



Azure OpenAI Service Intelligent Apps Workshop

Kevin Harris

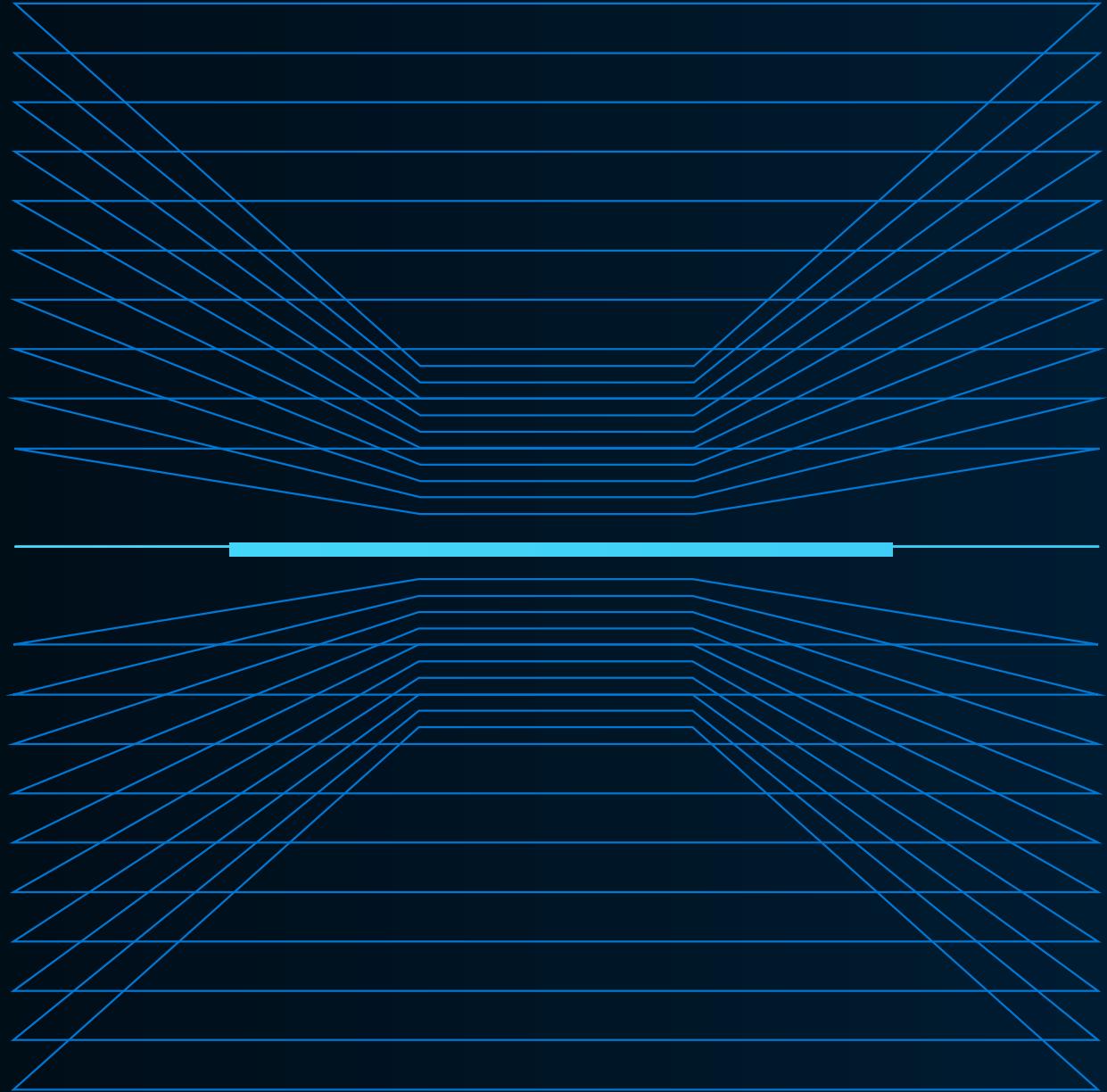
Kevin.Harris@microsoft.com

GBB App Innovation Specialist

Mark Whitby

Mark.Whitby@microsoft.com

GBB App Innovation Specialist



Plan for Today

Time (CEST)	Topic
09:00 – 09:15	Introductions & Setting Expectations
09:15 – 09:30	Use Case Ideation & Brainstorming
09:30 – 10:30	Intro to Azure OpenAI, Prompt Engineering & Demos
10:30 – 10:45	Break
10:45 – 11:15	Hands-on with Prompt Engineering
11:15 – 12:15	Intro to AI Orchestration & Demos
12:15 – 13:15	Lunch Break
13:15 – 15:15	Hands-on with AI Orchestration
15:15 – 15:30	Break
15:30 – 16:00	Use Case Validation
16:00 – 16:30	Q&A and Wrap-up

Expectations

- Key Learnings
 - What is Azure OpenAI?
 - Why Prompt Engineering is important?
 - Why is AI Orchestration needed?
 - What are some of the AI Orchestrators available?
 - Why Embeddings & Vector Stores are needed?
 - What are some of the Vector Stores available?

Use Case Ideation & Brainstorming

Use Case Ideation & Brainstorming

- Work with Coaches at your Table
 - Brainstorm on potential Use Cases / Scenarios (There is no bad idea/thought there ☺)
 - Capture Use Cases, will be used later

Intro to Azure OpenAI, Prompt Engineering

Artificial Intelligence

Machine Learning

Deep Learning

Generative AI

1956

Artificial Intelligence

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

1997

Machine Learning

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

2017

Deep Learning

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

2021

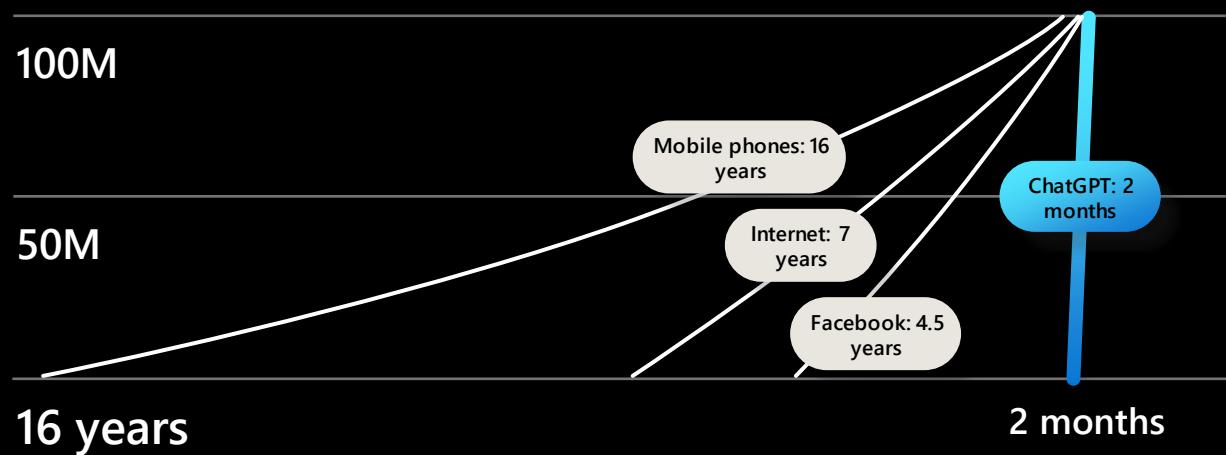
Generative AI

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

Why so much hype around ChatGPT?

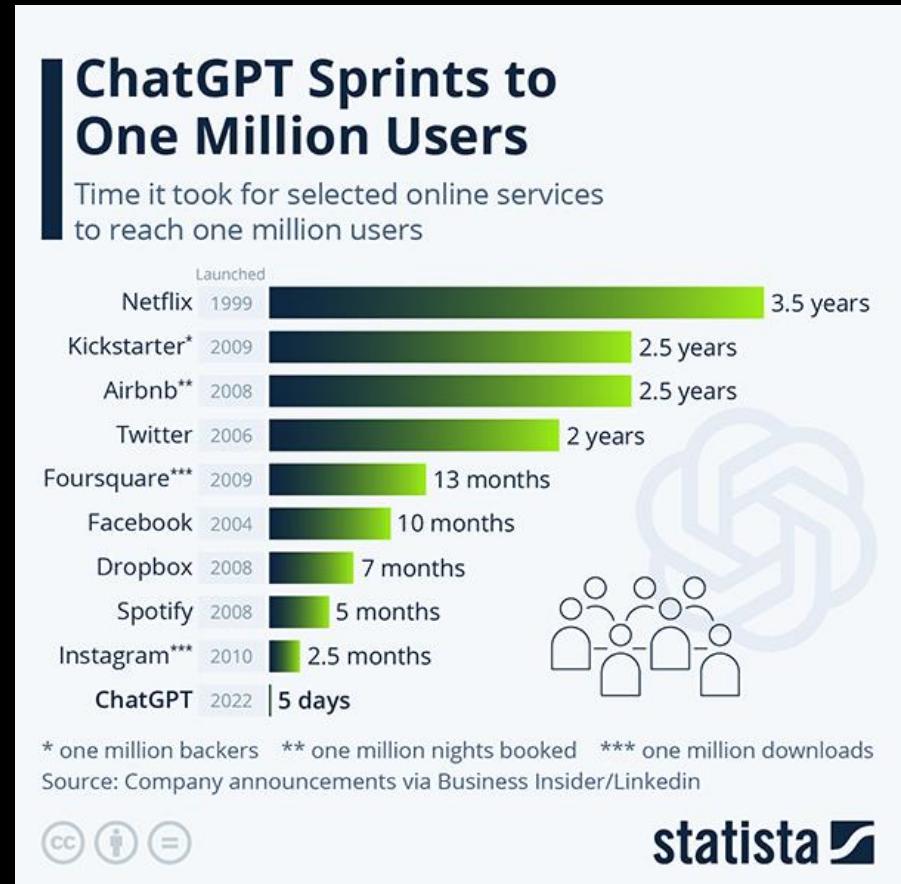
- ChatGPT
 - 100m users in 2.5 months
 - Most “human-like” interaction
- What’s new about it?
 - Can generate and do “things”
 - Answer knowledge questions directly
 - Answer coding questions
 - Can correct itself, if you ask it
- Prompt examples [here](#)

Time to reach
100M users



What's the verdict?

- ChatGPT is the fastest-growing app of all time - Swiss bank UBS
- ChatGPT reached 100 million users just two months after launching - The Guardian
- An average of about 13 million unique visitors had used ChatGPT per day in January 2023 - Similarweb



A **large language model** is a type of artificial intelligence that can process and produce natural language texts.

It **learns** from a massive amount of text data, such as books, articles, and web pages, and **discovers the patterns and rules of language** from them.

It can do various tasks, such as answering questions, summarizing texts, writing essays, and more.

Prompt:

In 100 words or less, please tell me what a large language model is

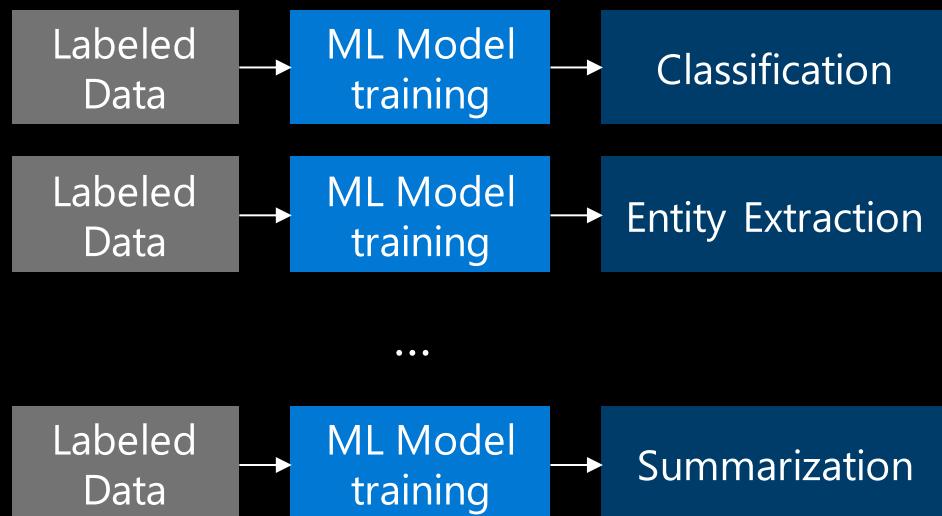
Prompt:

