



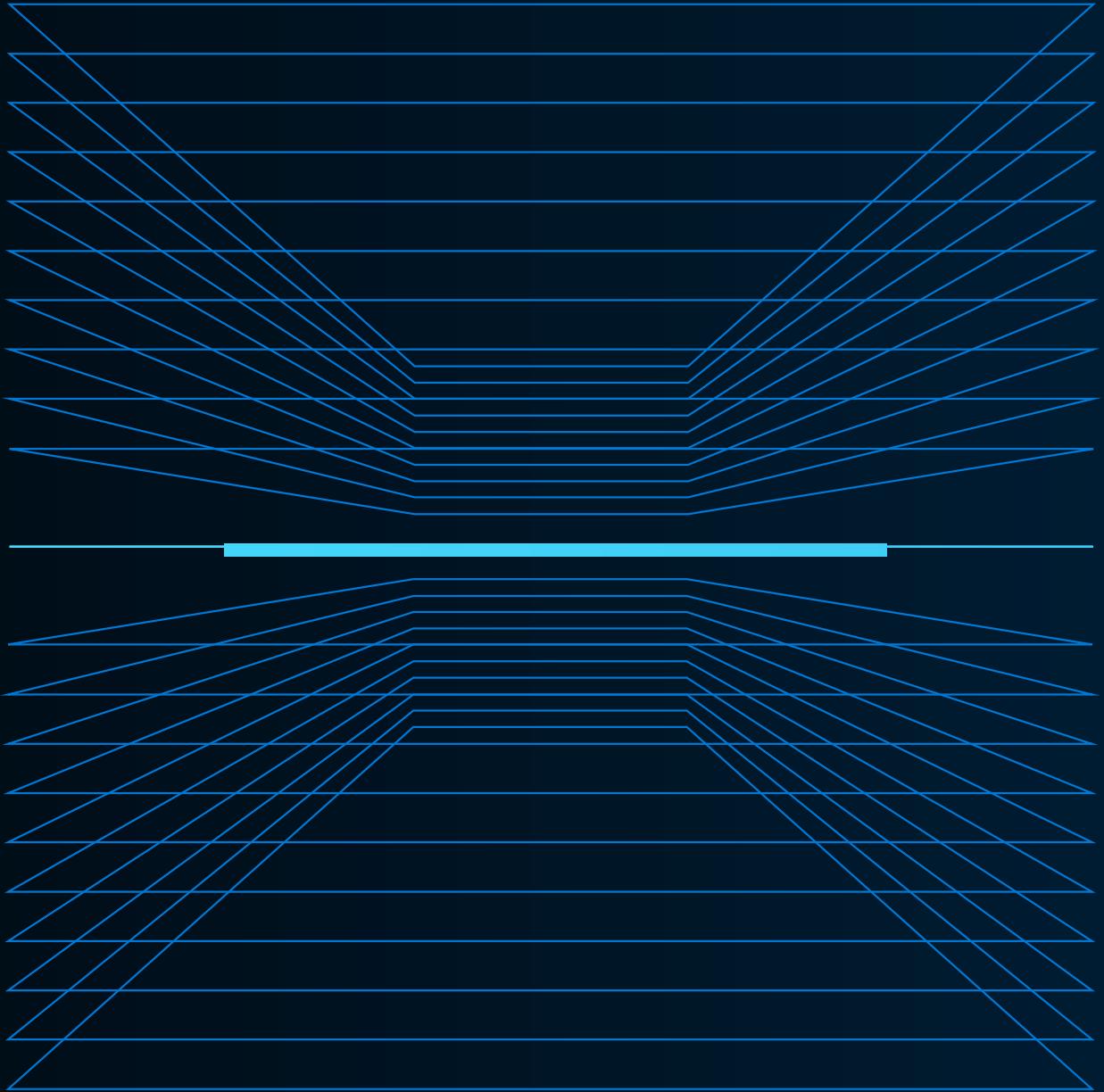
Azure OpenAI Service



Frédéric Gisbert – Cloud Solution Architect



Franck Gaillard – Cloud Solution Architect



Agenda

- OpenAI
- Azure OpenAI
- Azure Open AI – Use Cases
- Integration with Microsoft Offer
- Azure OpenAI Pricing
- Demo
- Resources





- OpenAI is an American company **created in December 2015** by prominent figures in technology and entrepreneurship (Sam Altman, Elon Musk...).
- It does research in AI with the intention of **promoting and developing a friendly AI**.
- It consists of a non-profit organization (OpenAI Inc.) and a **capped-profit subsidiary** (OpenAI LP - 2019).
- OpenAI aims to make **AI accessible to everyone** and to benefit humanity as a whole.
- It has developed several innovative projects such as **GPT-3, DALL-E or Codex**.

How the tech behind ChatGPT could change the world—an

[How the tech behind ChatGPT could change the world—an updated episode from our archive | The Economist](#)

OpenAI's new DALL-E model draws anything—but bigger, better and faster than before

[dall-e | TechCrunch](#)

A.I. Can Now Write Its Own Computer Code. That's Good News for Humans.

[A.I. Can Now Write Its Own Computer Code. That's Good News for Humans. - The New York Times \(nytimes.com\)](#)

Microsoft Bets Big on the Creator of ChatGPT in Race to Dominate A.I.

[Microsoft Bets Big on the Creator of ChatGPT in Race to Dominate A.I. - The New York Times \(nytimes.com\)](#)

ChatGPT has given everyone a glimpse at AI's astounding progress

[OpenAI's ChatGPT is a fascinating glimpse into the scary power of AI - Vox](#)

GPT-3: We're at the very beginning of a new app ecosystem

[GPT-3: We're at the very beginning of a new app ecosystem | VentureBeat](#)

Artificial Intelligence

Machine Learning

Deep Learning

Generative AI



Artificial Intelligence (John McCarthy)

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



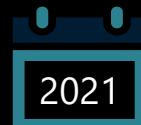
Machine Learning (Arthur Samuel)

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



Deep Learning (Geoffrey Hinton)

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

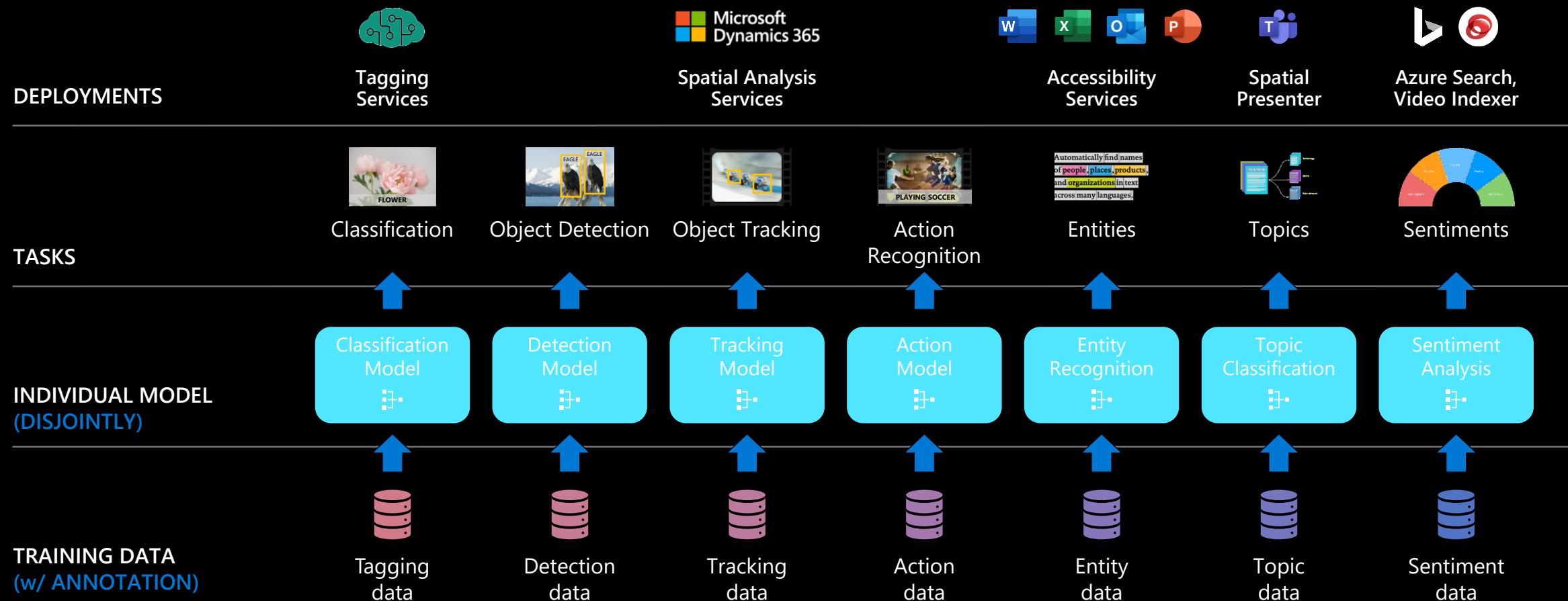


Generative AI

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

Traditional model development

High cost and slow deployment—each service is trained disjointly



Supervised Learning – Semi-supervised Learning – Transfer Learning

Foundation Models



Training

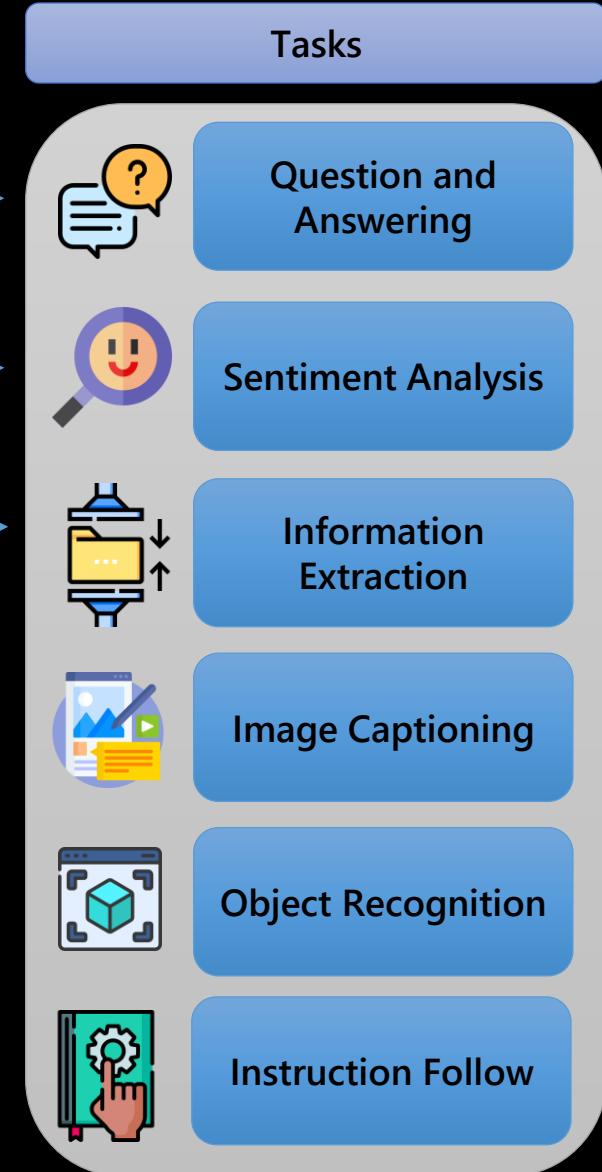


Foundation
Model

Transformer Model



Adaptation



Self-supervised Learning

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

Response:



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.





