



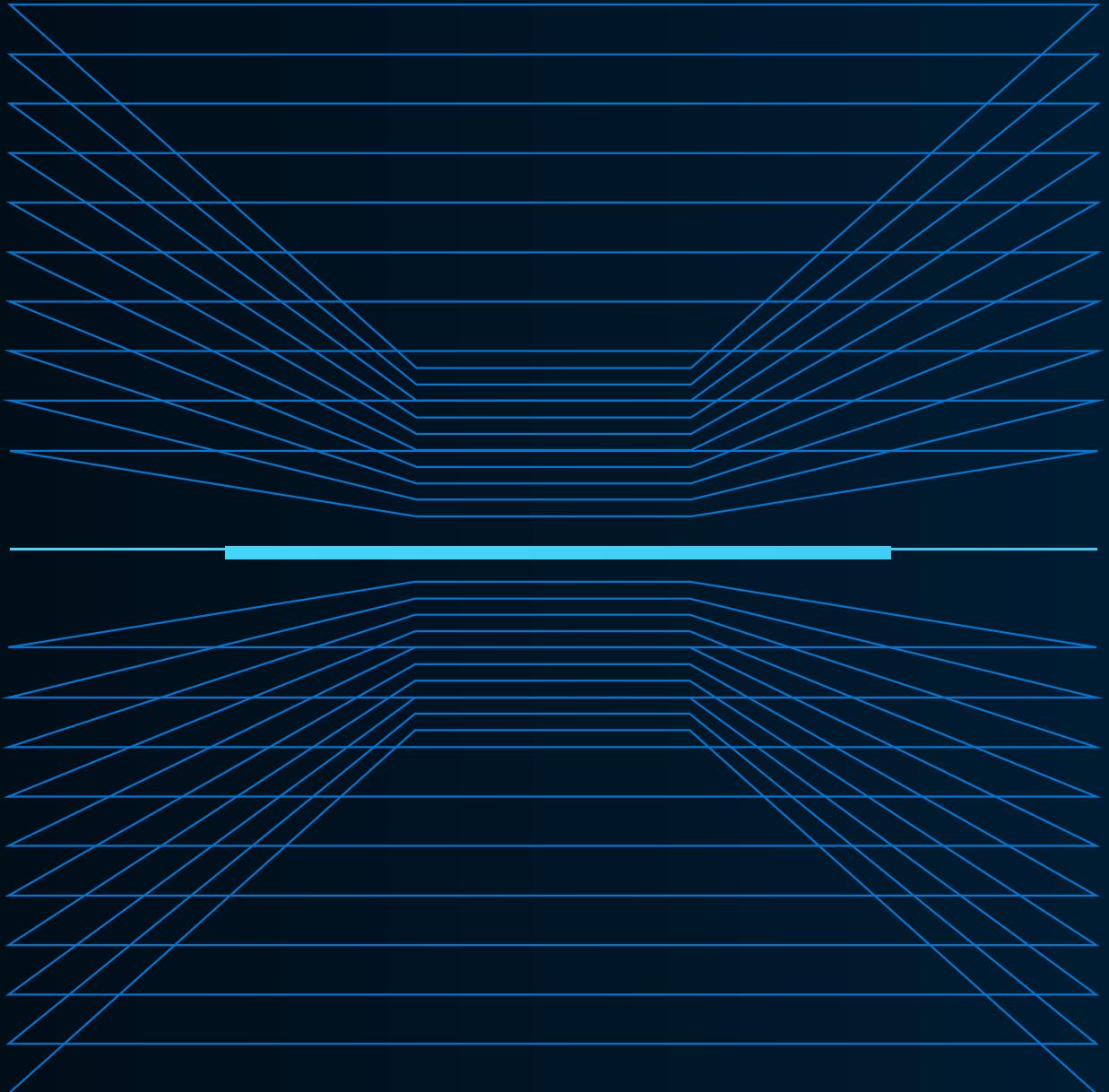
# Azure OpenAI Service Deep-Dive Workshop

**Andreas Hopfgartner**  
[andreas.hopfgartner@microsoft.com](mailto:andreas.hopfgartner@microsoft.com)

Cloud Solution Architect Data & AI

**Philipp Hinderberger**  
[philipp.hinderberger@microsoft.com](mailto:philipp.hinderberger@microsoft.com)

Cloud Solution Architect Data & AI



# Goal for today

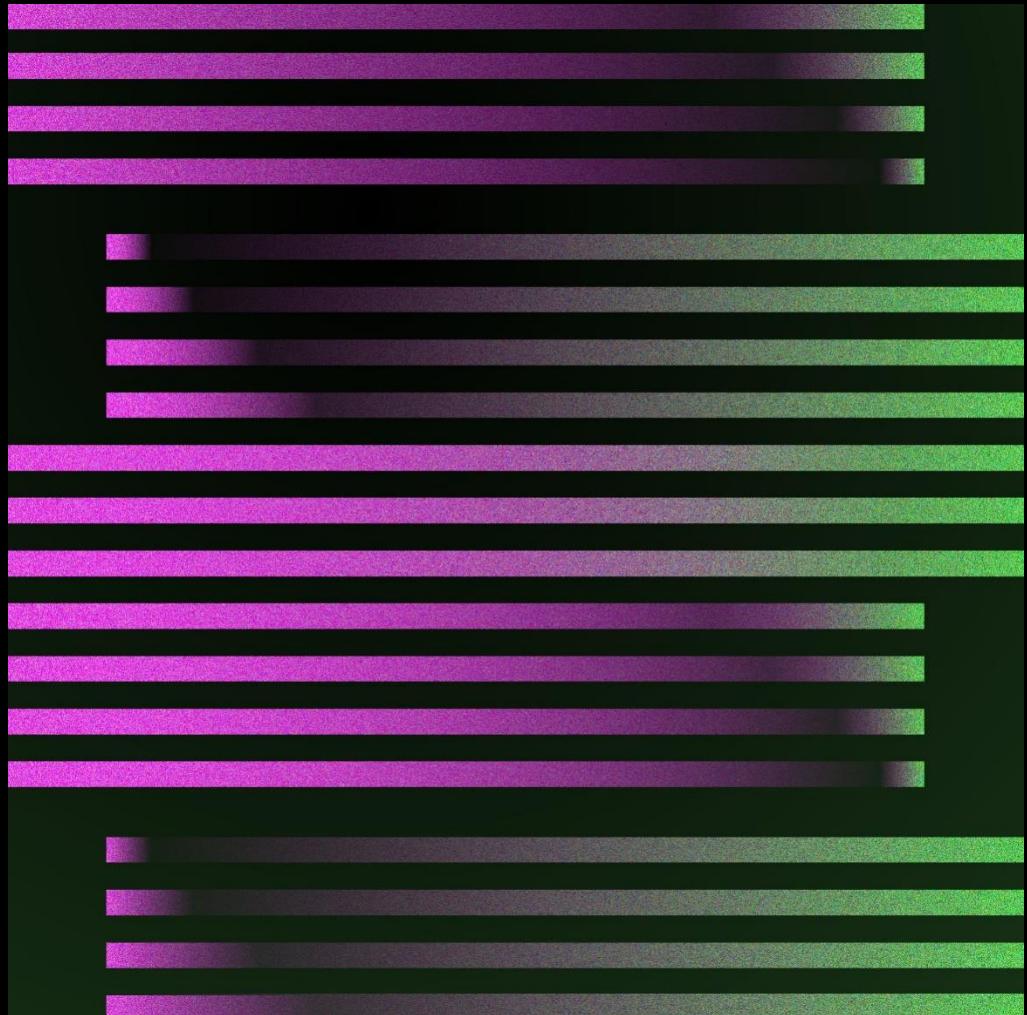
- Theory
  - Azure OpenAI Service
  - Completions
  - Embeddings
  - Few-shot learning & fine-tuning
  - MVP Development
- Hands-on
  - 10+ prompt engineering exercises
  - Q&A with embeddings
  - Movies unsupervised classification & recommendation system

# Plan for today

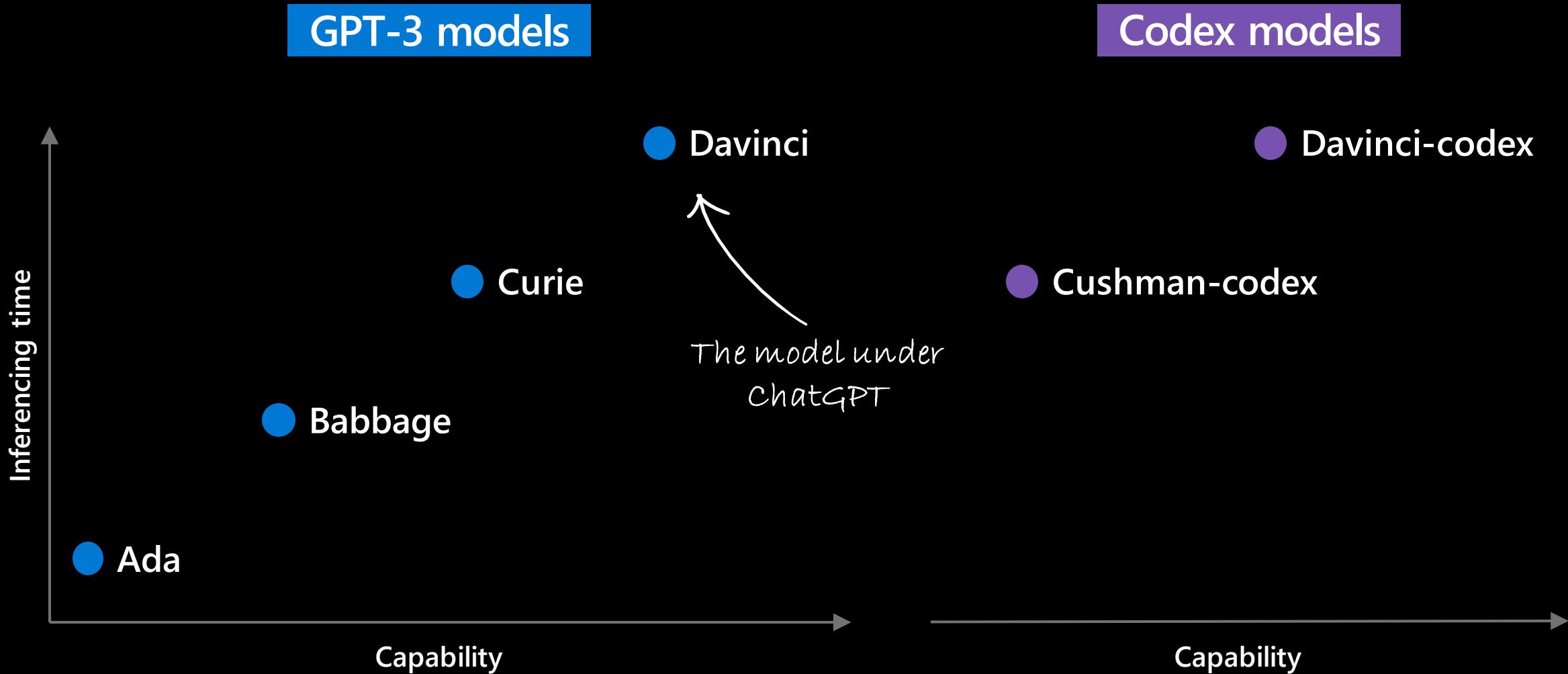
Time (CET)	Topic
09:00 – 09:45	Intro OpenAI
09:45 – 10:15	Azure OpenAI Studio
10:15 – 10:30	Break
10:30 – 12:00	Prompt Engineering Exercises
12:00 – 13:00	Lunch Break
13:00 – 13:30	Use Cases and Common Architectures
13:30 – 14:15	Using OpenAI to access company data
14:15 – 14:45	MVP Development
14:45 – 15:00	Break
15:00 – 16:30	Hands-on lab on two exemplary use-cases

# Why so much hype around ChatGPT?

- ChatGPT
  - 100m users in 2.5 months
  - Most “human-like” interaction
- What’s new about it?
  - Can **generate** and **do** “things”
  - Answer **knowledge** questions directly
  - Answer **coding** questions
  - Can **correct itself**, if you ask it
- Prompt examples [here](#)