A futuristic abstract rendering of a large language model



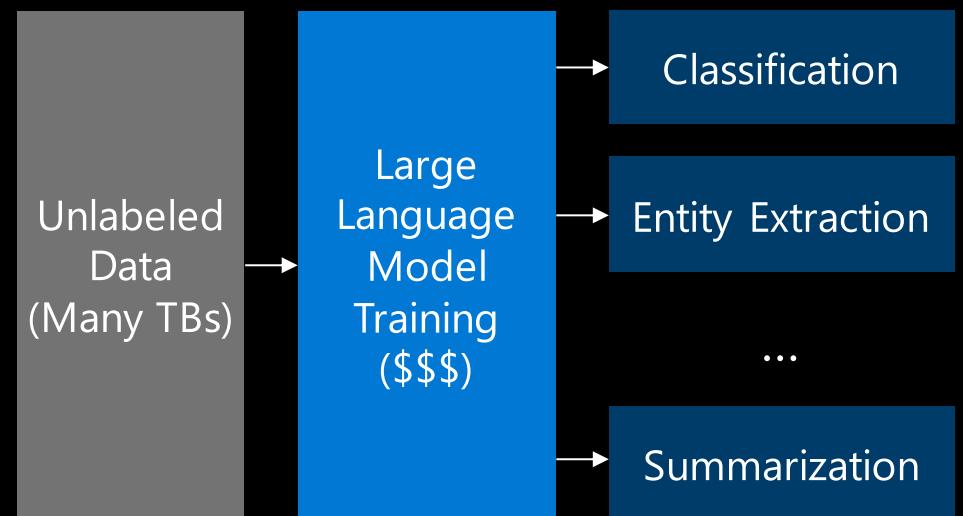
Large Language Models

Typical ML for NLP



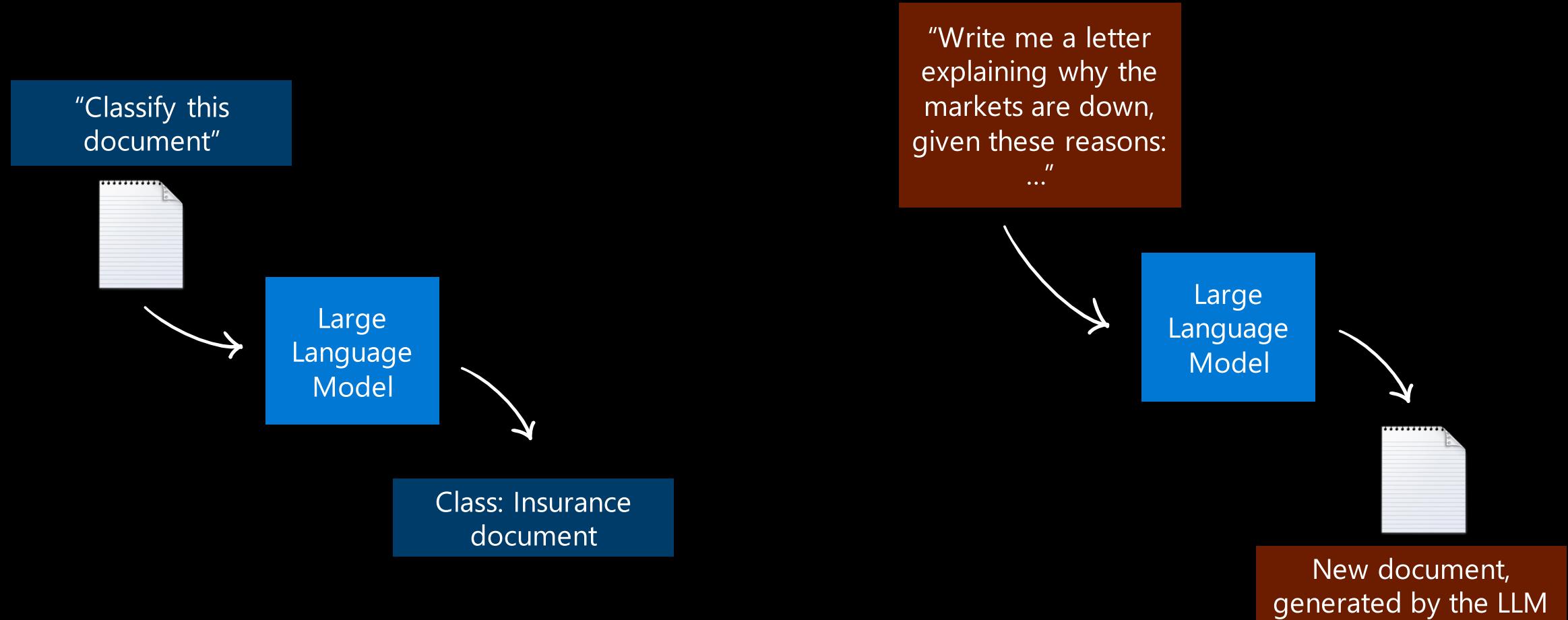
One model per capability
Labeled data to train
Highly optimized for use case

Large Language Models



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

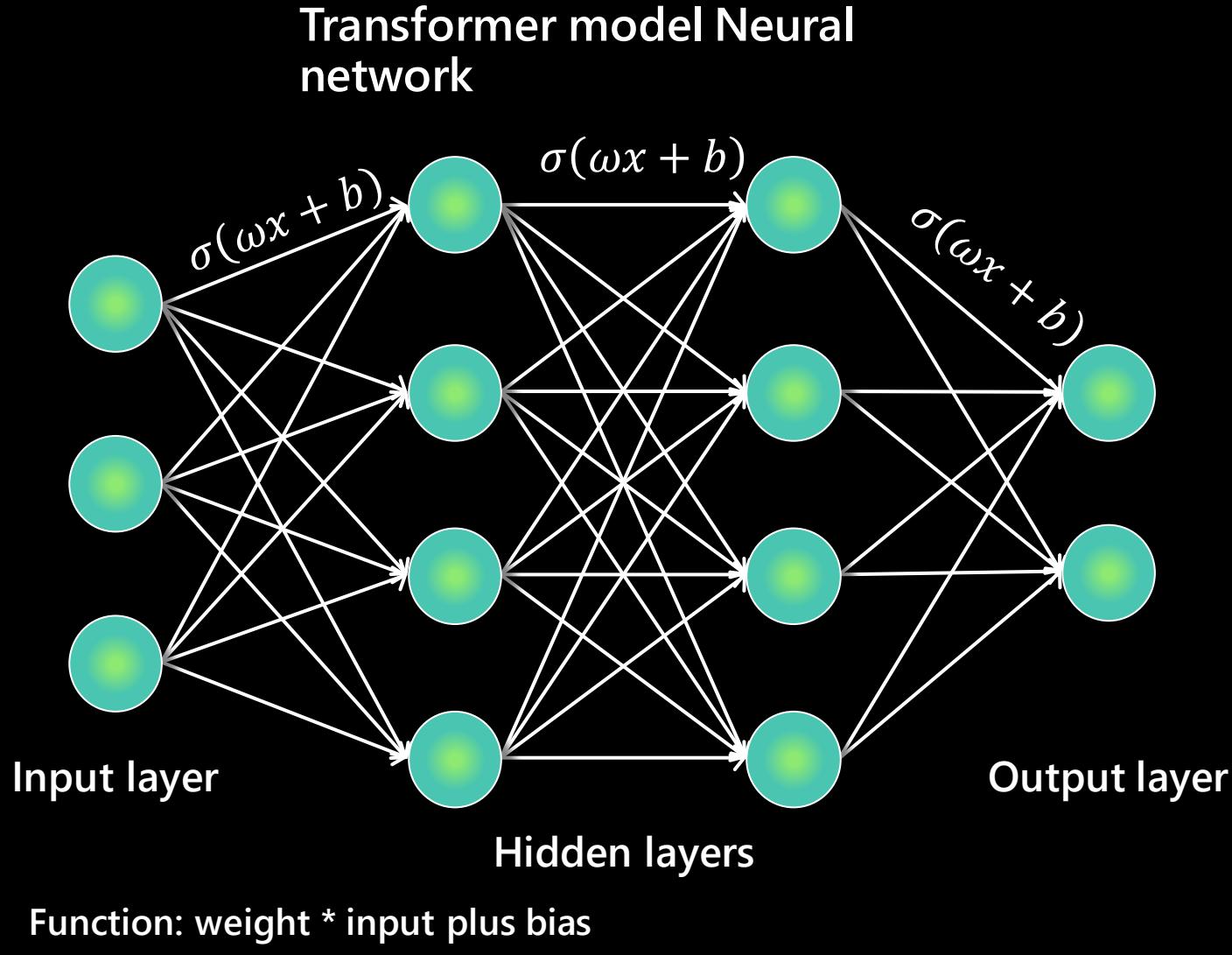
LLMs can also “generate things”



LLMs can do typical ML tasks easily (often zero-shot learned)

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

How large are they?



BERT Large - 2018

345M

GPT2 - 2019

1.5B

GPT3 - 2020

175B

Turing Megatron NLG
2021

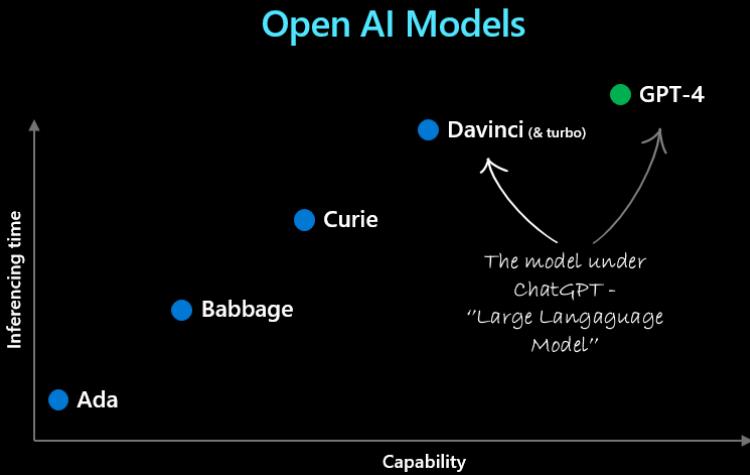
530B

GPT4 – 2023

1.4T (estimated)



- OpenAI is an AI research and deployment company
- Focus: Groundbreaking research, Latest models & Fast innovation
- *Mission is to ensure that artificial general intelligence benefits all of humanity.*



- In 2019, Microsoft invested over \$1 billion in OpenAI
- Develop advanced AI based super computing technologies on Microsoft Azure.
- Azure is OpenAI's exclusive cloud provider
- Azure will power all OpenAI workloads across research, products and API services.

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.





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

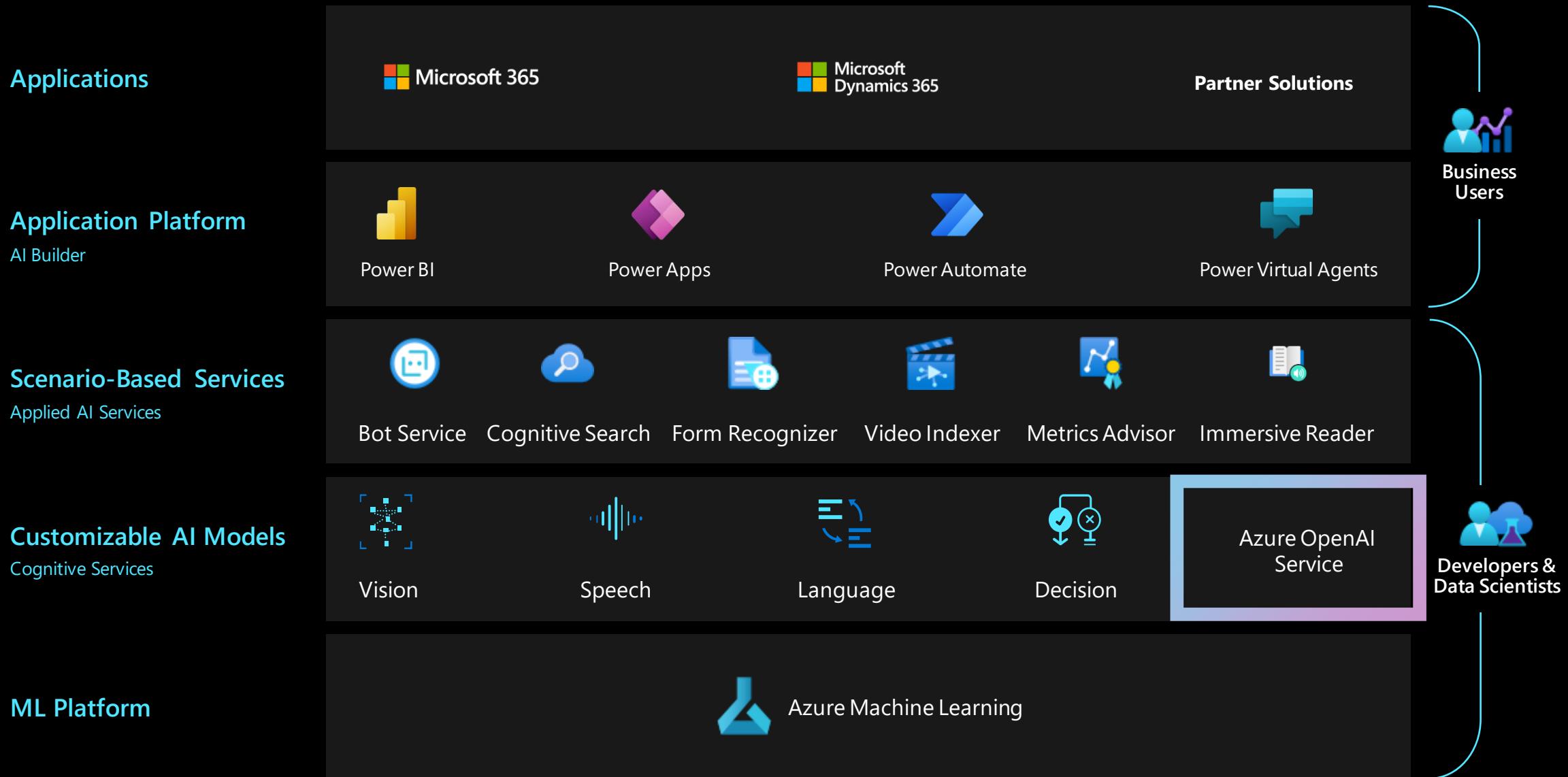
Groundbreaking research
Latest models
Fast innovation



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

Enterprise features
Production SLAs
Data privacy

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



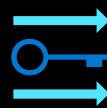
Large, pretrained AI models to unlock
new scenarios



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

Getting Started with Azure Open AI

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)

2

3. Make sure fields 1 to 6 are answered very short, e.g. for location output as JSON object called "results". Pretty print the JSON and

4

Phone conversation:
Hi I just had a car accident and wanted to report it. OK, I hope you up. That's understandable. Can you give me your full name? Sure accident take place? On the I-18 freeway. Was anyone else injured information? Sure, his name is John Radley. And your insurance the airbags went off. Are you going to be able to drive it? I don we'll get everything sorted out. Thank you.

```
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."  
}
```

Azure AI | Azure AI Studio

Azure OpenAI

Playground

Chat

Completions

DALL-E (Preview)

Azure OpenAI Service Instruction in English

machine-readable format, using the keys from above. Format the

Customer conversation in English (or other language)

18 and I hit a pothole. OK, where did the need to do an investigation. Can you give me the other drivers what type of damages has the car? Headlights are broken and I'll need to get it inspected. I'll go ahead and start the claim and

Output as JSON

Play in Azure OpenAI Studio

Cognitive Services | Azure OpenAI Studio

Azure OpenAI Studio > Completions playground

Completions playground

Deployments Examples

text-davinci-003 Load an example View code

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:
choose 1 if the question is about auto insurance.
choose 2 if the question is about home flood insurance.
choose 3 if the question is not relevant to insurance.

