Microsoft

Azure OpenAI

Laure Van Isterdael
Solution Specialist Data & AI

Kelly Vehent
Cloud Solution Architect Data & AI

Intro to Azure OpenAI

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

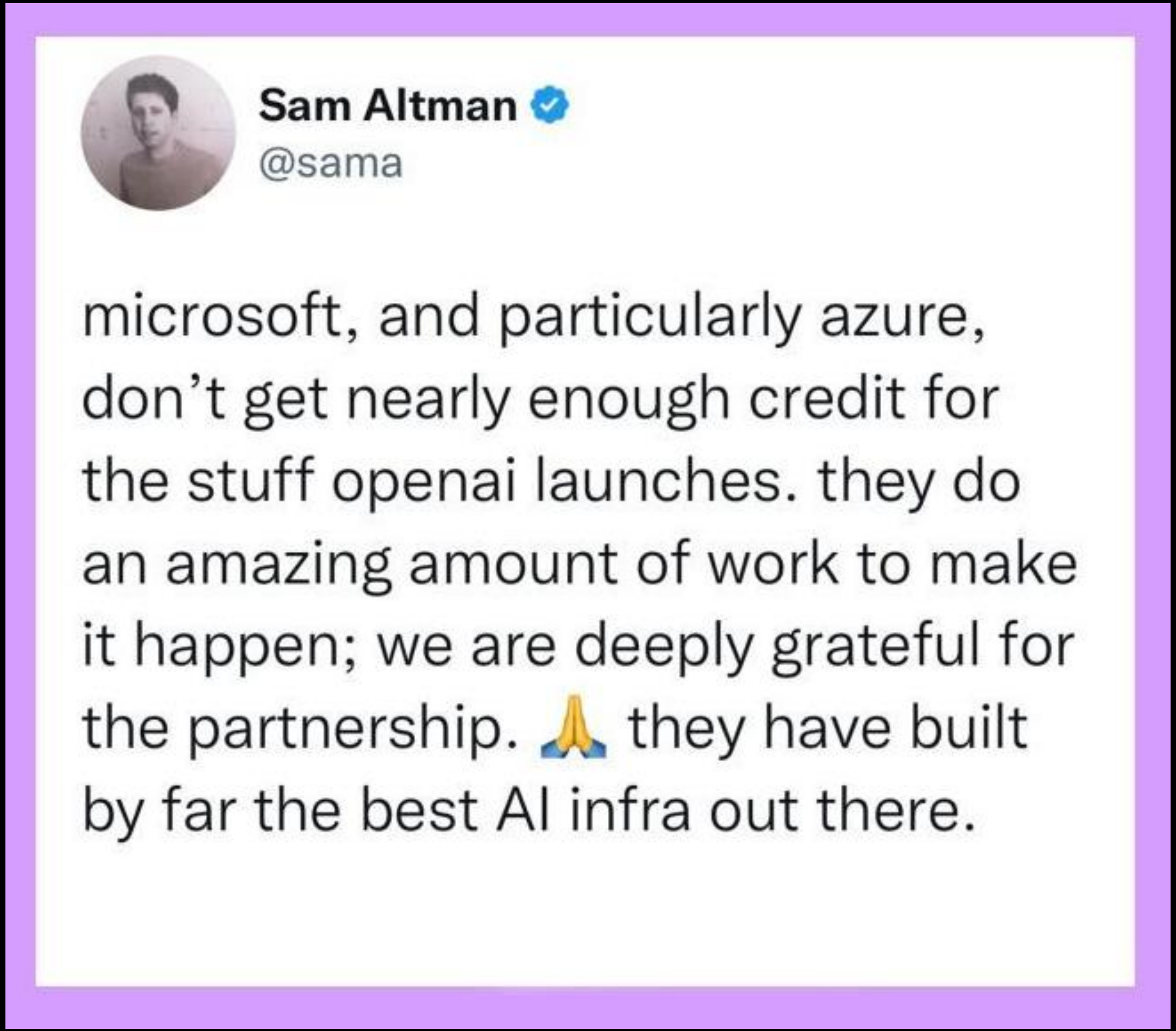
---

Building AI computing platform

Advancing AI Research

Making AI more accessible

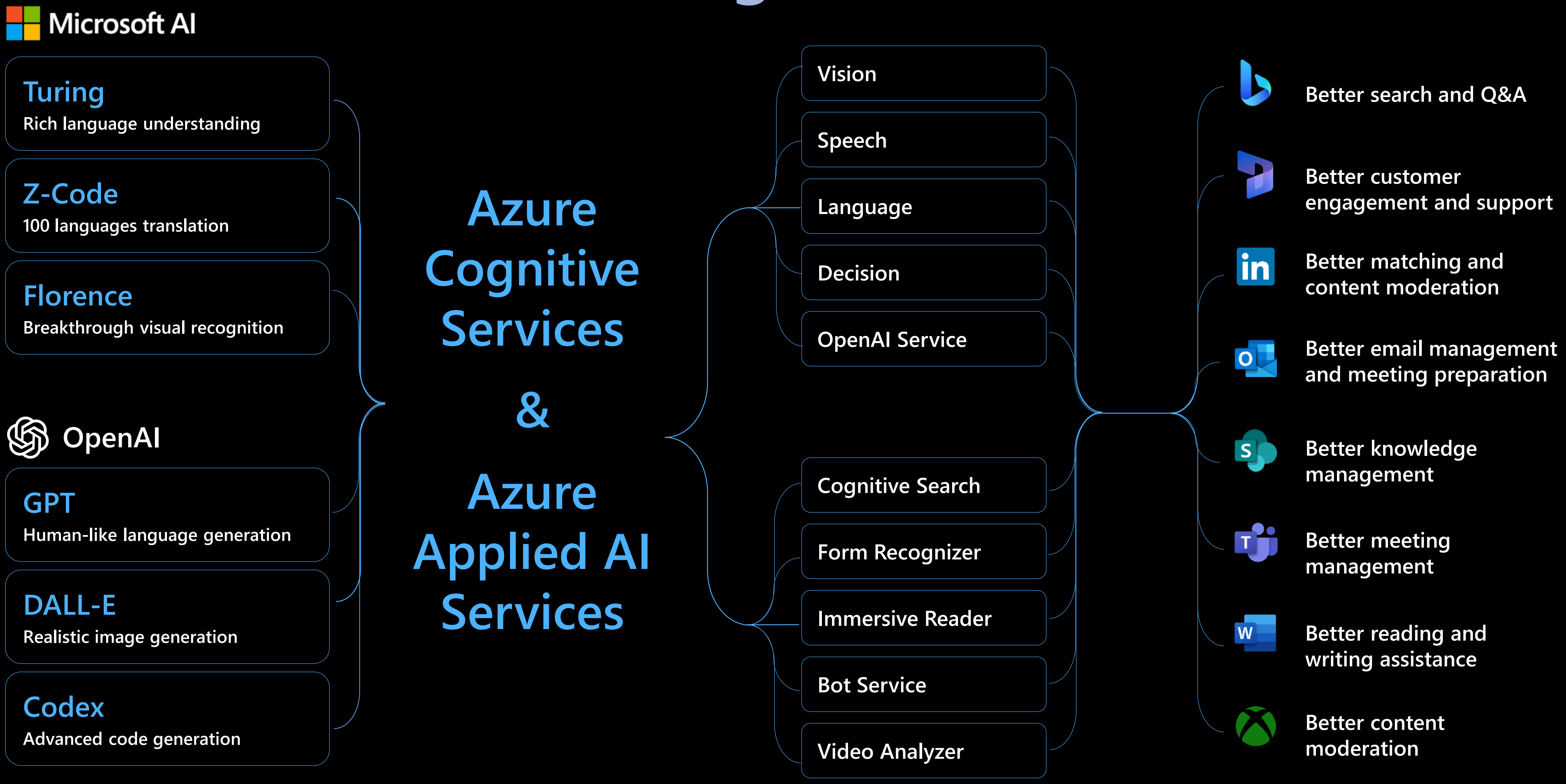
Building AI Computing Platform



Sam Altman
@sama

microsoft, and particularly azure, don’t get nearly enough credit for the stuff openai launches. they do an amazing amount of work to make it happen; we are deeply grateful for the partnership. 🙏 they have built by far the best AI infra out there.

Advancing AI Research



Microsoft AI

Turing
Rich language understanding

Z-Code
100 languages translation

Florence
Breakthrough visual recognition

OpenAI

GPT
Human-like language generation

DALL-E
Realistic image generation

Codex
Advanced code generation

Azure Cognitive Services & Azure Applied AI Services

- Vision
- Speech
- Language
- Decision
- OpenAI Service

- Cognitive Search
- Form Recognizer
- Immersive Reader
- Bot Service
- Video Analyzer

- Better search and Q&A
- Better customer engagement and support
- Better matching and content moderation
- Better email management and meeting preparation
- Better knowledge management
- Better meeting management
- Better reading and writing assistance
- Better content moderation

Making AI more accessible

Starting with…

Microsoft 365 Copilot
Viva Copilot
Bing
Microsoft Security Copilot
Dynamics Copilot
GitHub Copilot
Windows Copilot

And so much more!

