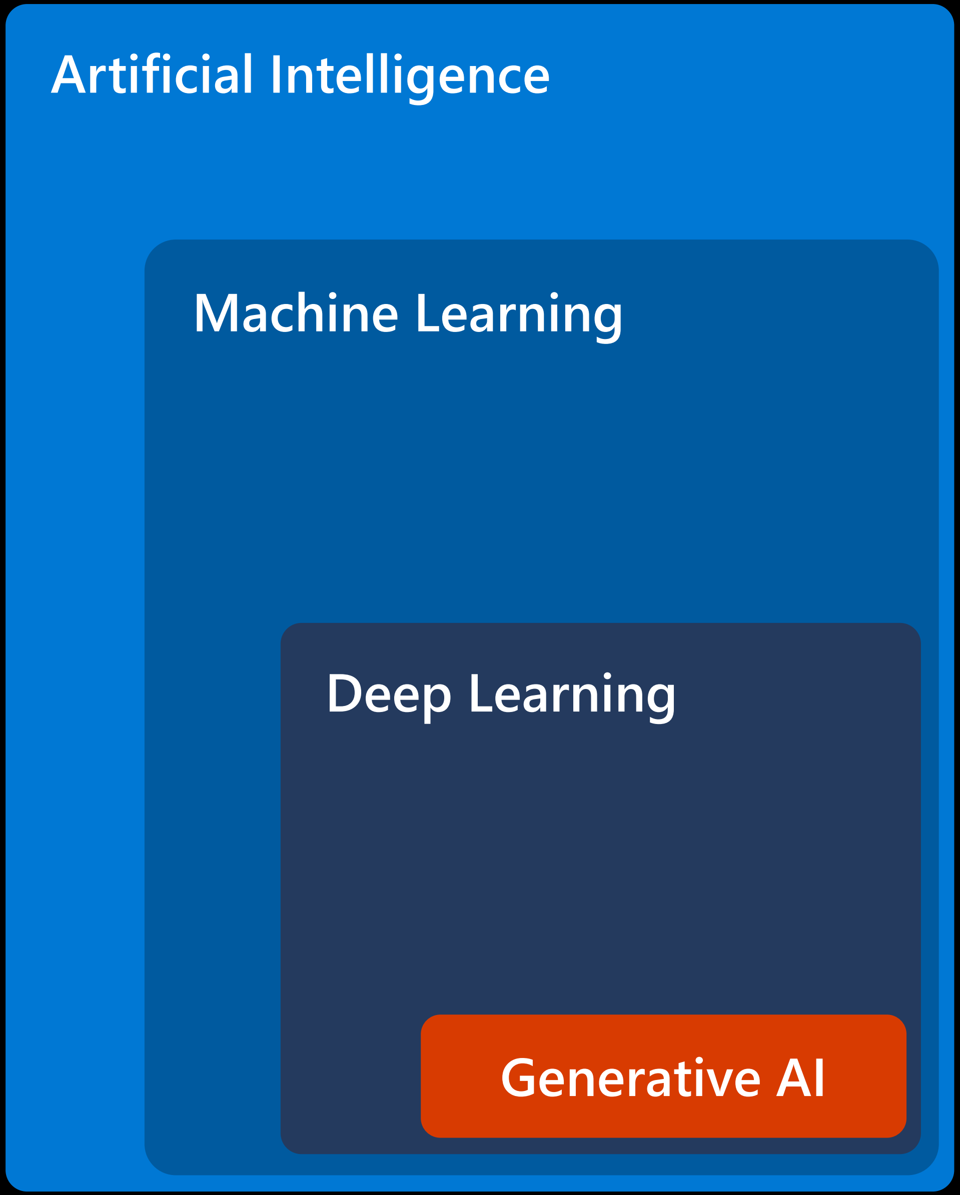
Microsoft

Azure OpenAI: Readiness Session

Roberta Bruno, CSA Data&AI



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

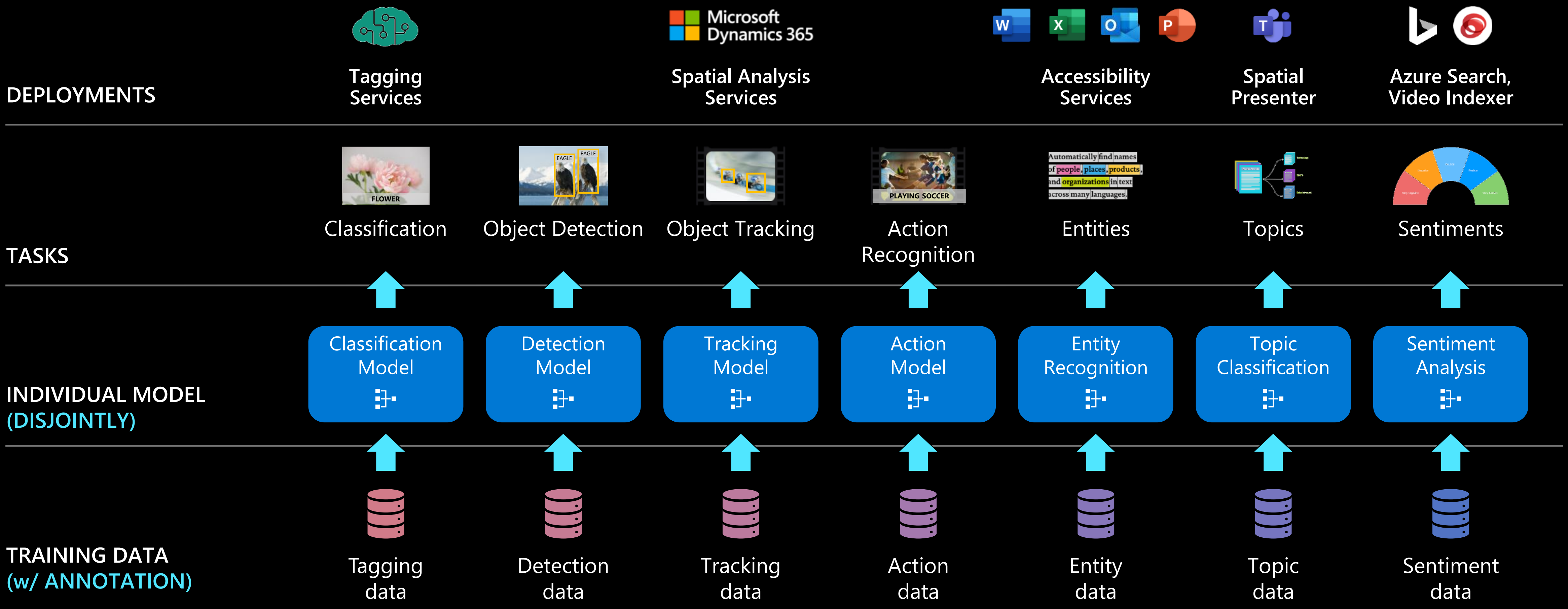
