Microsoft

Microsoft Azure AI Services: Introduction

Martin Abrle, Partner Solution Architect
Juan Manuel Servera, Sr. Partner Solution Architect

Partner Organization, Microsoft Switzerland

Agenda

Cognitive services

Azure Open AI

AI Discovery Cards (afternoon)

Azure AI Services

Microsoft AI

Turing
Rich language understanding

Z-Code
100 languages translation

Florence
Breakthrough visual recognition

OpenAI

GPT-3/GPT-4
Human-like language generation

DALL-E
Realistic image generation

Codex
Advanced code generation

ChatGPT
Conversation generation

Azure AI services

- Vision
- Speech
- Language
- Decision
- OpenAI Service
- AI Search
- Form Recognizer
- Immersive Reader
- Bot Service
- Video Analyzer

Microsoft AI portfolio

Business Users & Citizen Developers

Developers & Data Scientists

Applications And Copilot

Microsoft 365 | Microsoft Dynamics 365 | Microsoft Edge | Microsoft Bing | Windows | XBOX

Power Platform

Fabric | Power Apps | Copilot Studio | Power Automate | Power Virtual Agents

Azure AI

Azure AI services

Bot Service | Azure AI Search | Document Intelligence | Video Indexer | Content Safety | Decision | Metrics Advisor | Immersive Reader

Vision | Face | Speech | Translation | Language | Content Understanding | OpenAI Service

ML Platform

Azure Databricks | Azure Synapse | Azure Machine Learning | Fabric

Gen AI Platform

Azure AI Foundry

Dev Platform

GitHub Copilot

Innovate responsibly

Responsible AI principles

- Fairness
- Reliability & Safety
- Privacy & Security
- Inclusiveness
- Transparency
- Accountability

Microsoft improves facial recognition technology to perform well across all skin tones, genders

June 26, 2018 | John Roach

Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI

Michael Madaio, Luke Stark, Jennifer Wortman Vaughan, Hanna Wallach
CHI Conference on Human Factors in Computing Systems | March 2020
Organized by ACM
CHI 2020 Best Paper Award

Download BibTex

Many organizations have published principles intended to guide the ethical development and deployment of AI systems; however, their abstract nature makes them difficult to operationalize. Some organizations have therefore produced AI ethics checklists, as well as checklists for more specific concepts, such as fairness, as applied to AI systems. But unless checklists are grounded in practitioners' needs, they may be misused. To understand the role of checklists in AI ethics, we conducted an iterative co-design process with 48 practitioners, focusing on fairness. We co-designed an AI fairness checklist and identified desiderata and concerns for AI fairness checklists in general. We found that AI fairness checklists could provide organizational infrastructure for formalizing ad-hoc processes and empowering individual advocates. We discuss aspects of organizational culture that may impact the efficacy of such checklists, and highlight future research directions.

Download PDF

Groups

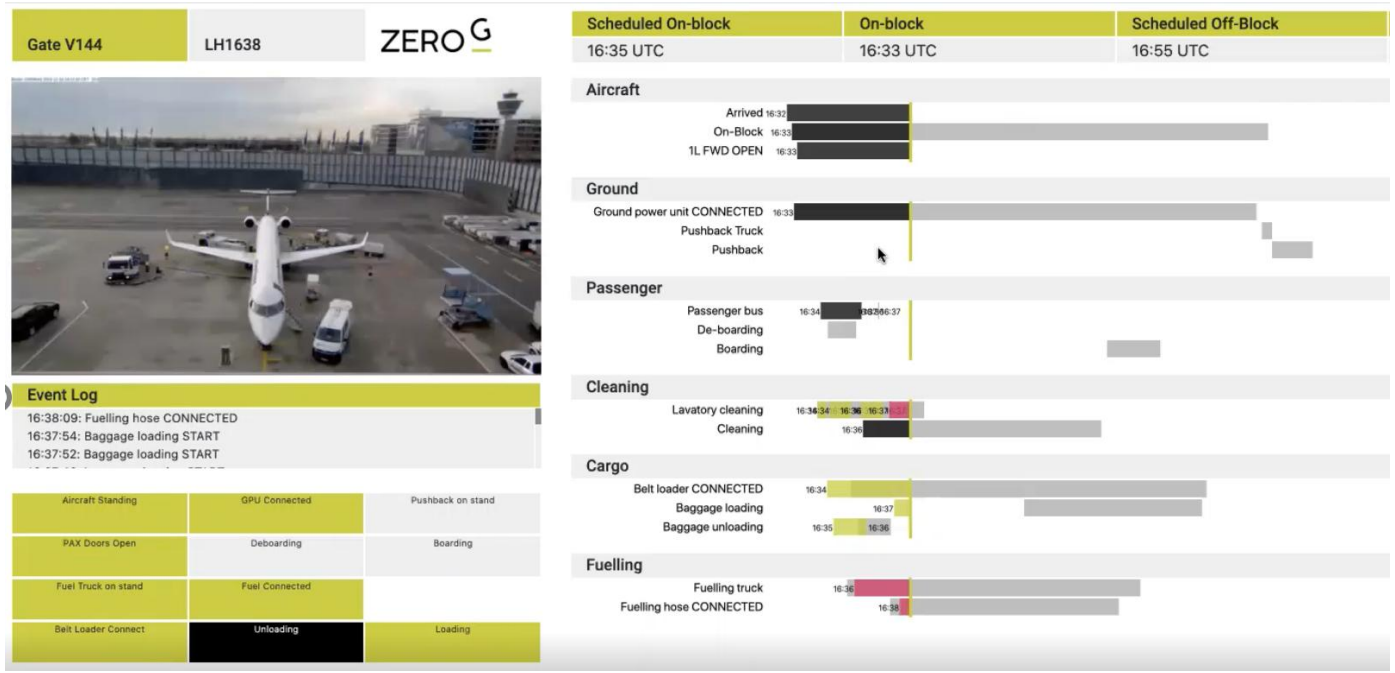
FATE: Fairness, Accountability, Transparency, and Ethics in AI
FATE | Montréal

Solving the challenge of securing AI and machine learning systems

Dec 6, 2019 | Valecia Maclin - General Manager, Engineering - Customer Security & Trust

Today, in collaboration with Harvard University's Berkman Klein Center, we at Microsoft are publishing a series of materials we believe will contribute to solving a major challenge to securing artificial intelligence and machine learning systems. In short, there is no common terminology today to discuss security threats to these systems and methods to mitigate them, and we hope these new materials will provide baseline language that will enable the research community to better collaborate.

