Microsoft

Microsoft AI Tour

Microsoft

Securing AI applications on Azure

Joylynn Kirui
Senior Cloud Security Advocate

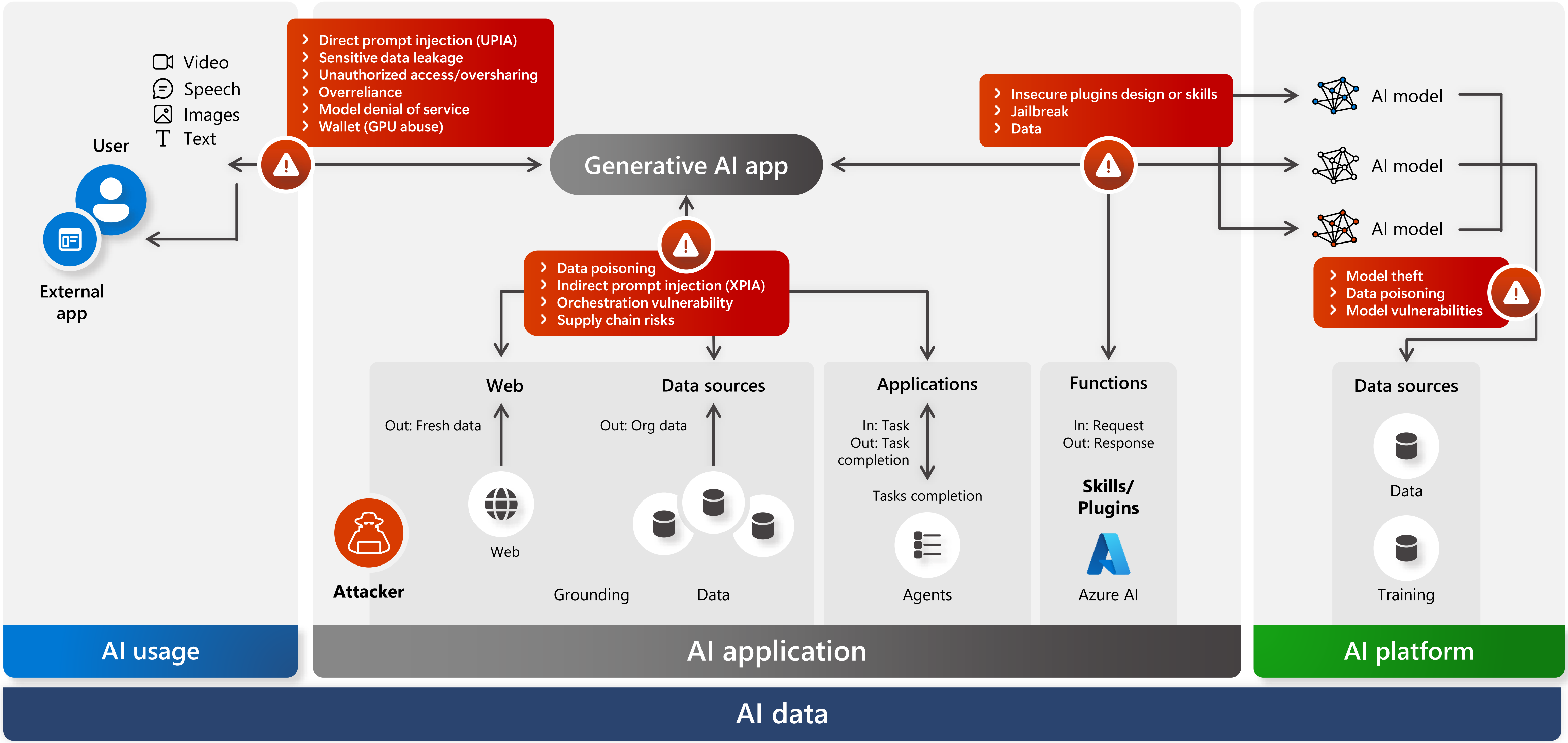
@joylynn_kirui

Agenda

1. Introduction
2. AI safety
3. Authentication and authorization with Microsoft Entra
4. Network security for AI apps
5. Continuous security for AI
6. Wrap up

Introduction

Generative AI threat landscape



Video
Speech
Images
Text

User

External app

- Direct prompt injection (UPIA)
- Sensitive data leakage
- Unauthorized access/oversharing
- Overreliance
- Model denial of service
- Wallet (GPU abuse)

