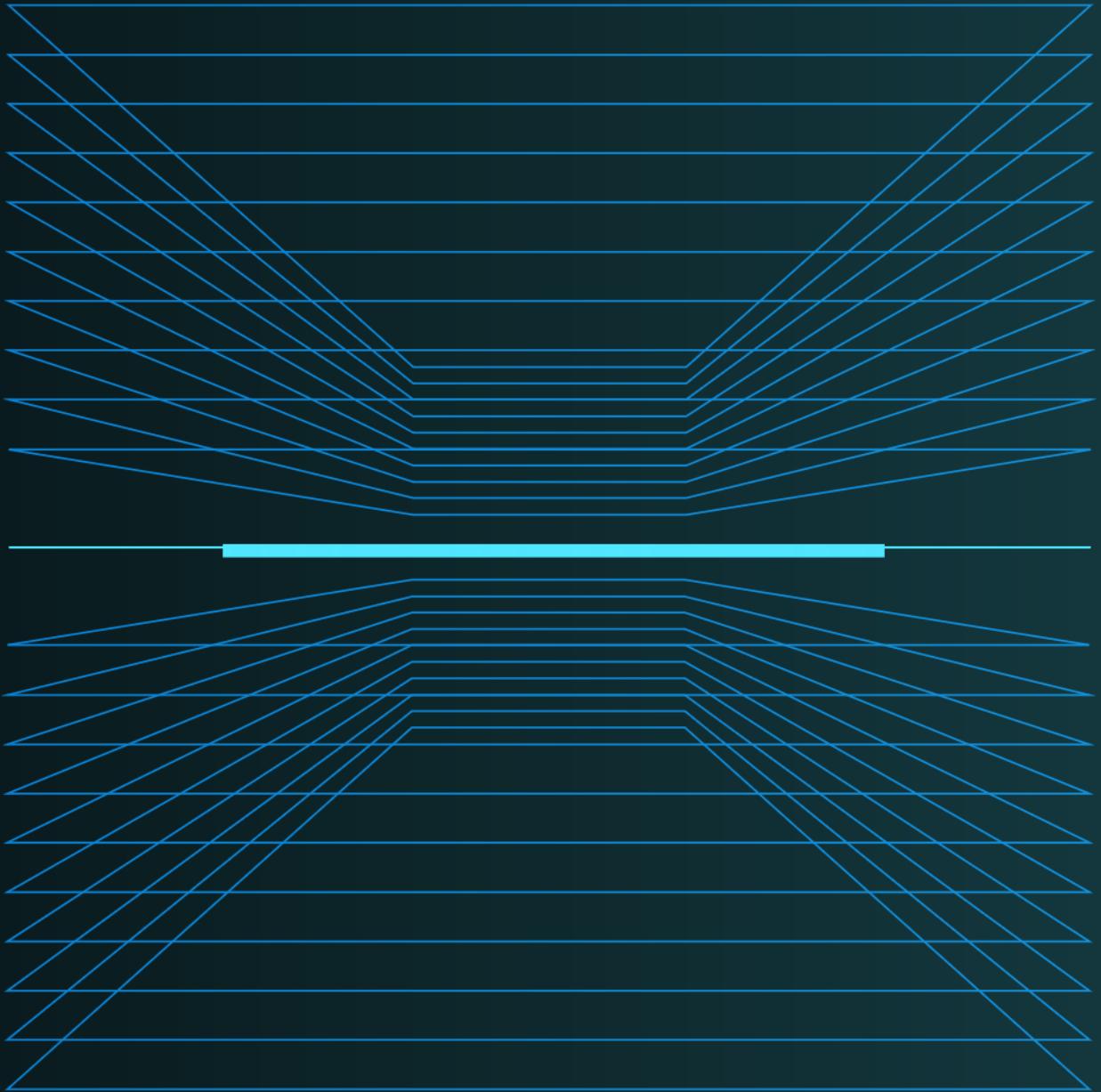




Azure OpenAI Workshop for Product Teams

GPS ISV Team



Agenda

10:00 am – 10:20 am	Keynote
10:20 am – 10:55 am	Opportunities of LLM targeted for product development teams
10:55 am – 11:00 am	Bio Break
11:00 am – 11:35 am	Partner showcases
11:35 am – 12:00 am	Code of Conduct for responsible AI
12:00 pm – 01:00 pm	Lunch
01:00 pm – 01:15 pm	Fundamentals of Azure OpenAi
01:15 pm – 01:40 pm	Best Practices for Prompt Engineering
01:40 pm – 02:00 pm	Embedding & FineTuning
02:00 pm – 02:05 pm	Bio Break
02:05 pm – 02:30 pm	Orchestration (Langchain & Semantic Kernels)
02:30 pm – 03:00 pm	Live Coding Demo
03:00 pm – 03:10 pm	AAAP – Microsoft Incentive Campaign for customers PoCs
03:10 pm – 03:30 pm	Q&A

The world is changing

Industries are transforming rapidly

The drivers for economic growth are evolving

The world is more connected than ever

The AI technology is here

Forbes

What ChatGPT And Generative AI Mean For Your Business?

COMPUTERWORLD

Microsoft's new Teams Premium tier integrates with OpenAI's GPT-3.5

MARKETS INSIDER

Nuance and Microsoft Announce the First Fully AI-Automated Clinical Documentation Application for Healthcare

VentureBeat

Microsoft gives Businesses a GPT boost In Teams and Viva Sales

TheVerge

ChatGPT is now available in Microsoft's Azure OpenAI service

USA TODAY

New Bing with ChatGPT brings the power of AI to Microsoft's signature search engine

VentureBeat

Microsoft announces generative AI-powered Copilot 365 to 'change work as we know it'

CNN BUSINESS.

Real estate agents say they can't imagine working without ChatGPT now

TechCrunch

Microsoft brings an AI-powered Copilot to its business app suite

And the impact is real



Carmax estimates an individual would take **11 years** to do what Azure OpenAI Service was able to do in **days**



Progressive is saving **\$10M** annually with AI-powered chatbots



EY is saving **250K hours** of manual work per client using intelligent document automation

The opportunity is yours to
lead the AI transformation



Microsoft
Cloud



Digital
Transformation



AI
Transformation

Artificial Intelligence

Machine Learning

Deep Learning

Generative AI



Artificial Intelligence

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



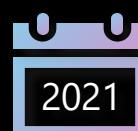
Machine Learning

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



Deep Learning

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



Generative AI

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

Modern apps: born in the cloud and intelligent

95% of all new digital workloads will deploy to cloud by 2025¹

85% of organizations will embrace cloud-first principals by 2025²

90% of enterprise apps will embed AI by 2025³



1. [Gartner press release](#)

2. ibid

3. IDC report



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

Text

ChatGPT

Conversation

Codex

Code

DALL·E 2

Images

GPT-3.5 and GPT-4

Text

ChatGPT

Conversation



DALL·E 2

Images

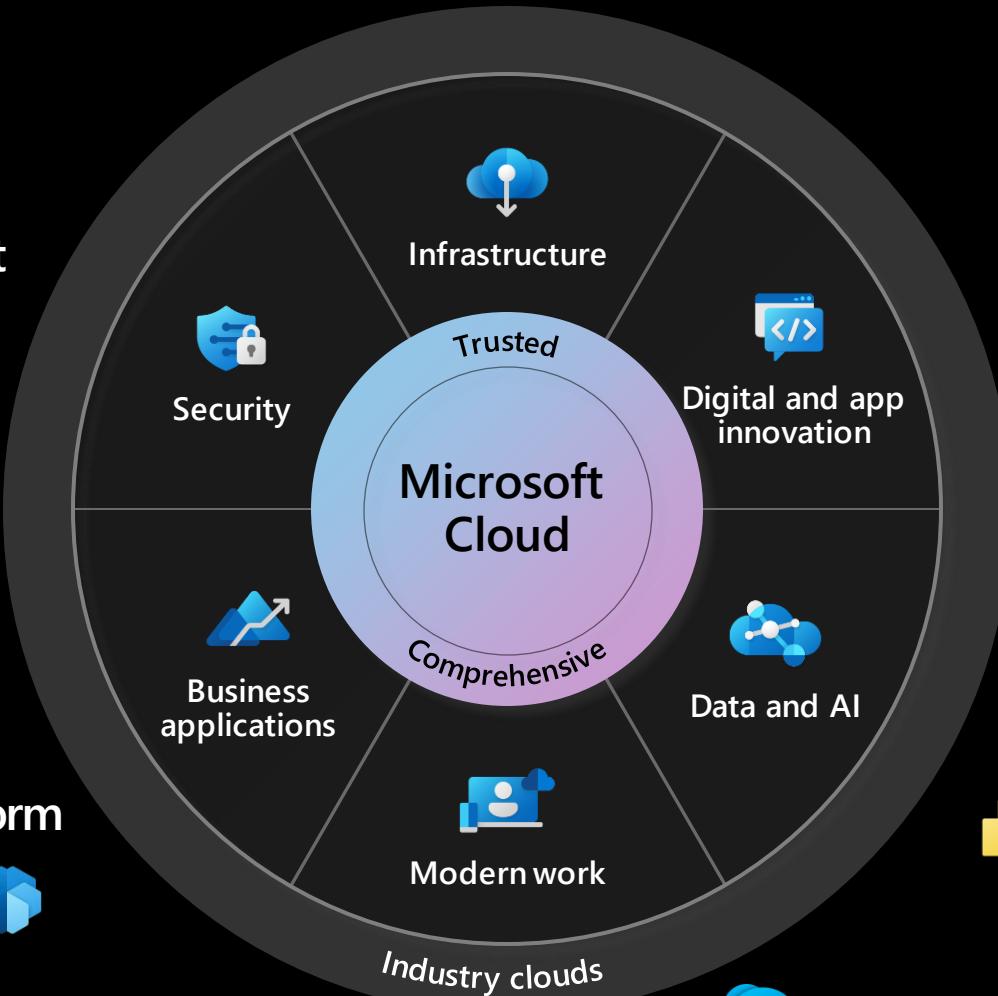


Codex

Code



Innovation powered by OpenAI models



Microsoft Security Copilot

Microsoft Dynamics 365 Copilot

NUANCE

Power Platform

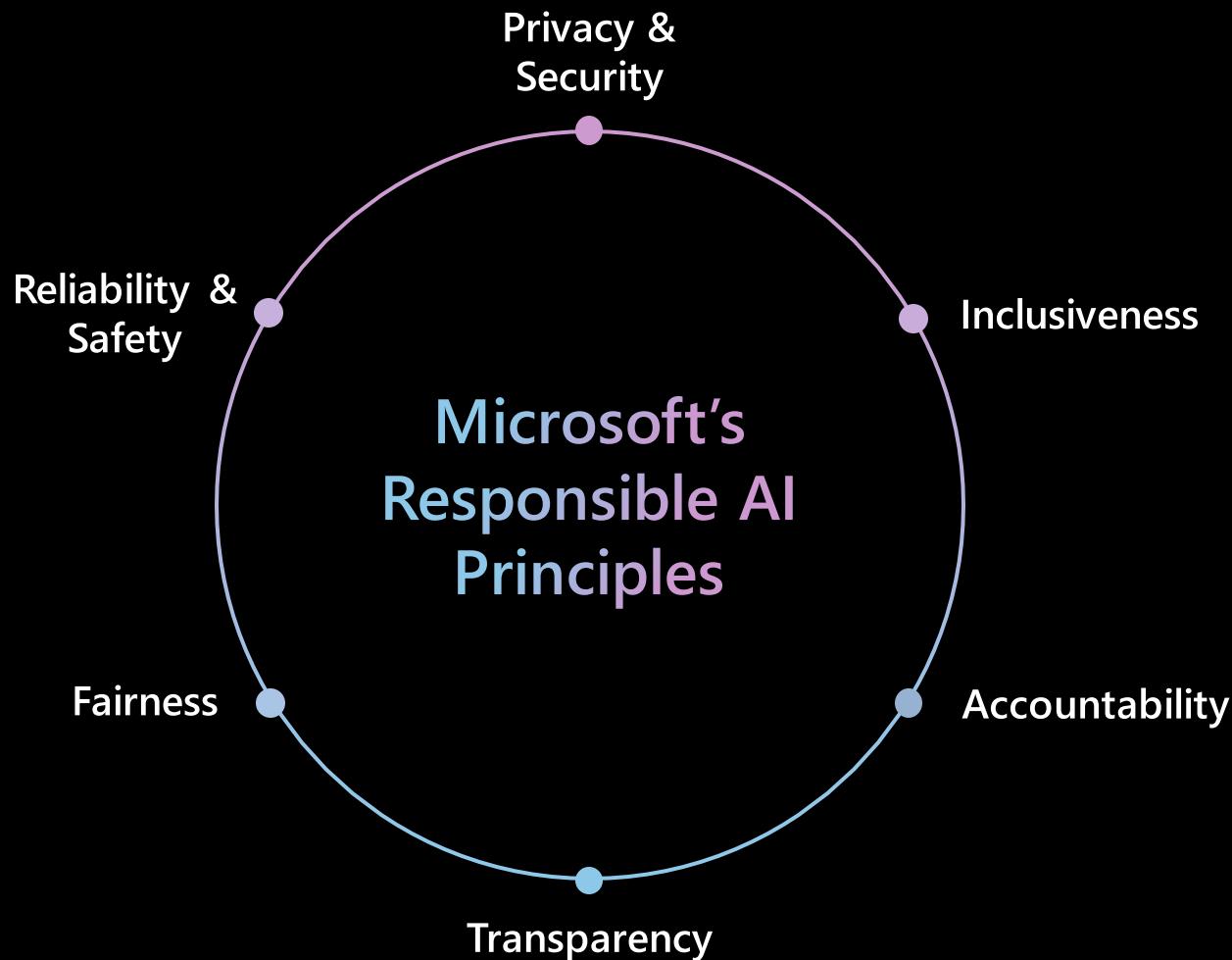
Microsoft 365 Copilot Windows 11



GitHub Copilot

Azure OpenAI Service

Power BI



Building blocks to enact principles



Tools and processes



Training and practices



Rules



Governance

| Microsoft Azure Cloud

Runs on trust

Your data is your data

Data is stored encrypted in your Azure subscription

Your data from any fine-tuning
is not used to train the
foundation AI models

Azure OpenAI Service provisioned in your Azure subscription

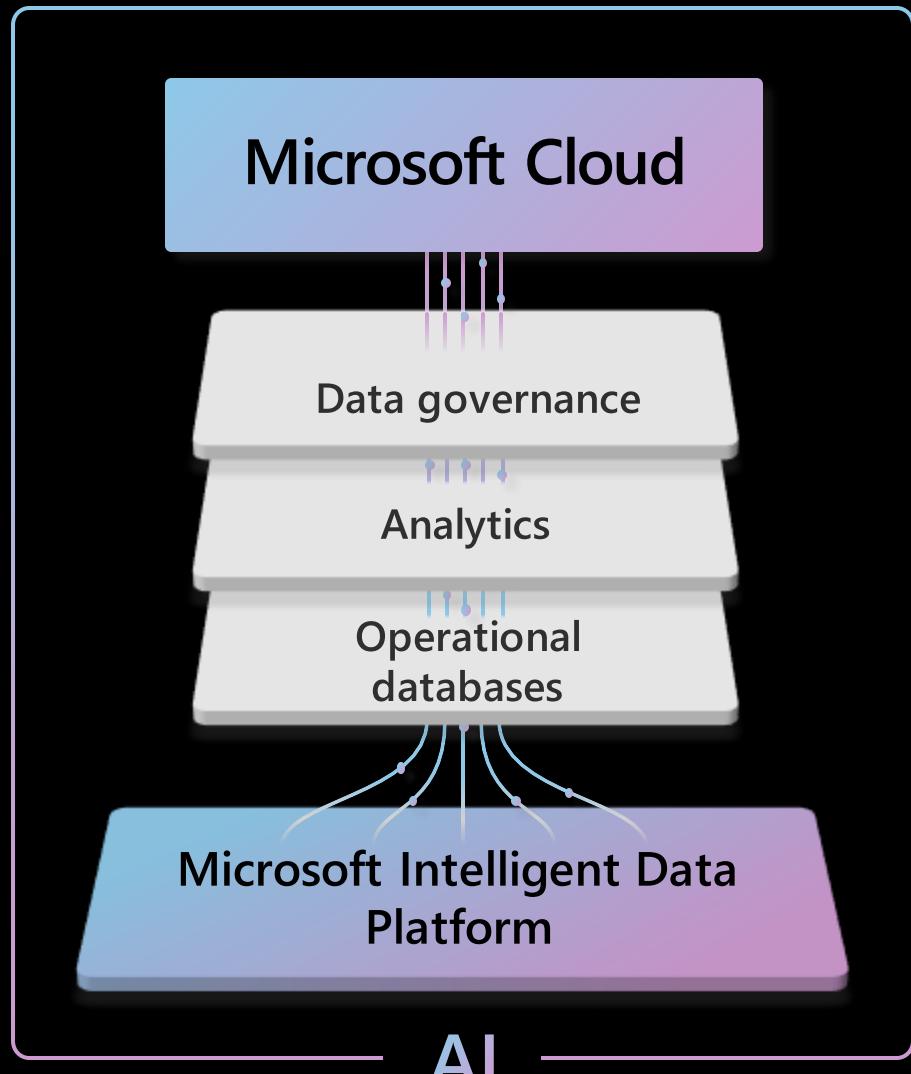
Model fine tuning stays in your Azure subscription
and never moves into the foundation AI models

Your data is protected by
the most comprehensive enterprise
compliance and security controls

Encrypted with Customer Managed Keys

Private Virtual Networks, Role Based Access Control

Soc2, ISO, HIPPA, CSA STAR Compliant



Deployed within your Azure subscription,
secured by you, accessed only by you, and
tied to your datasets and applications



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



Secure networking through private endpoints
and VNETs

| Azure OpenAI Service Capabilities

Essay outlines

Extracting insights

Creative Ideation

Subject Research

Summarizing text

Code generation

Semantic search

Virtual Assistants

Language Translation

Writing assistance

Limitless generation with a few lines of input

Comments from code

Code refactoring

Answering questions

Poem creation

Image Generation

Azure OpenAI

Top Capabilities and Use Cases



Content generation

Call center analytics: automatically generate responses to customer inquiries

Generate personalized UI for your website



Summarization

Call center analytics: summary of customer support conversation logs

Subject matter expert document: summarization (e.g. Financial reporting, analyst articles)

Social media trends summarization



Code generation

Convert natural language to SQL (or vice versa) for telemetry data

Convert natural language to query proprietary data models

Code documentation



Semantic search

Search reviews for a specific product/service

Information discovery and knowledge mining

Examples of multiple model use cases

End to end call center analytics: classification, sentiment, entity extraction, summarization and email generation

Customer 360: hyper-personalisation using timely summarization of customer queries & trends, search, and content generation

Business process automation: search through structured & unstructured documentation, generate code to query data models, content generation

A large, metallic Mercedes-Benz logo is positioned on the left side of the slide, casting a soft shadow on the dark blue background.