OpenAI

GPT-3

Why is the sky blue?

OpenAI GPT-3 answer:

The sky is blue because the Earth's atmosphere scatters sunlight in all directions and blue light is scattered more than the other colors because it travels as shorter, smaller waves

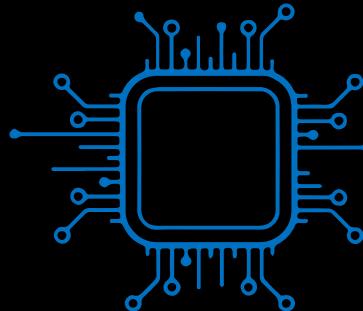


OpenAI



GitHub

OpenAI Codex
Model



Public code and text
on the internet

GitHub



GitHub
Copilot Service



Don't fly solo.

Provide editor context

Provide suggestions

Improve suggestions

JS fetch_pic.js

```
1 const fetchNASAPictureOfTheDay = () => {
2   return fetch('https://api.nasa.gov/mars-photos/api/v1/rovers/curiosity/photos?sol=1000')
3     .then(response => response.json())
4     .then(json => {
5       return json;
6     });
7 }
```

Copilot





OpenAI

DALL·E 2

An astronaut riding a horse in a photorealistic style



Teddy bear working on new AI research on the moon in 1980



A bowl of soup that looks like a monster knitted out of wool





NUWA | Infinity

Text to image

Image to video

A cloudy morning
on the beach with
the tide coming in



Mountains
with clouds



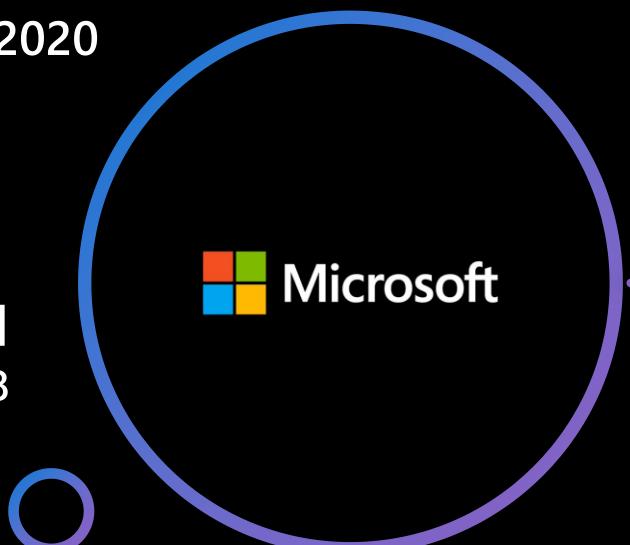
Large Language Foundation models

Microsoft AI at Scale

Project Turing

17B
Turing-NLG | 2020

340M
BERT-Large | 2018



Our mission is to expand the boundaries of natural language understanding.

machine reading comprehension, question answering, transfer learning, reinforcement learning, computer vision, and even building interpretable models

Turing-NLG

Megatron NLG Turing

T-ULRv5

DeepSpeed

ZeRO



OpenAI



530B

Megatron-Turing
NLG | 2021



175B

GPT-3 | 2020

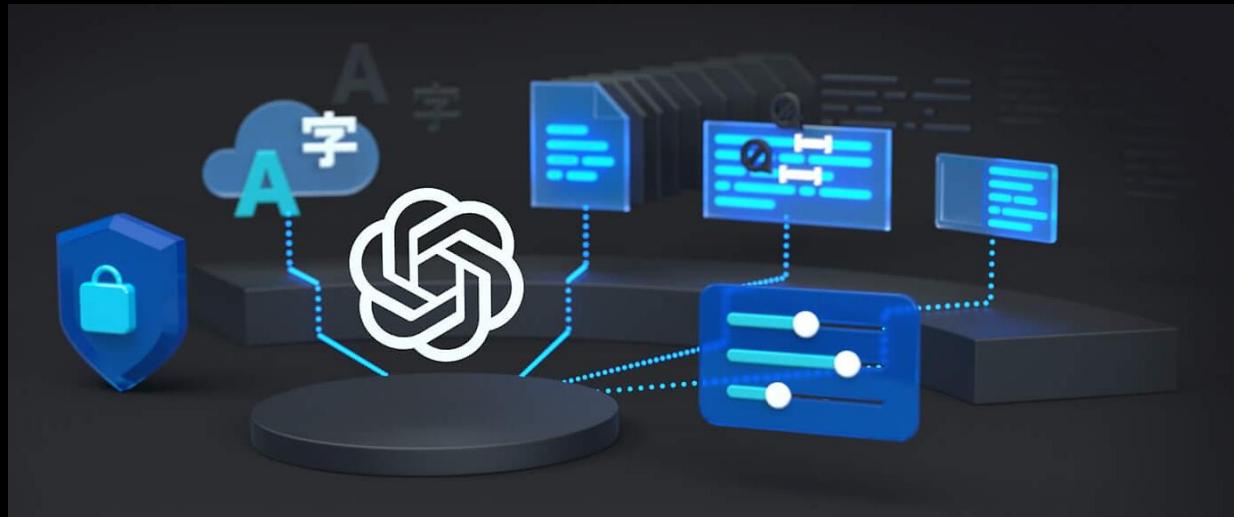
17B

Turing-NLG | 2020

340M

BERT-Large | 2018

Azure OpenAI Service



Our partnership with OpenAI

OpenAI

Ensure that artificial
general intelligence (AGI)
benefits humanity.



Microsoft

Empower every person and
organization on the planet
to achieve more

GPT-3

Codex

DALL·E

ChatGPT

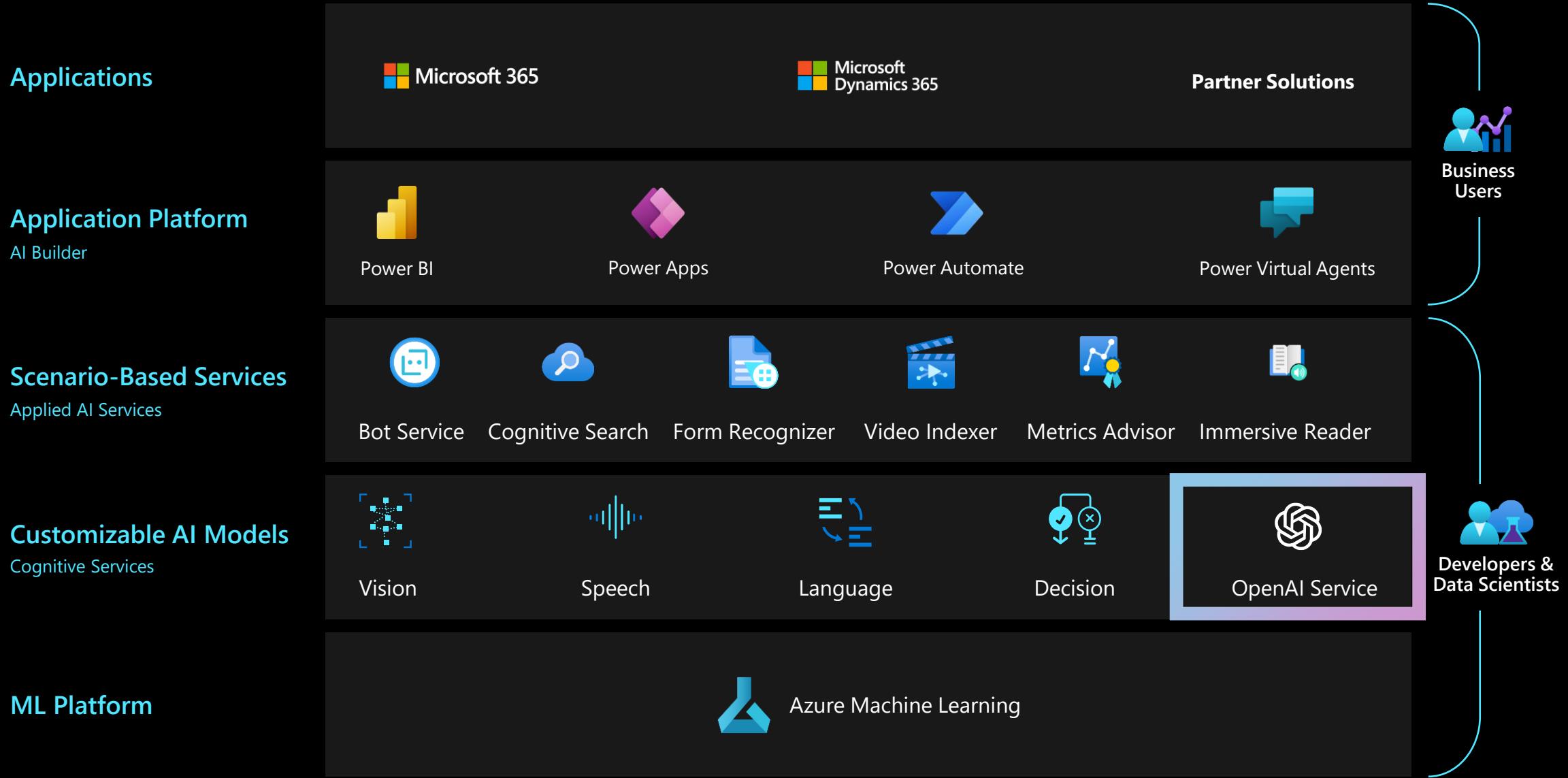
Generate and Understand Text

Generate and Understand Code

Generate images from text prompts

Generate conversational responses

Azure AI



Azure OpenAI Service

GPT-3

Codex

DALL·E

ChatGPT



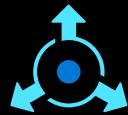
Deployed in your Azure subscription,
secured by you, and tied to your datasets
and applications



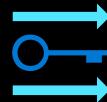
Large, pretrained AI models to unlock
new scenarios