Generative AI app

- Insecure plugins design or skills
- Jailbreak
- Data

AI model

AI model

AI model

- Data poisoning
- Indirect prompt injection (XPIA)
- Orchestration vulnerability
- Supply chain risks

- Model theft
- Data poisoning
- Model vulnerabilities

Web
Out: Fresh data
Web

Data sources
Out: Org data
Grounding
Data

Attacker

Applications
In: Task
Out: Task completion
Tasks completion
Agents

Functions
In: Request
Out: Response
Skills/ Plugins
Azure AI

Data sources
Data
Training

AI usage

AI application

AI platform

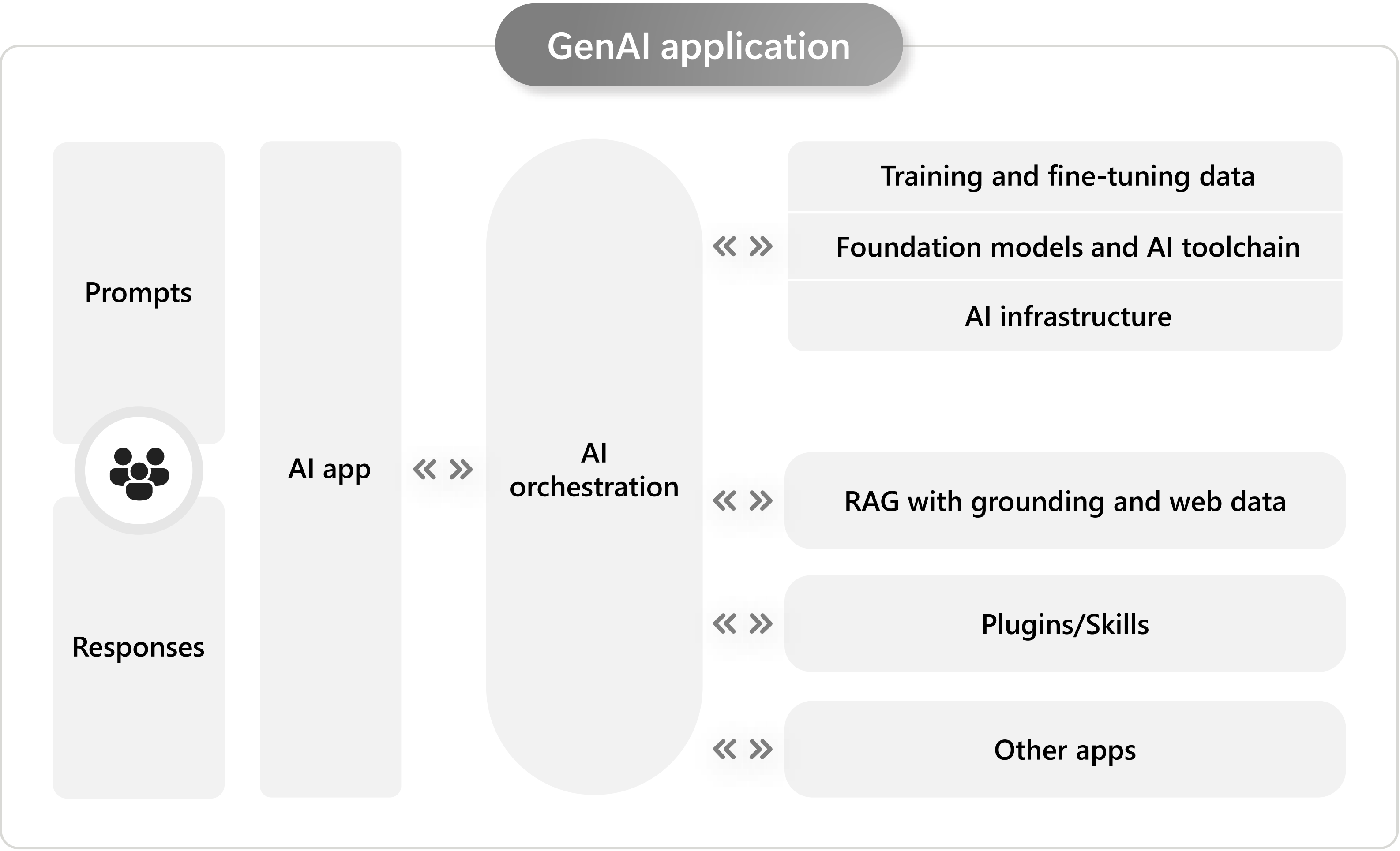
AI data

Generative AI introduces new attack surfaces

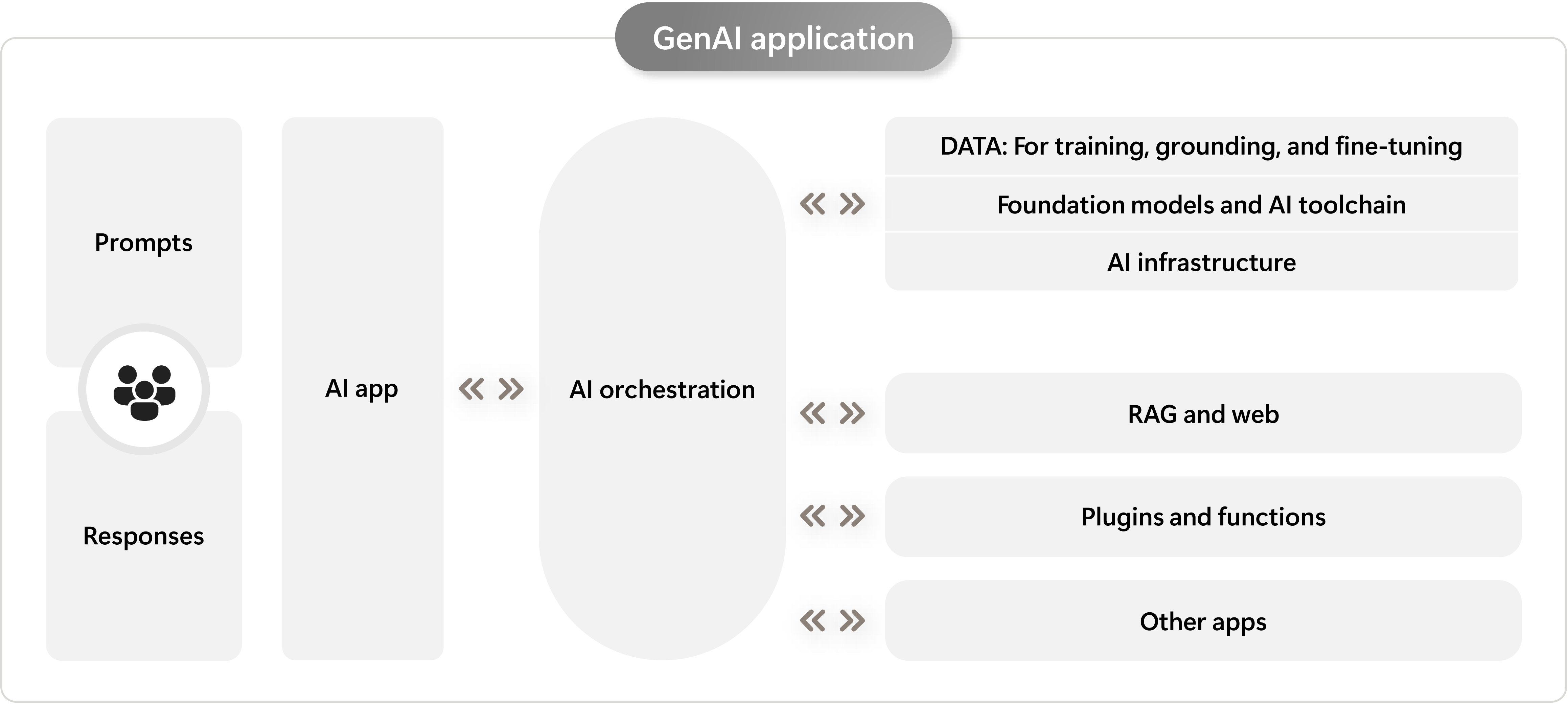
High connectivity to data

Natural languages

Non-deterministic



GenAI extends your attack surface



GenAI application

Prompts

Responses

AI app

AI orchestration

DATA: For training, grounding, and fine-tuning