Customer question: Hi there, do you know how to choose flood insurance?
Classified topic: 2

Customer question: Hi there, I have a question on my auto insurance.
Classified topic: 1

Customer question: Hi there, do you know how to apply for financial aid?
Classified topic: 3

Customer question: I am interested in getting another auto insurance for my new car.
Classified topic: 1

Parameters

Temperature: 0

Max length (tokens): 200

Stop sequences

Top probabilities: 1

Frequency penalty: 0

Presence penalty: 0

Best of: 1

Pre-response text

Post-response text

Learn more

Generate Undo Regenerate Tokens: 186

Intro to Prompt Engineering

Azure OpenAI Service | 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.

What is Prompt Engineering?

Prompt engineering is a concept in Natural Language Processing (NLP) that involves embedding descriptions of tasks in input to prompt the model to output the desired results.

Prompt typically includes problem descriptions, instructions on how to solve the problem, and examples of correct problem and solution pairs.

An example of Prompt Engineering– Topic Classifier

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Instructions

High level and detailed

Customer question: Hi there, do you know how to choose flood insurance?

Classified topic: 2

Customer question: Hi there, I have a question on my auto insurance.

Classified topic: 1

Customer question: Hi there, do you know how to apply for financial aid?

Classified topic: 3

Customer question: {insert new question here}

Classified topic:

Examples

Order of examples matter

Task and Prompting
answer

Two Types of Prompt Engineering

Zero-shot

The model predicts the answer given only a natural language description of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: {insert new question here}

Classified topic:

Few-shot

In addition to the task description, the model sees a few examples of the task.

Classify customer's question. Classify between category 1 to 3.

Detailed guidelines for how to choose:

choose 1 if the question is about auto insurance.

choose 2 if the question is about home flood insurance.

choose 3 if the question is not relevant to insurance.

Customer question: Hi there, do you know how to choose flood insurance?
Classified topic: 2

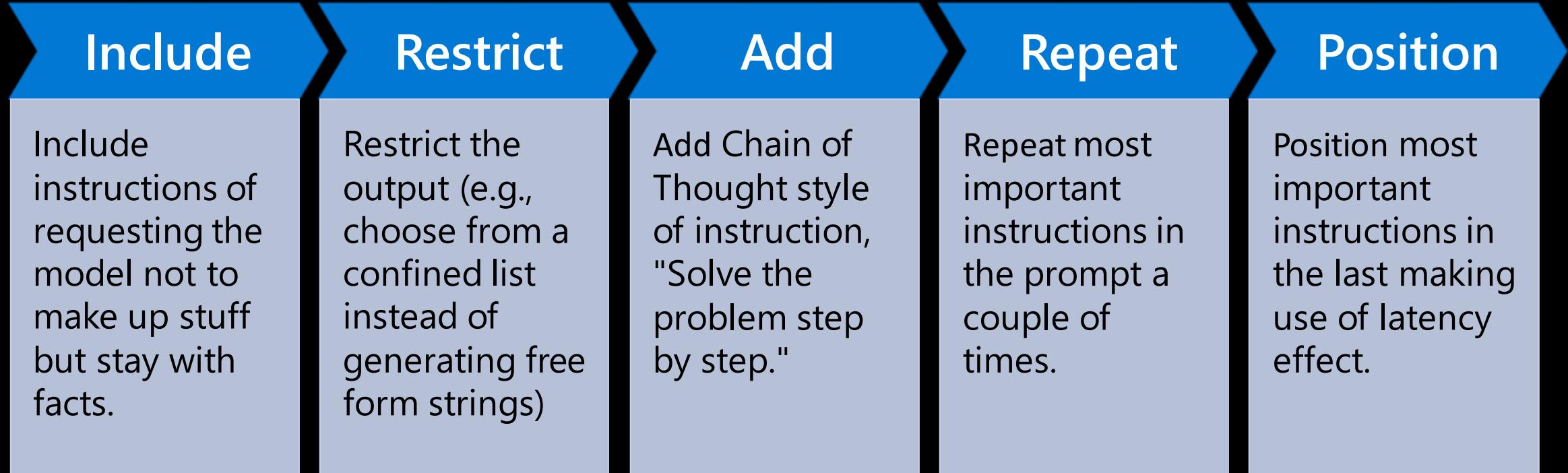
Customer question: Hi there, I have a question on my auto insurance.
Classified topic: 1

Customer question: Hi there, do you know how to apply for financial aid?
Classified topic: 3

Customer question: {insert new question here}

Classified topic:

Reduce Hallucination using Prompt Engineering



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. Custom-tune custom models to maximize performance



Prompt Chaining

Entity Extraction

Please **extract entities** from the following news article: 'The new iPhone model is set to be released next month. It has been highly anticipated by Apple fans and is expected to feature a larger screen and improved camera'

```
{'text': 'iPhone', 'type': 'PRODUCT'}, {'text': 'Apple', 'type': 'ORGANIZATION'}
```

Summarization

Please **summarize** the information about the product

```
[{'text': 'iPhone', 'type': 'PRODUCT'}, {'text': 'Apple', 'type': 'ORGANIZATION'}]
```

The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

Sentiment Analysis

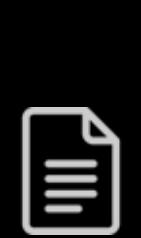
Please provide a **sentiment** for the following text:

The iPhone is a line of smartphones designed and marketed by Apple Inc. These devices combine a computer, camera, and internet connectivity in a single device, and are known for their sleek design and user-friendly interface.

Positive

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.



Large Text
Beyond
Token
Limit

Chunks

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 $100 \text{ liters} \times 200 \text{ days} = 20,000 \text{ liters per year}$.

→ **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.

Using Chain-of-Thought for Reasoning

1st prompt—Reason Extraction

On average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?

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

In one minute, Joe throws 25 punches.

In three minutes, Joe throws $3 * 25 = 75$ punches.

In five rounds, Joe throws $5 * 75 = 375$ punches.

2nd prompt—Answer Extraction

One average Joe throws 25 punches per minute. A fight lasts 5 rounds of 3 minutes. How many punches did he throw?

In one minute, Joe throws 25 punches.

In three minutes, Joe throws $3 * 25 = 75$ punches.

In five rounds, Joe throws $5 * 75 = 375$ punches.

Therefore, the answer (Arabic numerals) is

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)

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.

Add new fact to context



Demo Azure OpenAI Studio

- Sample Prompts in the repo, or feel free to use your own:
 - <https://github.com/Azure/intro-to-intelligent-apps/blob/main/presentations/prompts-demos.md>

Break

Hands-on with Prompt Engineering

- Create Azure OpenAI Service
 - <https://learn.microsoft.com/en-us/azure/cognitive-services/openai/how-to/create-resource?pivots=web-portal>
- Open AI Studio
 - Go to [Azure OpenAI Studio - Microsoft Azure](#)
 - Walk through Text Completion tutorial (if needed)
 - <https://learn.microsoft.com/en-us/azure/cognitive-services/openai/how-to/completions>
- Do Prompt Exercises in the repo:
 - <https://github.com/Azure/intro-to-intelligent-apps/blob/main/labs/01-prompts/README.md>

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.

Translate the following sentence into German.

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

German translation:

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

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

Decide whether the product review's sentiment is positive, neutral or negative. Show the probability for positive, neutral and negative sentiment.

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.

You are adding tag categories to movies, based on their descriptions.

###

Movie description: 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.

Tags: ['Action', 'Adventure', 'Science Fiction']

###

Movie description: A botched store robbery places Wonder Woman in a global battle against a powerful and mysterious ancient force that puts her powers in jeopardy.

Tags: ['Action', 'Adventure', 'Fantasy']

###

Movie description: 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.

Tags: ['Adventure', 'Science Fiction', 'Action']

###

Movie description: 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.

Tags: ['Drama', 'Family', 'Comedy']

###

Movie description: 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.

Tags:

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

Exercise - Bring Your Own Data

Exercise

Write a prompt that uses some of your own data and ask it questions.
It could be data copied from a website or document for example.

Input text

<<TEXT COPIED FROM WIKIPEDIA FOR EXAMPLE>>

Expected completion

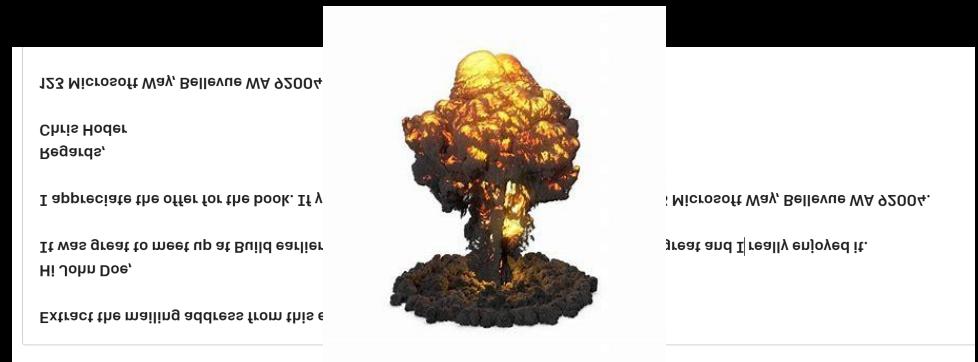
<<ARE YOU GETTING THE RESULT YOU EXPECTED? IF NOT, WHY?>>

Intro to Azure OpenAI, Prompt Engineering Q&A

Intro to AI Orchestration

Why do we need AI Orchestration?

Limits (aka Blowing up the Prompt)



Quotas and limits reference

The following sections provide you with a quick guide to the default quotas and limits that apply to Azure OpenAI:

Limit Name	Limit Value
OpenAI resources per region per Azure subscription	30
Default quota per model and region (in tokens-per-minute) ¹	Text-Davinci-003: 120 K GPT-4: 20 K GPT-4-32K: 60 K All others: 240 K
Default DALL-E quota limits	2 concurrent requests
Maximum prompt tokens per request	Varies per model. For more information, see Azure OpenAI Service models
Max fine-tuned model deployments	2
Total number of training jobs per resource	100
Max simultaneous running training jobs per resource	1
Max training jobs queued	20
Max Files per resource	30
Total size of all files per resource	1 GB
Max training job time (job will fail if exceeded)	720 hours
Max training job size (tokens in training file) x (# of epochs)	2 Billion
Max size of all files per upload (Azure OpenAI on your data)	16 MB

GPT-4

GPT-4 can solve difficult problems with greater accuracy than any of OpenAI's previous models. Like GPT-3.5 Turbo, GPT-4 is optimized for chat and works well for traditional completions tasks. Use the Chat Completions API to use GPT-4. To learn more about how to interact with GPT-4 and the Chat Completions API check out our [in-depth how-to](#).

To request access to GPT-4, Azure OpenAI customers can apply by filling out this form [↗](#)

- `gpt-4`
- `gpt-4-32k`

The `gpt-4` model supports 8192 max input tokens and the `gpt-4-32k` model supports up to 32,768 tokens.

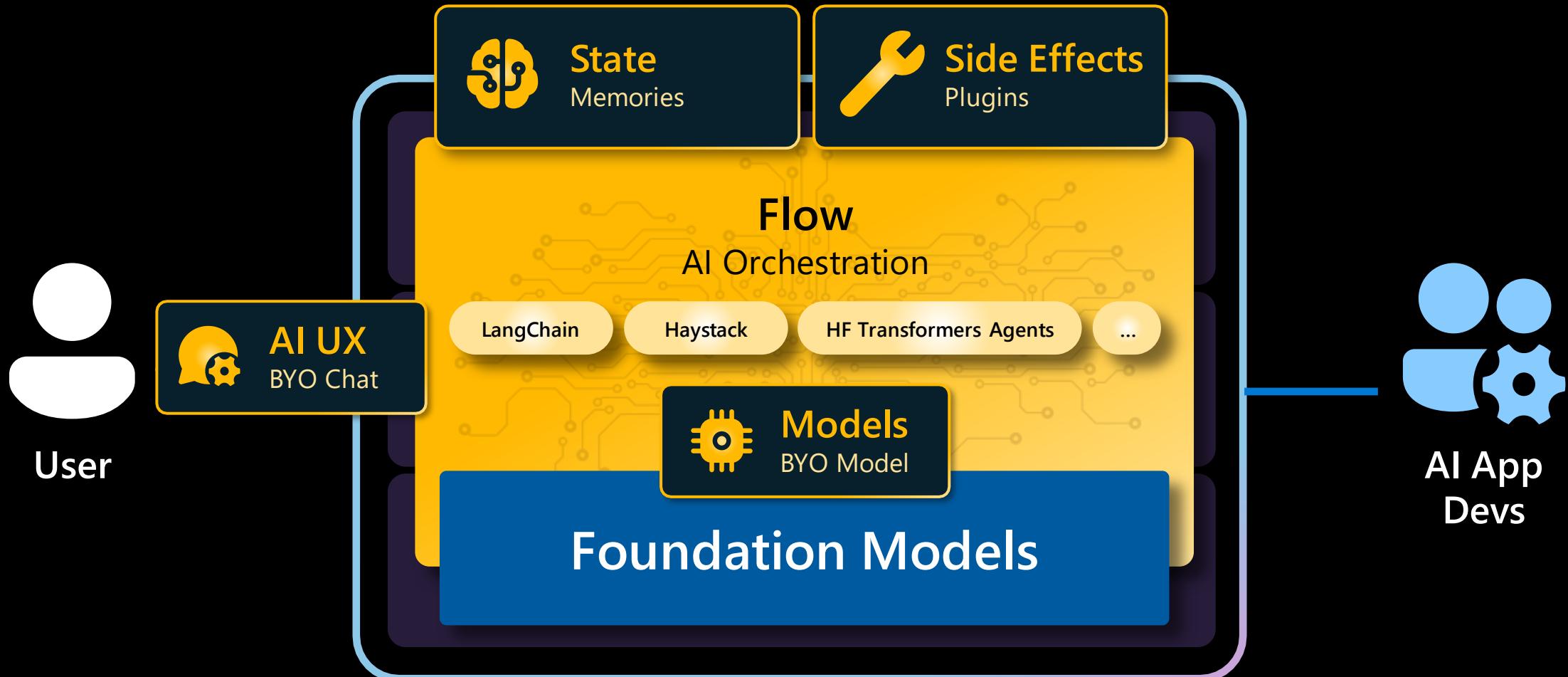
GPT-3.5

GPT-3.5 models can understand and generate natural language or code. The most capable and cost effective model in the GPT-3.5 family is GPT-3.5 Turbo, which has been optimized for chat and works well for traditional completions tasks as well. We recommend using GPT-3.5 Turbo over [legacy GPT-3.5 and GPT-3 models](#).