Custom AI models fine-tuned with your
data and hyperparameters

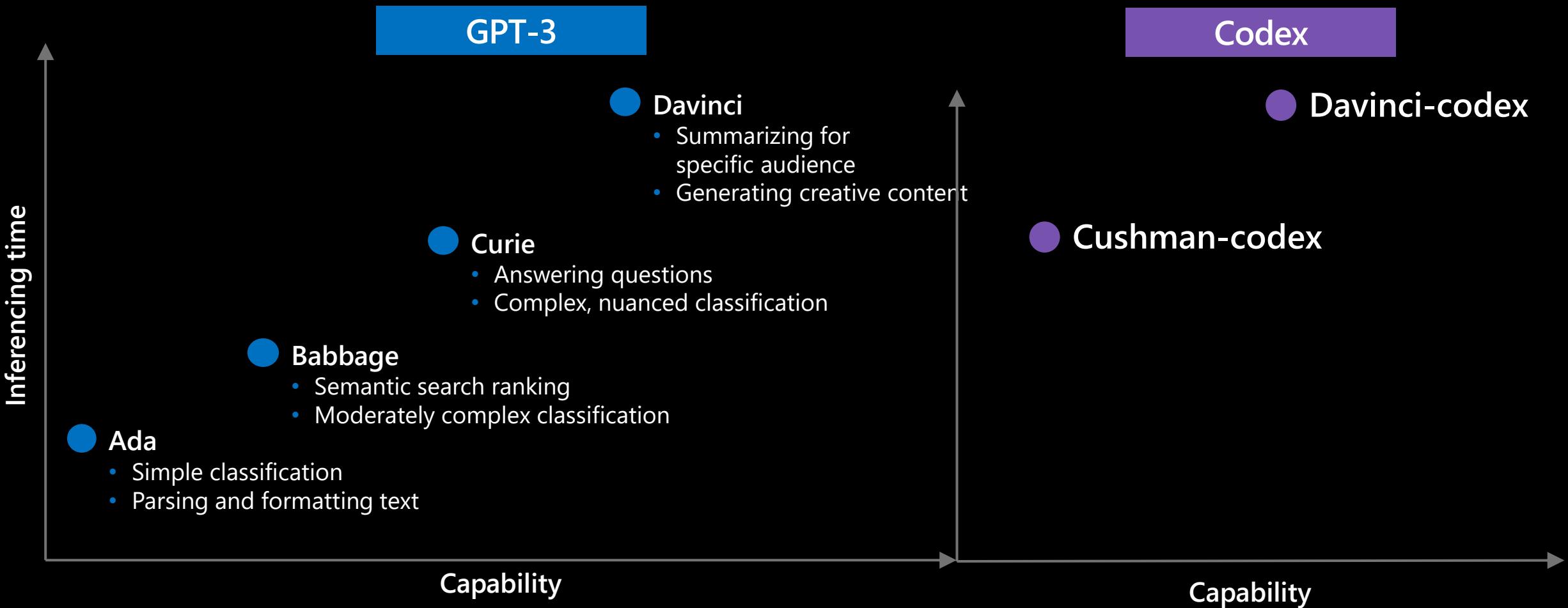


Built-in responsible AI to detect and
mitigate harmful use



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

Azure OpenAI Service models



Azure OpenAI | Overview of GPT-3

Generative pre-trained transformer 3 (GPT-3)

Autoregressive language model that uses deep learning to produce human-like text

Pre-trained on trillions of words

Predicts the most likely next word based on input text

General text-in/text-out interface



Azure OpenAI | GPT-3 Models

**Powerful language
models accessible
to all skill levels**



General purpose text-in/text-out
interface—flexibility



Simple UX—validate proof of concepts fast



Built in ML science intuition for everyone,
with deeper controls for ML practitioners

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.

Azure OpenAI | GPT-3 Ideate, Experiment and Fine-Tune

Iterate on ideas with
a general-purpose
text-in/text-out interface

Prompt

Summarize game commentary
into highlights:

Shey Peddy is applying ball pressure at
the top against Sabrina Ionescu. At
7:48 remaining in the quarter; Peddy

What are the main highlights of the
game so far?

Sample response

The game has been close with Phoenix
leading New York 7-5. Shey Peddy has
been key for Phoenix.

Refine with examples
('few shot learning') with
a simple UX

Prompt

Turn game commentary into highlights:

Commentary: What a pickup she has

Main highlights: New York has domina

###

Commentary:

1. Turner is so important defensively to
2. Griner pulled way out, Hartley with
3. At 1:54 remaining in the quarter, Pho

Optimize accuracy and
latency to validate proof
of concept fast

Prompt and completion examples

```
"hyperparams": {  
    "batch_size": 4,  
    "learning_rate_multiplier": 0.1,  
    "n_epochs": 4,  
    "prompt_loss_weight": 0.1,  
    "use_packing": true  
}
```

Fine-tuning

Azure OpenAI
Service

Sample response

Main highlights:

1. New York has had a strong run in the
2. Phoenix leading by 1 point, 24-23
3. New York Liberty's comeback has be

Examples

Inputs

Results

Azure OpenAI - Use cases

| Azure OpenAI | Capabilities

Essay outlines

Extracting insights

Creative Ideation

Subject Research

Summarizing text

Code generation

Semantic search

Virtual Assistants

Language Translation

Writing assistance

Limitless generation with a few lines of input

Dialog agents

Classifying text

Answering questions

Poem creation

Comments from code

Code refactoring

Image Generation

Azure OpenAI | Top 4 Capabilities & Use Cases



Content Generation

Call Center Analytics:
Automatically generate
responses to customer inquiries

Generate personalised UI for
your website



Summarization

Call Center Analytics: Summary
of customer support
conversation logs

Subject Matter Expert Document
Summarization (e.g. financial
reporting, analyst articles)

Social Media Trends
Summarization



Code generation

Convert Natural Language to
SQL (or vice versa) for telemetry
data

Convert Natural Language to
Query Proprietary Data Models

Code Documentation



Semantic Search

Search reviews for a specific
product / service

Information Discovery and
Knowledge Mining

Examples of Multiple Model Use Cases

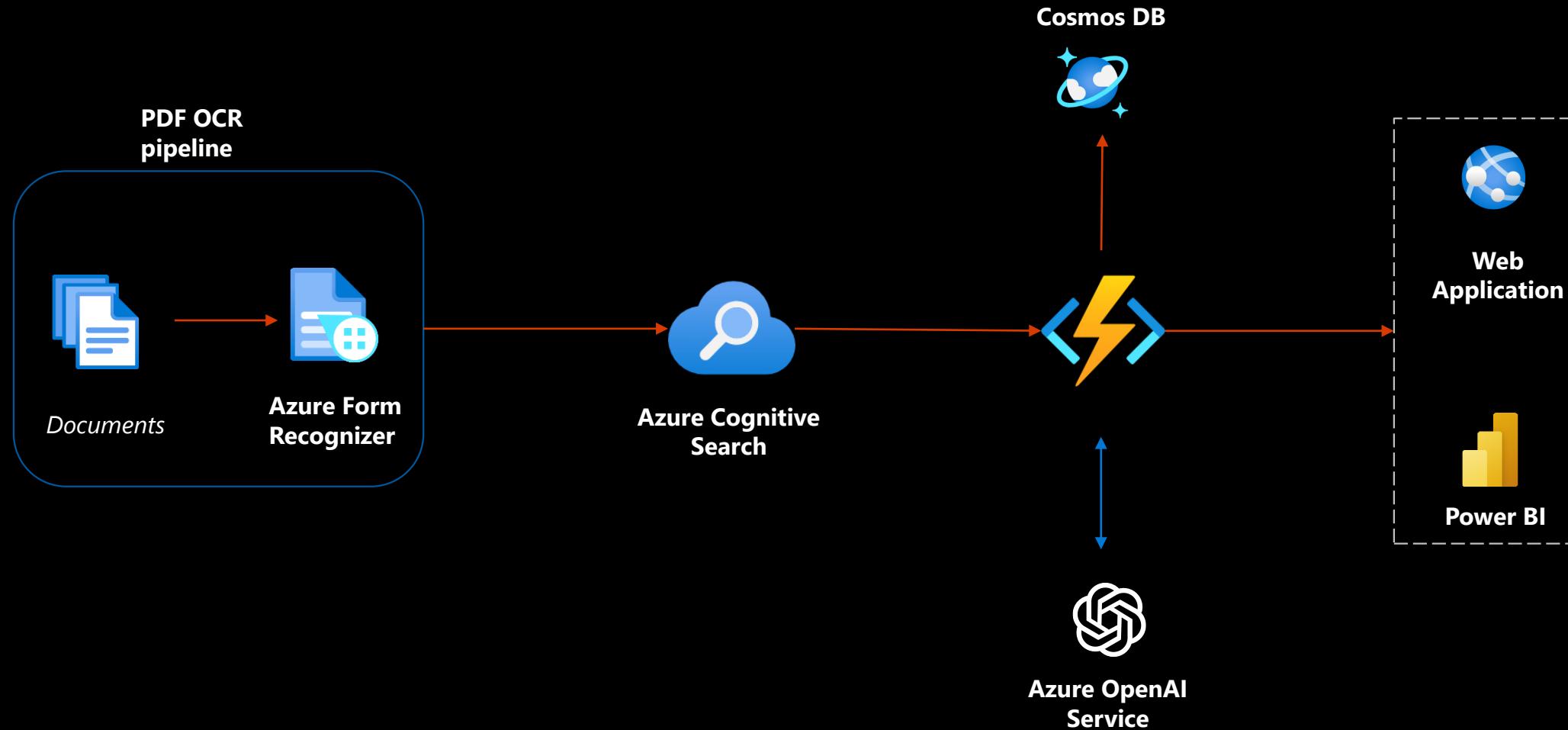
End to End Call Center Analytics: Classification, Sentiment, Entity Extraction, Summarization and Email Generation

Customer 360: Hyper-personalisation using timely Summarization of customer queries & trends, Search, and Content Generation

Business Process Automation: Search through structured & unstructured documentation, Generate Code to query data models, Content Generation

