large Language Models"
- ook wel afgekort met "LLM"



Wat is ChatGPT?

- ChatGPT is een chat **applicatie**



- ChatGPT is een finetuned **model**, gebaseerd op GPT3 en nu ook beschikbaar op basis van GPT4

Wat doet zo'n model?

- Genereer het eerstvolgende waarschijnlijke token

  Laten we eens kijken wat een token is:
  https://platform.openai.com/tokenizer

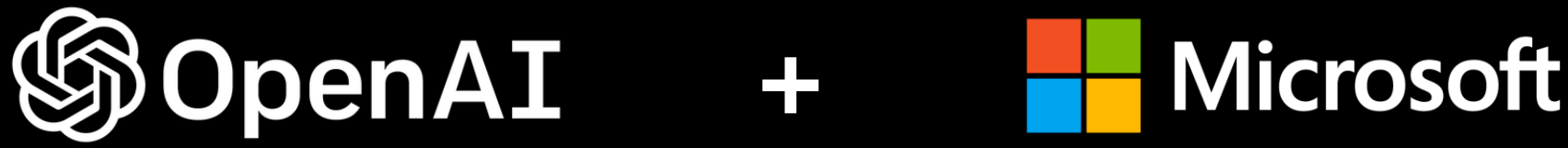
- OpenAI is niet de enige met dit soort modellen
- Zie https://huggingface.co/ voor veel meer modellen (Open Source)