Microsoft 365 Copilot

Embedded across Microsoft 365 apps

Copilot in Word

Write, edit, summarize, and create

Copilot in Excel

Analyze and explore your data

Copilot in PowerPoint

Turn ideas into stunning presentations

Copilot in Outlook

Clean up your inbox and messages

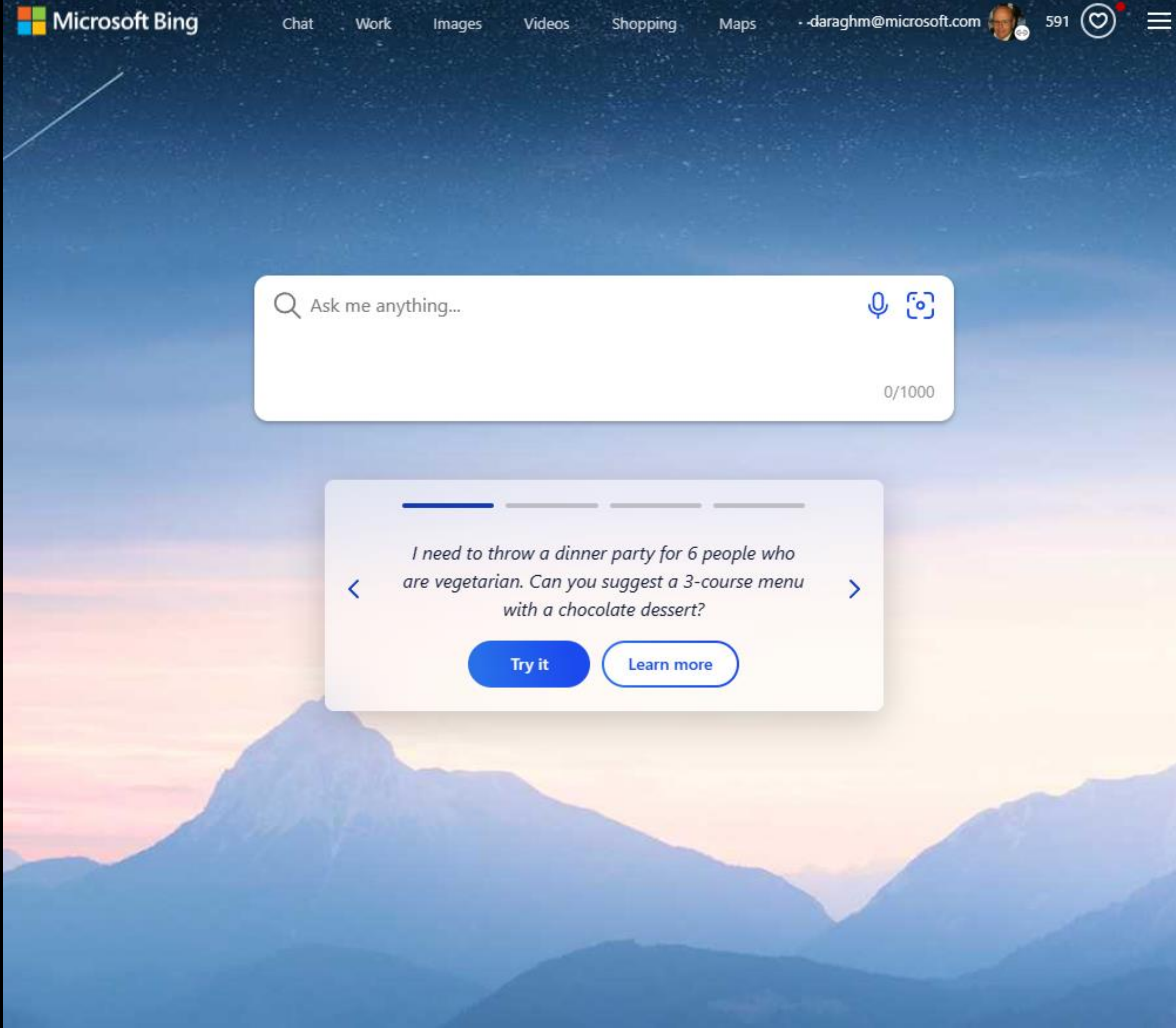
Copilot in Teams

Run more effective meetings

And so much more

Microsoft Bing

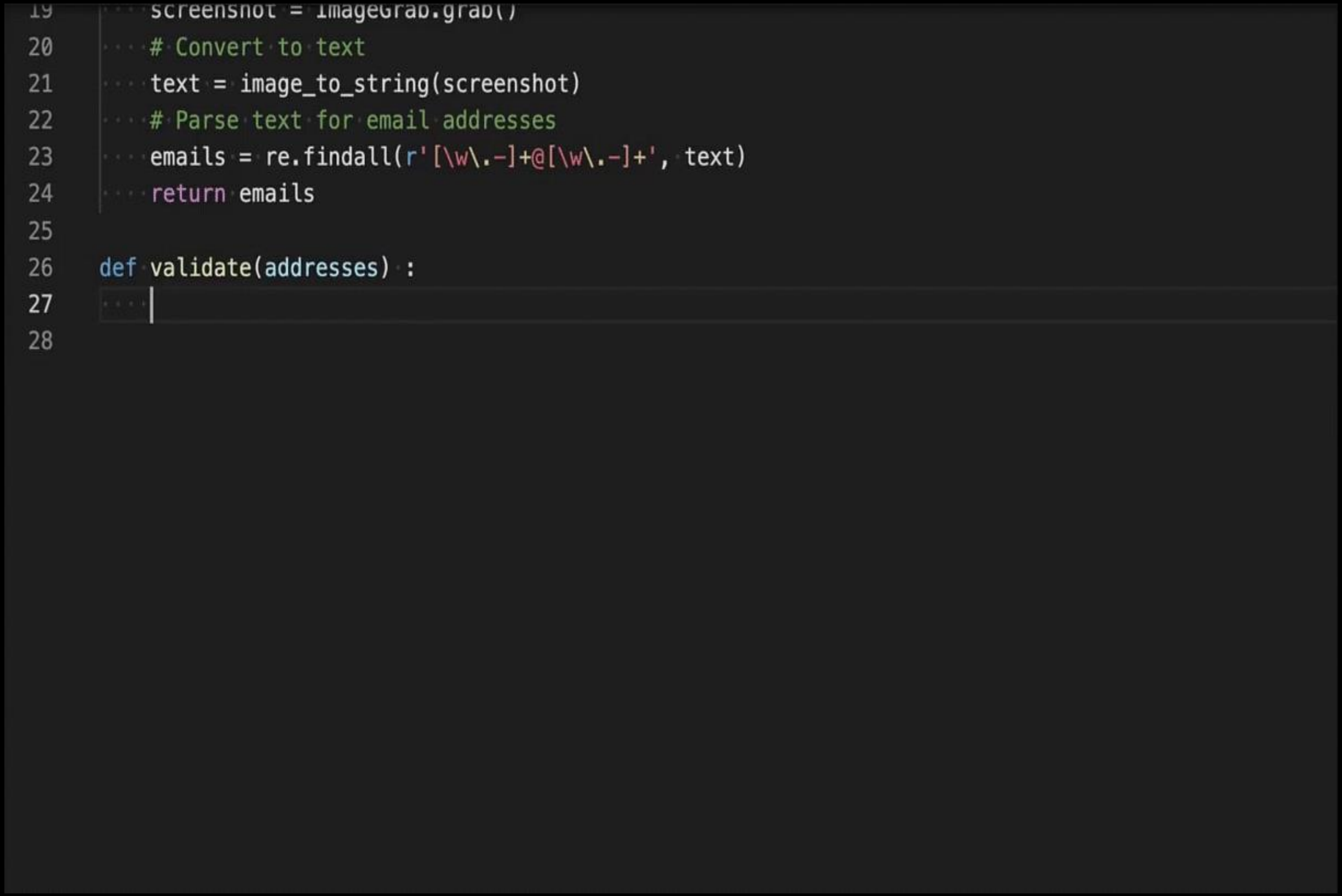
Your copilot for the web



I need to throw a dinner party for 6 people who are vegetarian. Can you suggest a 3-course menu with a chocolate dessert?

GitHub Copilot

Your AI pair programmer



```
    screenshot = ImageGrab.grab()
    # Convert to text
    text = image_to_string(screenshot)
    # Parse text for email addresses
    emails = re.findall(r'[\w\.-]+@[\w\.-]+', text)
    return emails

def validate(addresses) :
    
```

OpenAI Microsoft

Generative AI

GPT-3

Prompt:

Summarize this personal finance article in two sentences for my tax-filing customers:

...

Response:

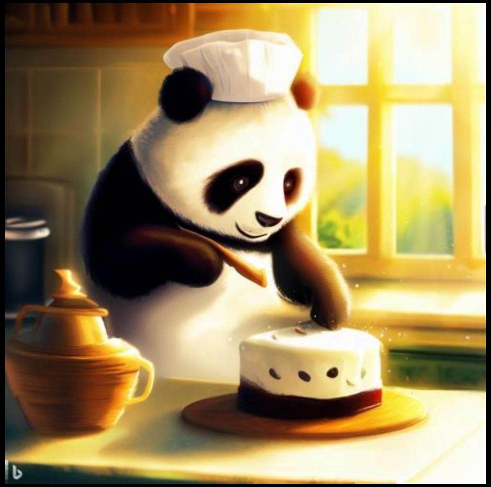
Tax pitfalls can be costly, so it's important to be aware of them. Some common tax pitfalls include failing to file or pay taxes on time, not reporting all income, claiming too many deductions, and not keeping accurate records.

DALL·E

Prompt:

a panda bear baking a cake in a sunny kitchen, digital art

Response:



ChatGPT, GPT-4

Prompt:

I'm having trouble getting my Xbox to turn on.

Response:

There are a few things you can try to troubleshoot this issue, First, …. Next, check the power cord …

Prompt:

Thanks! That worked. What games do you recommend for my 14-year-old?

