Microsoft

Azure OpenAI Academy
Session 2

Narjes Majdoub – nmajdoub@microsoft.com
Cloud Solution Architect

Franck Gaillard – franck.gaillard@microsoft.com
Cloud Solution Architect

26.04.2023

Agenda

1. Azure OpenAI Service Overview
2. Why Azure OpenAI vs OpenAI
3. Azure OpenAI models
4. Data privacy & Responsible AI
5. Model Adaptation
6. Expensya / Open AI Use Cases

Azure OpenAI Service Overview

Ensure that artificial general intelligence (AGI) benefits humanity

Empower every person and organization on the planet to achieve more

---

GPT-3.5 and GPT-4	ChatGPT	Codex	DALL·E 2
Text	Conversation	Code	Images

Generative AI

Content creation by API

Prompt

Write a tagline for an ice cream shop.

Response

We serve up smiles with every scoop!

Prompt

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

Response

There are a few things you can try to troubleshoot this issue … …

Prompt

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

Response

Here are a few games that you might consider: …

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'
```

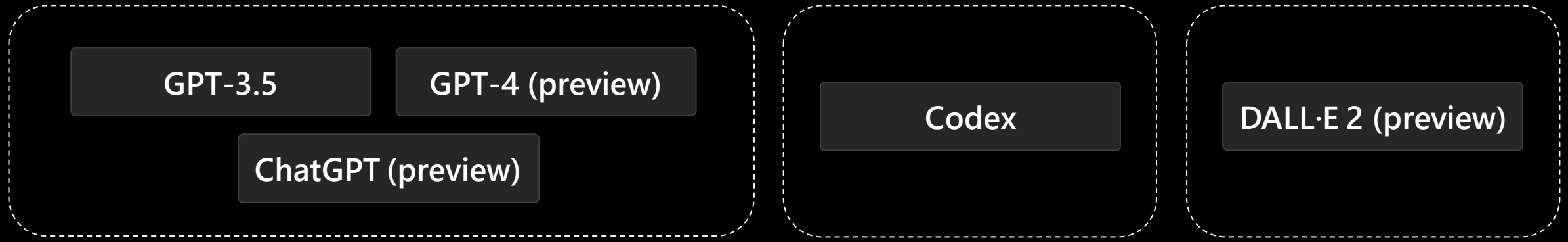
Prompt

A ball of fire with vibrant colors to show the speed of innovation at our media and entertainment company

Response

Azure OpenAI Service

Large pretrained foundation AI models
custom-tunable with your parameters and your data

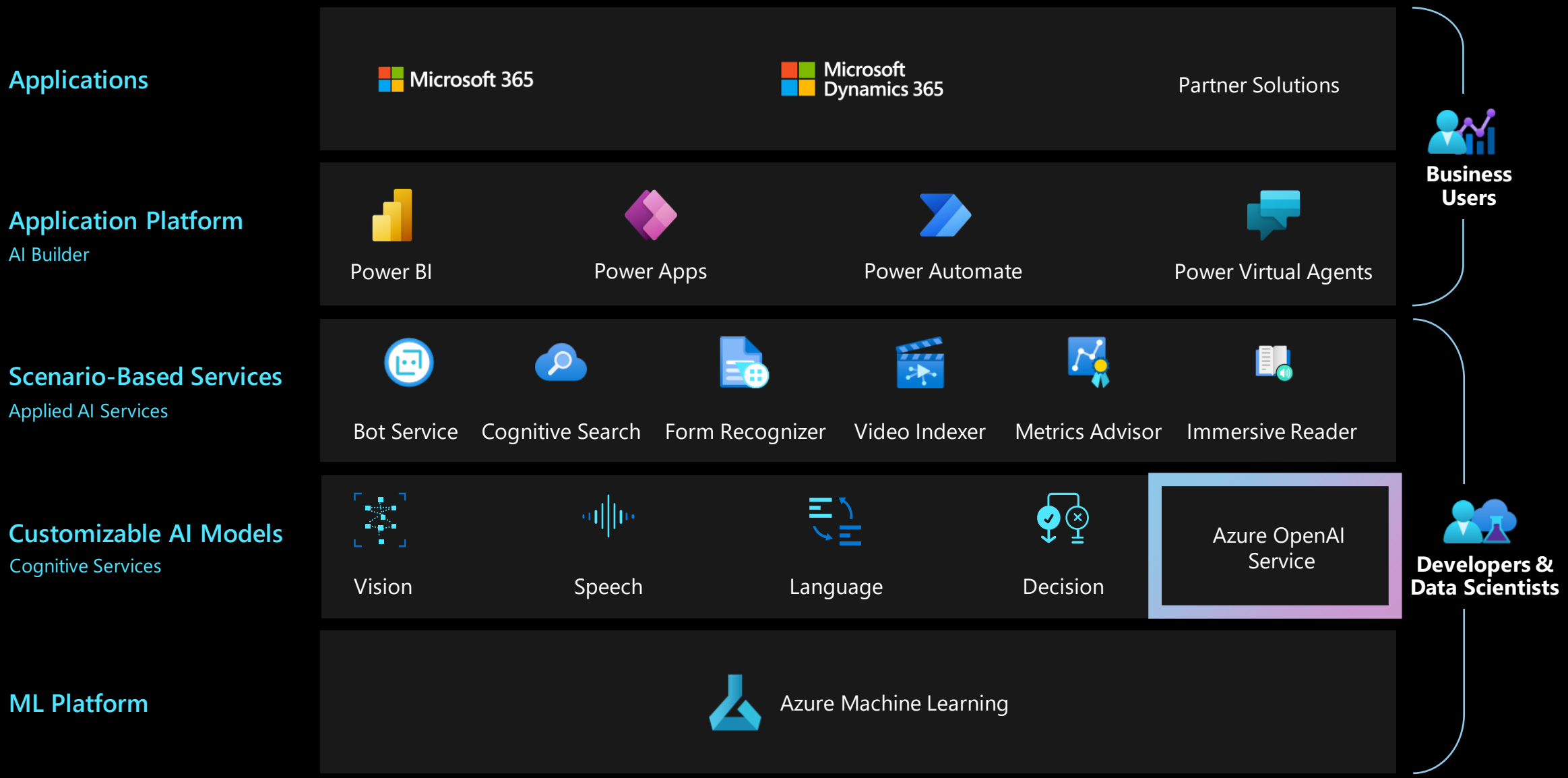


Generative Text Models, with varying capabilities and uses

Specialised Generative Coding Model

Generative Image Model

Azure AI

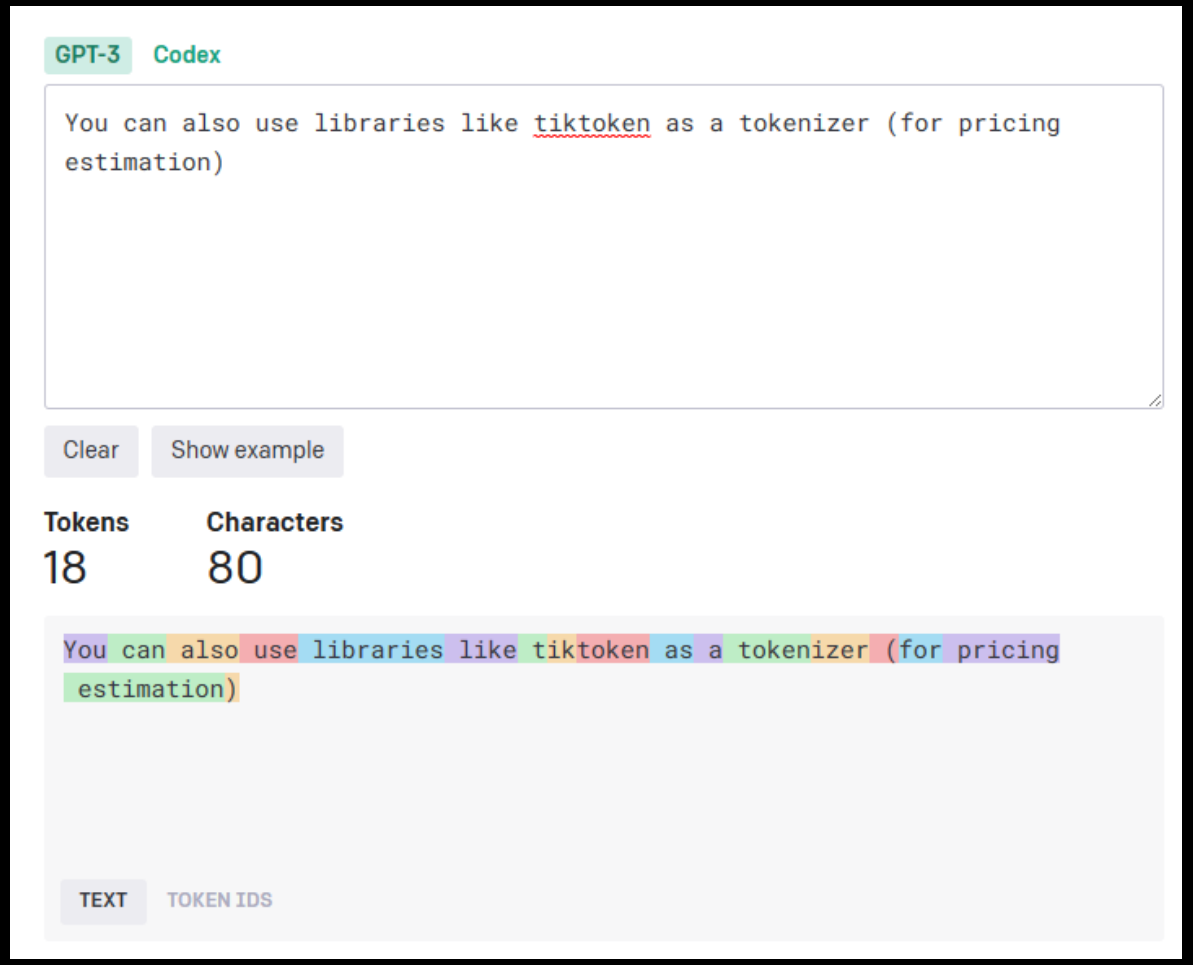

Applications
Microsoft 365
Microsoft Dynamics 365
Partner Solutions
Application Platform
AI Builder
Power BI
Power Apps
Power Automate
Power Virtual Agents
Business Users
Scenario-Based Services
Applied AI Services
Bot Service
Cognitive Search
Form Recognizer
Video Indexer
Metrics Advisor
Immersive Reader
Customizable AI Models
Cognitive Services
Vision
Speech
Language
Decision
Azure OpenAI Service
Developers & Data Scientists
ML Platform
Azure Machine Learning

Tokens

The GPT family of models process text using **tokens**, which are common sequences of characters found in text.

A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly ¾ of a word (so 100 tokens ~= 75 words).

[Tokenizer OpenAI]



Tokens - Pricing Estimation - Completion

Cost per completion = num_tokens (prompt) + max_tokens * max (n, best_of)