Hugging Face

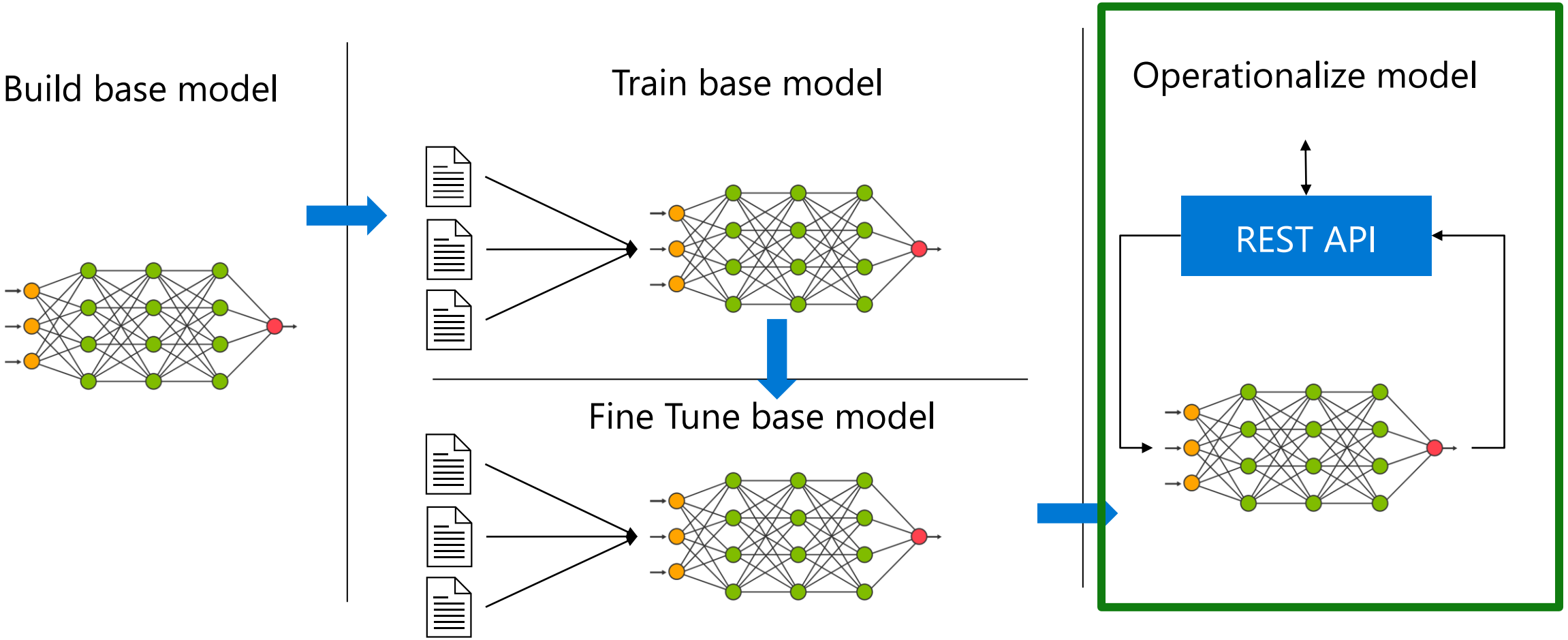
Tokens = $$$

Model	Input	Output
gpt-3.5-turbo-1106	$0.0010 / 1K tokens	$0.0020 / 1K tokens
gpt-3.5-turbo-instruct	$0.0015 / 1K tokens	$0.0020 / 1K tokens

Model	Input	Output
gpt-4-1106-preview	$0.01 / 1K tokens	$0.03 / 1K tokens
gpt-4-1106-vision-preview	$0.01 / 1K tokens	$0.03 / 1K tokens


OpenAI
+
Microsoft

Model lifecycle



ZU
YD

Microsoft

Prompts
Prompt Engineering

Wat is een prompt?

ChatGPT 3.5

User prompt

You

Wat is het adres van de campus hogeschool Zuyd in Heerlen

ChatGPT

Maar er is meer…

System Prompt (MetaPrompt)

Je bent een behulpzame AI Assistant

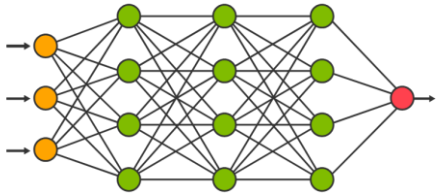
Je naam is Zuydbot

Je antwoord alleen in het Nederlands

+

User Prompt

Wat is het adres van de campus van de hogeschool Zuyd in Heerlen



Response

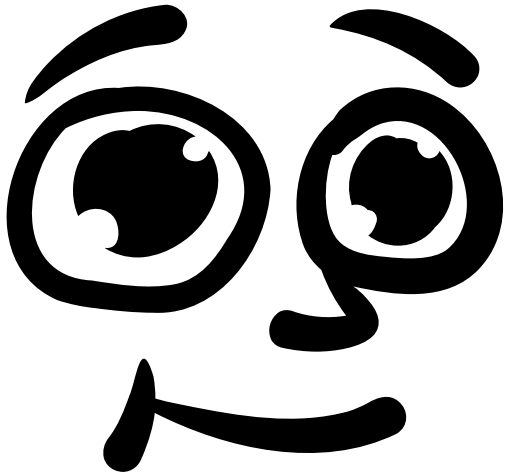
....

Hoe meer context en instructies – hoe beter (en voorspelbaarder) de response zal zijn