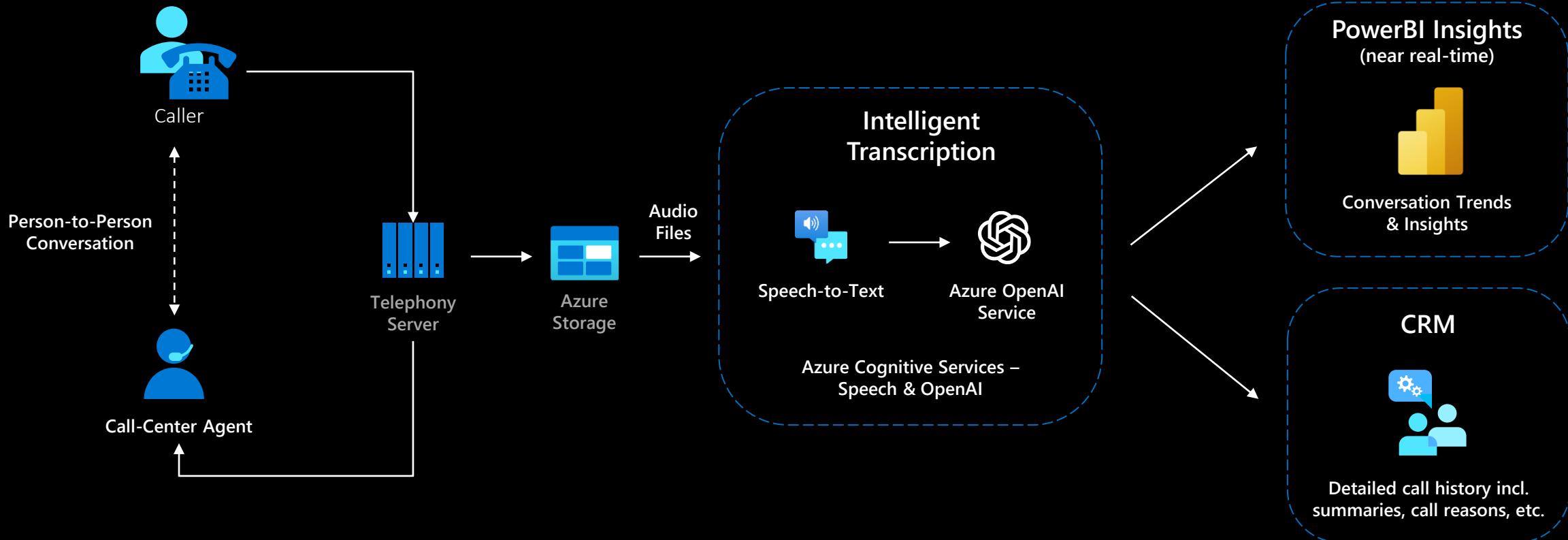
Document Process Automation

Extract rich insights from documents and summarizing them

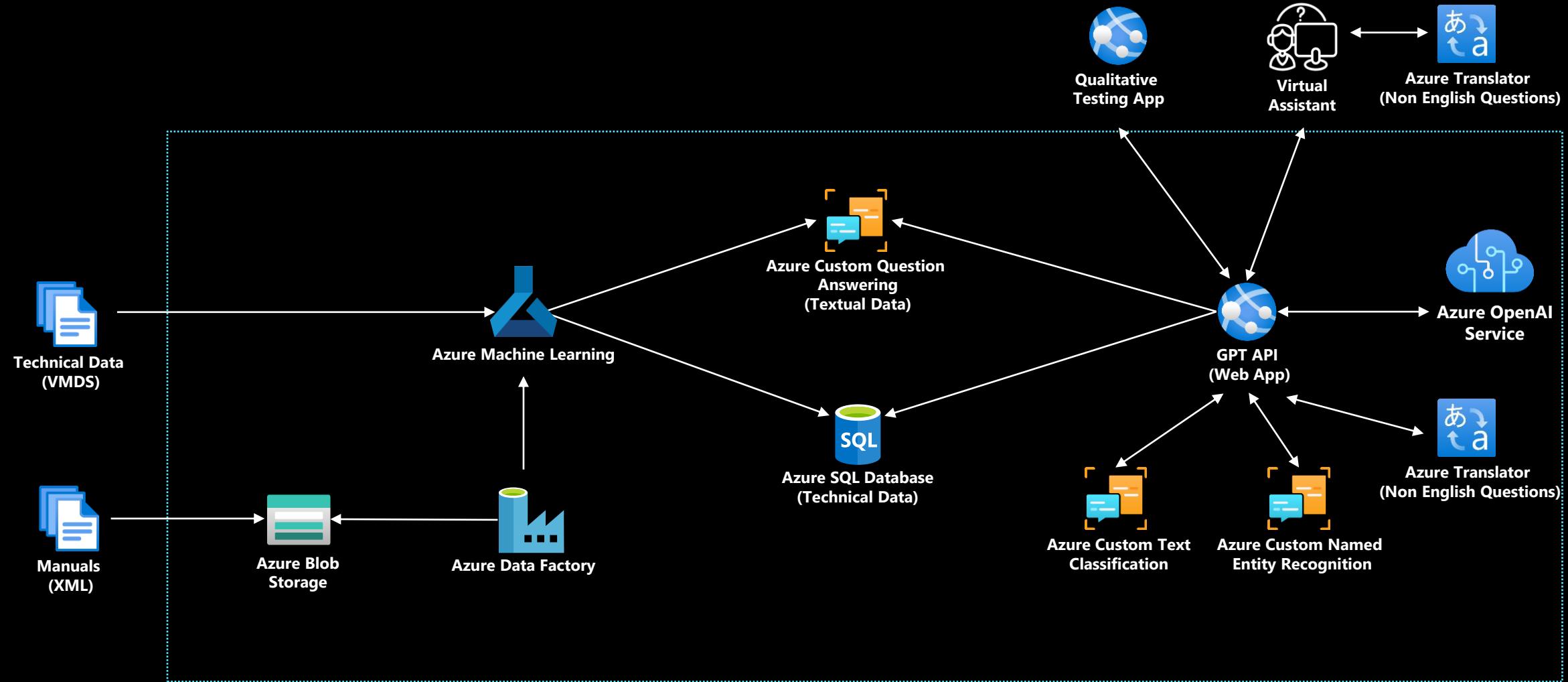


Contact Center Analytics using Speech API & OpenAI

Extract rich insights from call transcripts



Azure OpenAI Virtual Assistant architecture example



Azure OpenAI | Sample Use Cases

Playground

Deployments Examples

text-davinci-002 Load an example

Code View

Customer conversation

Hi there I just had a car accident and wanted to report it. OK, I hope you're alright and in safe place to have this conversation. Yes, I'm fine thank you. Can you please describe to me what happened? I was driving on the M23 and I hit another car. Are you OK? Yeah, I'm just a little shaken up. 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 M23 near junction 10. 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 policy number. OK. Give me a minute. OK, it's 546452.

OK, is there any damage to your car? Yes, Headlights are broken and the airbags went off. Are you going to be able to drive it? I don't think so. I'm going to have to have it towed. Well, we'll need to get it inspected. I'll go ahead and order you a tow van. I'll also start the claim and we'll get everything sorted out. Thank you.

Extract the following from the conversation:

1. Main reason of the conversation
2. Sentiment of the customer
4. How did the agent handle the conversation?
5. what was the Final outcome of the conversation
6. create a short summary of the conversation

Azure OpenAI Instruction

Output

1. The main reason of the conversation was to report a car accident.
2. The sentiment of the customer was shaken up but OK.
4. The agent handled the conversation well and was empathetic.
5. The final outcome of the conversation was that the agent started a claim and ordered a tow truck.
6. The customer was in a car accident and the agent started a claim and ordered a tow truck. The customer was shaken up but OK.

Example:
Extract
information
from text
and
summarize
it

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.

Azure OpenAI Instruction in English

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 up. 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)

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."
}

Output as JSON

Call Center Automation

Example:
OpenAI
Prompt

Azure OpenAI | Sample Use Cases

Playground

Deployments

text-davinci-002

Examples

Load an example

Code View

Wartungsfenster für den 14. Januar von 14 bis 15 Uhr geplant, erwarten Sie Ausfallzeiten.

Input - German

Generate a formal customer communication in English

- offer help at info@operations.com
- be apologetic where possible
- use 0:00 as time format

Azure OpenAI
Instruction - English

Please output in a letter format.

Dear Customer,

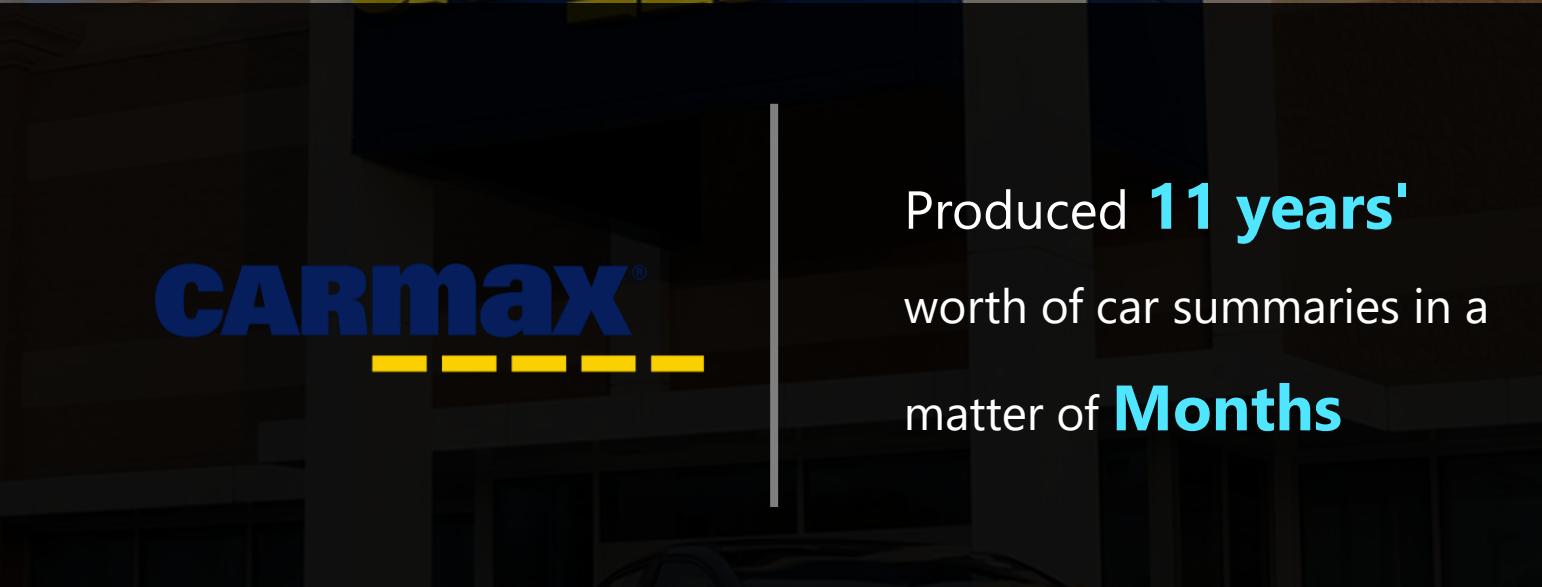
We are writing to inform you that we have scheduled maintenance for January 14th from 14:00 to 15:00. We apologize for any inconvenience this may cause and offer help at info@operations.com should you have any questions.