```
prompt = "Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large \
  corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method \
  still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans \
  can generally perform a new language task from only a few examples or from simple instructions - something which \
  current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves \
  task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning \
  approaches.\n\nTl;dr"

model = "text-curie-001"
response = openai.Completion.create(
  engine=model,
  prompt=prompt,
  temperature=0,
  max_tokens=60,
  n=1,
  best_of=1,
  stop=None)
```

Cost per completion = 148(prompt)+ 60 * max (n=1, best_of=2)
= 148 + 120 = 268

Azure OpenAI Service - Pricing | Microsoft Azure

Tokens - Pricing Estimation - Embeddings

```
[6]  1  df[['original_title', 'overview']]
```

✓ <1 sec

	original_title	overview
0	Hidden Figures	The untold story of Katherine G. Johnson, Dorothy Vaughan and Mary Jackson – brilliant African-American women working at NASA and serving as the brains behind one of the greatest operations in history – the launch of astronaut John Glenn into orbit. The visionary trio crossed all gender and race lines to inspire generations to dream big.
1	Gridlocked	Former SWAT leader David Hendrix and hard-partying movie star Brody Walker must cut their ride-along short when a police training facility is attacked by a team of mercenaries.
2	Joker	During the 1980s, a failed stand-up comedian is driven insane and turns to a life of crime and chaos in Gotham City while becoming an infamous psychopathic crime figure.
3	The Sand	Just when you thought it was safe to go back in the water again, you can't even get across the sand! BLOOD BEACH meets SPRING BREAKERS in an ace monster mash-up of smart nostalgia and up-to-the-minute visual effects. After an all-night graduation beach party, a group of hung-over students wake up under blazing sun to find their numbers somewhat depleted. An enormous alien creature has burrowed down deep and anyone foolish enough to make contact with the sand finds themselves at the mercy of a sea of flesh-eating tentacles. Will they ever be able to escape its carnivorous clutches?
4	America: The Motion Picture	A chainsaw-wielding George Washington teams with beer-loving bro Sam Adams to take down the Brits in a tongue-in-cheek riff on the American Revolution.
...	...	...
495	Lockout	Set in the near future, Lockout follows a falsely convicted ex-government agent , whose one chance at obtaining freedom lies in the dangerous mission of rescuing the President's daughter from rioting convicts at an outer space maximum security prison.
496	Black Panther	King T'Challa returns home to the reclusive, technologically advanced African nation of Wakanda to serve as his country's new leader. However, T'Challa soon finds that he is challenged for the throne by factions within his own country as well as without. Using powers reserved to Wakandan kings, T'Challa assumes the Black Panther mantle to join with ex-girlfriend Nakia, the queen-mother, his princess-kid sister, members of the Dora Milaje (the Wakandan 'special forces') and an American secret agent, to prevent Wakanda from being dragged into a world war.
497	Final Fantasy: The Spirits Within	Led by a strange dream, scientist Aki Ross struggles to collect the eight spirits in the hope of creating a force powerful enough to protect the planet. With the aid of the Deep Eyes Squadron and her mentor, Dr. Sid, Aki must save the Earth from its darkest hate and unleash the spirits within.
498	Boyhood	The film tells a story of a divorced couple trying to raise their young son. The story follows the boy for twelve years, from first grade at age 6 through 12th grade at age 17-18, and examines his relationship with his parents as he grows.
499	Last Man Standing	John Smith is a mysterious stranger who is drawn into a vicious war between two Prohibition-era gangs. In a dangerous game, he switches allegiances from one to another, offering his services to the highest bidder. As the death toll mounts, Smith takes the law into his own hands in a deadly race to stay alive.