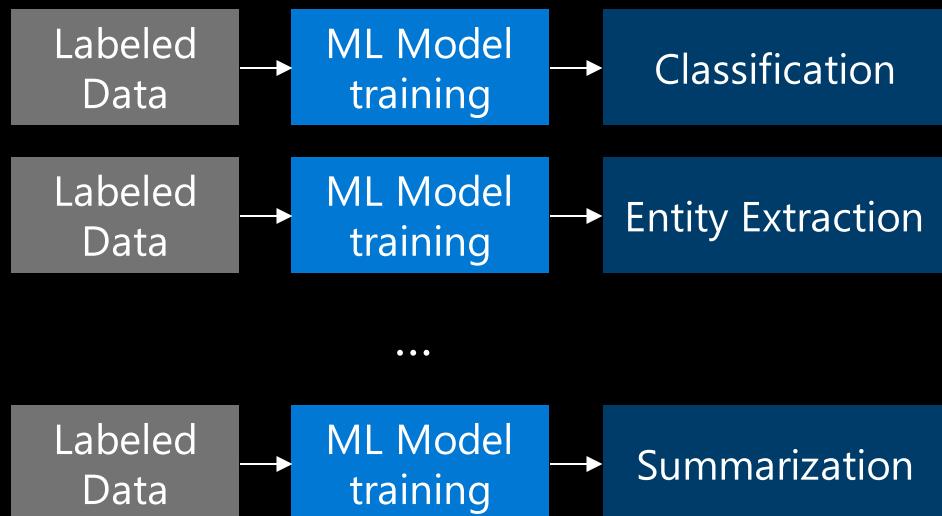


# OpenAI Models



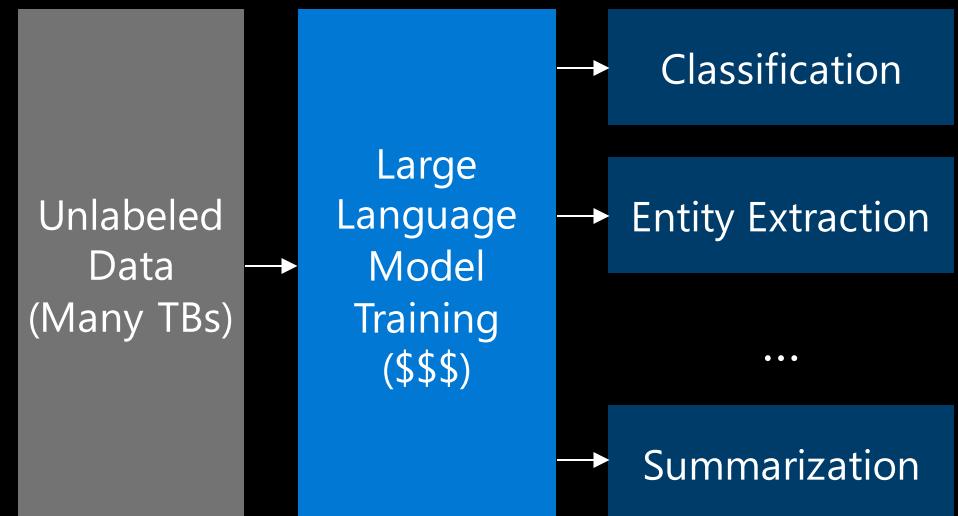
# Large Language Models

## Typical ML for NLP



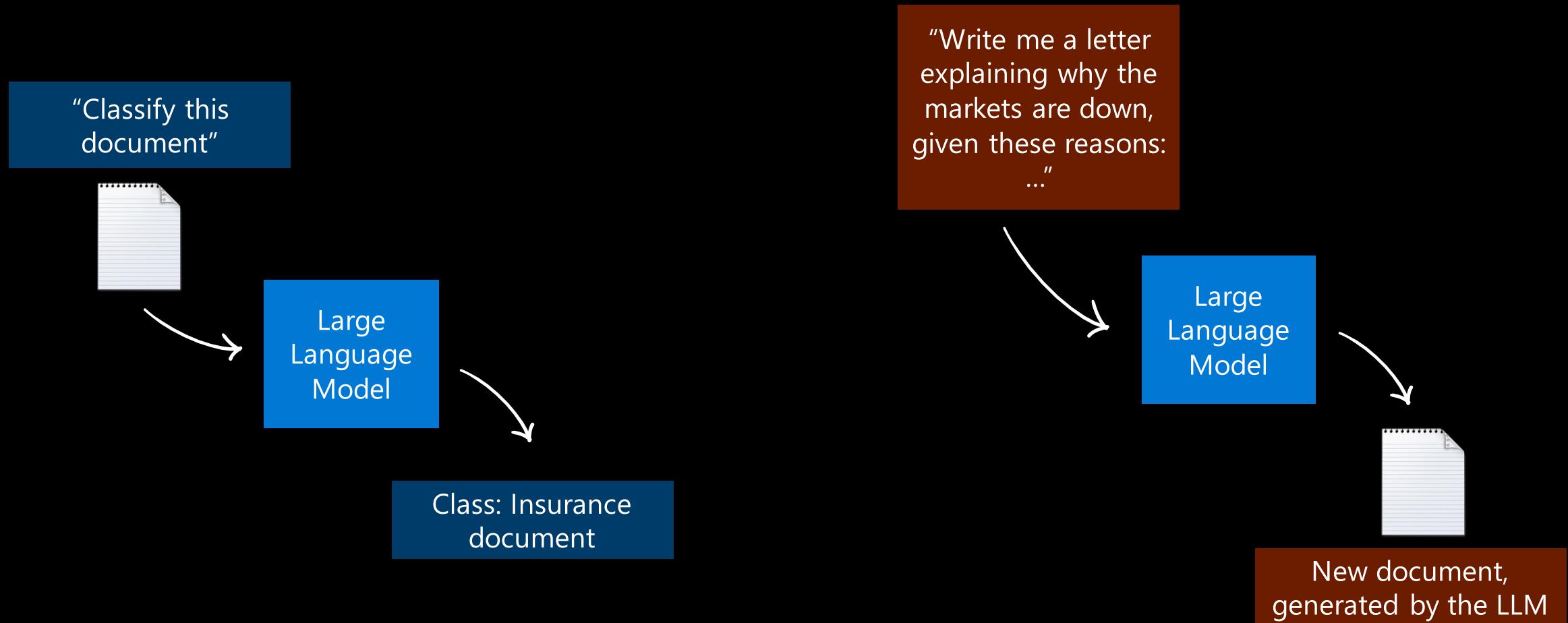
One model per capability  
Labeled data to train  
Highly optimized for use case

## Large Language Models



Single model for all use cases  
Describe in natural language what it should do

# LLMs can also “generate things”



LLMs can do typical ML tasks  
easily (often unsupervised)

LLMs can also machine-read (understand)  
and generate new documents

# Artificial Intelligence

## Machine Learning

## Deep Learning

## Generative AI



## Artificial Intelligence

the field of computer science that seeks to create intelligent machines that can replicate or exceed human intelligence



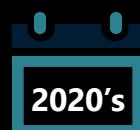
## Machine Learning

subset of AI that enables machines to learn from existing data and improve upon that data to make decisions or predictions



## Deep Learning

a machine learning technique in which layers of neural networks are used to process data and make decisions



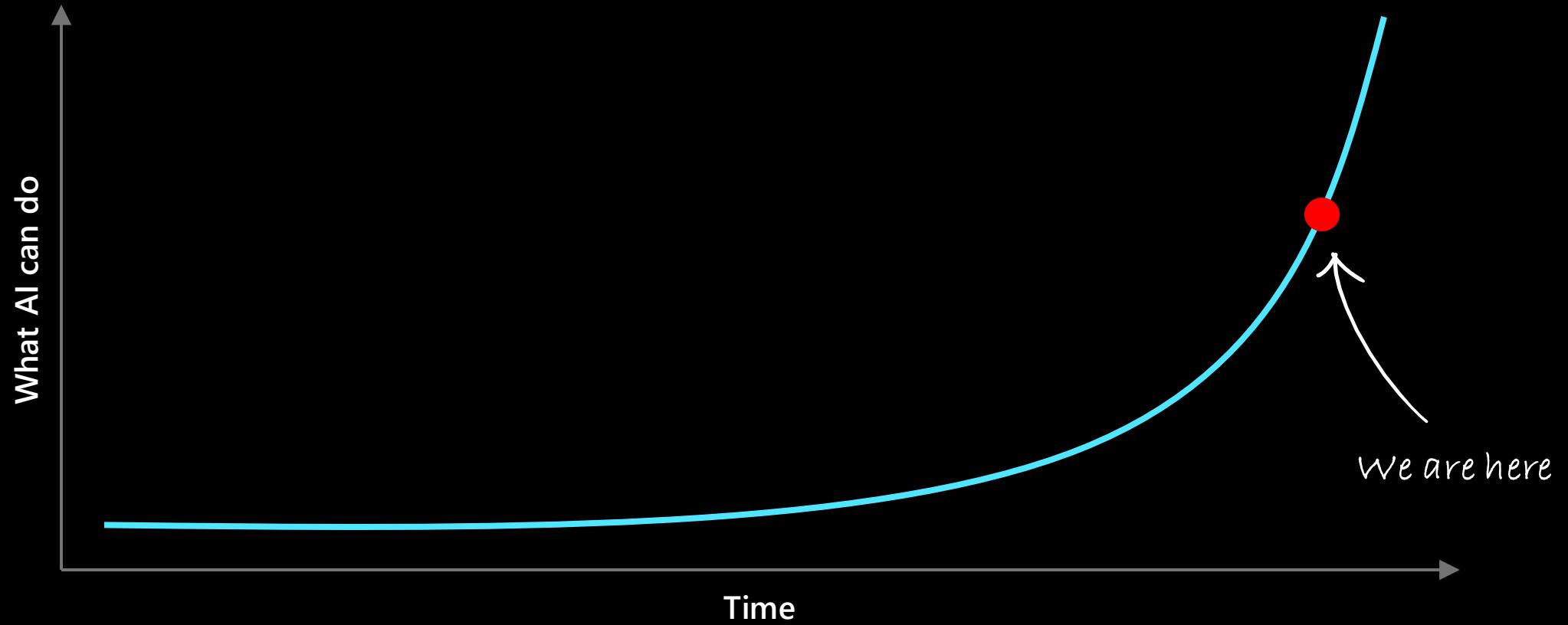
## Generative AI

Create new written, visual, and auditory content given prompts or existing data.

Exponential improvement

Exponential improvement

Exponential improvement



Tell me about Azure OpenAI Service from Microsoft!

Azure OpenAI Service is a product that allows you to access powerful language models from OpenAI, such as GPT-3, Codex and DALL-E, through a REST API [1](#). It is part of Azure Cognitive Services and runs on Azure global infrastructure [2](#). What are you interested in learning more about Azure OpenAI Service?

Learn more: [1. learn.microsoft.com](#) [2. azure.microsoft.com](#) [3. learn.microsoft.com](#) [+3 more](#) 1 of 6

How does Microsoft cooperate with OpenAI?

Microsoft and OpenAI have a long-term partnership that involves several aspects [1](#) [2](#). Microsoft invested \$1 billion in OpenAI in 2019 [1](#) [3](#) and became its exclusive cloud provider [4](#) [3](#). The two companies also collaborate on developing AI supercomputing technologies [4](#) [2](#) and licensing some of OpenAI's technology to Microsoft for commercialization [4](#). Azure OpenAI co-develops the APIs with OpenAI, ensuring compatibility and a smooth transition from one to the other [5](#). What else would you like to know about Microsoft and OpenAI?

Learn more: [1. forbes.com](#) [2. blogs.microsoft.com](#) [3. openai.com](#) [4. forbes.com](#) [+2 more](#) 2 of 6

Did Microsoft give more funding in 2023?

Yes, Microsoft confirmed a new multi-year, multi-billion dollar (reported to be \$10 billion) investment in OpenAI on January 23, 2023 [1](#) [2](#). This was the third phase of their long-term partnership and followed their previous investments in 2019 and 2021 [1](#) [2](#). Why are you curious about Microsoft's funding for OpenAI?

Learn more: [1. en.wikipedia.org](#) [2. blogs.microsoft.com](#) [3. en.wikipedia.org](#) [+2 more](#) 3 of 6



Ask me anything...



*Ensure that artificial general intelligence (AGI) benefits humanity.*

Groundbreaking research  
Latest models  
Fast innovation

**GPT-3**

Generate and Understand Text

**Codex**

Generate and Understand Code



*Empower every person and organization on the planet to achieve more*

Enterprise features  
Production SLAs  
Data privacy

**DALL·E** preview

Generate images from text prompts

**ChatGPT** preview

Chat version of GPT-3

## Generative AI

## GPT-3

Prompt:

Write a tagline for an ice cream shop.

Response:

We serve up smiles with every scoop!

## Codex

Prompt:

```
Table customers, columns =  
[CustomerId, FirstName,  
LastName, Company, Address,  
City, State, Country,  
PostalCode]
```

Create a SQL query for all customers in Texas named Jane  
query =

Response:

```
SELECT *  
FROM customers  
WHERE State = 'TX' AND  
FirstName = 'Jane'
```

## DALL·E

Prompt: A white Siamese cat



## ChatGPT

What is the fastest animal on land?

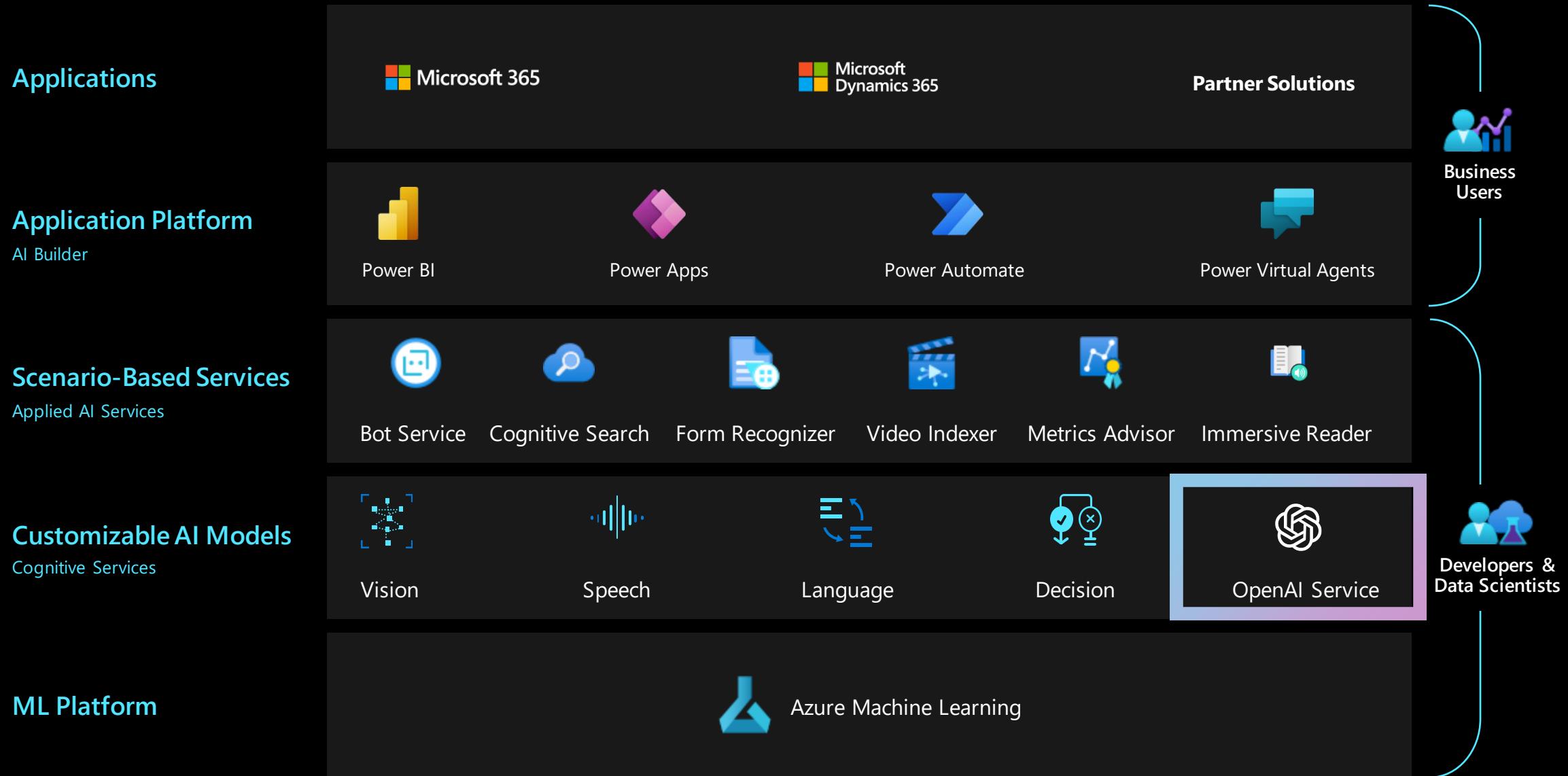
the cheetah (*Acinonyx jubatus*), which can reach speeds of up to 60 miles (97 kilometers) per hour.