- `gpt-35-turbo`
- `gpt-35-turbo-16k`

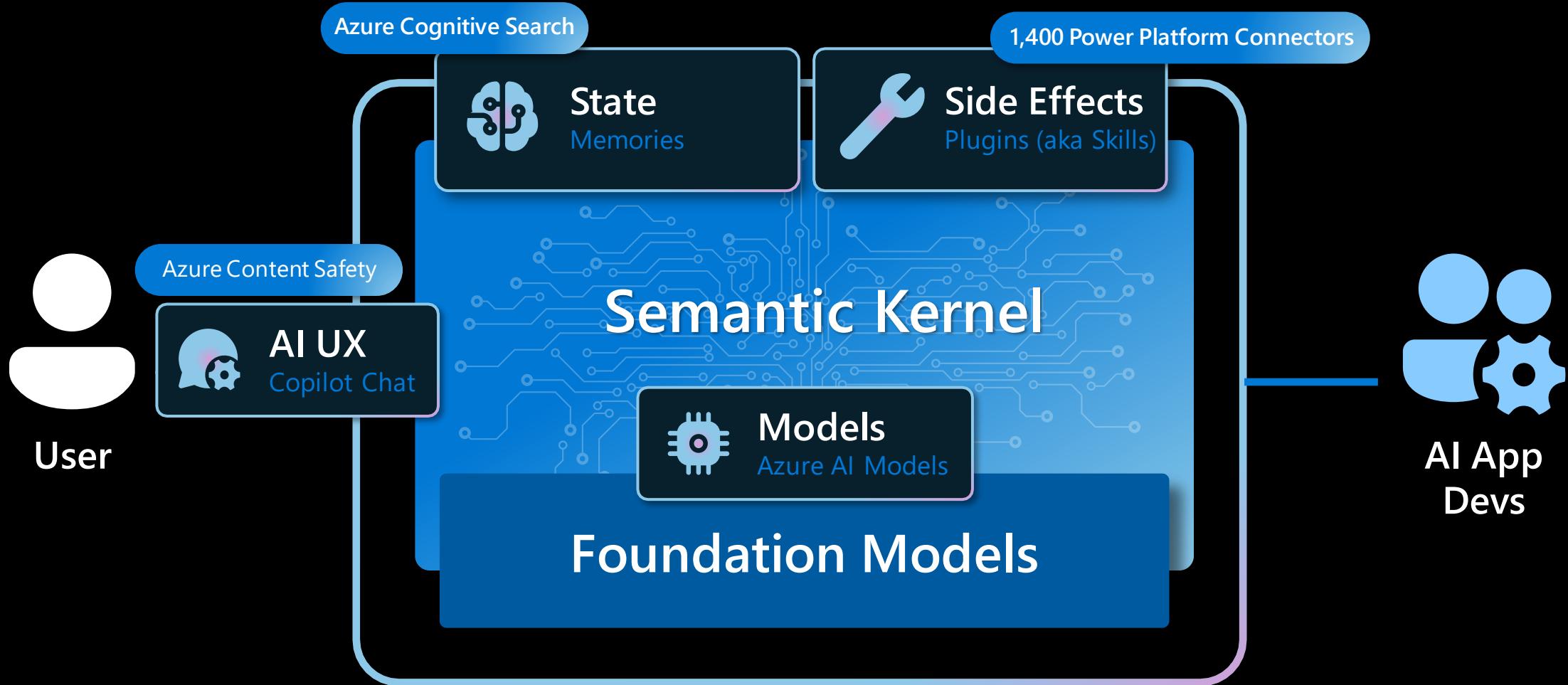
The `gpt-35-turbo` model supports 4096 max input tokens and the `gpt-35-turbo-16k` model supports up to 16,384 tokens.

AI Pipelines: Orchestration for All of the Moving Parts



Semantic Kernel

Integrate AI into your existing apps using C#, Python & Java



Intro to AI Orchestration Building Blocks

Foundation + Context



Tool

AI Pioneer Herbert Simon's Scissors (1970)



Cognition = The
'know-how'



Context =
'Specific world
knowledge'

Hallucinate & Ground:
Make-up or fill-in-the-blanks, when you
don't have context.
Provide context that fills in the blanks.

Prompt: the natural language text you send to the LLM.

Tokens: prompts encoded into tokens for the LLM to use

Embeddings: Vectors (arrays) of numbers, that represent the meaning & context of the tokens.

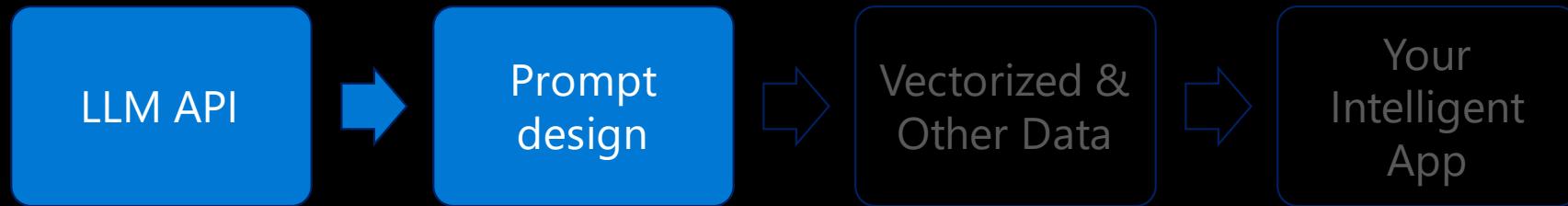
LLM API

Prompt
Design

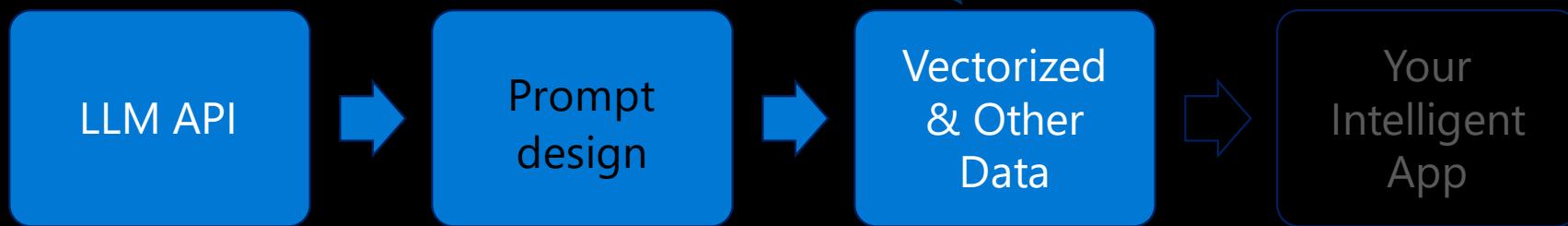
Vectorized &
Other Data

Your
Intelligent
App

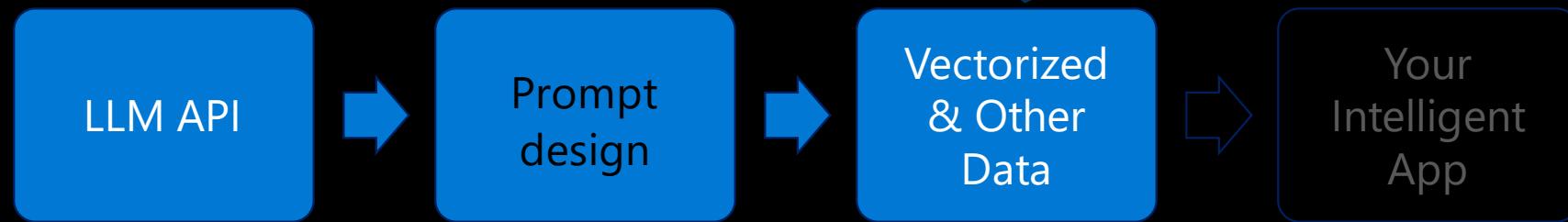
LLMs have limits to how many tokens they can process & charge based on the number of tokens you submit.

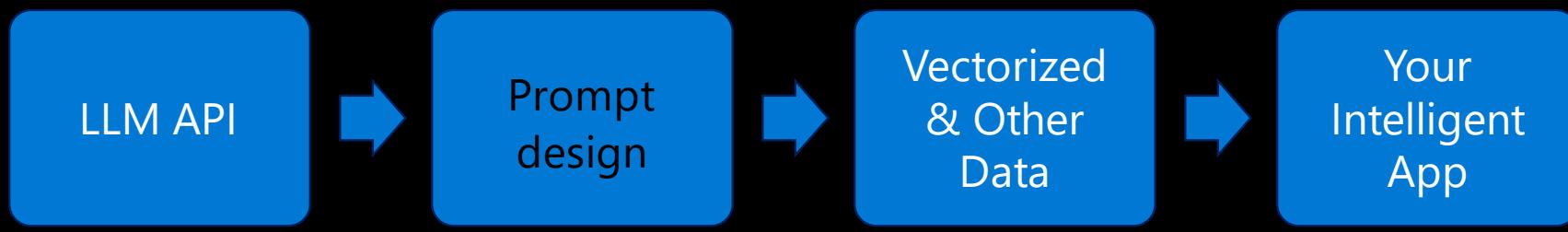


Embedding: An array of numbers, or vector, that represent the semantic meaning of a piece of text.



Vector DB: A database designed to store and retrieve data represented as vectors.
Querying in a vector db ('vectorized search') finds similar assets by searching for neighboring data points.





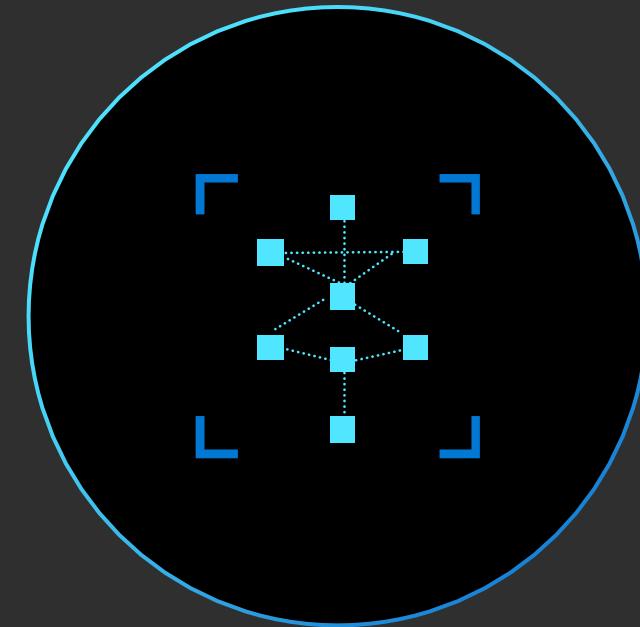
Embeddings

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.

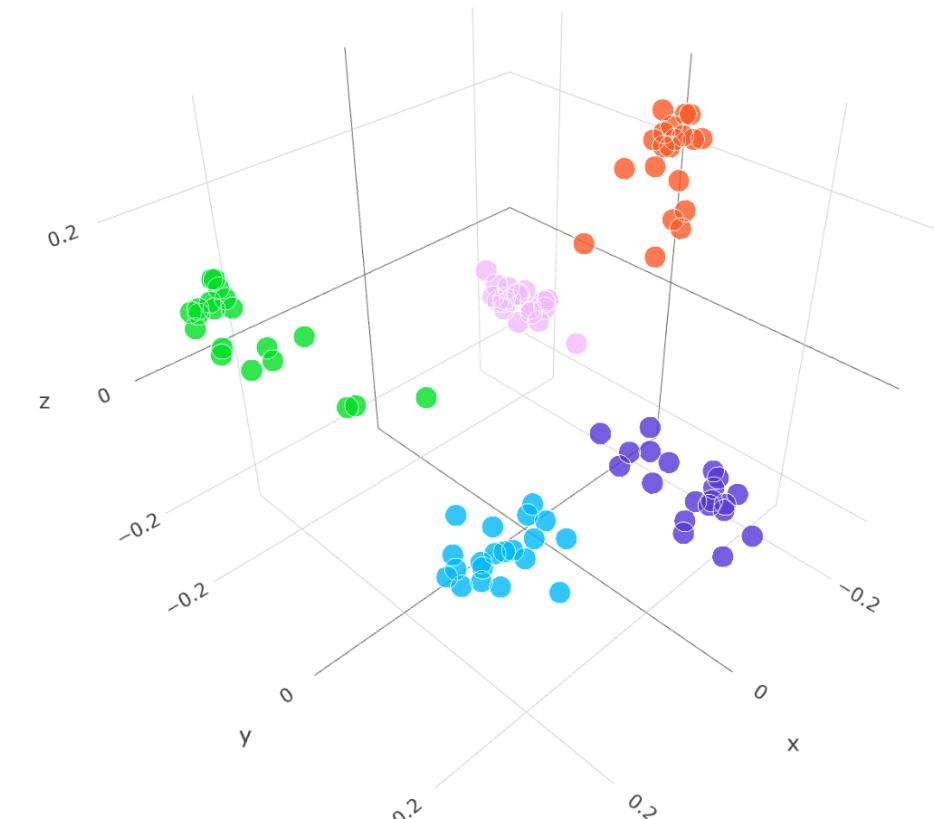
For example, if two texts are similar, then their vector representations should also be similar.