Mercedes-Benz pioneers use of Generative AI bots in Automotive

Challenge

How to offer customers with timely, reliable, and customized information and support for their vehicle-related needs and requests

Solution

Mercedes-Benz is leveraging conversational AI capabilities within Azure OpenAI Service to deliver best-in-class customer service with Mercedes Virtual Assistant

“... we have implemented two main features: user manuals and a configurator for customers without a vehicle, to give them the best possible experience. Although this bot is just a first step, we believe it will revolutionize the way Mercedes-Benz AG interacts with the customers and sets new standards. I am proud to see change being driven and boundaries being pushed!”

Automotive

Germany

Corporate
(10,000+ employees)

Azure OpenAI Service Customers





Opportunities of LLM for Product Development Teams

Pooja Patnaik
Partner Technology Strategist

LET'S START WITH THE BASICS.....

Artificial Intelligence



Artificial Intelligence

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

Machine Learning



Machine Learning

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

Deep Learning



Deep Learning

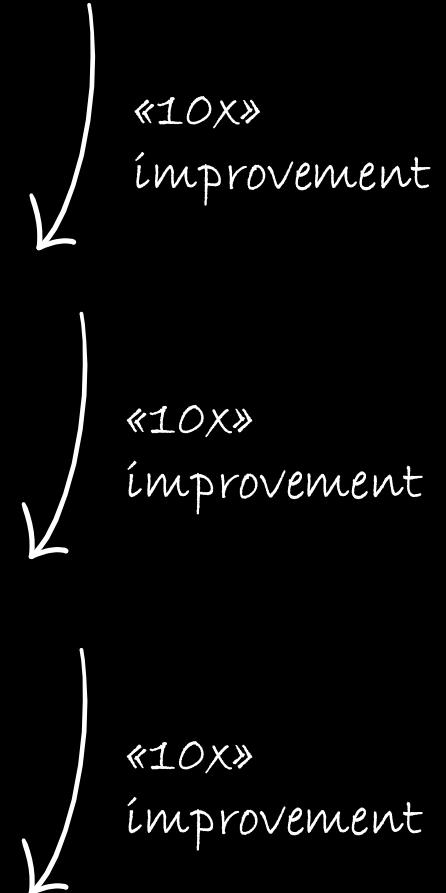
a machine learning technique in which layers of neural networks are used to process data and make decisions

Generative AI



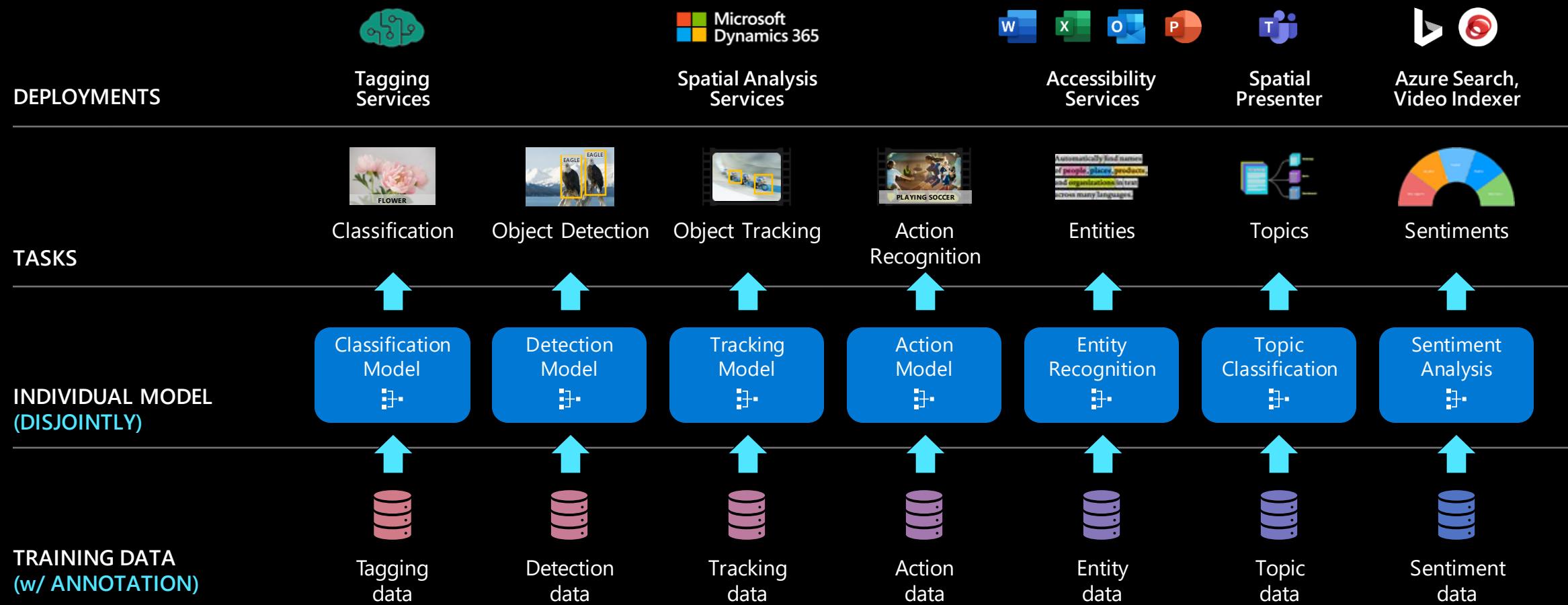
Generative AI

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

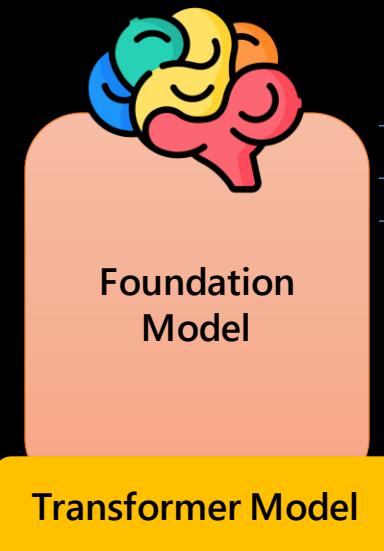


Traditional model development

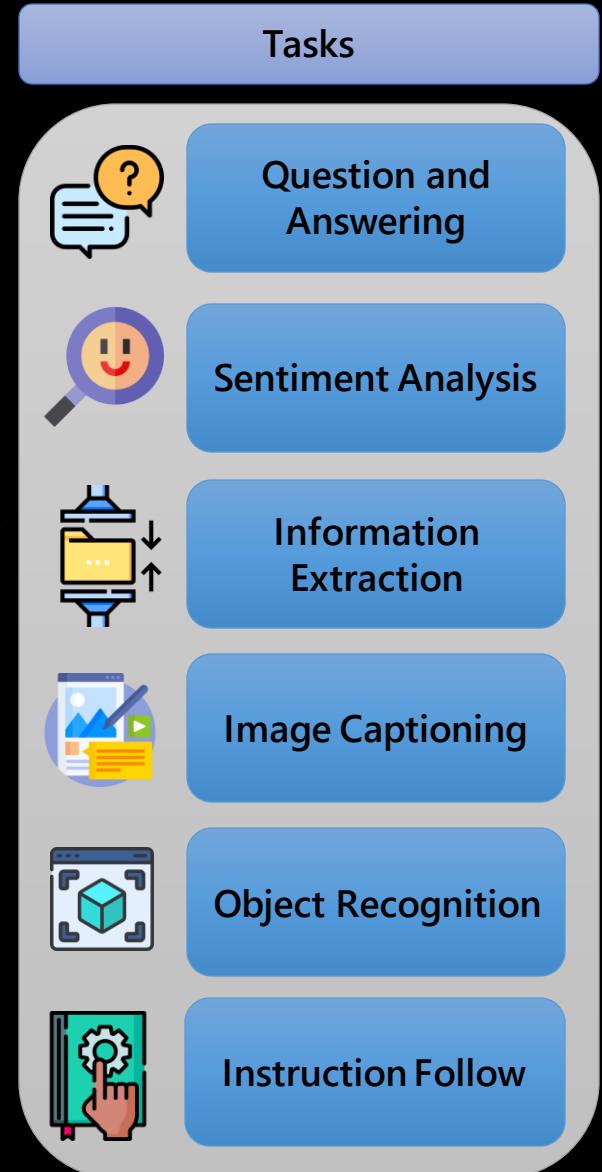
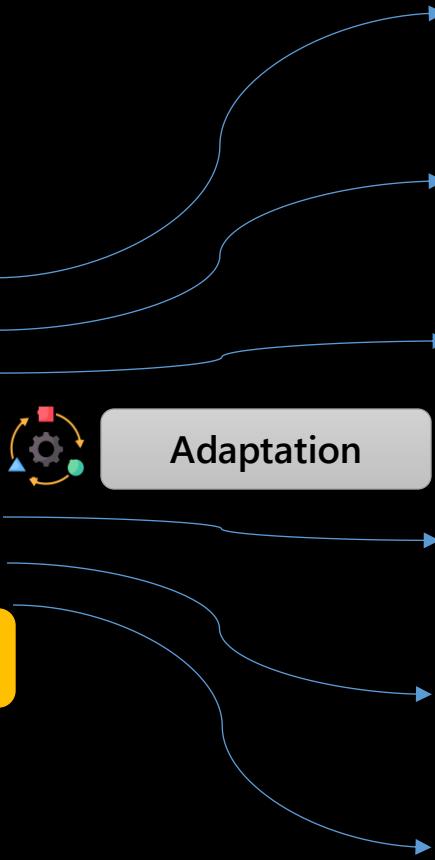
High cost and slow deployment—each service is trained disjointly



Foundation Models

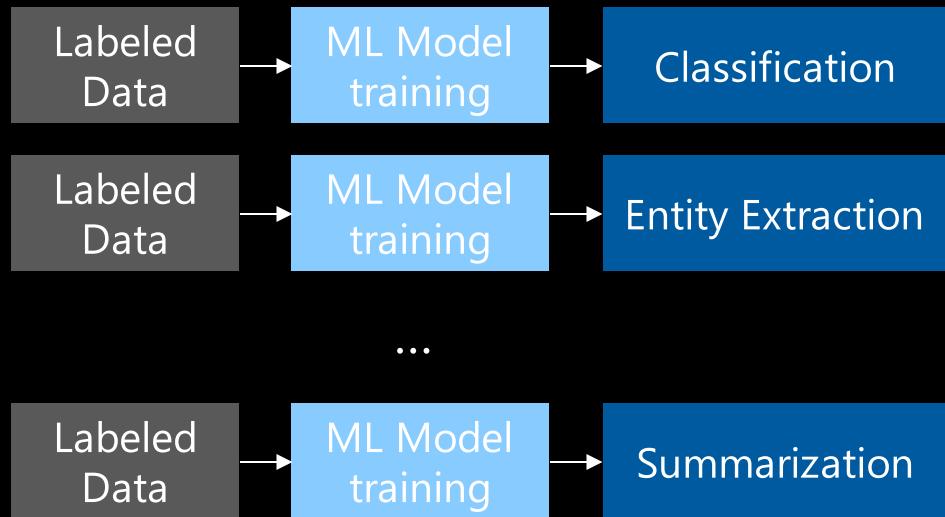


Adaptation

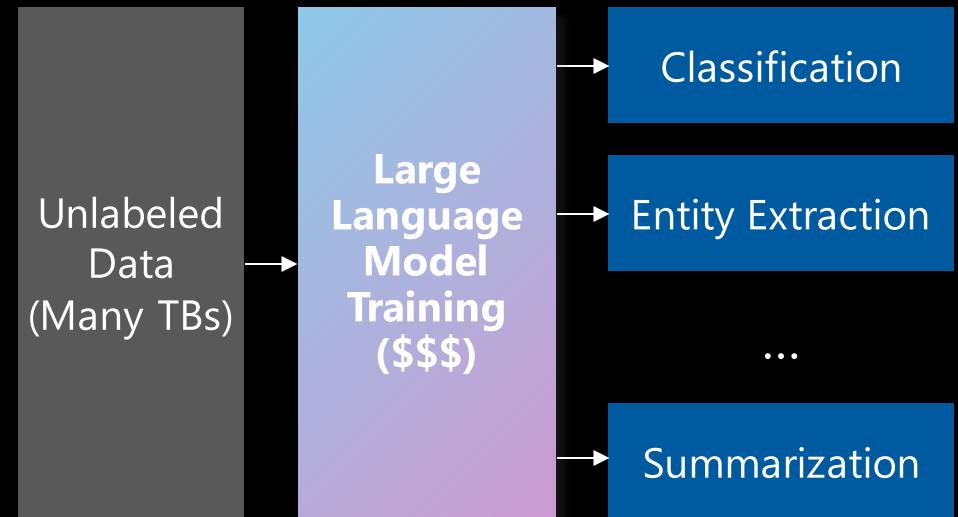


Large Language Models

Typical ML for NLP



Large Language Models



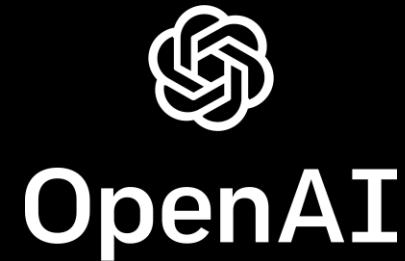
One model per capability
Labeled data to train
Highly optimized for use case

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

175B
GPT-3 | 2020

17B
Turing-NLG | 2020

340M
BERT-Large | 2018





OpenAI



*

GPT-4
2023

530B

Megatron
-Turing
NLG | 2021



175B
GPT-3 | 2020

17B
Turing-NLG | 2020

340M
BERT-Large | 2018

LLM | In a Nutshell

Large: More data than can be labeled (growing 10x/year!)

Language: Match context and words (e.g., tokens)

Model: Semi-supervised learning



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

LLMs can also generate things



Different versions produced by various companies LaMDA, Bard, LLaMa, RoBERTa



Microsoft + OpenAI LLMs →
GPT, Dall-E, Codex, Chat-GPT

GPT | Overview

A powerful type of LLM that can do a wide range of language related tasks, including...

*generating original content
that looks like a human wrote it*

**Generative pre-trained transformer
(GPT-3/4)**

Pre-trained on trillions of words, but NOT continuously learning

Predicts the most likely next word based on input text

GPT3 is unimodal, but GPT 4 is multimodal

Why is the sky blue?

OpenAI GPT-3 answer:

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

Main '*known*' Capabilities

Generation



Generate output that looks like the input, or novel text

e.g., Create Ideas, Generate a story, a song

Question answering



Answering questions from the prompts

e.g., Who is Satya Nadella?

Conversation



Taking on a role and responding over multiple turns in that style

e.g., You are a data scientist.....

Classification



Categorization, organization, sorting

e.g., Sort these notes into themes, label the results

Completion



Finish your sentence, paragraph, list.. or code

e.g., Complete this paragraph, this sentence, this function (github co-pilot)

Summarization



Summarizing text at different lengths & sophistications

e.g., Summarize this document in one sentence or at the level of a second grader

Transformation



Convert text (or data) from one form to another

e.g., Language translation, code translation, style/tone changes

Foundations for '*Product Thinkers*'

LLMs learn through *Prompts* (Few-shot) and not train

LLMs do not *understand* the meaning of the prompt input, nor the meaning of the words (or images) they are producing

LLMs are big *hallucinators*, and it is difficult to discern when

Big shift in a way a product is built → *Deterministic to Probabilistic*

They are not experts to rely on, they are probabilistic systems that are guidable.
Therefore, a *Collaborative UX* is necessary where the user is guiding the model towards their personal goals

LLMs in Industry | Use Cases

Industry	Application				
Marketing/ E-Commerce	Business Copywriting	Product Image Creation	Virtual Human Live-Streaming	Operation Data Analysis	...
Media/ Entertainment	News Writing	Playwriting	Virtual Idol	Music Composition	...
Smart Office	Meeting Summary	Coding Copilot	Brainstorming	Doc Creation	...
Game	Smart NPC Conversation	Game Plot Writing	Image/Video Creation	Game Map Generation	...
Healthcare	Disease Diagnosis	Drug Instruction	Health Assistant	Medical Image Analysis	...
Travel	Travel Plan Creation	Real-time Translation	Voice Tourist Guide	Car Rental/ Hotel Booking	...
Intelligent Industry	Smart Car	Finance	Education	Legal	...