2021

Generative AI

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

Traditional model development

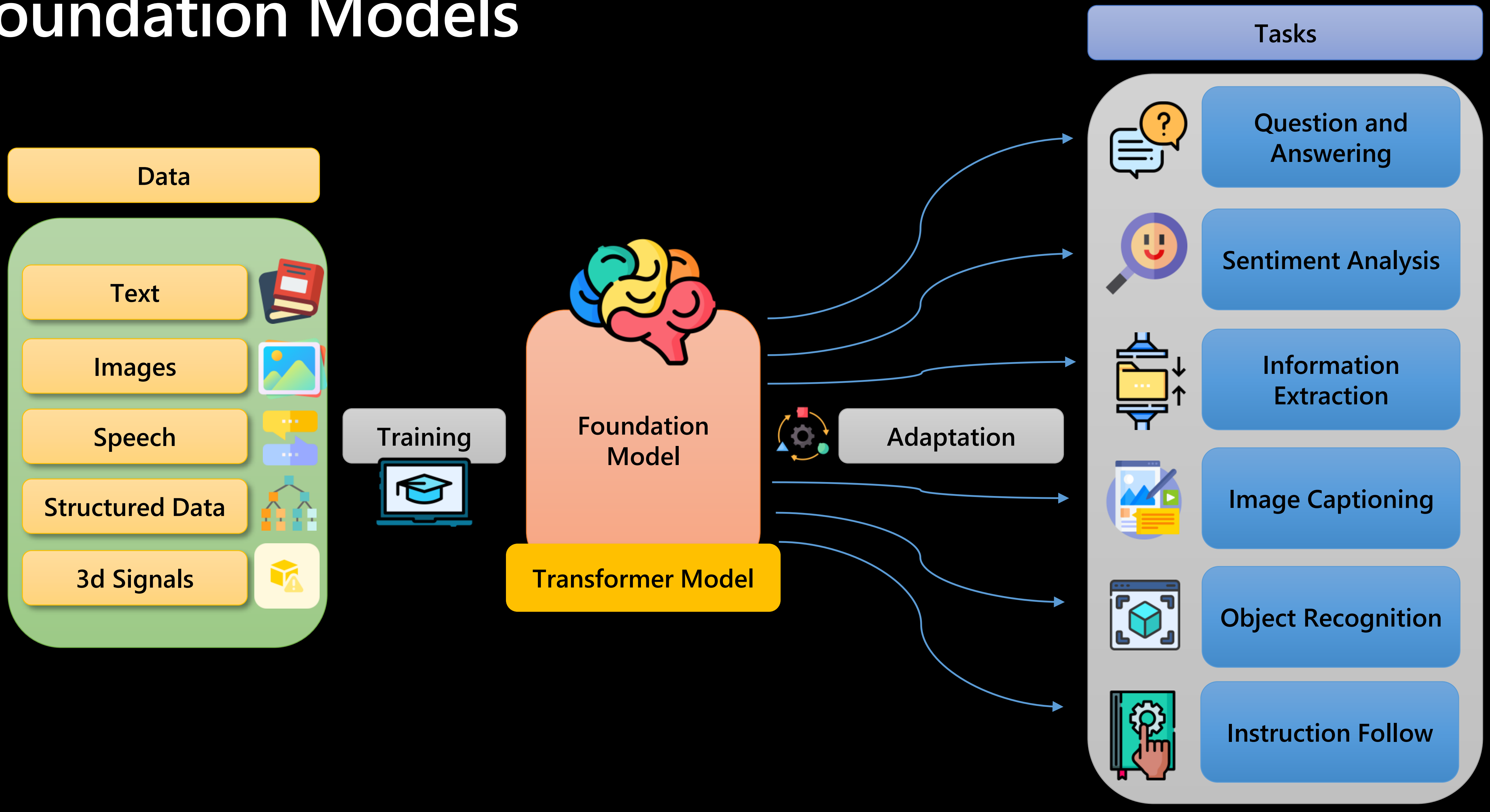
High cost and slow deployment—each service is trained disjointly