```python
# add a new column to the dataframe where you put the token count of the review
df = df.assign(token_count=df['overview'].apply(lambda x: len(encoding.encode(x))))

# print the first 5 rows of the dataframe, then also the total number of tokens
total_tokens = df['token_count'].sum()

cost_for_embeddings = total_tokens / 1000 * 0.0004
print(f"Test would cost ${cost_for_embeddings} for embeddings")
```

✓

```
Test would cost $0.0117564 for embeddings
```

Demo:

Azure OpenAI playground tour

Why Azure OpenAI vs OpenAI

Azure OpenAI Service

Why do customers choose Azure OpenAI?

Security, Compliance & Governance

Simplified Billing

Fine Tuning & Inference

Content Filtering & Moderation

GPT-3 models (language):
Ada, Babbage, Curie, Davinci

Future Models

	Azure OpenAI	OpenAI
Security, Compliance & Governance	✓	●
Simplified Billing	✓	●
Fine Tuning & Inference	✓	✓
Content Filtering & Moderation	✓	●
GPT-3 models (language): Ada, Babbage, Curie, Davinci / Future Models	✓	✓

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, HIPAA, CSA STAR Compliant
Responsible AI built-in	Fairness, Explainability, Counterfactuals, Casual Analysis, Error Analysis, and Responsible AI Scorecard

Regions Availability

GPT-3

Azure OpenAI Service models - Azure OpenAI | Microsoft Learn

Introduction to Azure OpenAI models

Azure OpenAI Models

GPT-4