What makes them so fast?

Cheetahs are built for speed and have several adaptations that make them the fastest land animal: lean body, long legs, flexible spine, large nostrils and claws that don't retract.



# Azure AI



# Azure OpenAI Service

GPT-3

DALL-E preview

Codex

ChatGPT preview



Deployed within your Azure subscription, secured by you, accessed only by you, and tied to your datasets and applications



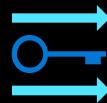
Large, pretrained AI models to unlock new scenarios



Custom AI models fine-tuned with your data and hyperparameters



Built-in responsible AI to detect and mitigate harmful use



Enterprise-grade security with role-based access control (RBAC) and private networks

# Demo Azure OpenAI Studio

# Break

# Azure OpenAI | GPT-3 Prompt Design

Extract the mailing address from this email:

Hi John Doe,

It was great to meet up at Build earlier this week. I thought the AI platform talk was great and I really enjoyed it.

I appreciate the offer for the book. If you are OK, you can mail it to me at home, or 123 Microsoft Way, Bellevue WA 92004.

Regards,

Chris Hoder

**Prompt** – Text input that provides some context to the engine on what is expecting.

**Completion** – Output that GPT-3 generates based on the prompt.

# Hands-on

- Signup
  - Go to [Azure OpenAI Studio - Microsoft Azure](#)
  - Walk through Text Completion tutorial (if needed)
    - <https://beta.openai.com/docs/guides/completion>
- Then do the exercises on the following slides
  - Each slide has some text and a desired completion (output)
  - Try to write a prompt, that achieves that completion
  - You'll find the solution for each exercise in the speaker notes
  - You'll also find the exercises here:
    - [openai-in-a-day/exercises.md at main · anderl80/openai-in-a-day \(github.com\)](https://github.com/anderl80/openai-in-a-day)

# Exercise 1 – German translation

Exercise

Write a prompt that generates the expected completion

Input text

I was enjoying the sun, but then a huge cloud came and covered the sky.

Expected completion

Ich genoss die Sonne, aber dann kam eine riesige Wolke und bedeckte den Himmel.

# Exercise 2 – Negation

Exercise

Write a prompt that generates the expected completion

Input text

I was enjoying the sun, but then a huge cloud came and covered the sky.

Expected completion

I was not enjoying the sun, and then a huge cloud did not come and cover the sky.

# Exercise 3 – Classification

Exercise

Write a prompt that generates the expected completion

Input text

"Not much to write about here, but it does exactly what it's supposed to. filters out the pop sounds. now my recordings are much more crisp. it is one of the lowest prices pop filters on amazon so might as well buy it, they honestly work the same despite their pricing"

Expected  
completion  
(or similar)

Positive: 0.75  
Neutral: 0.20  
Negative: 0.05

# Exercise 4 – E-Mail Summarization

Exercise

Write a prompt that generates the expected completion

Input text

Your own long email thread

Expected  
completion  
(or similar)

Summary: XYZ  
Open Questions: XYZ  
Action Items: XYZ

# Exercise 5 – Rewriting

## Exercise

Write a prompt that generates the expected completion

## Input text

I was enjoying the sun, but then a huge cloud came and covered the sky.

## Expected completion

She was enjoying the sun, but then a huge cloud came and covered the sky.

# Exercise 6 – Multiple Tasks

Exercise

Write a prompt that generates the expected completion

Input text

I was enjoying the sun, but then a huge cloud came and covered the sky.

Expected completion

```
{  
  "translated": "Ich genoss die Sonne, aber dann kam eine riesige Wolke und bedeckte den Himmel.",  
  "negated": "I was not enjoying the sun, and no huge cloud came and covered the sky.",  
  "third_person": "She was enjoying the sun, but then a huge cloud came and covered the sky."  
}
```

# Exercise 7 – Data extraction to JSON

## Exercise

Write a prompt that generates the expected completion

## Input text

Hello, my name is Mateo Gomez. I lost my Credit card on August 17th, and I would like to request its cancellation. The last purchase I made was of a Chicken parmigiana dish at Contoso Restaurant, located near the Hollywood Museum, for \$40. Below is my personal information for validation:

Profession: Accountant  
Social Security number is 123-45-6789  
Date of birth: 9-9-1989  
Phone number: 949-555-0110  
Personal address: 1234 Hollywood Boulevard Los Angeles CA  
Linked email account: mateo@contosorestaurant.com  
Swift code: CHASUS33XXX

## Expected completion

```
{  
  "reason": "Lost card",  
  "classified_reason": "lost_card",  
  "name": "Mateo Gomez",  
  "ssn": "123-45-6789",  
  "dob": "09/09/1989"  
}
```

# Exercise 8 – Fashion product description

Exercise

Write a prompt that generates the expected completion

Input text

Season: Winter  
Style: Sweater  
Gender: Female  
Target group: Teenager  
Material: Cotton

Expected  
completion  
(or similar)

Stay warm and stylish this winter with our cozy cotton sweaters, perfect for the fashion-forward teenager. Refresh your wardrobe with the latest winter styles from our collection.

# Exercise 9 – Write a Blog

## Exercise

Write a blog post about a topic of your choice

## Input text

Step 1: I want you to act as a social media manager. You will be helping me to brainstorm blog post outline ideas for the topic <Topic 1>:

Step 2: Write 3 engaging and informative paragraphs about <Idea 1 description>

Step 3: Write 3 engaging and informative paragraphs about <Idea 2 description>

Step 4: Tags <List of relevant #hashtags>

# Exercise 10 – Call Center

## Exercise

Analyze a call center conversation

### Input text

Employee: "Hello, this is Julia Schreider from Contoso Company. How can I help you today?"  
Customer: "Hi, I am Carsten Mueller. I ordered a package 10 days ago, on February 10th, and it was supposed to arrive in maximum 5 business days. I have called three times already and nobody could provide any more information. I want to know where the package is and I want the problem to be solved immediately. This is the worst service I had for a long time!"  
Employee: "I apologize for the inconvenience, Mr. Mueller. I understand your frustration and I'm here to help. Can you please provide me with your order number so I can look into this for you?"  
Customer: "Yes, it's ACZ456789."  
Employee: "Thank you. I'm looking into it now. Can you please hold for a few minutes while I check the status of your package?"  
Customer: "Okay."  
Employee: "Thank you for your patience. I am sorry to inform you that I am unable to find the status of your package. It appears to have left the sending address, but no up-to-date status on the current location. I will further investigate your case and get back to you as soon as possible via phone call. Could you please provide me your contact information?"  
Customer: "Ah not again. Anyway, my phone number is +4911112223344."  
Employee: "I apologize again for the inconvenience. Is there anything else I can help you with today?"  
Customer: "No."  
Employee: "Thank you. Have a great day!"

### Expected Completion

```
{  
  "classified_reason": "lost_package",  
  "resolve_status": "unresolved",  
  "call_summary": "Customer ordered package 10 days ago and has not received it yet.",  
  "customer_name": "Carsten Mueller",  
  "employee_name": "Julia Schreider",  
  "order_number": "ACZ456789",  
  "customer_contact_nr": "+4911112223344",  
  "new_address": "N/A",  
  "sentiment_initial": ["angry", "frustrated"],  
  "sentiment_final": ["calm"],  
  "satisfaction_score_initial": 0,  
  "satisfaction_score_final": 5,  
  "eta": "N/A",  
  "action_item": ["track_package", "inquire_package_status",  
  "contact_customer"]  
}
```

# Exercise 11 – Few-shot learning

## Exercise Data samples

Write a few-shot learned prompt that classifies a movie summary.

Paul Atreides, a brilliant and gifted young man born into a great destiny beyond his understanding, must travel to the most dangerous planet in the universe to ensure the future of his family and his people. As malevolent forces explode into conflict over the planet's exclusive supply of the most precious resource in existence-a commodity capable of unlocking humanity's greatest potential-only those who can conquer their fear will survive.  
['Action', 'Adventure', 'Science Fiction']

A botched store robbery places Wonder Woman in a global battle against a powerful and mysterious ancient force that puts her powers in jeopardy.  
['Action', 'Adventure', 'Fantasy']

After the devastating events of *Avengers: Infinity War*, the universe is in ruins due to the efforts of the Mad Titan, Thanos. With the help of remaining allies, the Avengers must assemble once more in order to undo Thanos' actions and restore order to the universe once and for all, no matter what consequences may be in store.  
['Adventure', 'Science Fiction', 'Action']

A widowed new dad copes with doubts, fears, heartache and dirty diapers as he sets out to raise his daughter on his own. Inspired by a true story.  
['Drama', 'Family', 'Comedy']

New data:  
Harry, Ron and Hermione walk away from their last year at Hogwarts to find and destroy the remaining Horcruxes, putting an end to Voldemort's bid for immortality. But with Harry's beloved Dumbledore dead and Voldemort's unscrupulous Death Eaters on the loose, the world is more dangerous than ever.

# Exercise 12 – NL to SQL with Codex

## Exercise

Write a prompt that generates the expected SQL statement

## Table details

Table: customer // Columns: firstname, name, customer\_id, address  
Table: orders // Columns: order\_id, customer\_id, product\_id, product\_amount  
Table: products // Columns: product\_id, price, name, description

## Query

A query that returns the top 10 orders and show the customer name

# Break

# Top 3 OpenAI Use Cases



## Contact Center

Summarization, information extraction from call center calls and chat conversations

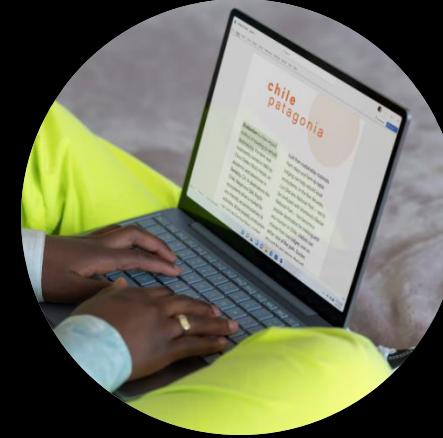
**Proof:** Telco, banking, insurance



## ChatGPT for my Enterprise

Index knowledge base and give generative answers to arbitrary questions

**Proof:** Nearly every industry



## Document Intelligence

Information extraction, classification, summarization  
Extract complex entities from documents and their relationship (e.g., in media or research)

**Proof:** Nearly every industry

# Azure OpenAI Service – Implemented Use Case

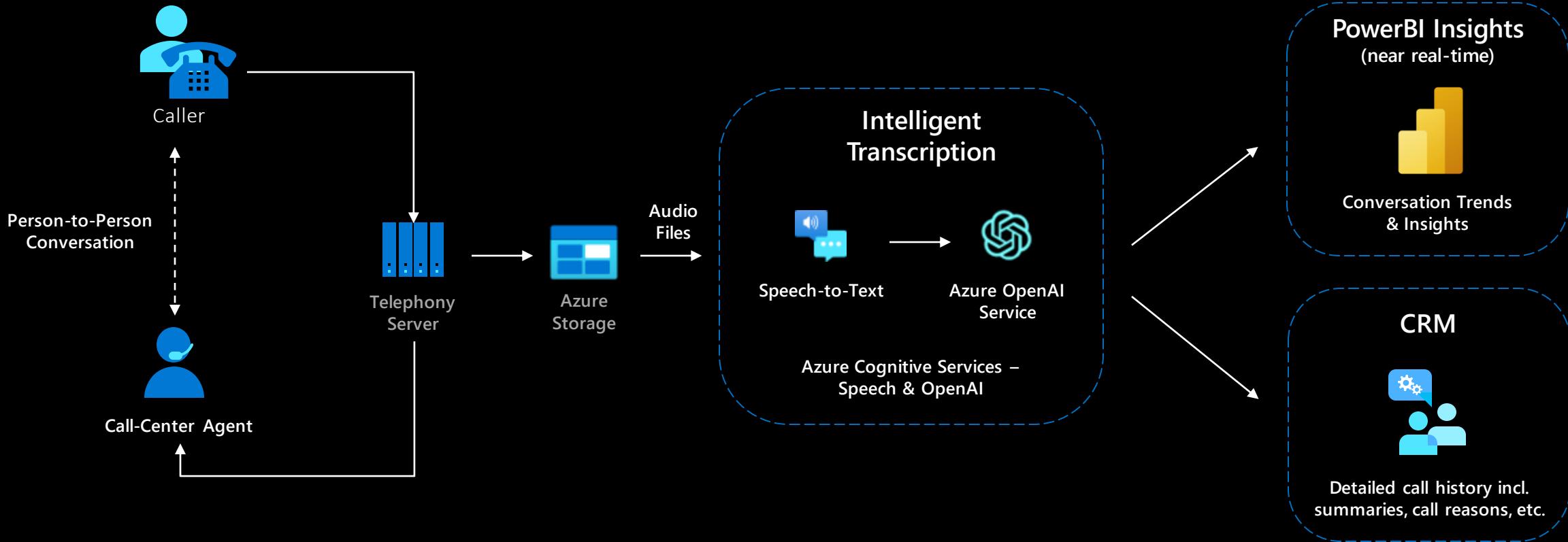
- **Contact Center & Customer Support**
  - Summarization, information extraction from call center calls and chat conversations
  - **Proof:** Several telcos
- **Question & Answering (Q&A)**
  - Index knowledge base and give generative answers to arbitrary questions
  - **Proof:** Several insurances
- **Product Description Generation**
  - Generate product taglines from metadata
  - **Proof:** Several fashion retailers
- **Code generation**
  - Code creation, summarization, explanation
  - **Proof:** Several Professional Service Companies

## Document Intelligence

- Information extraction, classification, summarization on documents, e.g., maintenance logs
- Extract complex entities from documents and their relationship (e.g., in media or research)
- **Proof:** Largely used in Energy sector
- **Text Rewriting / Paraphrasing**
  - News article rewriting for SEO optimization
  - **Proof:** Several Media Companies
- **Event Summarization**
  - Post-game summaries from transcripts & automated reviews
  - Meeting summarization & action item extraction
  - **Proof:** Several Sports Companies

# Contact Center Analytics using Speech API & OpenAI

Extract rich insights from call transcripts



# Azure OpenAI | Sample Use Cases

## Playground

Deployments Examples

text-davinci-002 Summarize Text

Code View

You must extract the following information from the phone conversation below:

1. Call reason (key: reason)  
2. Cause of the incident (key: cause)  
3. Names of all drivers as an array (key: driver\_names)  
4. Insurance number (key: insurance\_number)  
5. Accident location (key: location)  
6. Car damages as an array (key: damages)  
7. A short, yet detailed summary (key: summary)

Make sure fields 1 to 6 are answered very short, e.g. for location just say the location name. Please answer in JSON machine-readable format, using the keys from above. Format the output as JSON object called "results". Pretty print the JSON and make sure that is properly closed at the end.