Azure Video Analyzer



Gather and manage video insights

Tailor to your needs

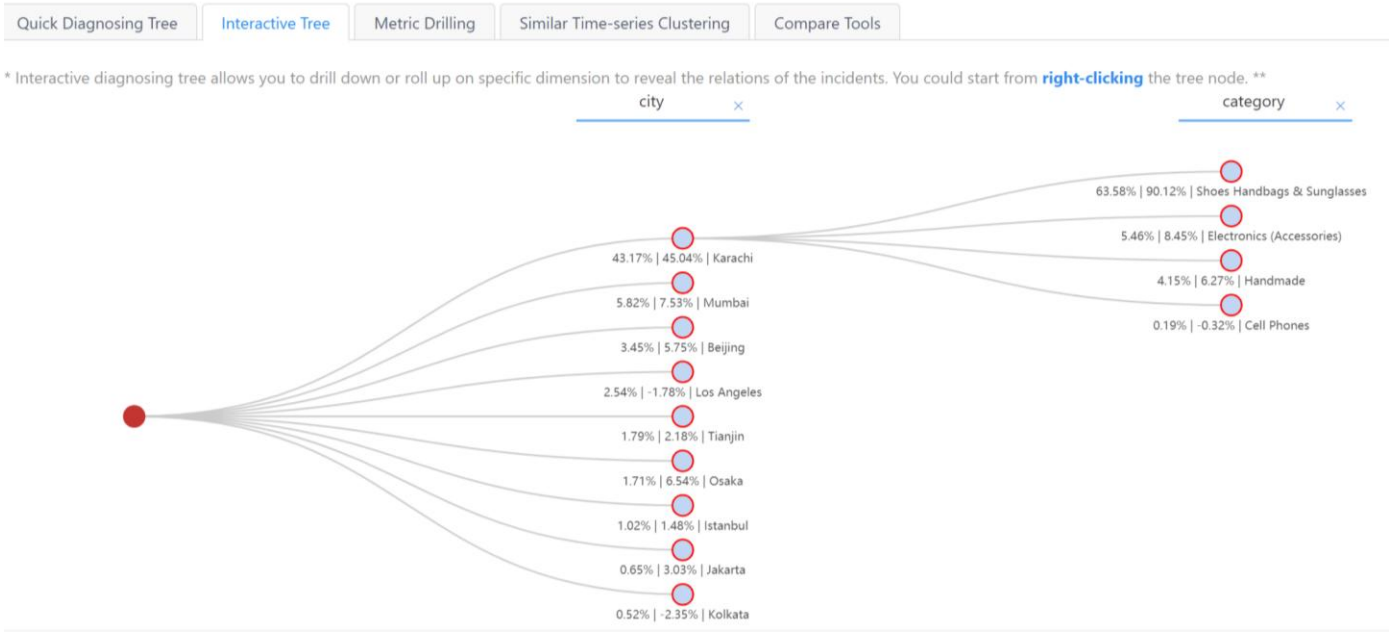
Build in days, not months

USE CASES

Process optimization	Workplace safety	Digital asset management

Azure Metrics Advisor

Proactively monitor metrics and diagnose issues

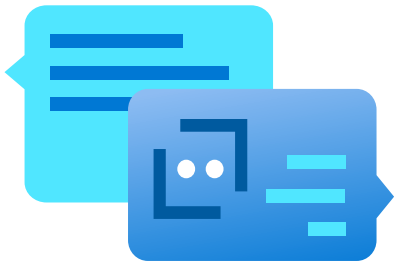


Detect anomalies for a variety of scenarios

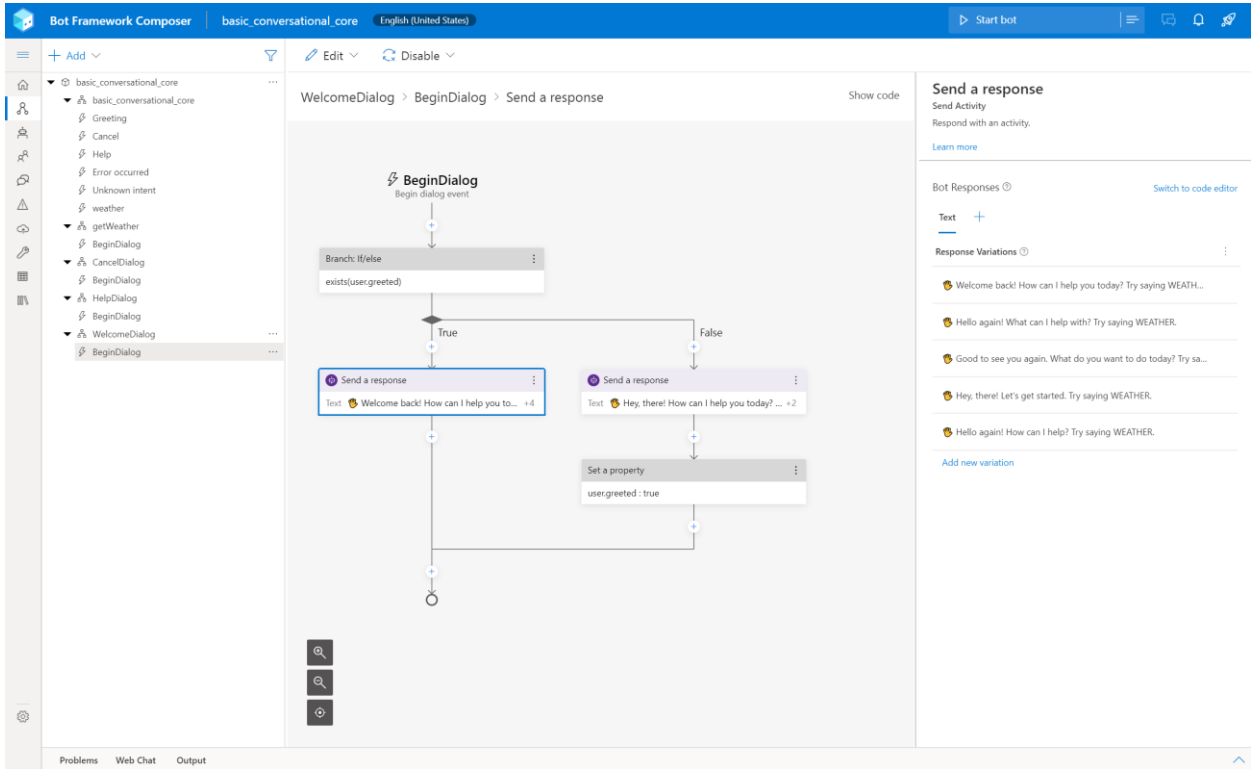
Focus on fixing rather than monitoring

Accelerate time to value

USE CASES		
Business metrics monitoring	AIOps	Predictive maintenance



Azure Bot Service



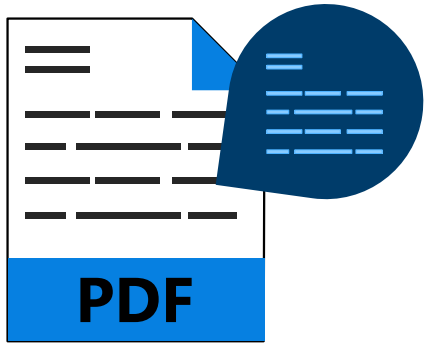
Accelerate bot building

Deploy to channels with minimal code changes

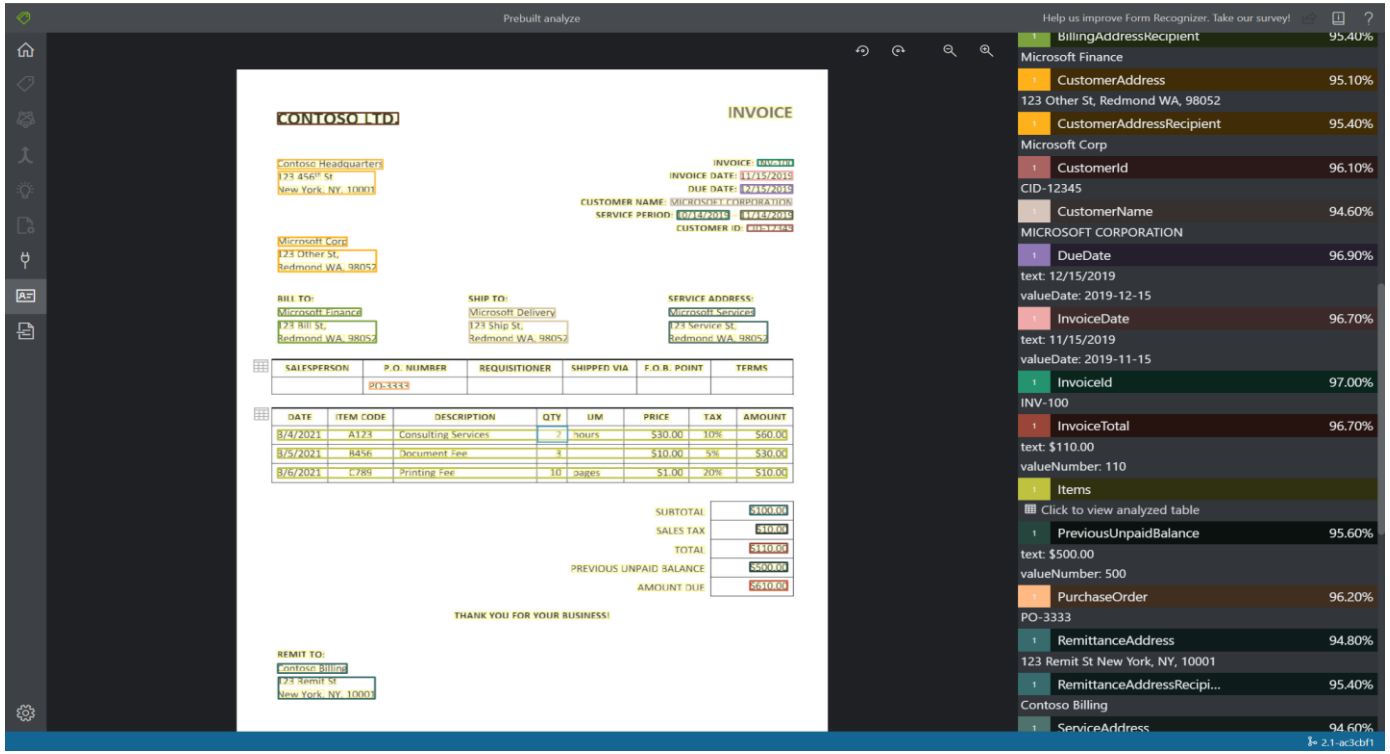
Conversational AI for healthcare

USE CASES

Customer service assistant	**Smart personalized care**	**Enable voice control** (voice assistant)



Azure AI Document Intelligence



Extract text and structure with simple API

Customize extraction to your forms

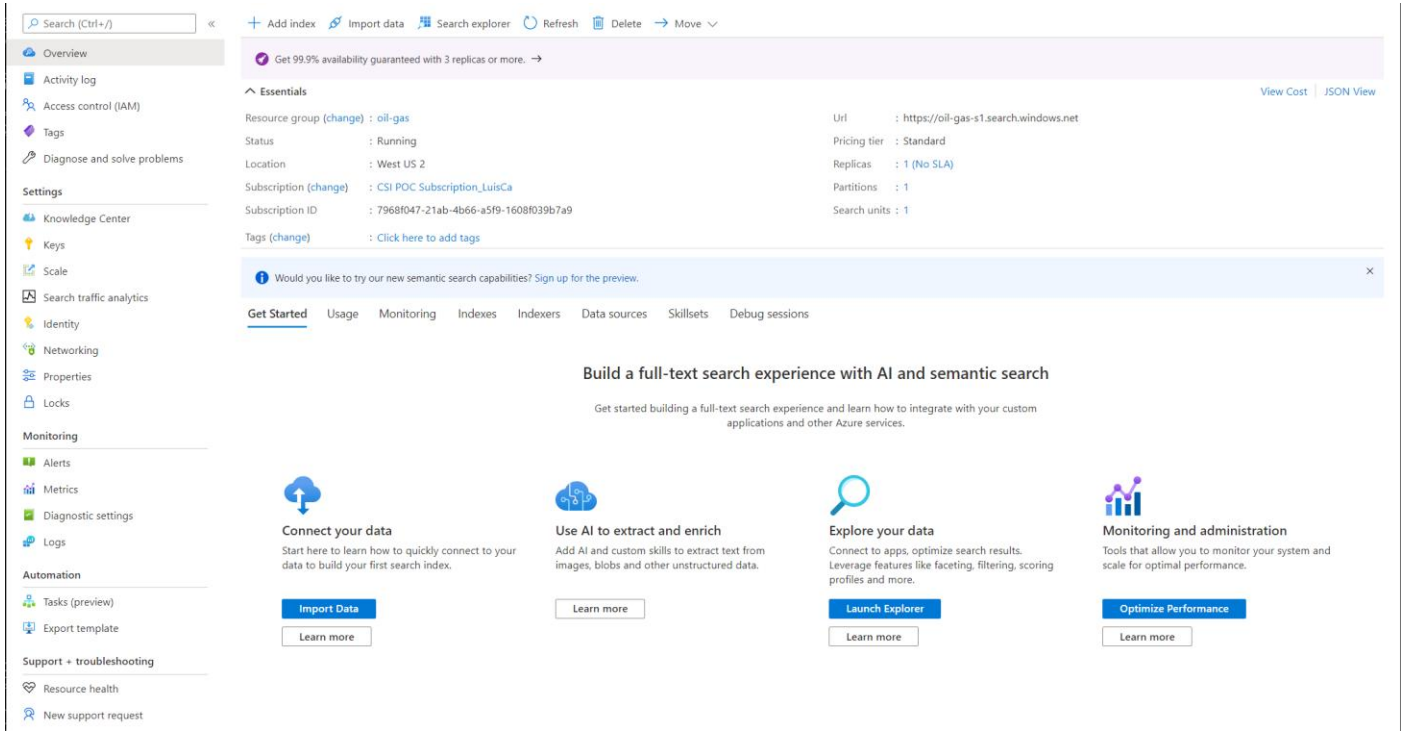
Deploy anywhere, from cloud to edge

Export to Markdown

USE CASES

Claims management and automation	Document process automation	Government forms automation

Azure AI Search



Start, maintain, and scale with minimal investment

Create searchable content using integrated AI

Search with user intent and not just keywords

Hybrid Search for RAG

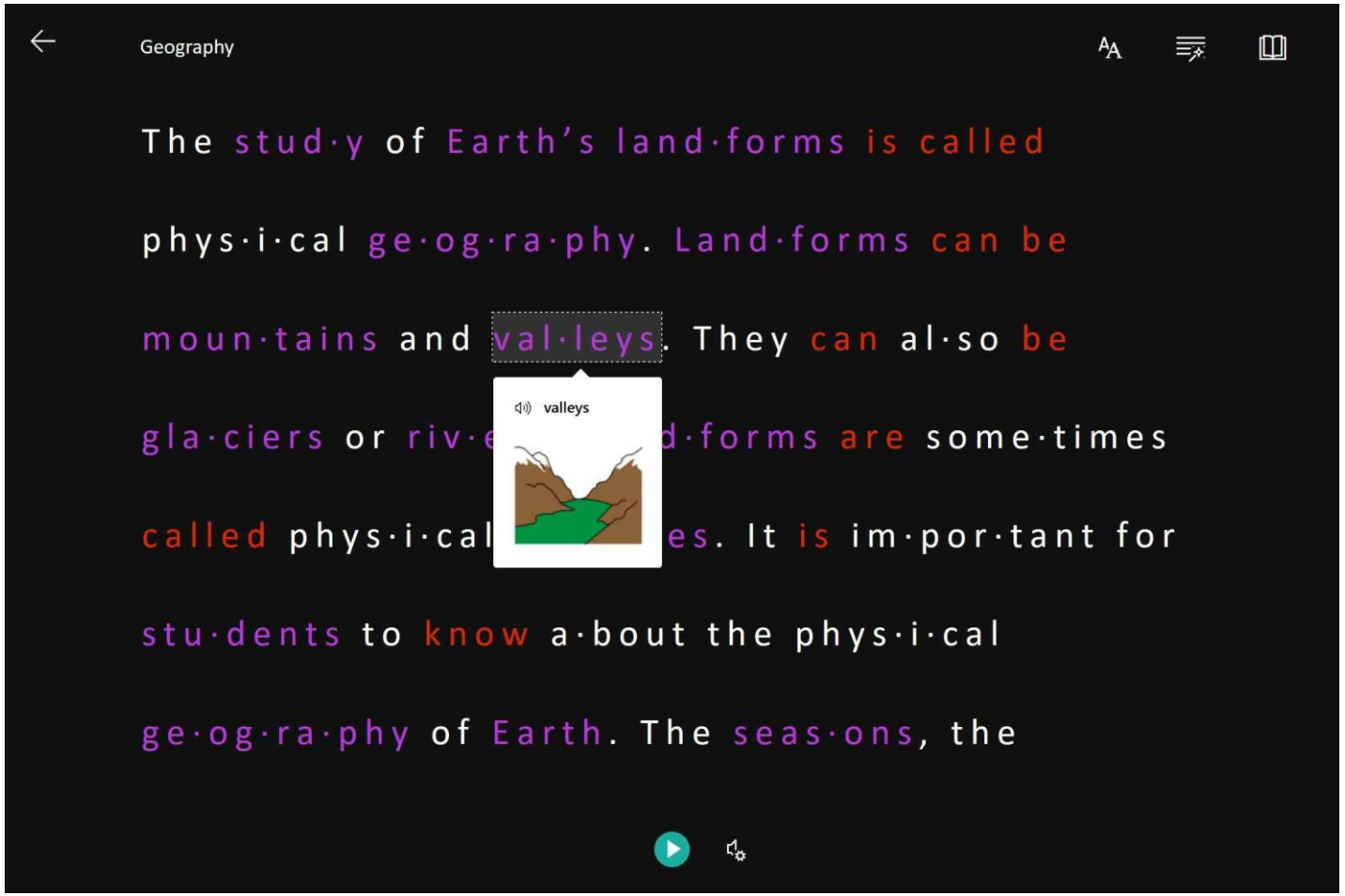
USE CASES

Content search	Product discovery optimization	Digital asset management

Azure Immersive Reader

Empower users of all ages and abilities

Get started quickly



USE CASES

Enhance reading comprehension	Multilingual learning	Accessible learning

Azure Immersive Reader

State-of-the-art Computer Vision Features

Extract robust insights from images and video content across multiple industry domains