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

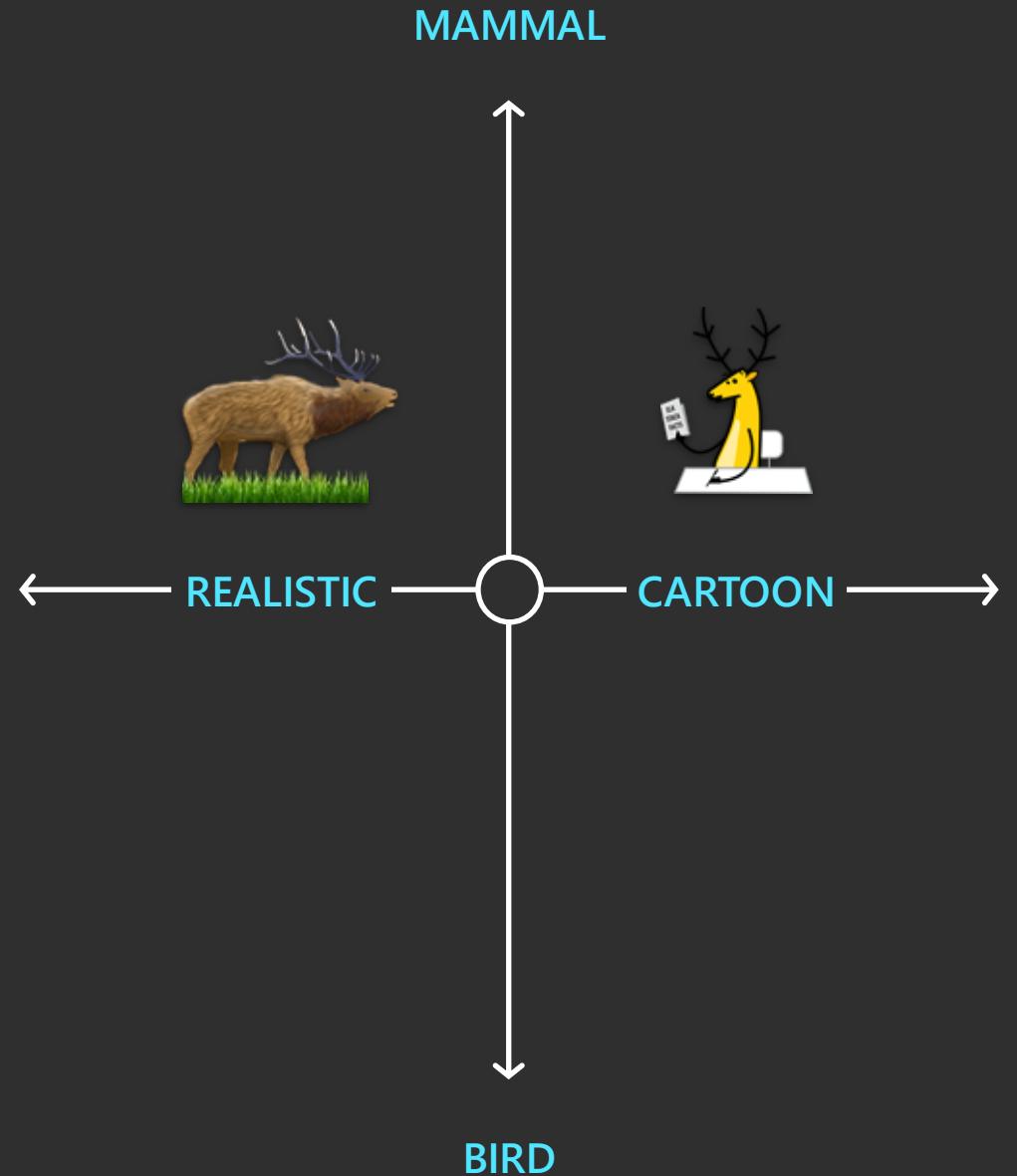
● animal ● athlete ● film ● transportation ● village



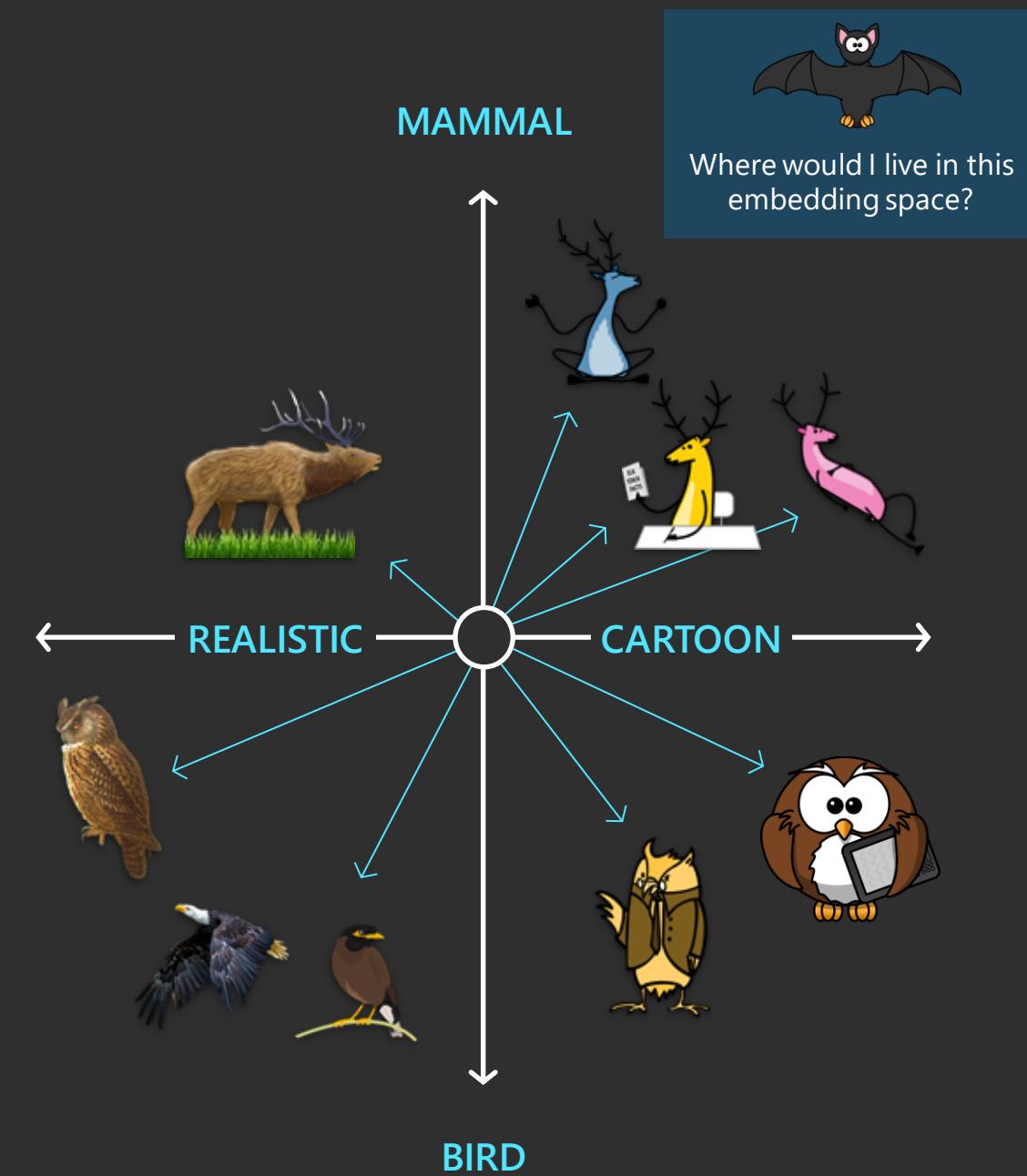
Embeddings represent your data
Example: 1-dimensional vector



Multiple dimensions represent different aspects of data



In the “embedding space,”
similar data are grouped together



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

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

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

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



[15 34 24 13 ...]

[16 22 89 26 ...]

[20 13 31 89 ...]

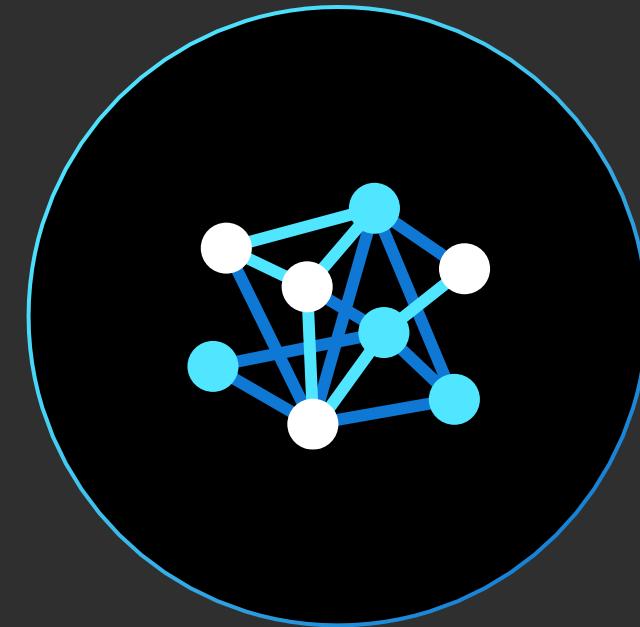
Embedding models

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

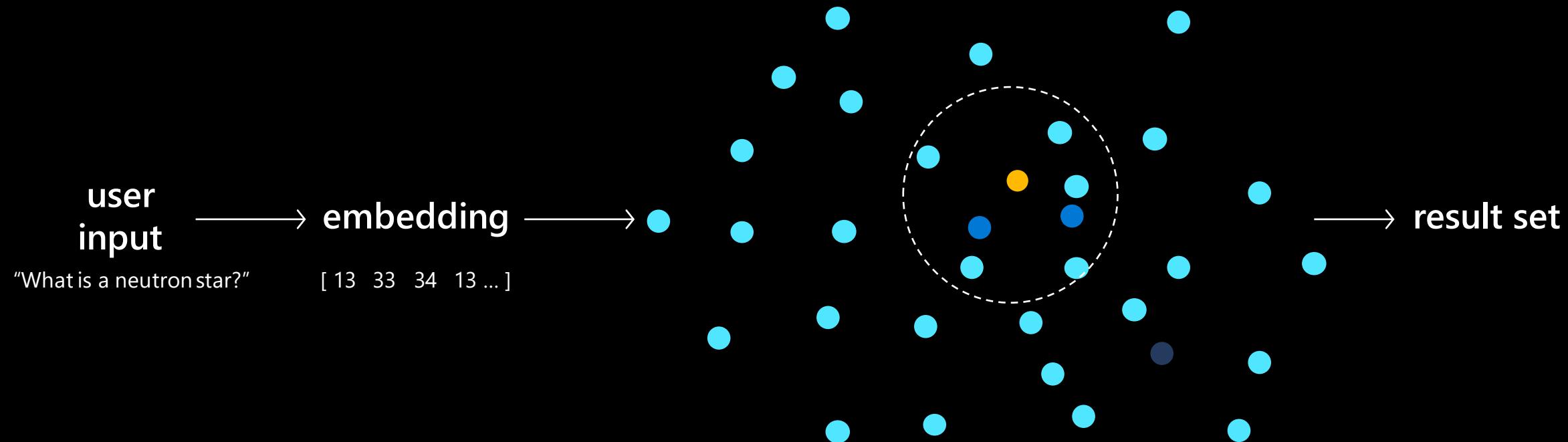
Embeddings make it easier to do machine learning on large inputs representing words by capturing the semantic similarities in a vector space.