Phone conversation:

Hi I just had a car accident and wanted to report it. OK, I hope you're alright, what happened? I was driving on the I-18 and I hit a pothole. That's understandable. Can you give me your full name? Sure, it's Sarah Standl. Do you know what caused the accident? I think I might have hit a pothole. OK, where did the accident take place? On the I-18 freeway. Was anyone else injured? I don't think so. But I'm not sure. OK, well we'll need to do an investigation. Can you give me the other drivers information? Sure, his name is John Radley. And your insurance number. OK. Give me a minute. OK, it's 546452. OK, what type of damages has the car? Headlights are broken and the airbags went off. Are you going to be able to drive it? I don't know. I'm going to have to have it towed. Well, we'll need to get it inspected. I'll go ahead and start the claim and we'll get everything sorted out. Thank you.

Customer conversation in English (or other language)

Output as JSON

```
results = {  
  "reason": "car accident",  
  "cause": "hit a pothole",  
  "driver_names": ["Sarah Standl", "John Radley"],  
  "insurance_number": 546452,  
  "location": "I-18 freeway",  
  "damages": ["headlights broken", "airbags went off"],  
  "summary": "Sarah Standl was driving on the I-18 when she hit another car. She thinks she hit a pothole. John Radley was the other driver. There were no injuries but both cars have damage."  
}
```

## Call Center Automation

### Example: OpenAI Prompt

# How do I build something like ChatGPT that uses my own data?

1  
You can use ChatGPT out of the box

2  
You have to fine-tune the model with your own data

3  
Inject the relevant information into ChatGPT along with the question

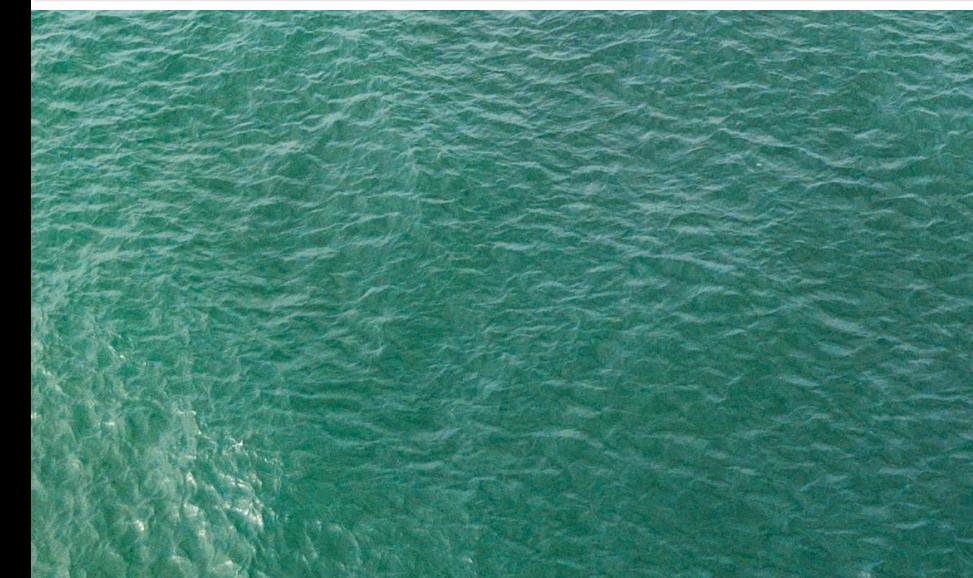
# Should I **fine-tune GPT-3** with my data?

- Yes, but only for these use cases:
  - Classification
  - Conditional Generation
    - Answering in a specific "style"
  - Both are supervised training tasks!
- Just throwing in your data (unsupervised) likely **won't** work
  - Even if you had a lot of data, it won't change the model's weights enough...



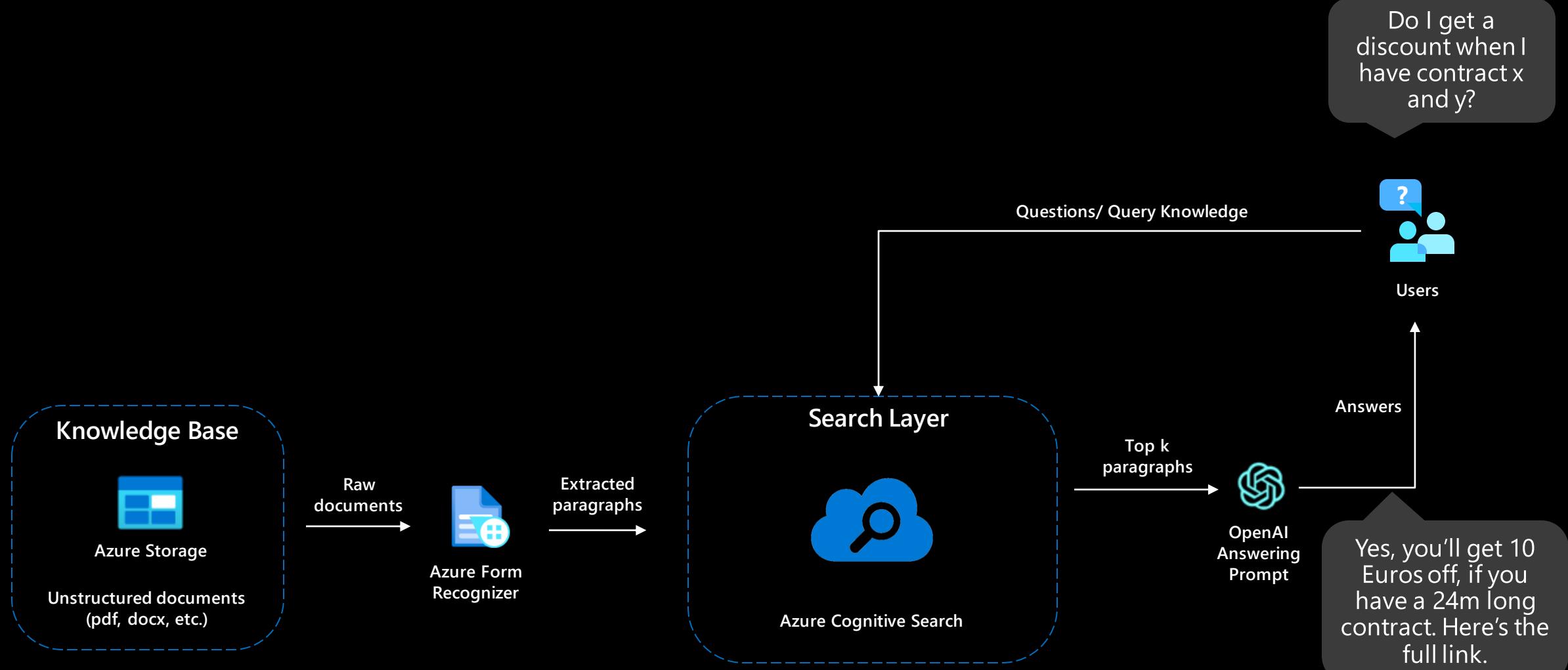
JSON

```
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}  
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}  
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}
```



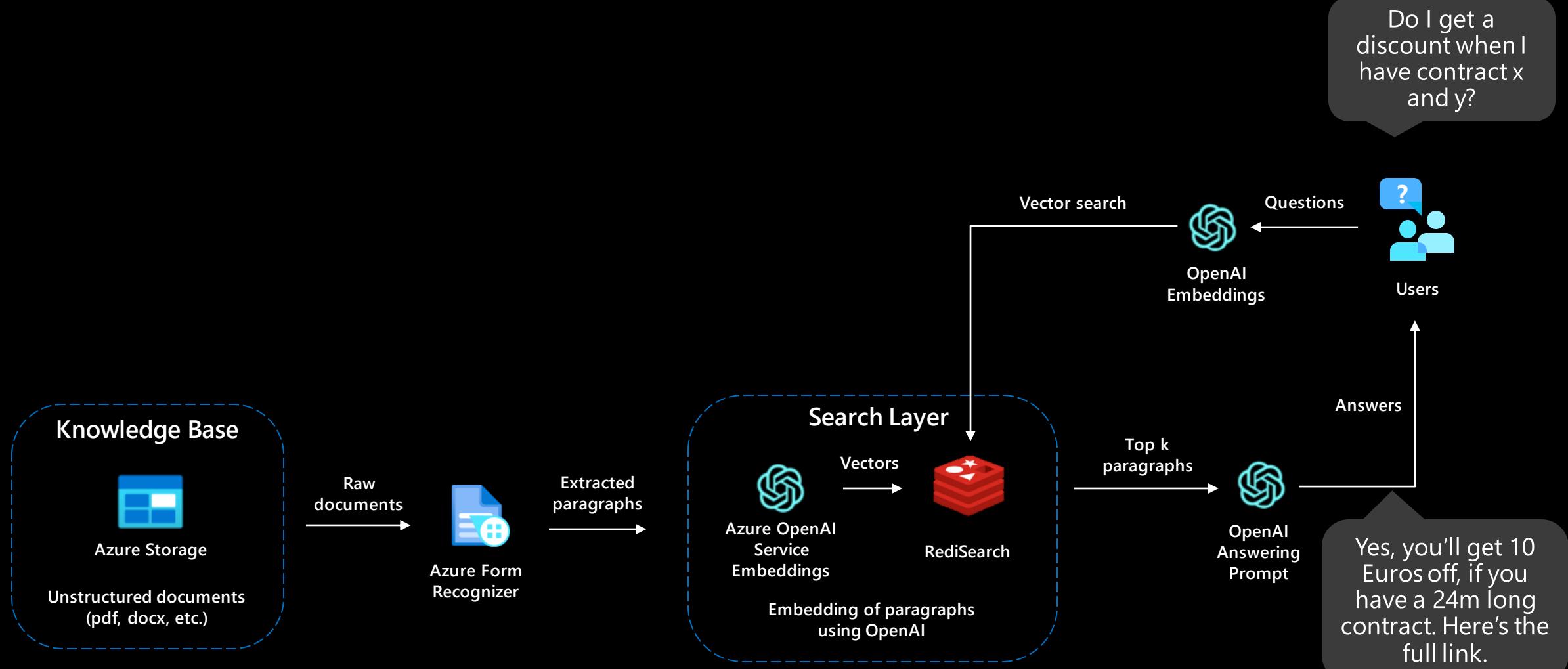
# ChatGPT with Enterprise Data

Extract precise answers from your unstructured documents



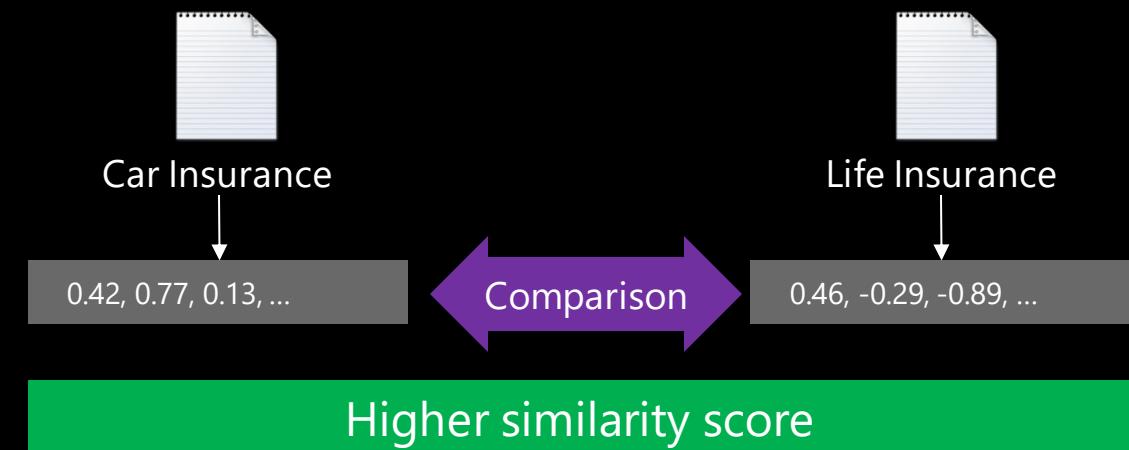
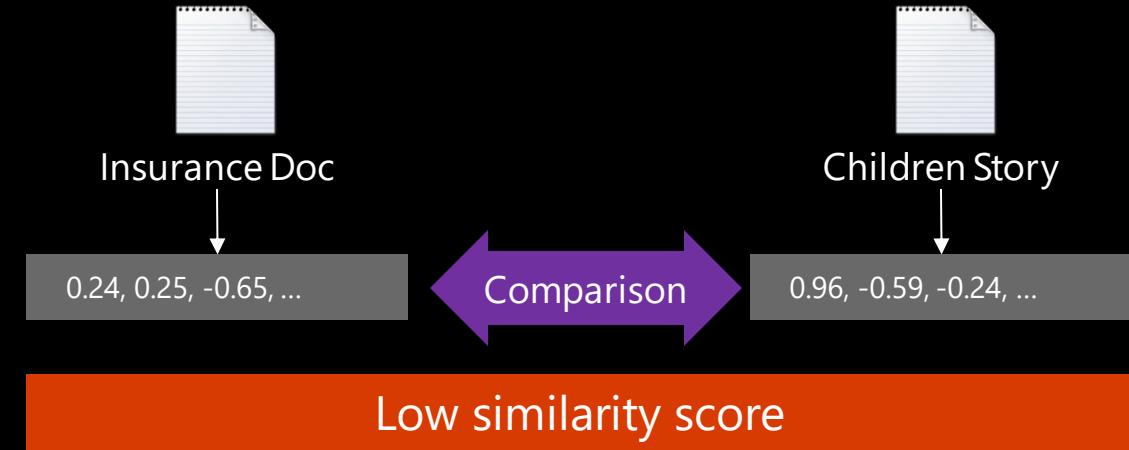
# ChatGPT with Enterprise Data