ZU
YD

Microsoft

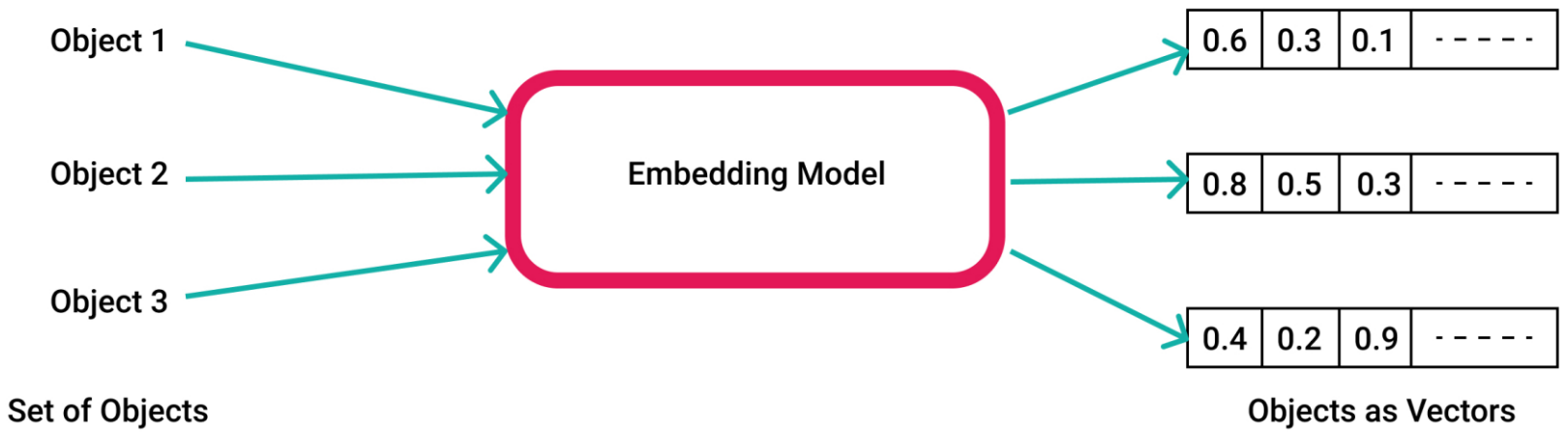
Demo time...



