Microsoft

Microsoft Ignite

Microsoft

An IT Pro's guide to Deploying and managing AI applications

Rick Claus
Vinicius Apolinario
Pierre Roman
Steven Murawski
Orin Thomas
Rob Hindman
Shriram Natarajan
Bilal Amjad

We'll start at 1:10pm

Microsoft

Thank you for your feedback!

On the Ignite page:
Please give us feedback

https://aka.ms/ignite25-feedback

An IT Pro's guide to Deploying and managing AI applications

Your feedback is very important to us!

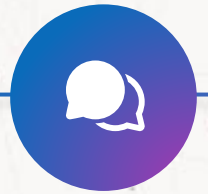
Agenda

- Setting the Stage: AI in the Real World
- Architecting for Secure AI: Identity, Access, Secrets, Keys, and Data Protection
- Networking for AI Workloads
- Monitoring and governing AI Apps
- AI Landing Zones

Setting the Stage: AI in the Real World

Why are we infusing AI into our apps?

Enrich employee experiences



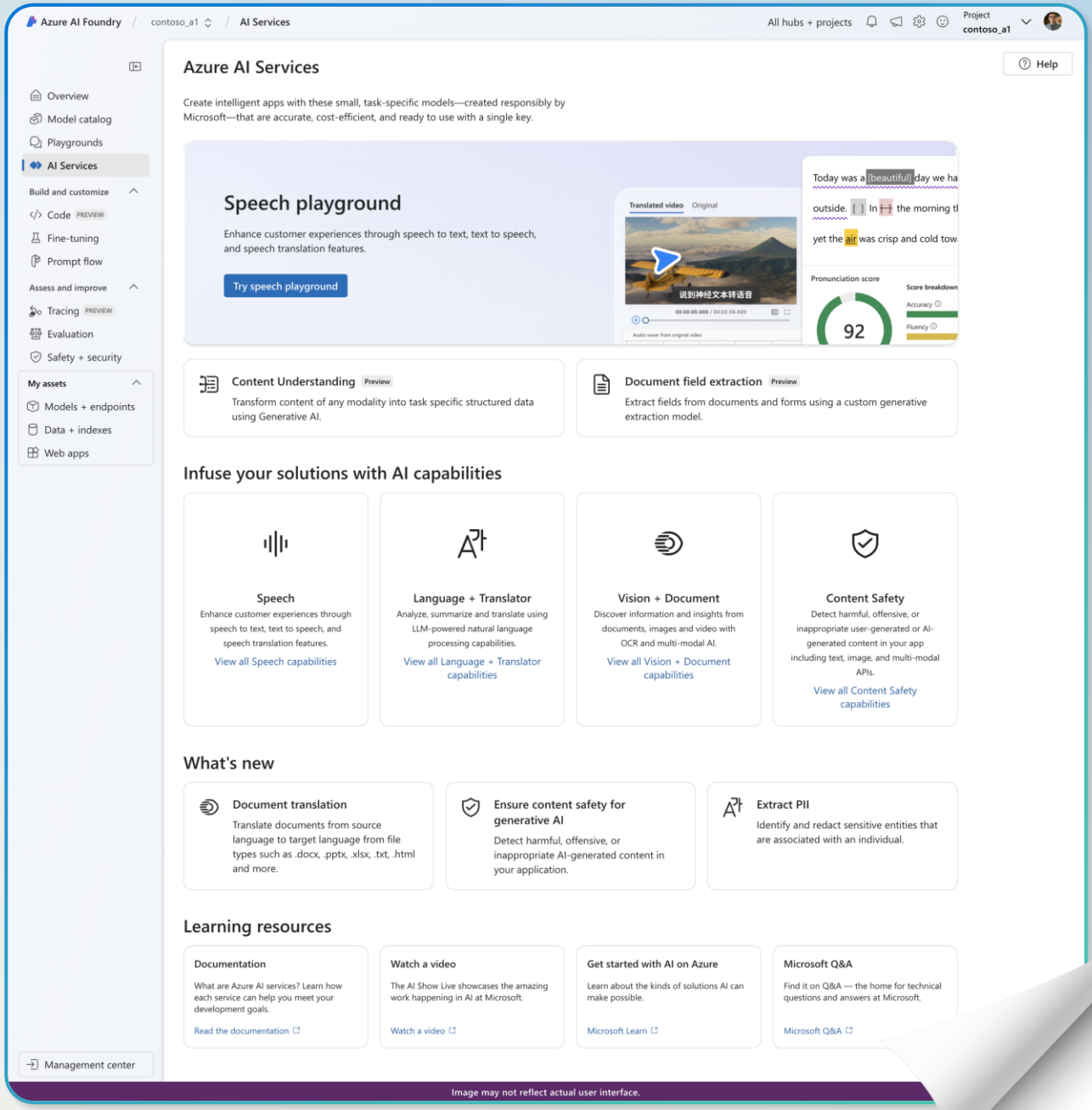
Reinvent customer engagement

Reshape business processes

Bend the curve on innovation

Azure AI services

Azure provides multiple AI services for different scenarios. As an IT Pro/Ops, you should understand the infrastructure related tasks to support these services.



Azure AI services

Azure OpenAI Service

- Access to powerful AI models
- Scalable development
- Compliance & security
- Integration with other Azure Services

Azure AI Search

- AI enrichment & semantic ranking
- Generative AI content creation
- Vector search for data organization

Azure AI Speech

- Speech to text (including the Whisper model on Azure OpenAI Service)
- Text to speech
- Speech translation
- Speaker recognition

Azure AI Vision

- Image and face analysis
- Custom model training
- Face detection and recognition
- Document text extraction

Azure AI Content Safety

- AI-driven content moderation for enhanced safety
- Customize safety thresholds for diverse user types
- Detect and prevent Jailbreak Risk from XPIA attacks

Azure AI Document Intelligence

- Automated documentation generation
- Documentation quality analysis
- Interactive documentation experiences
- Natural language understanding for documentation

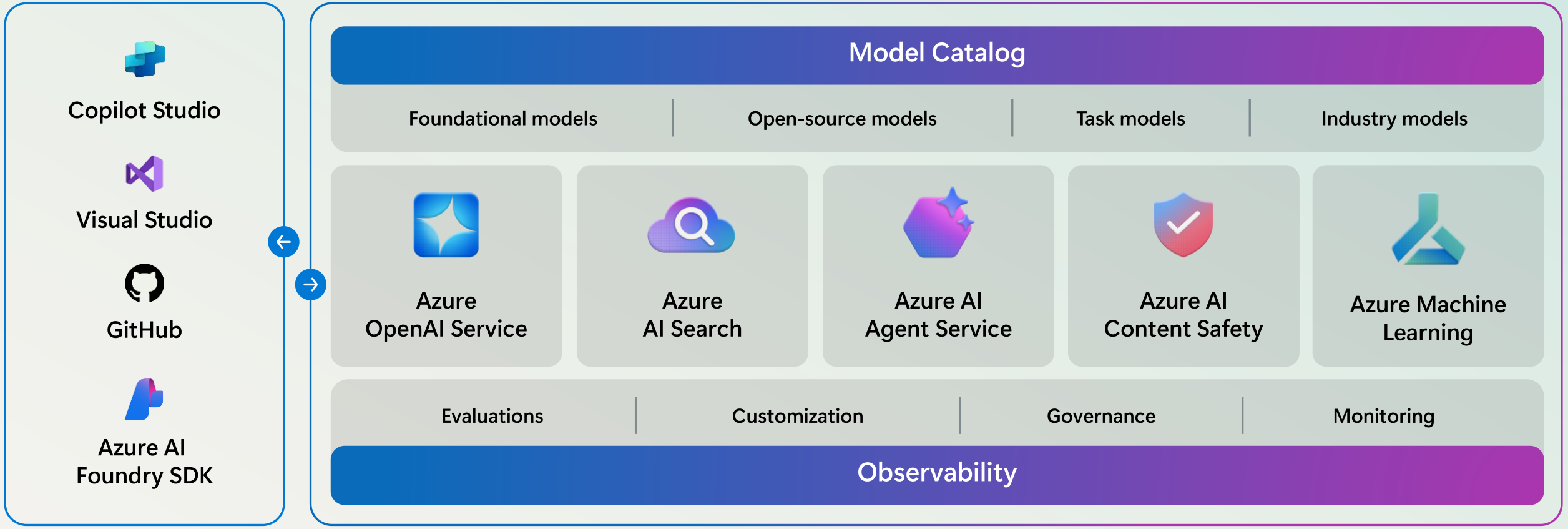
Azure AI Language

- Task-optimized AI models for text analytics
- Custom industry-specific AI for healthcare
- Custom, industry-specific models

Azure AI Translator

- Multilingual text and speech translation
- Synchronous and asynchronous translation request support
- Native translation of documents and manuals