Extract precise answers from your unstructured documents



# OpenAI Embeddings

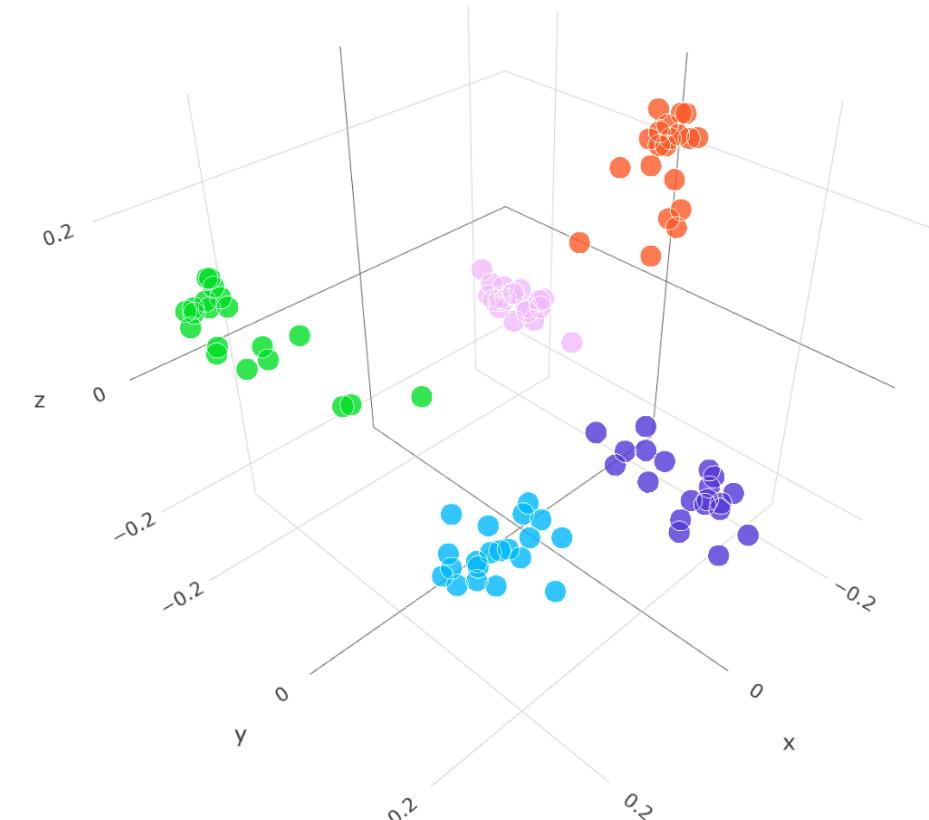
- Dense representation of **semantic meaning** of a piece of text
  - Allows to compare how similar two pieces of text are
  - Same idea as word2vec and others, but significantly better at semantic capturing
- A tool to build ChatGPT-like experiences on private data
- \$1 for 3000-5000 pages of text



# Use Cases for Embeddings

- **Search**
  - where results are ranked by relevance to a query string
- **Clustering**
  - where text strings are grouped by similarity
- **Recommendations**
  - where items with related text strings are recommended
- **Anomaly detection**
  - where outliers with little relatedness are identified
- **Diversity measurement**
  - where similarity distributions are analyzed
- **Classification**
  - where text strings are classified by their most similar label

● animal ● athlete ● film ● transportation ● village





# MVP Development

# Partnering with Microsoft to tackle the “Use Case Flood”

1

## Build up OpenAI CoE

- Central point of contact for OpenAI use cases
- Helps internal teams to get started

2

## Enablement

- Microsoft can provide workshops for CoE (incl. train the trainer)
- Scale internal knowledge

3

## Fast evaluation

- Goal: figure out which of the use cases is a low-hanging fruit & has high ROI
- “Speed dating”
- Microsoft can help

# Azure OpenAI MVP Development

## • Procedure

- Acquire data for project
  - Typically, mostly evaluation data
  - Training data only required for few use cases
  - Timeline: fully depends on customer
- Validate idea at small scale as quickly as possible
  - In Azure OpenAI Studio or Jupyter notebook
  - Timeline: 1-5 days (depends on use case)
- Evaluate idea at larger scale
  - Use annotated data for evaluation or human review
  - Timeline: 5-10 days (depends on use case)
- Build MVP around it
  - UI, API, security, user access, etc.
  - Timeline: 5-20 days (depends on use case)

## • Skills needed

- Data Scientist/Developer
  - Understanding of OpenAI
  - Can write Python
- Developer for UI
  - Typical skillset for webdev (use streamlit for MVPs)
- Azure developer
  - Azure service configuration, devops automation

## • Best practices

- Re-use existing code from samples
  - Azure OpenAI documentation, OpenAI cookbooks
- Complete idea validation & at scale testing as quickly as possible
  - Fail early if needed

# GPT Customization

- Try **zero-shot** with Davinci (prompt engineering)
  - If it works, great, try smaller models to save \$
  - Consider fine-tuning if use case allows it to save \$ (might require bigger scale)
- Try **few-shots** with Davinci (if use case allows it)
  - If it works, great, try smaller models to save \$
  - Consider fine-tuning if use case allows it to save \$
- Try **embeddings** or different model for **fine-tuning**

[Reference: How to customize a model with Azure OpenAI - Azure OpenAI | Microsoft Learn](#)

# Q&A

# Break

# Hands-On

- Quickstart
  - First, do the short quickstart example
  - [openai-in-a-day/01\\_minimal\\_example.ipynb](https://github.com/anderl80/openai-in-a-day/blob/main/01_minimal_example.ipynb) at main · [anderl80/openai-in-a-day \(github.com\)](https://github.com/anderl80/openai-in-a-day)
- Q&A
  - You'll be completing a simple Q&A example in Python
  - [openai-in-a-day/03\\_qa\\_semantic\\_answerering\\_embeddings.ipynb](https://github.com/anderl80/openai-in-a-day/blob/main/03_qa_semantic_answerering_embeddings.ipynb) at main · [anderl80/openai-in-a-day \(github.com\)](https://github.com/anderl80/openai-in-a-day)
- Embeddings
  - You'll be building an unsupervised movie clustering example and recommendation engine in Python
  - [openai-in-a-day/04\\_movie\\_classification\\_unsupervised.ipynb](https://github.com/anderl80/openai-in-a-day/blob/main/04_movie_classification_unsupervised.ipynb) at main · [anderl80/openai-in-a-day \(github.com\)](https://github.com/anderl80/openai-in-a-day)



Thank you!



Microsoft AI

# Few-shot learning example

- Helps to solve use cases where zero-shot does not give the correct results
  - Provide prompt and desired completion
  - Use a stop sequence (e.g., ###) to split up examples
  - Start with 3-5 examples
- Limitations
  - For long input/output this is not feasible (token limit!)
  - For very hard problems, 3-5 examples might not be enough
  - Cost (prompts might become very long)

→ Fine-tuning might help overcome those limitations

Extract the airport codes from this text:

###

Text: "I want to fly from Los Angeles to Miami."

Airport codes: LAX, MIA

###

Text: "I want to fly from Orlando to Boston"

Airport codes: MCO, BOS

###

Text: "I want to travel from London City to Newark"

Airport codes:

Text: "I want to travel from London City to Newark"

Airport codes: LCY, EWR

Convert movie titles into emoji.

Back to the Future: 🧑‍🤝‍🧑 🚗 🕒

Batman: 🕵️‍♂️ 🦇

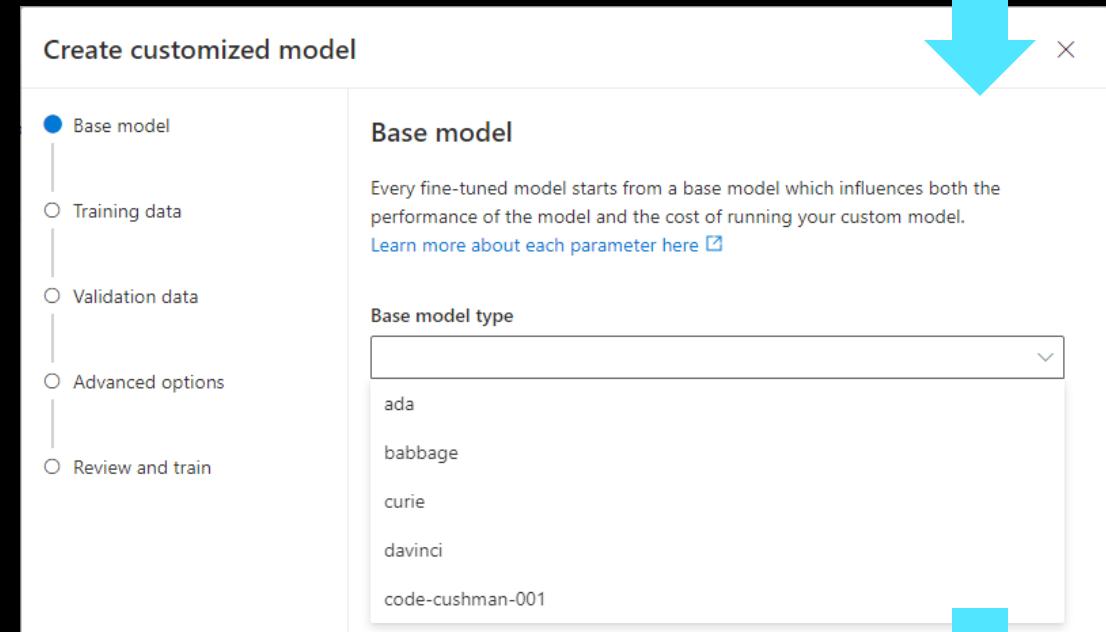
Transformers: 🚗 🤖

Star Wars: 🌟 🌟 🚀

# Fine-Tuning

- Idea of fine-tuning
  - Dataset - Set of given prompts and completions
  - Goal - Adapt GPT-3 to learn expected completions (transfer learning)
- Use cases
  - Tailored classification
  - Custom text generation
    - article writing, contracts, NER, chatbot
- Supported engines
  - By default - ada, babbage, curie
  - Upon request - davinci & codex
- Practical
  - JSONL dataset (prompts & completions)
  - Training and validation set - [details](#)

```
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}  
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}  
{"prompt": "<prompt text>", "completion": "<ideal generated text>"}
```



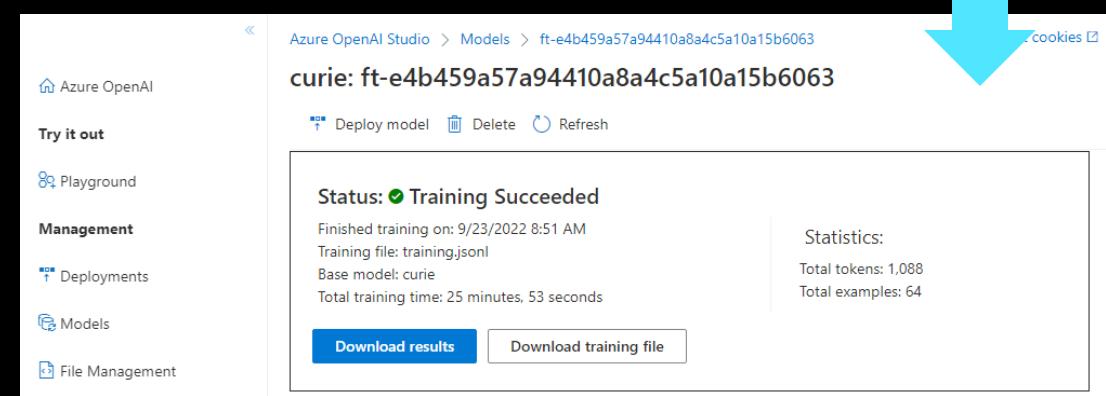
Create customized model

Base model

Every fine-tuned model starts from a base model which influences both the performance of the model and the cost of running your custom model. [Learn more about each parameter here](#)

Base model type

- ada
- babbage
- curie
- davinci
- code-cushman-001



Azure OpenAI Studio > Models > ft-e4b459a57a94410a8a4c5a10a15b6063

curie: ft-e4b459a57a94410a8a4c5a10a15b6063

Status: Training Succeeded

Finished training on: 9/23/2022 8:51 AM

Training file: training.jsonl

Base model: curie

Total training time: 25 minutes, 53 seconds

Statistics:

Total tokens: 1,088

Total examples: 64

Download results   Download training file

# I want to fine-tune GPT-3 with my company data!

- Sure, you can do this, but most likely this will
  - Be very expensive
  - Require a lot of data
  - Most likely does not give you the result you are looking for
- Rational
  - The original GPT-3 was trained on 300bn tokens ([paper](#))
  - Let's say you add 100k documents, each with 10k tokens each
  - This is still 300000 smaller than the original training dataset
  - How will you know that the model now uses that little information to answer your question?
- Better approach
  - Use fine-tuning for use cases where it makes sense (classification, conditional generation)
  - For retrieving/working with internal company data, leverage embeddings