Foundation models and AI toolchain

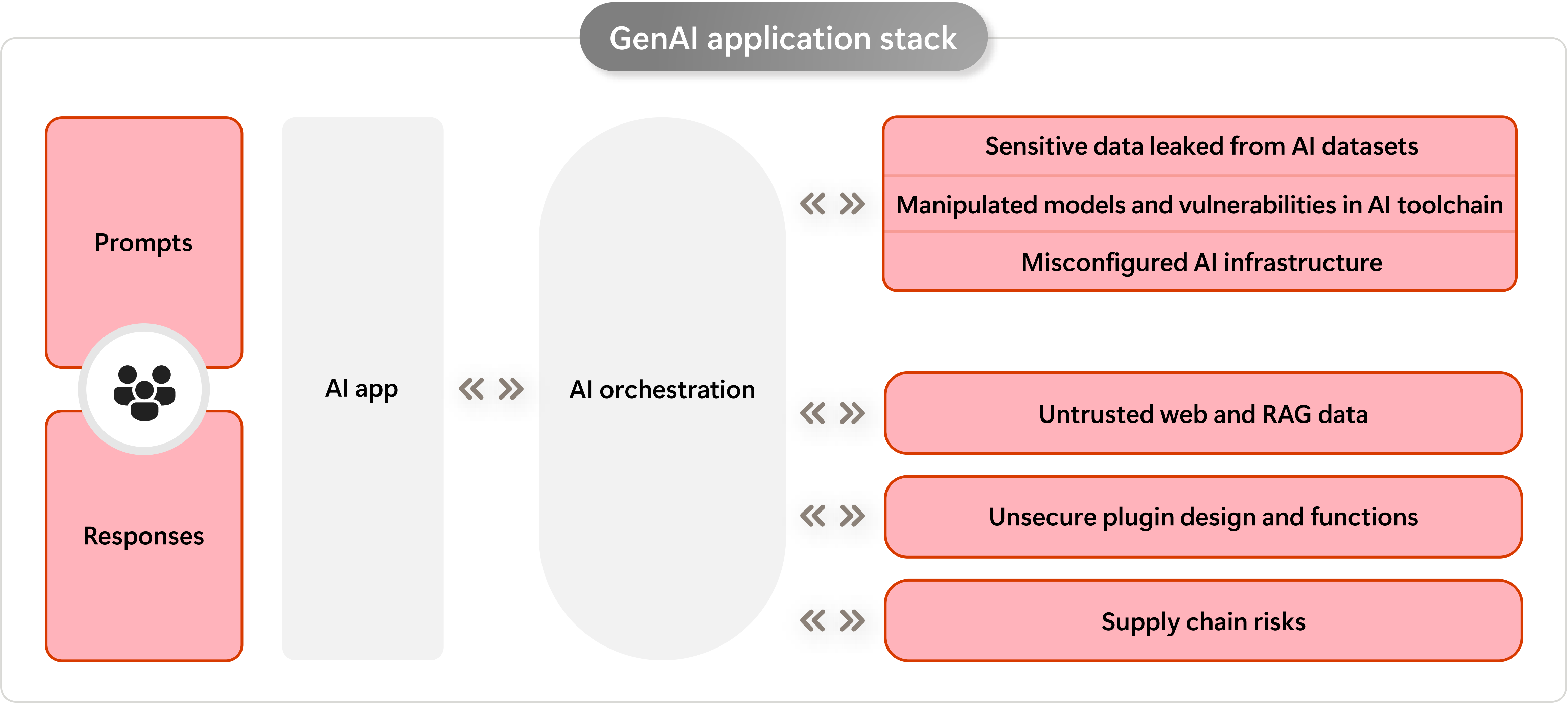
AI infrastructure

RAG and web

Plugins and functions

Other apps

The new GenAI stack extends your attack surface



GenAI application stack

Prompts

Responses

AI app

AI orchestration

Sensitive data leaked from AI datasets

Manipulated models and vulnerabilities in AI toolchain

Misconfigured AI infrastructure

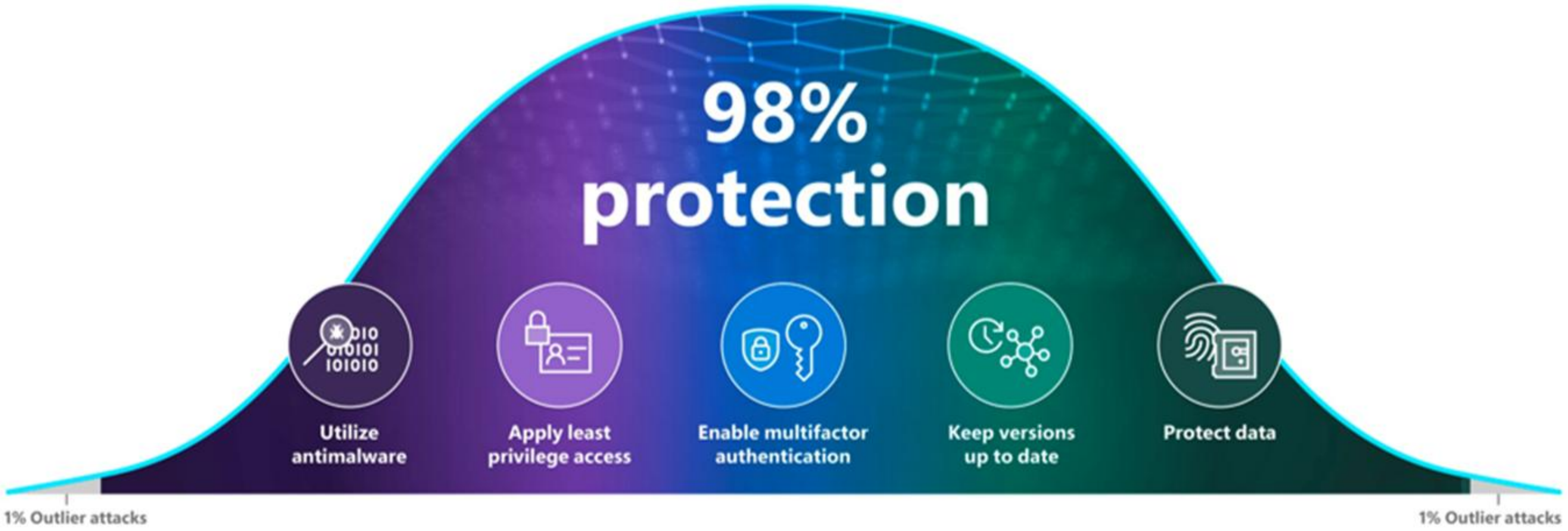
Untrusted web and RAG data

Unsecure plugin design and functions

Supply chain risks

The cybersecurity bell curve

Basic security hygiene still protects against 98% of attacks[1]



Enable multifactor authentication

Make it harder for bad actors to utilize stolen or phished credentials by enabling multifactor authentication. Always authenticate and authorize based on all available data points, including user identity, location, device health, service or workload, data classification, and anomalies.