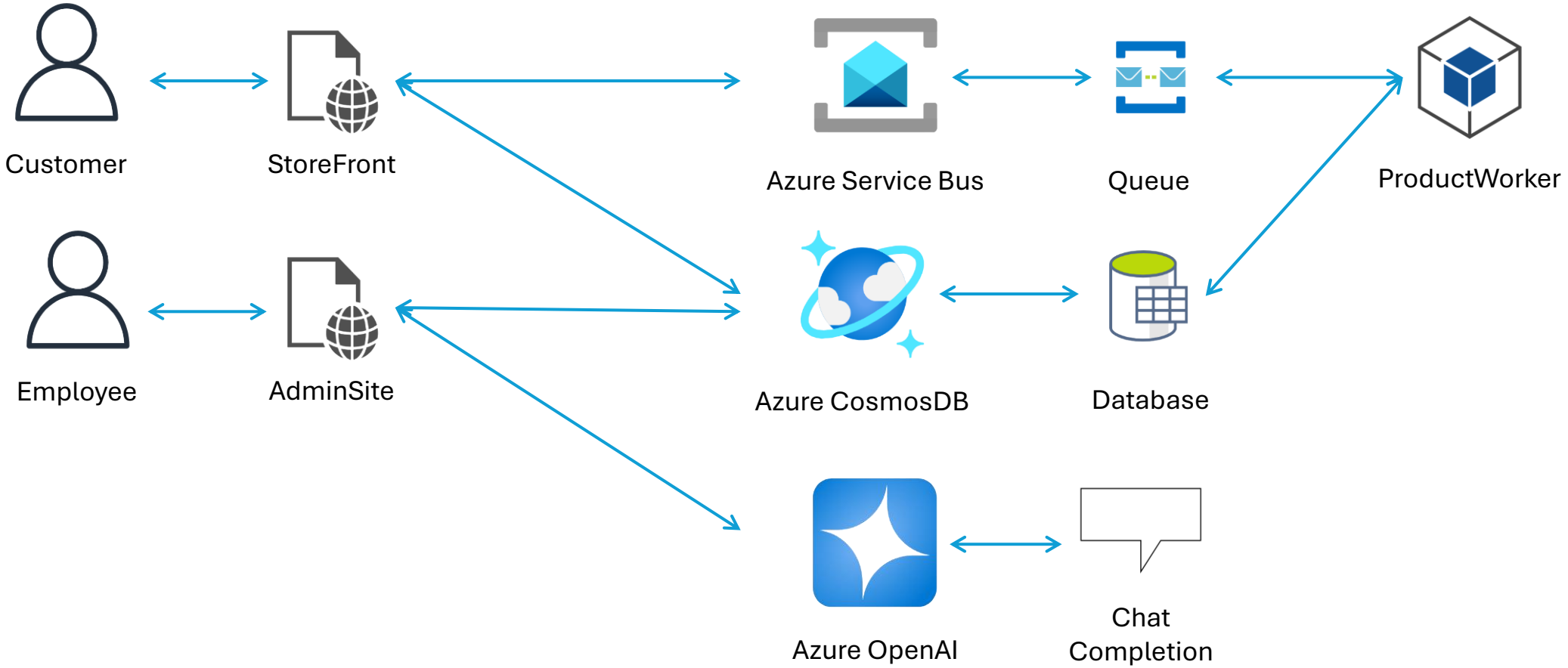
Azure AI Foundry



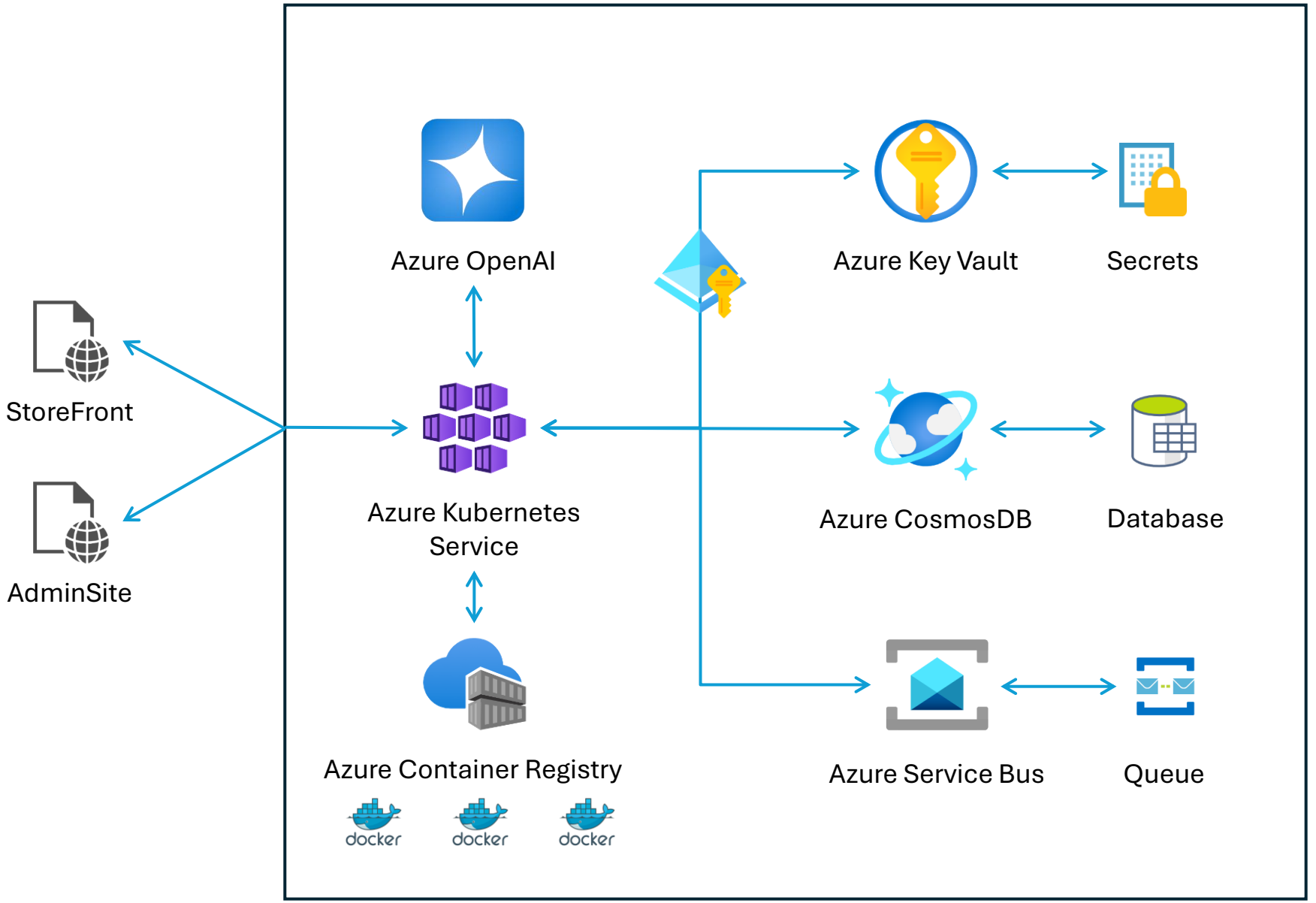
Microsoft

A closer look at an AI-enabled app

Sample App

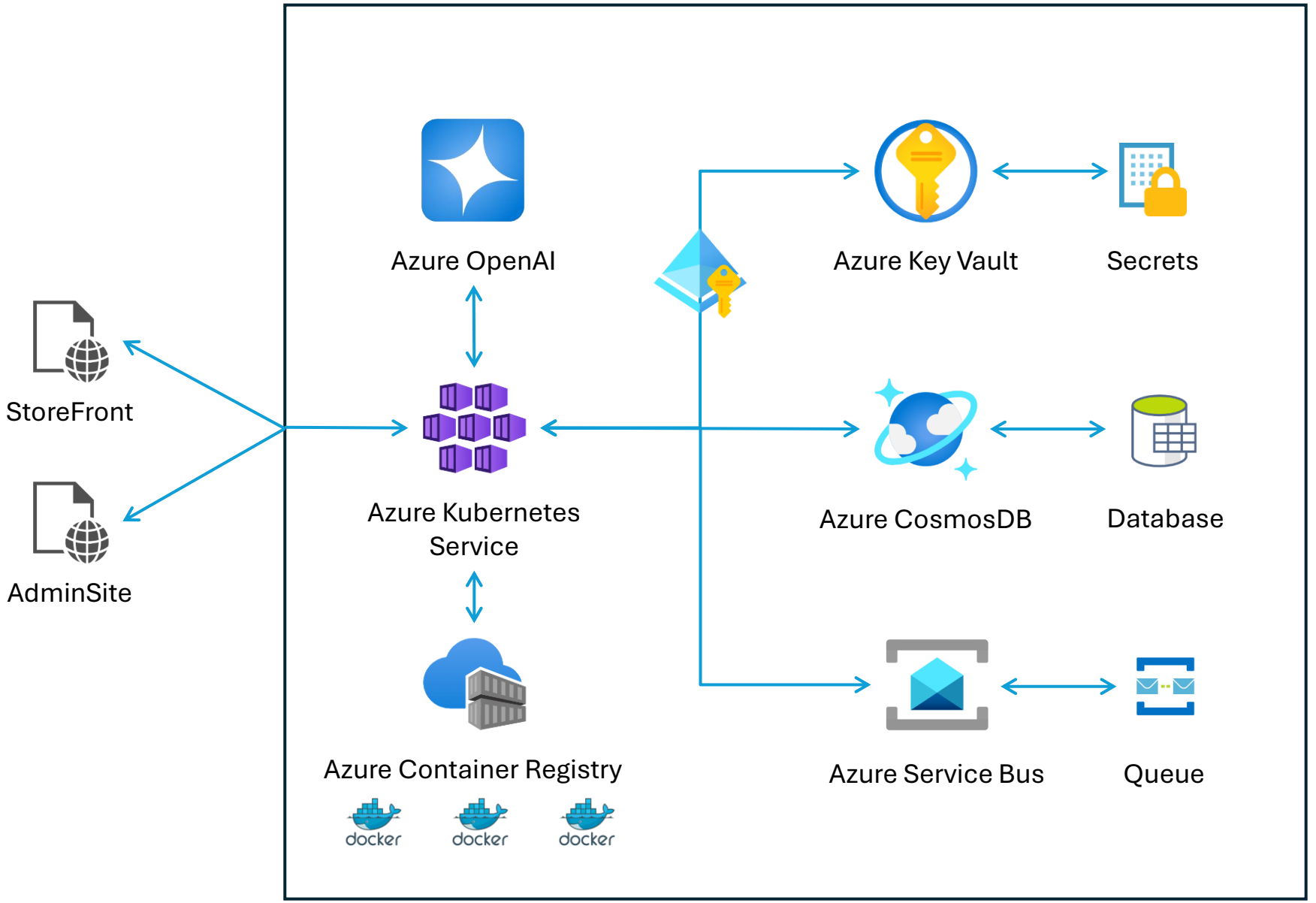

Customer
StoreFront
Azure Service Bus
Queue
ProductWorker
Employee
AdminSite
Azure CosmosDB
Database
Azure OpenAI
Chat
Completion