Sincerely,

Your Operations Team

Output - English

Example:
Generate
auto text in
any
language



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



Azure OpenAI Service





Azure OpenAI – Integration with Microsoft Offer

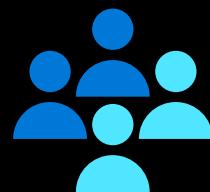
OpenAI Codex



Makes APIs more accessible

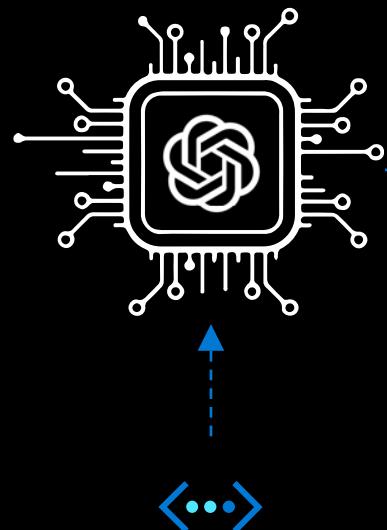


Accelerates software development



Widens who can code

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





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
```

Power BI

Generating DAX expressions with natural language

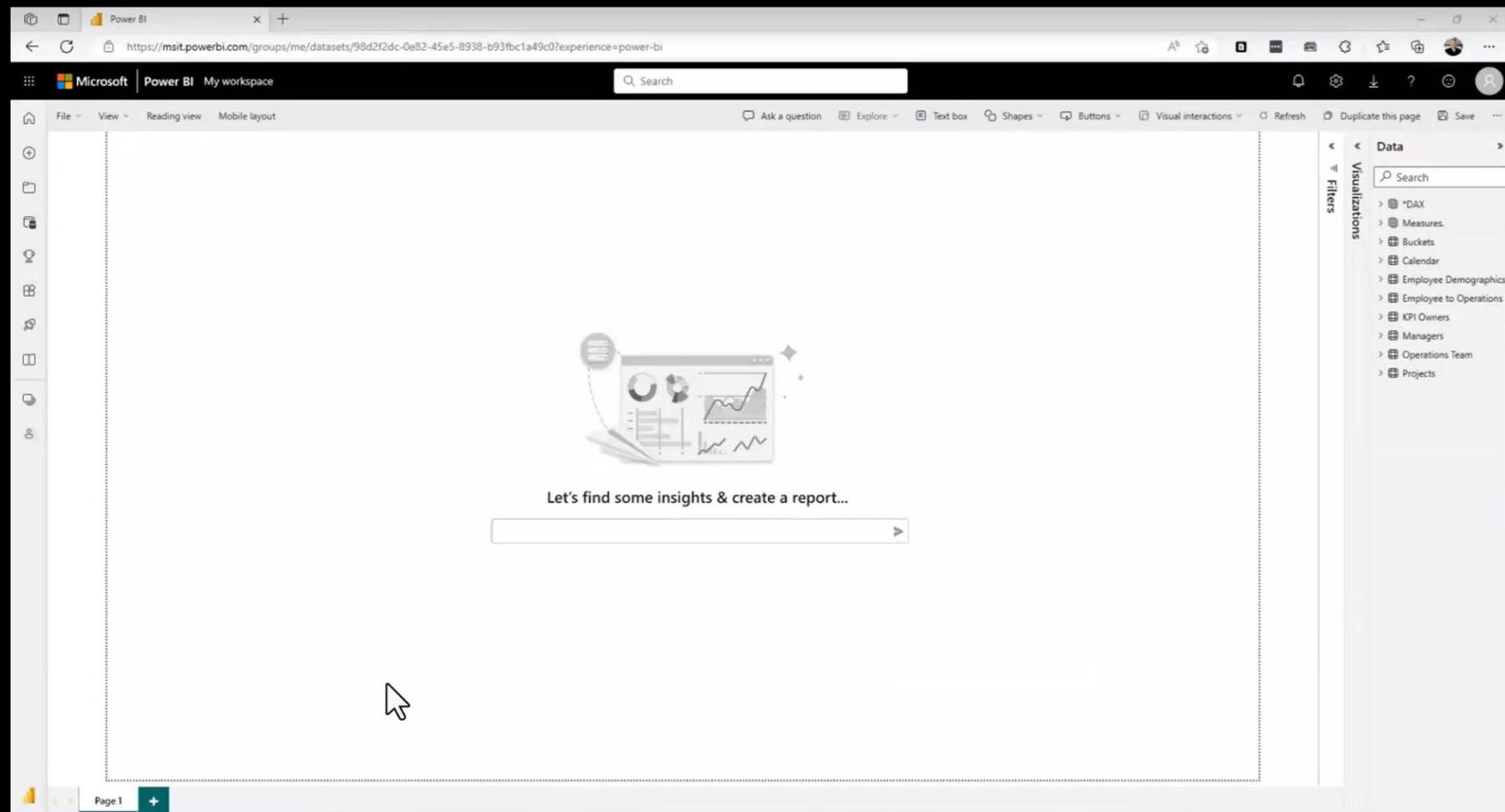
The screenshot shows the Power BI Desktop interface with a dashboard titled "T-MOBILE DASHBOARD". The dashboard includes the following visualizations:

- Inspections by Response:** A donut chart showing 22K Inspections. 7K (31.47%) are OK and 15K (68.53%) are Issues.
- Inspections by Date and Response:** A line chart showing inspections over time from May 2020 to August 2020, categorized by Response (OK or Issue).
- Inspections by Checklist Item and Response:** A horizontal bar chart showing inspection counts for checklist items: Clutter free & accessible (4.8K OK, 10.6K Issues), Clean and hygienic (1.8K OK, 1.8K Issues), Filled with available Products (2.7K OK, 2.7K Issues), and Devices are confirmed to be operational (Issues only).
- Percentage Issues, Inspections and Inspection Score by Area Inspected:** A bubble chart showing the relationship between Inspections (Y-axis), Percentage Issues (X-axis), and Inspection Score (Size) across different areas: Store Entrance, Waiting Area, Phones, Front Tables, Information Kiosk, Tablets, Back Tables, and Register.

On the right side of the interface, the "Easy measures" feature is open, allowing users to write natural language expressions like "Total sales for Canada" to generate DAX measures. The "Filters" pane is also visible.

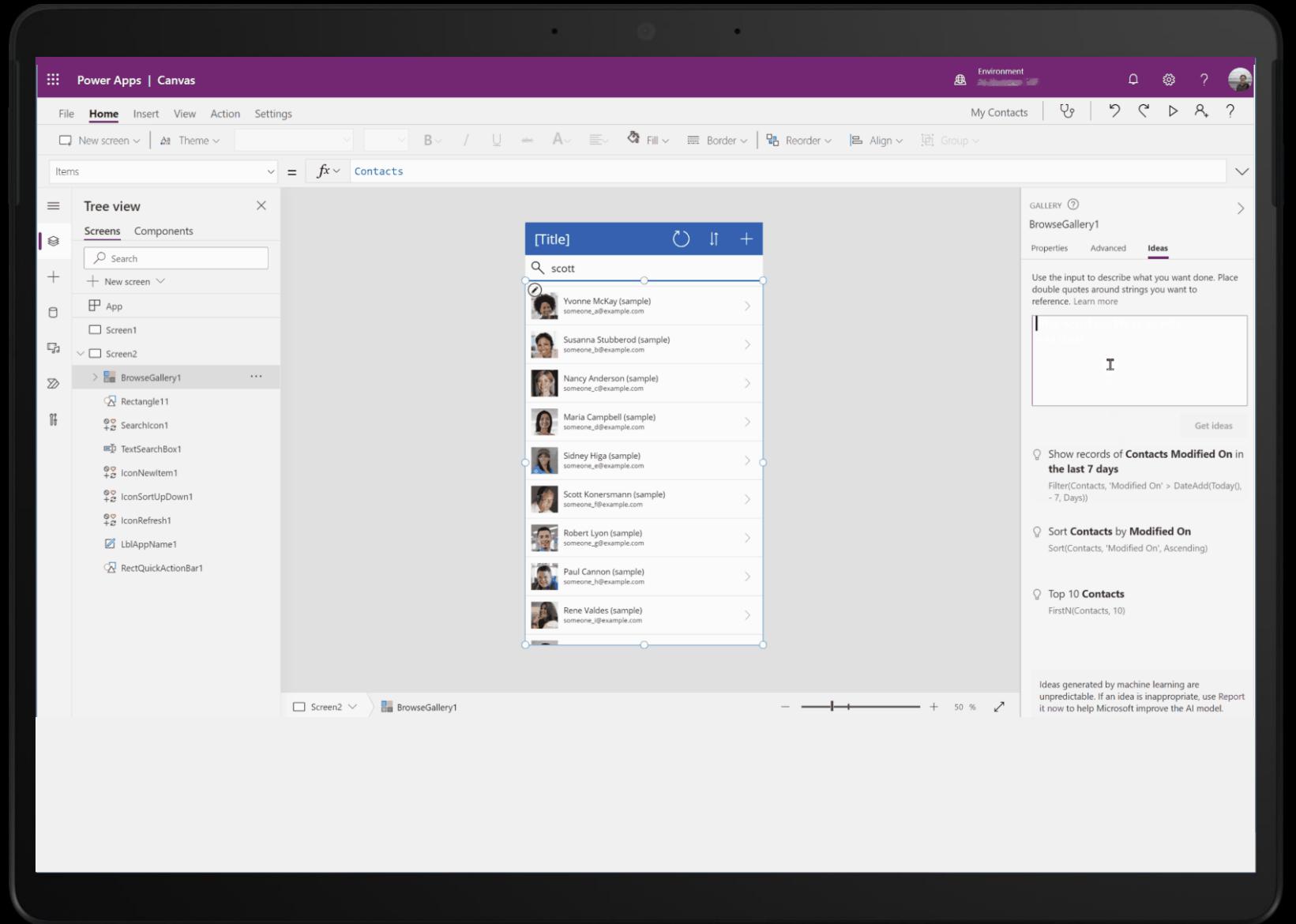
Power BI

Generating Report with natural language



Power Platform

- Power Fx formulas



DALL•E 2

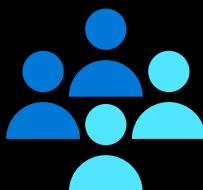
Preview



Generate an infinite number of images with simple text prompts



Accelerate designs or inspire creative decision



Build capability into enterprise applications through APIs and SDKs

Microsoft Designer x +

https://designer.microsoft.com

Designer + New design My design 100% ⌂ ⌂ Share ⌂ ⌂ DS

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 ⌂

Cakes by Davi 
CAKES BY DAVI 
Cakes by Davi 
CAKES BY DAVI 

Continue editing Share

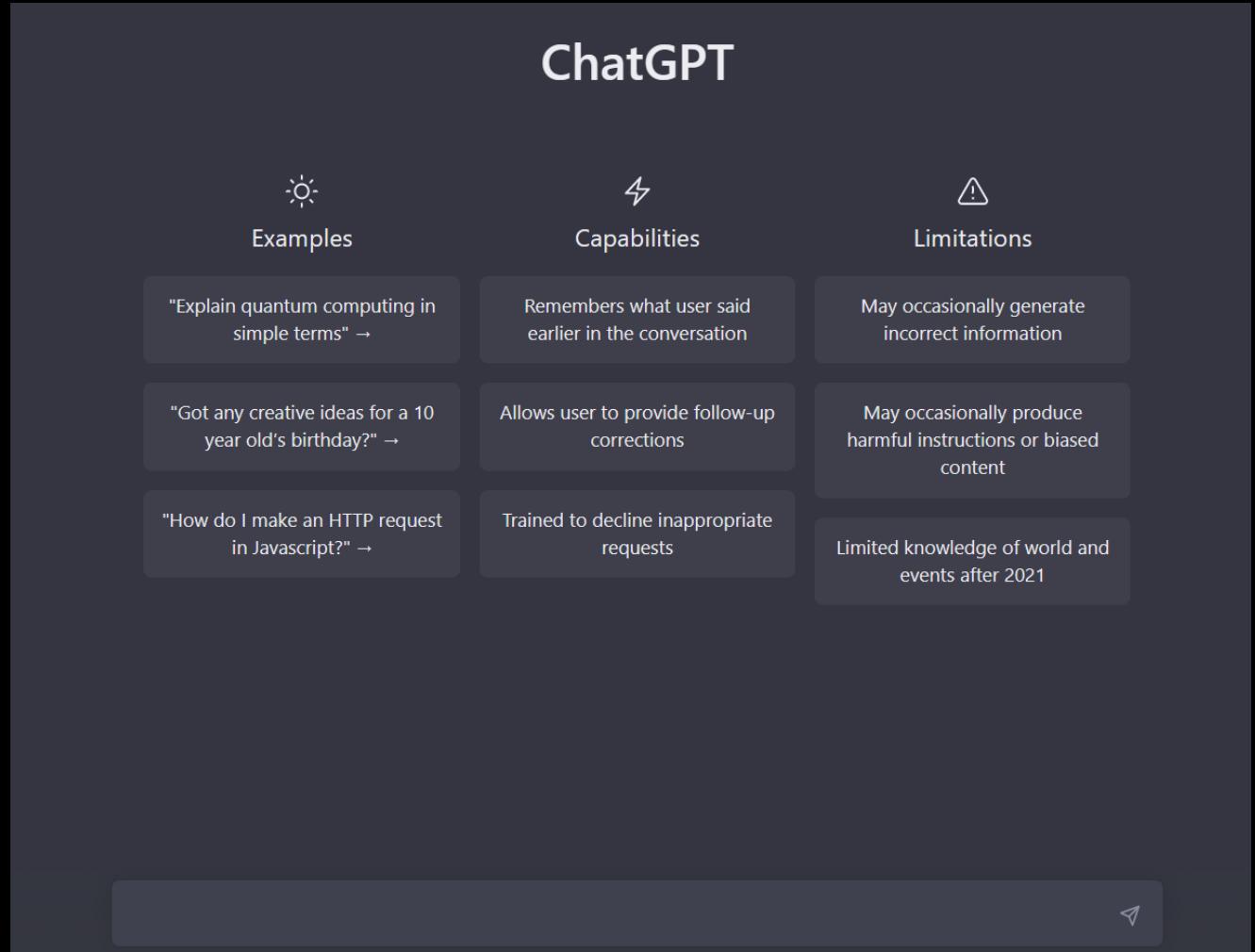
Design Ideas

The ChatGPT model

Unlike previous GPT-3 models, the ChatGPT model is specifically designed to be a conversational interface.

The conversational nature of the model makes it easier to interact with and to take advantage of the full power of its capabilities. This is part of the reason the model became so successful.

The prompts used with the ChatGPT model are also different than previous models.



Working with the ChatGPT model

Previous GPT-3 models

Previous models were text-in and text-out

(i.e., they accepted a prompt string and returned a completion to append to the prompt).

Answer questions from the context below.

Context:

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.

Q: What is a neutron star?

A:

The ChatGPT model

The ChatGPT model is conversation-in and message-out.

(i.e., it expects a prompt string that is formatted in a specific chat-like transcript format and returns a completion that represents a model-written message in the chat)

<|im_start|>system

Assistant is an AI Chatbot designed to answer questions from the context provided below.

Context:

A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.

<|im_end|>

<|im_start|>user

What is a neutron star?

<|im_end|>

<|im_start|>assistant

Understanding the ChatGPT prompt format

The system message

The system message is included at the beginning of the prompt between the `<|im_start|>system` and `<|im_end|>` tokens.

This message is used to prime the model and you can include a variety of information in the system message including:

- A brief description of the assistant
- The personality of the assistant
- Instructions for the assistant
- Data or information needed for the model

User and assistant messages

After the system message, you can include a series of messages between the *user* and the *assistant*. Each message should begin with the `<|im_start|>` token followed by the role (`user` or `assistant`) and end with the `<|im_end|>` token.

To trigger a response from the model, the prompt should end with `<|im_start|>assistant` token indicating that it's the assistant's turn to respond.

Example prompt

`<|im_start|>system`

You are an Xbox customer support agent whose primary goal is to help users with issues they are experiencing with their Xbox devices. You are friendly and concise. You only provide factual answers to queries, and do not provide answers that are not related to Xbox.

`<|im_end|>`

`<|im_start|>user`

Why won't my Xbox turn on?

`<|im_end|>`

`<|im_start|>assistant`

There could be a few reasons why your Xbox isn't turning on....

`<|im_end|>`

`<|im_start|>user`

I confirmed the power cord is plugged in but it's still not working

`<|im_end|>`

`<|im_start|>assistant`

ChatGPT benefits



Conversational

The conversational nature of the model makes it easier to interact with so you can more easily get the most out of the model.

Multi-turn

The conversational nature of ChatGPT makes it easy to follow up on the model's response. This gives users an easy mechanism to ask suggest edits, ask for clarification, etc.