Response:

Here are a few games that you might consider:

(bulleted list) …

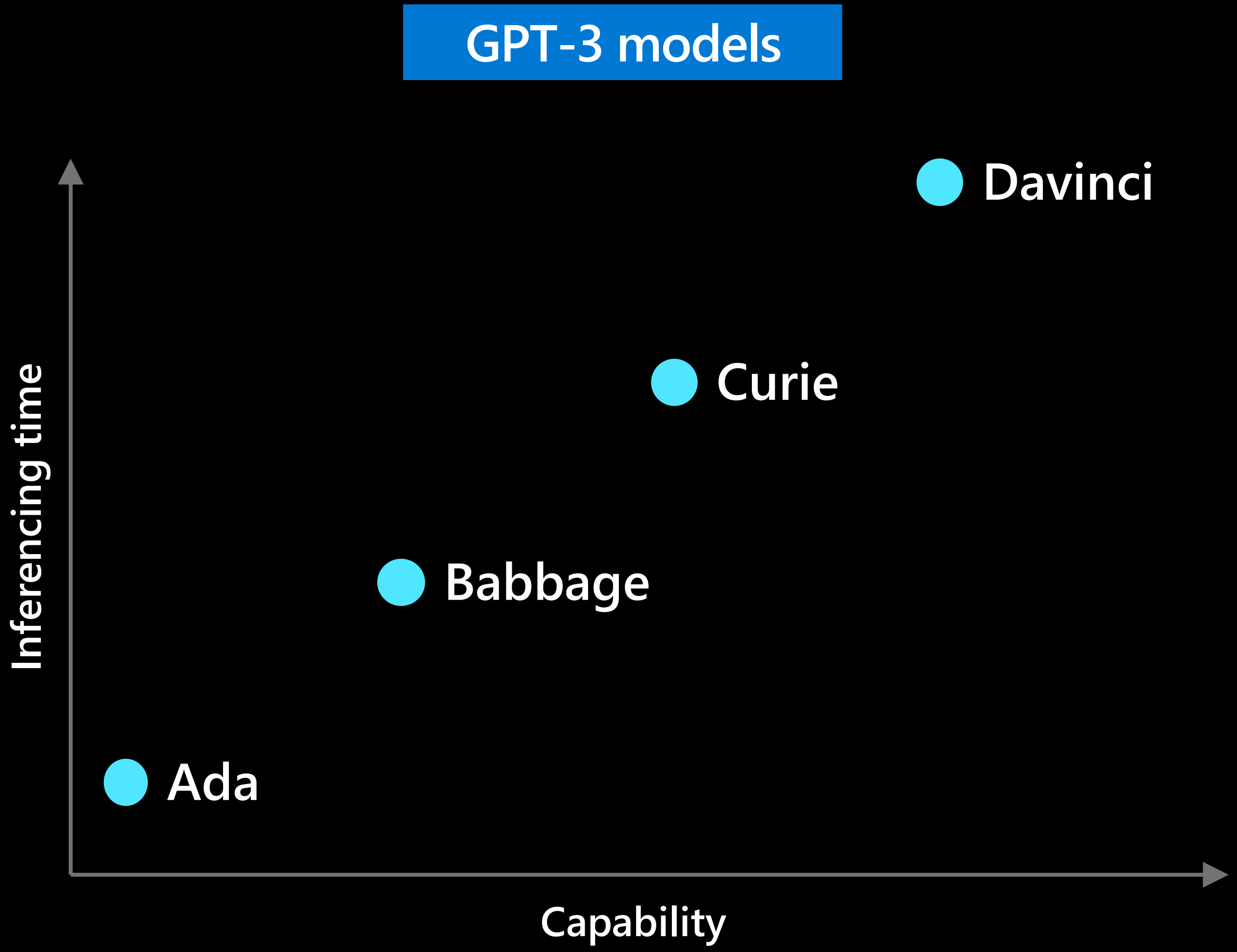
Microsoft Cloud

Runs on trust

Your data is your data	Data is stored encrypted in your Azure subscription
Your data from any fine-tuning is not used to train the foundation AI models	Azure OpenAI Service provisioned in your Azure subscription Model fine tuning stays in your Azure subscription and never moves into the foundation AI models
Your data is protected by the most comprehensive enterprise compliance and security controls	Encrypted with Customer Managed Keys Private Virtual Networks, Role Based Access Control Soc2, ISO, HIPPA, CSA STAR Compliant

Overview of Azure OpenAI models

Azure OpenAI | Family of Models



Azure OpenAI | GPT-3 Family of Models

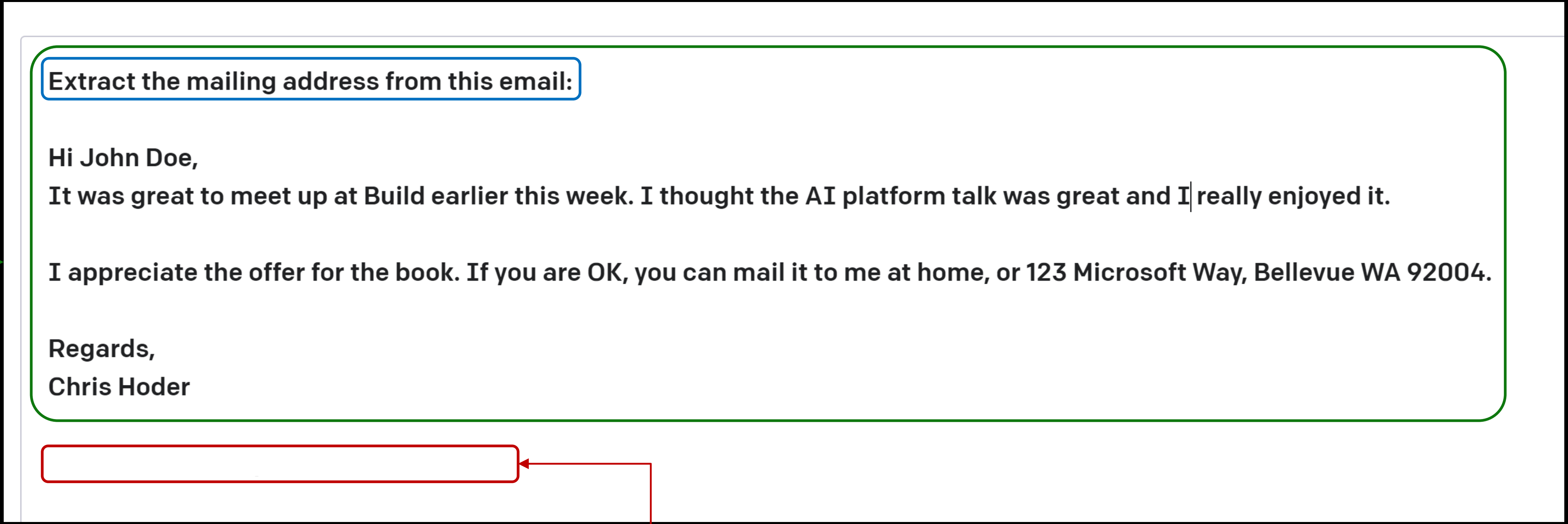
Model	Max # Tokens per Request	Description, performance, cost	Use cases
Davinci	4,096 tokens	**Most capable** GPT-3 model. Can do any task the other models can do, often with *higher quality*, *longer output* and *better instruction-following*.	Complex intent, cause and effect, summarization for audience
Curie	2048 tokens	**Very capable**, but *faster* and *lower cost* than Davinci.	Language translation, complex classification, text sentiment, summarization
Babbage	2048 tokens	**Capable** of straightforward tasks, *very fast*, and *lower cost*.	Moderate classification, semantic search classification
Ada	2048 tokens	**Capable** of very simple tasks, usually the *fastest* model in the GPT-3 series, and lowest cost.	Parsing text, simple classification, address correction, keywords

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.

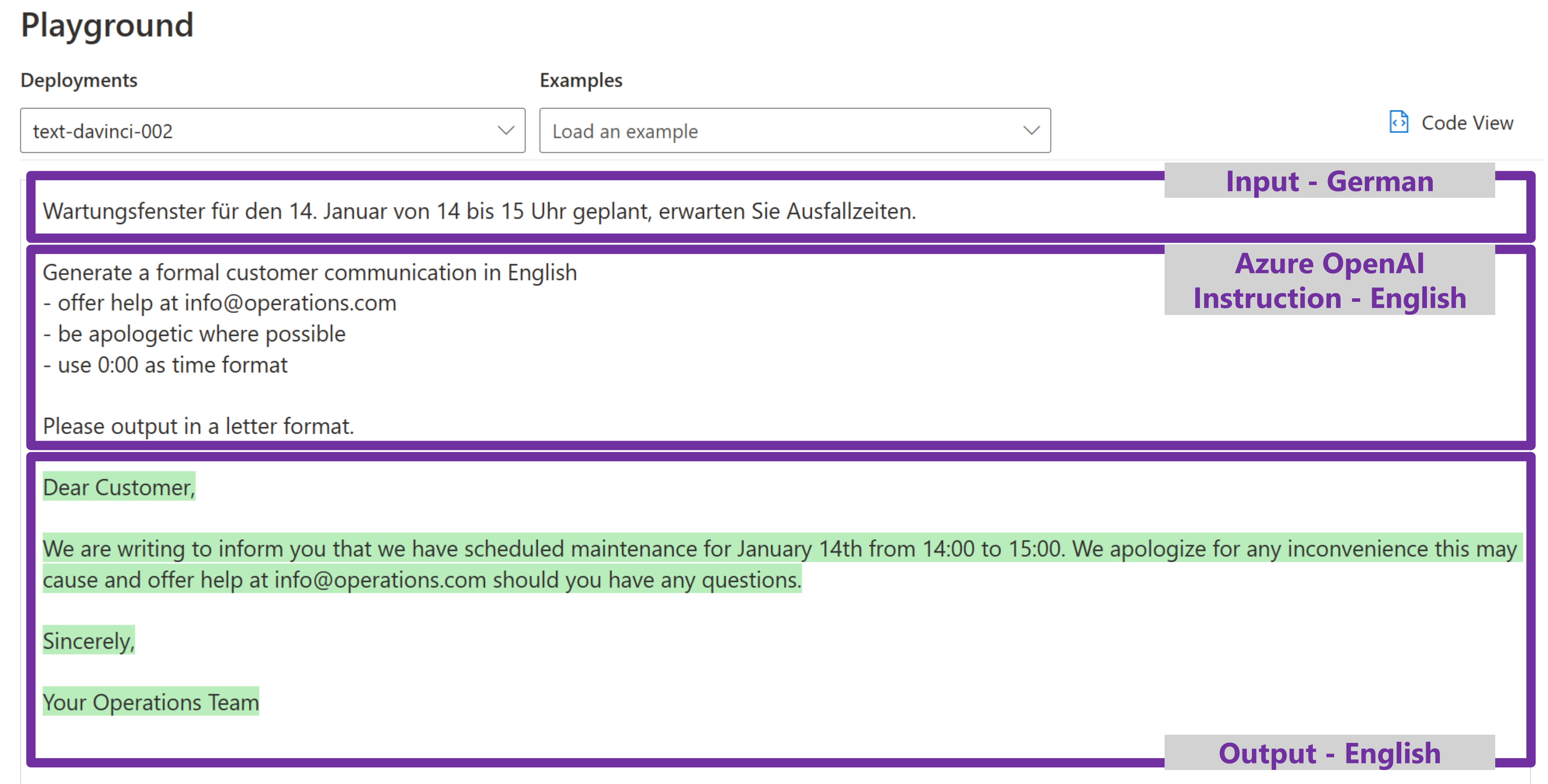
I appreciate the offer for the book. If you are OK, you can mail it to me at home, or 123 Microsoft Way, Bellevue WA 92004.