Sample App service architecture

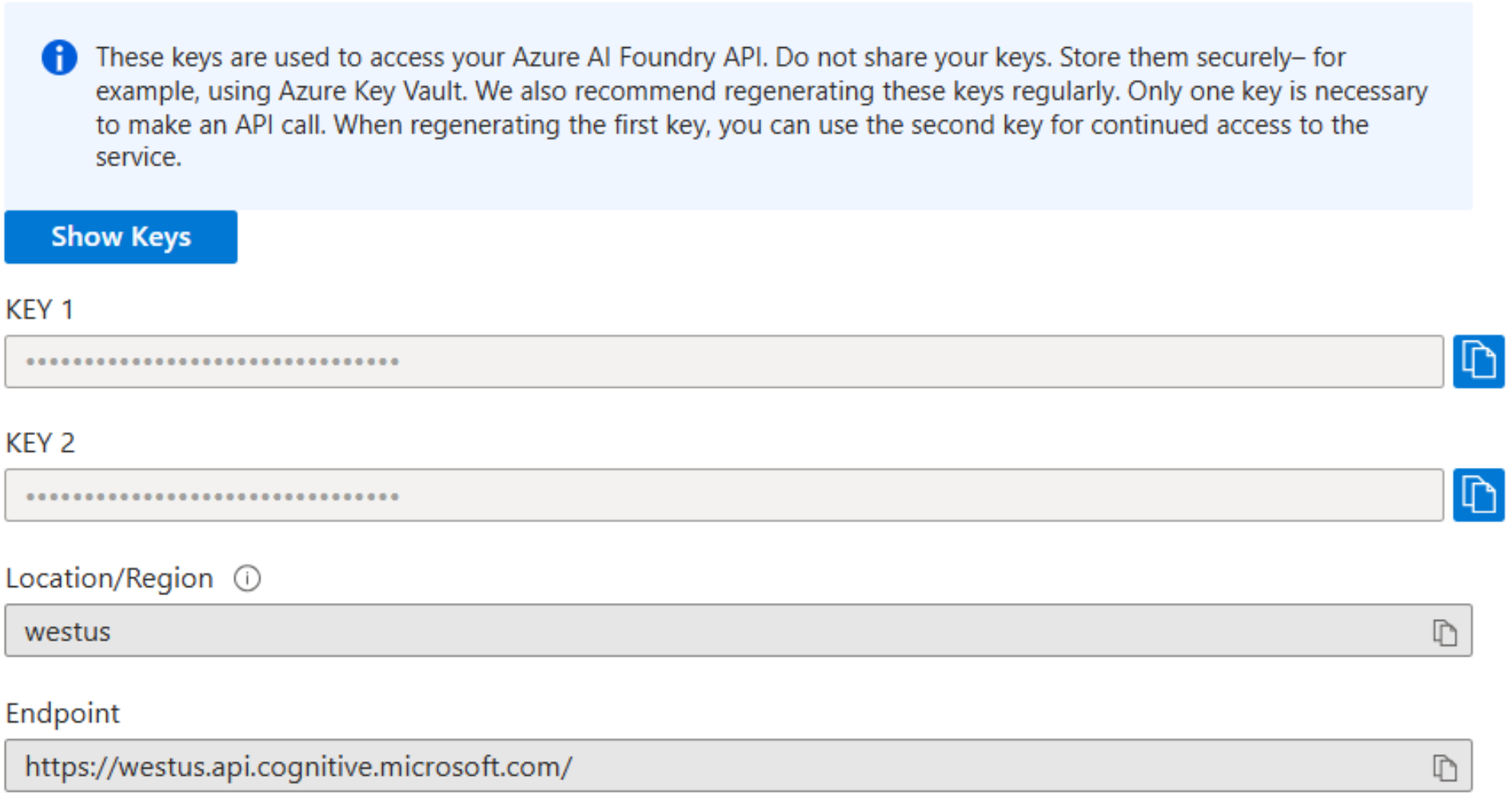


Architecting for Secure AI: Identity, Access, Secrets, Keys, and Data Protection

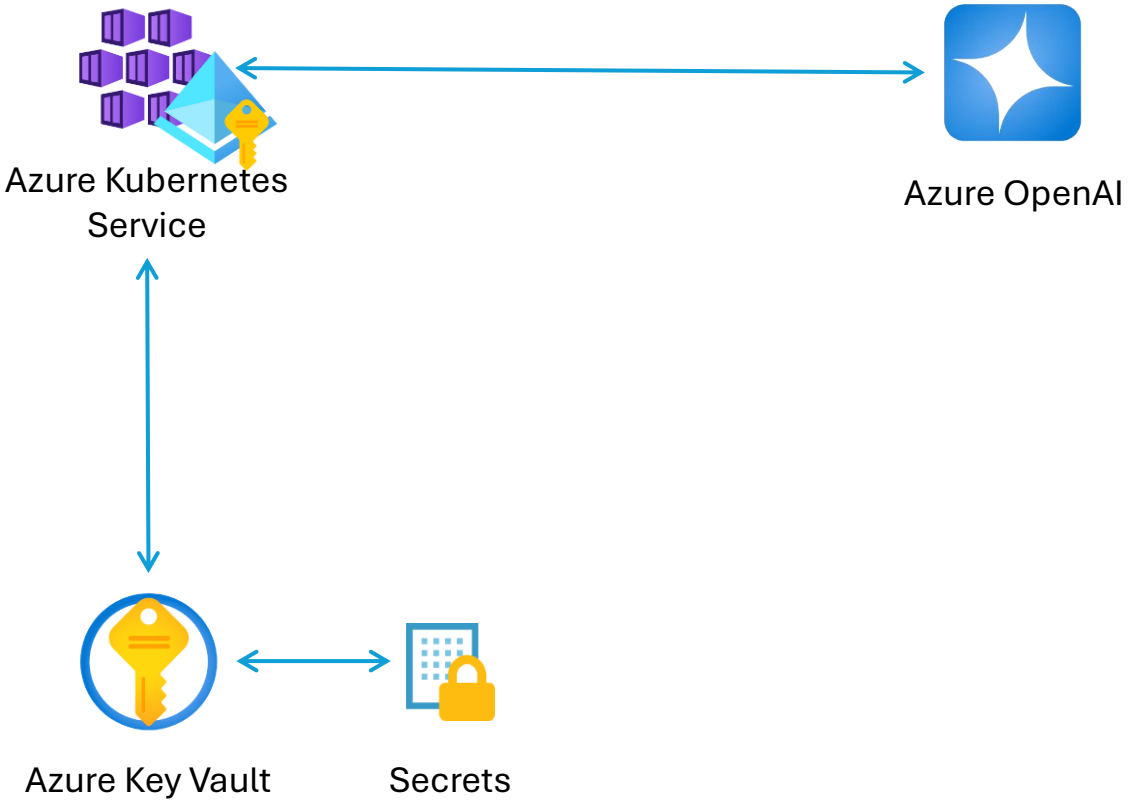
AI service HTTP endpoint



AI services HTTP endpoint



AI service HTTP endpoint



1. Azure OpenAI services have HTTP endpoint
2. App on AKS cluster needs that info to access AI endpoint
3. Instead of hardcoding AI endpoint in app, info is stored in Azure Key Vault
4. A Managed Identity is assigned to the AKS cluster and nodes
5. When AKS cluster access Azure Key Vault, the MI is used as credential
6. AKV checks if MI credential can retrieve the secrets (RBAC)
7. App on AKS is loaded with HTTP endpoint at start/runtime
8. When user requests AI endpoint interaction, the app in AKS can access the AI service endpoint

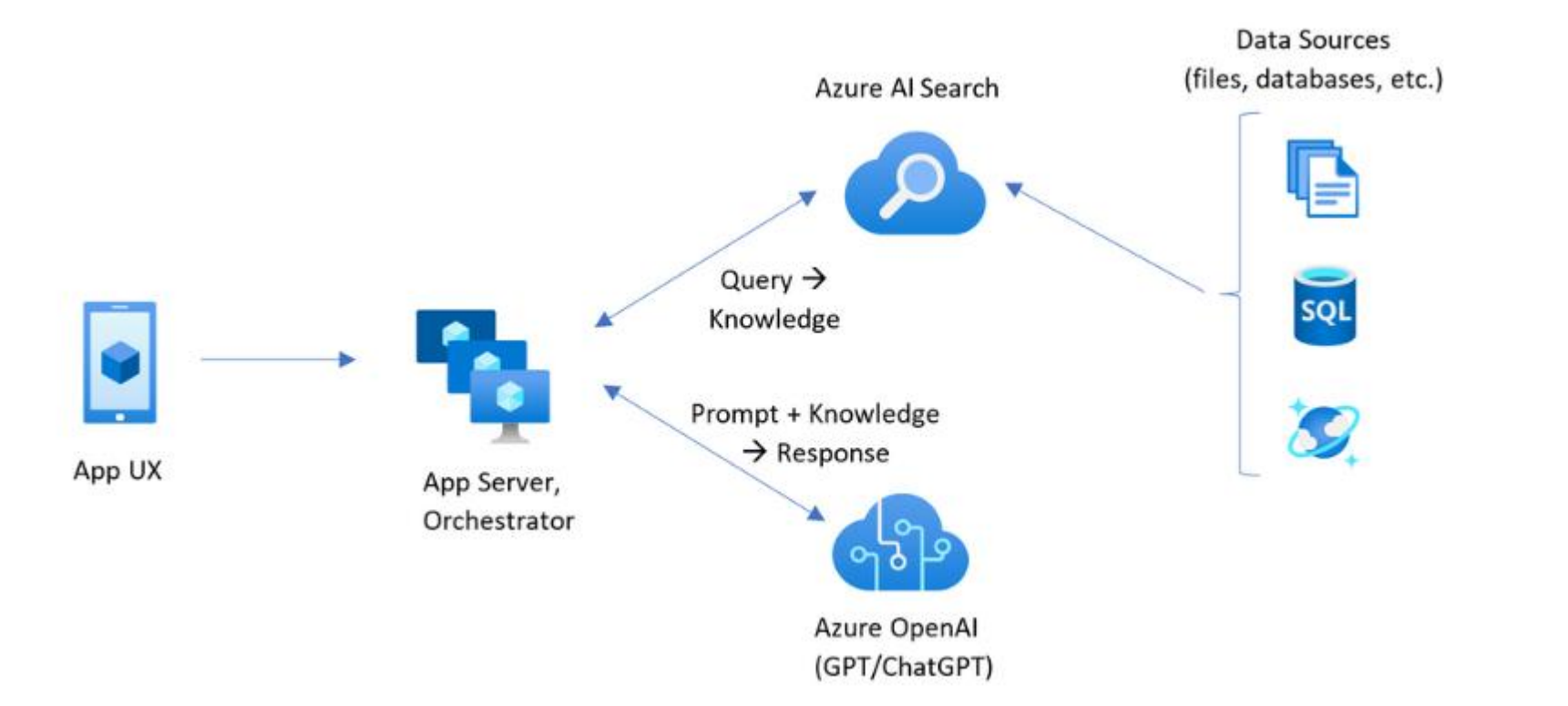
Additional IT/Ops considerations

- Azure Key Vault Security
- Identity and access management
- Access model
  - Managing administrative access to Key Vault
  - Controlling access to Key Vault data
- Managed identity
  - User-assigned
  - System-assigned
- AI services can use Managed Identities to access other Azure resources

Microsoft

What about AI service access to your data?

RAG - Retrieval-Augmented Generation



Purview enables Gen AI productivity

Tools to secure and govern your Copilot use

Address oversharing concerns

Protect against data loss & insider risks

Govern AI use to meet regulations & policies

Secure

Govern

Microsoft Purview

https://purview.microsoft.com/purviewforai/unifiedoverview

Microsoft Purview | Search | Try the new Microsoft Purview

Home
Solutions
Learn
Settings
DSPM for AI

DSPM for AI

- Overview
- Recommended actions
- Reports
- Data assessments (Preview)
- Policies
- Activity explorer

Solutions

Related

- Information Protection
- Data Loss Prevention
- Insider Risk Management

Data Security Posture Management for AI

Discover and secure all AI activity in Microsoft Copilot and other AI apps. Keep your data safe and stay on track with industry regulations. Learn more about DSPM for AI

Recommendations

View all recommendations →

Data security

Protect your data from potential oversharing risks

Use data assessments to identify potential oversharing risks in your organization. They also provide fixes to limit access to sensitive data.

View details

Sensitivity labels on data of top 100 sites

Labeled 16.6K

Not labeled 12.5K

- No sensitive information types detected
- Sensitive information types detected
- Data not scanned

Data security

10% of users have been detected with risky AI usage (preview)

In the last 30 days, risky AI usage has been detected from 1,000 users in your organization. Extend your insights to calculate user risk by detecting risky prompts and responses in Copilot and other AI apps.

Get started

Users with risky AI usage
Last 30 days

Received sensitive prompts in Copilot 594 / 100K

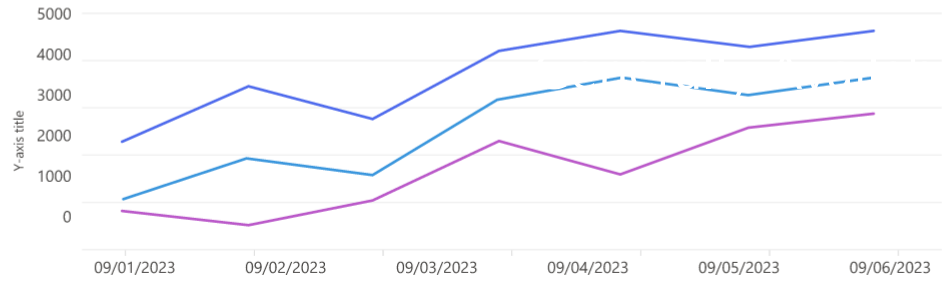
Entered risky prompts in Copilot 456 / 100K

Reports

View all reports →

Total interactions over time (Microsoft Copilot)

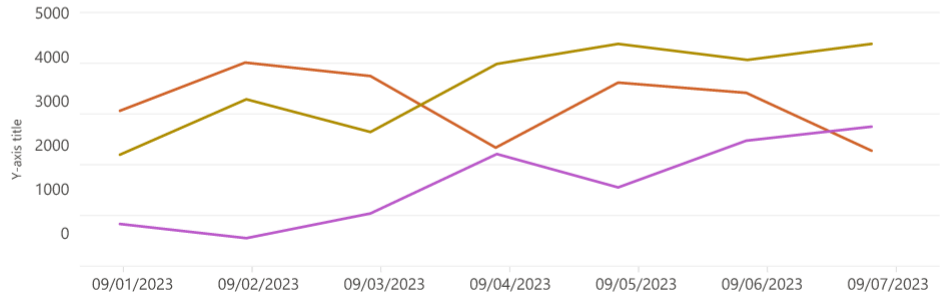
Up 20% in the last 30 days



Microsoft 365 Copilot Studio Microsoft Teams (AI notes in chat)

Total interactions over time (other AI apps)

Up 14% in the last 30 days



Google Gemini Open AI ChatGPT Copilot for Bing

Lab 1 – Deploy & Identity

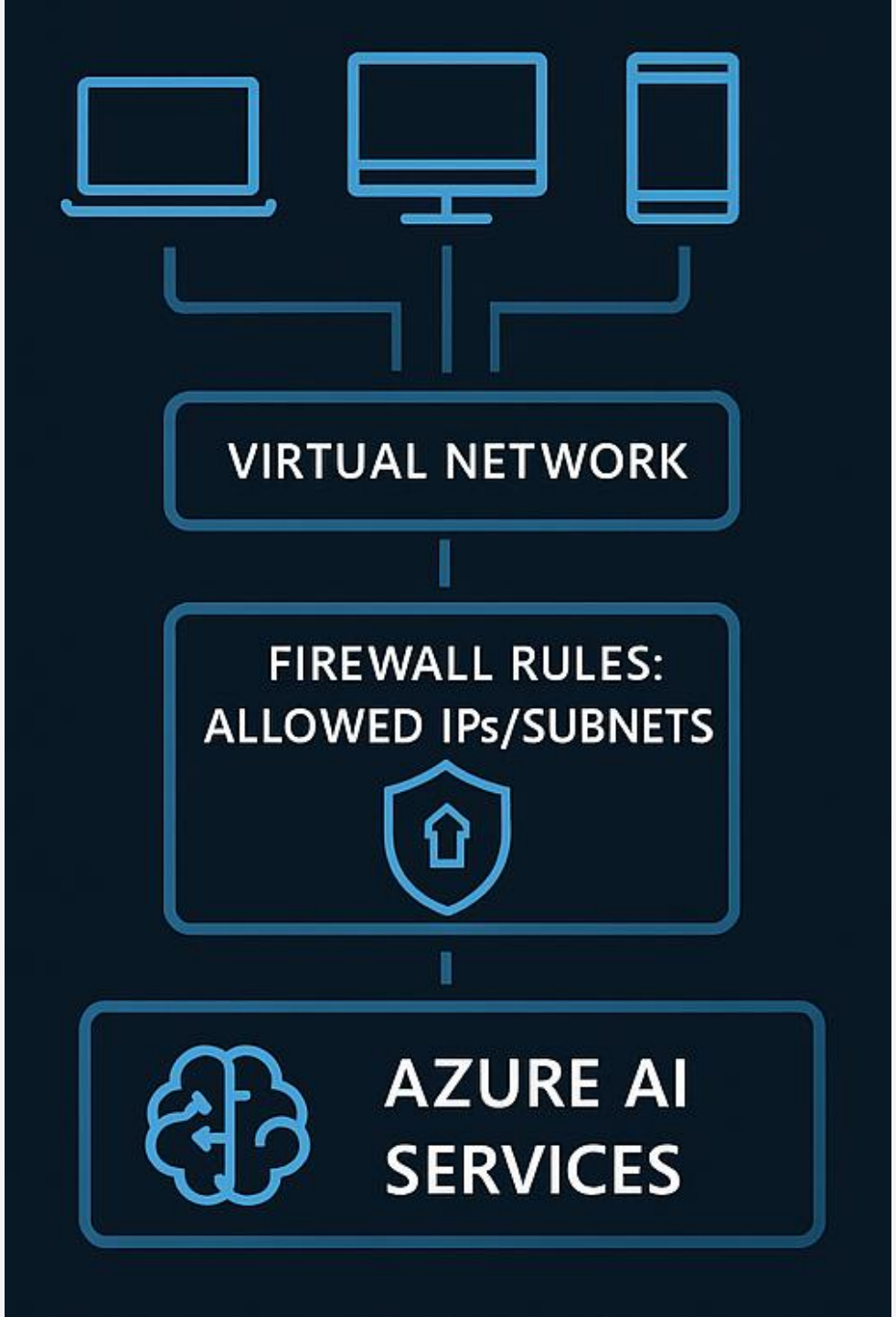
Networking for AI Workloads


Control ALL network traffic


East – West & North – South

Networking for AI Workloads

- **Limit exposure** by specifying allowed IP addresses and virtual networks.
- **Use Azure's virtual network service endpoints to restrict access**:
  - Only requests from permitted networks or IPv4 address ranges reach Azure AI services.
  - Requests from outside these sources are denied.
  - Ensures only trusted sources interact with your AI workloads.