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

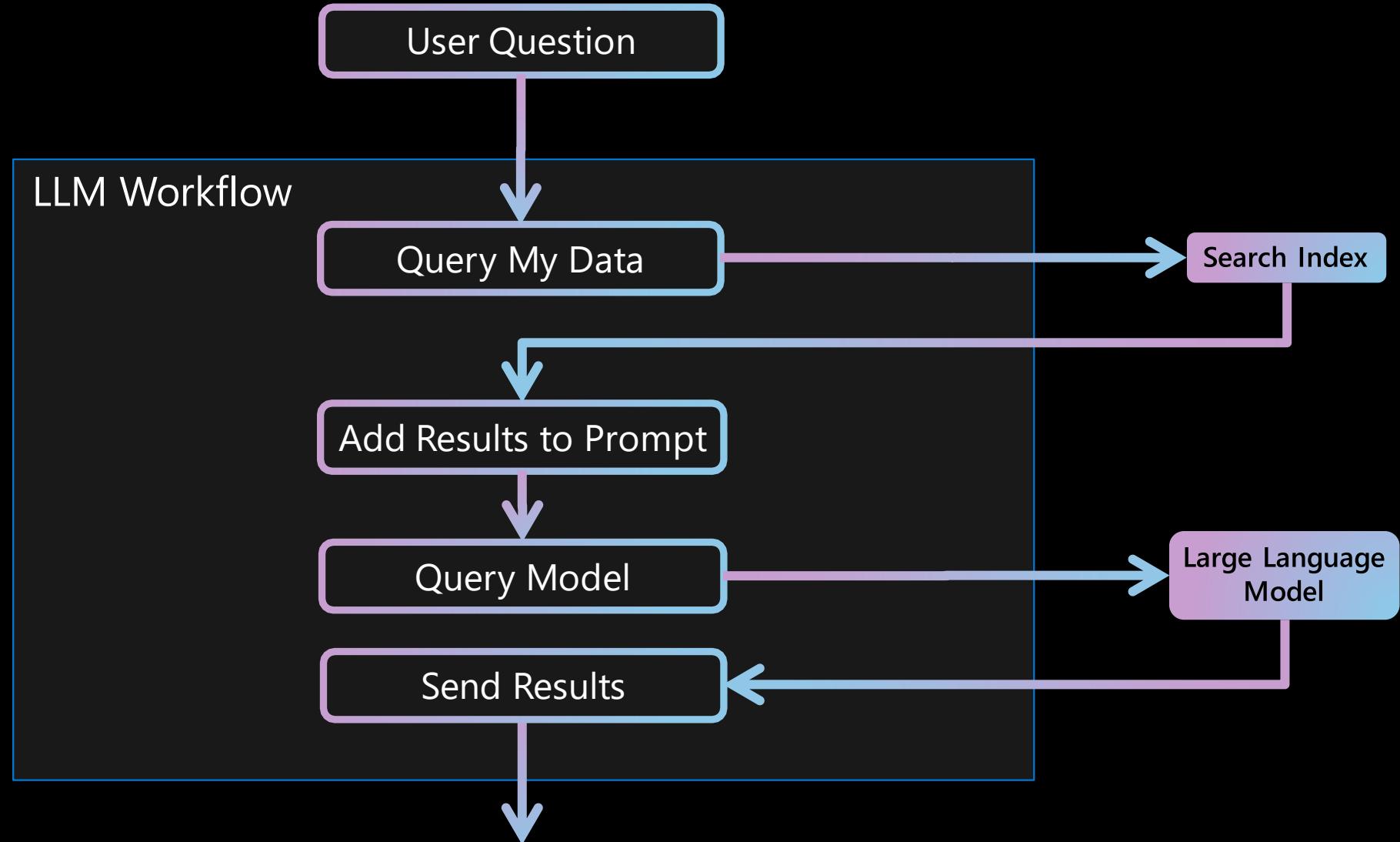


Similarity Search with embeddings

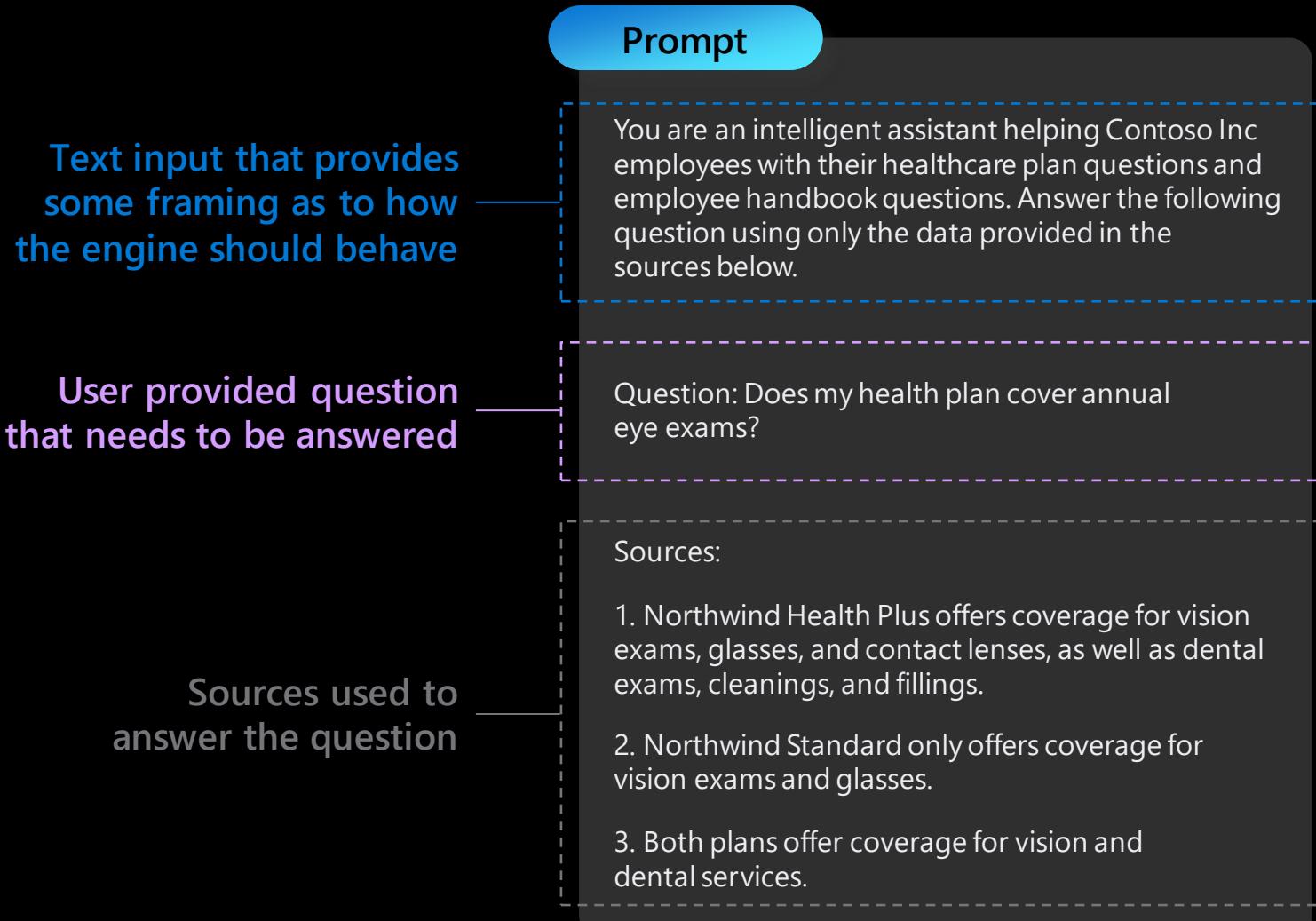
Once you encode your content as embeddings, you can then get an embedding from the user input and use that to find the most semantically similar content



Retrieval Augmented Generation



Retrieval augmented generation: bring your data to the prompt



Response

Based on the provided information, it can be determined that both health plans offered by Northwind Health Plus and Northwind Standard provide coverage for vision exams. Therefore, your health plan should cover annual eye exams.

General Recommendations



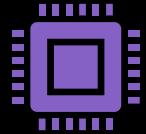
Direct use of foundation models with prompt engineering is preferred. For prompt engineering, try out a few different prompts and pick the best one.



Utilizing similarity search to create dynamic prompts can be beneficial in cases where the total content is too extensive, as it allows for the incorporation of the most pertinent information for each query.



For fine-tuning, only use the highest quality dataset. Monitor validation metrics and adjust parameters to avoid overfitting/underfitting.



Experiment with prompt engineering on advanced models and fine-tune less capable models, compare the performance metrics, operational latency and cost.

Top 3 Azure OpenAI Service Use Cases



Contact Center

Survey Summarization and Analyse chat conversations

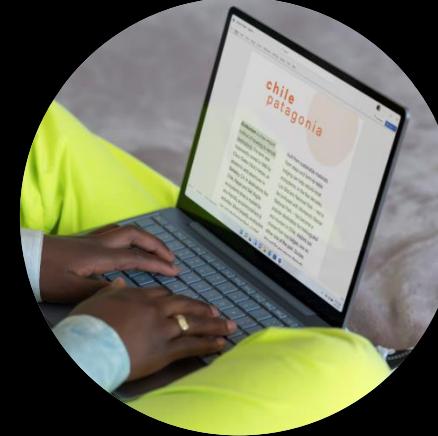
Proof: Telco, banking, insurance



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

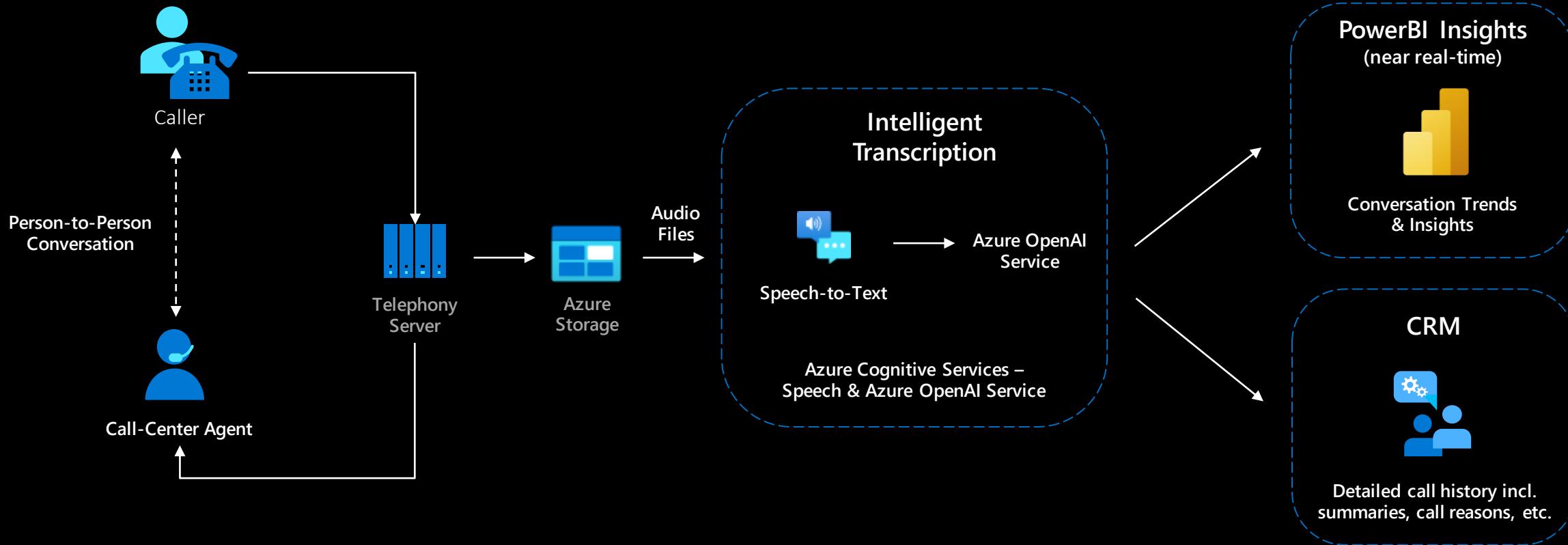
Azure OpenAI Service – Implemented Use Case

- **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 & Azure OpenAI Service

Extract rich insights from call transcripts



Azure OpenAI Service | 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 driver's 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