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 | Sample Use Cases



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

Generate a formal customer communication in English
- offer help at info@operations.com
- be apologetic where possible
- use 0:00 as time format

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

Example: Generate auto text in any language

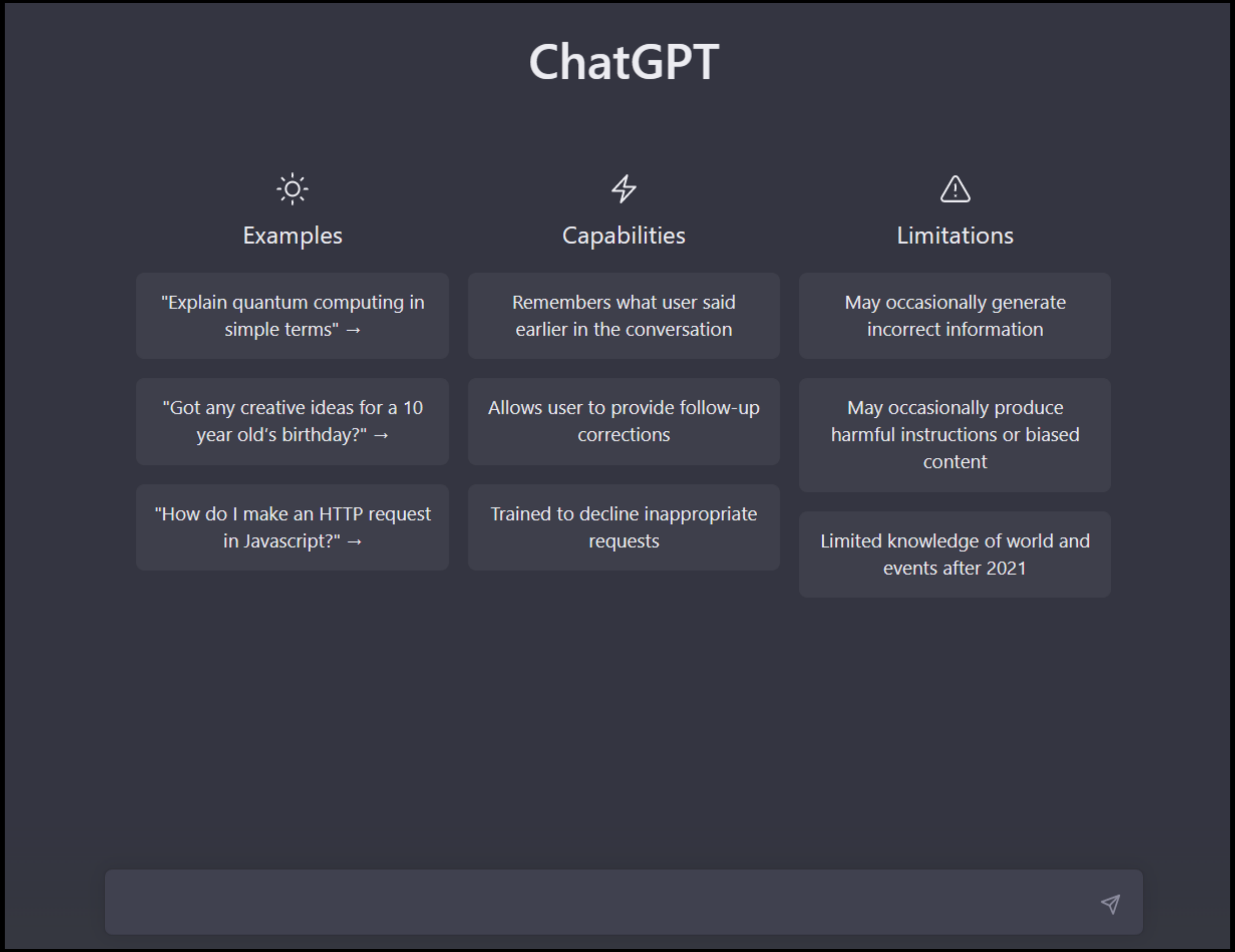
Demo

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.



ChatGPT

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

GPT-3 vs ChatGPT

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)

```
[
{"role":"system",
"content":"Assistant is an AI Chatbot designed to answer questions from the context provided below. \n\nContext: \nA 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."},

{"role":"user",
"content":"What is a neutron star?"}
]
```

Understanding the ChatGPT prompt format

The system message

The system message is included at the beginning of the prompt using the "system" role.

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 include the role (user or assistant) and content (user question/prompt + assistant reply)

For any follow-up messages, previous responses from the assistant should be included with the "assistant" role.

Example prompt

```
[
{"role":"system",
"content":"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."},

{"role":"user",
"content":"Why won't my Xbox turn on?"},

{"role":"assistant",
"content":"There could be several reasons why your Xbox is not turning on…"}
]
```

Azure OpenAI GPT-4 Overview

GPT-4 is a large multimodal model (accepting text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic benchmarks.

GPT-4 is more reliable, creative, and able to handle much more nuanced instructions than GPT-3.5.

Can do everything that GPT 3.5 can do

+ Visual Question and Answering (VQA)
+ Steerability
+ context length of 8,192 tokens

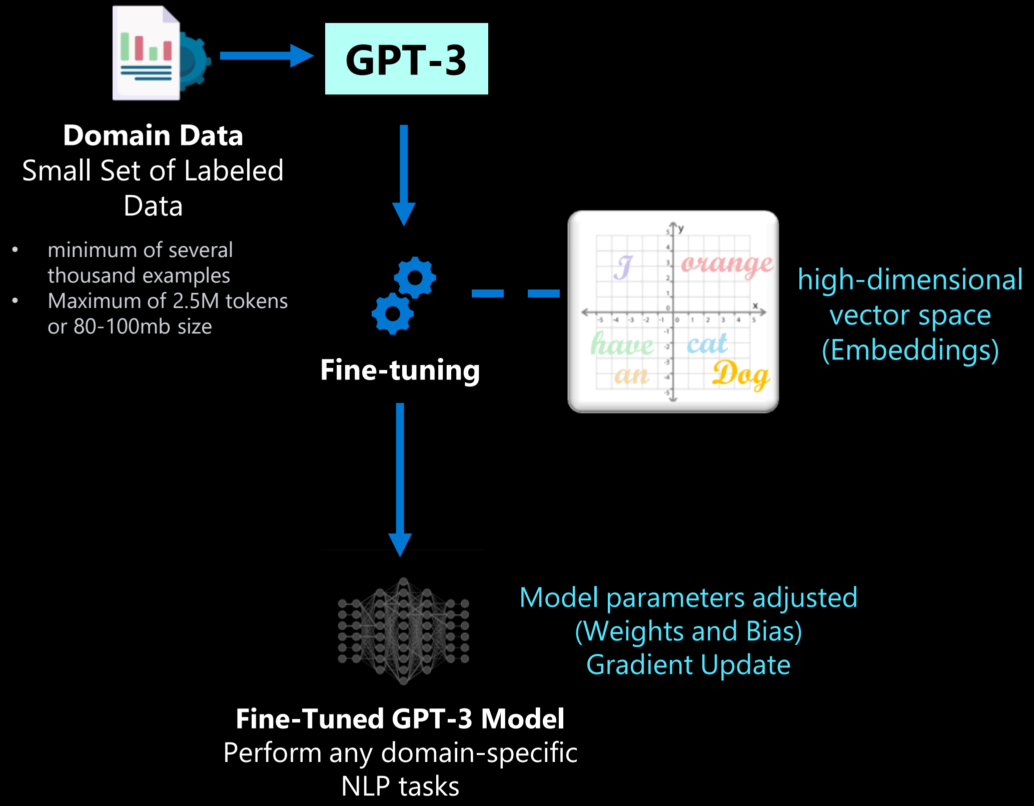
Apply here for GPT-4 access: [applying for access to GPT-4 today]()

Demo

Model Fine-Tuning

Model Adaptation with specifically styled data

Fine-Tuning



Domain Data
Small Set of Labeled Data

- minimum of several thousand examples
- Maximum of 2.5M tokens or 80-100mb size

Fine-tuning

high-dimensional vector space (Embeddings)

Model parameters adjusted
(Weights and Bias)
Gradient Update

Fine-Tuned GPT-3 Model
Perform any domain-specific NLP tasks

Fine-tuning results is a new model being generated with updated weights and biases.

This is in contrast to few-shot learning in which model weights and biases are not updated.

Best practices of Fine-Tuning

Fine-tuning data set must be in JSON format

A set of training examples that each consist of a single input ("prompt") and its associated output ("completion")

For classification task, the prompt is the problem statement, completion is the target class

For text generation task, the prompt is the instruction/question/request, and completion is the text ground truth

Best practices of Fine-Tuning

Fine-tuning data size: Advanced model (Davinci) performs better with limited amount of data; with enough data, all models do well.

Fine-tuning performs better with more high-quality examples.

To fine-tune a model that performs better than using a high-quality prompt with base models, you should provide at least a few hundred high-quality examples, ideally vetted by human experts.

From there, performance tends to linearly increase with every doubling of the number of examples. Increasing the number of examples is usually the best and most reliable way of improving accuracy.

Embeddings

Embeddings

An embedding is a special format of data representation that can be easily utilized by machine learning models and algorithms.

The embedding is an information dense representation of the semantic meaning of a piece of text.

Each embedding is a vector of floating-point numbers, such that the distance between two embeddings in the vector space is correlated with semantic similarity between two inputs in the original format.

For example, if two texts are similar, then their vector representations should also be similar.

Embeddings make it possible to map content to a “semantic space”

A neutron star is the collapsed core of a massive supergiant star

[15 34 24 13 …]

A star shines for most of its active life due to thermonuclear fusion.

[16 22 89 26 …]

The presence of a black hole can be inferred through its interaction with other matter

[20 13 31 89 …]

Embedding models

Different Azure OpenAI embedding models are specifically created to be good at a particular task.