Manage Digital Assets

Image Retrieval

Automatic, dense captions

Background Removal

Enhance Safety & Security

Video Search

Video summarization

Automate Retail Operations

Model Customization

Product Recognition

Try it Out at
https://aka.ms/VisionStudio

Azure Vision AI

Extract common tags from images

Use an AI model to automatically assign one or more labels to an image.

Try it out

Detect common objects in images

Recognize the location of objects of interest in an image and assign them a label.

Try it out

Extract text from images

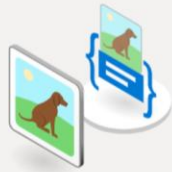
Extract printed and handwritten style text from images and documents for supported languages.

Try it out

Search photos with natural language
Preview

Retrieve specific moments within your photo album. For example, you can search for: a wedding you attended last summer, your pet, or your favorite city.

Try it out



Add captions to images
Preview

Generate a human-readable sentence that describes the content of an image.

Try it out

Dense captioning
Preview

Generate human-readable captions for all important objects detected in your image

Try it out

Customize models

Train custom image classification and object detection models using Vision Studio and Azure ML.

Start a project

Video summary and frame locator
Preview

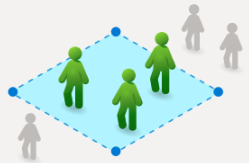
Generate a brief summary of the main points shown in video. Locate specific keywords and jump to the relevant section.

Try it out

Scan cameras for defects

Scan cameras for defects like blurry, under and over exposed cameras.

Learn more



Count people in an area

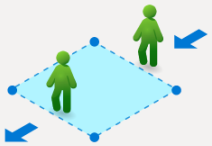
Analyze real-time video to count the number of people in a designated zone in a camera's field of view.

Learn more

Detect when people cross a line

Analyze real-time streaming video to detect when a person crosses a line in the camera's field of view.

Learn more



Detect when people enter/exit a zone

Analyze real-time streaming video to detect when a person enters or exits a zone in the camera's field of view.

Learn more

Video Analysis

Image Retrieval APIs

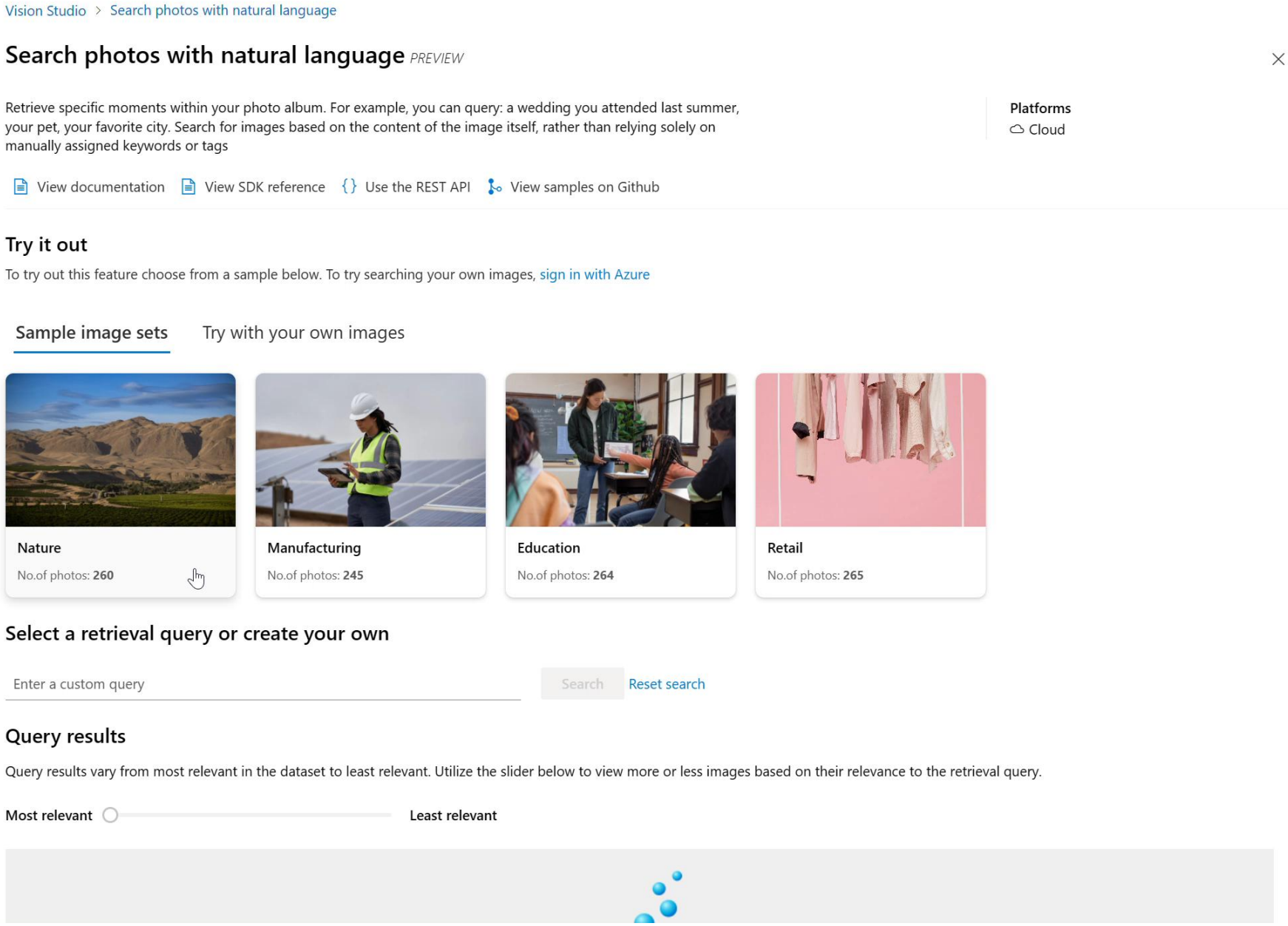
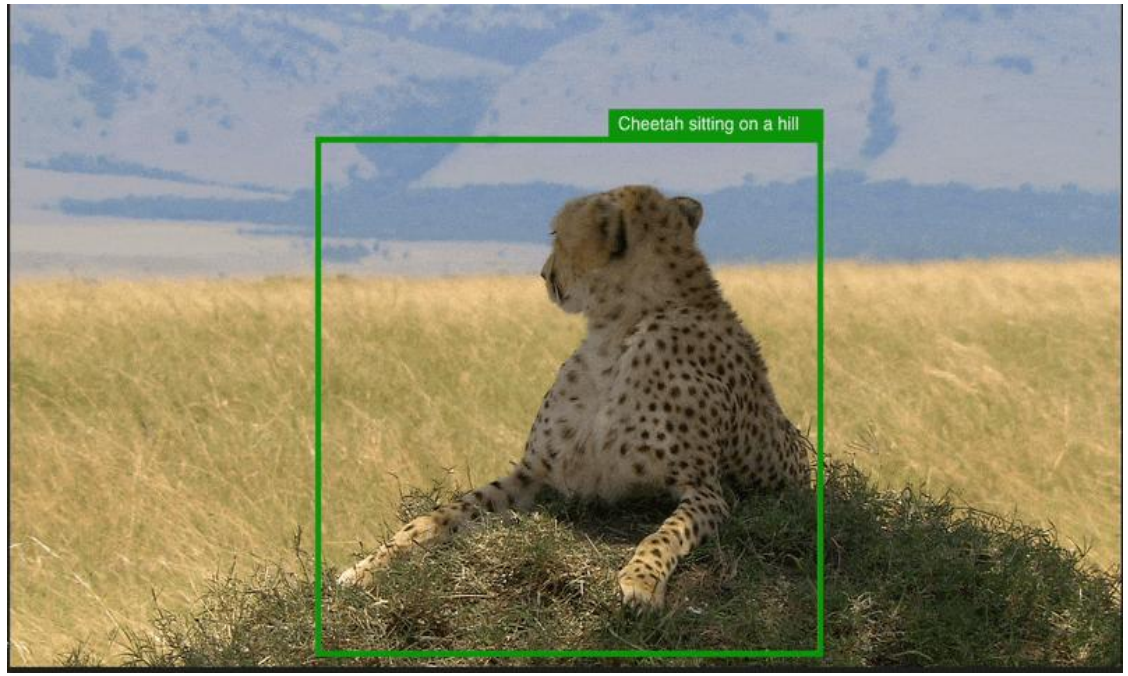


Image Analysis APIs

REST API

https://{endpoint}/computervision/imageanalysis:**analyze?features=caption**
https://{endpoint}/computervision/imageanalysis:**analyze?features=densecaptions**