Networking Security Overview - Azure OpenAI

Default posture (out of the box)

- Public endpoint; any client with endpoint + key can call the service
- Convenient for quick starts, not for production

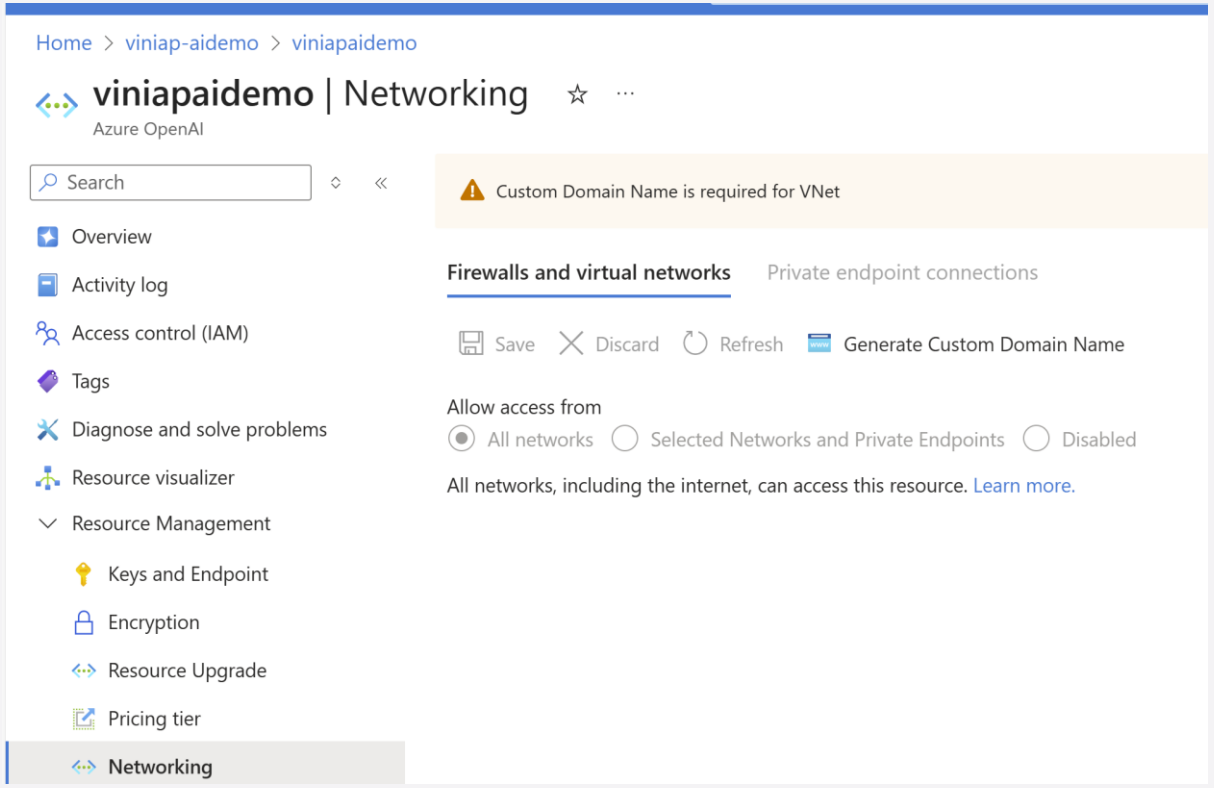
Firewalls & VNets

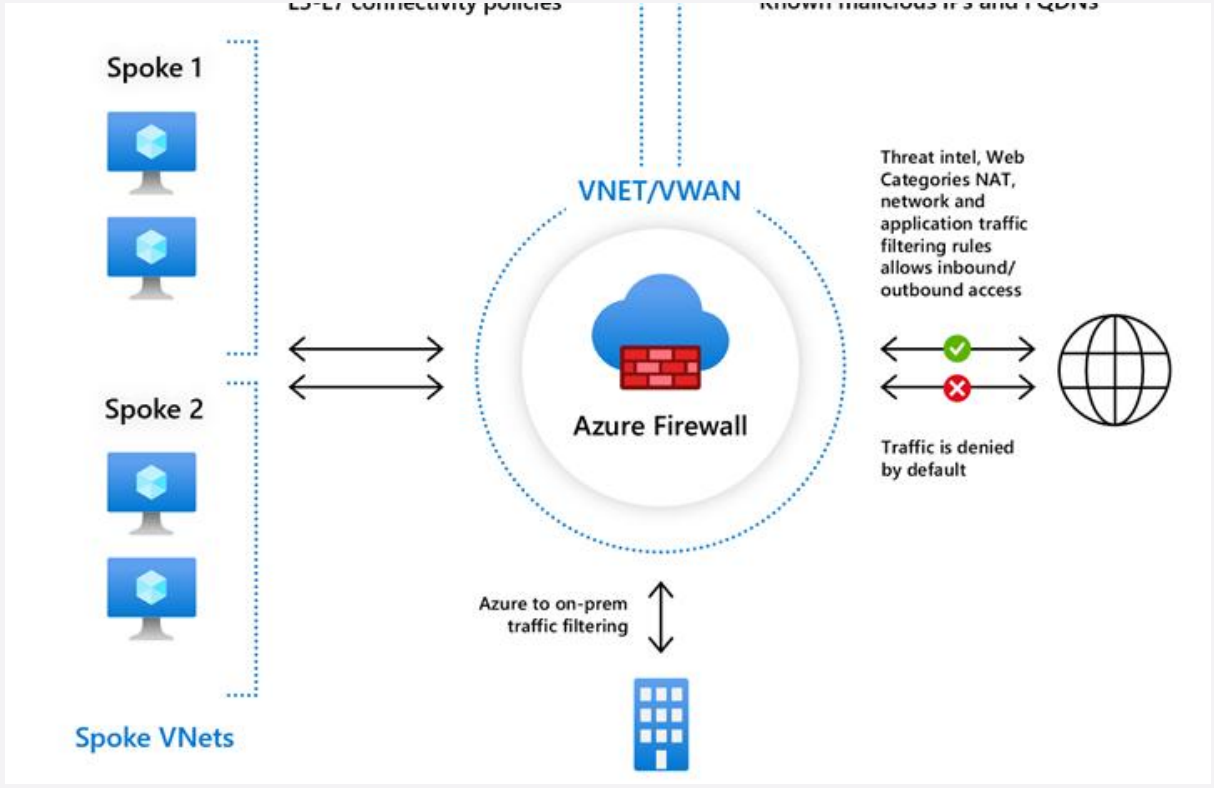
- Restrict access to specific VNets/subnets or IP ranges
- Unauthorized traffic → 403 Access Denied
- Ideal for dev/test environments

Private Endpoints

- Assigns private IP inside your VNet; no public exposure
- Traffic stays on Microsoft backbone
- Best for compliance and production workloads

Paas AI workloads vs IaaS AI workloads





Key Takeaways

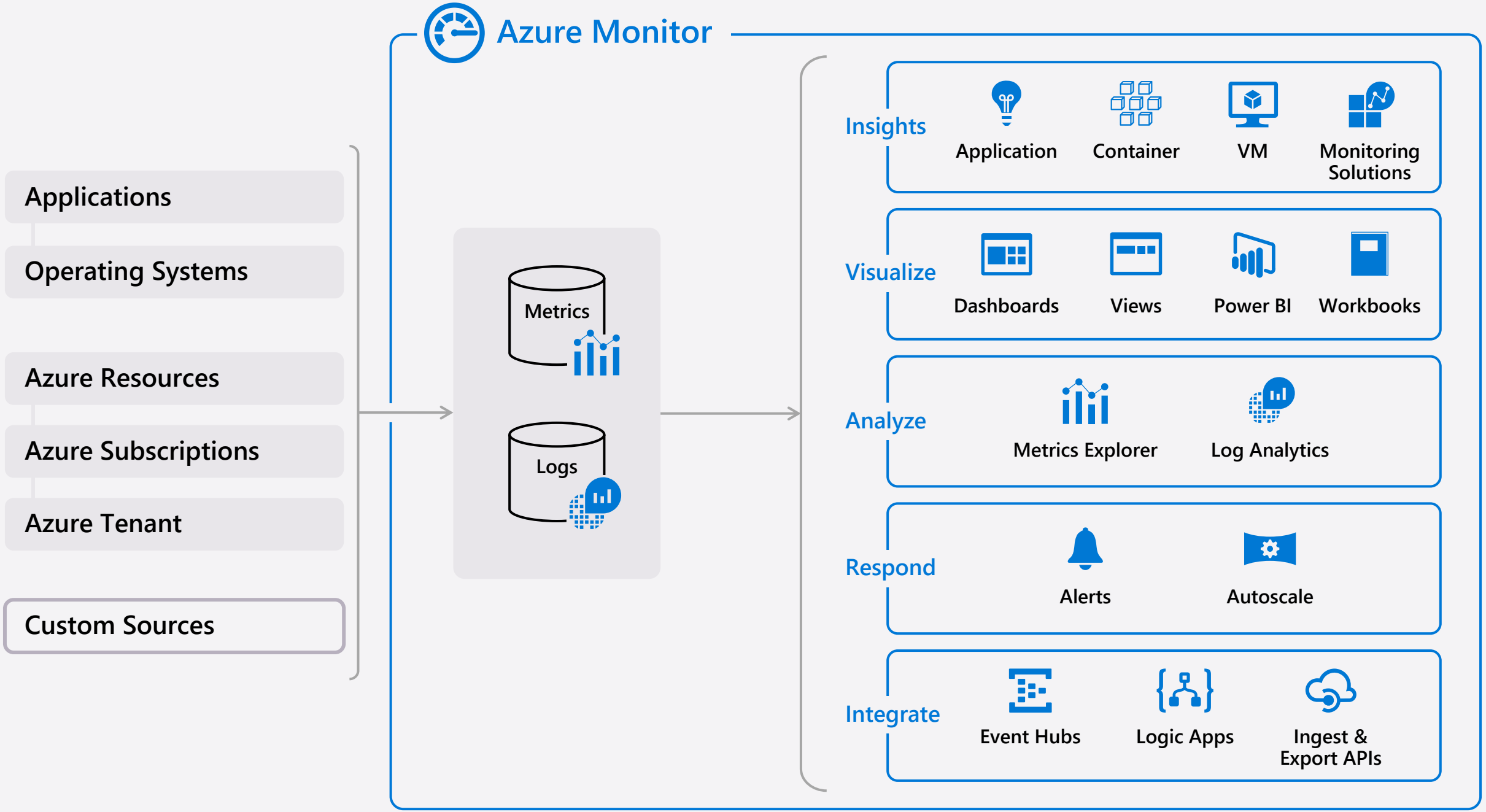
- **Security Layers:**
  - Secrets management + Networking controls = Stronger posture.
- **Best Practices:**
  - Always restrict access to trusted networks.
  - Use Private Endpoints for production workloads.