- A set of models that improve on GPT-3.5 and can understand as well as generate natural language and code. **These models are currently in preview.** GPT-4 is a large multimodal model that receives image and text inputs and then outputs correct text responses.

GPT-3

- A series of models that can understand and generate natural language.

Codex

- A series of models that can understand and generate code, including translating natural language to code.

Embeddings

- A set of models that can understand and use embeddings.
- The embedding is an information dense representation of the semantic meaning of a piece of text.

DALL E

- DALL·E is an AI system that can create realistic images and art from a description in natural language.

Model Capabilities – GPT-4

gpt-4 [1,2] (preview)
gpt-4-32k [1,2] (preview)

- "GPT-4 is a large multimodal model (**accepting image and 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 (openai.com)]

User What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

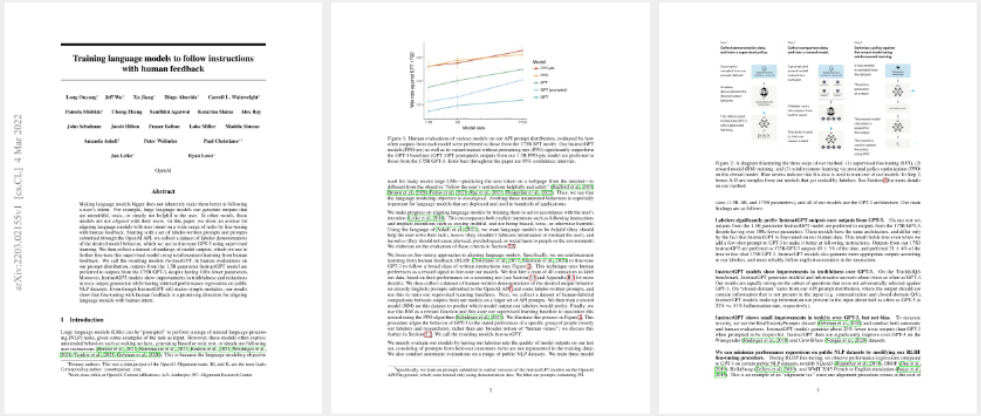
Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

User Below is part of the InstuctGPT paper. Could you read and summarize it to me?

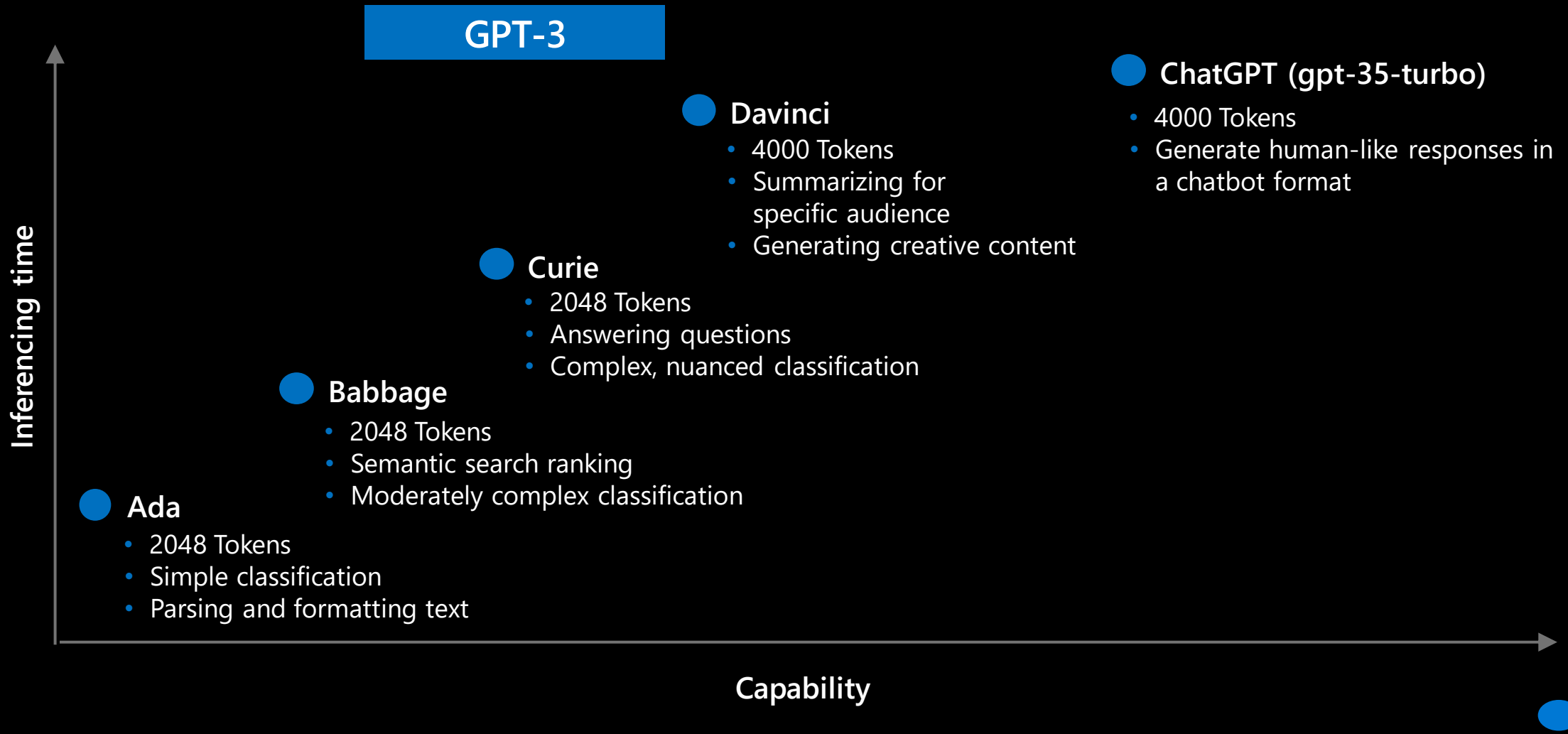


GPT-4 The InstructGPT paper focuses on training large language models to follow instructions with human feedback. The authors note that making language models larger doesn't inherently make them better at following a user's intent. Large models can generate outputs that are untruthful, toxic, or simply unhelpful.