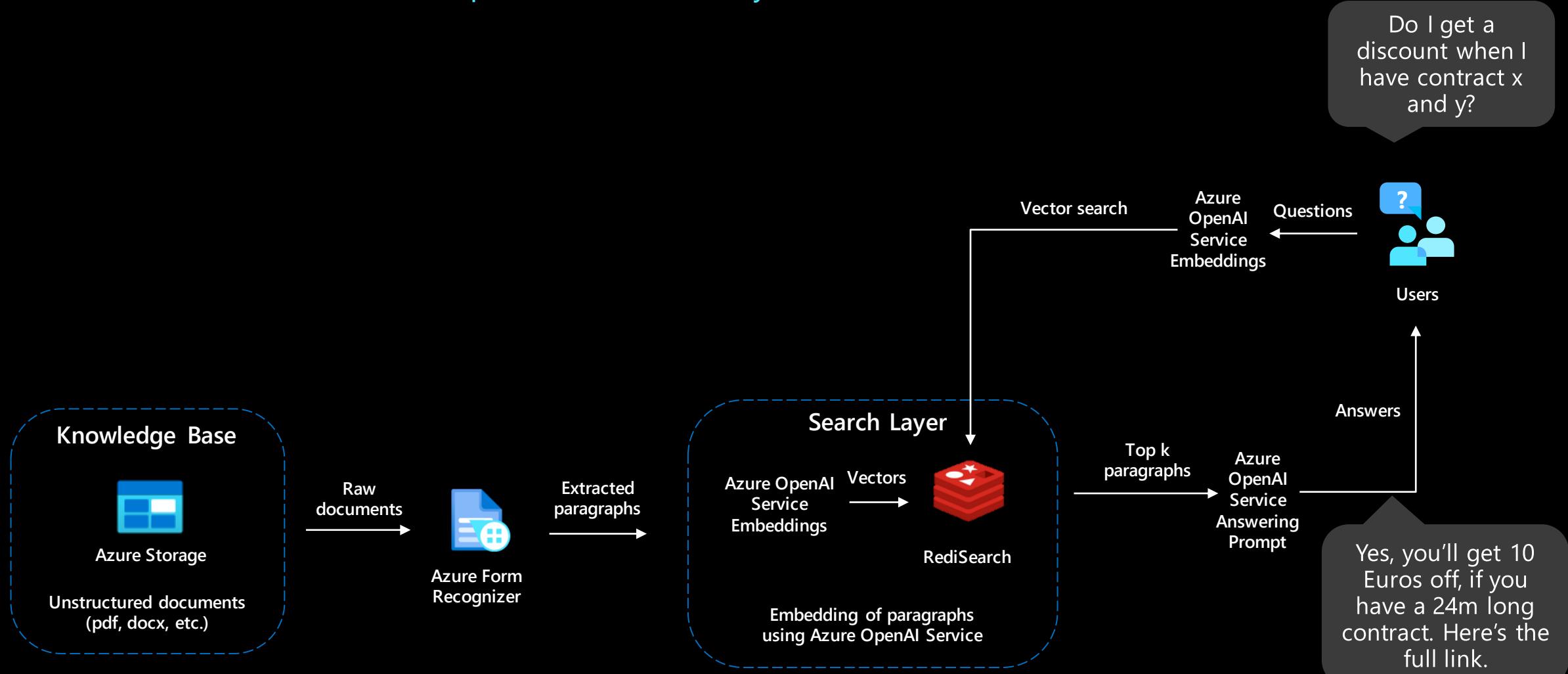
```
results = {  
    "reason": "car accident",  
    "cause": "hit a pothole",  
    "driver_names": ["Sarah Standl", "John Radley"],  
    "insurance_number": 546452,  
    "location": "I-18 freeway",  
    "damages": ["headlights broken", "airbags went off"],  
    "summary": "Sarah Standl was driving 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:
Azure OpenAI
Service Prompt

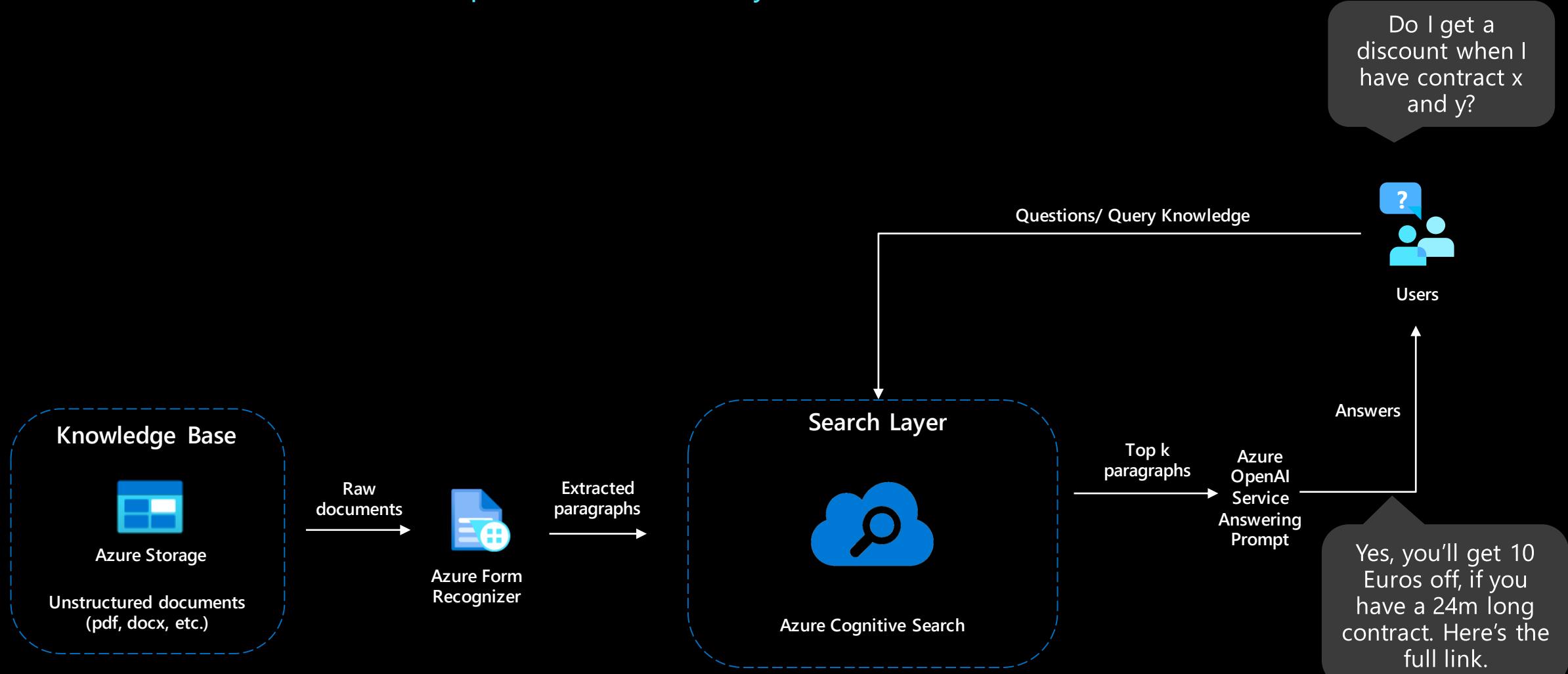
ChatGPT with Enterprise Data

Extract precise answers from your unstructured documents



ChatGPT with Enterprise Data

Extract precise answers from your unstructured documents



2 announcements and public cases in Germany

PRESS RELEASE

Siemens and Microsoft drive industrial productivity with generative artificial intelligence

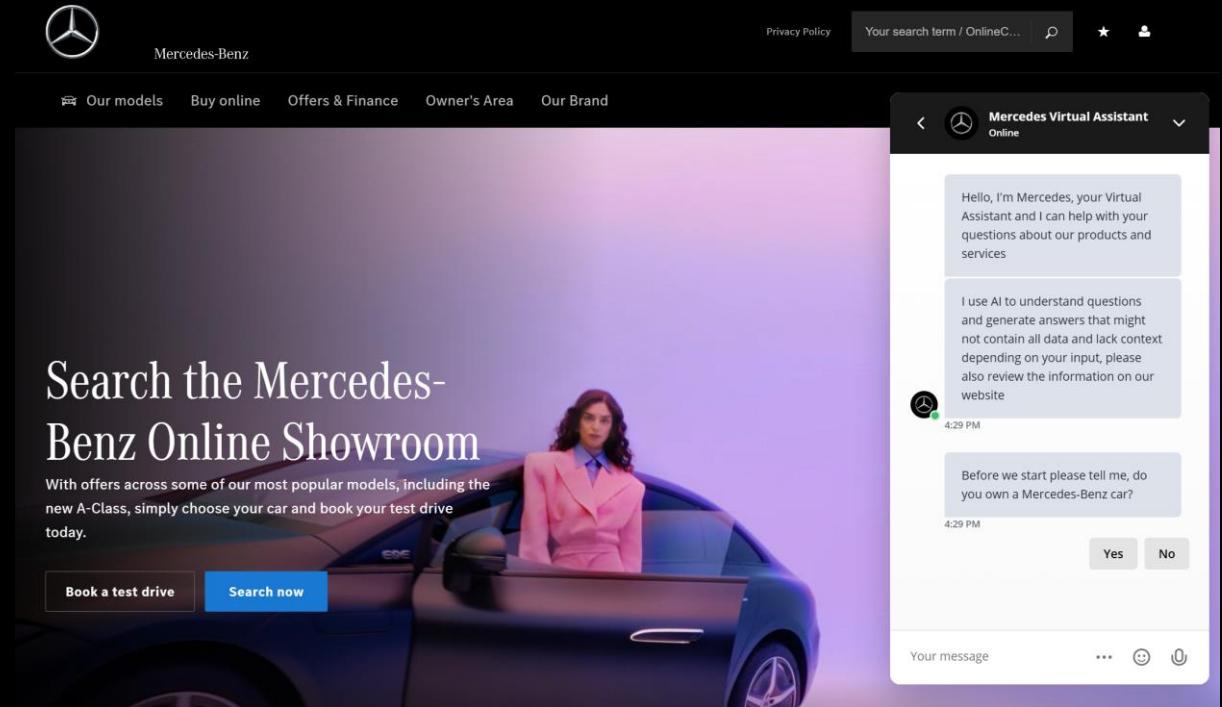
12 April 2023
Hannover Messe, Germany



Siemens and Microsoft drive industrial productivity with generative artificial intelligence

Siemens' Teamcenter® software for product lifecycle management (PLM) with Microsoft's collaboration platform Teams and the language models in Azure OpenAI Service as well as other Azure AI capabilities.

[Siemens & Microsoft drive industrial productivity with AI | Siemens Software](#)



Mercedes-Benz

Privacy Policy Your search term / OnlineC... 🔎

Our models Buy online Offers & Finance Owner's Area Our Brand

Mercedes Virtual Assistant Online

Hello, I'm Mercedes, your Virtual Assistant and I can help with your questions about our products and services

I use AI to understand questions and generate answers that might not contain all data and lack context depending on your input, please also review the information on our website

Before we start please tell me, do you own a Mercedes-Benz car?

4:29 PM Yes No

Your message

First generative AI-based bot in the automotive industry! Just the beginning!
(Car Configurator search & voice assistant)

[LinkedIn](#)



SIEMENS

Microsoft

Demo AI Apps

- Sample Demos in the repo, or feel free to use your own:
 - <https://github.com/Azure/intro-to-intelligent-apps/blob/main/presentations/aiapps-demos.md>

Lunch

Hands-on with AI Orchestration

- Setup for Labs
 - <https://github.com/Azure/intro-to-intelligent-apps/blob/main/labs/00-setup/README.md>
- Do AI Orchestration Exercises in the repo:
 - <https://github.com/Azure/intro-to-intelligent-apps/blob/main/labs/02-integrating-ai/README.md>
 - Lab 2: Integrating AI
 - Lab 3: AI Orchestration
 - Lab 4: Deployment

Break – Put this slide up around 14:45

Continue Hands-on with AI Orchestration

Intro to AI Orchestration Q&A

Use Case Validation

- Work with Coaches
 - Revisit Use Case from Morning, are they still valid?
 - Discuss amongst Team how you might go about implementing Use Cases based on what you learned from the day?
 - Capture Use Cases
 - Discuss amongst Team what it would take to get it into Production?

Q&A

Closing Remarks

- Key Learnings
 - What is Azure OpenAI?
 - Why Prompt Engineering is important?
 - Why is AI Orchestration needed?
 - What are some of the AI Orchestrators available?
 - Why Embeddings & Vector Stores are needed?
 - What are some of the Vector Stores available?

Closing Remarks Cont.

- How to get started?
 1. What is the business problem trying to be solved?
 2. Do I have access to the data needed to solve/answer the question(s)?
 3. Add needed data to a sample Prompt in Azure OpenAI Studio and ask Question(s)
(Remember, a Prompt is only so big ☺)
 1. If you did not get the right response, iterate again with a different Prompt or add additional data
 2. If you got a response that was in the general right direction, time to think about AI Orchestration