Apply least privilege access

Prevent attackers from spreading across the network by applying least privilege access principles, which limits user access with just-in-time and just-enough-access (JIT/JEA), risk-based adaptive polices, and data protection to help secure both data and productivity.

Keep up to date

Mitigate the risk of software vulnerabilities by ensuring your organization's devices, infrastructure, and applications are kept up to date and correctly configured. Endpoint management solutions allow policies to be pushed to machines for correct configuration and ensure systems are running the latest versions.

Utilize antimalware

Stop malware attacks from executing by installing and enabling antimalware solutions on endpoints and devices. Utilize cloud-connected antimalware services for the most current and accurate detection capabilities.

Protect data

Know where your sensitive data is stored and who has access. Implement information protection best practices such as applying sensitivity labels and data loss prevention policies. If a breach does occur, it's critical that security teams know where the most sensitive data is stored and accessed.

AI Safety

Microsoft's Responsible AI Principles

Fairness
AI systems should treat all people fairly.

Reliability and safety
AI systems should perform reliably and safely.

Privacy and security
AI systems should be secure and respect privacy.

Inclusiveness
AI systems should empower everyone and engage people.

Transparency
AI systems should be understandable.

Accountability
People should be accountable for AI systems.

Microsoft's Responsible AI Principles

Fairness
AI systems should treat all people fairly.

Reliability and safety
AI systems should perform reliably and safely.