To address this issue, the authors fine-tune language models on a wide range of tasks using human feedback. They start with a set of labeler-written prompts and responses, then collect a dataset of labeler demonstrations of the desired model behavior. They fine-tune GPT-3 using supervised learning and then use reinforcement learning from human feedback to further fine-tune the model. The resulting model, called InstructGPT, shows improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets.

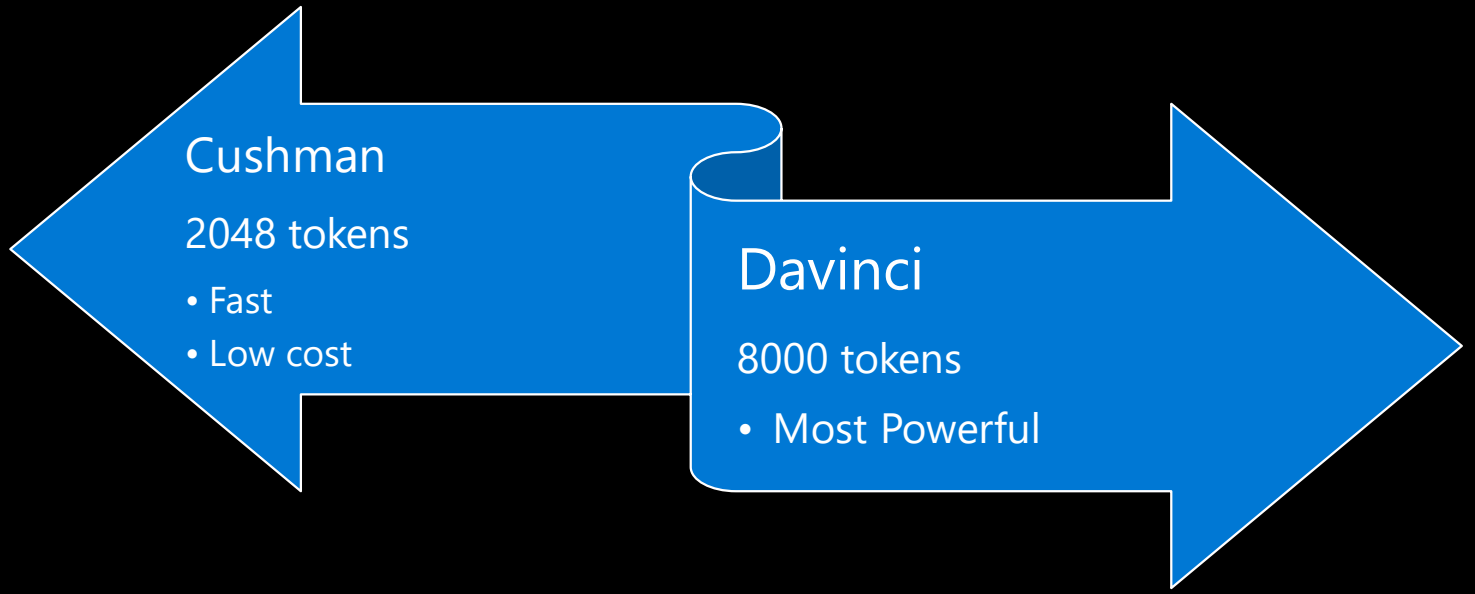
The authors conclude that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

Model Capabilities – GPT-3



Model Capabilities – Codex

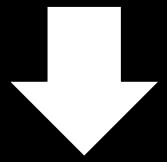
- Descendants of our base GPT-3 models
- Understand and generate code
- Most capable in Python, efficient in C#, JavaScript, Go, Perl, PHP, Ruby, Swift, TypeScript, SQL, and even Shell



Embeddings

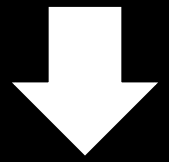
Embeddings make it possible to map content to a “semantic space”

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



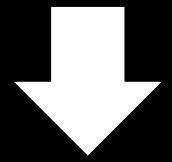
[15 34 24 13 …]

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



[16 22 89 26 …]

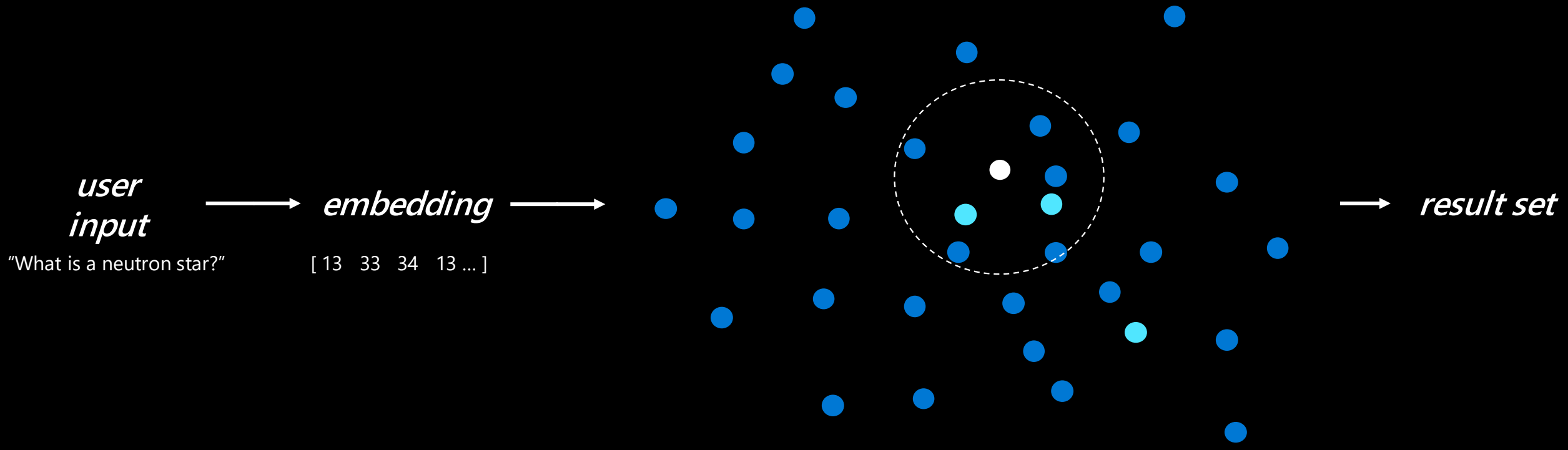
The presence of a black hole can be inferred through its interaction with other matter



[20 13 31 89 …]

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.



Azure OpenAI embeddings tutorial - Azure OpenAI | Microsoft Learn

Model Capabilities – Embeddings models