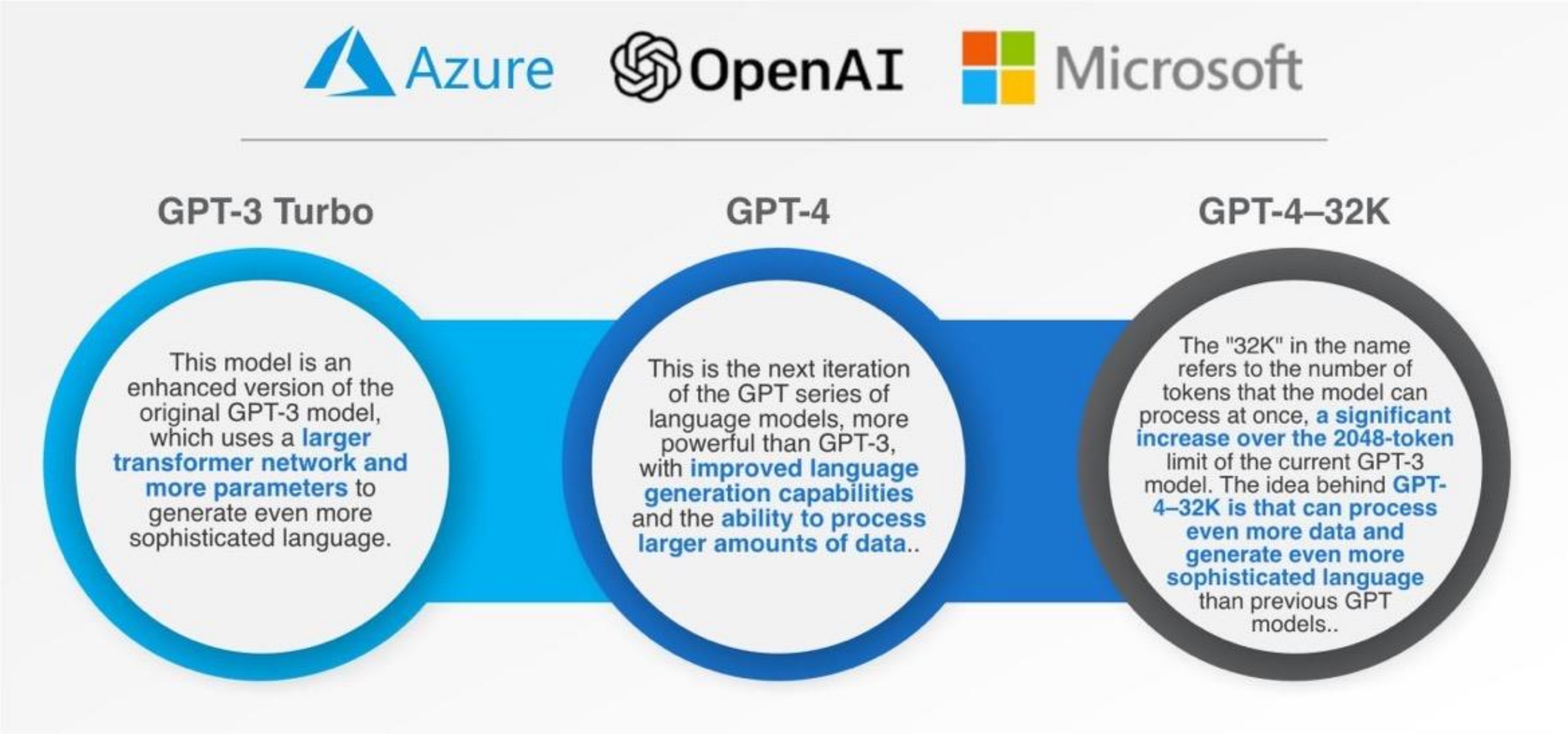


FLOWER
EAGLE
EAGLE
PLAYING SOCCER
Automatically find names of people, places, products and organizations in text across many languages.

DEPLOYMENTS	Tagging Services		Microsoft Dynamics 365 Spatial Analysis Services		Accessibility Services	Spatial Presenter	Azure Search, Video Indexer
TASKS	Classification	Object Detection	Object Tracking	Action Recognition	Entities	Topics	Sentiments
INDIVIDUAL MODEL (DISJOINTLY)	Classification Model	Detection Model	Tracking Model	Action Model	Entity Recognition	Topic Classification	Sentiment Analysis
TRAINING DATA (w/ ANNOTATION)	Tagging data	Detection data	Tracking data	Action data	Entity data	Topic data	Sentiment data

Foundation Models





Azure OpenAI Microsoft

GPT-3 Turbo

This model is an enhanced version of the original GPT-3 model, which uses a **larger transformer network and more parameters** to generate even more sophisticated language.