Additional IT/Ops considerations

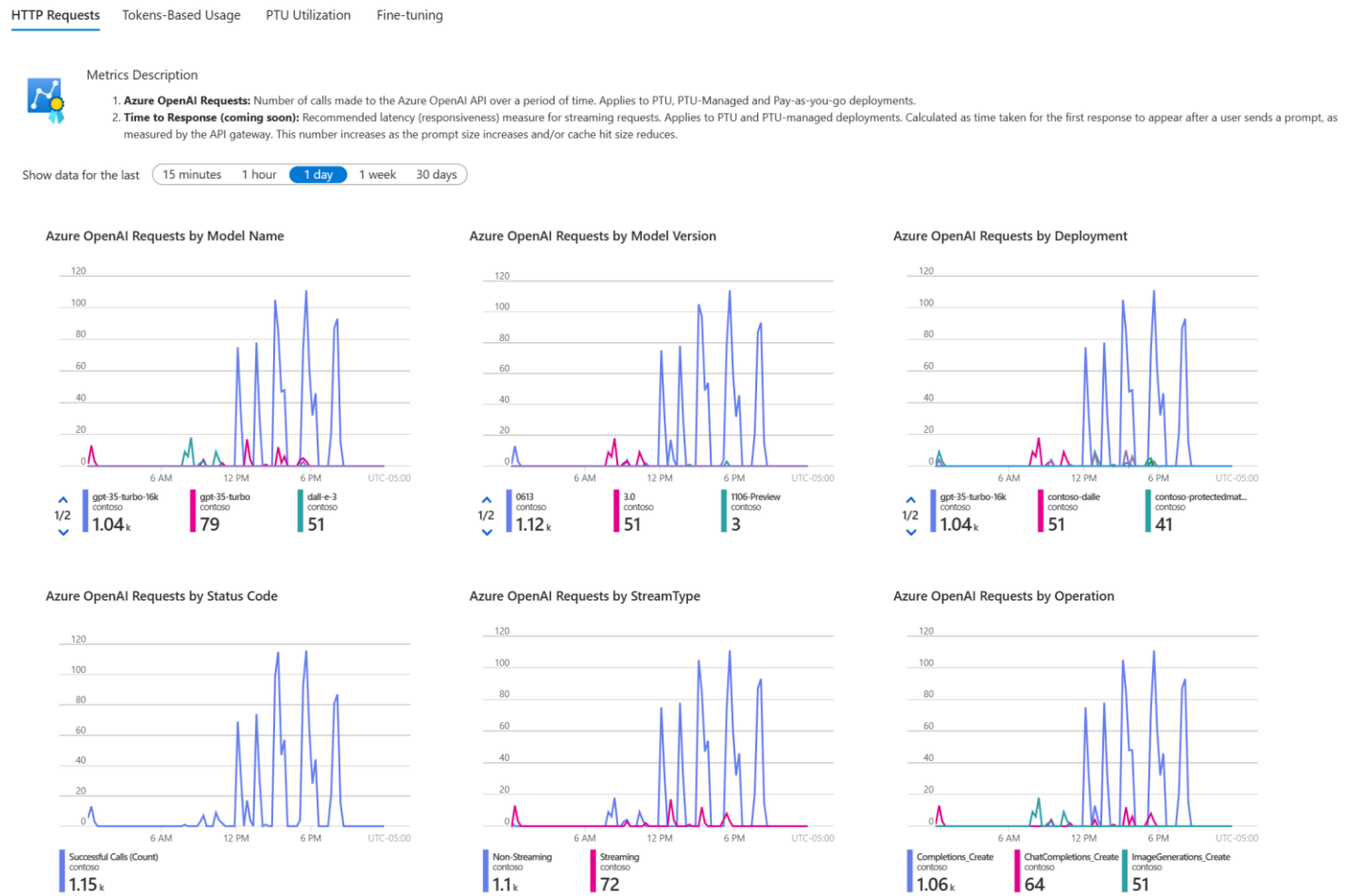
- No network policy by default
- Firewall rules provide additional security layer, but Private Endpoint has higher (more restrict) security requirements
- PE require usage of dedicated HTTP endpoint and DNS translation
- Explore advanced options (Azure Policy, Private DNS Zones).

Lab 2 - Networking

Monitoring and governing AI Apps


Azure Monitor
Applications
Operating Systems
Azure Resources
Azure Subscriptions
Azure Tenant
Custom Sources
Metrics
Logs
Insights
Application
Container
VM
Monitoring Solutions
Visualize
Dashboards
Views
Power BI
Workbooks
Analyze
Metrics Explorer
Log Analytics
Respond
Alerts
Autoscale
Integrate
Event Hubs
Logic Apps
Ingest & Export APIs

Monitoring AI services

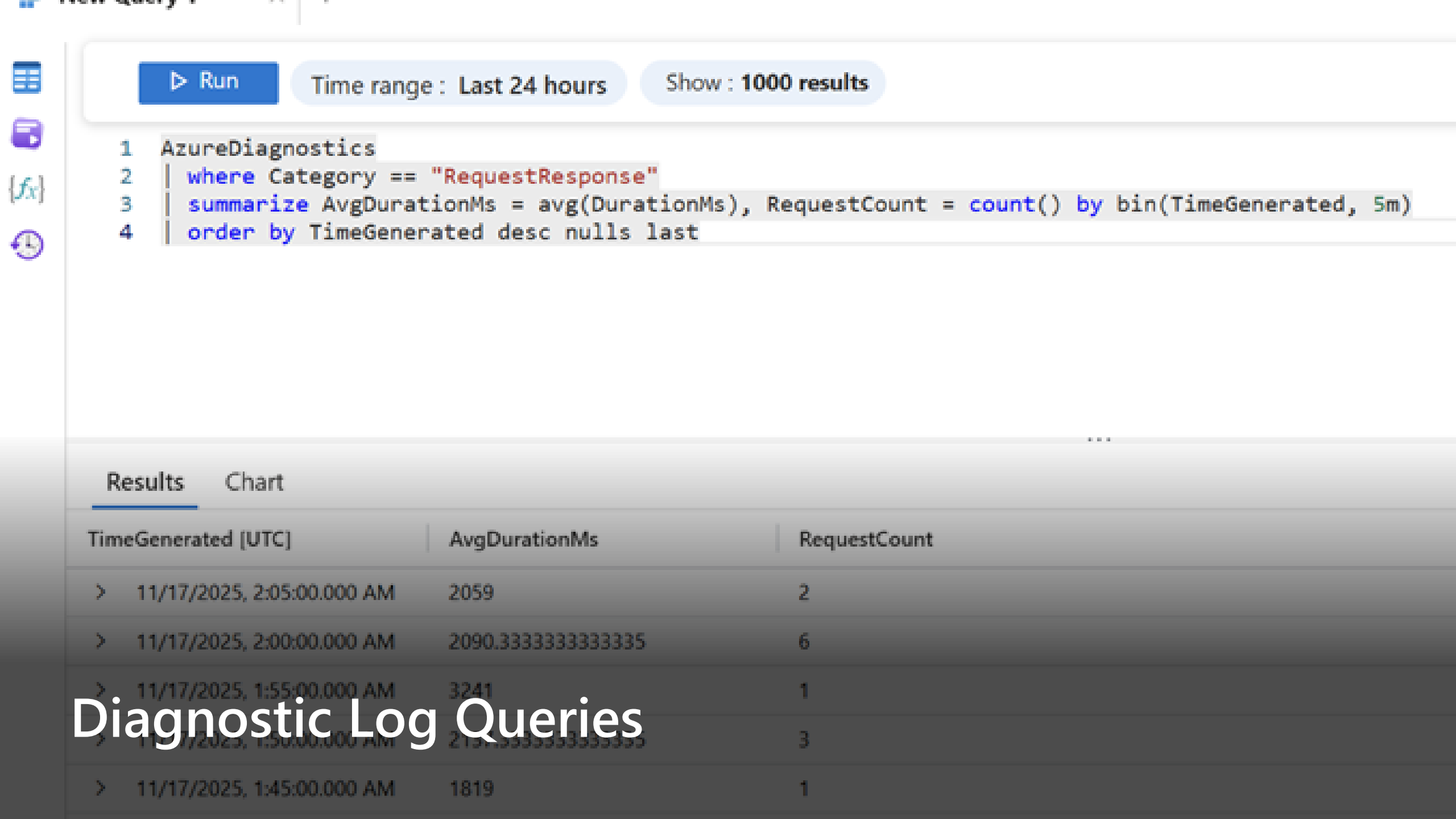


Additional considerations

- Metrics are enabled by default on per-service basis
- Logs are not enabled by default
- AI services have their own set of specific characteristics

From Signal to Action – Service Level Objectives

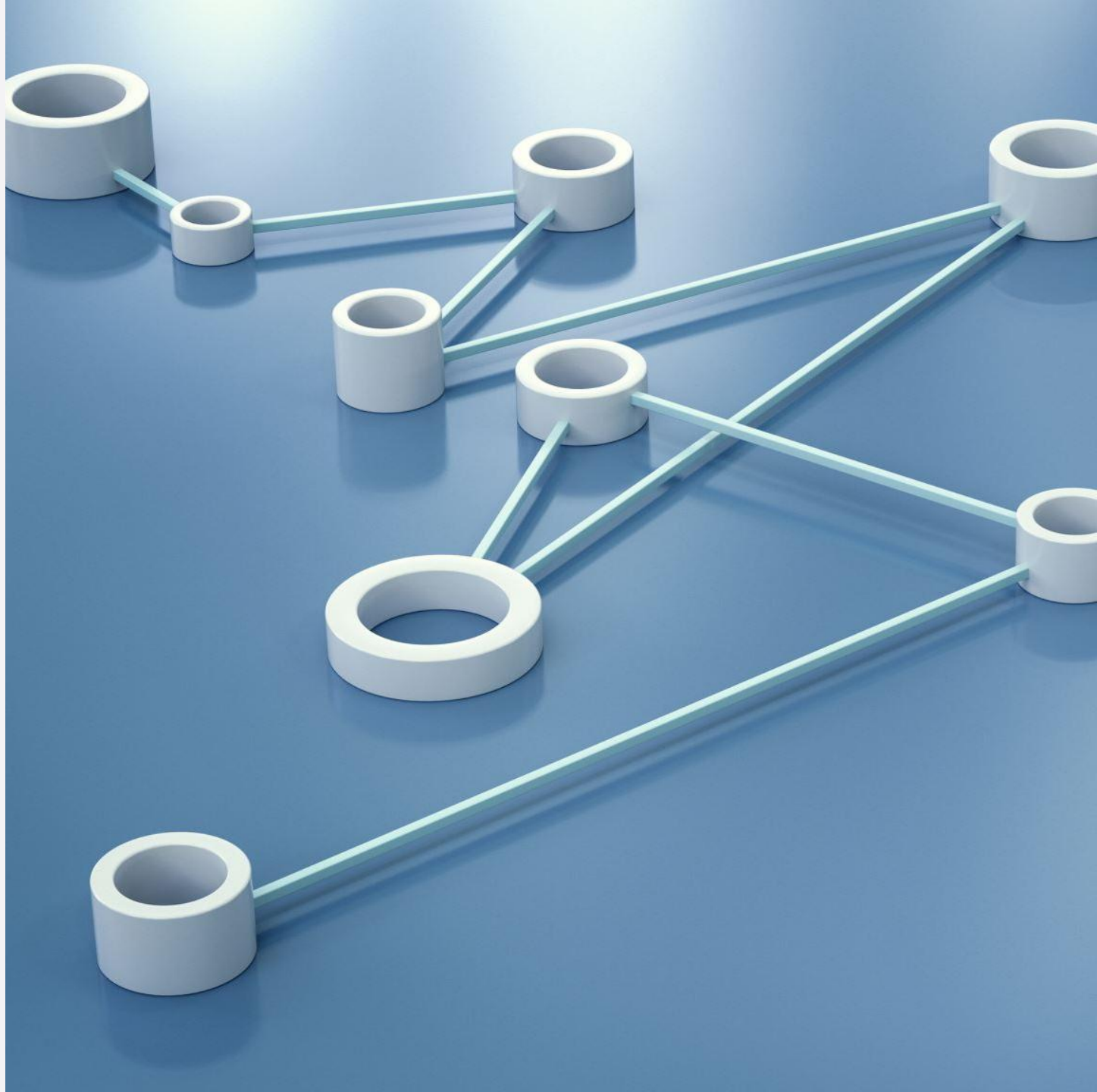
Metric/Log	Alert	Action
• Open AI Availability	• < 99% for 5 minutes	• Alert – Teams or PagerDuty
• Diagnostic Logs	• > 3000ms response time average over 5 minute window	• Open an Azure DevOps ticket with the performance information and time window.