Image captions

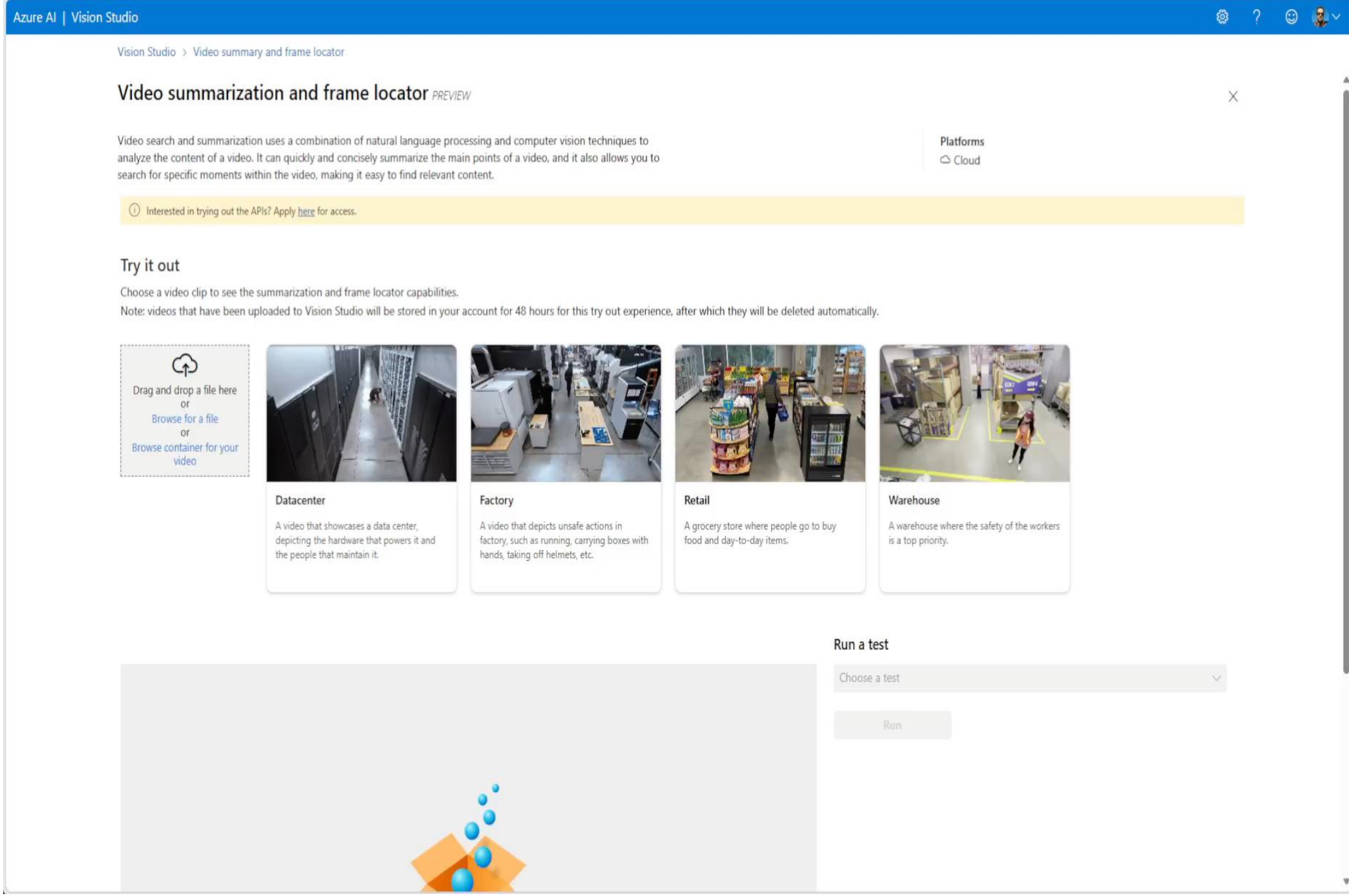


Dense captions

Segment API: Background Removal

REST API
https://{endpoint}/computervision/imageanalysis:**segment?mode=backgroundremoval**

Video Retrieval



Model customization

Fast path for customization for specific use cases

Few Shot learning and model bootstrapping with single image per label

Active Learning Loop with production data

Continuous learning for new use cases

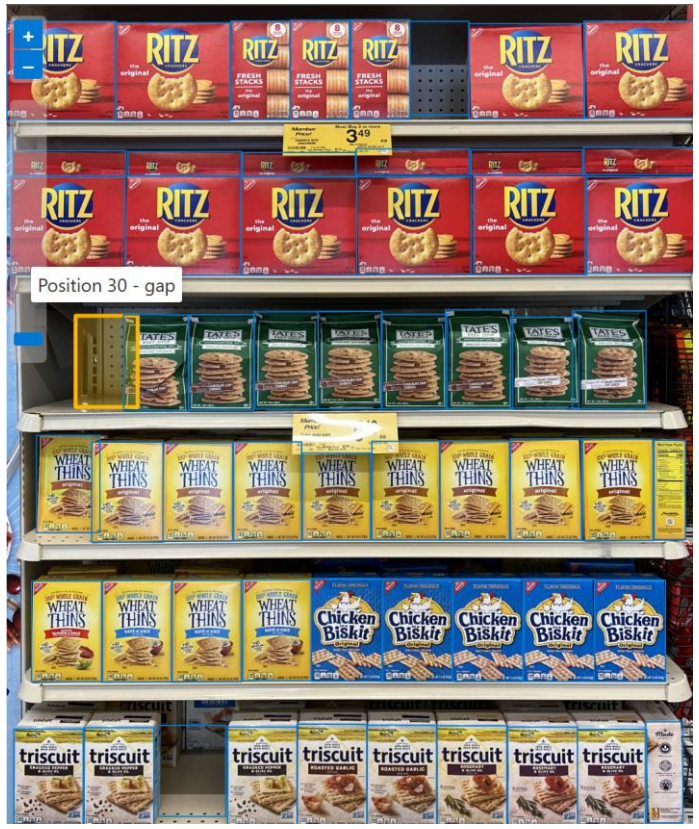
Custom Model

Product Recognition APIs

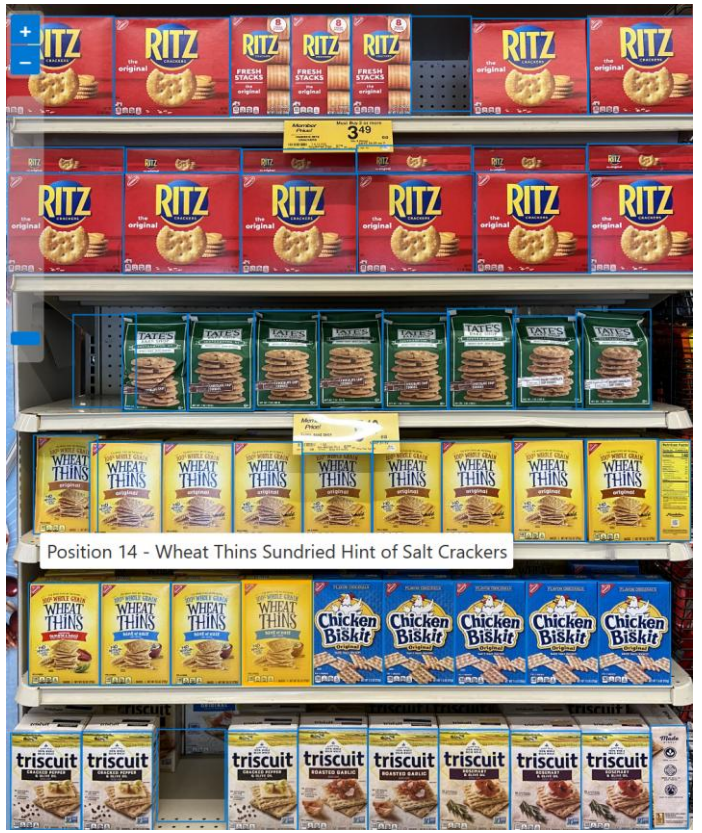
Product Recognition

https://{endpoint}/computervision/operations/**shelfanalysis-productunderstanding:analyze**

https://{endpoint}/computervision/operations/**shelfanalysis-productunderstanding:analyze**?product-classifier-model-name=**custommodel**



*



*

Demos:
No code, Try out experience available

Vision Studio – https://aka.ms/visionstudio

Azure Cognitive Services
Speech Service

Speech to Text

Language Comprehension

Speech Transcription

Versatile Speech Models

Text to Speech

Neural Voices

Custom Neural Voices

Cross lingual Adaptation

Speech Translation

Language Detection

Text-to-Text Translation

Speech-to-Text Translation

Speaker ID

Speaker Identification

Speaker Verification

Diarizations

SPEECH SDK

DEVICES SDK

AI CONTAINERS

LANGUAGES

CUSTOM MODELS

Batch versus Real Time Playbook

How does one decide on the operation mode

- Understand the latency requirements of your solution
- Understand the cost delta (real time has higher set costs)
- Understand development support requirements (You are likely to require development resources)

Batch

Faster to set up through V3 REST API

More features (diarization)

Lower response times (a 10min file is returned within 5mins + queueing time)

Real Time

Requires ingress management (SDK instances on VMs at a size relative to the audio volume)

Faster than real time for file streaming (a 10 min file is transcribed in 5mins)

Higher maintenance costs as well as up front development costs

Custom neural voices

Brand identity

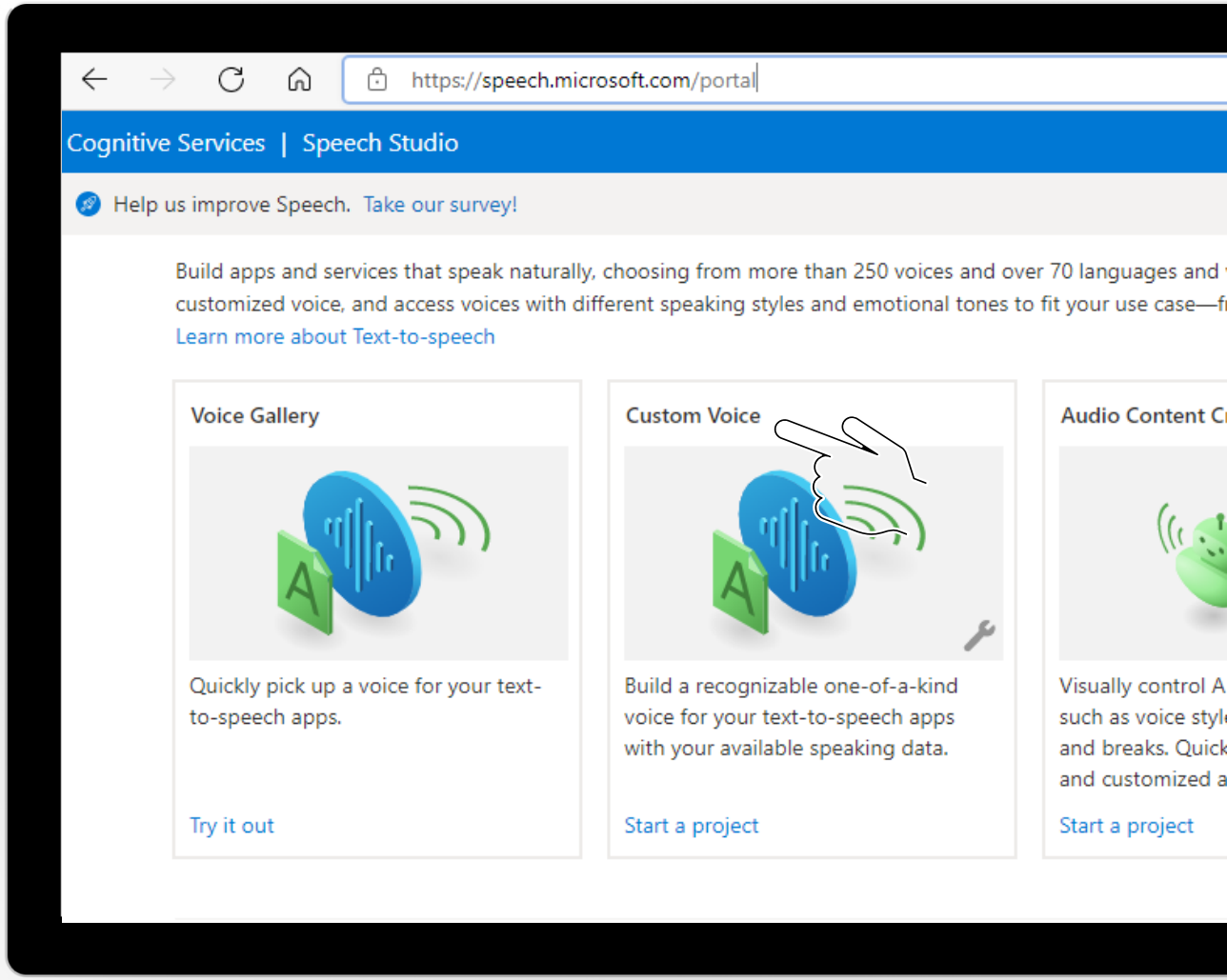
Design & implement a voice model that complements or augments your brand strategy