# MVP Development

# Partnering with Microsoft to tackle the “Use Case Flood”

1

## Build up OpenAI CoE

- Central point of contact for OpenAI use cases
- Helps internal teams to get started

2

## Enablement

- Microsoft can provide workshops for CoE (incl. train the trainer)
- Scale internal knowledge

3

## Fast evaluation

- Goal: figure out which of the use cases is a low-hanging fruit & has high ROI
- “Speed dating”
- Microsoft can help

# Azure OpenAI MVP Development

## • Procedure

- Acquire data for project
  - Typically, mostly evaluation data
  - Training data only required for few use cases
  - Timeline: fully depends on customer
- Validate idea at small scale as quickly as possible
  - In Azure OpenAI Studio or Jupyter notebook
  - Timeline: 1-5 days (depends on use case)
- Evaluate idea at larger scale
  - Use annotated data for evaluation or human review
  - Timeline: 5-10 days (depends on use case)
- Build MVP around it
  - UI, API, security, user access, etc.
  - Timeline: 5-20 days (depends on use case)

## • Skills needed

- Data Scientist/Developer
  - Understanding of OpenAI
  - Can write Python
- Developer for UI
  - Typical skillset for webdev (use streamlit for MVPs)
- Azure developer
  - Azure service configuration, devops automation

## • Best practices

- Re-use existing code from samples
  - Azure OpenAI documentation, OpenAI cookbooks
- Complete idea validation & at scale testing as quickly as possible
  - Fail early if needed

# GPT Customization

- Try **zero-shot** with Davinci (prompt engineering)
  - If it works, great, try smaller models to save \$
  - Consider fine-tuning if use case allows it to save \$ (might require bigger scale)
- Try **few-shots** with Davinci (if use case allows it)
  - If it works, great, try smaller models to save \$
  - Consider fine-tuning if use case allows it to save \$
- Try **embeddings** or different model for **fine-tuning**

[Reference: How to customize a model with Azure OpenAI - Azure OpenAI | Microsoft Learn](#)

# Hands-On

- You'll build a simple web app that allows to summarize and answer emails
- Technology
  - Python
  - Streamlit library (web frontend)
  - OpenAI (summarization, answering)

## Email summarization demo app

Email

Hey all,

hope you are all doing well and are excited for tomorrow's [OpenAI](#) workshop! We have a good mix of theory, prompt engineering and some hacking planned, so I think it'll be a good mix of everything!

Apologies for the last-minute email, I am still not 100% recovered from being sick, so I'll be running the workshop remotely tomorrow.

Here are some short prep items:

Preparation for everybody:

You will need to do this today (takes 5min) - Relevant for: [@Isabel Grund](#) [@Hendrik Juelich](#) [@Jan](#)

[Summarize email](#)

[Generate answer](#)

Summary:

This email is to remind everyone to prepare for the upcoming OpenAI workshop. Everyone needs to register their Microsoft alias in the Azure tenant and techies need to make sure they have a working Python environment. The workshop will include a mix of theory, engineering and hacking.

# Codex Deep Dive

# Codex Series

- Capabilities
  - Natural language to code
  - Code to natural language
  - Natural language to query (SQL, etc.)
  - Query to natural language
  - Rewriting of code (refactoring, replacing things, changing things)
- Mainly two models
  - `code-davinci-002` and `code-cushman-001`
  - Cushman is fine-tunable on Azure (requires allowlisting)
- Language support
  - Python, C#, SQL, Java, JavaScript, TypeScript, Go, Perl, PHP, Ruby, Swift, Shell (bash)

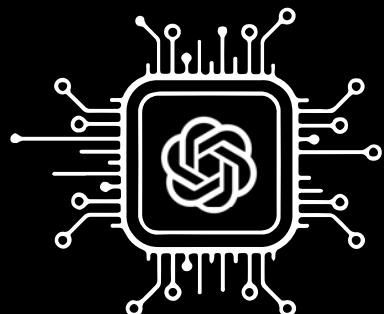


# GitHub Copilot

Your AI pair programmer

```
19     screenshot = ImageGrab.grab()  
20     ...# Convert to text  
21     text = image_to_string(screenshot)  
22     ...# Parse text for email addresses  
23     emails = re.findall(r'[\w\.-]+@[\\w\.-]+', text)  
24     ...return emails  
25  
26     def validate(addresses):  
27     ...|  
28
```

## OpenAI Codex Model



## GitHub



GitHub  
Copilot Service

Public code and text  
on the internet

Provide editor context

Provide suggestions

Improve suggestions

## Private code

```
JS fetch_pic.js push_to
1 const fetchNASAPictureOfTheDay = () => {
2   return fetch('https://api.nasa.gov/mars/rovers/curiosity/images/latest?api_key=DEMO_KEY')
3     .then(response => response.json())
4     .then(json => {
5       return json;
6     });
7 }
8
9
10
11
12 }
```



# Copilot or Codex? Which one is right one for me?



- **Copilot**

- **Main use:** during everyday coding
- Makes users more efficient by helping them to write code in their editor
- Ready made, speaks typical modern coding languages



- **Codex**

- **Target:** Give apps natural language capabilities
- Allows to enrich apps with new capabilities, e.g., allows users to perform tasks using natural language
- Might be useful if you want to build your own Copilot for any non-supported language

# Power BI

## Generating DAX expressions with natural language

The screenshot shows the Power BI Desktop application interface. The main area displays a dashboard titled "T-MOBILE DASHBOARD" with the following visualizations:

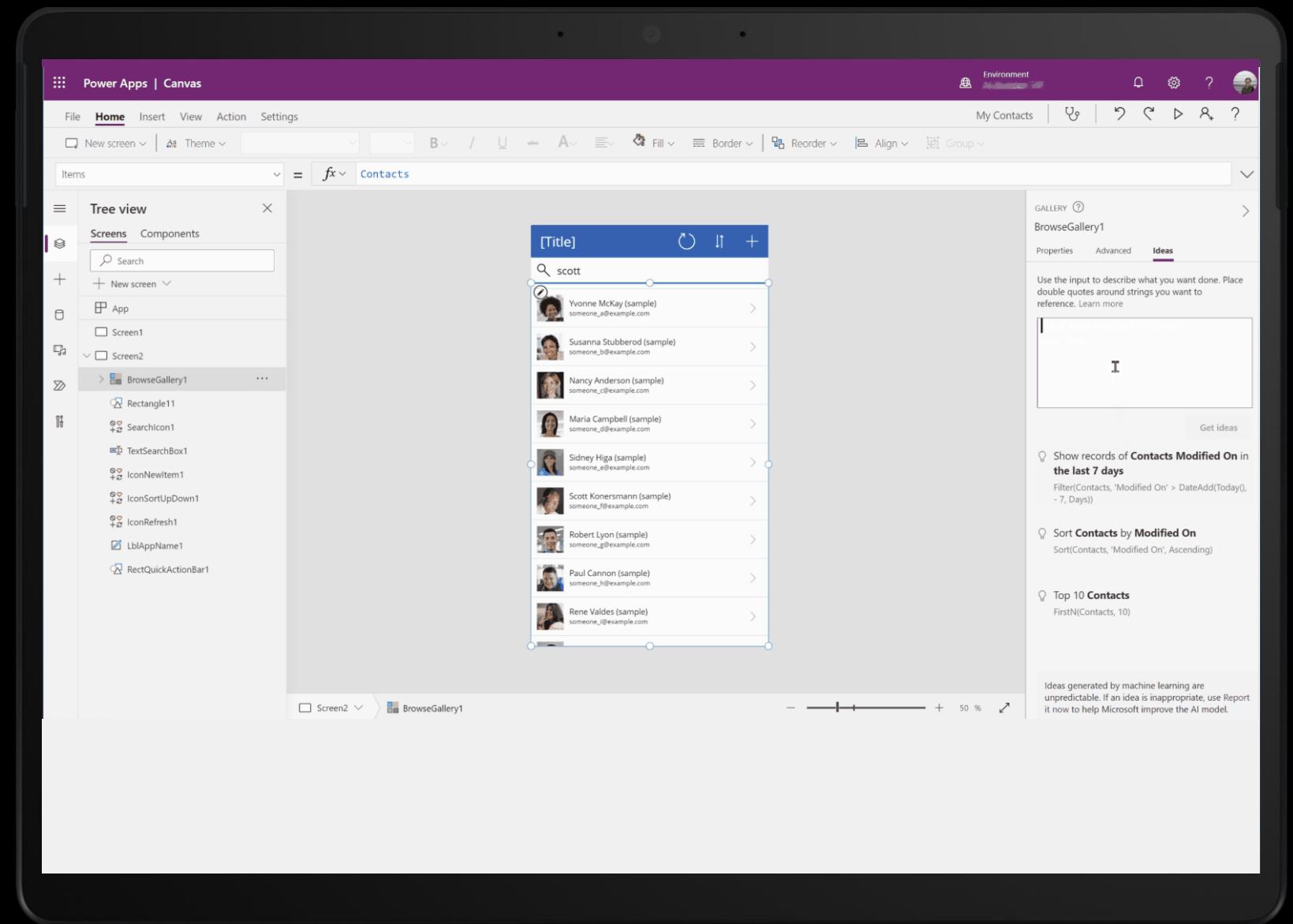
- Inspections by Response:** A donut chart showing 22K Inspections. The legend indicates "Response" with "OK" (light gray) and "Issue" (pink). The data is split into 7K (31.47%) Issues and 15K (68.53%) OKs.
- Inspections by Checklist Item and Response:** A horizontal bar chart showing the count of inspections for different checklist items. The items and their counts are: Clutter free & accessible (6.8K), Clean and hygienic (1.8K), Filled with available Products (2.7K), and Devices are confirmed to be operational (0.8K).
- Inspections by Date and Response:** A line chart showing the number of inspections over time from May 2020 to August 2020, categorized by Response (Issue in pink, OK in gray).
- Percentage Issues, Inspections and Inspection Score by Area Inspected:** A bubble chart showing the relationship between Inspections (Y-axis, 0K to 10K), Percentage Issues (X-axis, 0% to 100%), and inspection areas. The areas and their approximate data points are: Store Entrance (10K, 10%), Waiting Area (5K, 30%), Phones (2K, 35%), Front Tables (1.5K, 40%), Information Kiosk (1K, 45%), Tablets (1.5K, 55%), Back Tables (1K, 60%), and Register (1K, 80%).

The ribbon bar at the top includes File, Home, Insert, Modeling, View, Help, Format, Data / Drill, Get data, Refresh, New visual, More visuals, New measure, Sensitivity (preview), Publish, and a Search bar.

The right side of the interface shows the "Easy measures" pane, which allows users to create measures using natural language or built-in calculations. A "Filters" pane is also visible on the far right.

# Power Platform

## Power Fx formulas





# T-Shirt Sizing for Azure OpenAI projects

# Estimating cost of an OpenAI project

- Need to figure out
  - Initial ingest cost (if applicable, e.g., for Q&A document ingest)
  - Infrastructure running cost (Web App, storage, etc.)
  - Transactional cost per operation
- The **transactional cost** is typically the main cost driver in OpenAI projects
  - However, it is important to put it in relation to the potential saving
  - A typical use case that assists humans in a certain task should have an ROI of 2x-10x