 4. Decide on AI Orchestration, Data, Embedding & Vector Store Strategies
 5. Implement Enterprise Grade API + Data + Azure OpenAI
- MS Learning
 - <https://learn.microsoft.com/en-us/training/paths/develop-ai-solutions-azure-openai/>
- Next Steps in repo:
 - https://github.com/Azure/intro-to-intelligent-apps/blob/main/docs/next_steps.md
- Engage Microsoft, we are here to help.



Thank you!



it's like a...
Language Calculator

Prompt:

A realistic photo of a less futuristic language calculator on a desk in a classroom with a pencil and a spiral notebook



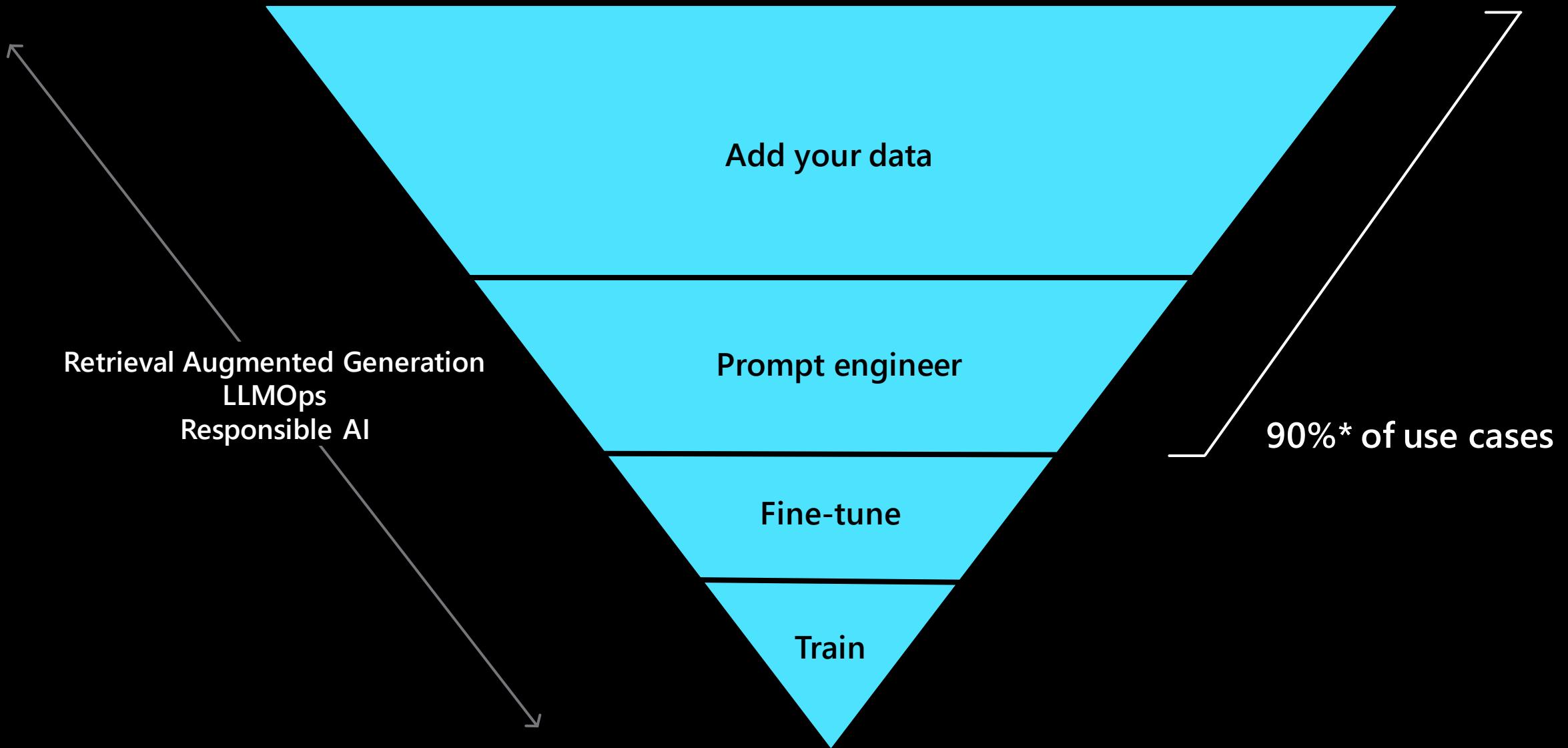
Language Calculator



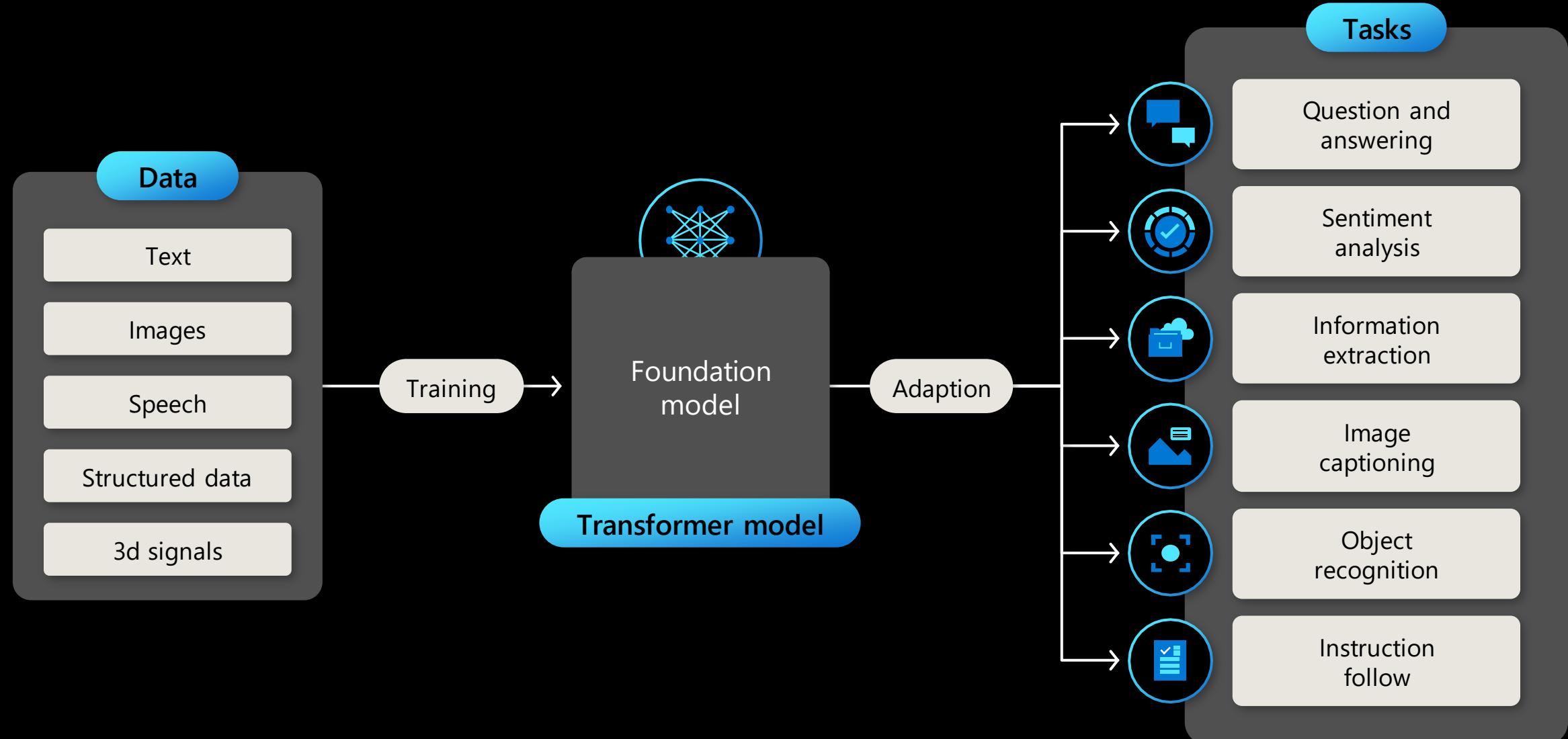
Language Calculator

Given a user input of N tokens,
return the $(N+1)$ th token,
recursive.

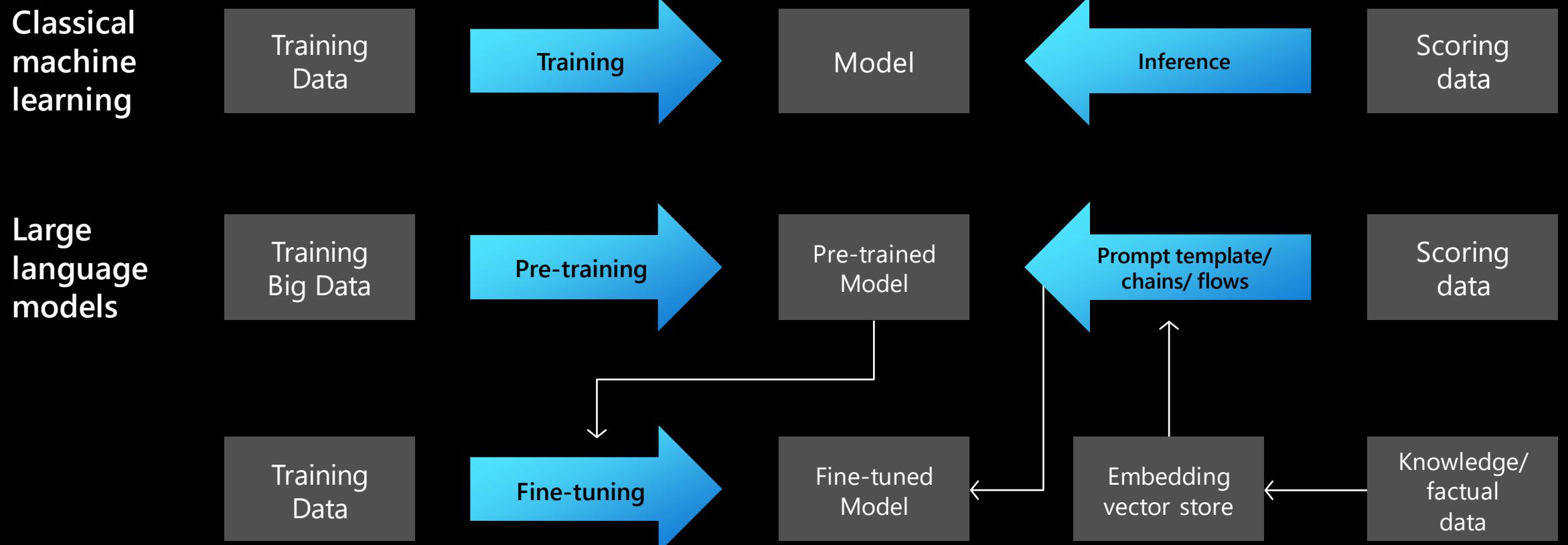
Hierarchy of language model customization



What are foundation models?



What are foundation models?

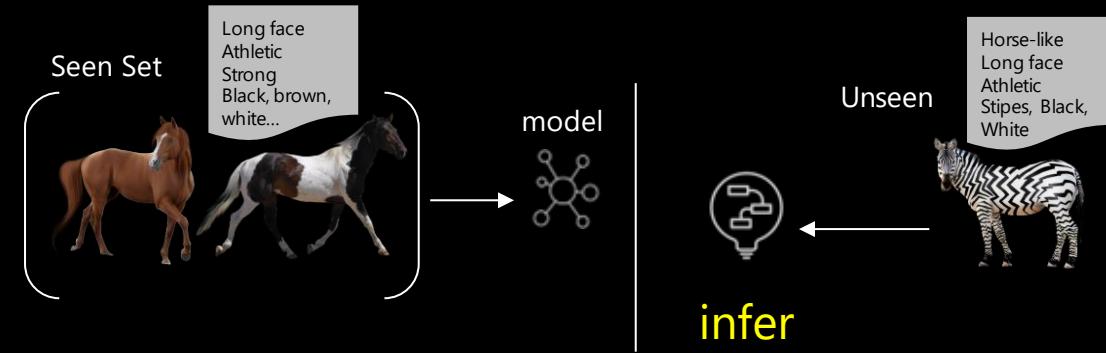


Adapting foundation models for your task

Zero-Shot

Method of training a model on data that has not been seen before.

Type of transfer learning that uses a pre-trained model to learn new task.



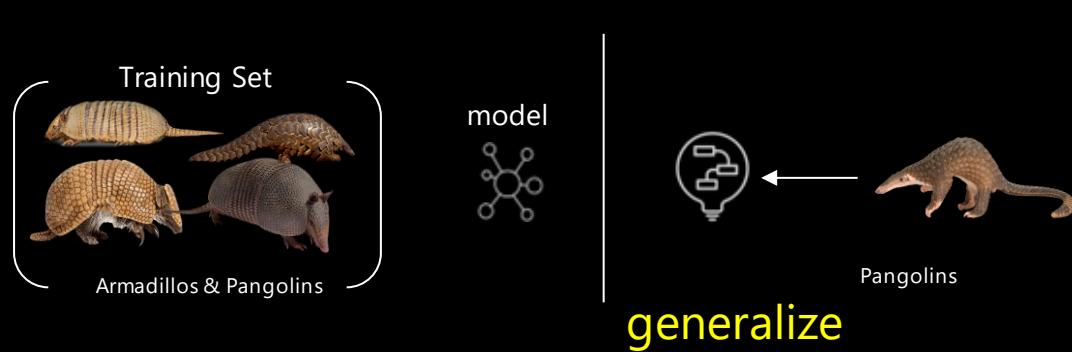
One-Shot

Describe a method of learning that requires a large dataset. It is used to train a model and make predictions based on the data provided.



Few-Shot

Learning from a small number of samples. Very useful when you have a small dataset or when it is required to generalize to a new domain.



Adapting foundation models for your task

No Gradient Updates

Zero-Shot

The model predicts the answer given only a natural language description of the task.

One-Shot

In addition to the task description, the model sees a single example of the task

Few-Shot

In addition to the task description, the model sees a few examples of the task.

Fine Tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

Prepare and upload training data



Train a new fine-tuned model



Use your fine-tuned model

1.

Potentially higher quality results than prompt engineering

2.

Ability to train on more examples than can fit in a single prompt

3.

Token savings due to shorter prompts

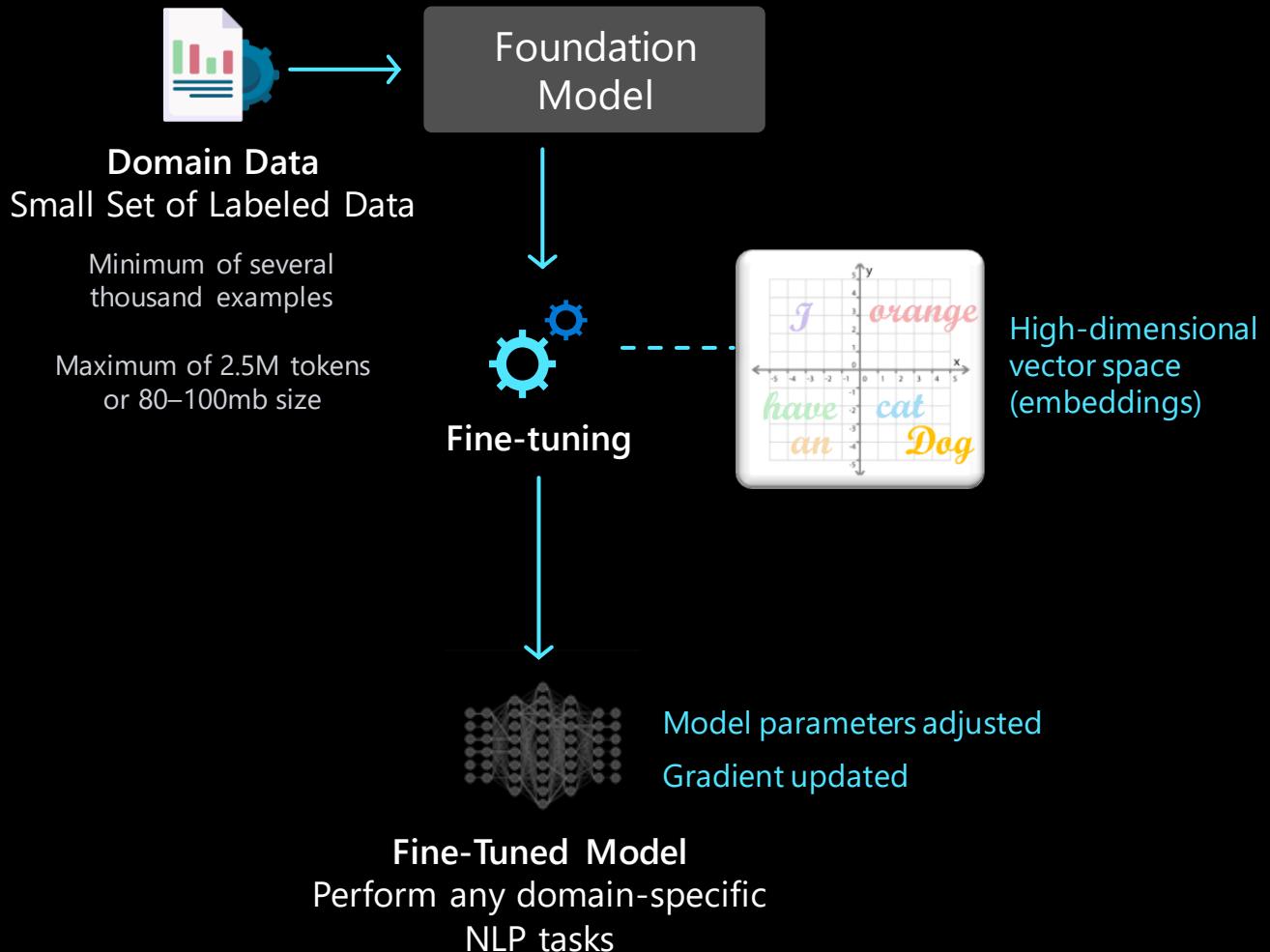
4.

Lower latency requests

Evolving to fine-tuning

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

This contrasts with **few-shot learning** in which model weights and biases are not updated.



What is Fine-Tuning?

Fine-tuning is a way of utilizing transfer learning. Specifically, fine-tuning is a process that takes a model that has already been trained and tune it using a labeled dataset for a specific task.

Fine-tuning results in a new model being generated with updated weights and biases. This contrasts with few-shot learning in which model weights and biases are not updated.

To fine-tune a model, you'll need a set of training examples that each consist of a single input ("prompt") and its associated output ("completion").

When to use Fine-Tuning?



For some use cases, fine-tuned less capable models (Ada, Babbage, Curie) performs well but more cost effective and faster than Davinci.



Latency is important for use case. Because no need to include examples in the inputs to fine-tuned models, tokens on few-shot examples are removed, this enables lower-latency requests and saves inferencing costs.



Customer has a lot of high-quality data and ground truth labels (at least a couple hundred for each use case), and customer committed to maintain high-quality data set overtime.



Use case is very domain specific, so model needs to learn from a lot of examples to know how to provide correct answers, and the number of examples needed don't fit in prompt due to token limit (less likely with 32K token limit).