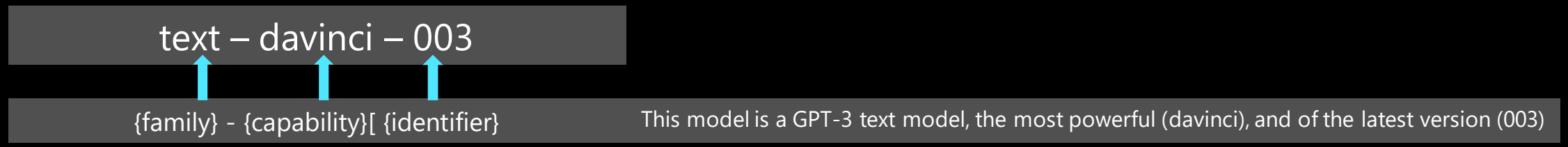
Text Similarity	Capturing semantic similarity between two or more pieces of text	Clustering, regression, anomaly detection, visualization
Text Search	Measure long documents are relevant to a short query	Search, context relevance, information retrieval
Code Search	Useful for embedding code snippets and embedding nature language search queries	Code search and relevance

Naming Convention

Azure OpenAI's model names typically correspond to the following standard naming convention:

{family} - {capability} [-{input-type}] - {identifier}

Element	Description
{family}	The model family of the model. For example, GPT-3 models uses text, while Codex models use code.
{capability}	The relative capability of the model. For example, GPT-3 models include ada, babbage, curie, and davinci.
{input-type}	(Embeddings models only) The input type of the embedding supported by the model. For example, text search embedding models support doc and query.
{identifier}	The version identifier of the model.

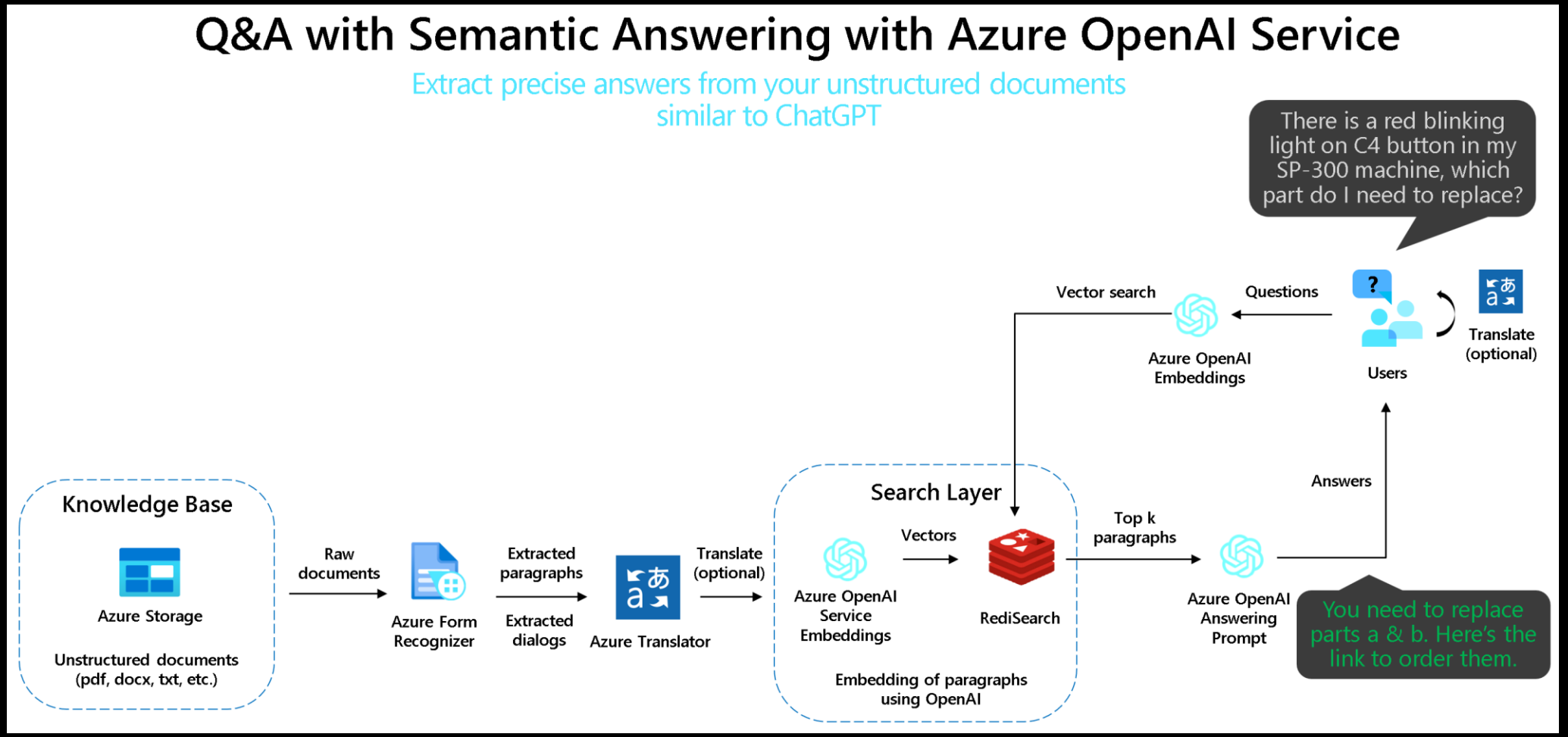


*https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/models

Demo:

Q&A semantic answering with Azure OpenAI

OpenAI Embeddings & Semantic Search

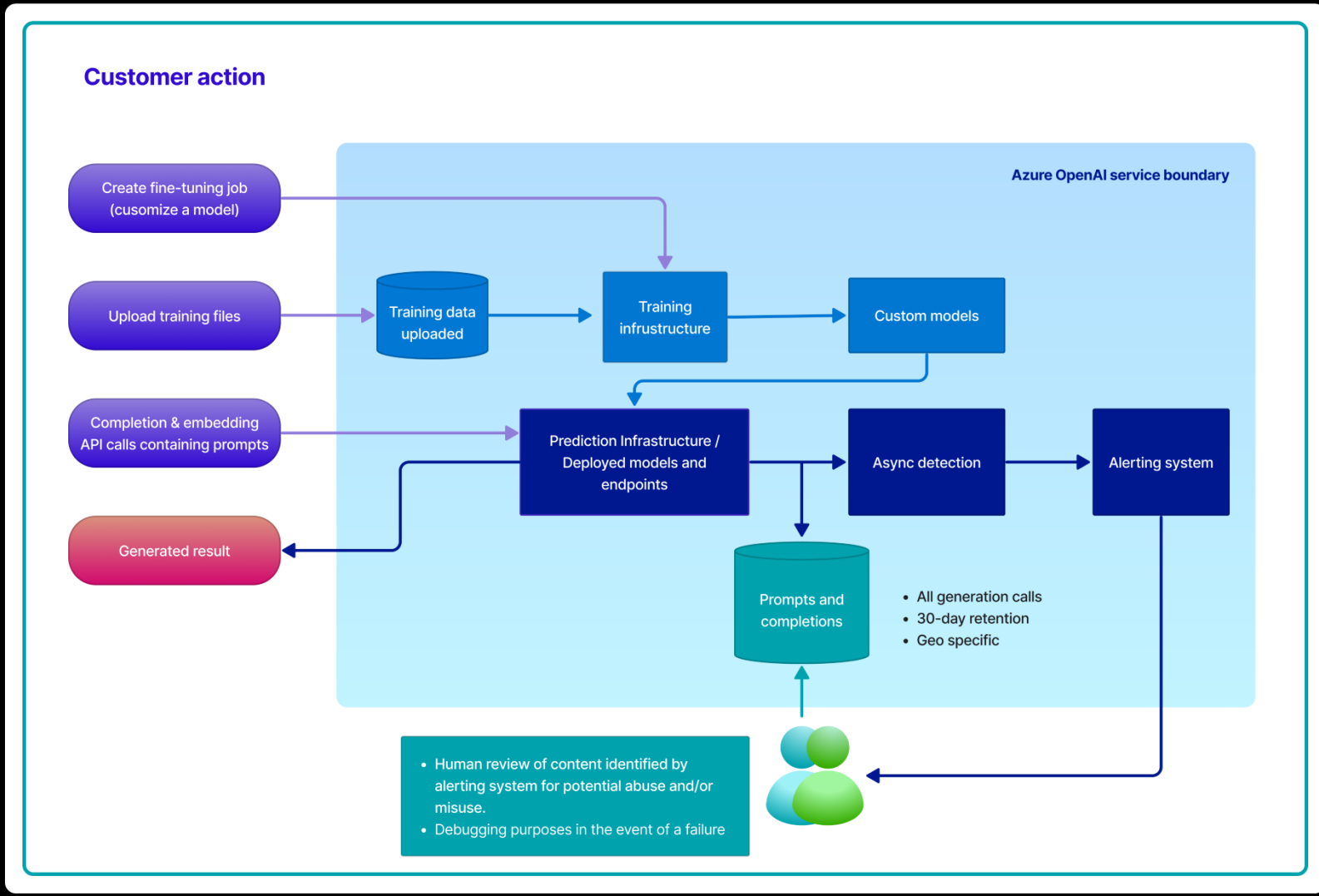


ruoccofabrizio/azure-open-ai-embeddings-qna

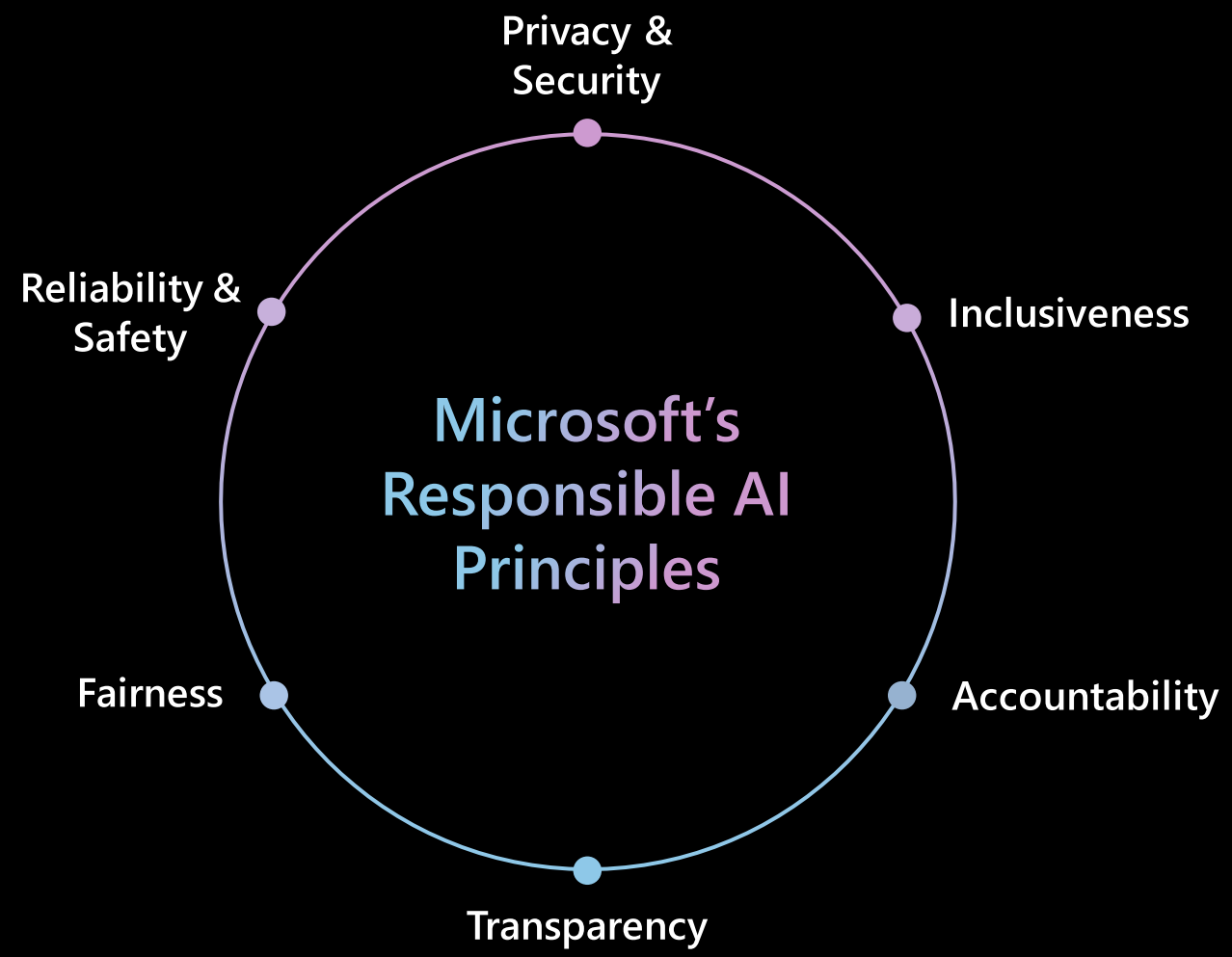
Microsoft Azure

Data privacy & Responsible AI

How does the AOAI service process data?