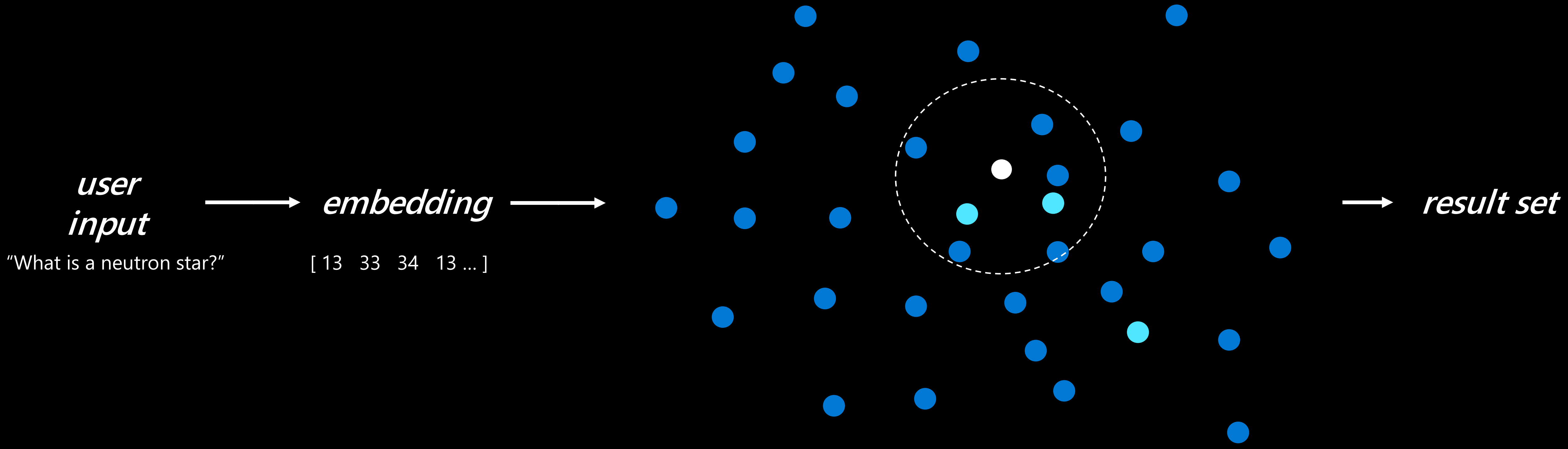
- **Similarity embeddings** are good at capturing semantic similarity between two or more pieces of text.
- **Text search embeddings** help measure long documents are relevant to a short query.
- **Code search embeddings** are useful for embedding code snippets and embedding nature language search queries.

Embeddings make it easier to do machine learning on large inputs representing words by capturing the semantic similarities in a vector space.

Therefore, we can use embeddings to determine if two text chunks are semantically related or similar, and provide a score to assess similarity.

Similarity Search with embeddings

Once you encode your content as embeddings, you can then get an embedding from the user input and use that to find the most semantically similar content.



user input

"What is a neutron star?"

embedding

[13 33 34 13 …]

result set

[Azure OpenAI embeddings tutorial - Azure OpenAI | Microsoft Learn]

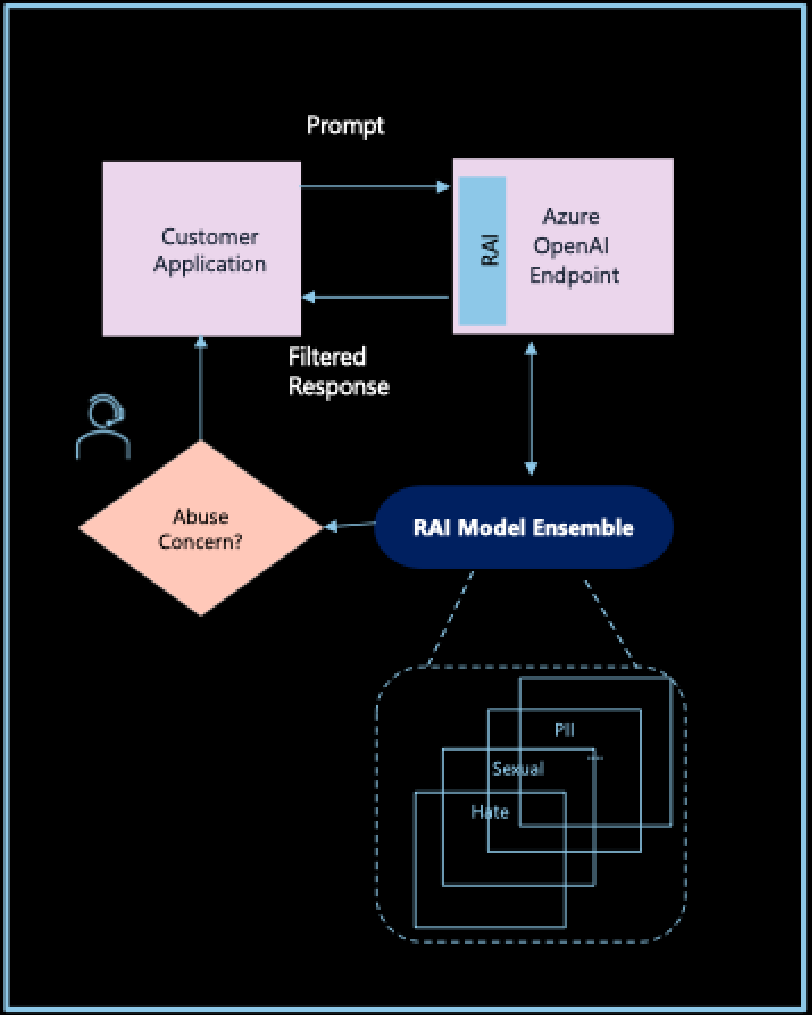
Responsible AI

Azure OpenAI Service Responsible AI

Content filters—
prevent harmful or inappropriate content

Abuse detection—
ensure responsible use of your application

Implementation guidelines, patterns,
and best practices



Prompt

Customer Application

RAI

Azure OpenAI Endpoint

Filtered Response

Abuse Concern?

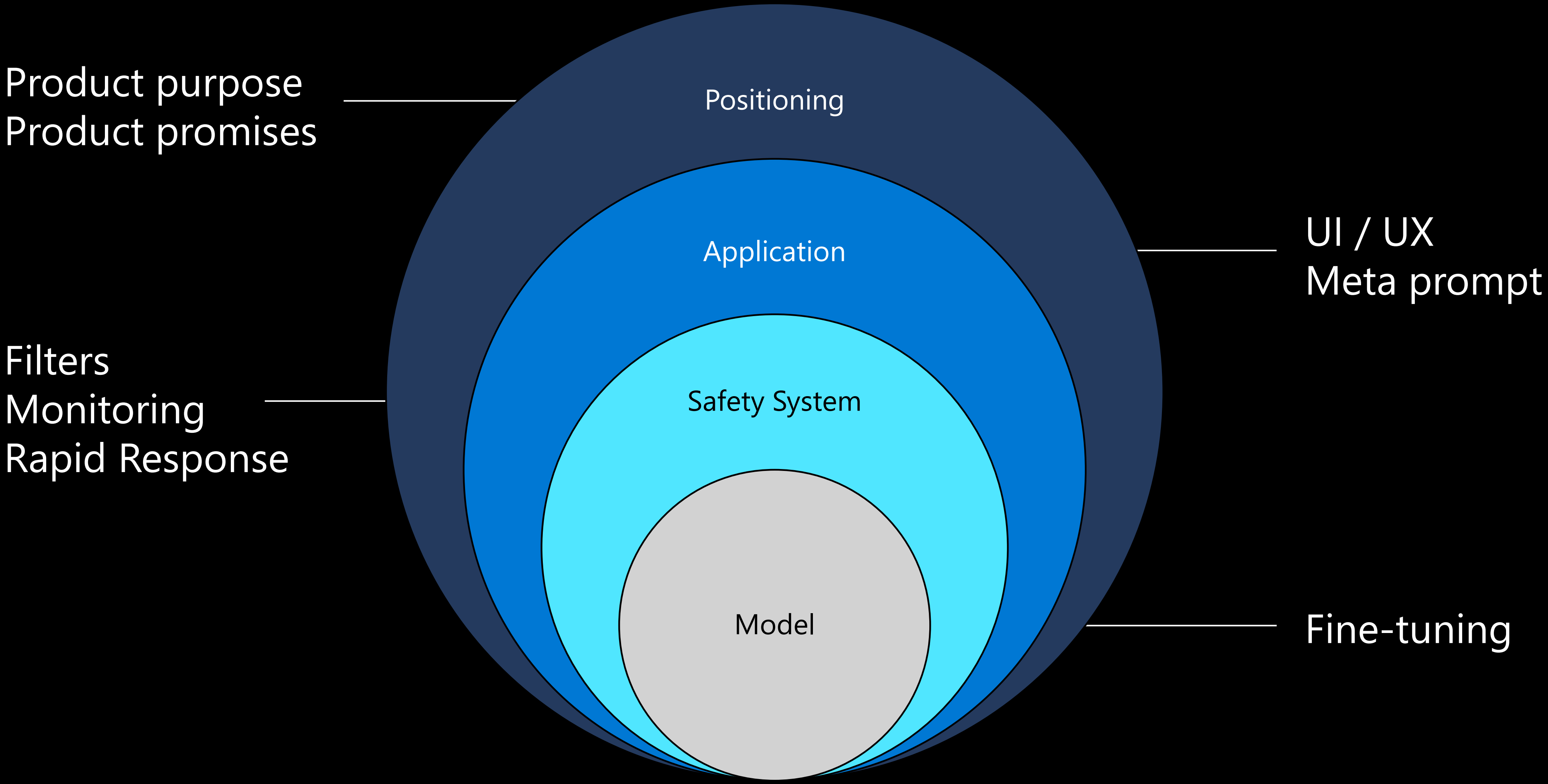
RAI Model Ensemble

PII

Sexual

Hate

Mitigation Layers



Responsible AI Mitigations

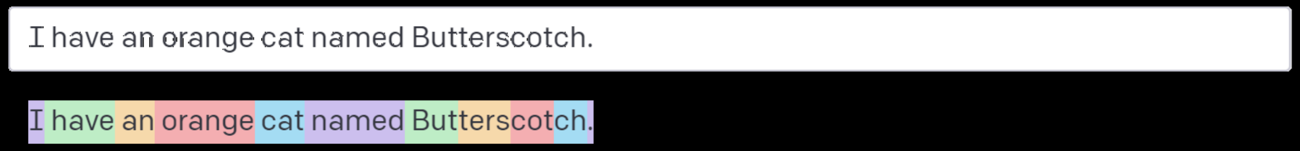
Customer	Structure user interactions. Limit the length, structure, and source of inputs and outputs Control user access Transparency and overreliance mitigations in UI/UX
Technical	Content Filtering Asynchronous abuse detection User-based throttling User-based shutdown
Process and Policy	Limited Access Abuse reporting channel Feedback channel Incident Response
Documentation and legal	Terms of use Transparency Note Design Guidelines

Prompt Engineering

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.