Custom persona

Extract value from your analytics to work by accommodating customer sentiment with custom voice characteristics

Natural interaction

Increase your customer's emotional connection and interaction with your applications



Modernize business processes

Azure
Cognitive Services

Task-specific
AI models

Bot Service

Language Understanding and Speech

Custom language models

Azure Video Analyzer

Vision and Speech

Automatic captioning models

Azure Cognitive Search

Vision - optical char recognition
Translation, and Language

Contextual search models

Form Recognizer

Vision - optical char recognition

Custom document models

Azure Metrics Advisor

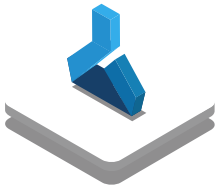
Anomaly detection

Root cause analysis model

Azure Immersive Reader

Translation and Speech

Custom readability model



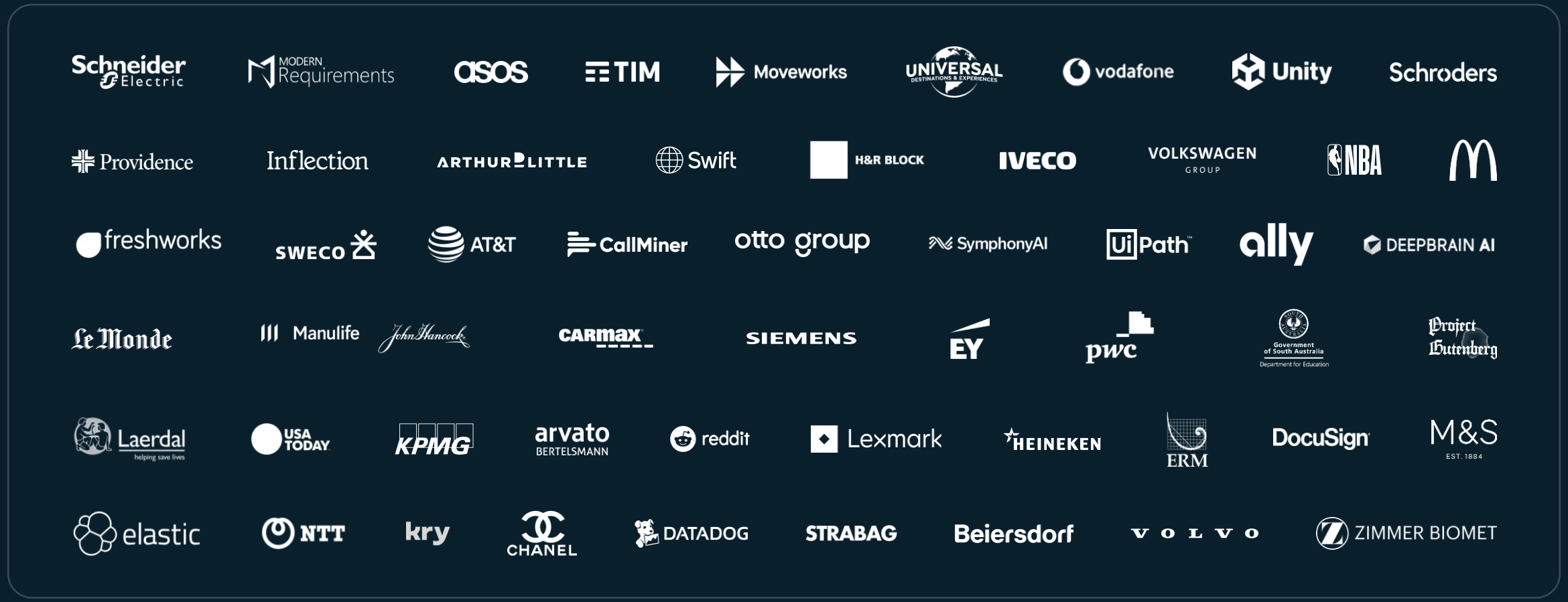
Extensible with Azure
Machine Learning

Azure OpenAI

Azure OpenAI

Azure OpenAI Service is a REST API service, which provides access to OpenAI's powerful language models including o1, o1-mini, GPT-4o, GPT-4o mini, GPT-4 Turbo with Vision, GPT-4, GPT-3.5-Turbo, and Embeddings model series. These models can be easily adapted to your specific task including but not limited to content generation, summarization, image understanding, semantic search, and natural language to code translation. Users can access the service through REST APIs, Python SDK, or in the Azure AI Foundry. (link)

60,000+ organizations using Azure AI today



Azure OpenAI Service

99.9%

Reliability
service wide

28

Regions
for data residency

99%

Latency SLA
for Provisioned

EU & US

Data Zones
for regional control

Using Azure OpenAI Service

When you use Azure OpenAI Service, your prompts (inputs) and completions (outputs), your embeddings, and your training data

Are **NOT** available to other customers.

ARE **NOT** available to OpenAI.

Are **NOT** used to improve OpenAI models.

Are **NOT** used to improve any Microsoft or 3rd party products or services.

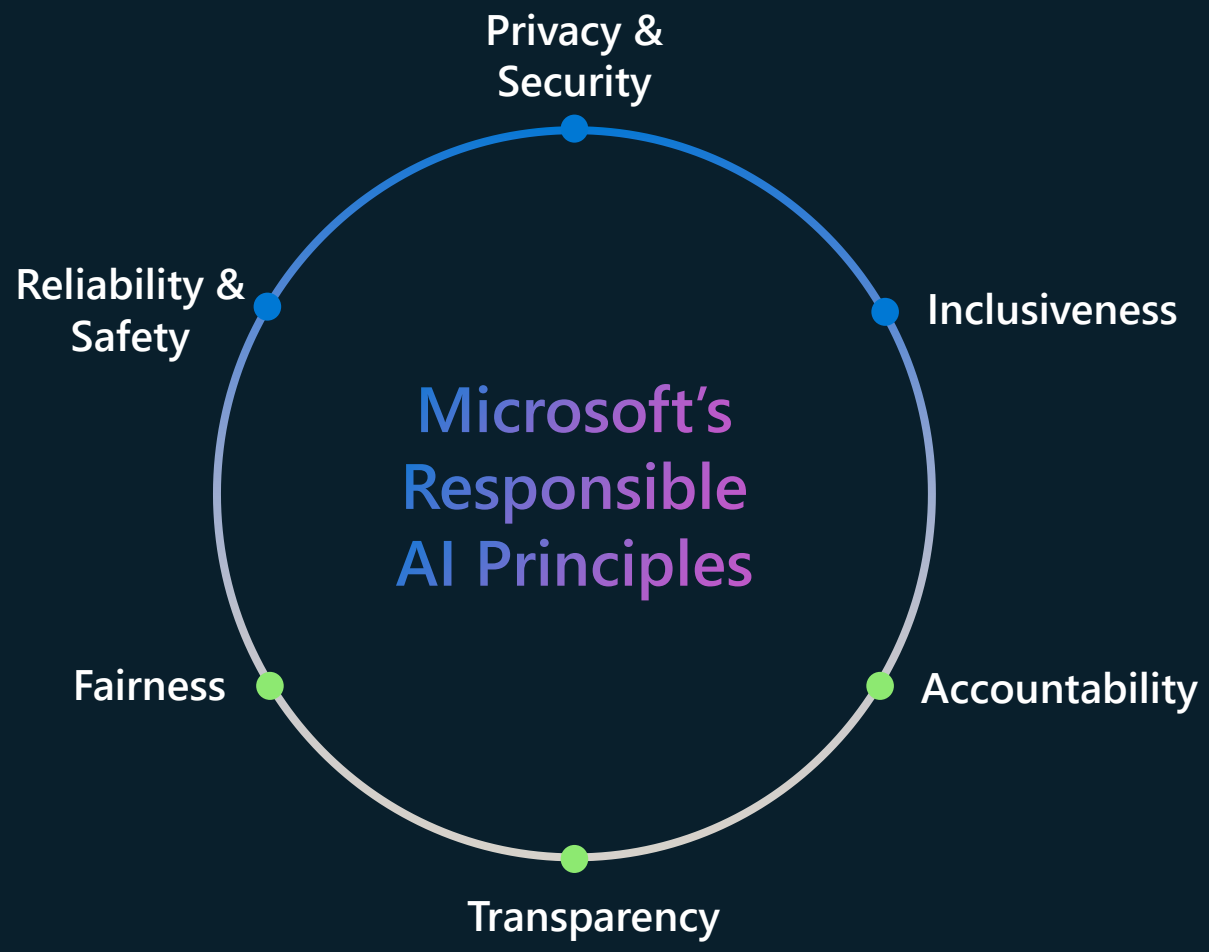
Are **NOT** used for automatically improving Azure OpenAI models for your use in your resource (The models are stateless, unless you explicitly fine-tune models with your training data).

Your fine-tuned Azure OpenAI models are available **exclusively** for your use.

The Azure OpenAI Service is fully controlled by Microsoft.

Microsoft hosts the OpenAI models in Microsoft's Azure environment and the Service does **NOT** interact with any services operated by OpenAI (e.g., ChatGPT, or the OpenAI API).

Responsible AI protection turned on by default

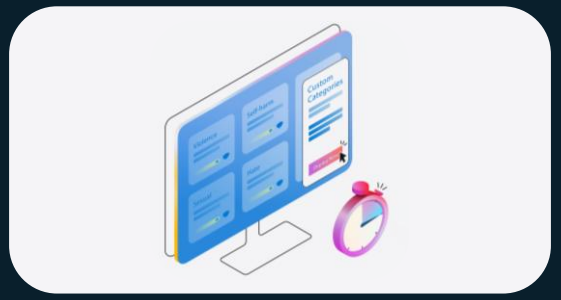


Building blocks to enact principles

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

Azure AI Content Safety

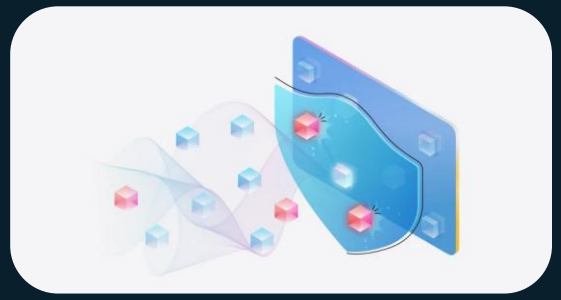
Safety is built-in with content filters for Aure OpenAI Service



Harms, Fairness & Bias

Content filters for inputs + outputs:

- Violence
- Hate and Fairness
- Sexual
- Self-Harm
- Custom categories PREVIEW
- Custom blocklists



Security

Prompt shields for inputs:

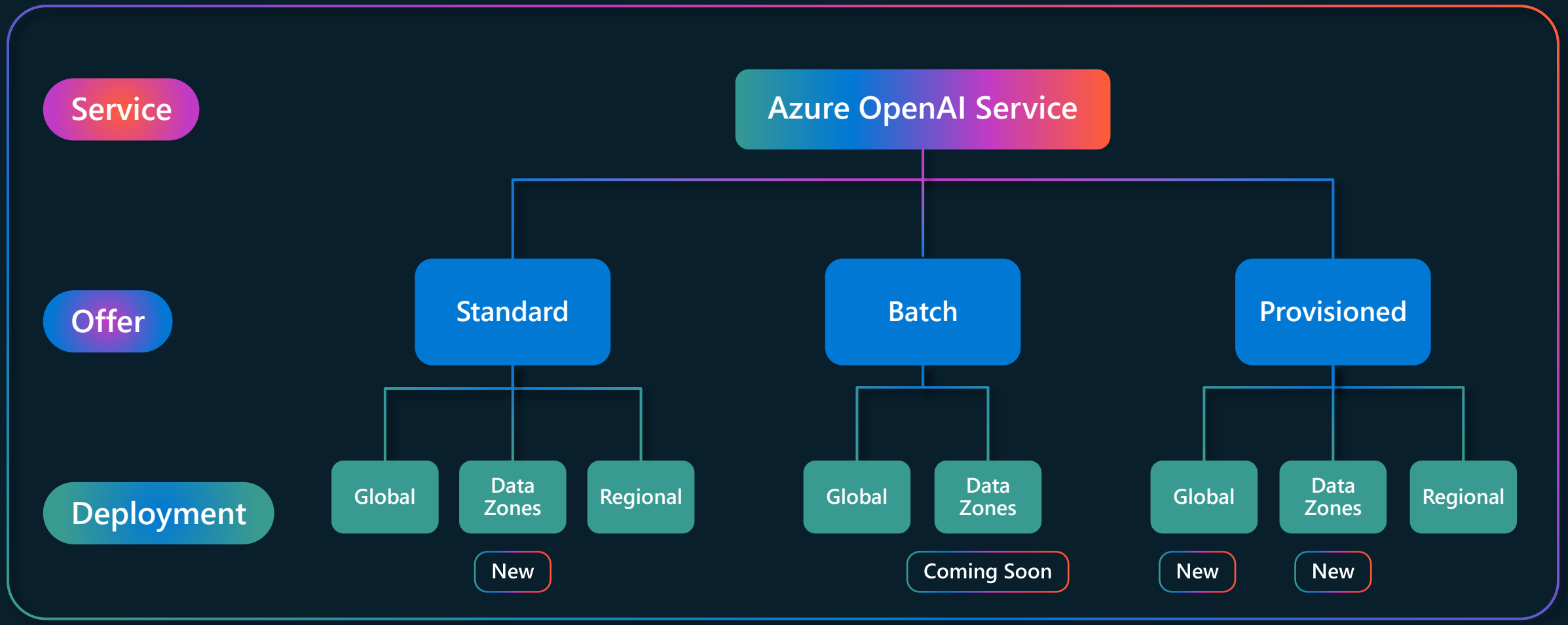
- Direct prompt injection attacks
- Indirect prompt injection attacks

Quality & Transparency

Content filters for inputs + outputs:

- Protected text
- Protected code PREVIEW

Azure OpenAI Service Offerings



Choosing the best offer

Standard

Benefits

- Flexible to scale up and down upon demand
- Optimized for low to medium volume workloads
- Ease of management

RECOMMENDED FOR

- ✓ Production workloads
- ✓ Development and testing
- ✓ Prototyping and proof of concept

Batch

Benefits

- Efficient Processing of large volumes of data
- Scalable to handle varying workloads
- Cost-effective for large-scale deployments

RECOMMENDED FOR

- ✓ Large-scale data processing
- ✓ Generate large volumes of content Transforming data at scale
- ✓ Evaluate LLM models and assess

Provisioned

Benefits

- Easy access with high and predictable throughput
- Real-time scoring for large consistent volume
- Cost effective for large-scale deployments

RECOMMENDED FOR

- ✓ Production workloads
- ✓ High volume data processing
- ✓ Throughput heavy workloads
- ✓ Real-time

Choosing the best deployment

Global

Overview

- Lowest price with highest throughput
- Broadest model availability
- Broadest capacity availability

BEST FOR

- ✓ Applications requiring a consistent experience across multiple regions
- ✓ Services needing to be available globally with low latency
- ✓ Cost savings is a priority
- ✓ Data residency independent of deployment

Data zone

Overview

- Cross region load balancing within a geographic boundary (US or EU)
- Broader model availability
- Broader capacity availability

BEST FOR

- ✓ Applications needing more processing capacity with data residency required
- ✓ Cost saving with meeting compliance requirements
- ✓ Optimal access to latest AI models and innovations

Regional

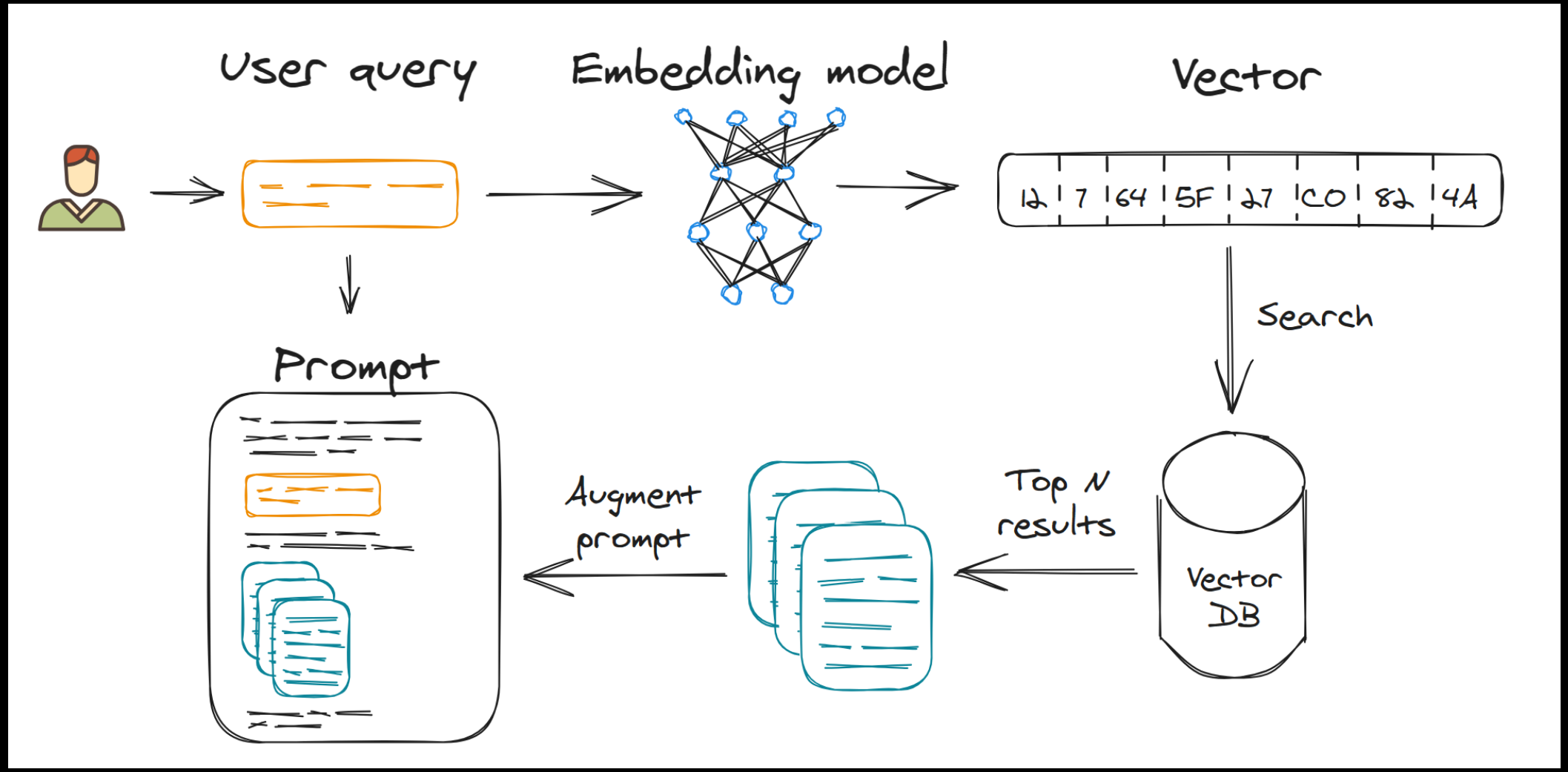
Overview

- Strictest levels of data control
- Limited model availability
- Limited capacity availability

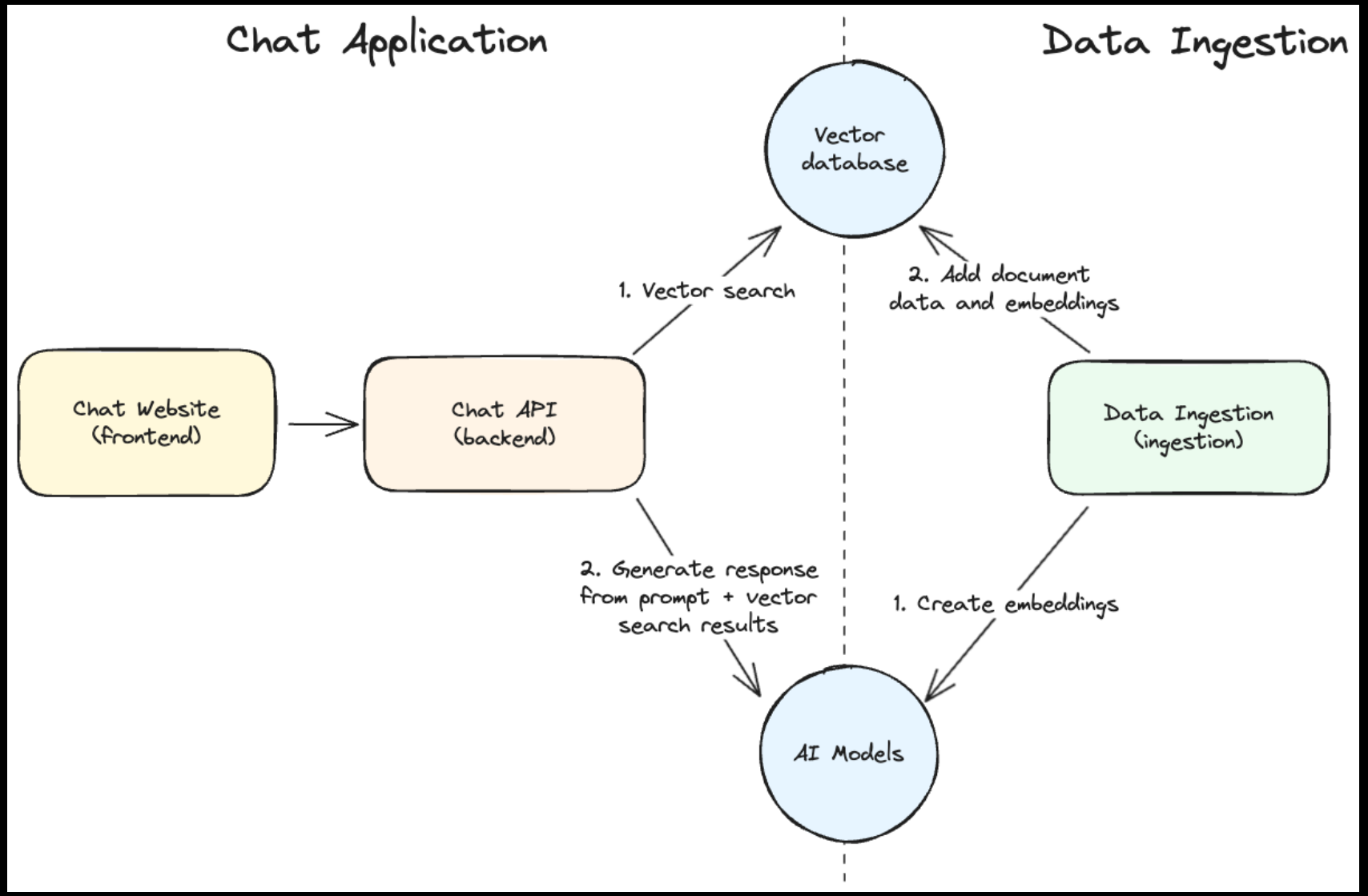
BEST FOR

- ✓ Applications required to meet data residency compliance globally with low latency
- ✓ Services needing to be closer to the end-users to reduce latency
- ✓ Applications requiring localized data processing and storage