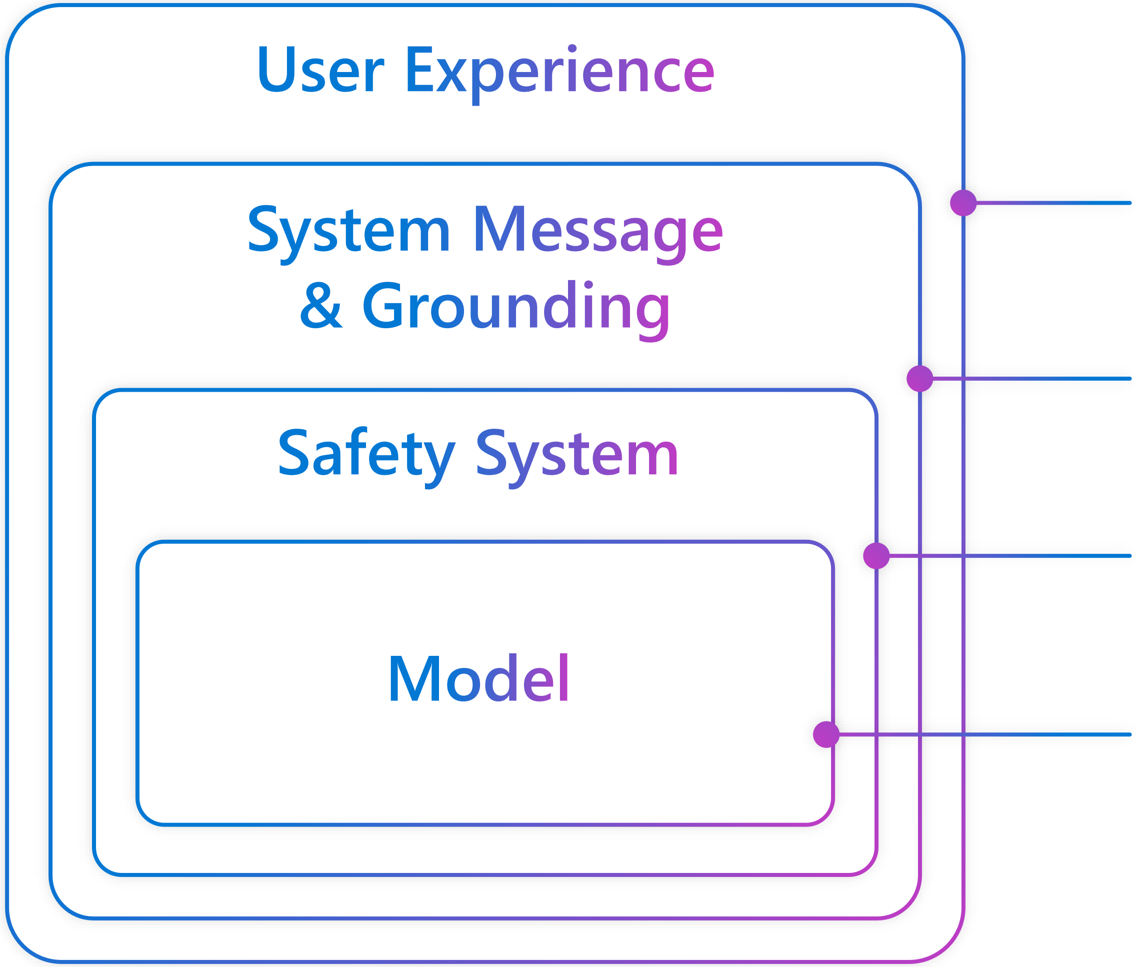
Privacy and security
AI systems should be secure and respect privacy.

Inclusiveness
AI systems should empower everyone and engage people.

Transparency
AI systems should be understandable.

Accountability
People should be accountable for AI systems.