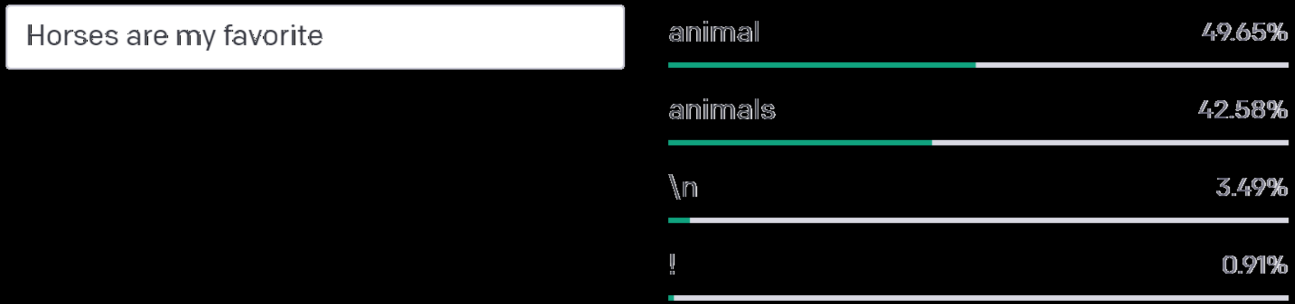
As a point of reference, the collected works of Shakespeare are about 900,000 words or 1.2M tokens.



I have an orange cat named Butterscotch.

I have an orange cat named Butterscotch.

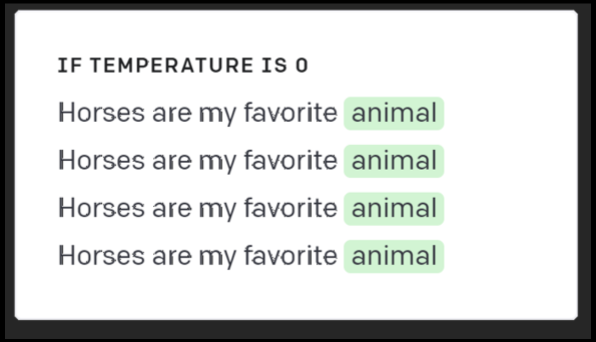
Temperature & probabilities



Horses are my favorite

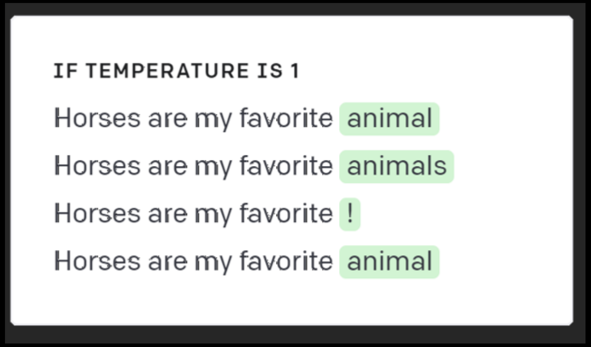
animal	49.65%
animals	42.58%
\n	3.49%
!	0.91%

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

Text classification

Custom Named Entities (Few-Shot)

Prompt:
Extract job titles from the following sentences.

Sentence: John Doe has been working for Microsoft for 20 years as a Linux Engineer.
Job title: Linux Engineer
###
Sentence: John Doe has been working for Microsoft for 20 years and he loved it.
Job title: none
###
Sentence: Marc Simoncini | Director | Meetic
Job title: Director
###
Sentence: Damien is the CTO of Platform.sh, he was previously the CTO of Commerce Guys, a leading ecommerce provider.

Completion:
Job title: CTO

Text classification

Multi-Label Text Classification (Few-Shot)

Prompt:
The following is a list of movies and the categories they fall into:

Topgun: As students at the United States Navy's elite fighter weapons school compete to be best in the class, one daring young pilot learns a few things from a civilian instructor that are not taught in the classroom.
Category: Drama, Action

Hustle: A basketball scout discovers a phenomenal street ball player while in Spain and sees the prospect as his opportunity to get back into the NBA.
Category: Comedy, Drama, Sport

The Northman: From visionary director Robert Eggers comes The Northman, an action-filled epic that follows a young Viking prince on his quest to avenge his father's murder.
Category:

Completion :
Action, Adventure, Drama

Reason over unstructured text

Question and answering

Prompt:
On Monday, Microsoft launched an OpenAI service as part of its Azure cloud platform, offering businesses and start-ups the ability to incorporate models like ChatGPT into their own systems. The company has already been building AI tools into many of its consumer products, such as a DALL-E 2 feature in its Bing search engine that can create images based on a text prompt, and the Information reported recently that it’s working to bring more of them to Microsoft Office as well.

What will Microsoft incorporate in its own systems?

Completion:
Microsoft plans to incorporate AI tools into its own systems, such as a DALL-E 2 feature in its Bing search engine that can create images based on a text prompt, and the Information reported recently that it’s working to bring more of them to Microsoft Office as well

Summarization

Extractive vs Abstractive

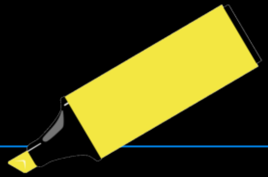
Extractive

Select parts **of the original text to form a summary**

- **Easier**
- **Restrictive (no paraphrasing)**

Peter and Elizabeth took a taxi to attend the night party in the city. While in the party, Elizabeth collapsed and was rushed to the hospital.

Peter and Elizabeth attend party city. Elizabeth rushed hospital



Abstractive (Semantic)

Generates new text **using natural language generation techniques**

- **More difficult**
- **More flexible (more human)**

Peter and Elizabeth took a taxi to attend the night party in the city. While in the party, Elizabeth collapsed and was rushed to the hospital.

Elizabeth was hospitalized after attending a party with Peter.

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).

ChatGPT—Enterprise Knowledge Bot

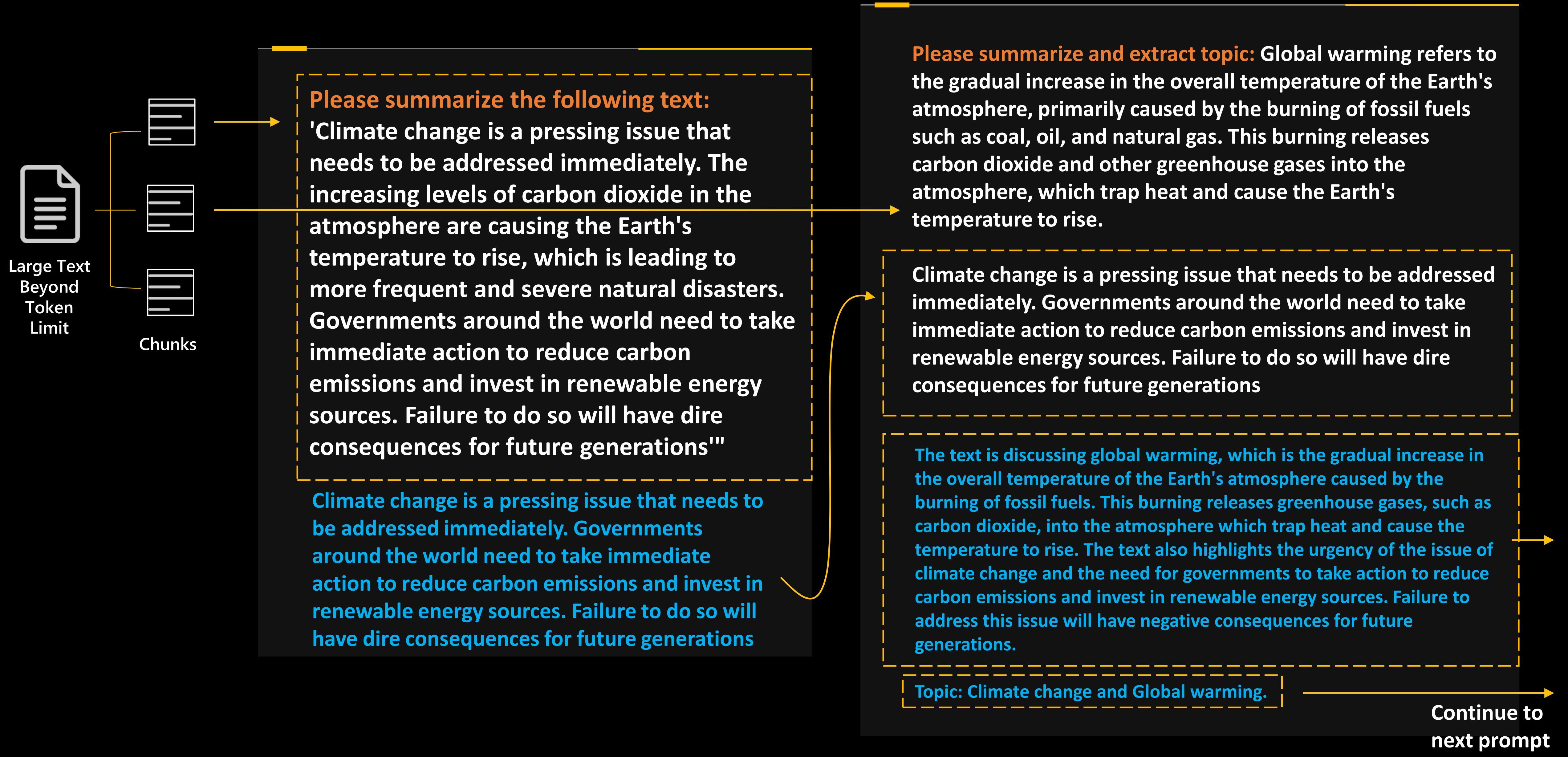
```
[
{"role":"system",
"content":"Assistant is an intelligent chatbot designed to help internal users answer questions from an enterprise knowledge base. Relevant information for you will be included in the context section below. Only answer questions using the context below and if you're not sure of 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"},

{"role":"user",
"content":"What is the Azure OpenAI Service?"}
]
```

Prompt Engineering Guide

1. Give clearer instructions
2. Split complex tasks into simpler subtasks
3. Structure the instruction to keep the model on task
4. Prompt the model to explain before answering
5. Ask for justifications of many possible answers, and then synthesize
6. Generate many outputs, and then use the model to pick the best one
7. Fine-tune custom models to maximize performance

Prompt Chunking (token limit)

Then all summaries are passed to GPT again to summarize, i.e. second round of summarization. The sum of the summaries has to fit within the token limit.