Use Cases that stand out | Microsoft

Content generation

Call Center Analytics: Automatically generate responses to customer inquiries

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

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

Code generation

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

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

Semantic search

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

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

Summarization

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

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

Media example

News Analyses & Article Creations



News Broadcast
(Global Warming)



Azure
Speech
Transcription

Azure OpenAI Service
Text Completion

Search
SEO
Virtual Agent
Analytics
Reporting
Knowledge Mining

Content Analysis

Content Creation

Ideation
Productivity
Personalization
Accessibility
A/B Testing

Topic Classification
Global warming, Deforestation, Carbon footprint

Entity Extraction
Organizations: IPCC, UNFCCC, Green Peace
Geography: Canada, USA

Key Word Extraction
Human activities, fossil fuels, earth atmosphere

Question and Answer
What is the Intergovernmental Panel on Climate Change (IPCC)?
The IPCC is an international organization that studies climate change and the effectiveness ...

Video summarization
The article discusses about global warming and its effects on the Earth's atmosphere, wildlife, and human communities. It states that the primary cause of global warming is

News article generation (or blogs, social media)
Global warming is the gradual increase in the overall temperature of the Earth's atmosphere, primarily caused ...

Script Generation
Act 1: The show opens with a shot of a beautiful coastal town
Act 2: As the town struggles to cope with the crisis ...

Personalized Content generation (or Advertising)
Simon, as someone passionate about global warming, you are aware of the urgent threat it poses to our planet ...

LLMs within Microsoft | AI Copilot

The screenshot shows a Microsoft Excel spreadsheet titled "Q1 Performance - Saved". The spreadsheet contains data for various products across different countries, including manufacturing and sales details. A green border highlights the main data area. In the top right corner of the Excel window, there is a "Copilot" button. The Excel interface includes a ribbon bar with tabs like File, Home, Insert, Formulas, Data, Review, View, Automate, and Help. On the left side, there is a vertical sidebar with icons for Activity, Chat, Teams, Calls, Files, and Apps. The status bar at the bottom indicates "Calculation Mode: Automatic", "General", and "Workbook Statistics".

	A	B	C	D	E	F	G	H	I	J	K
1	Country	Customer	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales	COGS
2	France	VanArtsdel	Biodegradable Cleaning Products	Low	4345	\$10	\$352	\$1,529,440	\$15,294	\$1,514,146	\$1,451,648
3	USA	VanArtsdel	Biodegradable Cleaning Products	Low	4245	\$10	\$352	\$1,494,240	\$14,942	\$1,479,298	\$1,451,648
4	Australia	VanArtsdel	Biodegradable Cleaning Products	None	4124	\$10	\$352	\$1,451,648	\$0	\$1,451,648	\$1,451,648
5	Mexico	VanArtsdel	Biodegradable Cleaning Products	Medium	3954	\$10	\$352	\$1,391,808	\$83,508	\$1,308,300	\$1,265,800
6	France	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	4952	\$10	\$127	\$628,904	\$6,289	\$622,615	\$561,648
7	France	Contoso Pharmaceuticals	Recycled Can	Low	4903	\$5	\$127	\$622,681	\$6,227	\$616,454	\$561,648
8	Canada	Proseware	Natural Personal Care Products	Low	1954	\$270	\$302	\$590,108	\$5,901	\$584,207	\$558,207
9	Canada	Proseware	Biodegradable Cleaning Products	High	1990	\$220	\$302	\$600,980	\$84,137	\$516,843	\$482,843
10	USA	VanArtsdel	Compostable Food Packaging	Low	1254	\$3	\$380	\$476,520	\$4,765	\$471,755	\$437,188
11	USA	Proseware	Natural Personal Care Products	Medium	1465	\$224	\$302	\$442,430	\$26,546	\$415,884	\$389,336
12	USA	Proseware	Sustainable Snack Packaging	Medium	1410	\$120	\$302	\$425,820	\$25,549	\$400,271	\$389,336
13	Australia	Proseware	Natural Personal Care Products	Medium	1369	\$224	\$302	\$413,438	\$24,806	\$388,632	\$368,676
14	USA	VanArtsdel	Compostable Food Packaging	Medium	1052	\$3	\$380	\$399,760	\$23,986	\$375,774	\$352,188
15	Canada	VanArtsdel	Compostable Food Packaging	Low	980	\$312	\$380	\$372,400	\$3,724	\$368,676	\$336,992
16	Australia	Proseware	Sustainable Snack Packaging	Medium	1298	\$120	\$302	\$391,996	\$23,520	\$368,476	\$345,992
17	Australia	VanArtsdel	Compostable Food Packaging	None	954	\$3	\$380	\$362,520	\$0	\$362,520	\$336,992
18	Canada	Contoso Pharmaceuticals	Biodegradable Cleaning Products	Low	2785	\$110	\$127	\$353,695	\$3,537	\$350,158	\$336,992

Choosing Use Case | Other Risks

- I need a general-purpose model that can handle multiple tasks.
e.g., Use cases involving text analysis, summarization, forms or content generation are the sweet spot
- Not designed to process up-to-date factually accurate information
e.g., avoid use cases involving regulatory impact or oversight

Additional Considerations

-  Lacks interpretability and explainability
-  Limited input size
-  More computationally expensive
-  Risk of hallucinations
-  Can amplify bias, & publish offensive and copyrighted content

Trends & Potential

Market Size

- By 2030, Grand View Research¹ anticipates the global generative AI market will reach \$109.37 billion at a CAGR of 34.6% from 2022 to 2030
- Growing demand to modernize workflow across industries is expected to drive demand for generative AI applications.
- The fields that generative AI addresses—knowledge work and creative work—comprise billions of workers. The technology can make these workers at least 10% more efficient and/or creative

Significant growth is expected for the generative AI market

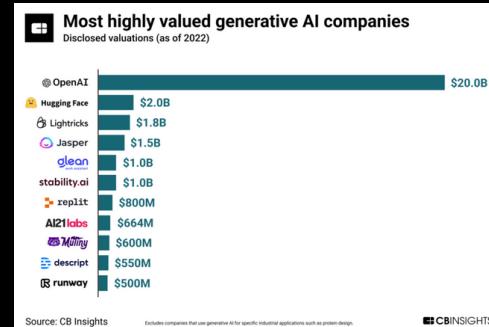
Use Cases

- Copywriters, designers, coders, photo and video editors, law firms, video game creators, and strategists now have access to generative AI tools.²
- Gartner³ states that generative AI will have a direct impact on the pharmaceutical, manufacturing, media, architecture, interior design, engineering, automotive, aerospace, defense, medical, electronics and energy industries by augmenting core processes with AI models.

Most generative AI use cases have less than 1% of target market adoption

Key Players

- In January 2023, Microsoft announced⁴ the third phase of a long-term partnership with OpenAI



<https://www.cbinsights.com/research/generative-ai-funding-top-startups-investors/>

- In February 2023, the company announced⁵ Bard, its LLM-based conversational AI;

Big technology companies are investing heavily in generative AI

Business Impact

- LLMs are expected to motivate a new wave⁶ of generative AI applications.
- Search is a good bet for disruption Microsoft & Google headed for a showdown over generative AI.⁷
- Microsoft Chief Technology Officer Kevin Scott⁸ believes that "2023 is going to be the most exciting year that the AI community has ever had." He predicts that generative AI models will "democratize access to AI" so a more diverse group of people will be able to "participate in the creation of technology."

Generative AI will be a game-changer

A Sneak Peek



Microsoft's Approach to AI



“

We have made huge investments in AI because we're **optimistic about what it can do to help people, industry and society**, and because we're committed to **bringing technology and people together to realize the promises of AI responsibly**.



AI is built and used responsibly and ethically



AI advances international competitiveness and national security



AI serves society broadly, not narrowly

Meeting the AI moment: advancing the future through responsible AI

Feb 2, 2023 | [Brad Smith - Vice Chair & President](#)

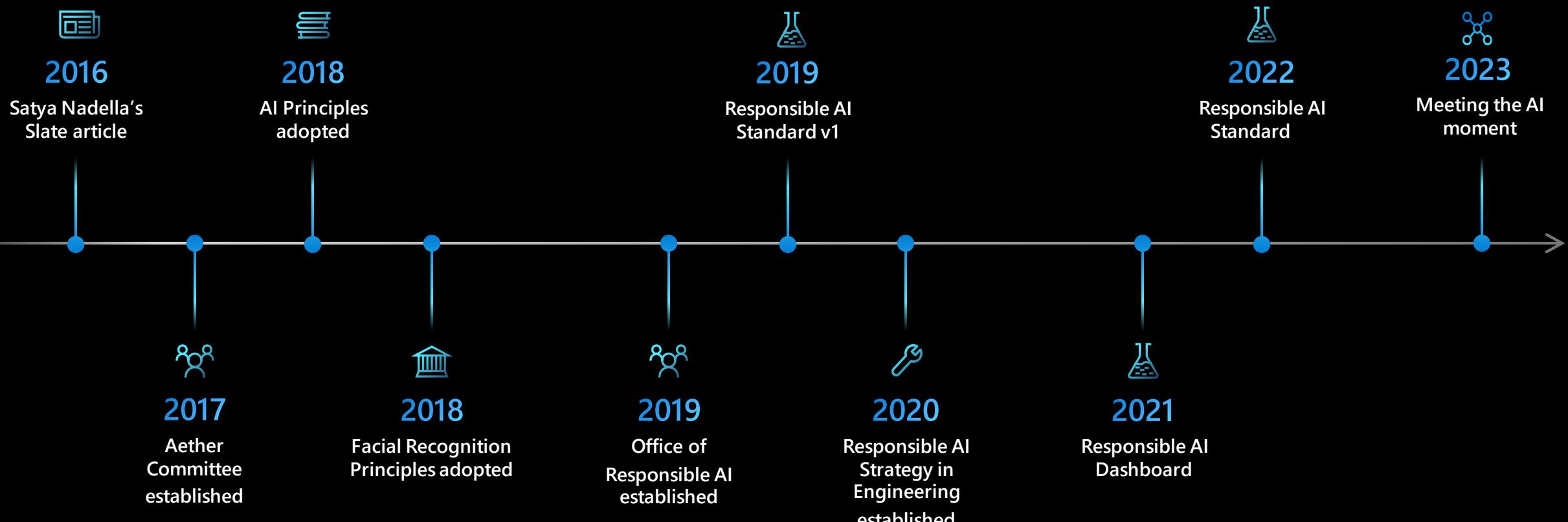


Early last summer, a small group of senior leaders and responsible AI experts at Microsoft started using technology from OpenAI similar to what the world now knows as ChatGPT. Even for those who had worked closely with the developers of this technology at OpenAI since 2019, the most recent progress seemed remarkable. AI developments we had expected around 2033 would arrive in 2023 instead.

Looking back at the history of our industry, certain watershed years stand out. For example, internet usage exploded with the popularity of the browser in 1995, and smartphone growth accelerated in 2007 with the launch of the iPhone. It's now likely that 2023 will mark a critical inflection point for artificial intelligence. The opportunities for people are huge. And the responsibilities for those of us who develop this technology are bigger still. We need to use this watershed year not just to launch new AI advances, but to responsibly and effectively address both the promises and perils that lie ahead.

The stakes are high. AI may well represent the most consequential technology advance of our lifetime. And while that's saying a lot, there's good reason to say it. Today's cutting-edge AI is a powerful tool for advancing critical thinking and stimulating creative expression. It makes it possible not only to search for information but to seek answers to questions. It can help people uncover insights amid complex data and processes. It speeds up our ability to express what we learn more quickly. Perhaps most important, it's going to do all these things better and better in the coming months and years.

Our Responsible AI journey



Microsoft's AI Principles



Fairness



Reliability
& Safety



Privacy &
Security



Inclusiveness



Transparency



Accountability

Responsible AI mitigation requirements

Integrations with Azure OpenAI Service must:

- Implement meaningful human oversight
- Implement strong technical limits on inputs and outputs to reduce the likelihood of misuse beyond the application's intended purpose
- Test applications thoroughly to find and mitigate undesirable behaviors
- Establish feedback channels
- Implement additional scenario-specific mitigations

To learn more, see the Azure OpenAI transparency note.

Integrations with Azure OpenAI Service must not:

- be used in any way that violates Microsoft's [Acceptable Use Policy](#), including but not limited to any use prohibited by law, regulation, government order, or decree, or any use that violates the rights of others;
- be used in any way that is inconsistent with this code of conduct, including the Limited Access requirements, the Responsible AI mitigation requirements, and the Content requirements;
- exceed the use case(s) you identified to Microsoft in connection with your request to use the service;
- interact with individuals under the age of consent in any way that could result in exploitation or manipulation or is otherwise prohibited by law or regulation;
- generate or interact with content prohibited in this Code of Conduct;
- be presented alongside or monetize content prohibited in this Code of Conduct;
- make decisions without appropriate human oversight if your application may have a consequential impact on any individual's legal position, financial position, life opportunities, employment opportunities, human rights, or result in physical or psychological injury to an individual;
- infer sensitive information about people without their explicit consent unless if used in a lawful manner by a law enforcement entity, court, or government official subject to judicial oversight in a jurisdiction that maintains a fair and independent judiciary; or
- be used for chatbots that (i) are erotic, romantic, or used for companionship purposes, or which are otherwise prohibited by this Code of Conduct; (ii) are personas of specific people without their explicit consent; (iii) claim to have special wisdom/insight/knowledge, unless very clearly labeled as being for entertainment purposes only; or (iv) enable end users to create their own chatbots without oversight.

Code of Conduct

23

The Azure OpenAI Service is subject to the applicable [Azure Legal Terms](#) (including the Product Terms), and preview features are also subject to the [Supplemental Terms for Preview Features](#) for the Azure OpenAI Service.

As provided in the Product Terms, your use of the Azure OpenAI Service is governed by the [Code of Conduct for the Azure OpenAI Service](#), which outlines minimum standards of behavior that all applications must adhere to in good faith. Microsoft may update the Code of Conduct from time to time. Applications that fail to comply with the Code of Conduct may lose access to the Azure OpenAI Service. *

Yes, I attest

Azure OpenAI Service Responsible AI



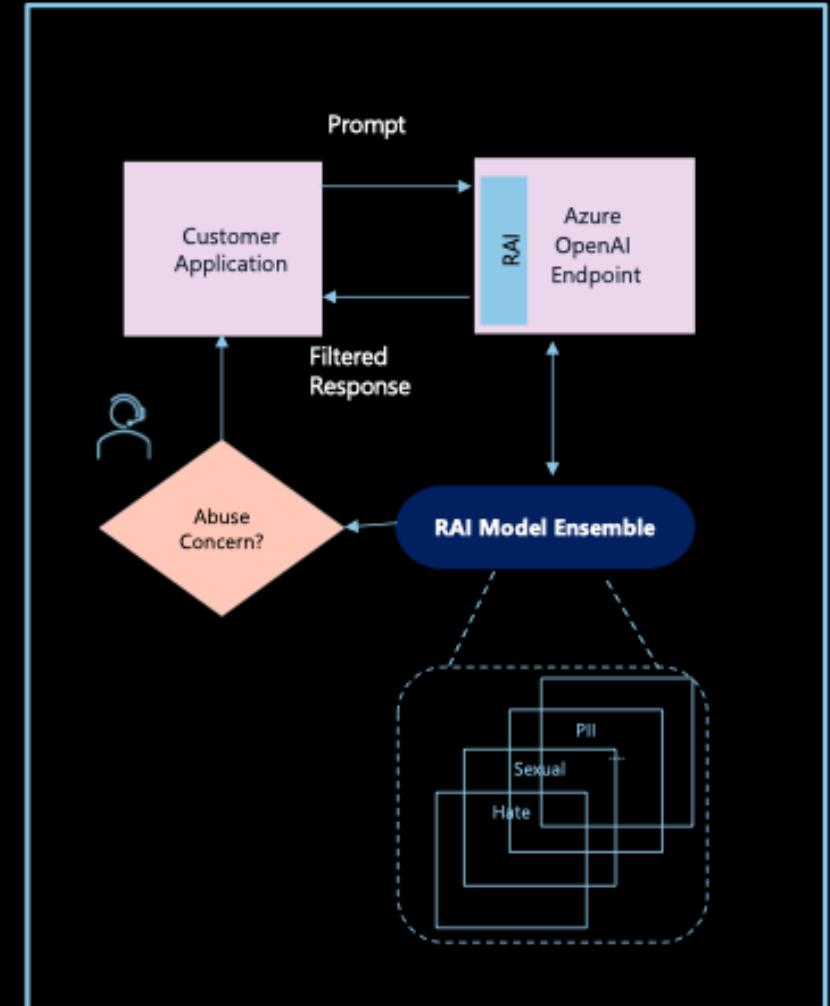
**Custom content filters—
tailor tone and topics to your application**