Risk mitigation layers



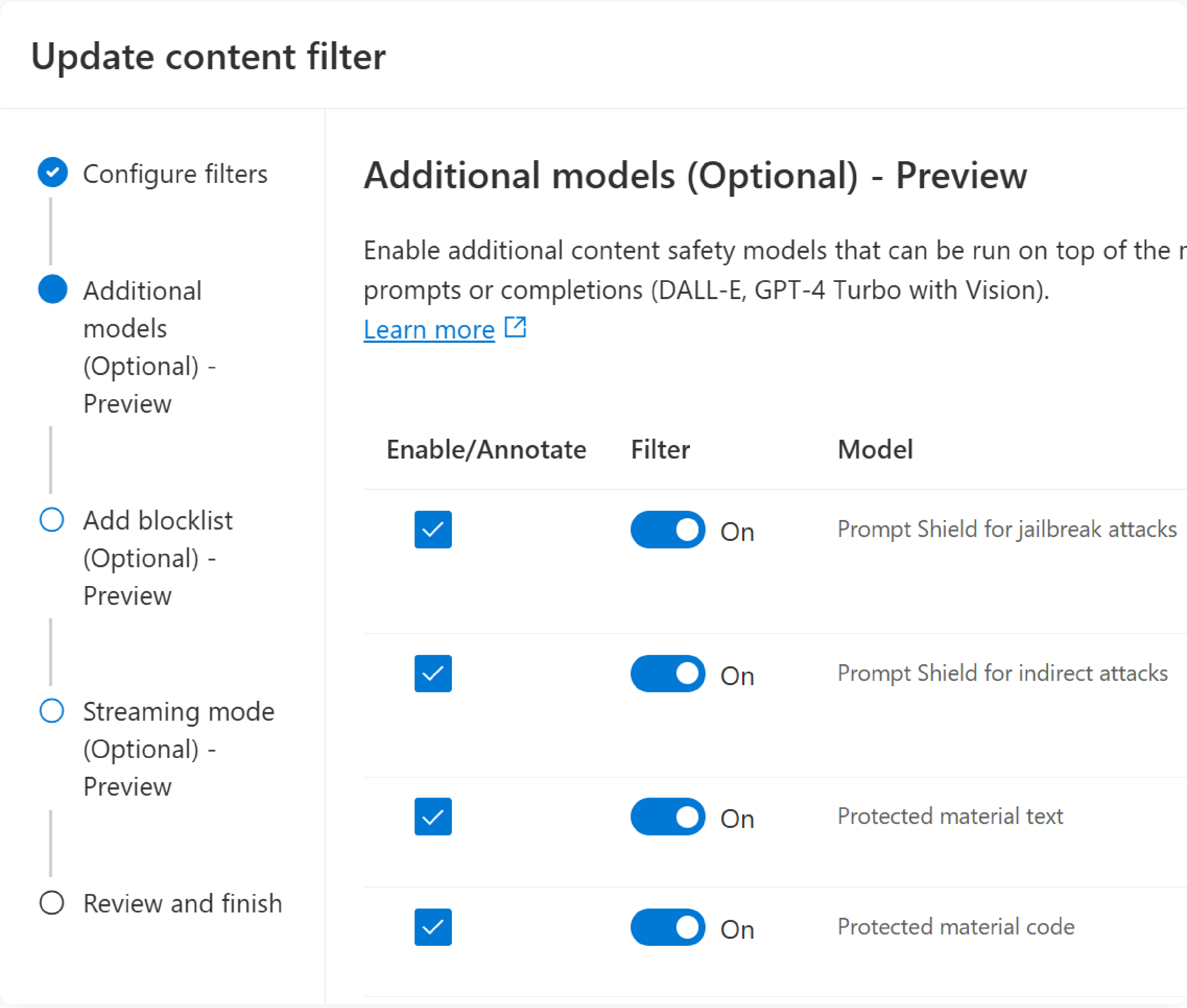
Design for responsible human-AI interaction