Data, privacy, and security for Azure OpenAI Service - Azure Cognitive Services | Microsoft Learn



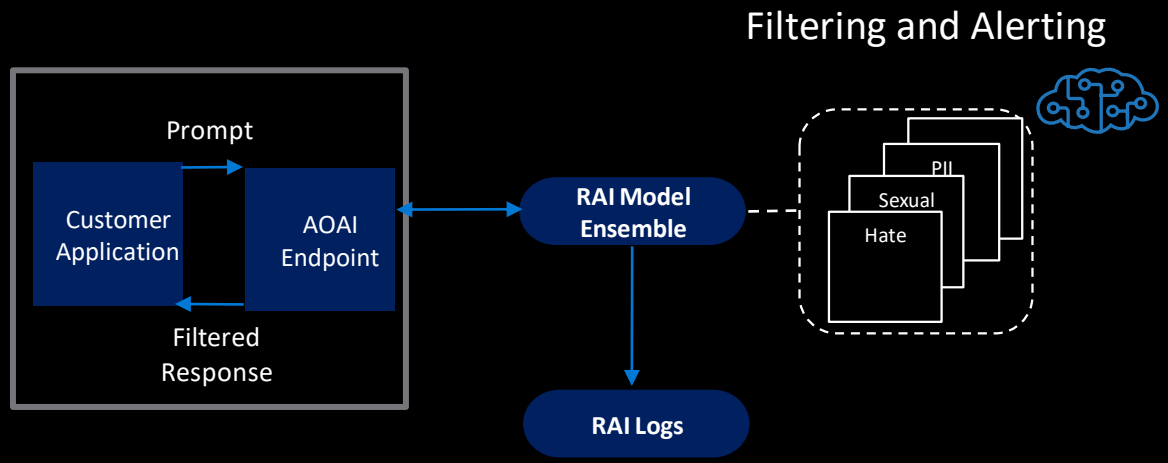
Building blocks to enact principles

- Tools and processes
- Training and practices
- Rules
- Governance

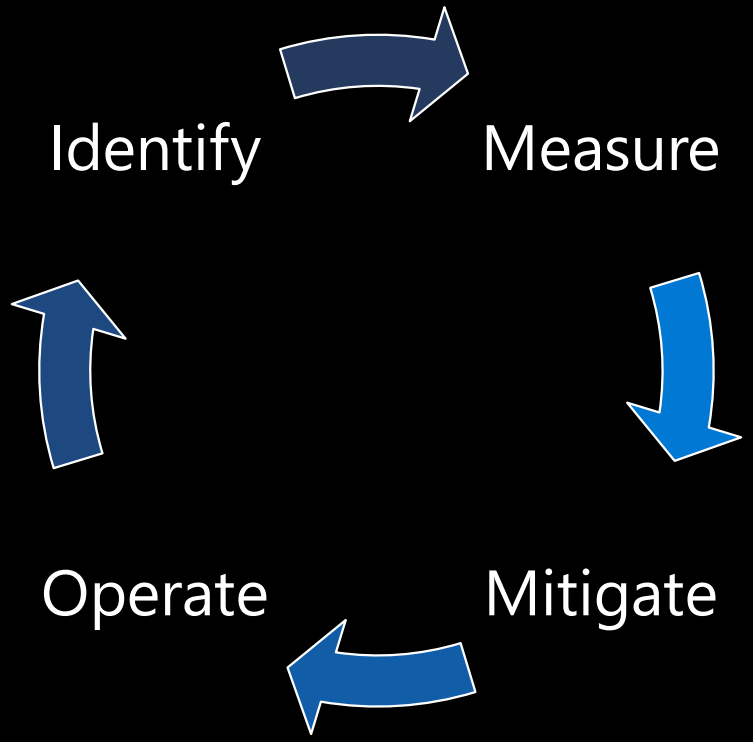
[Transparency Note for Azure OpenAI]

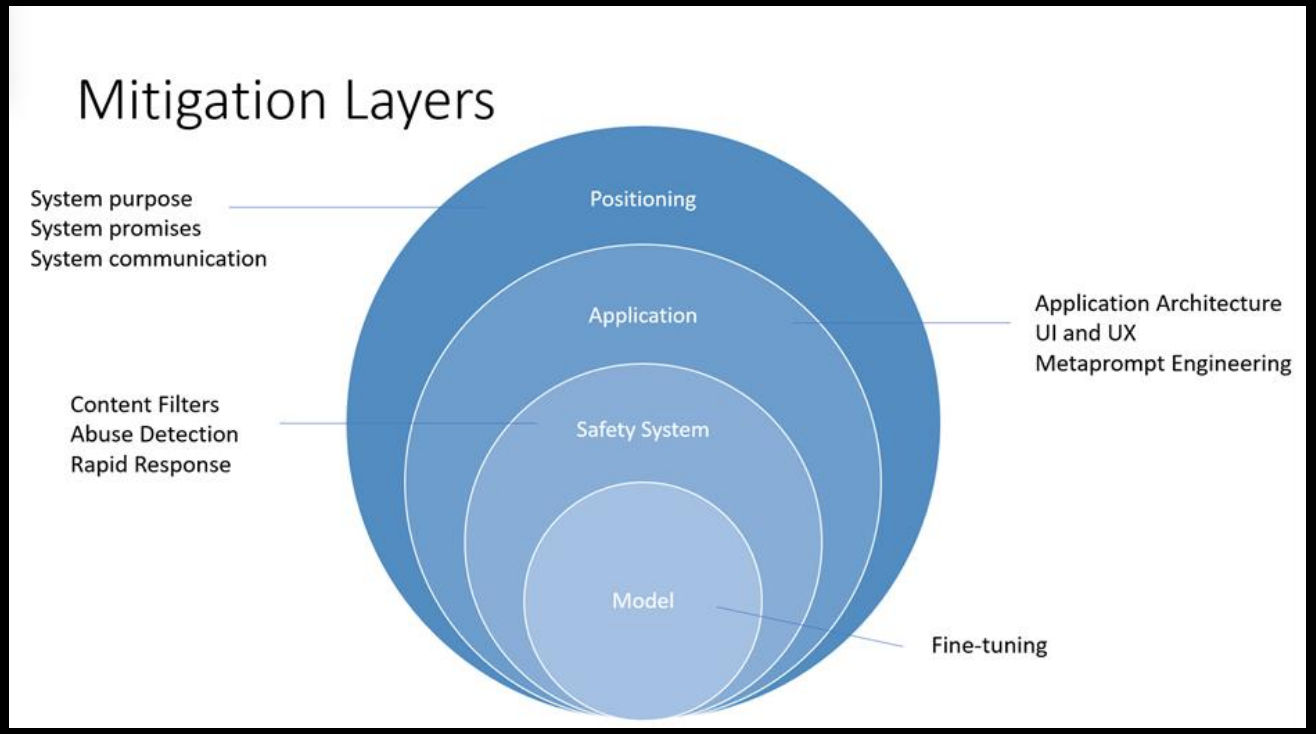
Azure Responsible AI

- Determine appropriate use cases
- Content filters and abuse detection
- Implementation guidelines, patterns, and best practices



Evaluating and integrating Azure OpenAI for your use





Model Adaptation

GPT-3 Ideate, Experiment and Fine-Tune

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

Prompt

Summarize game commentary into highlights:

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

What are the main highlights of the game so far?

Sample response

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

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

Prompt

Turn game commentary into highlights:

Commentary: What a pickup she has

Main highlights: New York has domina

###

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

Examples

Inputs

Sample response

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

Results

Optimize accuracy and latency to validate proof of concept fast

Prompt and completion examples

Fine-tuning

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

Azure OpenAI Service

How to adapt GPT-3 model for your task

No Gradient Updates