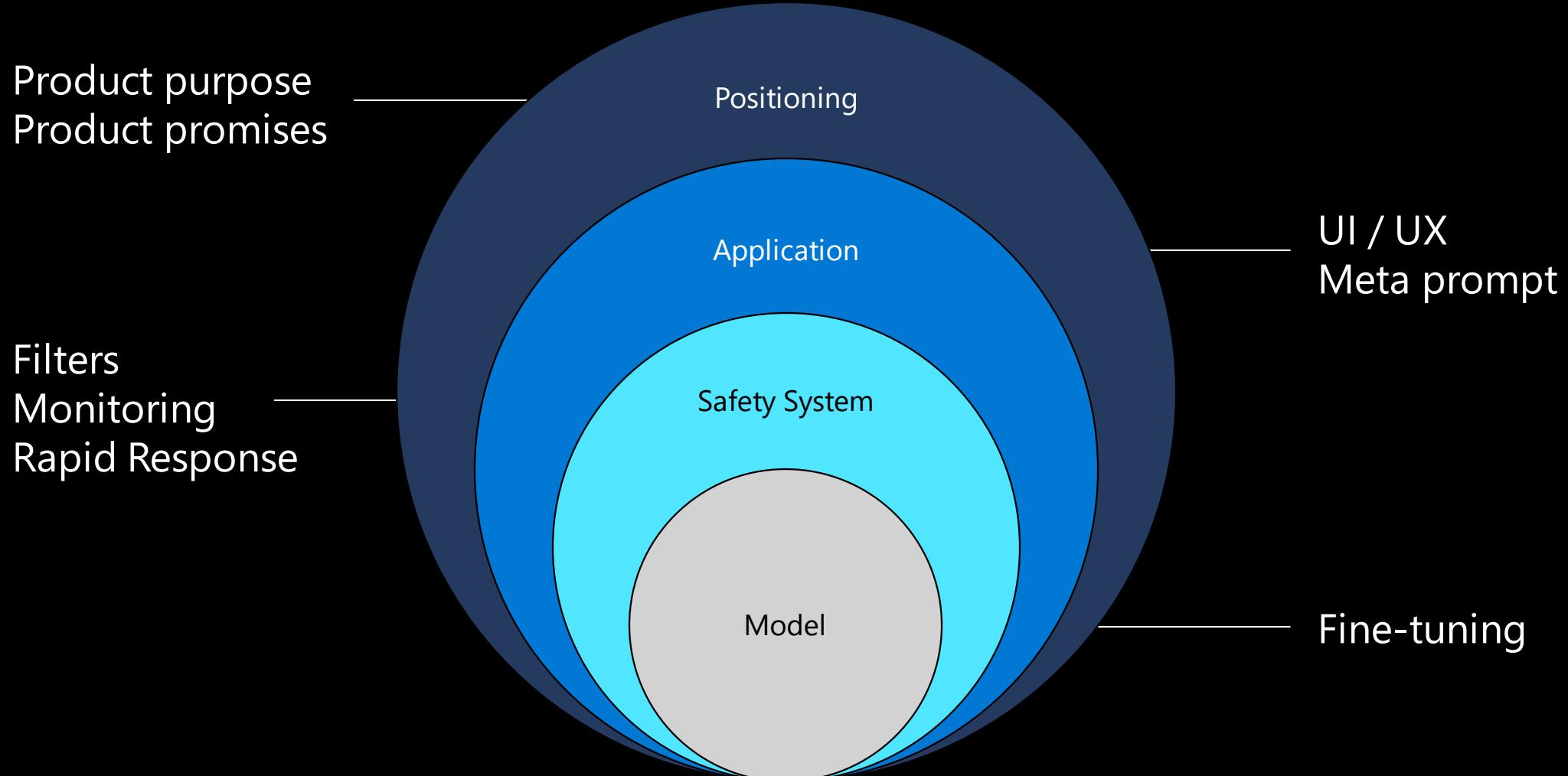
**Abuse detection—
ensure responsible use of your application**



**Implementation guidelines, patterns,
and best practices**



Mitigation Layers



RAI Mitigations

Customer

- Structure user interactions. Limit the length, structure, and source of inputs and outputs
 - Control user access
 - Transparency and overreliance mitigations in UI/UX
-

Technical

- Content Filtering
 - Asynchronous abuse detection
 - User-based throttling
 - User-based shutdown
-

Process and Policy

- Limited Access
 - Abuse reporting channel
 - Feedback channel
 - Incident Response
-

Documentation and legal

- Terms of use
- Transparency Note
- Design Guidelines

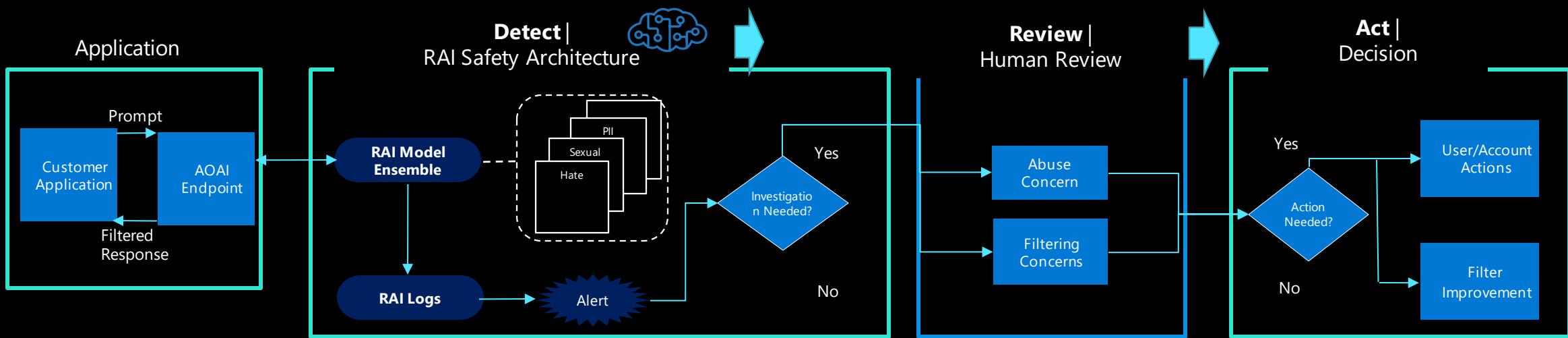
Quality

Harm to individuals or businesses due to unintended outputs or overreliance

INCIDENT CLASS SHORTHAND	INCIDENT CLASS EXTENDED DESCRIPTION
Inaccurate Text	API generates misleading, inaccurate, or poorly-contextualized content on high-stakes topics
Incorrect/insecure code	API generates incorrect or insecure code that is used unknowingly by users
PII	API generates responses that contain email addresses, SSNs, and other PII that is
Proprietary Info/plagiarism	API generates based on content or code that is proprietary
Demeaning, stereotyping, hate	API generates content that is offensive toward members of social groups
Inequitable allocation	API outputs lead to inequitable allocation of resources (e.g., likelihood of receiving job interview based on automated resume screener)
Quality of service harms	API systematically performs worse on text by, for, and about different social groups
Violence or Self-Harm	API instructs, affirms, or radicalizes a human to commit direct harm to themselves or others
Profane, sexual, inappropriate, or sensitive content	API generates contextually inappropriate, offensive, or sensitive content

Safety Execution Workflow

□ People & Policy



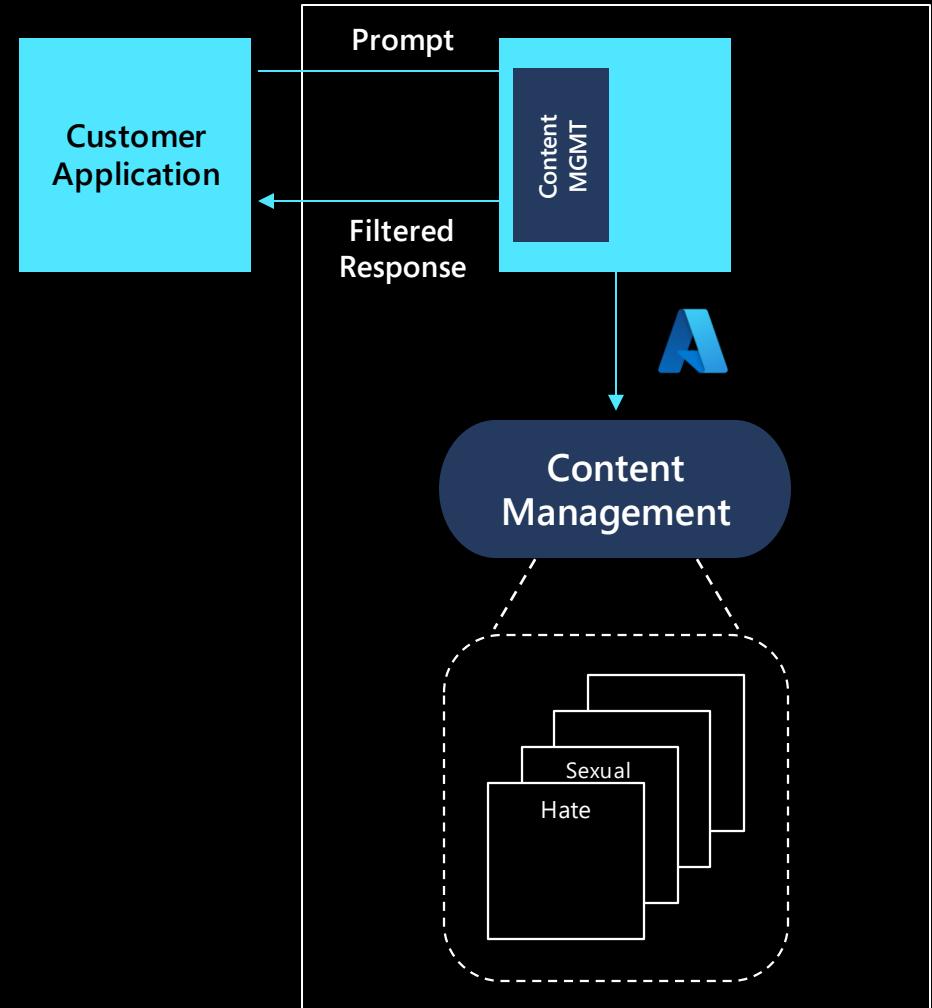
Azure OpenAI Service Responsible AI



**Content filtering—
can filter out abuse and misuse**

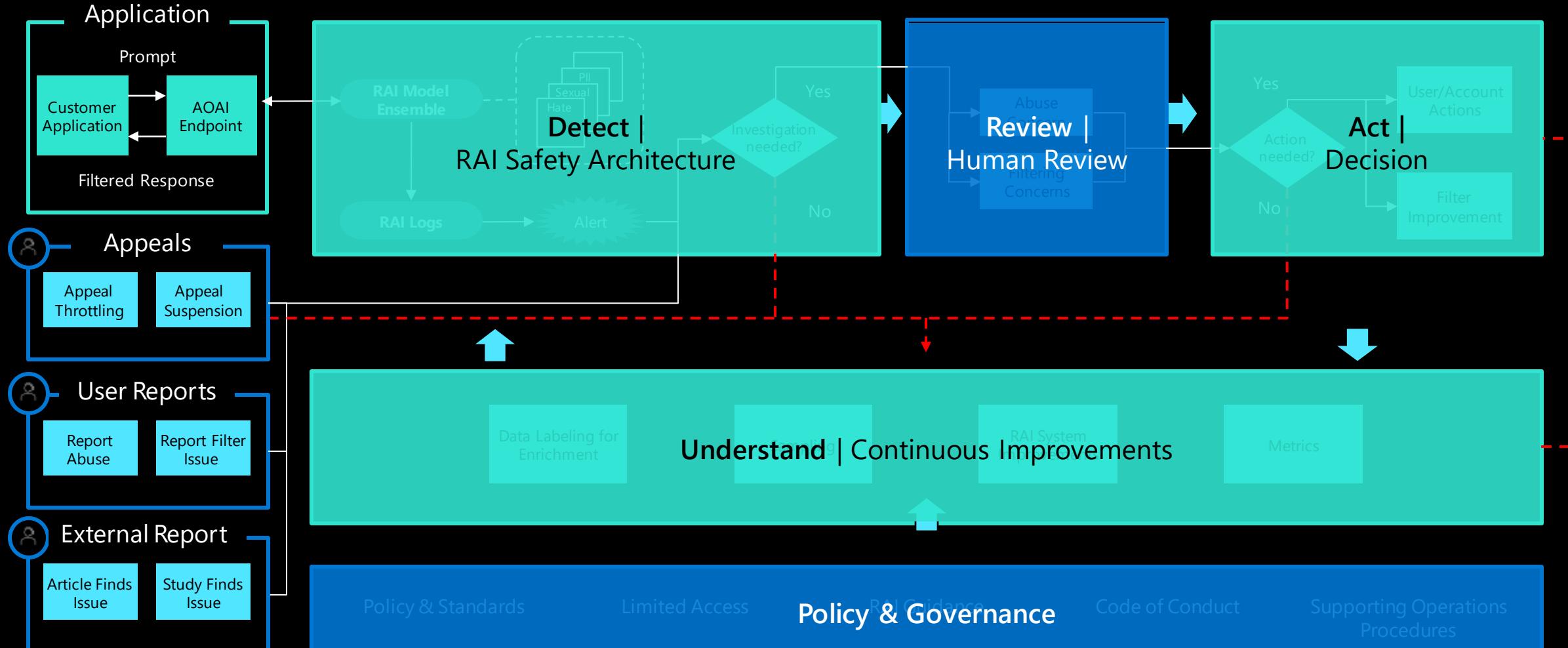
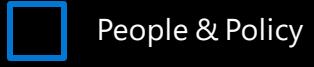


**Implementation guidelines, patterns,
and best practices**

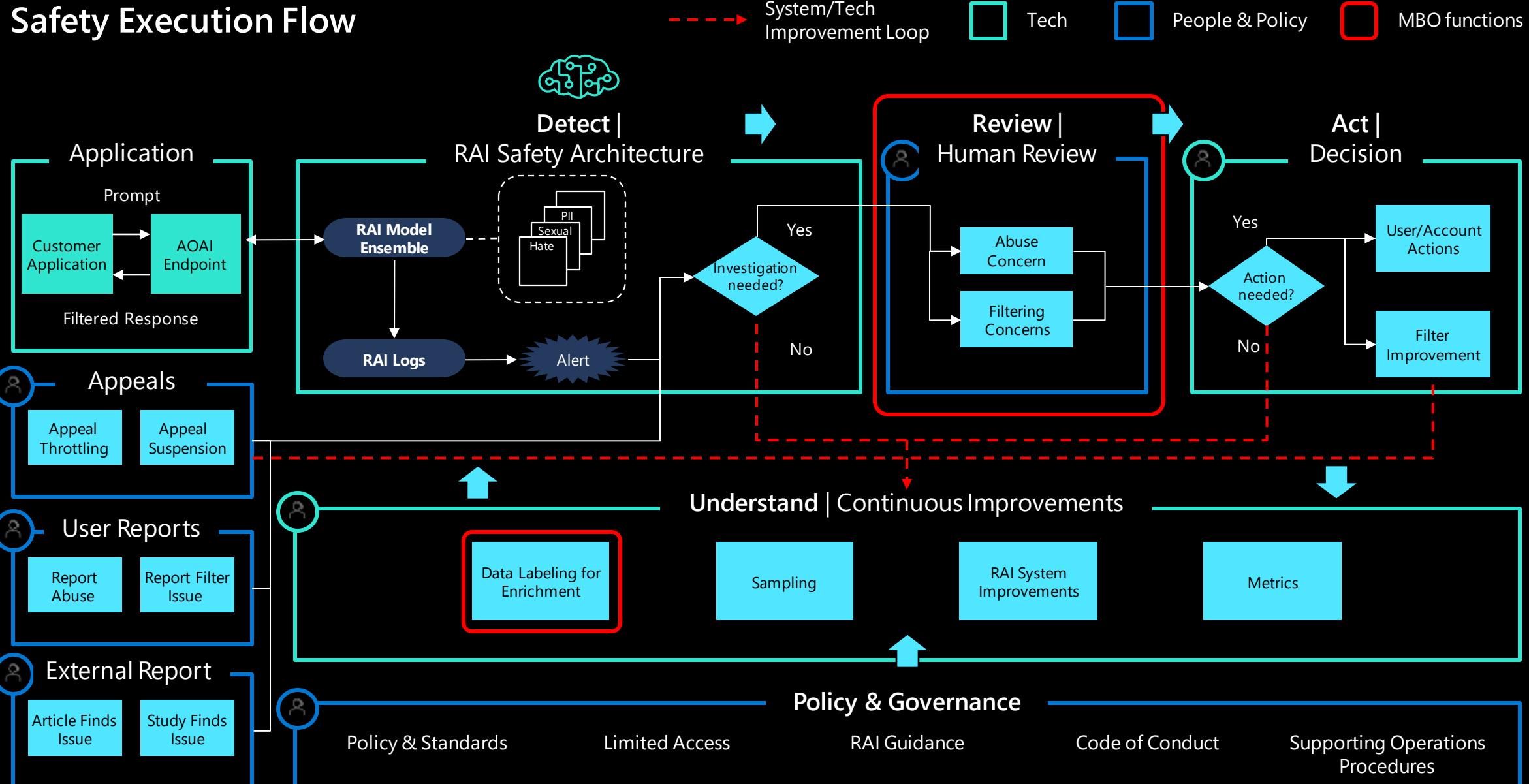


Safety Execution Flow

System/Tech Improvement Loop



Safety Execution Flow



Responsible AI resources

Content filtering:

<https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/content-filter>

Responsible AI resources (aka.ms/RAIResources)

Human and AI Interaction Toolkit

(<https://www.microsoft.com/en-us/haxtoolkit/workbook/>)

Management tools

Afternoon agenda

01:00 pm – 01:15 pm Fundamentals of Azure OpenAi

01:15 pm – 01:40 pm Best Practices for Prompt Engineering

01:40 pm – 02:00 pm Embedding & FineTuning

02:00 pm – 02:05 pm Bio Break

02:05 pm – 02:30 pm Orchestration (Langchain & Semantic Kernels)