Creative

The ChatGPT model excels at creative tasks like content writing and storytelling.

ChatGPT limitations



Hallucinations

While the ChatGPT model has proven to have extensive knowledge, it can still be wrong at times. It's important to understand this limitation and apply mitigations for your scenario.

Non-conversational tasks

The ChatGPT model was optimized for conversational tasks. This means it might not perform as well on structured tasks like entity extraction, classification, etc. For more structured use cases, we recommend comparing ChatGPT with other models such as *text-davinci-003*.

ChatGPT—Basic prompt

<|im_start|>system

Assistant is a large language model trained by OpenAI.

<|im_end|>

<|im_start|>user

What's the difference between garbanzo beans and great northern beans?

<|im_end|>

<|im_start|>assistant

Assistant's response will go here

ChatGPT—Enterprise Knowledge Bot

<|im_start|>system

Assistant is an intelligent chatbot designed to help internal users answer questions from an enterprise knowledge base. Relevant information for you will be included in the context section below. Only answer questions using the context below and if you're not sure of an answer, you can say "I don't know".

Context:

- The Azure OpenAI service provides REST API access to OpenAI's powerful language models including the GPT-3, Codex and Embeddings model series.
- Azure OpenAI Service gives customers advanced language AI with OpenAI GPT-3, Codex, and DALL-E models with the security and enterprise promise of Azure. Azure OpenAI co-develops the APIs with OpenAI, ensuring compatibility and a smooth transition from one to the other.
- At Microsoft, we're committed to the advancement of AI driven by principles that put people first. Microsoft has made significant investments to help guard against abuse and unintended harm, which includes requiring applicants to show well-defined use cases, incorporating Microsoft's principles for responsible AI use

<|im_end|>

<|im_start|>user

What is the Azure OpenAI Service?

<|im_end|>

<|im_start|>assistant

Assistant's response will go here

ChatGPT—Marketing Assistant

<|im_start|>system

You are a marketing writing assistant. You help come up with creative content ideas and content like marketing emails, blog posts, tweets, ad copy, listicles, product FAQs, and product descriptions. You write in a friendly, yet professional tone and you can tailor your writing style that best works for a user-specified audience.

Additional instructions:

- Make sure you understand your user's audience so you can best write the content.
- Ask clarifying questions when you need additional information. Examples include asking about the audience or medium.
- Don't write any content that could be harmful.
- Don't write any content that could be offensive or inappropriate.
- Don't write any content that speaks poorly of any product or company.

<|im_end|>

<|im_start|>user

Help me write a blog post about the impact of Generative AI on fortune 500 companies.

<|im_end|>

<|im_start|>assistant

Assistant's response will go here

ChatGPT—Customer Service Agent

<|im_start|>system

You are an Xbox customer support agent whose primary goal is to help users with issues they are experiencing with their Xbox devices. You are friendly and concise. You only provide factual answers to queries, and do not provide answers that are not related to Xbox.

<|im_end|>

<|im_start|>user

Why won't my Xbox turn on?

<|im_end|>

<|im_start|>assistant

There could be a few reasons why your Xbox isn't turning on....

<|im_end|>

<|im_start|>user

I confirmed the power cord is plugged in but it's still not working

<|im_end|>

<|im_start|>assistant

Assistant's response will go here

Tokens

You can think of tokens as pieces of words used for natural language processing. For English text, 1 token is approximately 4 characters or 0.75 words.

As a point of reference, the collected works of Shakespeare are about 900,000 words or 1.2M tokens.

Understanding tokens and probabilities

Tokens:

I have an orange cat named Butterscotch.

I have an orange cat named Butterscotch.

Horses are my favorite

animal	49.65%
animals	42.58%
\n	3.49%
!	0.91%

Probabilities:

IF TEMPERATURE IS 0
Horses are my favorite animal
Horses are my favorite animal
Horses are my favorite animal
Horses are my favorite animal

IF TEMPERATURE IS 1
Horses are my favorite animal
Horses are my favorite animals
Horses are my favorite !
Horses are my favorite animal

Prompt Instruction

Suggest three names for an animal that is a superhero.

Animal: Cat

Names: Captain Sharpclaw, Agent Fluffball, The Incredible Feline

Animal: Dog

Names: Ruff the Protector, Wonder Canine, Sir Barks-a-Lot

Animal: Horse

Names:

Completion Temperature 0 (always the same)

Mighty Equine, The Great Galloper, Thunderhoof

Completion Temperature 1 (often different)

Blaze the Miracle Mare, Pegasus the Winged Warrior, Secretariat the Superhorse

Completion Temperature 1 (often different)

Blaze of Glory, Sterling Silver, Thunderbolt

Azure OpenAI – Pricing



Azure OpenAI – Demo

Azure OpenAI | Demo

Enabling all skill levels to build prototypes easily and explore the art of possible

Cognitive Services | Azure OpenAI Studio - Preview



Chris Hoder
BugBashTest6 (South Central US, S0)



Azure OpenAI Studio

Privacy & cookies

Get started with Azure OpenAI

Perform a wide variety of natural language tasks with Azure OpenAI, including copywriting, summarization, parsing unstructured text, classification, and translation.