Embeddings
Vectoren en search

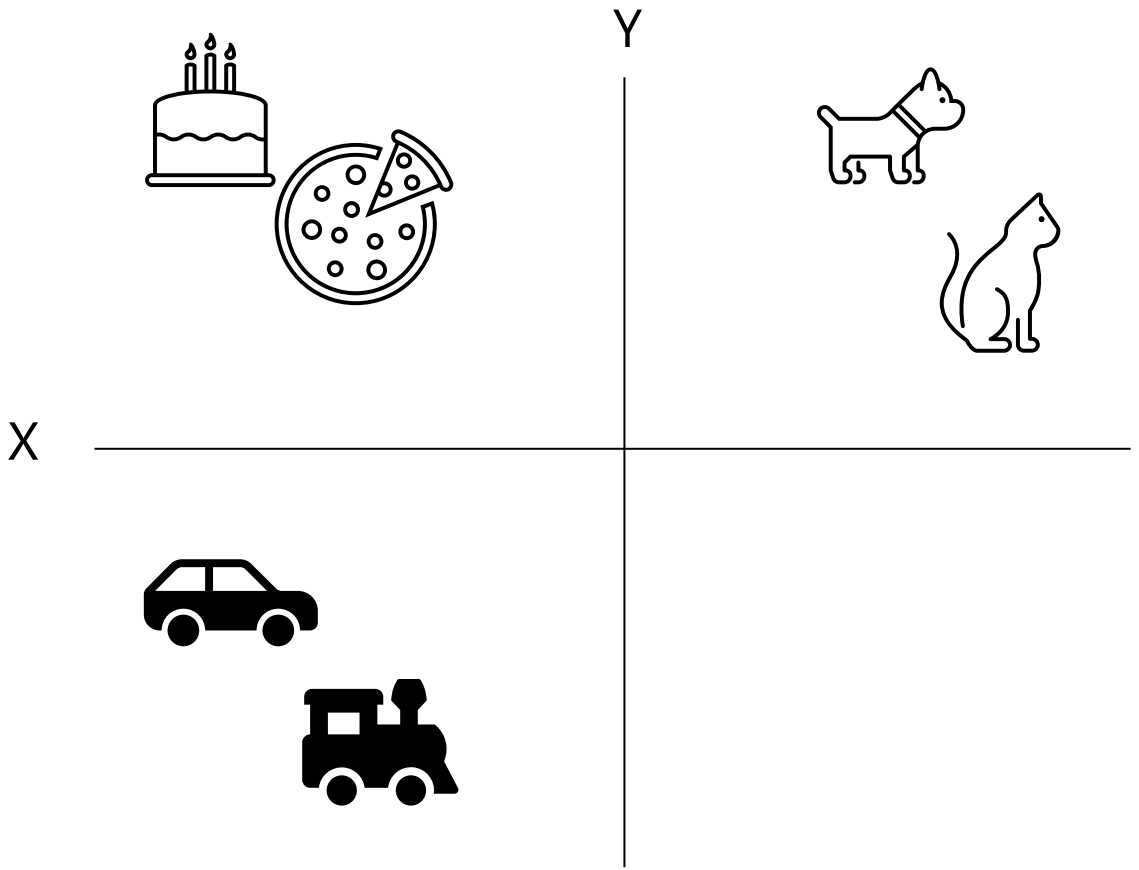
Semantische gelijkheid

- Kat
- Pizza
- Trein
- Hond
- Taart
- Auto



Semantische gelijkheid

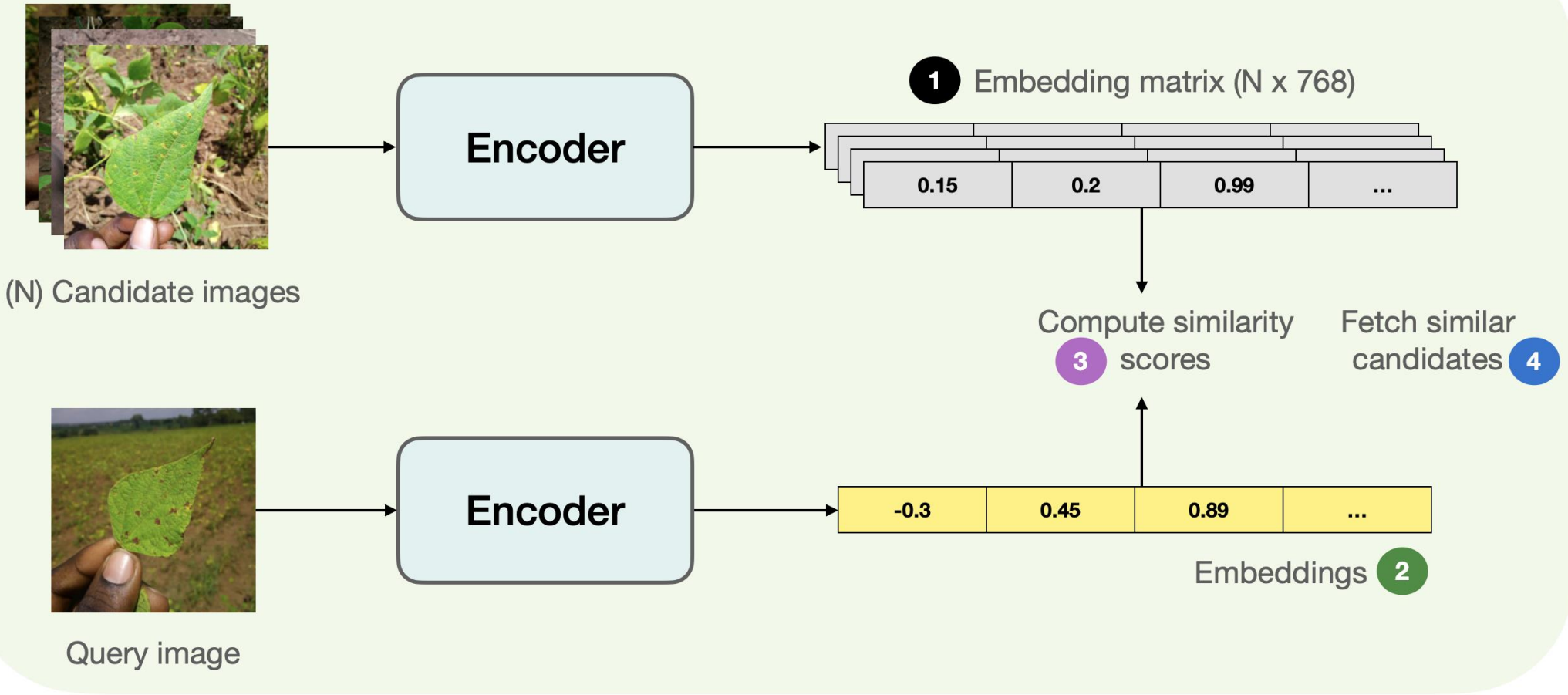
- Kat
- Pizza
- Trein
- Hond
- Taart
- Auto