02:30 pm – 03:00 pm Live Coding Demo

03:00 pm – 03:10 pm AAAP – Microsoft Incentive Campaign for customer PoCs

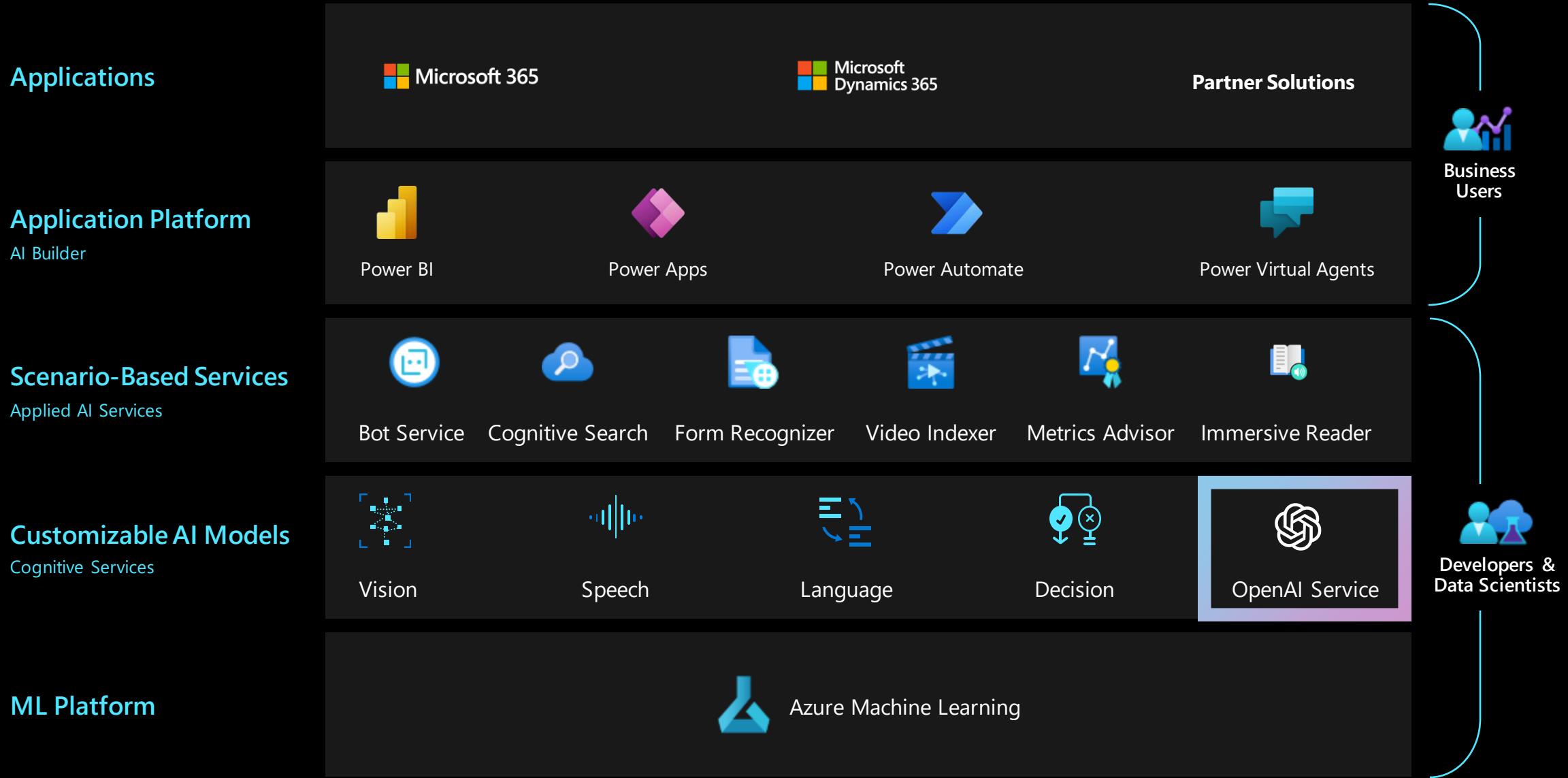
03:10 pm – 03:30 pm Q&A



Fundamentals of Azure OpenAI

Florian Lautenschlager
Partner Technology Strategist

Azure AI



Azure OpenAI Service

GPT-3 / 4

Codex

DALL·E

ChatGPT



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



Compatibility and a smooth transition
as Azure OpenAI co-develops the APIs
with OpenAI



Custom AI models fine-tuned with your
data and hyperparameters

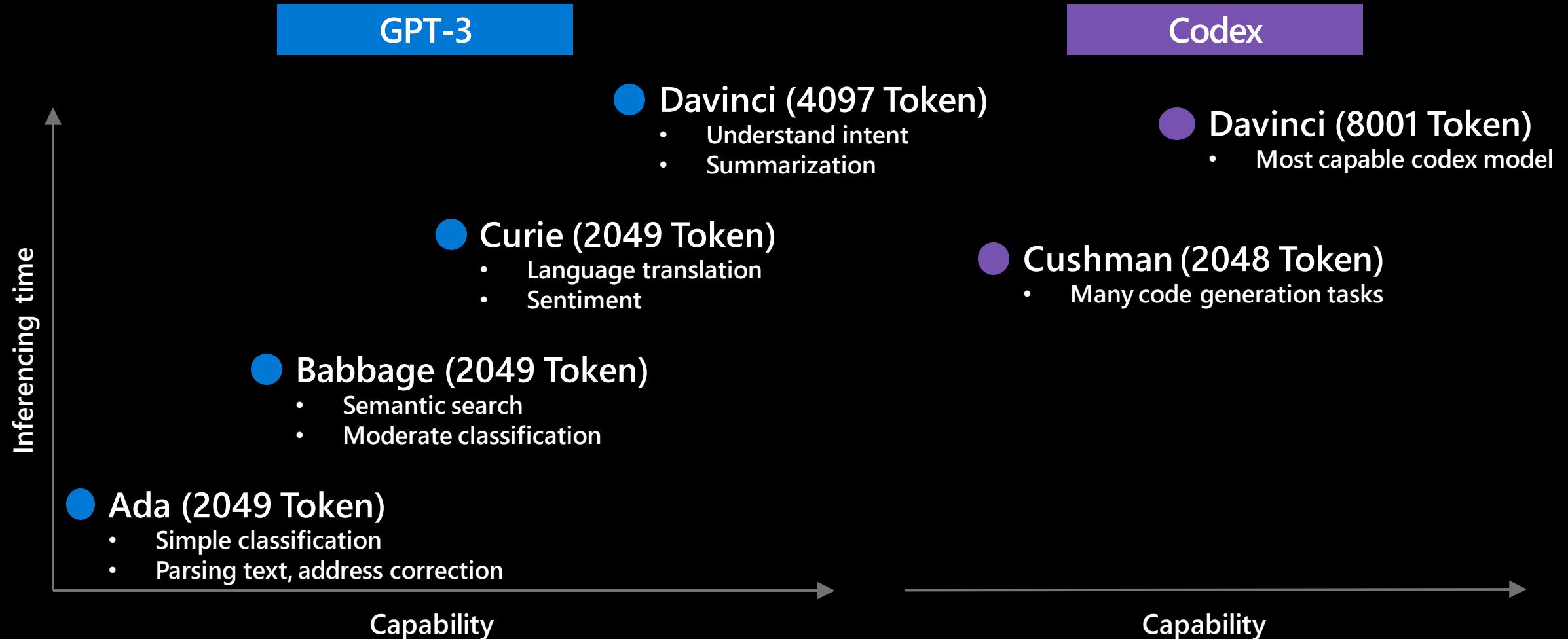


Built-in responsible AI to detect and
mitigate harmful use



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

Azure OpenAI | Family of Models



GPT-3 & Codex Naming Convention

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

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

Example: text – davinci – 003

Family:
text (GPT-3), code
(Codex)

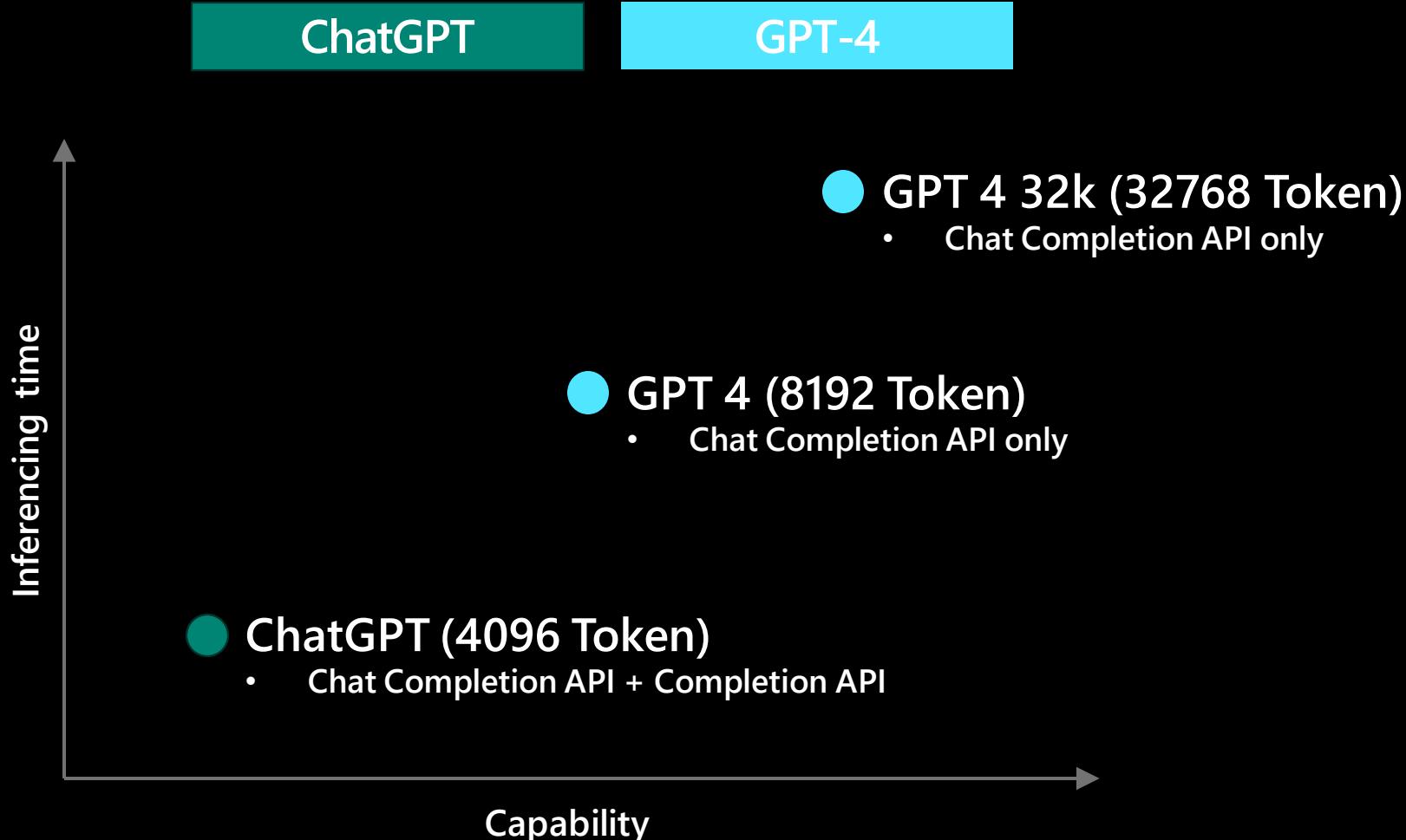
Capability:
Ada, Babbage, Curie,
Davinci

Input-Type (optional):
Only embeddings
models

Identifier:
Version of the model

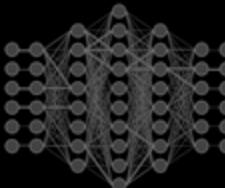
*<https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/models>

Azure OpenAI | Family of Models



Common Terms & Parameters

In:



Out:

Prompt: Should I paint my house blue or

Completion:... green.

Conversation: Messages

Conversation: Message

Parameters:

Token: Number of tokens per request including input, output and parameters.

- Roughly 4 characters for a typical English text, e.g., Hamburger => "ham", "bur", "ger"
- Also affects the response latency and throughput for the models

Temperature: Controls randomness / creativity of the response (value between 0 (low) and 2 (high))

Top P: Like Temperature (different method) (value between 0 (low: likelier tokens) and 1 (high: high & low likely tokens)).

Frequency / Presence penalty: Control the occurrence of a token in the response

- Reduce the chance of repeating a token proportionally based on how often it has appeared so far.
- Reduce the chance of repeating any token that has appeared in the text at all so far.

Prompt Engineering, Fine Tuning, Embeddings

Prompt Engineering: The prompt acts to configure the model weights to complete the desired task, but it's more of an art than a science, often requiring experience and intuition to craft a successful prompt.

Fine Tuning: Tailor models to your personal datasets to get higher quality result than what you can get just from prompt design, plus the ability to train on more examples than can fit into a prompt.

Embeddings: Special format of data representation that can be used to correlate with semantic similarity between two texts. Models are used to calculate the embeddings.

Completion Endpoint: Text in. Text Out

GPT 3, Codex + ChatGPT

```
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/completions?api-version=2022-12-01\
```

```
  -H "Content-Type: application/json" \
  -H "api-key: YOUR_API_KEY" \
  -d "{"
    \"prompt\": \"Once upon a time\",
    \"max_tokens\": 5
}"
```

=>", a dark line crossed" (simplified)

Chat Completion Endpoint: Conversation in. Message out.

ChatGPT + GPT 4

```
https://YOUR_RESOURCE_NAME.openai.azure.com/openai/deployments/YOUR_DEPLOYMENT_NAME/  
chat/completions?api-version=2023-03-15-preview \  
-H "Content-Type: application/json" \  
-H "api-key: YOUR_API_KEY" \  
-d '{
  "messages": [
    {"role": "system", "content": "You are a helpful assistant."},
    {"role": "user", "content": "Does Azure OpenAI support customer managed keys?"},
    {"role": "assistant", "content": "Yes, customer managed keys are supported by Azure OpenAI."},
    {"role": "user", "content": "Do other Azure Cognitive Services support this too?"}
  ]
}'  
=> [{"message": {"role": "assistant", "content": "Yes, ..."}]} (simplified)
```

Azure OpenAI Studio

Let's get started

The screenshot shows the Azure OpenAI Studio interface with the 'Chat' tab selected. On the left, a sidebar lists 'Azure OpenAI', 'Playground', 'Completions', 'Management', 'Deployments', 'Models', and 'Data Files'. The main area is titled 'Chat playground (Preview)' and contains two panels: 'Assistant setup' and 'Chat session'. The 'Assistant setup' panel includes sections for 'Specify how the chat should act' (using a template or writing your own system message), 'Use a system message template' (set to 'Marketing Writing Assistant'), and 'System message' (a text input field containing a template for a marketing writing assistant). The 'Chat session' panel shows a conversation history with the AI suggesting marketing slogans. It includes buttons for 'Clear chat', 'View code', and 'Show raw JSON'. On the right, there are 'Parameters' and 'Session settings' sections, showing deployment details ('chat-gpt') and session statistics ('Past messages included: 10', 'Current token count: 0', 'Input tokens progress indicator: 258/4000'). A bottom 'User message' input field is ready for the user's query.



Azure OpenAI Service Workshop Prompt Engineering and Best Practices

Edwin Huber
Sr. Partner Technology Strategist

When GPT fails on a task, what should you do?



When GPT-3 fails on a task, what should you do?

1. Give clearer instructions, also say what not to do!

2. Split complex tasks into simpler subtasks

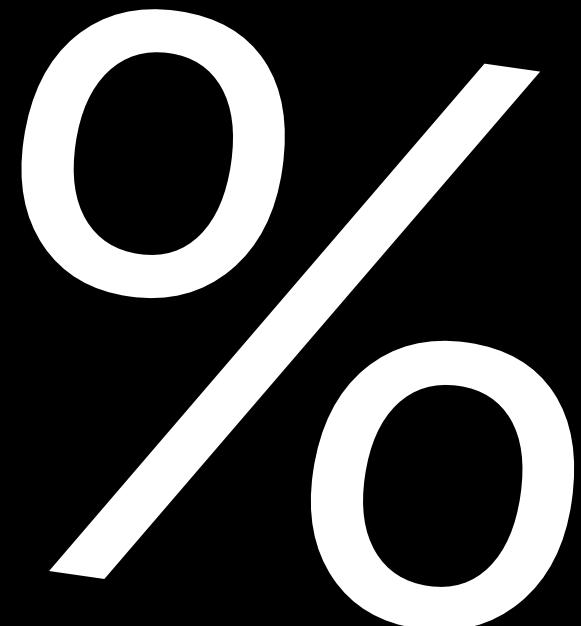
3. Structure the instruction to keep the model on task

4. Prompt the model with logical path and to explain before answering

5. Ask for justifications of many possible answers, and then synthesize

6. Generate many outputs, and then use the model to pick the best one

7. Fine-tune custom models to maximize performance



What are clearer instructions?

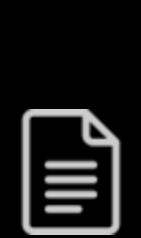


Splitting complex instructions
and large inputs...



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



Structuring complex instructions...

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

Explain what you want or how to do it...

“Few-Shot”



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.



A quick word about
demoing something
and expectations
based on
probabilities...



DEMO – Few-Shot reasoning

Azure OpenAI

Playground

Chat

Completions

Management

Deployments

Models

Data Files

Completions playground

Deployments

text-davinci-002

Examples

Load an example

View code

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: The answer is 11.