Please summarize the following text: 'Climate change is a pressing issue that needs to be addressed immediately. The increasing levels of carbon dioxide in the atmosphere are causing the Earth's temperature to rise, which is leading to more frequent and severe natural disasters. Governments around the world need to take immediate action to reduce carbon emissions and invest in renewable energy sources. Failure to do so will have dire consequences for future generations'"

Climate change is a pressing issue that needs to be addressed immediately. Governments around the world need to take immediate action to reduce carbon emissions and invest in renewable energy sources. Failure to do so will have dire consequences for future generations

Please summarize and extract topic: Global warming refers to the gradual increase in the overall temperature of the Earth's atmosphere, primarily caused by the burning of fossil fuels such as coal, oil, and natural gas. This burning releases carbon dioxide and other greenhouse gases into the atmosphere, which trap heat and cause the Earth's temperature to rise.

Climate change is a pressing issue that needs to be addressed immediately. Governments around the world need to take immediate action to reduce carbon emissions and invest in renewable energy sources. Failure to do so will have dire consequences for future generations

The text is discussing global warming, which is the gradual increase in the overall temperature of the Earth's atmosphere caused by the burning of fossil fuels. This burning releases greenhouse gases, such as carbon dioxide, into the atmosphere which trap heat and cause the temperature to rise. The text also highlights the urgency of the issue of climate change and the need for governments to take action to reduce carbon emissions and invest in renewable energy sources. Failure to address this issue will have negative consequences for future generations.

Topic: Climate change and Global warming.

Continue to next prompt

Few-Shot Reasoning (Human version)

1st prompt

Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Answer: The answer is 11.

The cafeteria has 23 apples. If they used 20 to make lunch and bought 6 more, how many do they have?

The answer is 27

✗

2nd prompt—provide reasoning

Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Answer: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5+6 = 11. The answer is 11.

The cafeteria has 23 apples. If they used 20 to make lunch and bought 6 more, how many do they have?

The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23-20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9.

✓

Model Reasoning—Chain-of-Thought

Before

What is the annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day?

The annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day is 8,000,000 liters per year.

After

What is the annual water demand of a single-family household containing four people who are at home an average of 200 days per year and use an average of 100 liters of water per day?

Let's think step by step and explain the calculation step by step.

Step 1: Calculate the total number of days in a year:
There are 365 days in a year.

Step 2: Calculate the total number of days the family is at home:
The family is at home an average of 200 days per year.

Step 3: Calculate the total water demand of the family:
The family uses an average of 100 liters of water per day. Therefore, the total water demand of the family is 100 liters x 200 days = 20,000 liters per year.

Step 4: Calculate the annual water demand of the single-family household:
The annual water demand of the single-family household is 20,000 liters per year.

Few-Shot—Chain-of-Thought—Examples

Math Word Problems (free response)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11.

The answer is 11.

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3, which is less than water. Thus, a pear would float

So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn’t a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500? Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. 9 +90(2) + 401(3) = 1392.

The answer is (b).

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943.

So the answer is 05/23/1943.

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya".

So the answer is ya

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go? Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do.

So the answer is (b).

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship.”

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer.

So the answer is no.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up.

So the answer is no.

Selection-inference prompting

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

14 times

✗

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

To solve "How many times can she slide before it closes" we need to first solve: **Chain-of-Thought**

How much time does it take for Amy to climb and slide?

Decomposes problem into a smaller task (Reasoning)

Add new fact to context

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

How much time does it take for Amy to climb and slide?

It takes Amy 5 minutes to climb and slide

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slides close in 15 minutes. How many times can she slide before it closes?

It takes Amy 5 minutes to climb and slide.

She can slide 3 times before the water slides close.

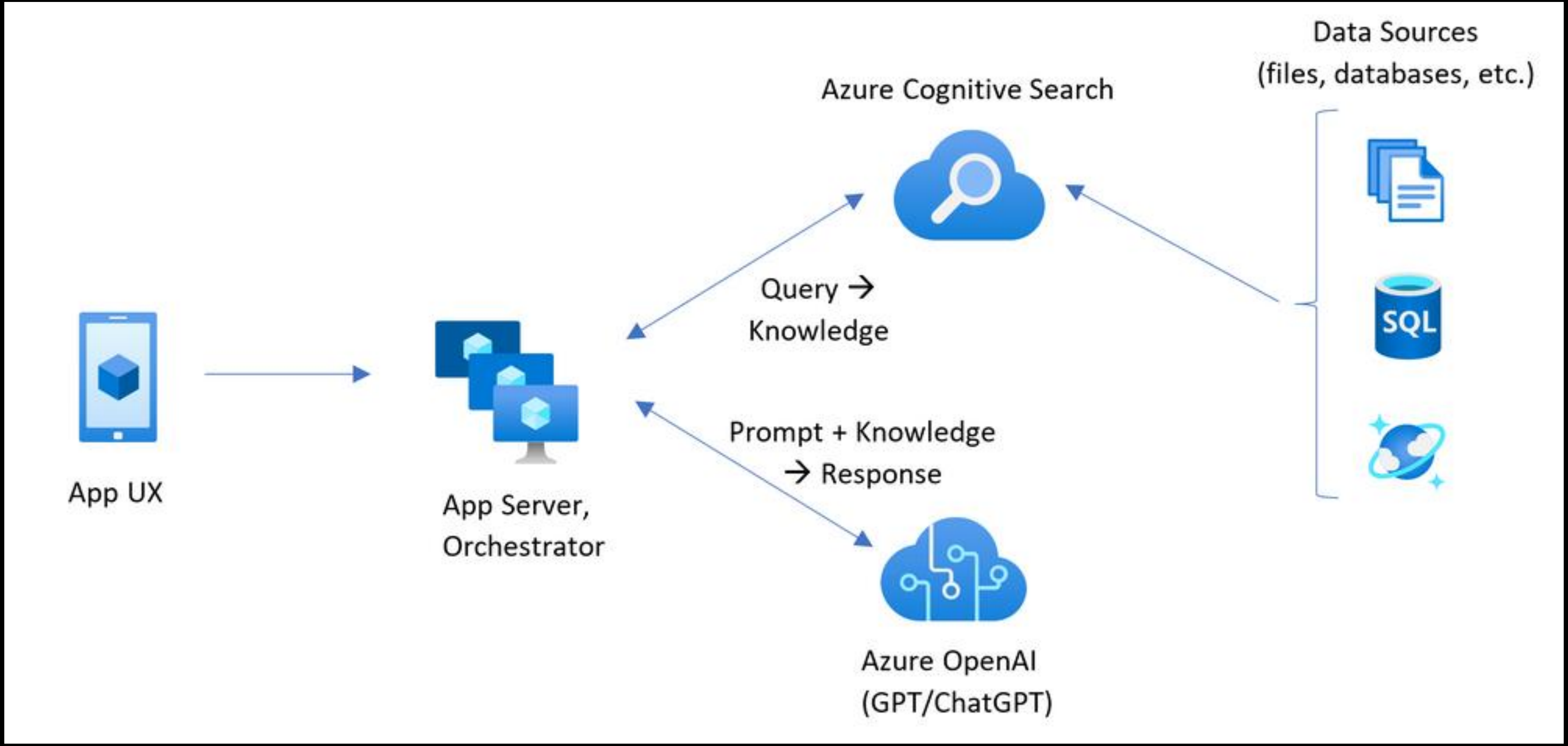
✓

Use Cases + Solution Accelerators

Industry examples and uses for GPT-3

Customer Service	GPT-3 is used to automate response to customer inquiries
Reporting	Generate reports with GPT-3, making it easier for companies to get the most up-to-date information
Knowledge & Information	GPT-3 helps organizations track knowledge and information, making it easier for employees to find and use information they need
Content Generation	GPT-3 automatically generates content on various topics
Sentiment Analysis	GPT-3 analyzes the sentiment of text data to understand customer sentiment and make better decisions
Natural Language Processing	GPT-3 processes and understand natural language data, making it easier to extract information from text data
Machine Translation	GPT-3 automatically translates text from one language to another
Predictive Modeling	GPT-3 builds predictive models that can be used to make better decisions about the future.
Analytics	GPT-3 can extract insights from data that facilitate informed decision making about their business
Security	GPT-3 can protect data and prevent unauthorized access

Search your own data



Try it here: https://github.com/Azure-Samples/azure-search-openai-demo

LangChain 101

What is LangChain?