GPT-4

This is the next iteration of the GPT series of language models, more powerful than GPT-3, with **improved language generation capabilities** and the **ability to process larger amounts of data**..

GPT-4–32K

The "32K" in the name refers to the number of tokens that the model can process at once, **a significant increase over the 2048-token** limit of the current GPT-3 model. The idea behind **GPT-4–32K is that can process even more data and generate even more sophisticated language** than previous GPT models..

OpenAI

Ensure that artificial general intelligence (AGI) benefits humanity.

+

Microsoft

Empower every person and organization on the planet to achieve more

GPT-3

Generate and Understand Text

Codex

Generate and Understand Code

DALL·E

Generate images from text prompts

ChatGPT

Chat Bot, Avatar, Contact Center Reply

OpenAI Microsoft

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 =

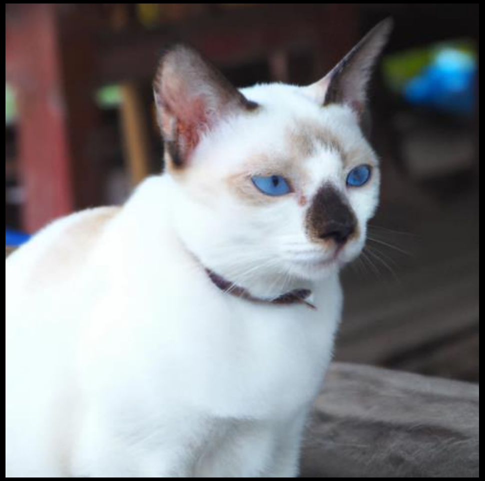
Response:

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

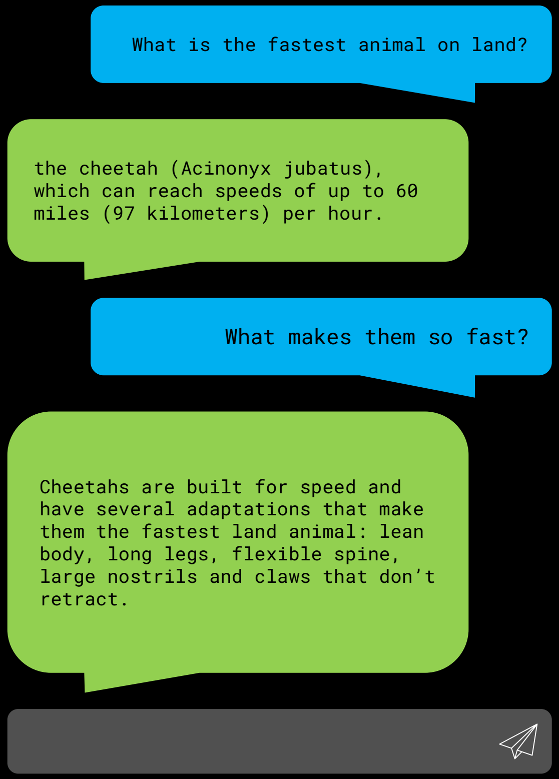
DALL·E

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.