Q: A juggler has 16 balls. Half of the balls are golf balls and half of the golf balls are blue. How many blue golf balls are there?

A:

Generate

Undo

Regenerate

Tokens: 81

Parameters

Temperature

1

Max length (tokens)

100

Stop sequences

Stop sequences

Top probabilities

0.5

Frequency penalty

0

Presence penalty

0

Best of

1

Pre-response text

Enter text

Post-response text

Enter text

[Learn more](#)

Azure OpenAI Studio > Completions playground

Completions playground

Deployments Examples

basic-gpt35-turbo Load an example

View code

I want to develop a comprehensive marketing strategy for a new product launch. The product is a smart home security system targeting tech-savvy homeowners. I need ideas on how to effectively position the product in the market, identify target customer segments, and outline key marketing channels and tactics.

Parameters

Temperature 0.7

Max length (tokens) 600

Stop sequences

Top probabilities 0.5

Frequency penalty 0

Presence penalty 0

Pre-response text

Post-response text

Learn more

Generate Undo Regenerate Tokens: 59

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

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.

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.

DEMO – Chain of thought reasoning



Azure OpenAI

Playground

Chat

Completions

Management

Deployments

Models

Data Files

Completions playground

Deployments

Examples

basic-gpt35-turbo

Load an example

View code

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:

Generate



Tokens: 43

Parameters

Temperature

0.7

Max length (tokens)

90

Stop sequences

Stop sequences

Top probabilities

0.5

Frequency penalty

0

Presence penalty

0

Pre-response text

 Enter text

Post-response text

 Enter text[Learn more](#)

Selection-Inference Prompting

- First, a prompt selects a relevant subset of facts from the text ('selection prompt').
- Then, a second prompt infers a conclusion from the selected facts ('inference prompt').
- These prompts are then alternated in a loop to generate multiple steps of reasoning and eventually land on a final answer.

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 – Selection-Inference prompting

Selection-Inference Demo

```
In [30]: text_prompt = '\nHere are a collection of stories about people carrying\none room to another . You will be asked where any obj\nanswer this question you need to figure out who last\nobject and which room they have the object in by the\nstory . Here are some examples :\nStory :\n\nat t =0 mary grabbed the football there\nat t =1 daniel got the apple there\nat t =2 mary went to the kitchen\nat t =3 daniel journeyed to the office\nat t =4 daniel went to the bedroom\nat t =5 mary moved to the garden\nQuestion: where is the apple?\n'
```

VS

```
In [32]: text_prompt="Here are a collection of stories about people carrying objects from\nnone room to another .
```



Selection-inference Prompting

Here we shall show a small example of how we provide the language model with a more structured and constrained input, which can improve its ability to generate accurate and relevant outputs.

```
In [ ]: import os
import openai
from dotenv import load_dotenv
load_dotenv()
openai.api_type = "azure"
# openai.api_version = "2023-03-15-preview"
openai.api_version = "2022-12-01"

API_KEY = os.getenv("OPENAI_API_KEY","").strip()
assert API_KEY, "ERROR: Azure OpenAI Key is missing"
openai.api_key = API_KEY

RESOURCE_ENDPOINT = os.getenv("OPENAI_API_BASE","").strip()
assert RESOURCE_ENDPOINT, "ERROR: Azure OpenAI Endpoint is missing"
assert "openai.azure.com" in RESOURCE_ENDPOINT.lower(), "ERROR: Azure OpenAI Endpoint should be in the"
openai.api_base = RESOURCE_ENDPOINT
```

This prompt fails to be reasoned correctly

There are 2 prompts, which differ only in their formatting, which sometimes leads to different answers and behaviors of the model's response.

Again, we are working with probabilities, and small changes sometimes dramatically affect the probabilities of a response.

In such a case, we need to work with both narrowing response options and guiding the model to respond correctly:

Justify...



Ask the model to justify its' reasoning

- Given a list of potential solutions to a problem.
- Request that the model generates content that synthesizes the justifications for each solution
- Finally request an explanation for the choice of solution.



A close-up photograph showing a person's hands holding a fan deck of color swatches. The swatches are fanned out, displaying a variety of colors including shades of pink, red, orange, and white. The person is wearing a silver watch on their left wrist. In the background, a large, flat surface holds several other color palettes and a blue pen. The overall scene suggests a design or color selection process.

Choose...

Choose...

- Use the model to generate many possible answers
- Use another model or function to pick the ones that look best

Generate a list of 5 different office products

1. Desk Lamp
2. Computer Monitor
3. Wireless Keyboard and Mouse
4. Standing Desk Converter
5. Noise-Cancelling Headphones

Recommend the best one for a knowledge worker based on their needs.

A knowledge worker is someone who primarily uses their brainpower and knowledge to carry out their work. They spend most of their time working on a computer, reading and writing, and analyzing information

Therefore, a computer monitor is the best office product for a knowledge worker as it can help them to increase their productivity and improve the quality of their work.

Choose...

- Self-consistency
 - For tasks with a discrete set of answers, one simple way to improve reliability is to sample multiple explanations & answers from the model (using a positive temperature) and then pick the final answer that appears most often.
- Maieutic prompting
 - This approach uses the model to generate a tree of possible explanations (both correct and incorrect), and then analyzes their relationships to guess at which set is correct.
- Least-to-most prompting
 - In addition to doing poorly on long reasoning chains (where selection-inference shines), chain-of-thought prompting can especially struggle when the examples are short but the task is long.
 - Splits up reasoning tasks into smaller, more reliable subtasks. The idea is to elicit a subtask from the model by prompting it with something like To solve {question}, we need to first solve: ". Then, with that subtask in hand, the model can generate a solution. The solution is appended to the original question and the process is repeated until a final answer is produced.

DEMO – Justification & Choosing



Justification and choosing the most promising approach

Create a list of potential solutions to a problem.

Generate a paragraph that synthesizes the justifications for each solution. Identifies the most promising approach based on the synthesis.

```
In [ ]: import os
import openai
from dotenv import load_dotenv
load_dotenv()
openai.api_type = "azure"
# openai.api_version = "2023-03-15-preview"
openai.api_version = "2022-12-01"

API_KEY = os.getenv("OPENAI_API_KEY","").strip()
assert API_KEY, "ERROR: Azure OpenAI Key is missing"
openai.api_key = API_KEY

RESOURCE_ENDPOINT = os.getenv("OPENAI_API_BASE","").strip()
assert RESOURCE_ENDPOINT, "ERROR: Azure OpenAI Endpoint is missing"
assert "openai.azure.com" in RESOURCE_ENDPOINT.lower(), "ERROR: Azure OpenAI Endpoint should be in the"
openai.api_base = RESOURCE_ENDPOINT
```

```
In [ ]: potential_solutions="\
Solution 1: Increase funding for public schools.\n\
Solution 2: Implement a voucher system to allow parents to choose where their children attend school.\n\
Solution 3: Provide additional teacher training and resources to improve classroom instruction.\n\
Solution 4: Increase access to technology and online learning resources.\n\
"
```

Summary - before fine tuning and embedding...

Split

- Split complex tasks into sub tasks



Structure

- Structure instructions to guide the model through the execution of complex tasks



Explain

- Provide logical path / explanation and ask model to explain before answering



Justify

- Ask for justification of multiple answers and then synthesize



Choose

- Generate many outputs, and get model to choose the best

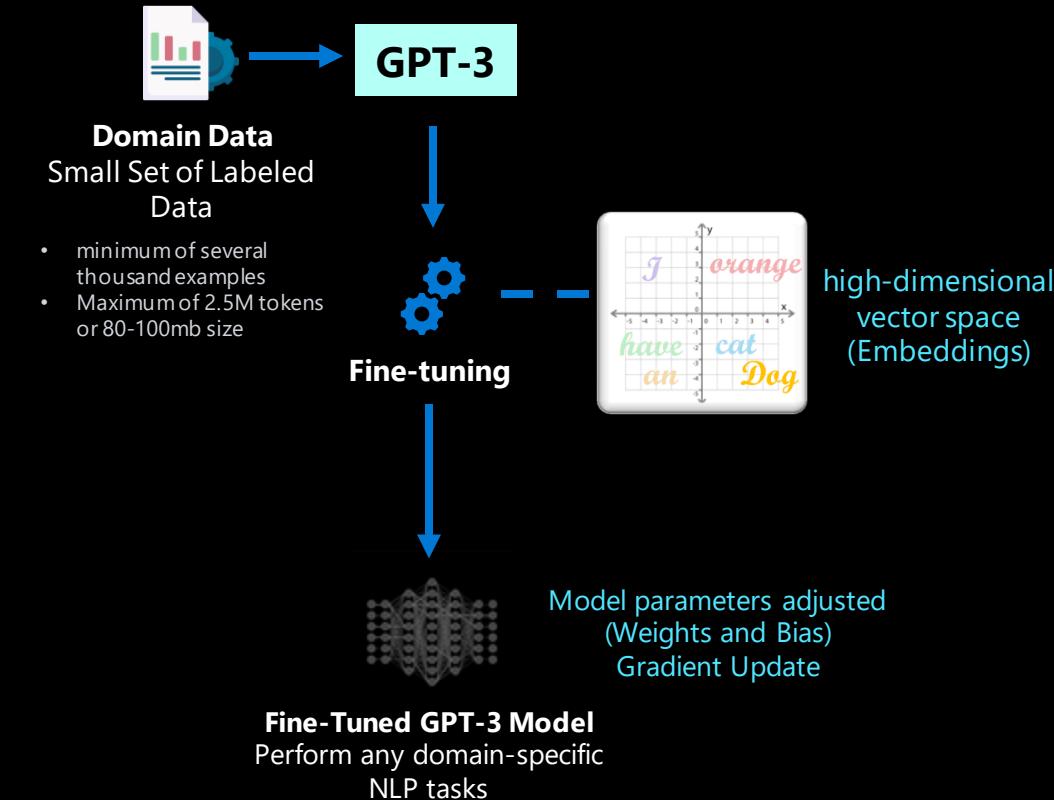


Appendix & Reference

- [openai-cookbook/techniques to improve reliability.md at main · openai/openai-cookbook · GitHub](#)
- [Best practices for prompt engineering with OpenAI API | OpenAI Help Center](#)
- <https://github.com/Azure/azure-openai-samples>
- [Azure OpenAI Service - Documentation, quickstarts, API reference - Azure Cognitive Services | Microsoft Learn](#)
- [\[2205.11916\] Large Language Models are Zero-Shot Reasoners \(arxiv.org\)](#)
- [Language Models Perform Reasoning via Chain of Thought – Google AI Blog \(googleblog.com\)](#)
- [\[2110.14168\] Training Verifiers to Solve Math Word Problems \(arxiv.org\)](#)
- [\[2207.10342\] Language Model Cascades \(arxiv.org\)](#)
- [\[2208.14271\] Faithful Reasoning Using Large Language Models \(arxiv.org\)](#)
- [\[2205.09712\] Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning \(arxiv.org\)](#)
- [\[2205.10625\] Least-to-Most Prompting Enables Complex Reasoning in Large Language Models \(arxiv.org\)](#)
- [\[2203.14465\] STaR: Bootstrapping Reasoning With Reasoning \(arxiv.org\)](#)
- [\[2205.11822\] Maieutic Prompting: Logically Consistent Reasoning with Recursive Explanations \(arxiv.org\)](#)
- [\[2203.11171\] Self-Consistency Improves Chain of Thought Reasoning in Language Models \(arxiv.org\)](#)

Model Adaptation with specific domain data

Fine-Tuning



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

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

When Fine-Tuning is needed



If model is making untrue statements ("hallucinations"), then mitigate the hallucinations

Accuracy of results of the model does not meet customer requirements

Fine-tuning lets you get more out of the models available through the API by providing:

- Higher quality results than prompt design
- Ability to train on more examples than can fit in a prompt
- Lower latency requests

Fine-tuning improves over few-shot learning by training on many more examples than can fit in the prompt, letting you achieve better results on a wide number of tasks.

Best practices of Fine-Tuning



Fine-tuning data set must be in JSON format

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

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

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

Best practices of Fine-Tuning



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

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

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

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

Best practices of Fine-Tuning



Make sure to completely remove wrong labels in fine-tuning dataset. If you are fine-tuning on a pre-existing dataset rather than writing prompts from scratch, be sure to manually review your data for offensive or inaccurate content if possible, or review as many random samples of the dataset as possible if it is large.

Fine-Tuning data formatting



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 the base models, where you might input detailed instructions or multiple examples in a single prompt.

Each prompt should end with a fixed separator to inform the model when the prompt ends and the completion begins.

A simple separator which generally works well is `\n\n###\n`.

The separator should not appear elsewhere in any prompt.

Fine-Tuning Data Formatting



Each completion should start with a whitespace due to tokenization, which tokenizes most words with a preceding whitespace.

Each completion should end with a fixed stop sequence to inform the model when the completion ends.

A stop sequence could be \n, ###, or any other token that does not appear in any completion.

For inference, you should format your prompts in the same way as you did when creating the training dataset, including the same separator.

Also specify the same stop sequence to properly truncate the completion.

Hyperparameters specific to Fine-Tuning

Parameter	Description	Recommendation
n_epochs controls how many times each example is trained on	The number of epochs to train the model for. An epoch refers to one full cycle through the training dataset.	Start from 4 and small datasets may need more epochs and large datasets may need fewer epochs. If you see low training accuracy (underfitting), try increasing n_epochs. If you see high training accuracy but low validation accuracy (overfitting), try lowering n_epochs.
batch_size controls the number of training examples used in a single training pass	The batch size to use for training. The batch size is the number of training examples used to train a single forward and backward pass.	We've seen good performance in the range of 0.01% to 4% of training set size. In general, we've found that larger batch sizes tend to work better for larger datasets.
learning_rate_multiplier controls rate at which the model weights are updated	The learning rate multiplier to use for training. The fine-tuning learning rate is the original learning rate used for pre-training, multiplied by this value.	We recommend experimenting with values in the range 0.02 to 0.2 to see what produces the best results. Empirically, we've found that larger learning rates often perform better with larger batch sizes. Empirically, we found learning_rate_multiplier has minor impact compared to n_epochs and batch_size.
prompt_loss_weight controls how much the model learns from prompt tokens vs completion tokens	The weight to use for loss on the prompt tokens. This value controls how much the model tries to learn to generate the prompt (as compared to the completion, which always has a weight of 1.0.) Increasing this value can add a stabilizing effect to training when completions are short.	When a model is fine-tuned, it learns to produce text it sees in both the prompt and the completion. In fact, from the point of view of the model being fine-tuned, the distinction between prompt and completion is mostly arbitrary. The only difference between prompt text and completion text is that the model learns less from each prompt token than it does from each completion token. This ratio is controlled by the prompt_loss_weight, which by default is 0.1. If prompts are extremely long (relative to completions), it may make sense to reduce this weight to avoid over-prioritizing learning the prompt. Empirically, we found prompt_loss_weight has minor impact compared to n_epochs and batch_size.

Guides and examples for Fine-Tuning

Getting started with Fine-Tuning:

[How to customize a model with Azure OpenAI - Azure OpenAI | Microsoft Learn](#)

Fine-tuning—OpenAI API

<https://platform.openai.com/docs/guides/fine-tuning/general-best-practices>

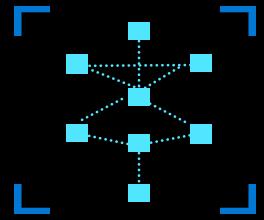
Detailed examples of fine-tuning

[\[PUBLIC\] Best practices for fine-tuning GPT-3 to classify text—Google Docs](#)

Using fine-tuning to improve truthfulness

[Fine-tuning a Classifier to Improve Truthfulness | OpenAI Help Center](#)

Embeddings



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

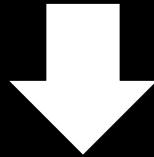
The embedding is an information dense representation of the semantic meaning of a piece of text.

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

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

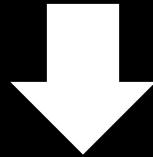
Embeddings make it possible to map content to a “semantic space”

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



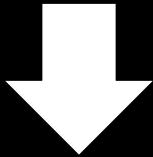
[15 34 24 13 ...]

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



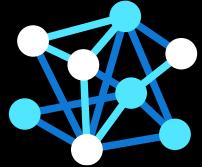
[16 22 89 26 ...]

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



[20 13 31 89 ...]

Embedding models



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

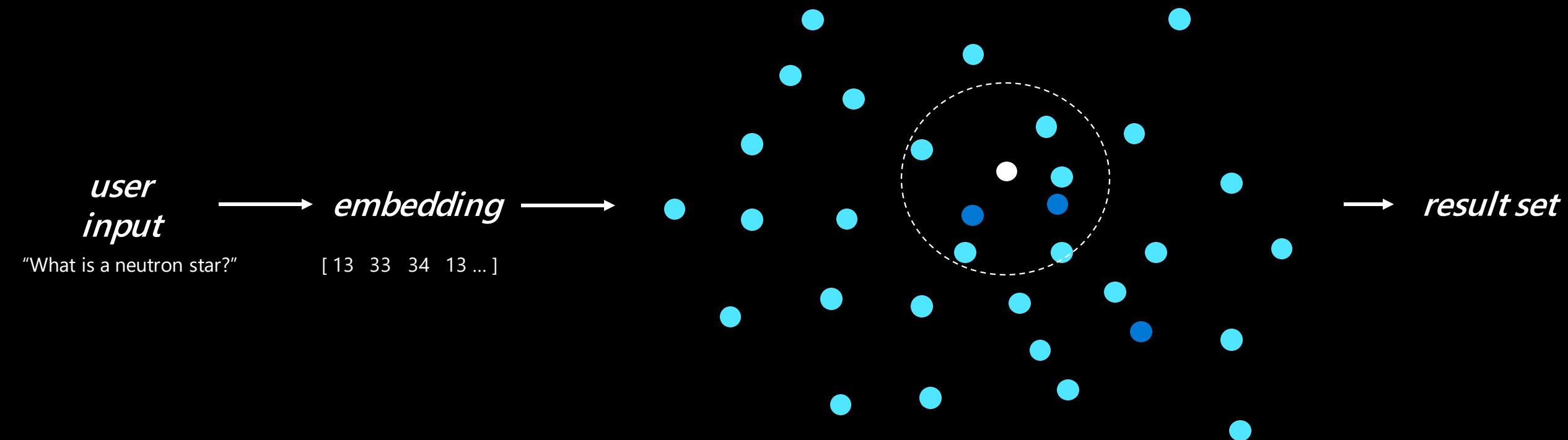
- **Similarity embeddings** are good at capturing semantic similarity between two or more pieces of text.
- **Text search embeddings** help measure long documents are relevant to a short query.
- **Code search embeddings** are useful for embedding code snippets and embedding nature language search queries.