Run
Time range : Last 24 hours
Show : 1000 results
1 AzureDiagnostics
2 | where Category == "RequestResponse"
3 | summarize AvgDurationMs = avg(DurationMs), RequestCount = count() by bin(TimeGenerated, 5m)
4 | order by TimeGenerated desc nulls last
Results
Chart
TimeGenerated [UTC]
AvgDurationMs
RequestCount
11/17/2025, 2:05:00.000 AM
2059
2
11/17/2025, 2:00:00.000 AM
2090.3333333333335
6
11/17/2025, 1:55:00.000 AM
3241
1
11/17/2025, 1:50:00.000 AM
3
11/17/2025, 1:45:00.000 AM
1819
1
Diagnostic Log Queries

Better Operational Awareness with Azure Monitor Workbooks

Governance for AI Apps with Azure Policy

Network and Access Controls
Identity and Key Management
Data Protection
Model Deployment
Monitoring and Compliance



Policy Examples

Policy Definition	Purpose and Enforcement	Applies To
• **Cognitive Services Deployments – Only Approved Models**	• Restricts deployment of AI models to a curated list. Enforces that only **approved or whitelisted AI models** can be deployed, preventing use of unvetted or non-compliant models	• Azure Cognitive Services (incl. Azure OpenAI)

Policy Examples

Policy Definition	Purpose and Enforcement	Applies To
• **Resources Should Disable Local Authentication (Key Access)**	• Requires that **local authentication keys are disabled**, forcing exclusive use of Azure AD (Microsoft Entra ID) tokens for access. Removing static API keys mitigates leaked credential risks and enables granular, identity-based access control	• Azure Cognitive Services (incl. Azure OpenAI)

Policy Examples

Policy Definition	Purpose and Enforcement	Applies To
• **Diagnostic Logs Should Be Enabled**	• Requires that **diagnostic logging** is turned on for cognitive services. This ensures that activities and access to the AI service are recorded for audit and monitoring	• Azure Cognitive Services (incl. Azure OpenAI)

Resources for Monitoring and Governance

Monitor Azure OpenAI	https://aka.ms/prel19/monitor-openai
Azure OpenAI monitoring data reference	https://aka.ms/prel19/monitor-openai-reference
Implement advanced monitoring for Azure OpenAI through a gateway	https://aka.ms/prel19/advanced-openai-monitoring
Govern Azure platform services (PaaS) for AI	https://aka.ms/prel19/govern-openai
Control AI model deployment with built-in policies in Azure AI Foundry portal	https://aka.ms/prel19/govern-model-deployment

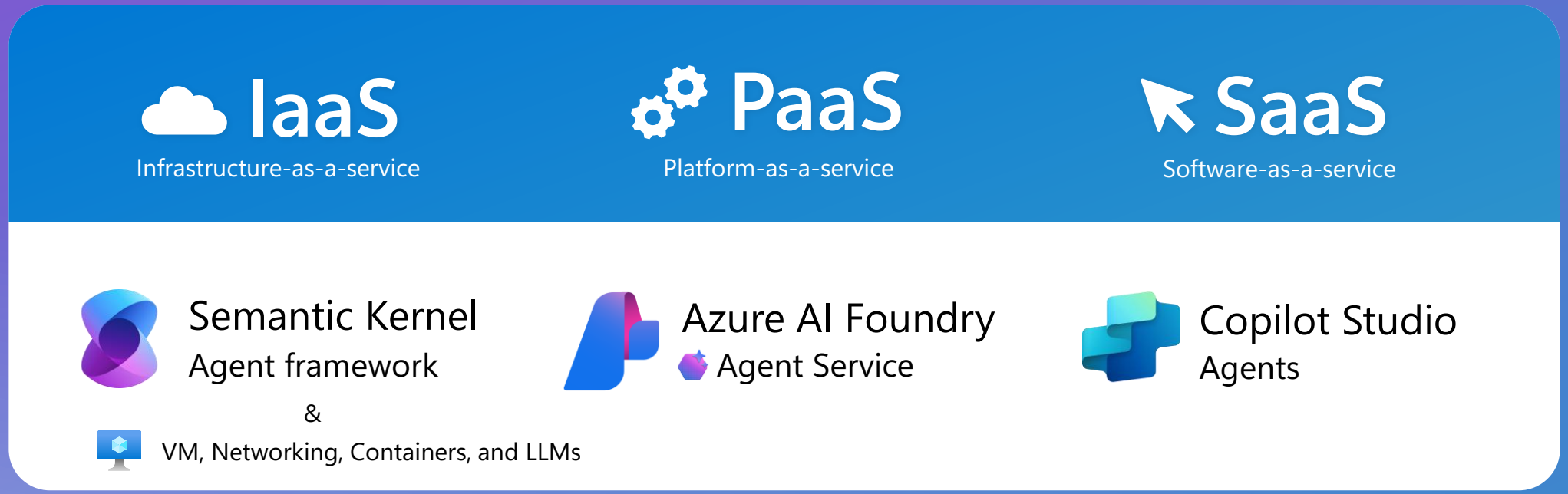
Lab 3 & 4 – Monitoring & Governance

AI Landing Zones

Bilal Amjad

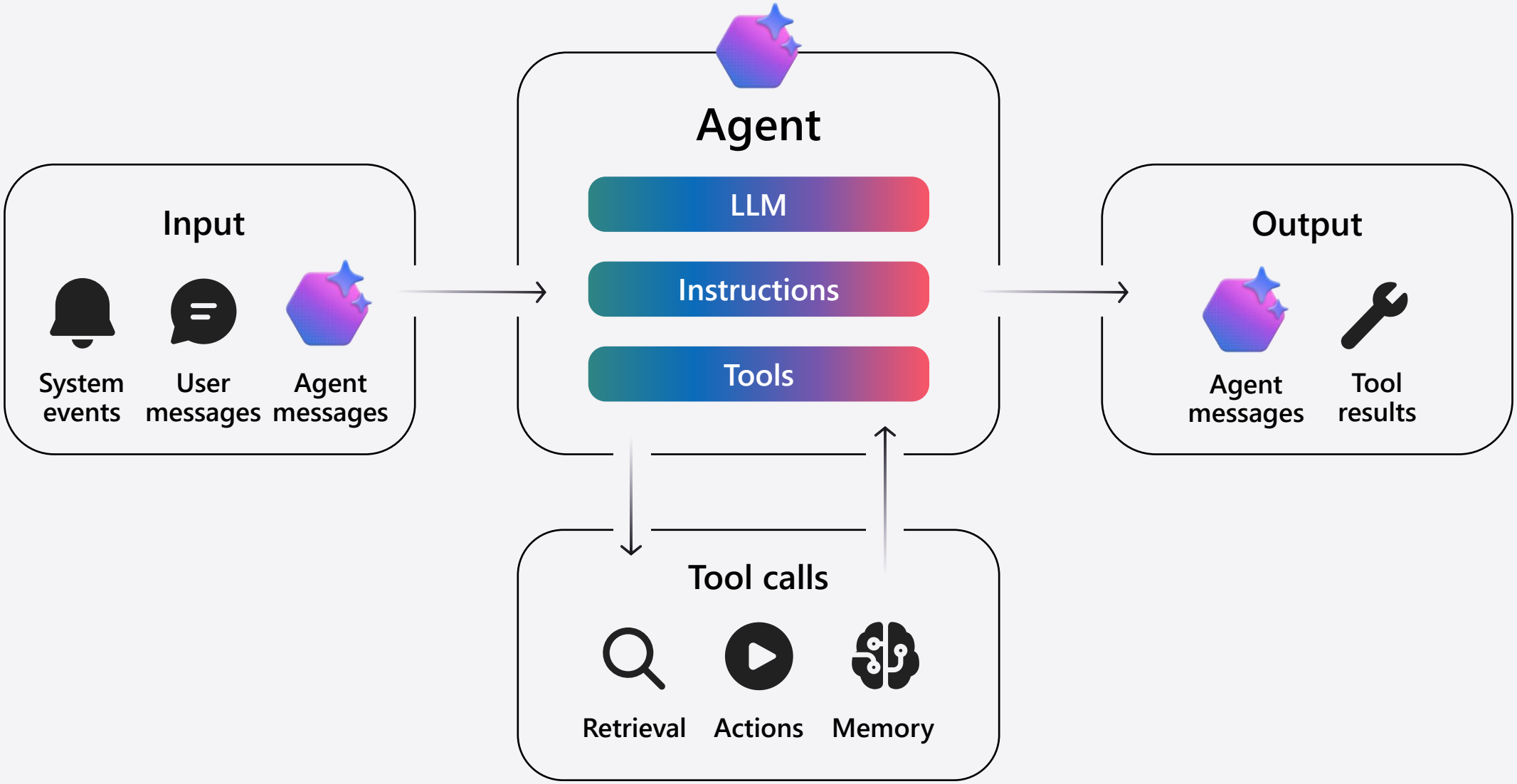
Microsoft has different ways of building agents

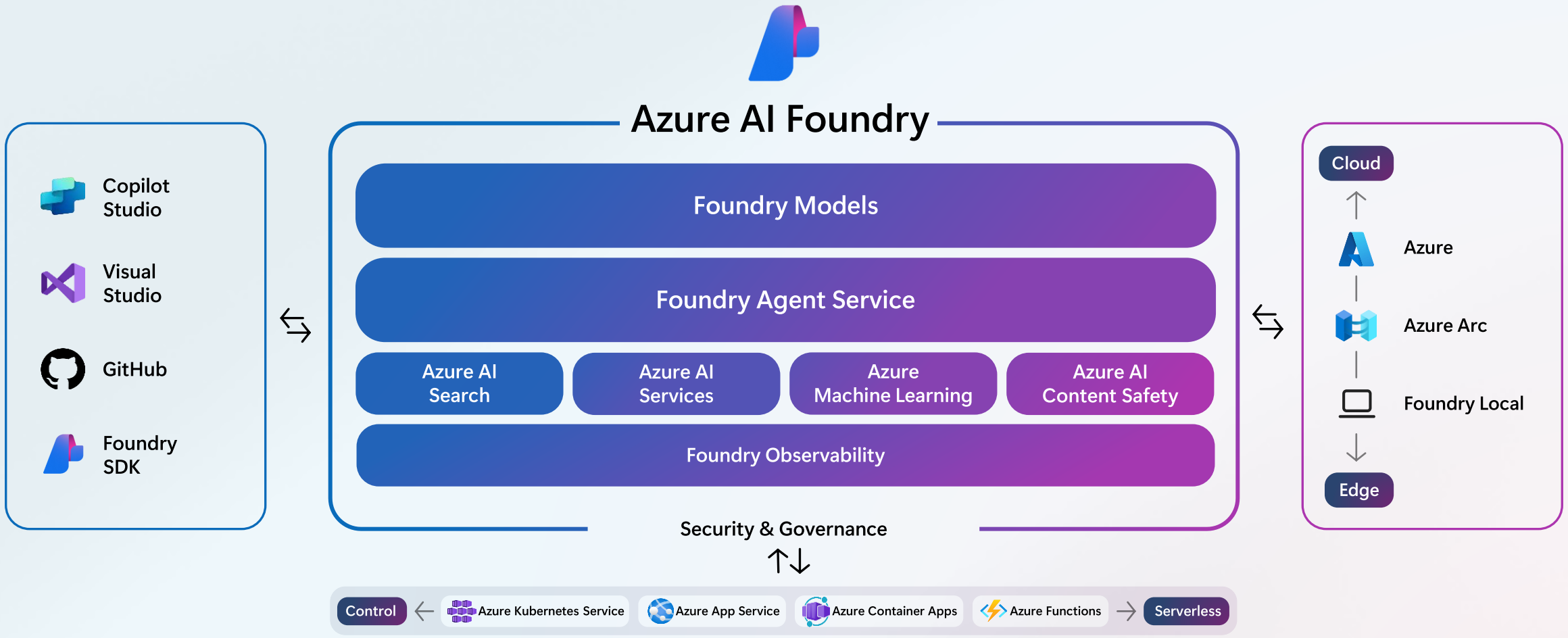
Platform Integrations, ease-of-use, and development speed



Control, visibility, and customization

What is an agent?