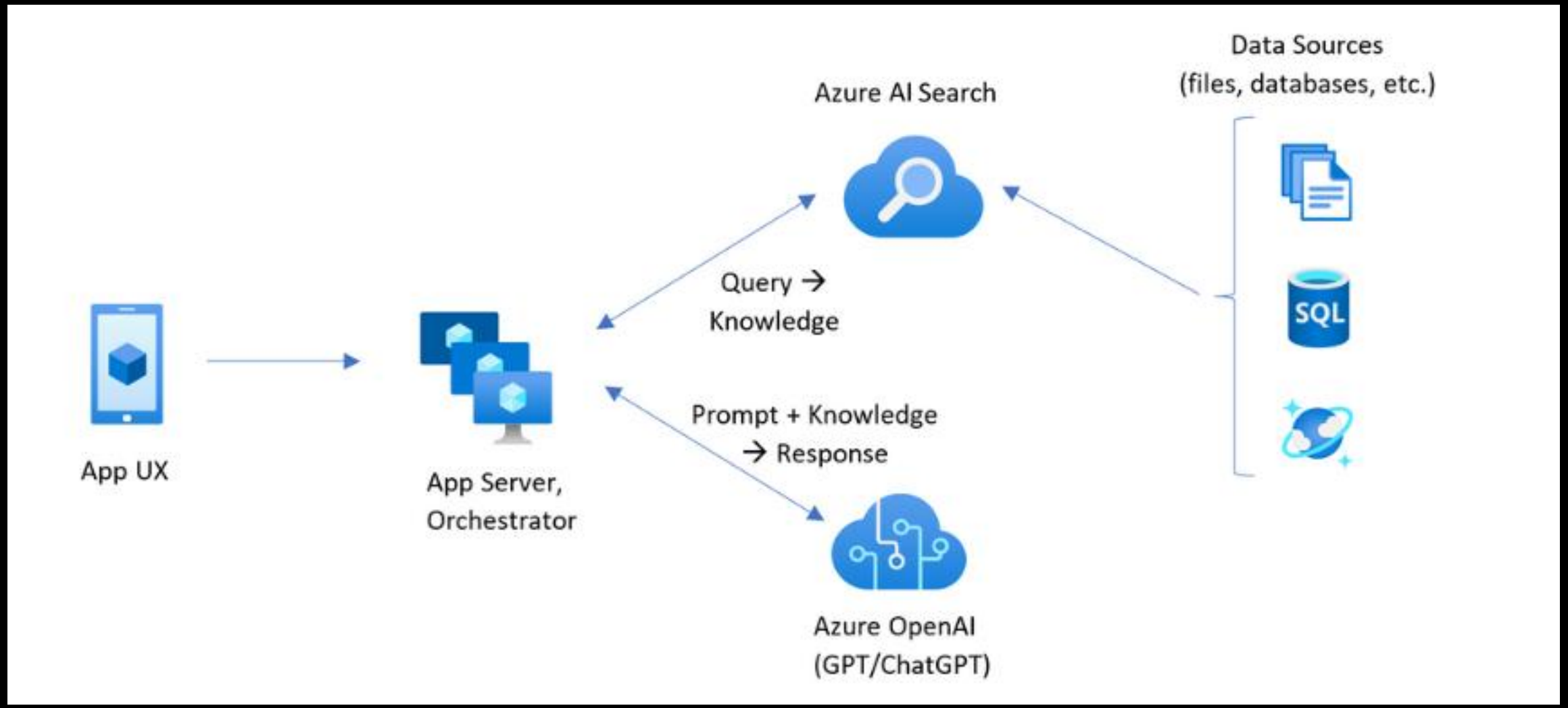
RAG Application: user query processing



RAG Application: architecture



RAG Application on Azure: architecture (link)



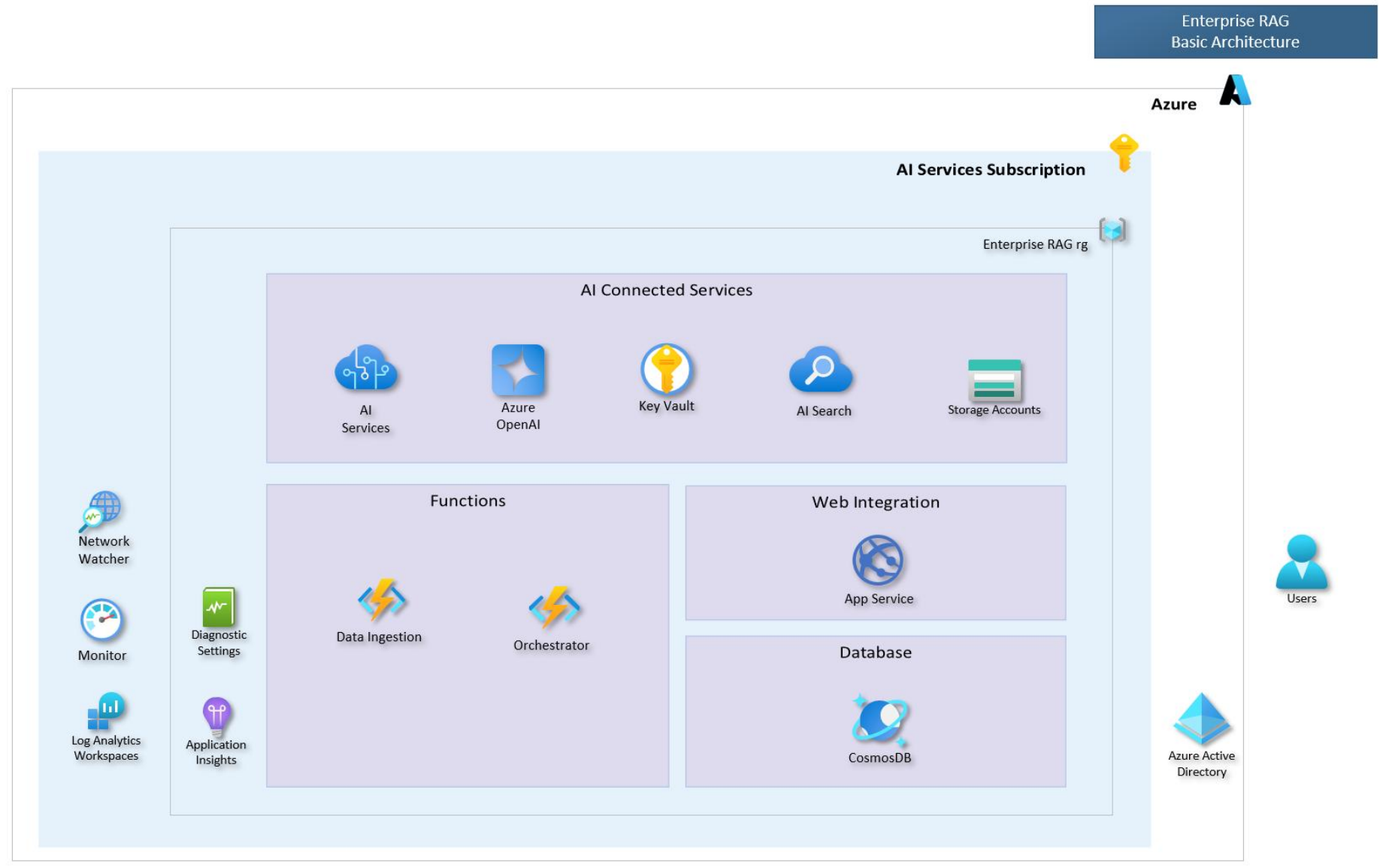
Microsoft runs on Azure AI Foundry

Microsoft Copilot	Microsoft 365 Copilot	Microsoft Security Copilot	GitHub Copilot	DAX Copilot