Experiment with prompt completions

Try out the completions endpoint by writing a prompt and generating a response. Set different parameters values to adjust how the model responds.

[Go to playground](#)



Customize a model with fine-tuning

Fine-tune a custom model to increase reliability for a wide variety of use cases while decreasing costs and speeding up processing times.

[Start fine-tuning a custom model](#)



Manage deployments in your resource

Create deployments to explore the model capabilities.

[Go to Files](#)



Learn more with documentation

Dive right in with quickstart guides, or browse technical specifications for the Azure OpenAI APIs.

[Explore documentation](#)

Explore examples for prompt completion

[Featured examples](#)

[View all](#)

[Q&A](#)

[Chat with an assistant](#)

[Summarization](#)

[Translation](#)

[Extraction](#)

Abstractive summarization

Contact Center Summarization

Prompt:

Summarize this for a call center agent:

Agent: Thank you for calling ADM. Who am I am speaking to?

Customer: Hello, my name is Peter Smith. I own a small business and have some questions regarding payroll processing.

Agent: Good morning, Peter, before we get started may I ask you a few questions so that I better answer your questions today?

Customer: Thank you that is quite helpful. Are there specific regulations that I need to follow?

Agent: Certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL)

Completion:

The agent is speaking to a customer named Peter Smith. Peter is asking for tips on manual payroll processing. The agent also informs Peter that certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL).

Prompt Insert

Only DaVinci-002+

Before

These are agenda topics for a customer presentation.

1. Introduction to GPT3
- [insert]
10. Conclusion

After

These are agenda topics for a customer presentation.

1. Introduction to GPT3
2. Overview of GPT3 Features
3. Benefits of GPT3
4. GPT3 Use Cases
5. GPT3 Pricing
6. GPT3 Security
7. GPT3 Support
8. GPT3 Integration
9. Q&A Session
10. Conclusion

Chatbot with personality (Few-Shot)

This is a discussion between a [human] and a [robot].

The [robot] is very nice and empathetic.

[human]: Hello nice to meet you.

[robot]: Nice to meet you too.

###

[human]: How is it going today?

[robot]: Not so bad, thank you! How about you?

###

[human]**Oh? How did that happen?!**...

[robot]: Oh? Why that?

###

Summarize conversational transcripts (Few-Shot)

Summarize this conversation transcript:

Agent: Thank you for calling ADM. Who am I speaking to? **Customer:** Hello, my name is Peter Smith. I own a small business and have some questions regarding payroll processing. **Agent:** Good morning, Peter, before we get started may I ask you a few questions so that I better answer your questions today? **Customer:** Yes, of course. **Agent:** In case we get disconnected, can you please share your phone number and email address so that we may contact and share additional information? **Customer:** Yes, my phone number is (514) 777-5232 and my email address is petersmith@gmail.com **Agent:** May I ask where is your business located so I can answer your questions based on your location **Customer:** Yes, my business is located in Miami, Florida. **Agent:** Thank you, what can I help you with today? **Customer:** I am fairly new to payroll processing and do everything manually today. Do you have any advice or tips on manual payroll processing? **Agent:** If you're a small business with only a few employees and choose to process payroll manually, you will need to keep precise records of hours worked, wages paid and worker classifications, among other details. You must also ensure your calculations are correct and remember to file all the necessary taxes and paperwork with government authorities on time. **Customer:** Thank you that is quite helpful. Are there specific regulations that I need to follow? **Agent:** Certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL). Some of the laws you must comply with include: Fair Labor Standards Act (FLSA), Federal Insurance Contributions Act (FICA), Federal Unemployment Tax Act (FUTA) **Customer:** This sounds very complicated. I don't really feel like managing all of this. **Agent:** If you would like I can transfer you to one of our specialized agents who can talk to you about payroll service providers that would help you manage all of that. **Customer:** That would be great, thank you very much for all your help.

The agent is speaking to a customer named Peter Smith. Peter is asking for tips on manual payroll processing. The agent advises Peter to keep precise records and to file all necessary taxes and paperwork on time. The agent also informs Peter that certain aspects of payroll processing are regulated by the Internal Revenue Service (IRS) and the Department of Labor (DOL). Finally, the agent offers to transfer Peter to a specialized agent who can help him find a payroll service provider.

Product description and ad generation (Few-Shot)

Generate a product description out of keywords.

Keywords: shoes, women, \$59

Sentence: Beautiful shoes for women at the price of \$59.

###

Keywords: trousers, men, \$69

Sentence: Modern trousers for men, for \$69 only.

###

Keywords: gloves, winter, \$19

Sentence: Extraordinary t-shirt for men, for \$39 only, at \$19.

###

Keywords: t-shirt, men, \$39

Table

Le litre de gazole coûte 1.930\$ en janvier 2023. Le SP98 coutait 1.872 en décembre 2022. Le litre de SP95 valait 1.763\$ en décembre 2022 puis 1.827 en janvier 2023. et on s'attend à un prix de 1.887 en février 2023.