# Notice

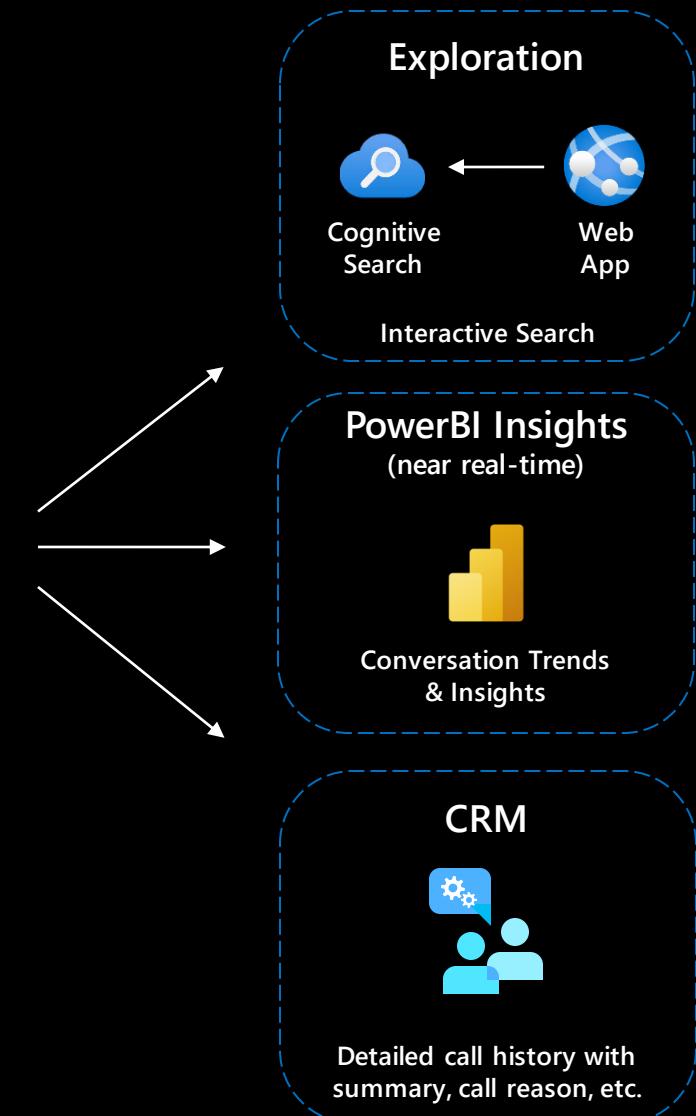
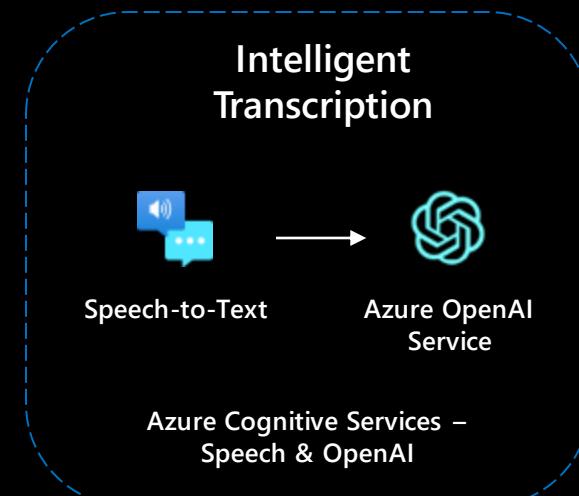
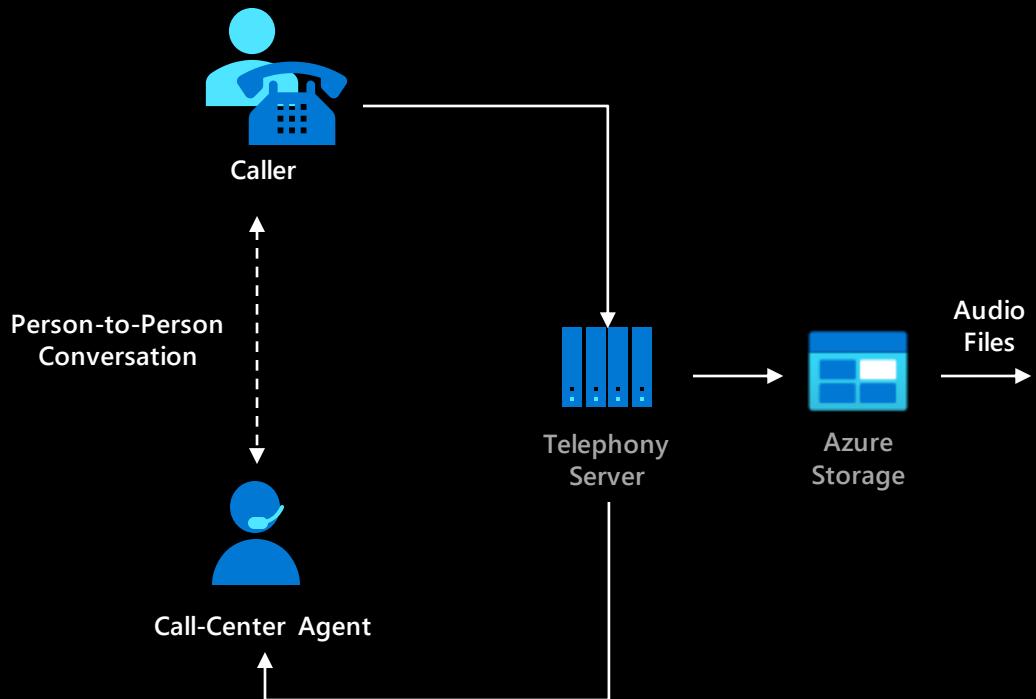
- These slides should give a **rough idea** how much AI ACR we can expect from Azure OpenAI customer projects
- When a customer approaches us with **50+ use cases**:
  - Do some generic sizing:
  - 40x Small
  - 5x Medium
  - 4x Large
  - 1x XL
- This deck uses Azure Pricing as of March 2<sup>nd</sup>
  - Once gpt-3.5-turbo is available on Azure, the T-Shirt sizes might change

# Generic OpenAI project sizing

	<b>S</b>	<b>M</b>	<b>L</b>	<b>XL</b>
Cost/transaction	\$0.02	\$0.02	\$0.02	\$0.02
Transactions/month	12.5k	50k	500k	2.5m
Transactions/sec	0.02	0.06	0.6	3.0
Reoccurring	Likely	Likely	Mostly yes	Yes
Examples	<b>Mostly use cases w/high human interaction</b>  Paraphrasing, creative writing, smaller LOB use cases, etc.	<b>Use Cases w/low human interaction (large scale)</b>  Document classification, information extraction from docs (<10k docs)	<b>OpenAI used fully automated, no human in loop</b>  Contact center analytics (small customer), Q&A internal (1k users), information extraction from docs (10k+ docs)	<b>OpenAI used fully automated, no human in loop</b>  Contact center analytics (big customer), Q&A internal/external (10k+ users)
<b>Monthly AI ACR</b>	<b>\$250</b>	<b>\$1000</b>	<b>\$10,000</b>	<b>\$50,000</b>

# Contact Center Analytics using Speech API & OpenAI

Extract rich insights from call transcripts



# Contact Center Analytics with OpenAI

Contact Center Size	S	M	L	XL
Avg call length	7 Minutes			
OpenAI cost	\$0.04 per call			
Speech cost	\$0.16 per call	\$0.12 per call	\$0.09 per call	\$0.08 per call
Calls/month	50000	100000	500000	1000000
<b>AI ACR/Month</b>	<b>\$10k</b>	<b>\$15k</b>	<b>\$65k</b>	<b>\$120k</b>

- **Notes**

- Speech commitment pricing is included
- Currently no commitment pricing for Azure OpenAI Service (EA discounts apply on top)
- Call length can vary, but 7min is a good average

# Real-world example - OpenAI in Call Center

- Azure Costs

- Transcription
  - Using Azure Speech-to-Text
  - \$0.10 per call transcript
- OpenAI
  - \$0.03 per call summary from transcript
- Total
  - \$0.13 per call

- Call volume

- 50k calls/day \* \$0.13 = \$6500/day
- 1 minute per call saved
- 50k minutes = 833 hours saved/day
- Cost per agent hour: \$50-\$100

- Customer savings per year

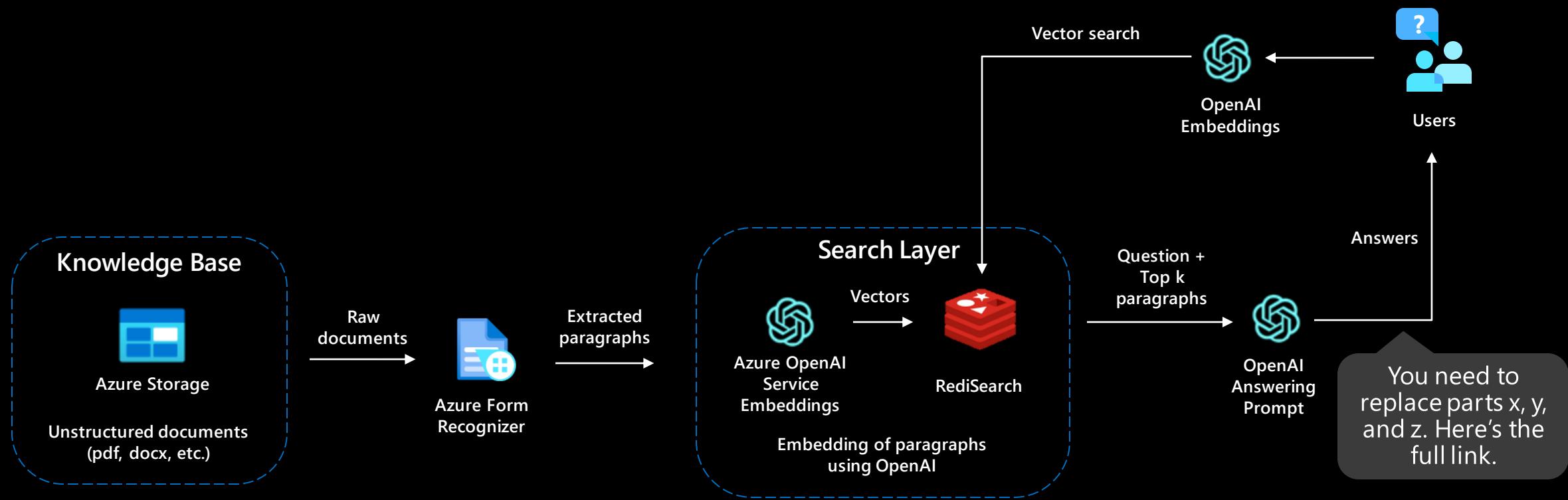
- Costs for Azure
  - ~\$2m (25 days \* 12 months)
- Savings from agents
  - \$12m-\$25m
- ROI
  - ~\$10m-\$23m saved per year

- Similar ROI for other use cases
  - Scale might differ!

# Q&A with Semantic Answering with Azure OpenAI Service

Extract precise answers from your unstructured documents  
similar to ChatGPT

The xyz is broken  
on machine 5243,  
which part do I  
need to replace?



# Q&A with OpenAI

Q&A Sizing	S	M	L	XL
OpenAI cost	\$0.04 per query			
Number of users	100	500	2000	100000
Daily queries/user	10	25	50	2
Use case	Internal KB for single LOB, low daily usage	Internal KB for single LOB, but more frequent usage	Internal KB for single LOB with larger scale and frequent usage. I.e agent assist in contact center, or advisor assist in FSI	External KB for single app, e.g., public chatbot for customer, or agent assist at contact center
<b>AI ACR/Month</b>	<b>\$1k</b>	<b>\$15k</b>	<b>\$120k</b>	<b>\$250k</b>

- **Notes**

- Running cost for vector database/frontend app not included
- One shot ingestion costs:
  - Initial indexing cost is fairly cheap with text-embedding-ada-002 (around \$1 per 5000 pages)
  - Form Recognizer AI ACR not included here (~10\$ per 1000 pages), i.e. 100k docs would be ~\$2k around
- Most of the revenue comes from the user running queries

# Example – Q&A

- **Azure Costs**

- Initial document ingest
  - \$0.01 per page (Form Recognizer Read)
  - OpenAI Embeddings negligible (\$0.0002/page)
- Prompt answering
  - ~\$0.02 per answer
- Redis
  - Fixed cost per month, e.g., \$1k

- **Example**

- 1000 users, 50 queries per day
- = \$1k/day for question/answer
- Saves person ~1h of time per day

- **Customer savings per year**

- Costs for Azure
  - ~\$312k (25 days \* 12 months)
- Benefits
  - Surely, we can quantify this again using the same formula (see call center example), but there are other benefits:
    - Less annoying work (no more searching through pdfs)
    - Quicker onboarding of new agents/users, especially in scenarios with high turnaround
    - More quality answers (depends obviously on the use case)



# Data Privacy

# Best practices

## Security is key

1. [Azure OpenAI Service encryption of data at rest - Azure Cognitive Services | Microsoft Learn](#)
2. [Data, privacy, and security for Azure OpenAI Service - Azure Cognitive Services | Microsoft Learn](#)
3. [Safety best practices - OpenAI API](#)

## Performance (fair limits to democratize service)

Current limits: davinci 10 rps, rest 20 rps (subject to change)

## Throttling in place

Use exponential backoff retry (manually) or libraries like tenacity

```
import openai
from tenacity import retry, stop_after_attempt, wait_random_exponential

@retry(wait=wait_random_exponential(min=1, max=60), stop=stop_after_attempt(6))
def completion_with_backoff(**kwargs):
    return openai.Completion.create(**kwargs)
completion_with_backoff(model="text-davinci-003", prompt="Once upon a time...")
```

# Azure OpenAI Service Use Cases



## Codex

- Natural Language to Code
- Natural Language to SQL
- Code to Natural Language
- Code documentation
- Refactoring

## DALL·E 2

- Creative ideation
- Podcast and music playlist images
- Content syndication
- Marketing campaign personalization
- Hyper-personalization

## Language

Reason over structured and unstructured data:  
**Classification, Sentiment, Entity Extraction, Search**

- Refined text-based search with embeddings
- Product feedback sentiment
- Customer and employee feedback classification
- Claims and risk analyses
- Support emails and call transcripts
- Social media trends

## Writing Assistance

- Marketing copy / email taglines
- Long format text
- Paragraphs from bullets

## Use Cases that use multiple model capabilities

- Contact Centers
  - Classification – route mails to appropriate team
  - Sentiment – prioritize angry customers
  - Entity extraction and search – analyze liability and risk
  - Mail and call transcript summarization
  - Customer response email generation
- Rapid response marketing campaigns: classification, sentiment, summarization, content generation, image generation

## Summarization

- Call center call transcripts
- Subject Matter Expert Documents
  - Competitive analysis
  - Peer Analysis
  - Technical reports
- Product and service feedback
- Social media trends

## Conversational AI

- Smart assists for call centers
- Tech support chat bots
- Virtual assistants
- Artificial humans

# Azure OpenAI Service – Implemented Use Case

- **Contact Center & Customer Support**
  - Summarization, information extraction from call center calls and chat conversations
  - **Proof:** Several telcos
- **Question & Answering (Q&A)**
  - Index knowledge base and give generative answers to arbitrary questions
  - **Proof:** Several insurances
- **Product Description Generation**
  - Generate product taglines from metadata
  - **Proof:** Several fashion retailers
- **Code generation**
  - Code creation, summarization, explanation
  - **Proof:** Several Professional Service Companies

## Document Intelligence

- Information extraction, classification, summarization on documents, e.g., maintenance logs
- Extract complex entities from documents and their relationship (e.g., in media or research)
- **Proof:** Largely used in Energy sector
- **Text Rewriting / Paraphrasing**
  - News article rewriting for SEO optimization
  - **Proof:** Several Media Companies
- **Event Summarization**
  - Post-game summaries from transcripts & automated reviews
  - Meeting summarization & action item extraction
  - **Proof:** Several Sports Companies

To learn more, please visit Azure OpenAI Service's [product page](#).