Azure AI Foundry
Copilot Studio
Visual Studio
GitHub
Foundry SDK
Foundry Models
Foundry Agent Service
Azure AI Search
Azure AI Services
Azure Machine Learning
Azure AI Content Safety
Foundry Observability
Security & Governance
Control
Azure Kubernetes Service
Azure App Service
Azure Container Apps
Azure Functions
Serverless
Cloud
Azure
Azure Arc
Foundry Local
Edge

Azure AI Foundry Agent Service

Securely customize, orchestrate, and deploy AI agents

Model choice

Model choice and flexibility with the model catalog

Azure OpenAI Service

o1, o3-mini, GPT-4.1, 4o, etc

Models-as-a-Service

- Llama 3.1-405B-Instruct
- Mistral Large
- Cohere-Command-R-Plus

AI tools

Richest set of enterprise connectivity

Knowledge

Actions

Logic Apps | Azure Functions | OpenAPI | MCP

Trust

Customer control over data, networking, and security

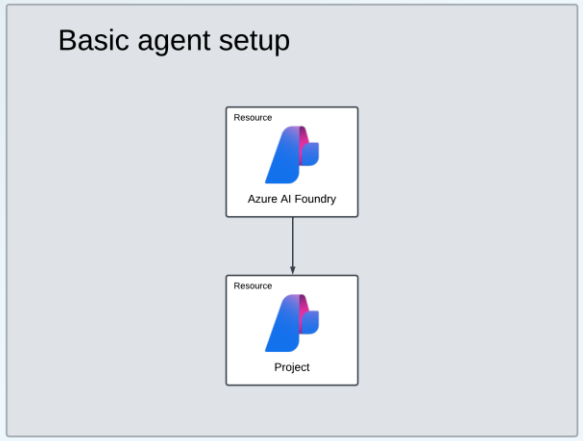
- BYO-file storage
- BYO-search index
- BYO-virtual network
- BYO-thread storage
- OBO authentication
- Content filtering

Built-in Enterprise Readiness

Azure AI Foundry Agent Service offers three different configurations to suit different customer needs:

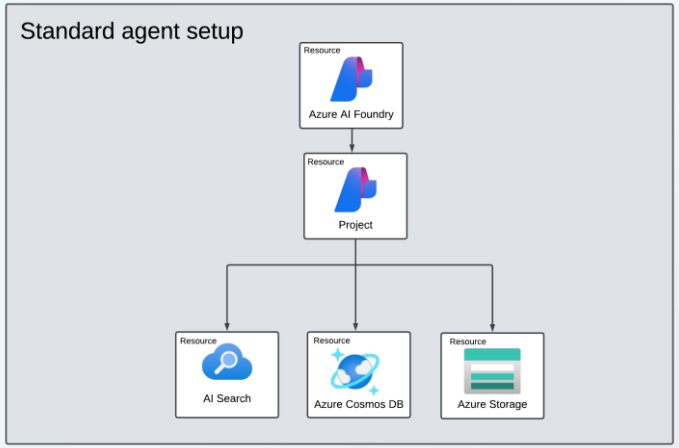
Basic Setup

Compatible with OpenAI Assistants and manages agent states using the platform's built-in storage.



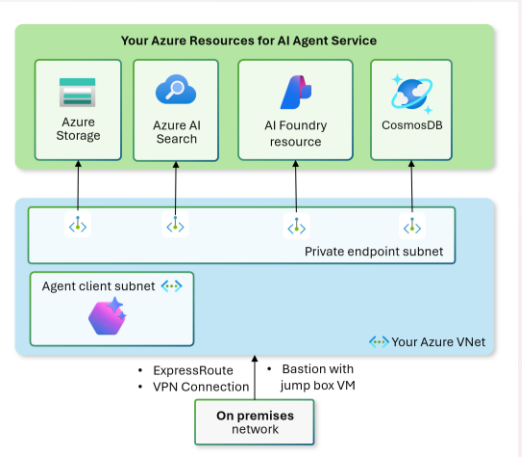
Standard Setup

Provides fine-grained data control by allowing customers to use their own resources. All customer data, including files, threads, and vector stores will remain in your own resources



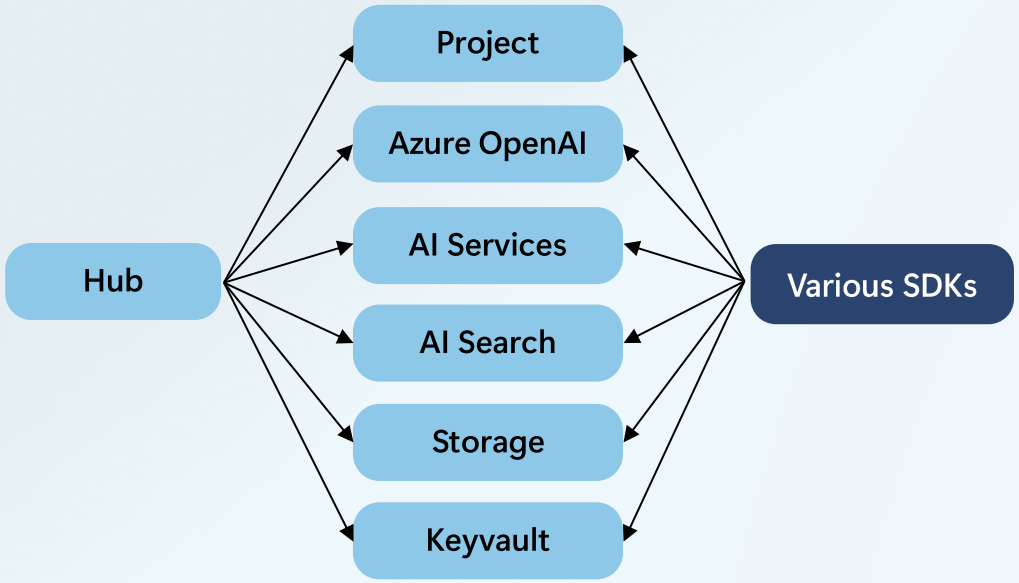
Standard Setup with BYO Virtual Network

Enables full operation within a customer's virtual network, ensuring strict data control and preventing exfiltration by keeping all traffic confined to your network environment



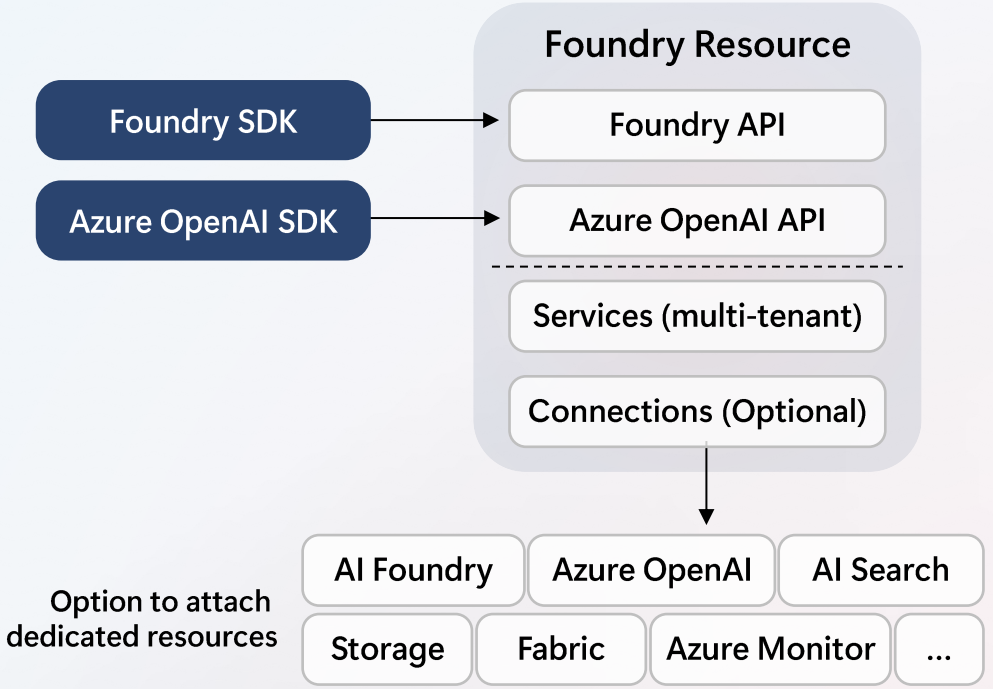
Simplifying the Service

Hub Projects



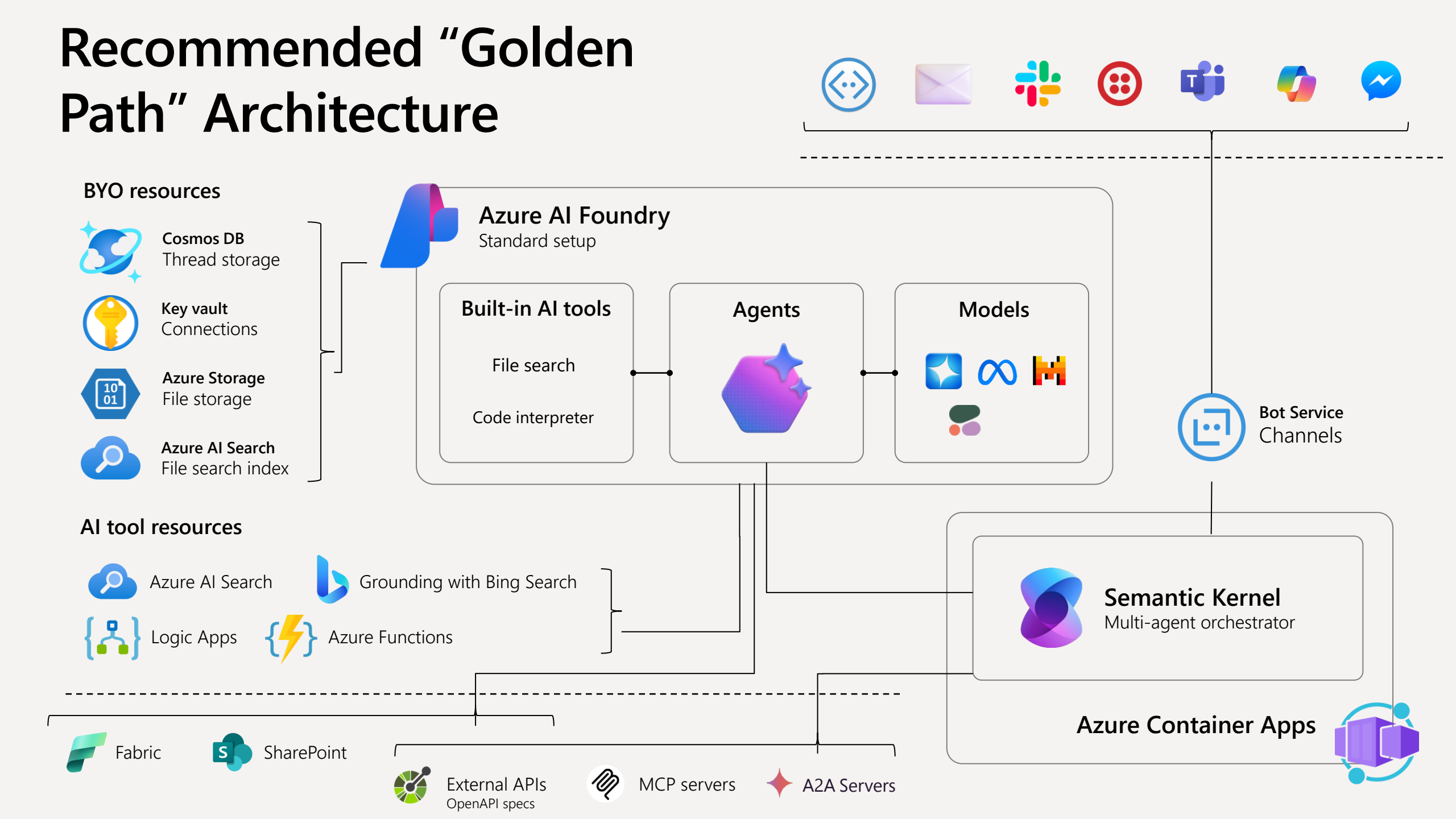
Many different resources and SDKs enabled scale, customization and enterprise controls but made setup and coding complicated

Foundry projects *New*



Simplified setup and coding, with ability to attach dedicated resources for enterprise scale and management