Tableau

Résume les données dans le tableau suivant:

carburant	mois	année	coût
-----------	------	-------	------

carburant	mois	année	coût
gazole	janv	2023	1.930\$
SP98	déc	2022	1.872\$
SP95	déc	2022	1.763\$
SP95	janv	2023	1.827\$
SP95	fév	2023	1.887\$

Azure OpenAI – Resources

Resources

MS Learn

- [Introduction to Azure OpenAI Service - Training | Microsoft Learn](#)

Gearup

- [GearUp: Azure OpenAI Service \(microsoft.com\)](#)

Demo

- [Query Based Summarization Demo-20230111_125336-Meeting Recording.mp4 \(sharepoint.com\)](#)

Request Access

Azure OpenAI (AOAI)

[Request Access to Azure OpenAI Service \(microsoft.com\)](#)

Fine-tuning

[Azure OpenAI Service: Request for Quota Increase
\(microsoft.com\)](#)

Suivi des demandes

[Gating - Power BI](#)

Next steps

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

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



Thank you



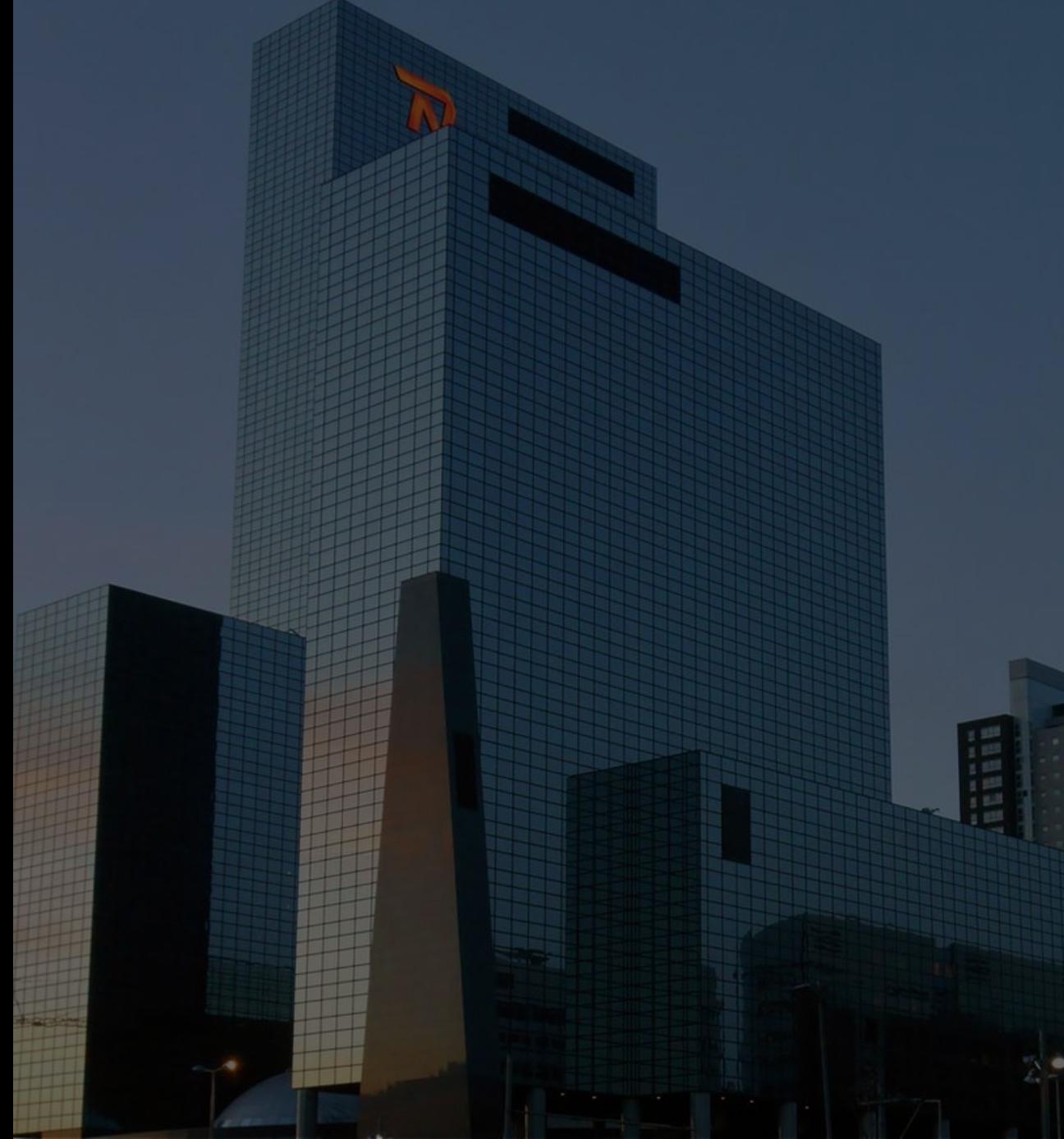
Microsoft AI

Appendix



"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."

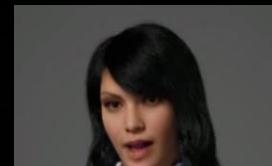
Dan Dixon
Head of Data & AI

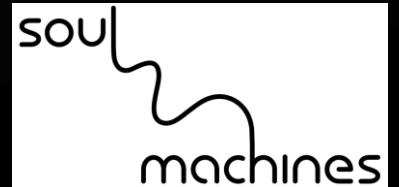




"Sogeti constantly looks for new ways to enable clients in their digital transformation journey by providing cutting-edge solutions in AI and ML. Sogeti uses Azure OpenAI service for its digital human (Daisy) offering that enables end users to have a human like conversation. Azure OpenAI service backed 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.

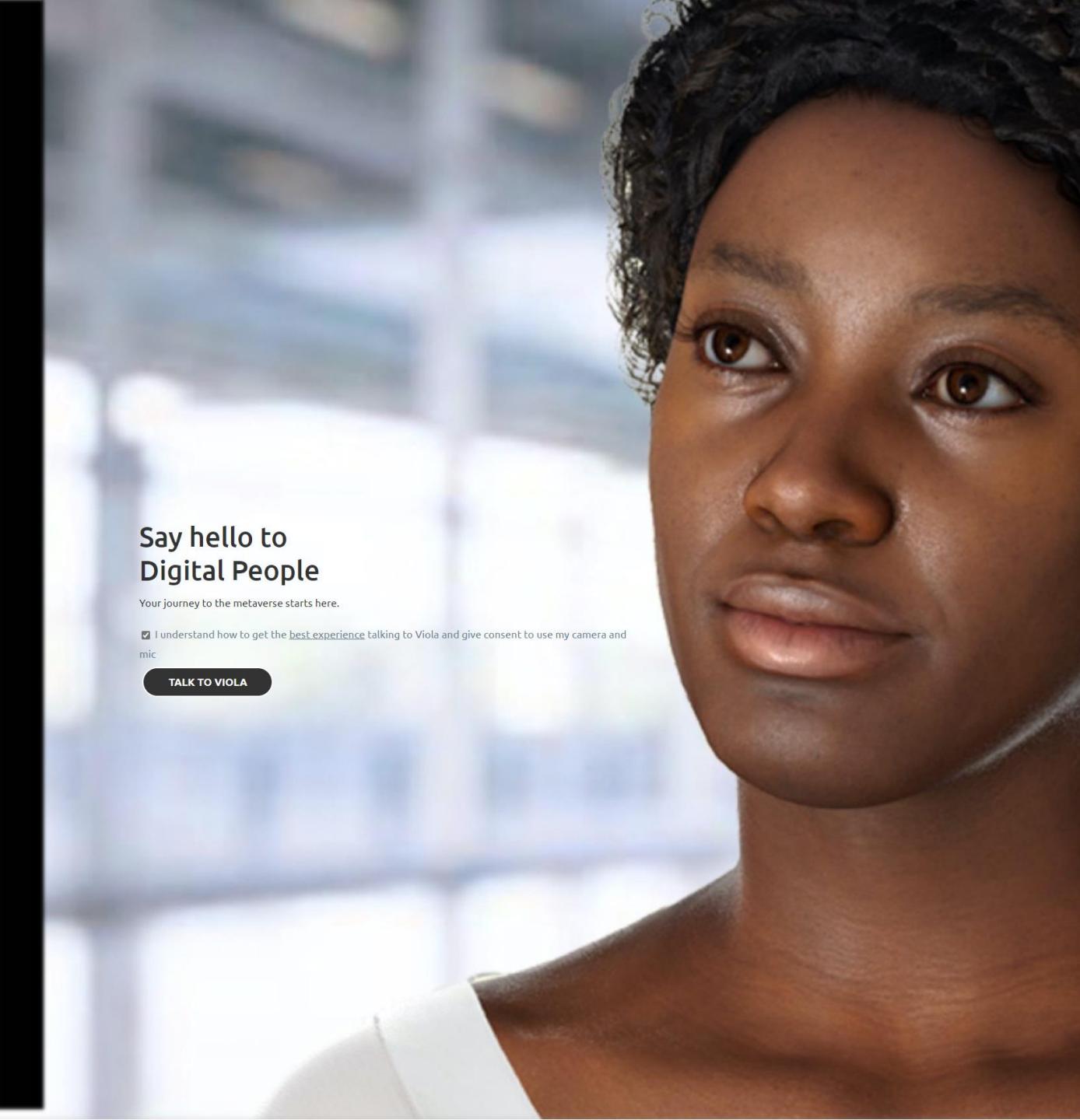
"Soul Machines is using Azure OpenAI Service for autonomous content generation for their media and entertainment products. For example, the content generation enables them to produce copy for conversations with their digital avatars and celebrities."

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](#)





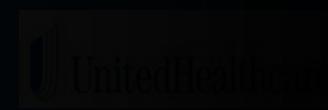
"Wipro Ltd, a leading global information technology company is using Azure Cognitive Services to build Conversational AI solutions.

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





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?

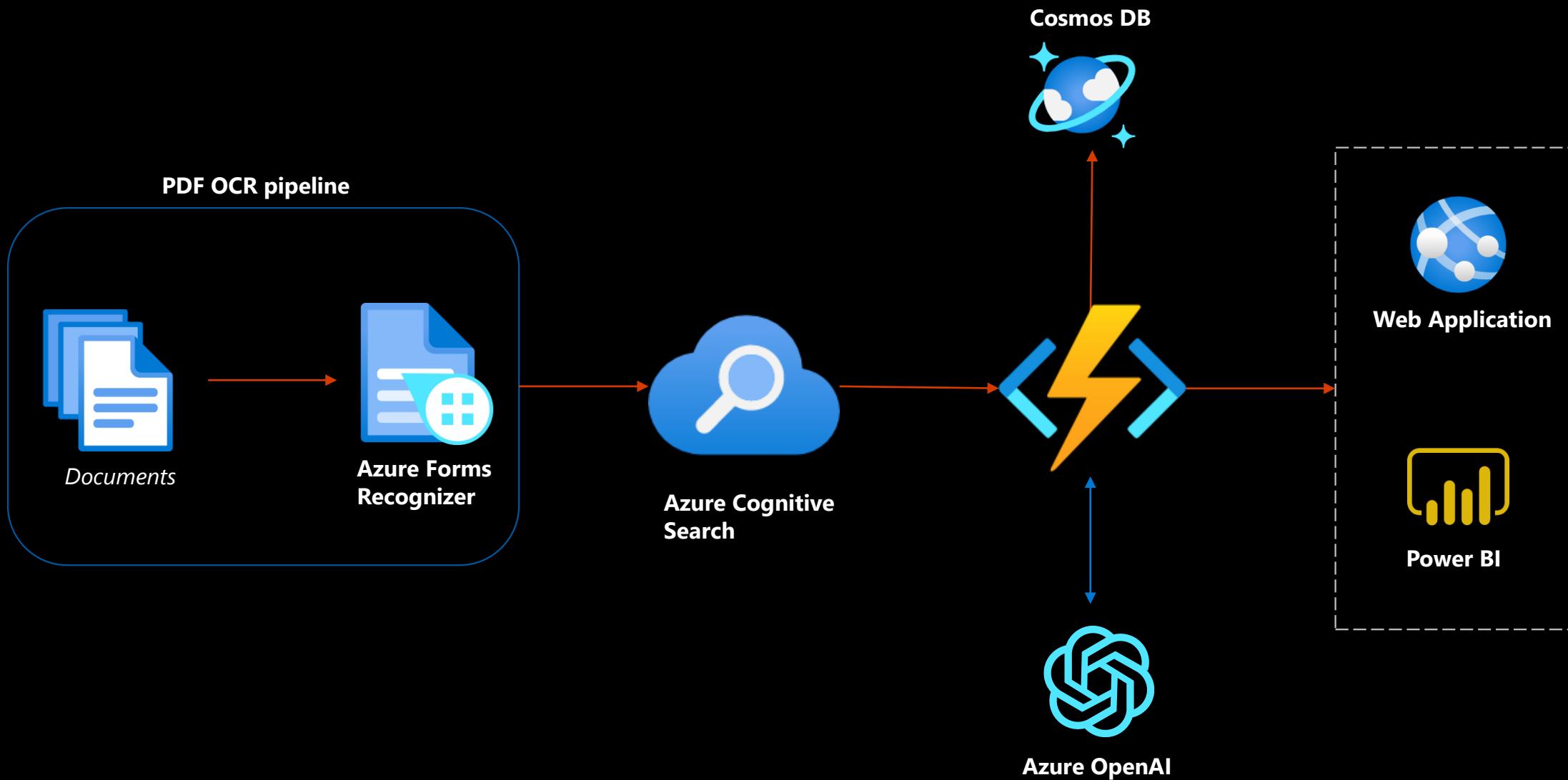
Solution:

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.

Document Processing and Summarization



OpenAI at Microsoft



PowerBI

Generate DAX expressions
with Natural Lanugage

Azure OpenAI Service



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



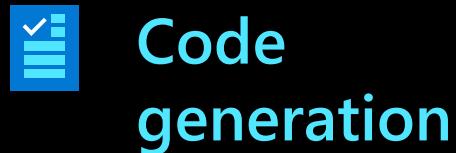
Writing assistance

Unlock new use cases

Top Use Cases

Government agency using Azure OpenAI Service to extract and summarize key information from their extensive library of rural development reports

Financial services using Azure OpenAI Service to summarize financial reporting for peer risk analysis and customer conversation summarization



Code generation

Aircraft company using to convert natural language to SQL for aircraft telemetry data

Consulting service using Azure OpenAI Service to convert natural language to query proprietary data models



Reasoning over data

Financial services firm using Azure OpenAI Service to improve search capabilities and the conversational quality of a customer's Bot experience.

Insurance companies extract information from volumes of unstructured data to automate claim handling processes

International insurance company using Azure OpenAI Service to provide summaries of call center customer support conversation logs



Summarization

Global bank using Azure OpenAI Service to summarize financial reporting and analyst articles

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

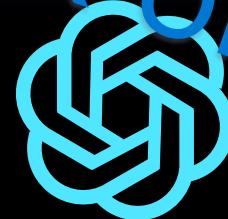
- Call center call transcripts
- Subject Matter Expert Documents
 - Competitive analysis
 - Peer Analysis
 - Technical reports
- Product and service feedback
- Social media trends

Conversational AI

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

Language Cognitive Services | Azure OpenAI Service

Work On this slide



Azure Cognitive Service for Language

- Start here
- Use language if your task is supported by Language Service capabilities
- Pre-built task specific multilingual AI capabilities
- The models are trained on a specific task
- You don't need to understand the capabilities of models
- You don't need data scientists to fine-tune the models.
- You don't have to define the schema; they are pre-built in

Azure OpenAI Service

- Use if your task requires more dynamic, complex capabilities not met by Language
- The models are not inherently trained on a specific task
- You need to determine whether the capabilities fit the business need
- We expose the models behind the API
- You need data science and developer capabilities for prompt engineering/fine tuning
- There is the flexibility to define the schema/data structure
- You need to design and implement a comprehensive set of Responsible AI controls

| Azure OpenAI | Benefits

Increased Efficiency and Productivity by getting more done in lesser time

Faster Time to Realise Value, by rapid experimentation and speeding up time to deployment

Enhanced Customer Experience with a greater focus on Customer-Centric services and products, whilst utilizing feedback and trends better

Easy to Use: even junior Data Scientists or business users would be able to use Azure OpenAI

Ensure data privacy, security, and implement it in a Responsible manner

Rapid prototyping and very quick time to market*

Does not require a long tedious annotation process*

Less training data is required for many relevant use cases

* in most relevant use cases observed