Azure OpenAI Service

GPT-3

Codex

DALL·E

ChatGPT

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

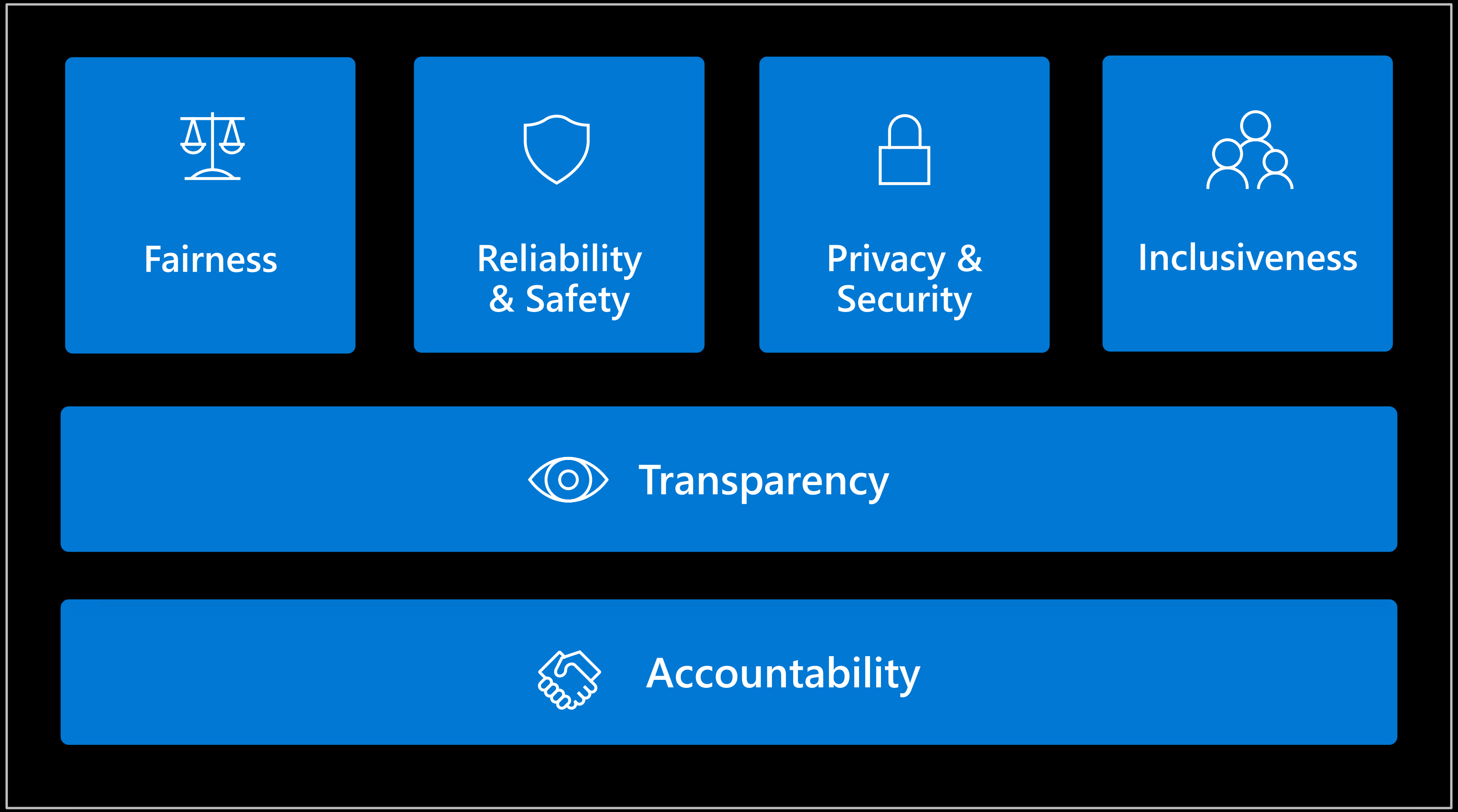
Large, pretrained AI models to unlock new scenarios

Custom AI models fine-tuned with your data and hyperparameters

Built-in responsible AI to detect and mitigate harmful use

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

Learn Microsoft's AI principles



Fairness

Reliability & Safety

Privacy & Security

Inclusiveness

Transparency

Accountability

Azure OpenAI | Overview of GPT-3

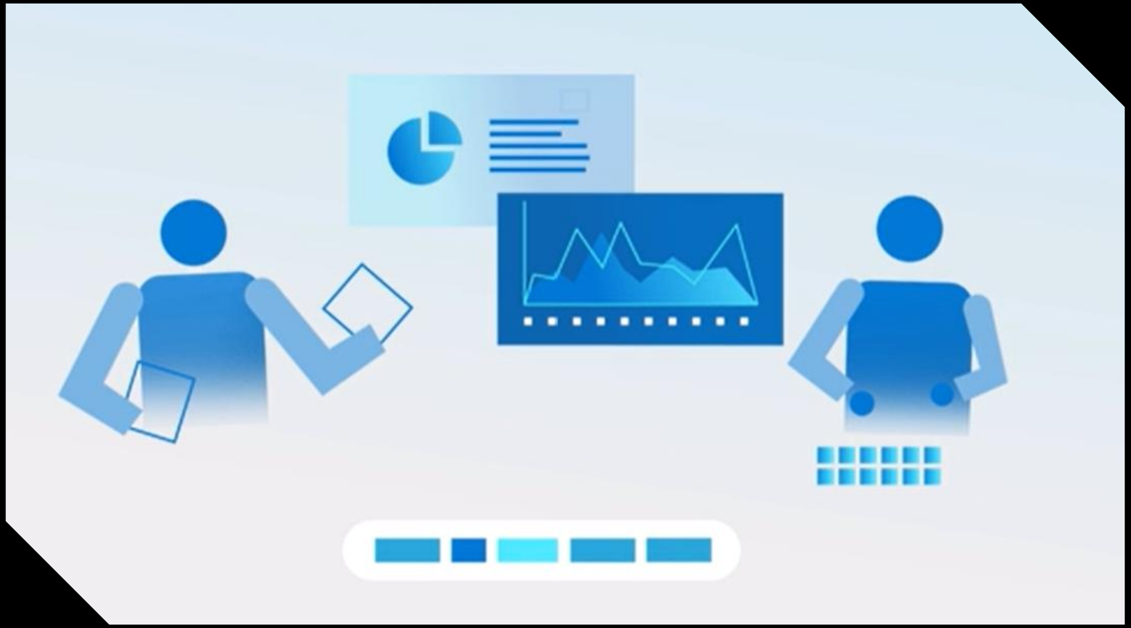
Generative pre-trained transformer 3 (GPT-3)