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.

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 fined tuned model → Use your fine-tuned model

1. Higher quality results than prompt design

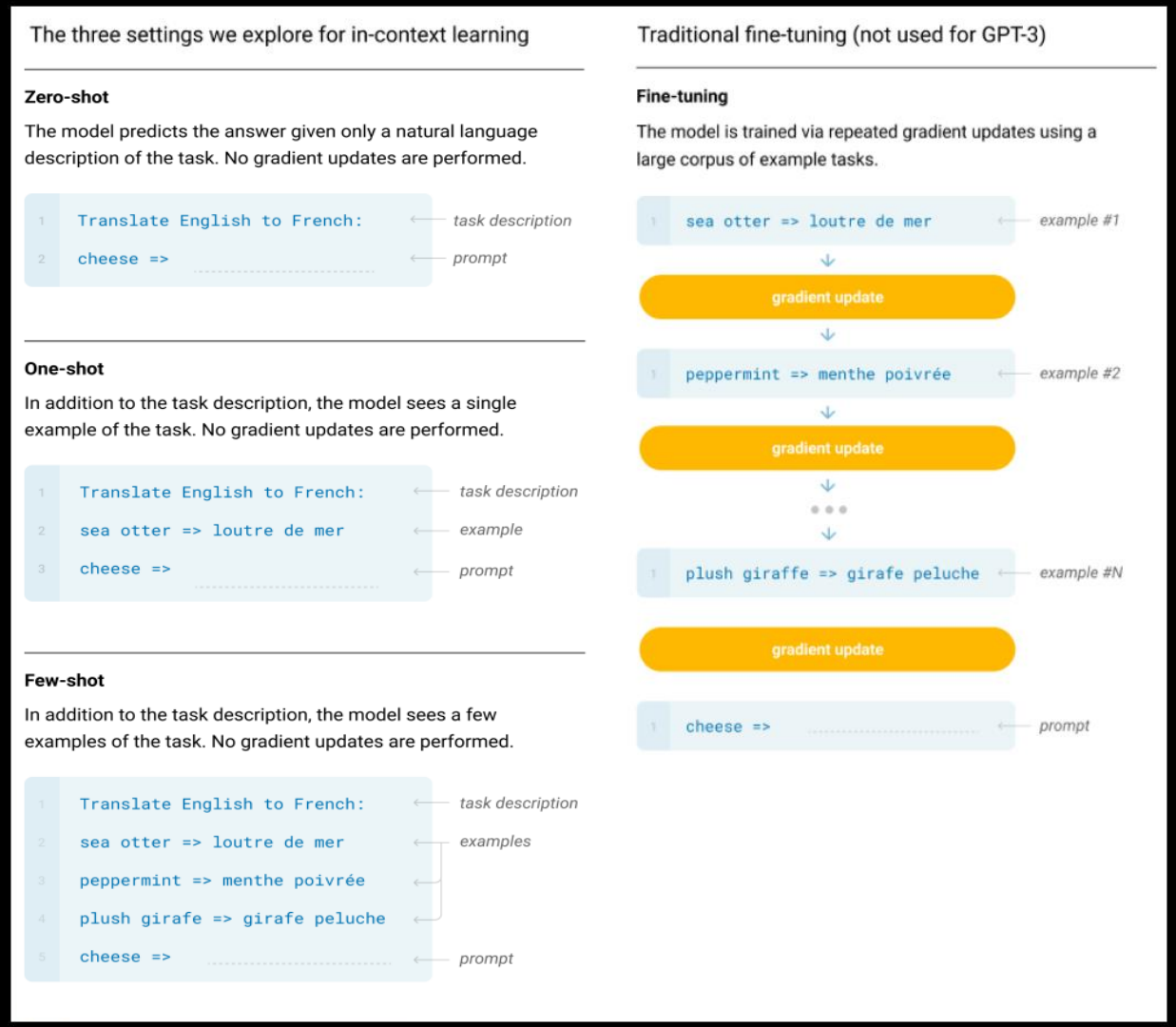
2. Ability to train on more examples than can fit in a prompt

3. Token savings due to shorter prompts

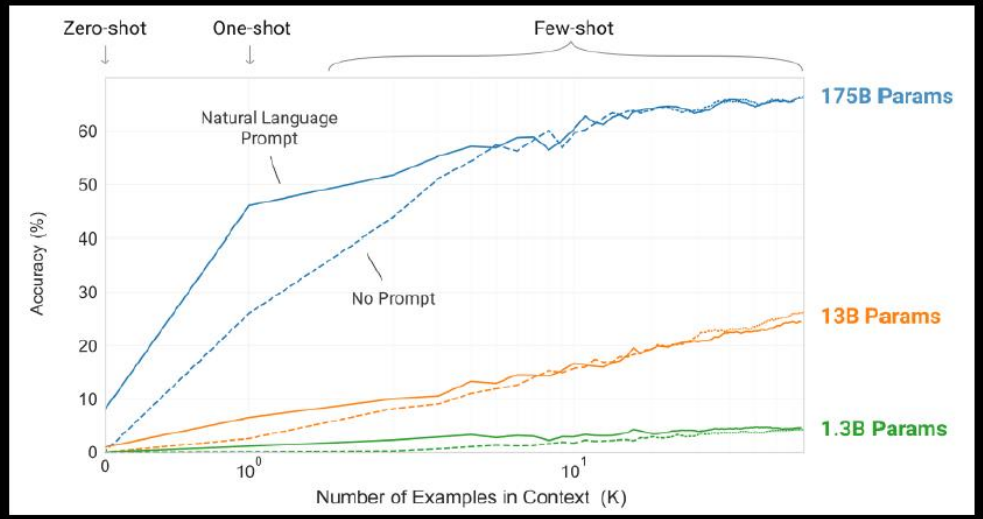
4. Lower latency requests

GPT focuses on zero-shot, one-shot and few-shot

GPT-3 focuses on zero-shot, one-shot and few-shot



Larger models make increasingly efficient use of in-context information



These "learning" curves involve no gradient updates or fine-tuning, just increasing numbers of demonstrations given as conditioning

Finetuning – Training Data

Training data is how you teach GPT-3 what you'd like it to say.

Your data must be a JSONL document, where each line is a prompt-completion pair corresponding to a training example.

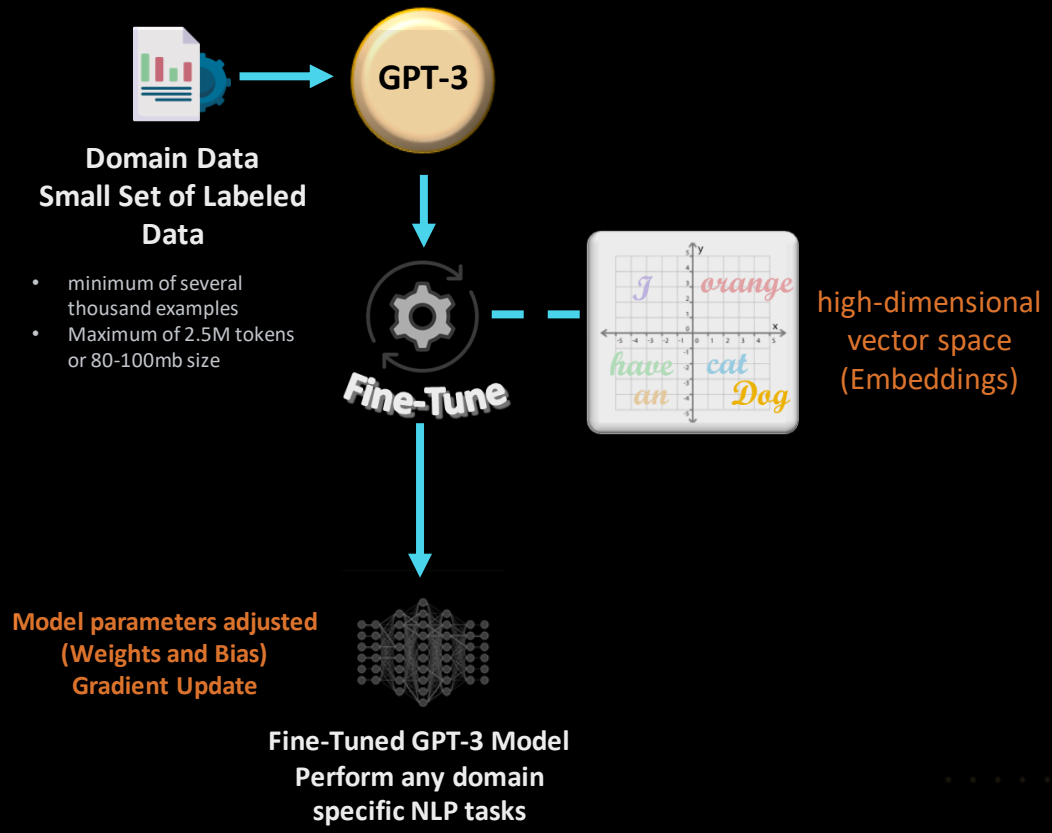
To fine-tune a model, you'll need a set of training examples that each consist of a **single input ("prompt")** and its **associated output ("completion").** This is notably different from using our base models, where you might input detailed instructions or multiple examples in a single prompt.

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

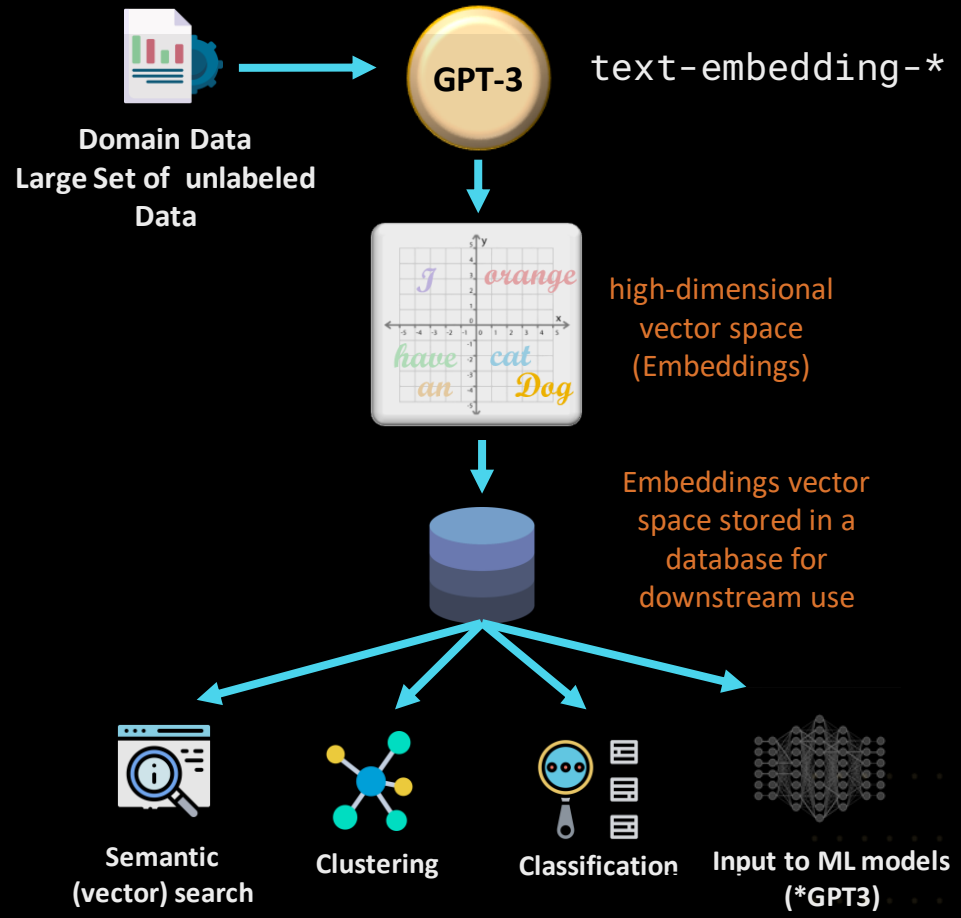
Microsoft Azure

Model Adaptation with specific domain data

Fine-Tuning



Embedding



Expensya / Open AI Use Cases