Embedding modellen

- Een embedding model – bijv. OpenAI text-embedding-ada-002 – is geen “chat model” maar heeft wel kennis van taal. Hiermee kun je alleen de vector van een tekst berekenen.
- text-embedding-ada-002 berekend een vector van een gegeven tekst met **1536** dimensies
- Er zijn verschillende embedding modellen, met verschillende hoeveelheid dimensies en getraind op verschillende datasets.
- Als je vectoren wil vergelijken moeten ze berekend zijn met hetzelfde model om “semantische gelijkheid” te bepalen
- Vectoren berekenen en vergelijken kan ook met plaatjes, video en audio

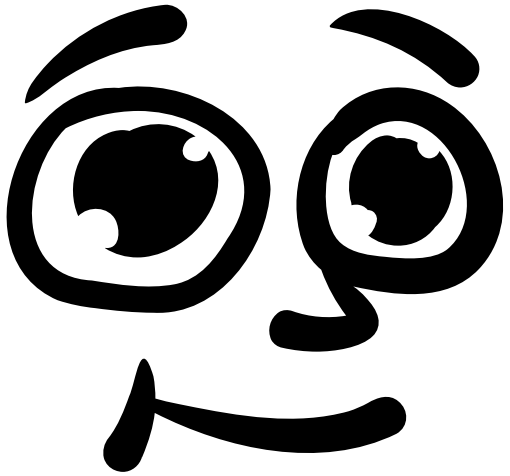
Hoe het werkt



ZU
YD

Microsoft

Demo time...