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

Pre-trained on trillions of words

Predicts the most likely next word based on input text

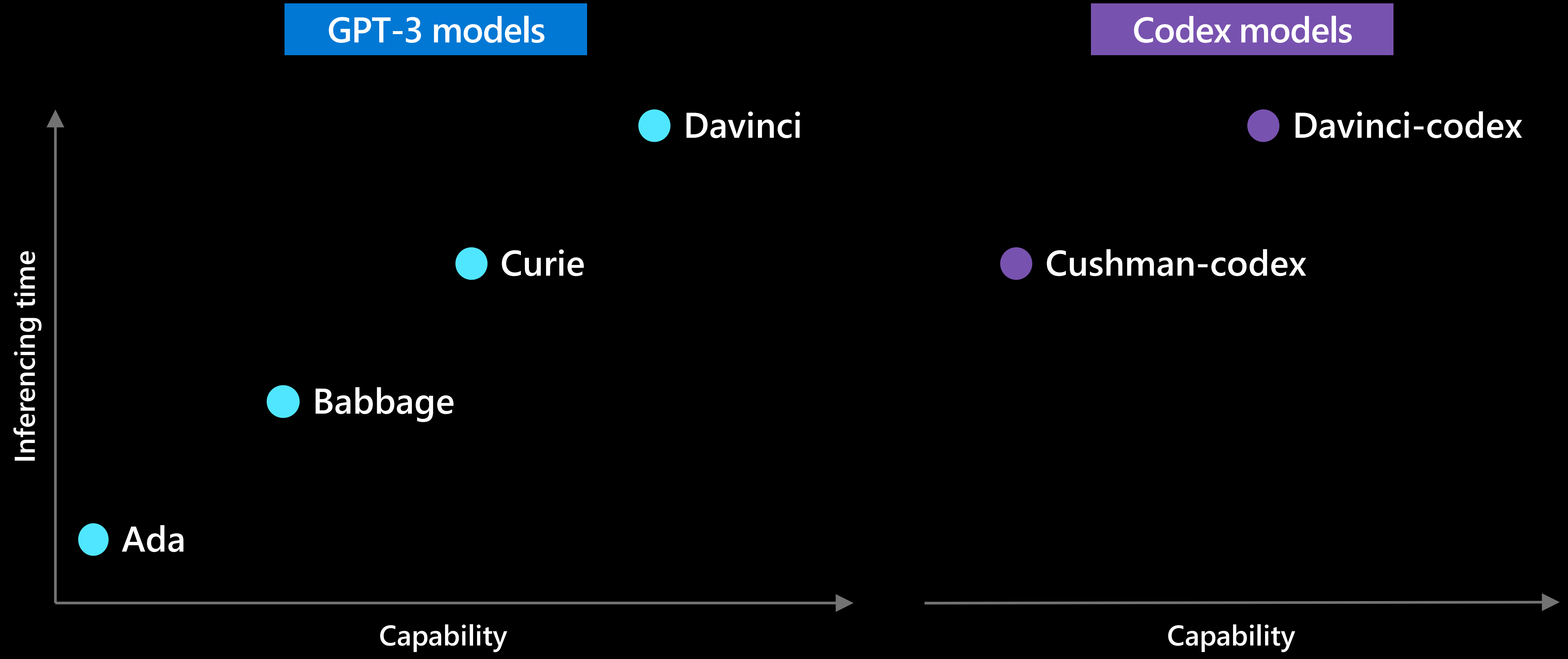
General text-in/text-out interface



Azure OpenAI | GPT-3 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 <u>lowest cost</u>.	Parsing text, simple classification, address correction, keywords

Azure OpenAI | Family of Models



Azure OpenAI | Comparing the GPT-3.5 models

Model	Description
text-davinci-002	A GPT-3.5 model that was fine-tuned on natural language instructions and can perform a variety of tasks including summarization, question answering, classification, and more.
text-davinci-003	An improvement over the text-davinci-002 model. The model is similar to its predecessor but generally more capable across all tasks.
ChatGPT model (gpt-3.5)	A model fine-tuned from text-davinci-002 that was optimized for working with dialogue. ChatGPT is a great model to use for conversational tasks and also excels at creative tasks.