## Next steps

Have a use case? Speak to your Specialist / AE and sign up for access here: [aka.ms/oai/access](#)

# Customer References



Templates



My media



Visuals



Text



Styles

## Let's start with your content

Generate an image using a description ⓘ

Ombre cake decorated with flowers and fall foliage



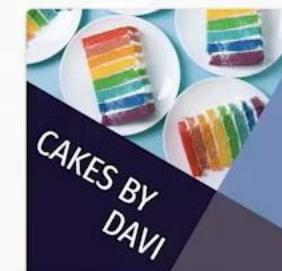
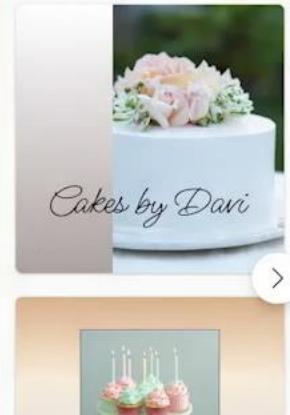
Tip

Explore different camera angles

"An aerial still of a seascape, Brazil"

Color ↕

Size ↕



Continue editing

Share

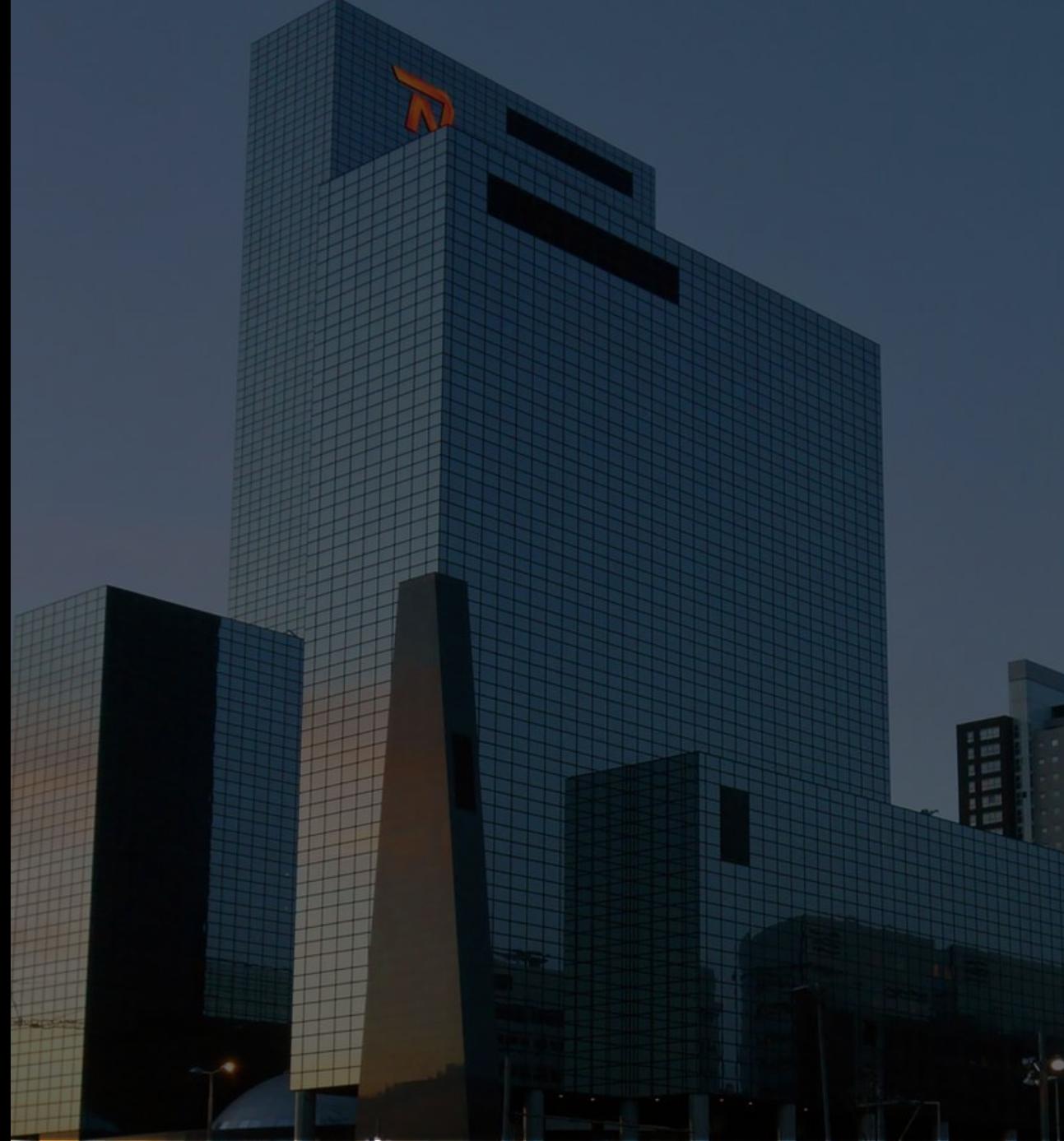


Design Ideas



*"Nationale-Nederlanden part of NN Group, the largest player in the Dutch insurance market is using the Azure OpenAI services to extract information from large volumes of unstructured data to automate their claim handling processes. The Azure OpenAI services provide the building blocks to quickly develop AI solutions that understand Dutch text. This opens the door to many other user cases."*

**Tjerrie Smit**  
**Head of Data & AI**





*"HSBC Bank is one of the largest banking and financial services institutions in the world, serving millions of customers globally. Microsoft's Azure OpenAI have been a key strategic partner to help us adopt and operationalize novel methods in the Natural Language space. As an organization we see huge potential in Large Models for summarization and classification of financial information; and also programming applications for our community of citizen scientist. Pro tem we continue to seek for opportunities to bring such benefits to 120K support function colleagues worldwide."*

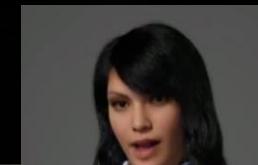
**Tjerrie Smit**  
**Head of Data & AI**

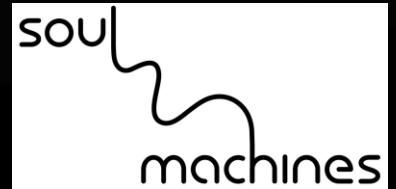




**“Sogeti constantly looks for new ways to enable clients in their digital transformation journey by providing cutting-edge solutions in AI and ML. Sogeti uses Azure OpenAI service for its digital human (Daisy) offering that enables end users to have a human like conversation. Azure OpenAI service backed up by a strong language model, has helped Daisy to learn custom data specific to a domain or client, accelerated response times, gained knowledge from the internet and much more. Azure OpenAI service has helped Sogeti to take its digital human offering to various Azure enabled clients easily and helped reduced implementation time.”**

*– Mark Oost - Sogeti Chief Technology Officer, Artificial Intelligence and Machine Learning*





**Soul Machines creates digital people that transform brand experience by creating engaging and emotional connections with customers.**

**Soul Machines is using OpenAI for sentiment to increase the empathy of their Digital People so that conversations with users are deeper and more meaningful.**

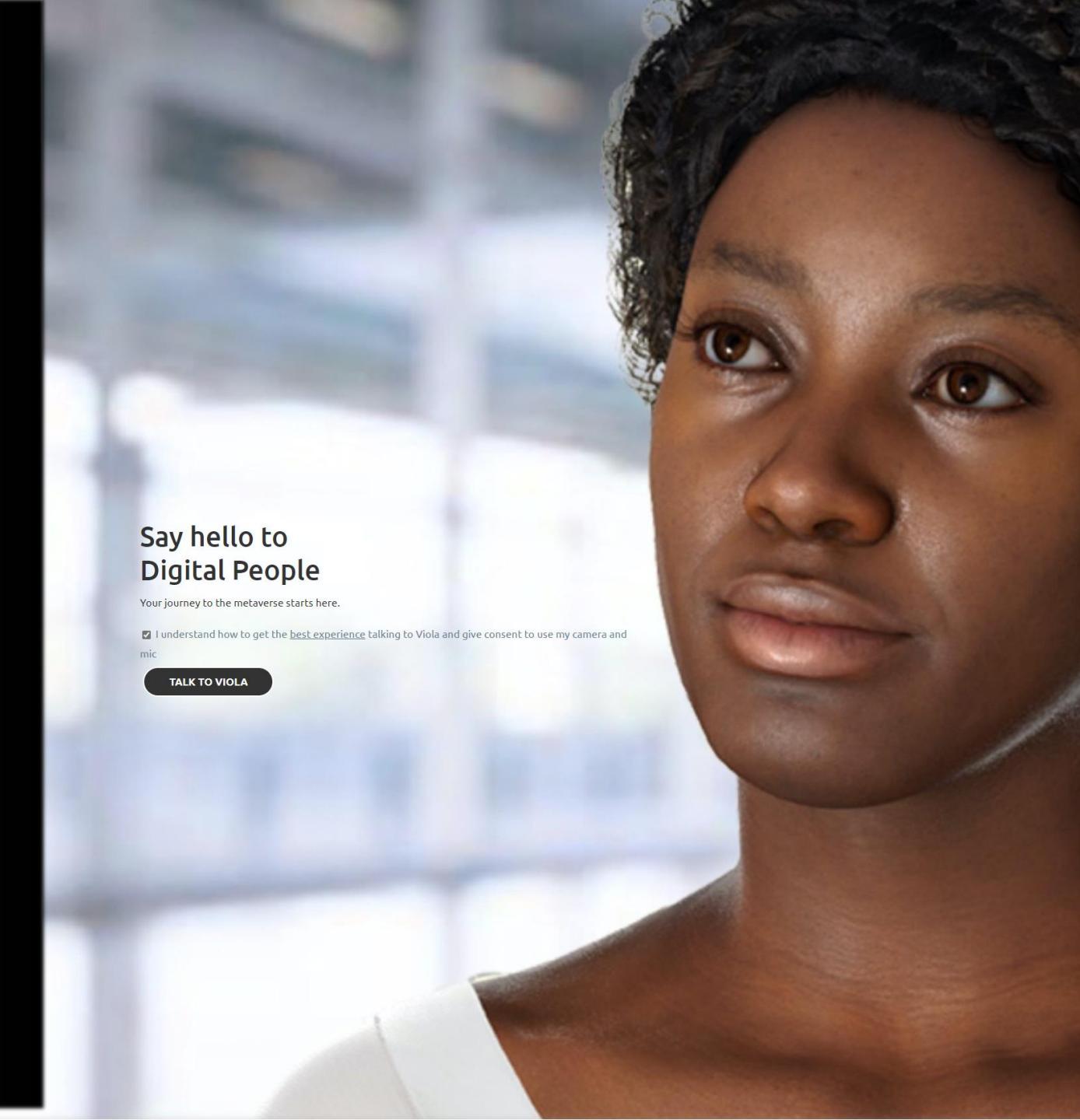
*“Soul Machines is using Azure OpenAI Service for autonomous content generation for their media and entertainment products. For example, the content generation enables them to produce copy for conversations with their digital avatars and celebrities.”*

### Say hello to Digital People

Your journey to the metaverse starts here.

I understand how to get the [best experience](#) talking to Viola and give consent to use my camera and mic

[TALK TO VIOLA](#)





"Wipro Ltd, a leading global information technology company is using Azure Cognitive Services to build Conversational AI solutions.

The solution, offers personalized IT support and enhanced employee engagement initiatives and is being used by 65,000 Wipro employees globally. The platform uses a combination of Azure natural language-understanding APIs along with customized Machine Language models.

These are powered by Azure Machine Learning Studio and Azure Open AI to foster personalized employee-experience within the organization."

– Anup Purohit, Global CIO, Wipro Ltd.





Farmlands is the largest agriculture products buying group in New Zealand. They have 450,000+ interactions with customers every year, documented in Dynamics 365.

They are using Azure OpenAI Service to gain an understanding of sentiment around key words or phrases that emerge in cases logged in Dynamics.

They are using Davinci and Curie to highlight currently trending key phrases; prepare summaries and enable users to drill down and look at case or interaction summaries.

*"It saves our call center staff time and improves our customers' experience"*

*– Gareth Pullar, insights and analytics manager at Farmlands.*

*"In terms of generating content, it's a game-changer. "It's just really exciting technology."*

*– Jean van Schalkwyk, Business Intelligence Lead*





CarMax is using Azure OpenAI Service for vehicle and store review summarization, creating marketing content, and preparing vehicle trim comparisons. Using OpenAI reduces their marketing freelancer spend and enables them to provide new information to their customers faster than their competitors

Related Articles

**Review** **Review** **Review**



2018 Porsche Cayenne: Reviews, Photos, and More

2015 Porsche Cayenne: Reviews, Photos, and More

2016 Porsche Cayenne: Reviews, Photos, and More

WED, DEC 01, 2021

Are you shopping for a luxury SUV and unsure what features to expect or which engine to choose? Let's make it easy (and fun) with a comprehensive look at the 2018 Porsche Cayenne, a luxury SUV with the heart of a sports car. Find out more in our 2018 Porsche Cayenne review.

→ [Read More](#)

MON, JAN 21, 2022

The Porsche Cayenne is an impressive luxury SUV, and the 2015 model adds technology, style, and features to a wide variety of engine options. The Cayenne is also available as a plug-in hybrid model. Take a look at our 2015 Porsche Cayenne review for key specs and features.

→ [Read More](#)

WED, DEC 01, 2021

Looking for performance and practicality? With the Porsche Cayenne you can have both. The Cayenne has powerful engine options and a luxury interior perfect for individuals or families. Learn more about perks and features in this 2016 Porsche Cayenne review.

→ [Read More](#)

*“With Azure OpenAI Service, we are able to create content that empowers our customers so they can be informed before they make a decision,”*

– Shamim Mohammad, CarMax executive vice president and chief information and technology officer. “

*“OpenAI Service is definitely playing a big part in our SEO improvement.”*

– Sean Goetz, Director, Application Systems





Produced **11 years'**  
worth of car summaries in a  
matter of **Months**



Customer:  
Trelent

Industry:  
Professional Services

Size:  
1-49 employees

Country:  
Canada

Products and services:  
Azure OpenAI Service

[Read full story here](#)



“With a product like OpenAI Service behind you, you can focus a lot more on what really matters, which is delivering a great experience, a great product, and a lot of value to your customers.”

—Calum Bird, CEO, Trelent

#### Situation:

Trelent, a pre-seed code documentation startup that uses the OpenAI Codex algorithm, found its audience in high-growth tech companies with large or distributed engineering teams. The problem: How does a two-person team provide enterprise-ready service?

#### Solution:

A Microsoft for Startups webinar provided the answer: Microsoft Azure OpenAI Service paired the powerful OpenAI algorithms Trelent was already using with Azure security, safety controls, and global availability—and worked with the startup’s existing solution.

#### Impact:

With Azure OpenAI Service, Trelent benefits from content filtering, increased security, and faster response times (from 1-3s down to an average of 750ms). Those enterprise-ready features free Trelent engineers up to focus on their core innovation.



Thank you