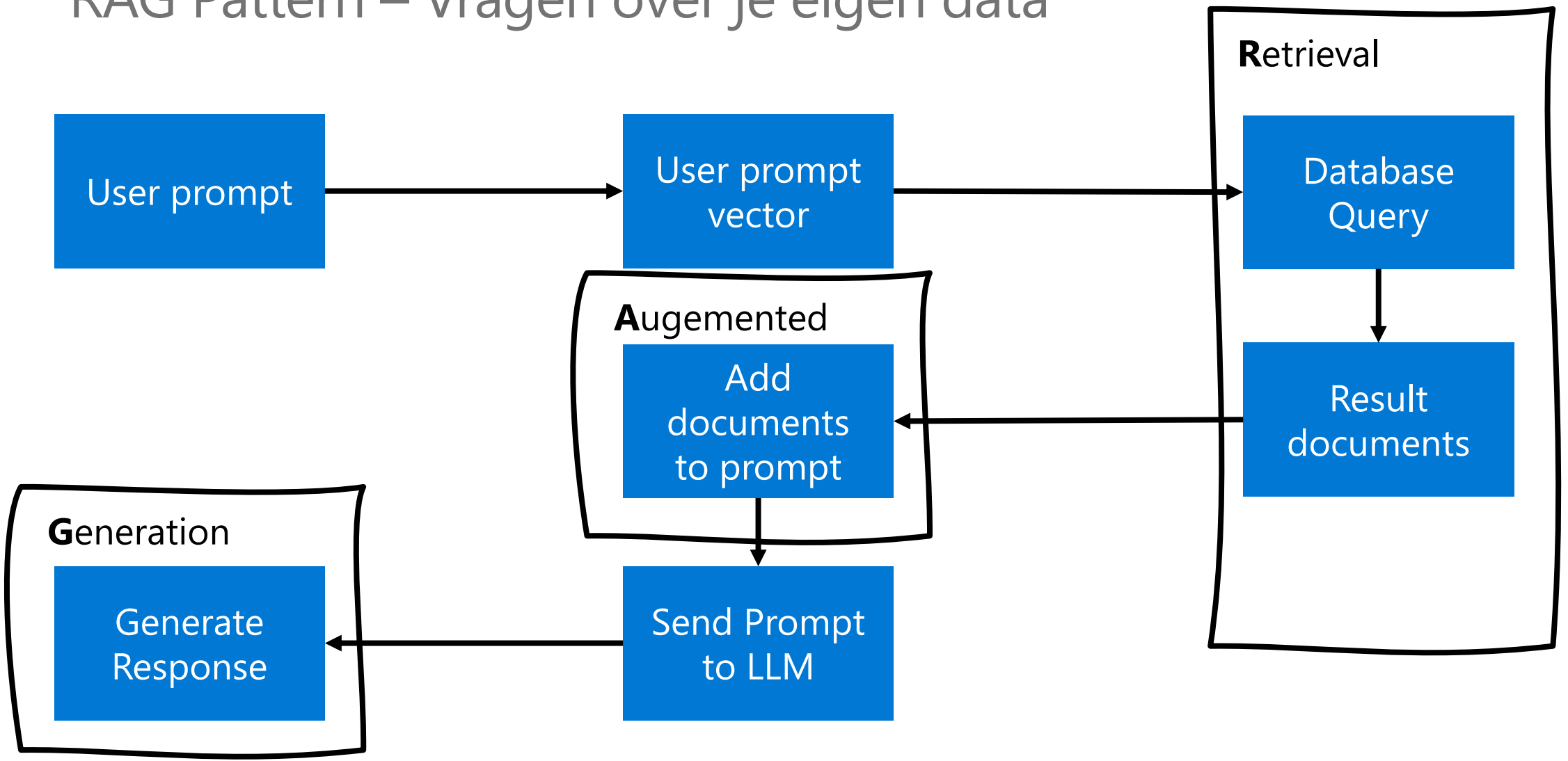


ZU
YD

Microsoft

RAG
Retrieval Augmented Generation

RAG Pattern – Vragen over je eigen data



Orchestrators

Alle onderdelen/bouwblokken van RAG hebben we al gezien

Orchestrators zijn libraries die helpen je om (o.a.) RAG te implementeren in je code

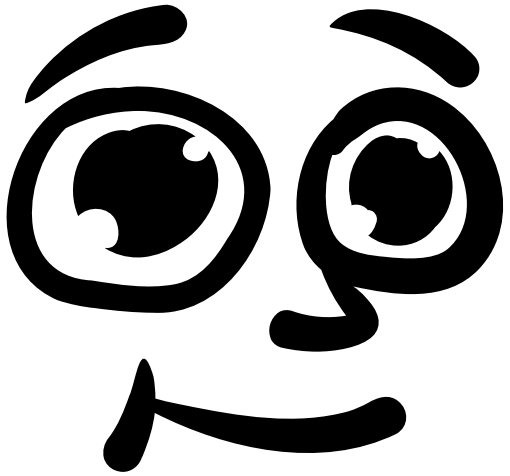
Voorbeelden van Orchestrators

- Langchain (Python)
- Semantic Kernel (C#, Python en Java)