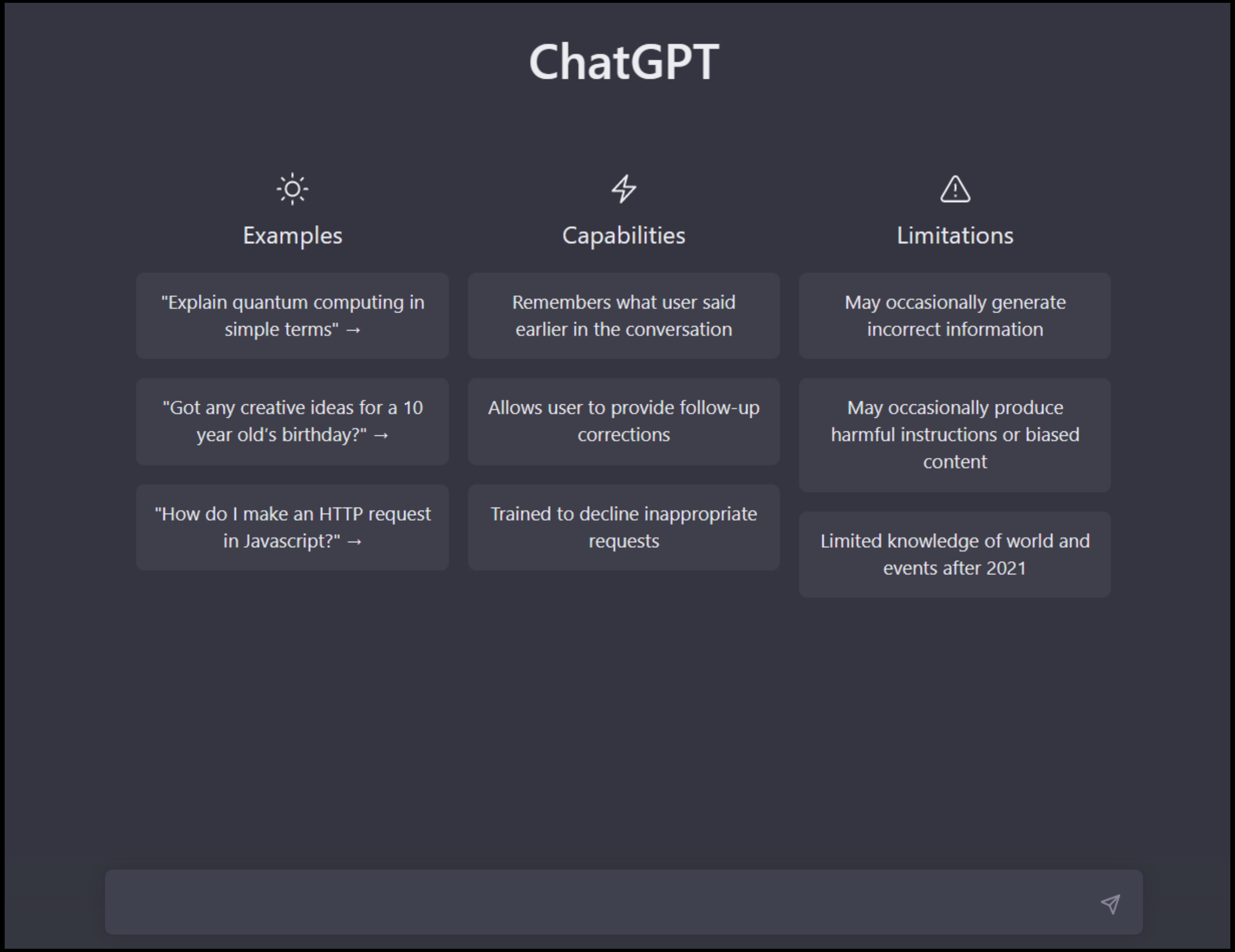
Visit the Azure OpenAI Service pricing page for pricing details

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

Working with the ChatGPT model

Previous GPT-3 models

Previous models were text-in and text-out

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

```
Answer questions from the context below.

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

Q: What is a neutron star?
A:
```

The ChatGPT model

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

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

```
<|im_start|>system
Assistant is an AI Chatbot designed to answer questions from the context provided below.

Context:
A neutron star is the collapsed core of a massive supergiant star, which had a total mass of between 10 and 25 solar masses, possibly more if the star was especially metal-rich.
<|im_end|>
<|im_start|>user
What is a neutron star?
<|im_end|>
<|im_start|>assistant
```

ChatGPT benefits

Conversational

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

Multi-turn

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

Creative

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

ChatGPT limitations

Hallucinations

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

Non-conversational tasks

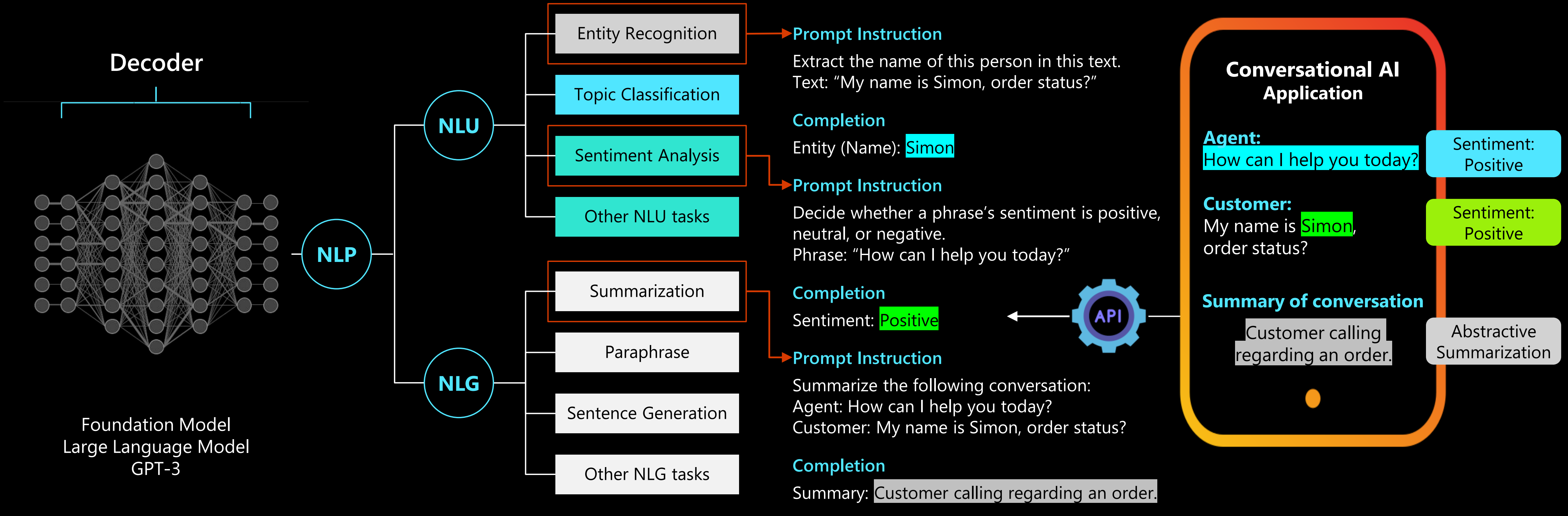
The ChatGPT model was optimized for conversational tasks. This means it might not perform as well on structured tasks like entity extraction, classification, etc. For more structured use cases, we recommend comparing ChatGPT with other models such as *text-davinci-003*.