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

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

Similarity Search with embeddings

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



bill_id	text	summary
0 110_hr37	SECTION 1. SHORT TITLE. This Act may be cited as the ``National Science Education Tax Incentive Act of 2010''.	National Science Education Tax Incentive for Business
1 112_hr2873	SECTION 1. SHORT TITLE. This Act may be cited as the ``Small Business Expansion and Hiring Act of 2011 - American Job Protection Act of 2011''.	Small Business Expansion and Hiring Act of 2011 - Am
2 109_s2408	SECTION 1. RELEASE OF DOCUMENTS CAPTURED IN IRAQ AND AFGHANISTAN. (a) In General. Requires the Director of National Intelligence to make available to the public documents captured in Iraq and Afghanistan.	Requires the Director of National Intelligence to make available to the public documents captured in Iraq and Afghanistan.
3 108_s1899	SECTION 1. SHORT TITLE. This Act may be cited as the ``National Cancer Act of 2003''.SEC. 2. I. National Cancer Act of 2003 - Amends the Public Health Service Act.	National Cancer Act of 2003 - Amends the Public Health Service Act.
4 107_s1531	SECTION 1. SHORT TITLE. This Act may be cited as the ``Military Call-Up Relief Act''.SEC. 2. W. Military Call-up Relief Act - Amends the Internal Revenue Code.	Military Call-up Relief Act - Amends the Internal Revenue Code.
5 107_hr4541	SECTION 1. RELIQUIDATION OF CERTAIN ENTRIES PREMATURELY LIQUIDATED IN ERROR. Requires the Customs Service to reliquidate certain entries.	Requires the Customs Service to reliquidate certain entries.
6 111_s1495	SECTION 1. SHORT TITLE. This Act may be cited as the ``Service Dogs for Veterans Act of 2009''.Service Dogs for Veterans Act of 2009 - Directs the Secretary of Veterans Affairs to...	Service Dogs for Veterans Act of 2009 - Directs the Secretary of Veterans Affairs to...
7 111_s3885	SECTION 1. SHORT TITLE. This Act may be cited as the ``Race to the Top Act of 2010''.SEC. 2. R. Race to the Top Act of 2010 - Directs the Secretary of Education to...	Race to the Top Act of 2010 - Directs the Secretary of Education to...
8 113_hr1796	SECTION 1. SHORT TITLE. This Act may be cited as the ``Troop Talent Act of 2013''.SEC. 2. FINI. Troop Talent Act of 2013 - Directs the Secretaries of the...	Troop Talent Act of 2013 - Directs the Secretaries of the...
9 103_hr1987	SECTION 1. SHORT TITLE. This Act may be cited as the ``Taxpayer's Right To View Act of 1993''.Taxpayer's Right to View Act of 1993 - Amends the Code of Federal...	Taxpayer's Right to View Act of 1993 - Amends the Code of Federal...
10 103_hr1677	SECTION 1. SHORT TITLE. This Act may be cited as the ``Full-Service Schools Act''.SEC. 2. FINI. Full-Service Schools Act - Establishes the Federal Interagency...	Full-Service Schools Act - Establishes the Federal Interagency...
11 111_s3149	SECTION 1. SHORT TITLE. This Act may be cited as the ``Wall Street Compensation Reform Act''.Wall Street Compensation Reform Act of 2010 - Amends the Code of...	Wall Street Compensation Reform Act of 2010 - Amends the Code of...
12 110_hr1007	SECTION 1. FINDINGS. The Congress finds the following: (1) The goal of the Marine Mammal Protection Act of 1972 to...	Amends the Marine Mammal Protection Act of 1972 to...
13 113_hr3137	SECTION 1. SHORT TITLE. This Act may be cited as the ``Freedom and Mobility in Consumer Banking Act''.Freedom and Mobility in Consumer Banking Act - Amends the...	Freedom and Mobility in Consumer Banking Act - Amends the...
14 115_hr1634	SECTION 1. SHORT TITLE. This Act may be cited as the ``Education and Training for Health Act''.Education and Training for Health Act of 2017 or the E...	Education and Training for Health Act of 2017 or the E...
15 103_hr1815	SECTION 1. SHORT TITLE. This Act may be cited as the ``Recreational Hunting Safety and Preservation Act''.Recreational Hunting Safety and Preservation Act of 2017 -...	Recreational Hunting Safety and Preservation Act of 2017 -...
16 113_s1773	SECTION 1. SHORT TITLE. This Act may be cited as the ``Andrew Prior Act'' or ``Andrew's Law''.Andrew Prior Act or Andrew's Law - Amends the Truth in...	Andrew Prior Act or Andrew's Law - Amends the Truth in...
17 106_hr5585	SECTION 1. SHORT TITLE. This Act may be cited as the ``Energy Independence Act of 2000''.S. Directs the President, in coordination with designated...	Energy Independence Act of 2000''.S. Directs the President, in coordination with designated...
18 114_hr2499	SECTION 1. SHORT TITLE. This Act may be cited as the ``Veterans Entrepreneurship Act of 2011''.This measure has not been amended since it was passed.	Veterans Entrepreneurship Act of 2011''.This measure has not been amended since it was passed.
19 111_hr3141	SECTION 1. SHORT TITLE. This Act may be cited as the ``Strengthening the Health Care Safety Net Act of 2009''.Strengthening the Health Care Safety Net Act of 2009 -...	Strengthening the Health Care Safety Net Act of 2009 -...

```
df_bills['ada_v2']
```

✓ 0.0s

```

0 [-0.022678455337882042, -0.00799493957310915, ...]
1 [-0.041596777737140656, -0.009042778052389622, ...]
2 [-0.0420723482966423, -0.0023871997836977243, ...]
3 [-0.013265267945826054, -0.007186719682067633, ...]
4 [-0.03917262703180313, -0.025464843958616257, ...]
5 [-0.016097435727715492, -0.015194901265203953, ...]
6 [-0.029804326593875885, -0.02193712815642357, ...]
7 [-0.020575817674398422, 0.0003083075280301273, ...]
8 [-0.030029239133000374, -0.0169113427400589, 0...]
9 [-0.018460797145962715, -0.024794351309537888, ...]
10 [-0.01014762744307518, 0.015381667762994766, ...]
11 [-0.047416988760232925, -0.007824325934052467, ...]
12 [-0.025547074154019356, -0.0012037496780976653...]
13 [-0.024552075192332268, 0.002101766411215067, ...]
14 [-0.0032339368481189013, -0.000158853552420623...]
15 [-0.022656632587313652, -0.007432988379150629, ...]
16 [-0.033216603100299835, -0.010047564283013344, ...]
17 [0.0029555223882198334, -0.024077415466308594, ...]
18 [-0.02254152111709118, -0.030212445184588432, ...]
19 [-0.022656632587313652, -0.007432988379150629, ...]
20 [-0.022656632587313652, -0.007432988379150629, ...]
21 [-0.022656632587313652, -0.007432988379150629, ...]
22 [-0.022656632587313652, -0.007432988379150629, ...]
23 [-0.022656632587313652, -0.007432988379150629, ...]
24 [-0.022656632587313652, -0.007432988379150629, ...]
25 [-0.022656632587313652, -0.007432988379150629, ...]
26 [-0.022656632587313652, -0.007432988379150629, ...]
27 [-0.022656632587313652, -0.007432988379150629, ...]
28 [-0.022656632587313652, -0.007432988379150629, ...]
29 [-0.022656632587313652, -0.007432988379150629, ...]
30 [-0.022656632587313652, -0.007432988379150629, ...]
31 [-0.022656632587313652, -0.007432988379150629, ...]
32 [-0.022656632587313652, -0.007432988379150629, ...]
33 [-0.022656632587313652, -0.007432988379150629, ...]
34 [-0.022656632587313652, -0.007432988379150629, ...]
35 [-0.022656632587313652, -0.007432988379150629, ...]
36 [-0.022656632587313652, -0.007432988379150629, ...]
37 [-0.022656632587313652, -0.007432988379150629, ...]
38 [-0.022656632587313652, -0.007432988379150629, ...]
39 [-0.022656632587313652, -0.007432988379150629, ...]
40 [-0.022656632587313652, -0.007432988379150629, ...]
41 [-0.022656632587313652, -0.007432988379150629, ...]
42 [-0.022656632587313652, -0.007432988379150629, ...]
43 [-0.022656632587313652, -0.007432988379150629, ...]
44 [-0.022656632587313652, -0.007432988379150629, ...]
45 [-0.022656632587313652, -0.007432988379150629, ...]
46 [-0.022656632587313652, -0.007432988379150629, ...]
47 [-0.022656632587313652, -0.007432988379150629, ...]
48 [-0.022656632587313652, -0.007432988379150629, ...]
49 [-0.022656632587313652, -0.007432988379150629, ...]
50 [-0.022656632587313652, -0.007432988379150629, ...]
51 [-0.022656632587313652, -0.007432988379150629, ...]
52 [-0.022656632587313652, -0.007432988379150629, ...]
53 [-0.022656632587313652, -0.007432988379150629, ...]
54 [-0.022656632587313652, -0.007432988379150629, ...]
55 [-0.022656632587313652, -0.007432988379150629, ...]
56 [-0.022656632587313652, -0.007432988379150629, ...]
57 [-0.022656632587313652, -0.007432988379150629, ...]
58 [-0.022656632587313652, -0.007432988379150629, ...]
59 [-0.022656632587313652, -0.007432988379150629, ...]
60 [-0.022656632587313652, -0.007432988379150629, ...]
61 [-0.022656632587313652, -0.007432988379150629, ...]
62 [-0.022656632587313652, -0.007432988379150629, ...]
63 [-0.022656632587313652, -0.007432988379150629, ...]
64 [-0.022656632587313652, -0.007432988379150629, ...]
65 [-0.022656632587313652, -0.007432988379150629, ...]
66 [-0.022656632587313652, -0.007432988379150629, ...]
67 [-0.022656632587313652, -0.007432988379150629, ...]
68 [-0.022656632587313652, -0.007432988379150629, ...]
69 [-0.022656632587313652, -0.007432988379150629, ...]
70 [-0.022656632587313652, -0.007432988379150629, ...]
71 [-0.022656632587313652, -0.007432988379150629, ...]
72 [-0.022656632587313652, -0.007432988379150629, ...]
73 [-0.022656632587313652, -0.007432988379150629, ...]
74 [-0.022656632587313652, -0.007432988379150629, ...]
75 [-0.022656632587313652, -0.007432988379150629, ...]
76 [-0.022656632587313652, -0.007432988379150629, ...]
77 [-0.022656632587313652, -0.007432988379150629, ...]
78 [-0.022656632587313652, -0.007432988379150629, ...]
79 [-0.022656632587313652, -0.007432988379150629, ...]
80 [-0.022656632587313652, -0.007432988379150629, ...]
81 [-0.022656632587313652, -0.007432988379150629, ...]
82 [-0.022656632587313652, -0.007432988379150629, ...]
83 [-0.022656632587313652, -0.007432988379150629, ...]
84 [-0.022656632587313652, -0.007432988379150629, ...]
85 [-0.022656632587313652, -0.007432988379150629, ...]
86 [-0.022656632587313652, -0.007432988379150629, ...]
87 [-0.022656632587313652, -0.007432988379150629, ...]
88 [-0.022656632587313652, -0.007432988379150629, ...]
89 [-0.022656632587313652, -0.007432988379150629, ...]
90 [-0.022656632587313652, -0.007432988379150629, ...]
91 [-0.022656632587313652, -0.007432988379150629, ...]
92 [-0.022656632587313652, -0.007432988379150629, ...]
93 [-0.022656632587313652, -0.007432988379150629, ...]
94 [-0.022656632587313652, -0.007432988379150629, ...]
95 [-0.022656632587313652, -0.007432988379150629, ...]
96 [-0.022656632587313652, -0.007432988379150629, ...]
97 [-0.022656632587313652, -0.007432988379150629, ...]
98 [-0.022656632587313652, -0.007432988379150629, ...]
99 [-0.022656632587313652, -0.007432988379150629, ...]
100 [-0.022656632587313652, -0.007432988379150629, ...]

```

```

# search through the reviews for a specific product
def search_docs(df, user_query, top_n=3, to_print=True):
    embedding = get_embedding(
        user_query,
        engine="text-embedding-ada-002" # engine should be set to the deployment name you chose when you deployed the text-embedding-ada-002 (Version 2) model
    )
    df["similarities"] = df.ada_v2.apply(lambda x: cosine_similarity(x, embedding))

    res = (
        df.sort_values("similarities", ascending=False)
        .head(top_n)
    )
    if to_print:
        display(res)
    return res

res = search_docs(df_bills, "Are Veteran entitled to be provided dogs", top_n=4)

```

✓ 0.4s

	text	summary	title	n_tokens	ada_v2	similarities
6	SECTION 1. SHORT TITLE. This Act may be cited ...	Service Dogs for Veterans Act of 2009 - Direct...	A bill to require the Secretary of Veterans Af...	880	[-0.029804326593875885, -0.02193712815642357, ...	0.843258
8	SECTION 1. SHORT TITLE. This Act may be cited ...	Troop Talent Act of 2013 - Directs the Secreta...	Troop Talent Act of 2013	2479	[-0.030029239133000374, -0.0169113427400589, 0...	0.780919