- Framework to integrate LLMs into applications
  - Supports Python and JavaScript/Typescript
- Two main ideas: components & (use-case specific) chains
- Helps remove restraint of public only training data
- Helps remove restraint of time-limited training data (Sep 2021)
  - Interact with anything that has an available tool (e.g. internet search, SQL databases,…)

- Create a chain of various steps to be executed

LangChain Components

- Models (language, chat, embedding): connection to Azure OpenAI
- Prompts: template for your model input
- Indexes: use unstructured data (e.g. text doc) with LLMs
- Memory: save information during and across conversations
- Chains: sequence of components (incl. other chains)
- Agents: orchestrate access to tools based on user input

More about Agents

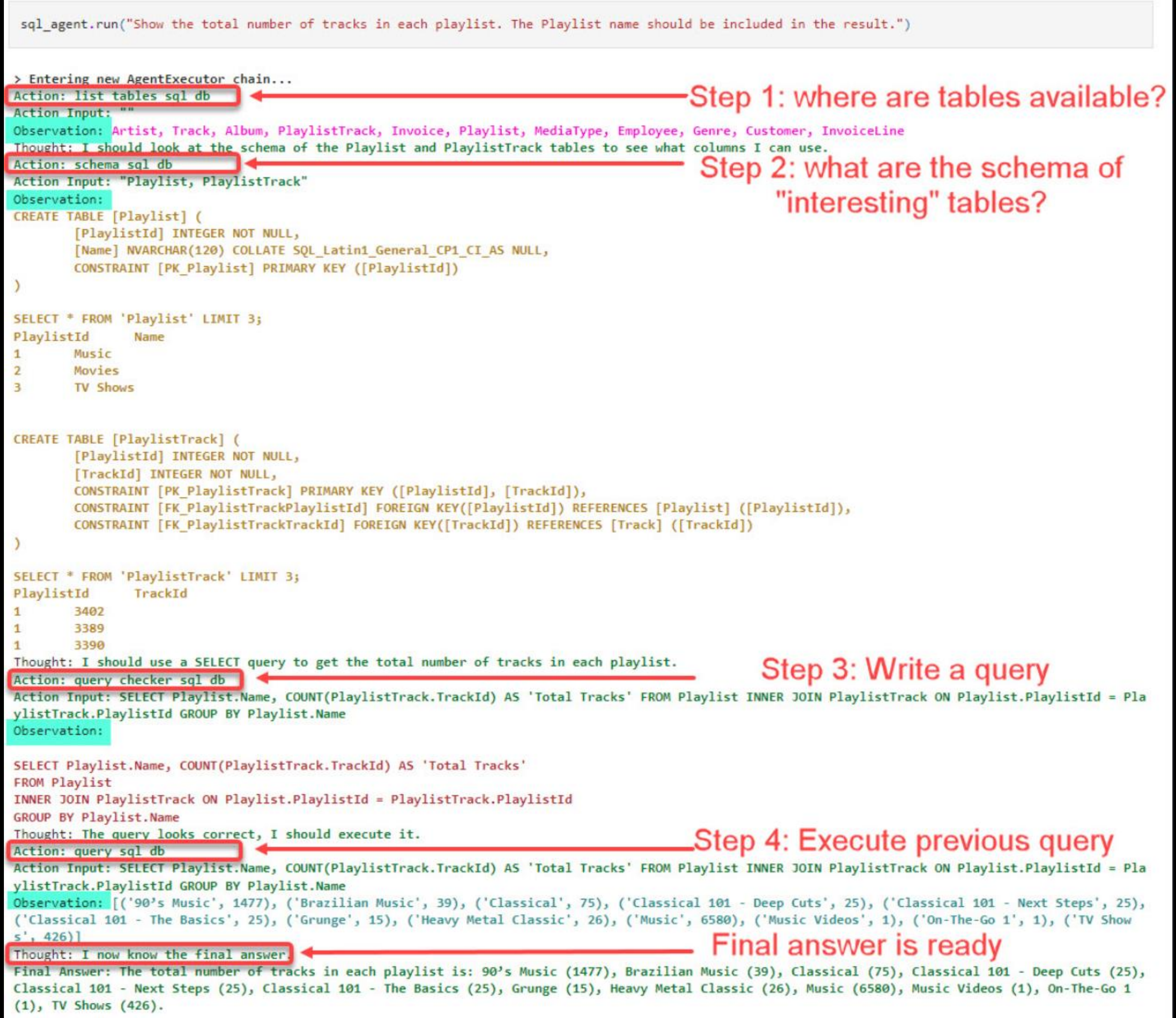
- Top-level Agent Executor combines an Agent with various Tools
- Uses a LLM to generate steps/decisions
  - Step-by-step reasoning as orchestrator
- Tools interact with various resources
- Group multiple tools into a Toolkit
- Tool examples: Search, file system, requests (web access),…
- 4 types of agents (+ custom) for different use case patterns

- The agent decides sequence and type of actions to be taken

LangChain + Azure OpenAI

- Watch out for differences between OpenAI and Azure OAI API
- Library of AzOAI samples using LangChain on GitHub: azure-openai-samples/fundamentals/langchain at main · Azure/azure-openai-samples (github.com)

- “Search your own data” solution accelerator also uses LangChain
  - Converting user query into keywords for index search



```
sql_agent.run("Show the total number of tracks in each playlist. The Playlist name should be included in the result.")
```

```
> Entering new AgentExecutor chain...
Action: list tables sql db
Action Input: ""
Observation: Artist, Track, Album, PlaylistTrack, Invoice, Playlist, MediaType, Employee, Genre, Customer, InvoiceLine
```

Step 1: where are tables available?

```
Thought: I should look at the schema of the Playlist and PlaylistTrack tables to see what columns I can use.
Action: schema sql db
Action Input: "Playlist, PlaylistTrack"
Observation:
CREATE TABLE [Playlist] (
        [PlaylistId] INTEGER NOT NULL,
        [Name] NVARCHAR(120) COLLATE SQL_Latin1_General_CP1_CI_AS NULL,
        CONSTRAINT [PK_Playlist] PRIMARY KEY ([PlaylistId])
)

SELECT * FROM 'Playlist' LIMIT 3;
PlaylistId      Name
1       Music
2       Movies
3       TV Shows


CREATE TABLE [PlaylistTrack] (
        [PlaylistId] INTEGER NOT NULL,
        [TrackId] INTEGER NOT NULL,
        CONSTRAINT [PK_PlaylistTrack] PRIMARY KEY ([PlaylistId], [TrackId]),
        CONSTRAINT [FK_PlaylistTrackPlaylistId] FOREIGN KEY([PlaylistId]) REFERENCES [Playlist] ([PlaylistId]),
        CONSTRAINT [FK_PlaylistTrackTrackId] FOREIGN KEY([TrackId]) REFERENCES [Track] ([TrackId])
)

SELECT * FROM 'PlaylistTrack' LIMIT 3;
PlaylistId      TrackId
1       3402
1       3389
1       3390
```

Step 2: what are the schema of "interesting" tables?

```
Thought: I should use a SELECT query to get the total number of tracks in each playlist.
Action: query checker sql db
Action Input: SELECT Playlist.Name, COUNT(PlaylistTrack.TrackId) AS 'Total Tracks' FROM Playlist INNER JOIN PlaylistTrack ON Playlist.PlaylistId = Pla
ylistTrack.PlaylistId GROUP BY Playlist.Name
Observation:

SELECT Playlist.Name, COUNT(PlaylistTrack.TrackId) AS 'Total Tracks'
FROM Playlist
INNER JOIN PlaylistTrack ON Playlist.PlaylistId = PlaylistTrack.PlaylistId
GROUP BY Playlist.Name
```

Step 3: Write a query

```
Thought: The query looks correct, I should execute it.
Action: query sql db
Action Input: SELECT Playlist.Name, COUNT(PlaylistTrack.TrackId) AS 'Total Tracks' FROM Playlist INNER JOIN PlaylistTrack ON Playlist.PlaylistId = Pla
ylistTrack.PlaylistId GROUP BY Playlist.Name
Observation: [('90’s Music', 1477), ('Brazilian Music', 39), ('Classical', 75), ('Classical 101 - Deep Cuts', 25), ('Classical 101 - Next Steps', 25),
('Classical 101 - The Basics', 25), ('Grunge', 15), ('Heavy Metal Classic', 26), ('Music', 6580), ('Music Videos', 1), ('On-The-Go 1', 1), ('TV Show
s', 426)]
```

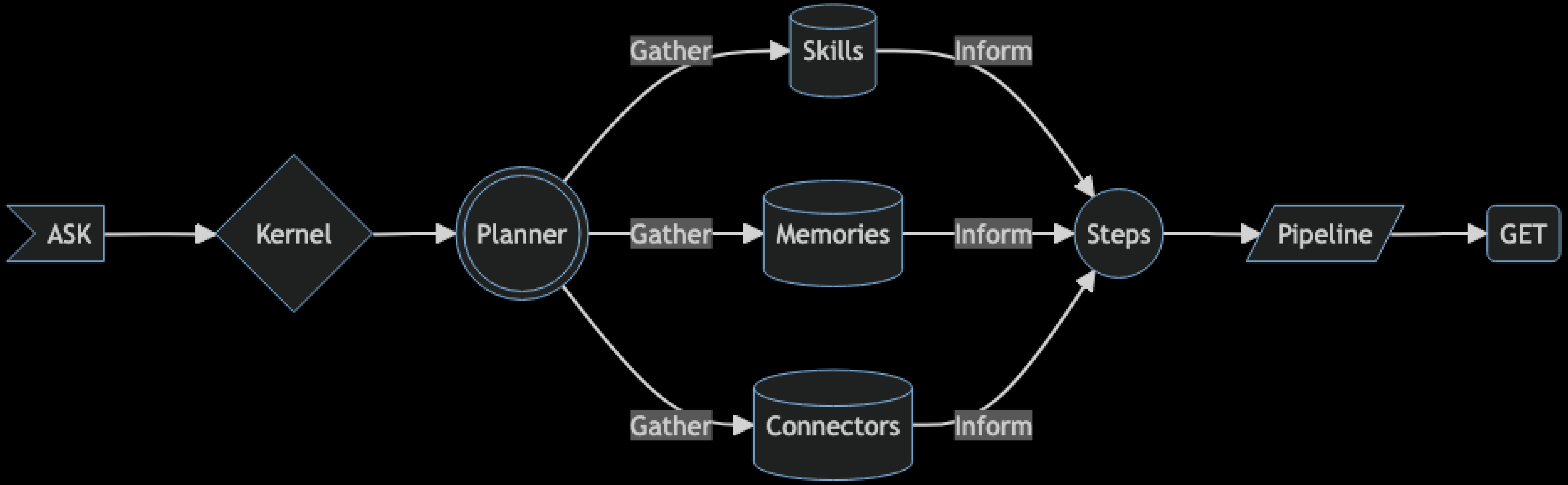
Step 4: Execute previous query

```
Thought: I now know the final answer.
Final Answer: The total number of tracks in each playlist is: 90’s Music (1477), Brazilian Music (39), Classical (75), Classical 101 - Deep Cuts (25),
Classical 101 - Next Steps (25), Classical 101 - The Basics (25), Grunge (15), Heavy Metal Classic (26), Music (6580), Music Videos (1), On-The-Go 1
(1), TV Shows (426).
```

Final answer is ready

Microsoft alternative: Semantic Kernel

- Open-source SDK for C# and Python
- Supports templating, chaining, and planning
- Early stages, open for feedback



Microsoft

Q&A