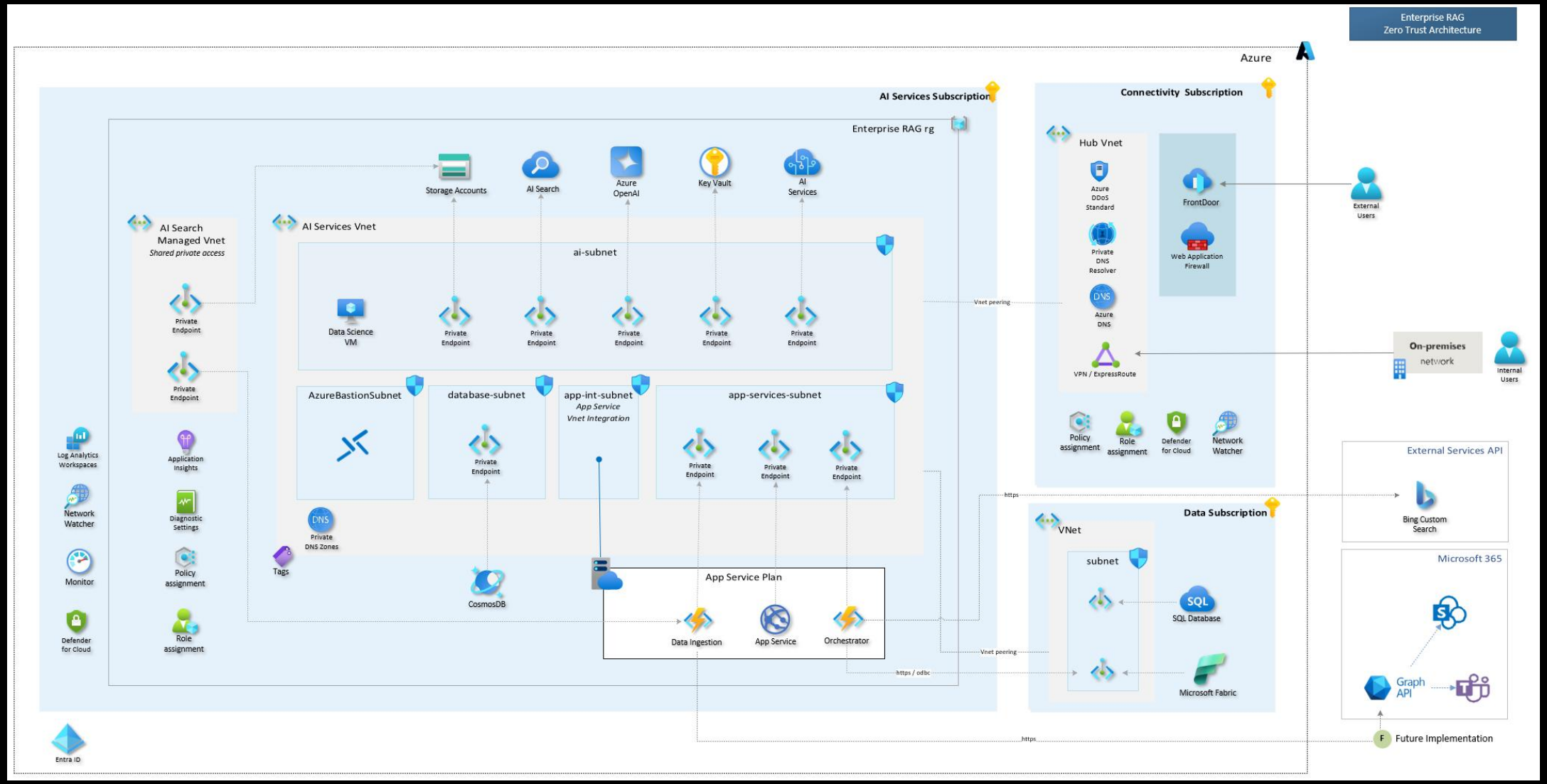
Demo

Azure Open AI | AI Foundry | AI Search | RAG

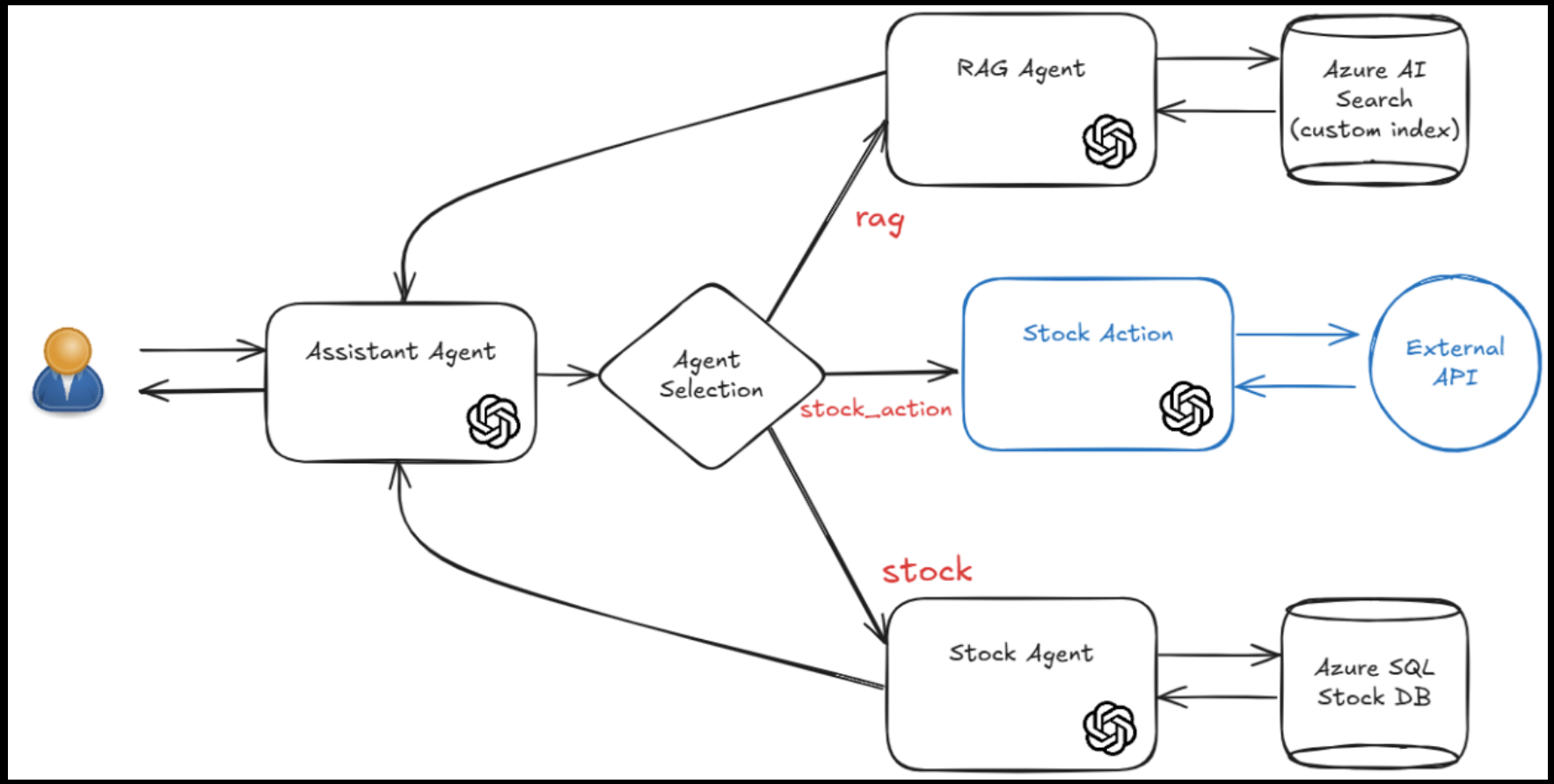
RAG Application on Azure: more realistic architecture



RAG Application on Azure: enterprise architecture



Agentic workloads



Framework examples:

- Microsoft Autogen
- Phidata
- OpenAI Swarm
- CrewAI
- LangGraph
- ....

phidata example

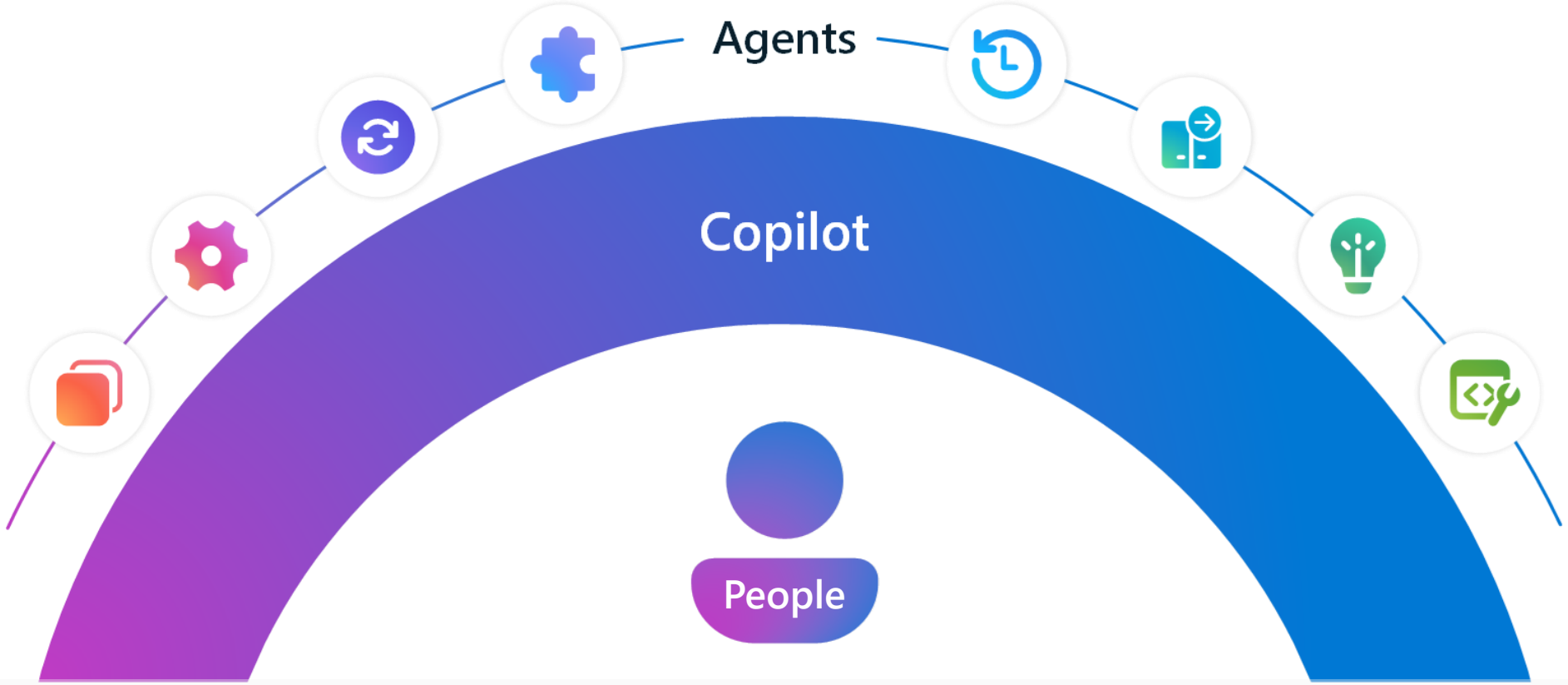
- GH Link

```python
from phi.agent import Agent
from phi.model.openai import OpenAIChat
from phi.tools.duckduckgo import DuckDuckGo
from phi.tools.yfinance import YFinanceTools

web_agent = Agent(
    name="Web Agent",
    role="Search the web for information",
    model=OpenAIChat(id="gpt-4o"),
    tools=[DuckDuckGo()],
    instructions=["Always include sources"],
    show_tool_calls=True,
    markdown=True,
)

finance_agent = Agent(
    name="Finance Agent",
    role="Get financial data",
    model=OpenAIChat(id="gpt-4o"),
    tools=[YFinanceTools(stock_price=True, analyst_recommendations=True, company_info=True)],
    instructions=["Use tables to display data"],
    show_tool_calls=True,
    markdown=True,
)

agent_team = Agent(
    team=[web_agent, finance_agent],
    model=OpenAIChat(id="gpt-4o"),
    instructions=["Always include sources", "Use tables to display data"],
    show_tool_calls=True,
    markdown=True,
)

agent_team.print_response("Summarize analyst recommendations and share the latest news for NVDA", strea
```

Copilot

Business Process

Transform the way you work with Copilot + Agents

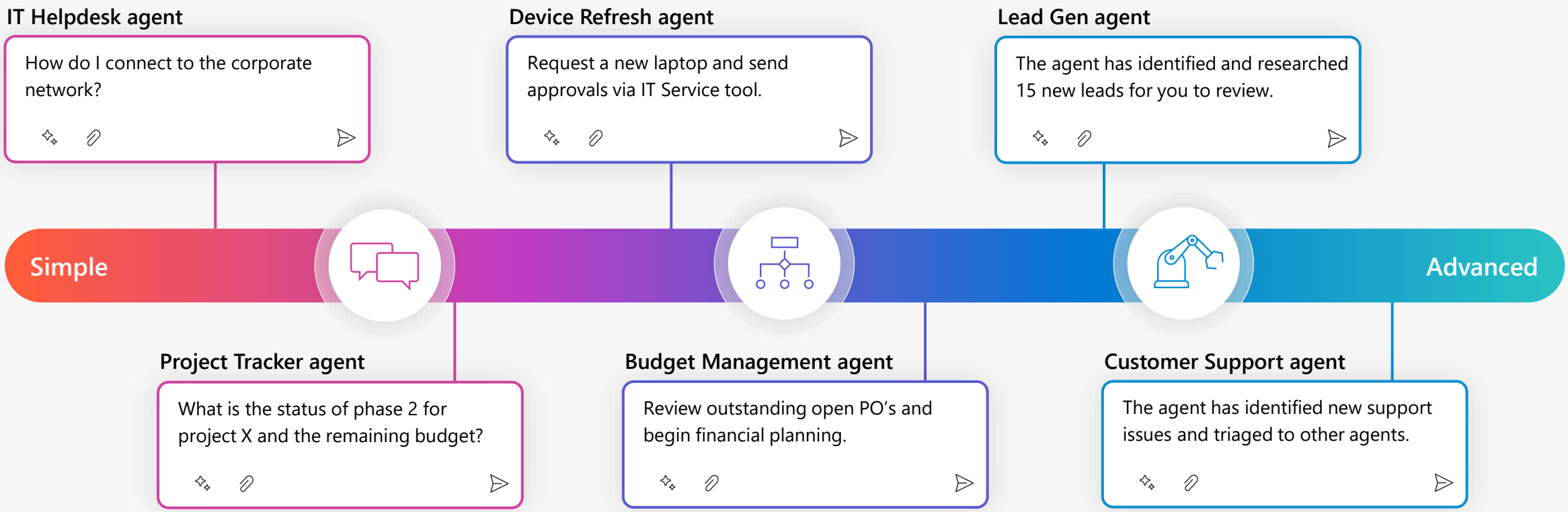


Copilot makes agents more personal to you

Agents make Copilot more capable

Together, they benefit from consistent security, governance & management

Explore a continuum of solutions



Links

- https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/provisioned-throughput-onboar
- https://learn.microsoft.com/en-us/legal/cognitive-services/content-safety/data-privacy
- https://learn.microsoft.com/en-us/azure/ai-studio/how-to/concept-data-privacy
- https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/abuse-monitoring

Microsoft

Thank you