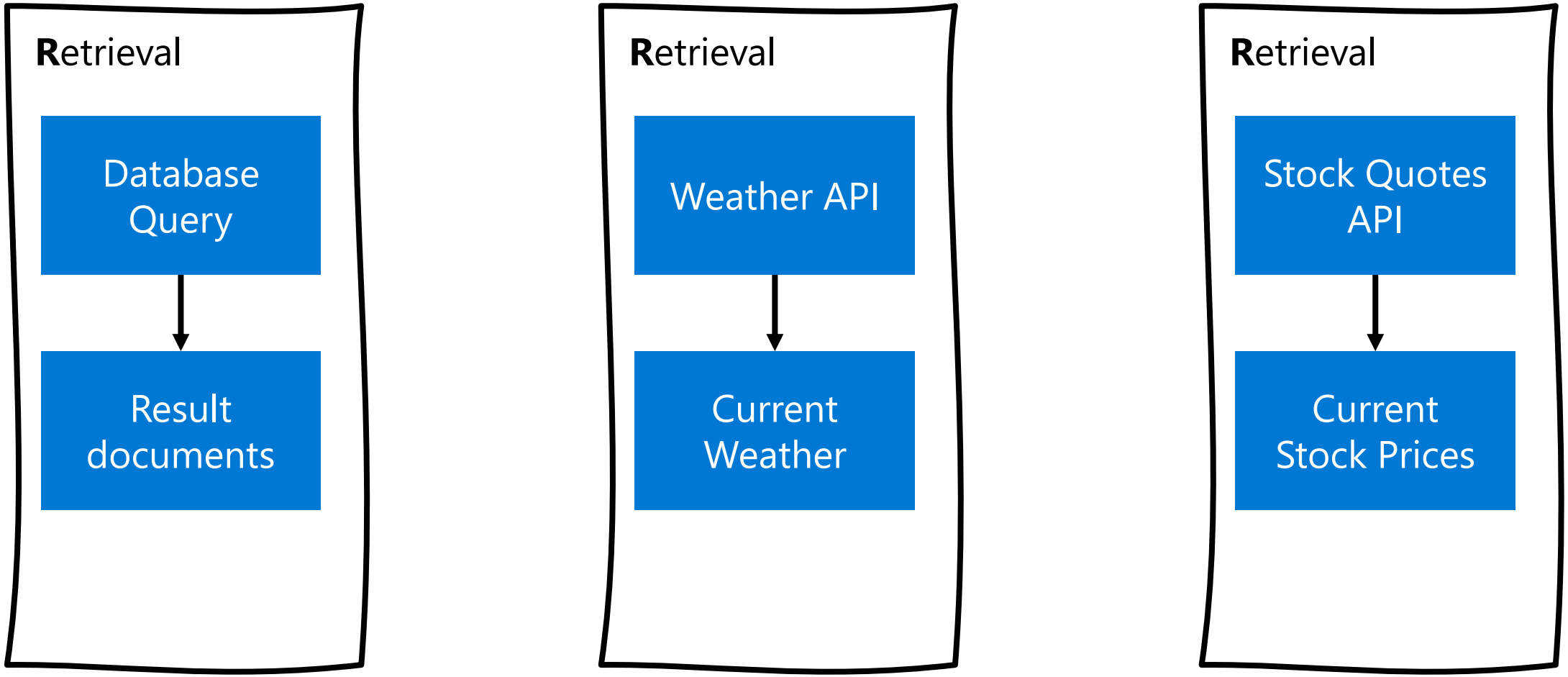
ZU
YD

Microsoft

Demo time...



Inspiratie – Je kunt nog veel meer data injecteren



Microsoft

Achieving more **together**

Referenties

- Azure AI Hardware - (https://www.youtube.com/watch?v=Rk3nTUfRZmo)
- Andere (Open Source) modellen - https://huggingface.co/
- AI Modellen op je eigen machine - https://ollama.ai/
- Langchain - https://python.langchain.com/docs/get_started/introduction
- Semantic Kernel - https://learn.microsoft.com/en-us/semantic-kernel/overview/
- Autogen - https://github.com/microsoft/autogen
- MongoDb - https://www.mongodb.com/products/platform/atlas-vector-search
- Azure AI Search - https://learn.microsoft.com/en-us/azure/search/search-what-is-azure-search