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.

Model use out of the box—prompting



Prompt Instruction

Extract the name of this person in this text.
Text: "My name is Simon, order status?"

Completion

Entity (Name): Simon

Prompt Instruction

Decide whether a phrase's sentiment is positive, neutral, or negative.
Phrase: "How can I help you today?"

Completion

Sentiment: Positive

Prompt Instruction

Summarize the following conversation:
Agent: How can I help you today?
Customer: My name is Simon, order status?

Completion

Summary: Customer calling regarding an order.

Conversational AI Application

Agent:
How can I help you today?

Customer:
My name is Simon, order status?

Summary of conversation

Customer calling regarding an order.

Zero-Shot	One-Shot	Few-Shot
The model predicts the answer given only a natural language description of the task.	**In addition to the task description, the model sees a single example of the task**	**In addition to the task description, the model sees a few examples of the task.**

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?

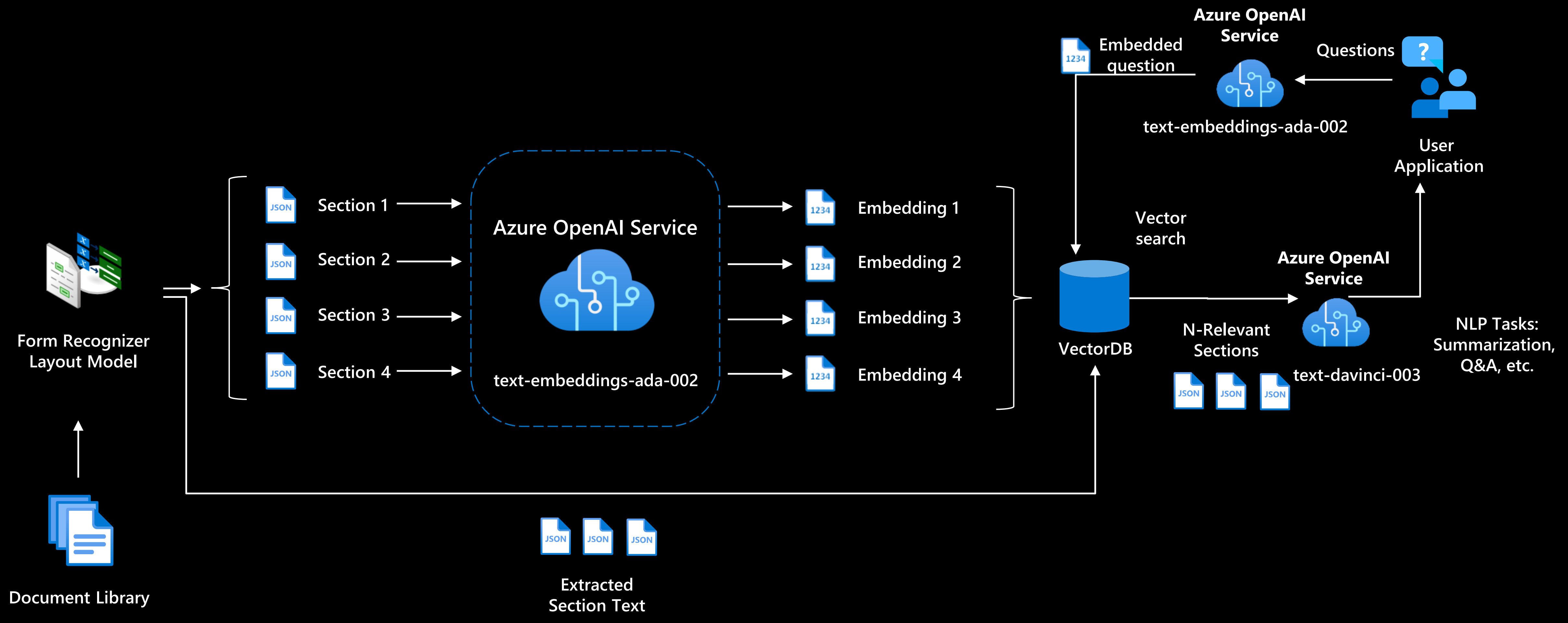
Answer: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5+6 = 11. The answer is 11.

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.

✓

Q&A with Semantic Answering over Document Library



Microsoft

Thank you