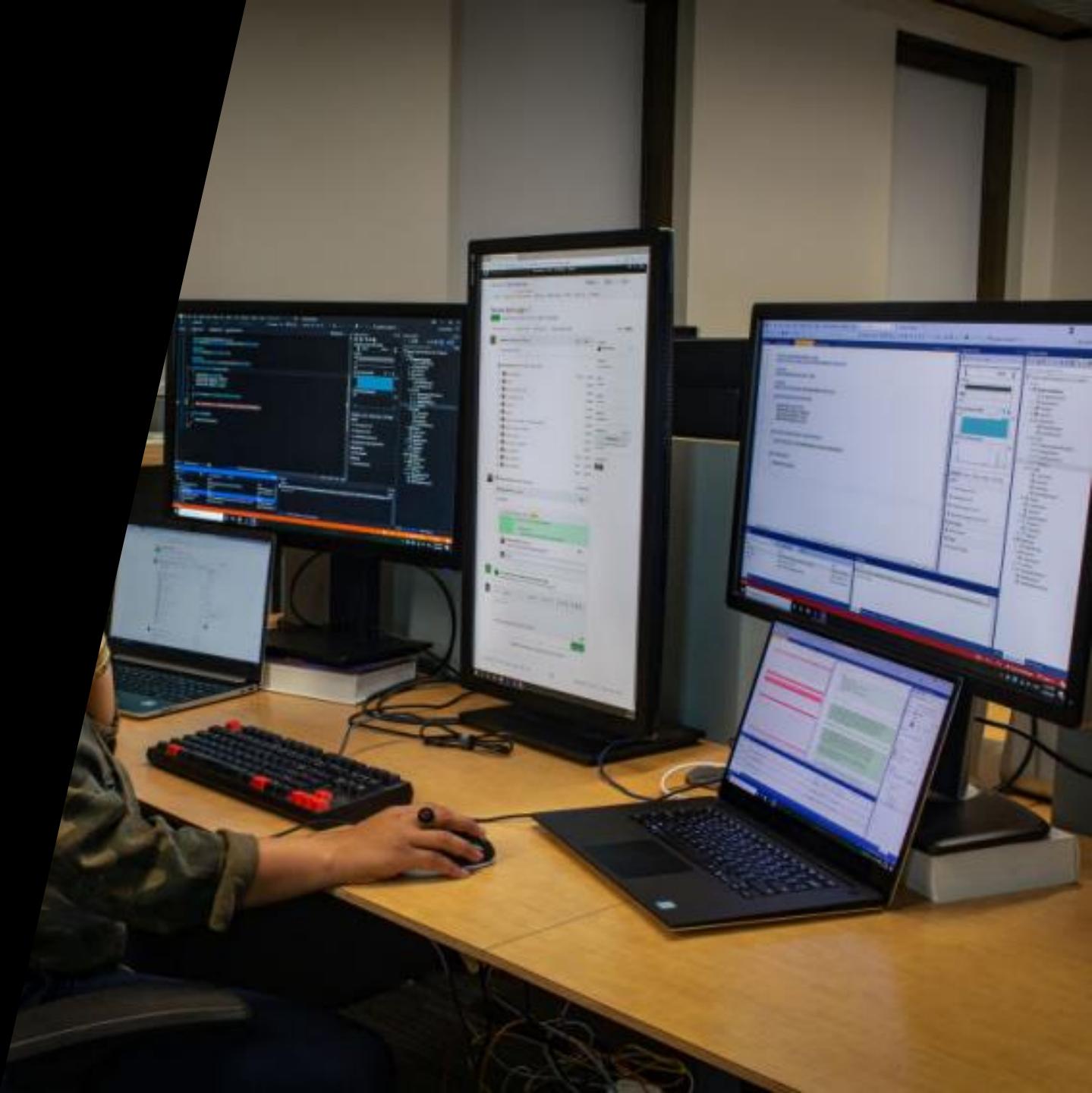
Orchestration



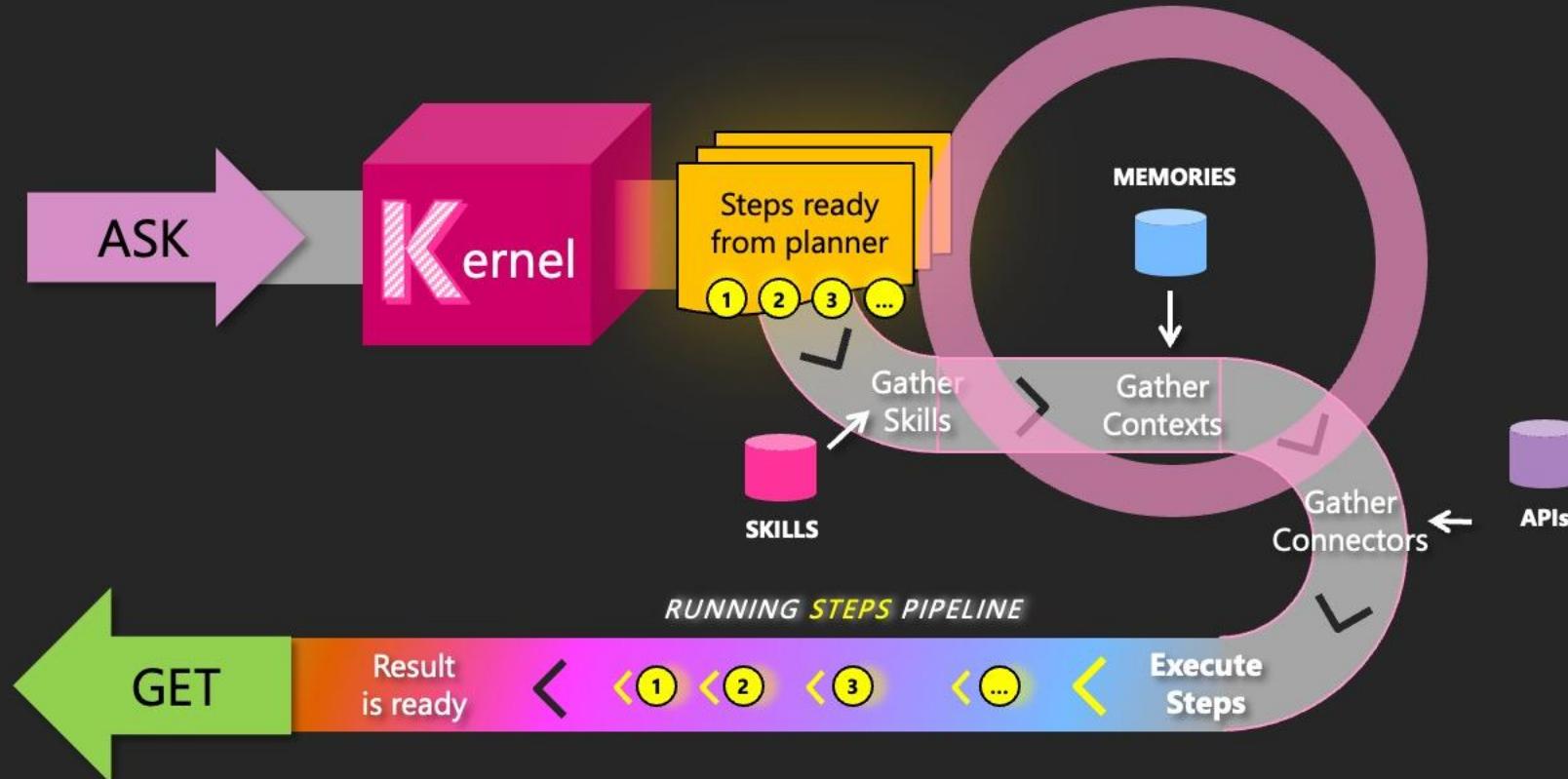
Leverage vs. Precision

Semantic Kernels

"is a core engine that processes a user's goal-oriented query (ASK) using natural language semantic functions, such as LLM AI prompts. The kernel also orchestrates the planning, execution, and fulfilment of the ASK using available skills, memories, and connectors (...) The semantic kernel represents a shift from **syntax to semantics** in programming.



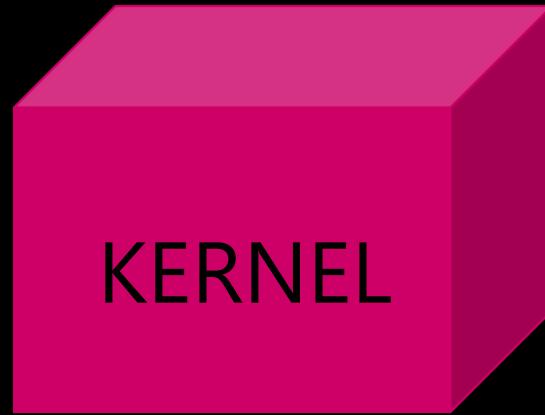
It all starts with
a user's AI ask ...



... resulting in
new productivity

aka.ms/semantic-kernel

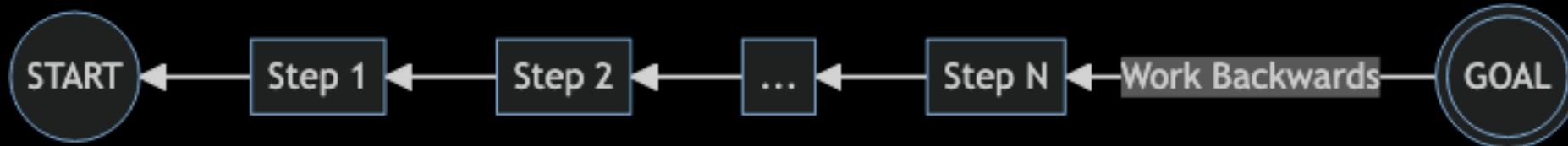
Kernel



The kernel orchestrates a user's ASK expressed as a goal

*An instance of the processing engine that processes an ASK
all the way through to fulfilment.*

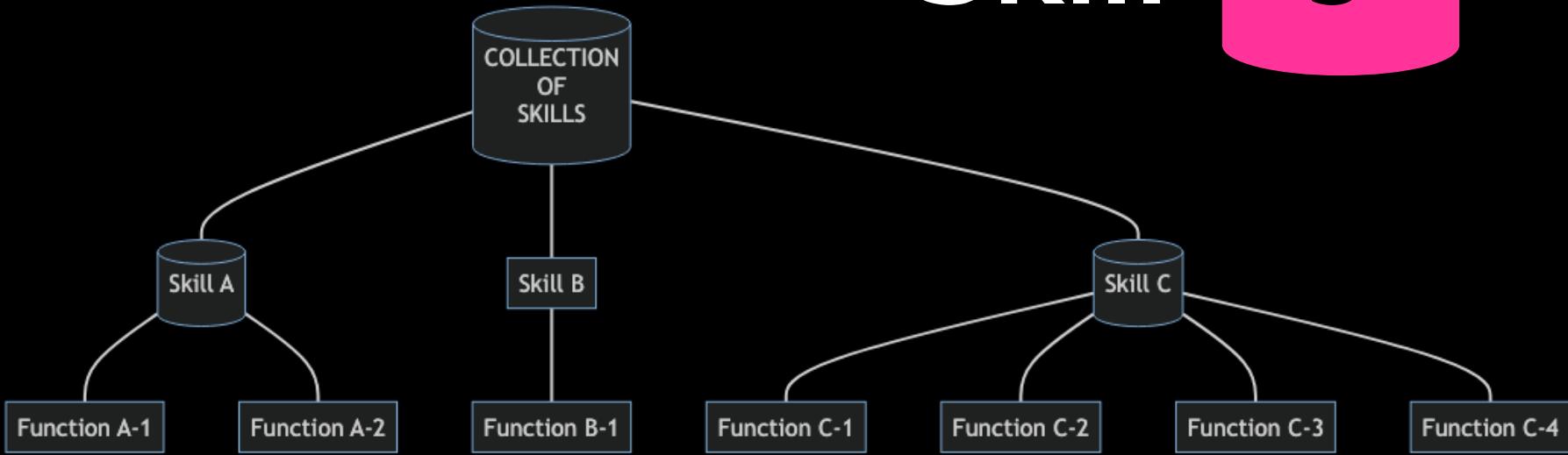
Planner



The planner breaks it down into steps based upon resources that are available

The planner works backwards from a goal that's provided from a user's ASK.

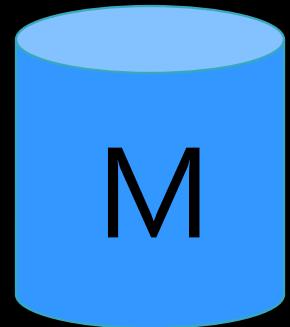
Skills



Skills are customizable resources built from LLM AI prompts and native code

A skill refers to a domain of expertise made available to the kernel as a single function, or as a group of functions related to the skill

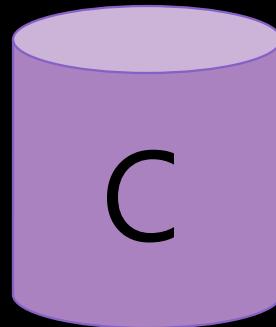
Memories



Memories are customizable resources that manage contextual information

1. *Conventional key-value pairs: Just like you would set an environment variable in your shell, the same can be done when using SK. The lookup is "conventional" because it's a one-to-one match between a key and your query.*
2. *Conventional local-storage: When you save information to a file, it can be retrieved with its filename. When you have a lot of information to store in a key-value pair, you're best off keeping it on disk.*
3. *Semantic memory search: You can also represent text information as a long vector of numbers, known as "embeddings." This lets you execute a "semantic" search that compares meaning-to-meaning with your query.*

Connectors



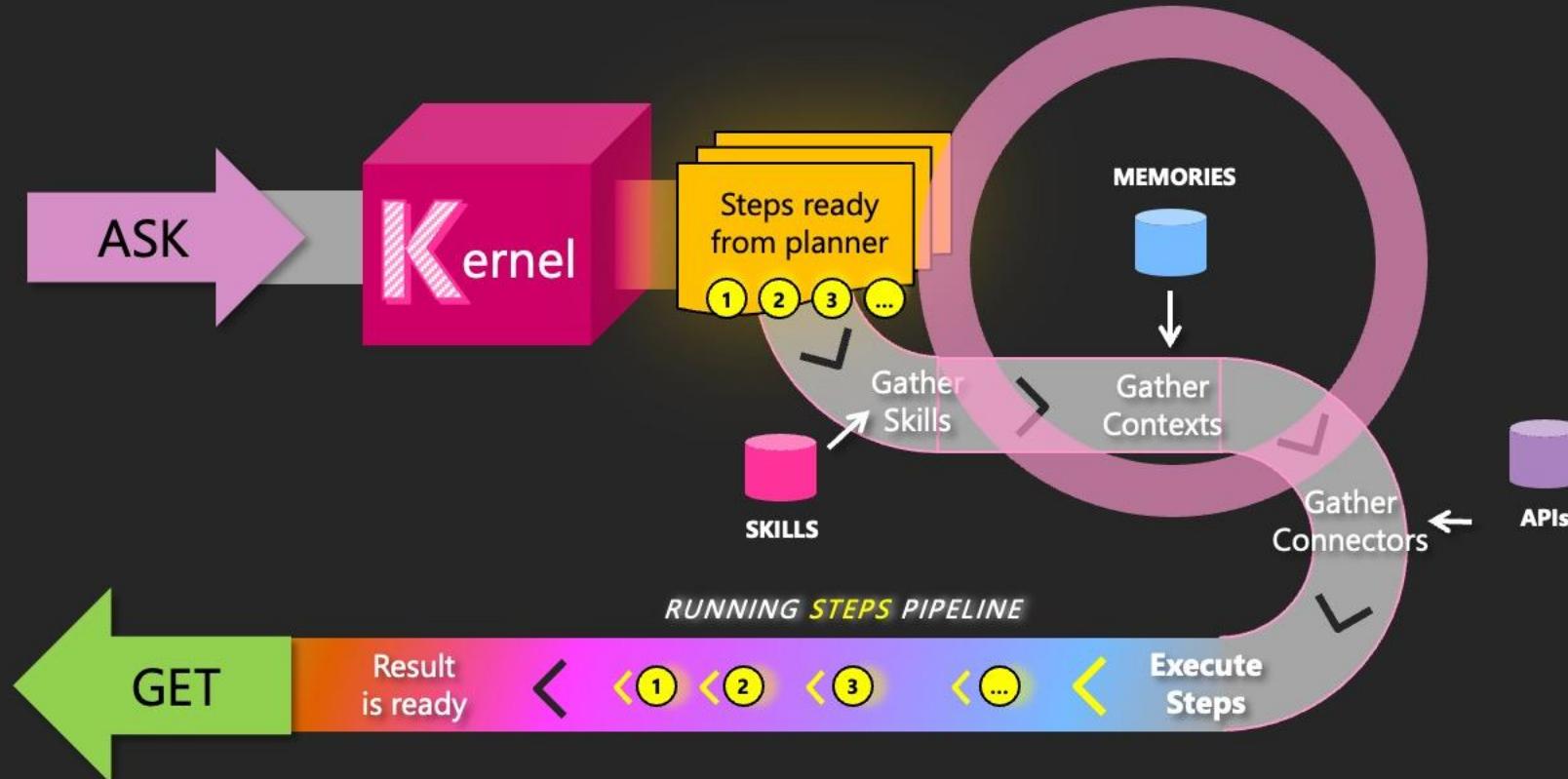
Connectors are customizable resources that enable external data access

Connectors let you reach outside of the skills universe to external APIs and whatever else you can imagine.

For instance you could:

- *Add an event to your calendar*
- *Send an email for you*
- *Add a file to your OneDrive*
- *Create a share link to a file in your OneDrive*
- *Query your organization hierarchy*
- *Issue a Bing search query*

It all starts with
a user's AI ask ...



... resulting in
new productivity

aka.ms/semantic-kernel

Schillace Laws

1. Don't write code if the model can do it; the model will get better, but the code won't.
2. Trade leverage for precision; use interaction to mitigate.
3. Code is for syntax and process; models are for semantics and intent.
4. The system will be as brittle as its most brittle part.
5. Ask Smart to Get Smart.
6. Uncertainty is an exception throw.
7. Text is the universal wire protocol.
8. Hard for you is hard for the model.
9. Beware pareidolia of consciousness; the model can be used against itself.

**Microsoft's Deputy CTO
Sam Schillace**

<https://aka.ms/semantic-kernel>

<https://devblogs.microsoft.com/semantic-kernel/recipes/>

LangChain

01 - SUMMARIZATION

the first time in the history of the world, the
whole of the human race has been gathered
together in one place, and that is the
present meeting of the World's Fair.

The World's Fair is a great exhibition of
the products of all the nations of the world.

02 - Q&A OVER DOCUMENTS



03 - EXTRACTION



04 - EVALUATION



05 – QUERY TABULAR DATA



06 – CODE UNDERSTANDING



07 – INTERACTING WITH APIs

1000

1000

08 – CHATBOTS

ANSWER

1. *What is the primary purpose of the U.S. Constitution?*

2. *What is the Bill of Rights?*

3. *What is the difference between a state and a nation?*

4. *What is the difference between a state and a nation?*

5. *What is the difference between a state and a nation?*

6. *What is the difference between a state and a nation?*

7. *What is the difference between a state and a nation?*

8. *What is the difference between a state and a nation?*

9. *What is the difference between a state and a nation?*

10. *What is the difference between a state and a nation?*

11. *What is the difference between a state and a nation?*

12. *What is the difference between a state and a nation?*

13. *What is the difference between a state and a nation?*

14. *What is the difference between a state and a nation?*

15. *What is the difference between a state and a nation?*

16. *What is the difference between a state and a nation?*

17. *What is the difference between a state and a nation?*

18. *What is the difference between a state and a nation?*

19. *What is the difference between a state and a nation?*

20. *What is the difference between a state and a nation?*

21. *What is the difference between a state and a nation?*

22. *What is the difference between a state and a nation?*

23. *What is the difference between a state and a nation?*

24. *What is the difference between a state and a nation?*

25. *What is the difference between a state and a nation?*

26. *What is the difference between a state and a nation?*

27. *What is the difference between a state and a nation?*

28. *What is the difference between a state and a nation?*

29. *What is the difference between a state and a nation?*

30. *What is the difference between a state and a nation?*

09 – AGENTS





10 – DEMO



三國志：真田義定の復讐

三國志：真田義定の復讐

A A A P

Let's build.

- 1** Your PTS & PDM will reach out to you after the workshop and set up 1:1 sessions with your product teams to discuss Azure OpenAI use cases for your products and customers.
- 2** You'll get invites for deep dive sessions on Azure OpenAI and your PTS will enable you to develop a PoC for your pilot customers.
- 3** There will be regular Office Hours to support you with technical questions and challenges.