Recommended “Golden Path” Architecture



AI landing zones

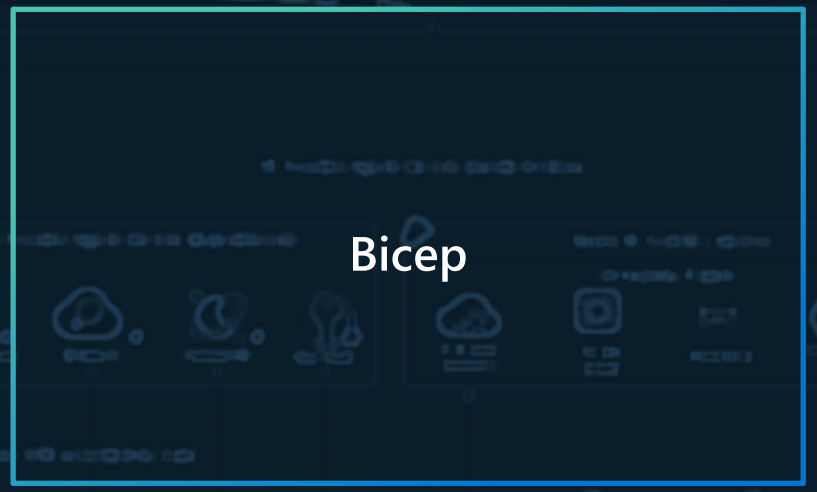
Design Framework

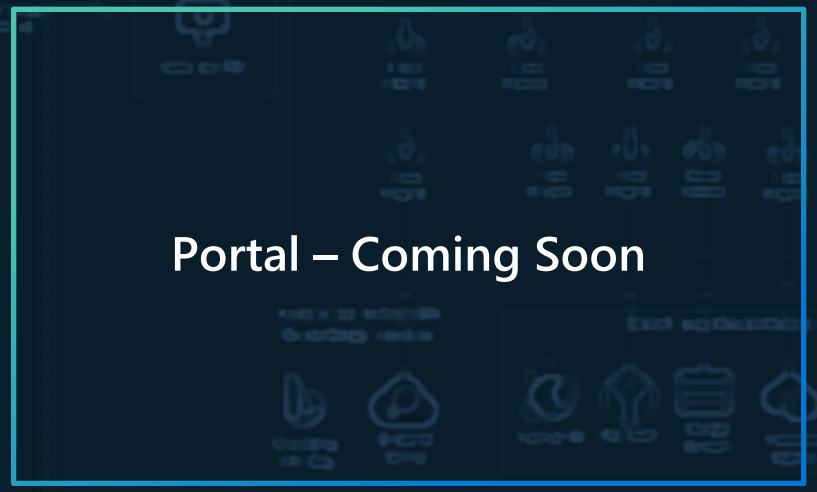
Reference Architectures

Extensble Implementations

Design Framework

Extensible Implementations

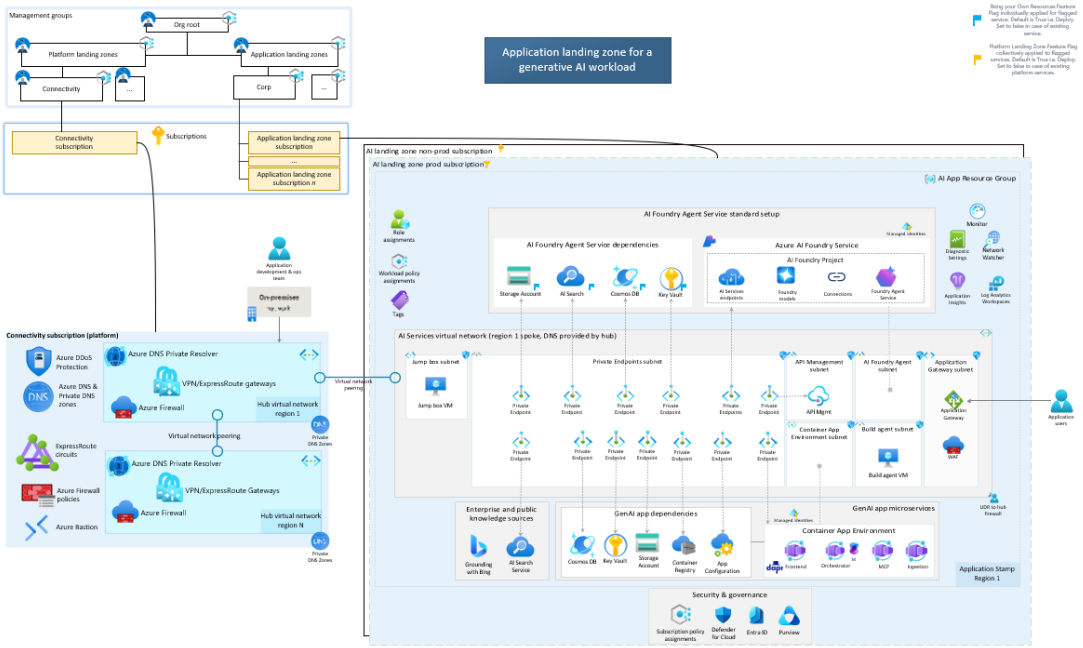




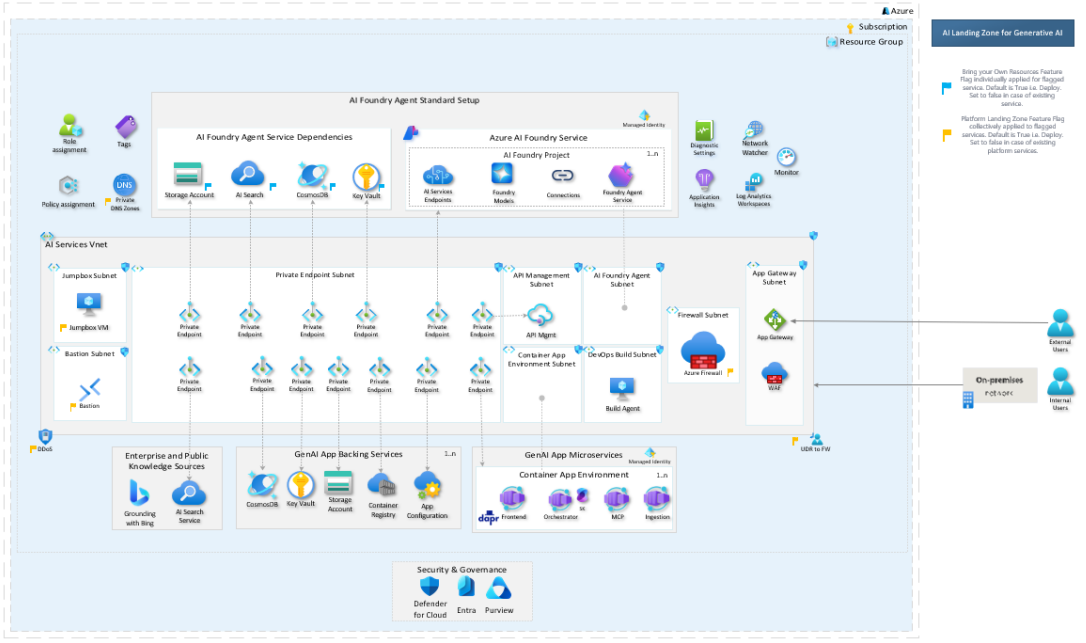
Reference Architectures

Enterprise-Scale and Production Ready to accelerate AI use cases

AI Landing Zone with Platform Landing Zone

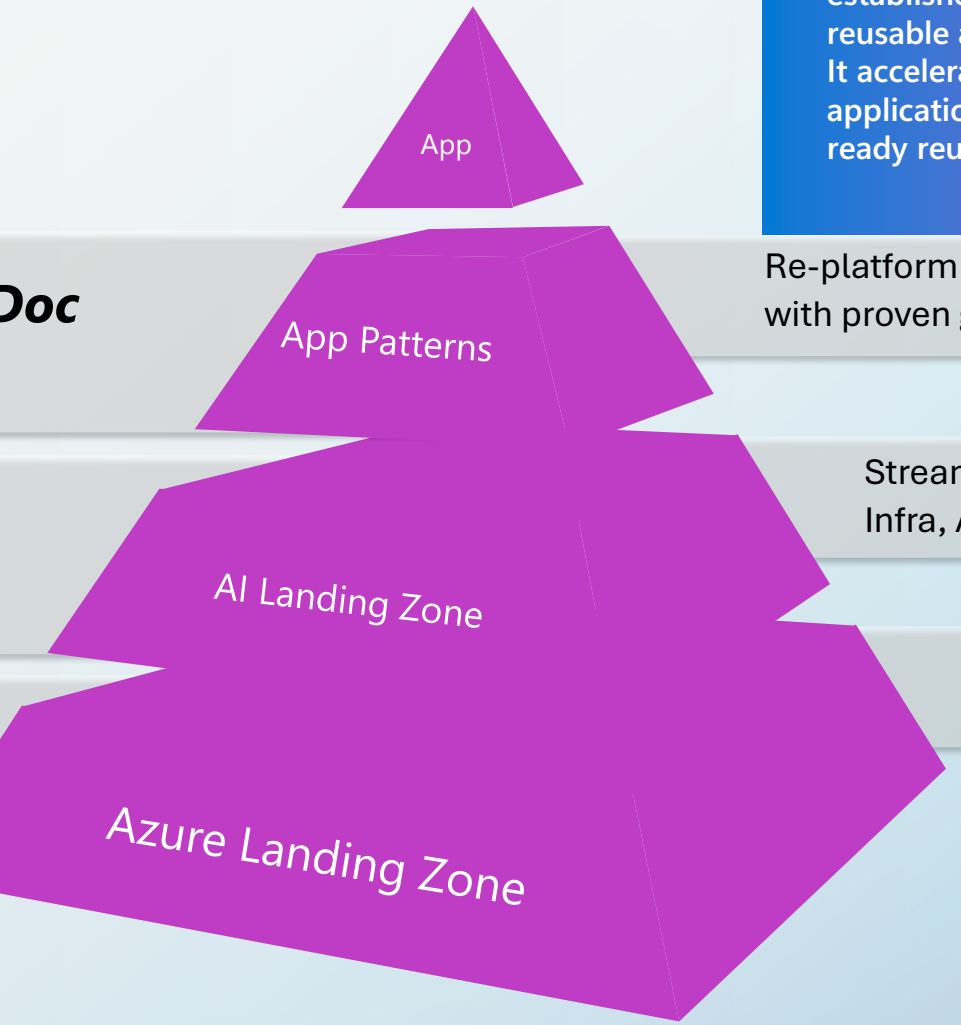


AI Landing Zone without Platform Landing Zone



AI Landing Zone Accelerator

AI Landing Zones offer prescriptive guidance, established cloud developer practices and offer reusable assets built upon Landing Zone Accelerators. It accelerates POC development and expedites application deployment by leveraging production ready reusable infrastructure files.



AI Use Cases (GPT-RAG, Doc Processing, Fine Tuning)

Re-platform .NET and Java apps with proven guidance.

Application Infrastructure

Streamline workload deployment for Infra, App, Data & Security

Platform Infrastructure

Set up secure, scalable well-governed cloud environment.

Authoritative resources for accelerating your app migration and modernization journey.

Microsoft

Demo

Explore AI Landing Zones

AI Landing Zone

https://aka.ms/AILZ

Wrap-up + Resources

Wrap-up

- While AI has its specificities, your knowledge in Ops will be leveraged
- You should plan to operate and manage AI services like any other Azure services
- Partner with your development team to establish a secure, robust operation approach

Resources

Microsoft Learn Plan	https://aka.ms/PrepForAIWorkloads-Learn
ITOpsTalk blog	https://aka.ms/ITOpsTalk
ITOpsTalk YouTube channel	https://youtube.com/@itopstalk

Microsoft

Thank you for your feedback!

On the Ignite page:
Please give us feedback

https://aka.ms/ignite25-feedback

An IT Pro's guide to Deploying and managing AI applications

Your feedback is very important to us!

Microsoft

AI for IT/Ops workshop

Continue Learning at Home
https://github.com/microsoft/AIforITOps

Get started and build Azure skills on MS Learn

PREL19

Coming Spring 2026

SQL AI Database Developer Associate Certification

Empower Your Future with AI-Ready SQL Development Skills:

- AI-augmented SQL development
- Building AI-powered LLM-integrated apps
- Security, governance and responsible AI

Microsoft