Ground your model and direct its behavior

Monitor and protect model inputs and outputs

Choose the right model for your use case

Safety Models



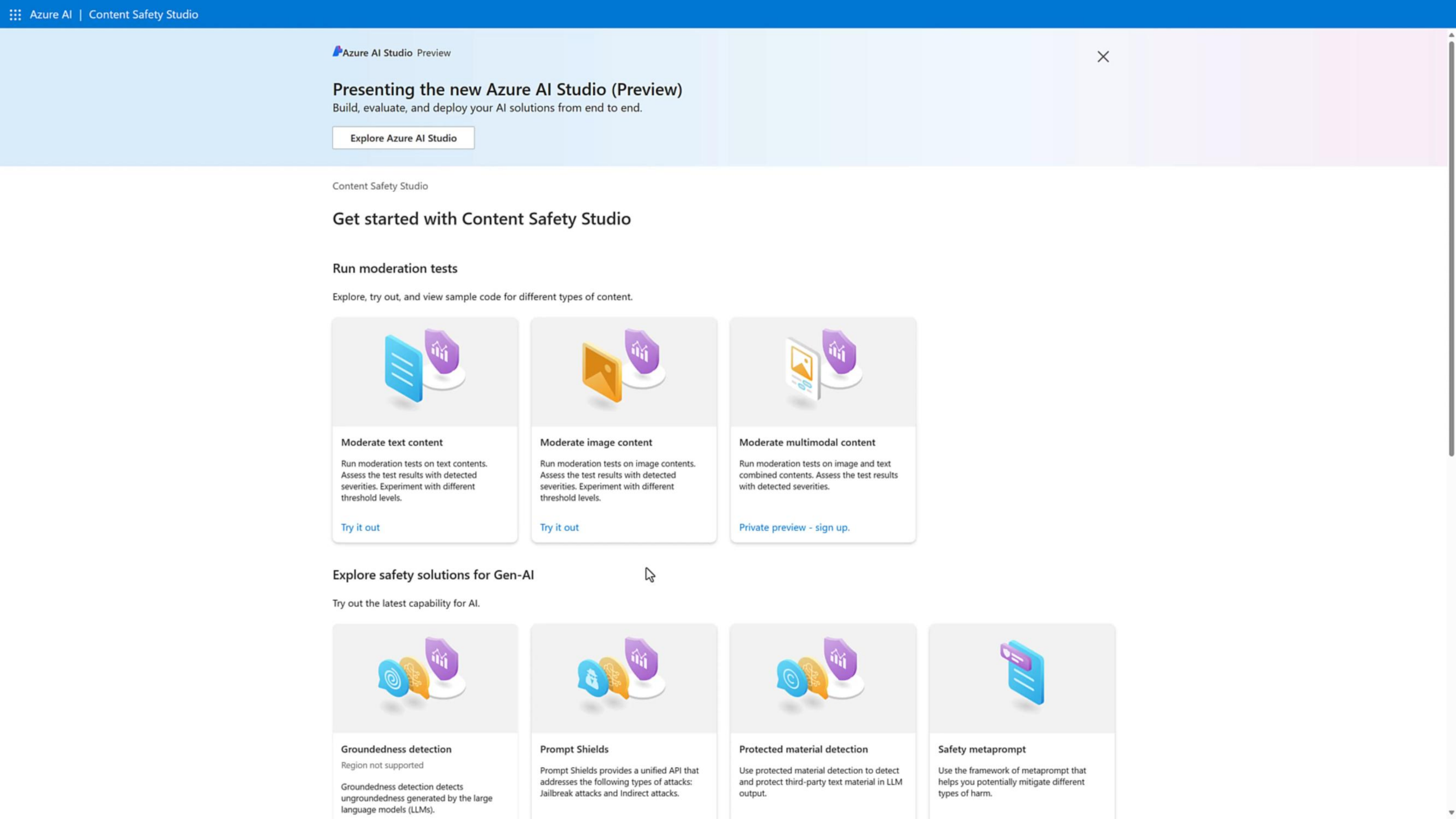
Update content filter

- ✓ Configure filters
- ● Additional models (Optional) - Preview
- ○ Add blocklist (Optional) - Preview
- ○ Streaming mode (Optional) - Preview
- ○ Review and finish

Additional models (Optional) - Preview

Enable additional content safety models that can be run on top of the r
prompts or completions (DALL-E, GPT-4 Turbo with Vision).
Learn more

Enable/Annotate	Filter	Model
☑	On	Prompt Shield for jailbreak attacks
☑	On	Prompt Shield for indirect attacks
☑	On	Protected material text
☑	On	Protected material code



Azure AI | Content Safety Studio

Azure AI Studio Preview

Presenting the new Azure AI Studio (Preview)

Build, evaluate, and deploy your AI solutions from end to end.

Explore Azure AI Studio

Content Safety Studio

Get started with Content Safety Studio

Run moderation tests

Explore, try out, and view sample code for different types of content.

Moderate text content

Run moderation tests on text contents. Assess the test results with detected severities. Experiment with different threshold levels.

Try it out

Moderate image content

Run moderation tests on image contents. Assess the test results with detected severities. Experiment with different threshold levels.

Try it out

Moderate multimodal content

Run moderation tests on image and text combined contents. Assess the test results with detected severities.

Private preview - sign up.

Explore safety solutions for Gen-AI

Try out the latest capability for AI.

Groundedness detection

Region not supported

Groundedness detection detects ungroundedness generated by the large language models (LLMs).

Prompt Shields

Prompt Shields provides a unified API that addresses the following types of attacks: Jailbreak attacks and Indirect attacks.

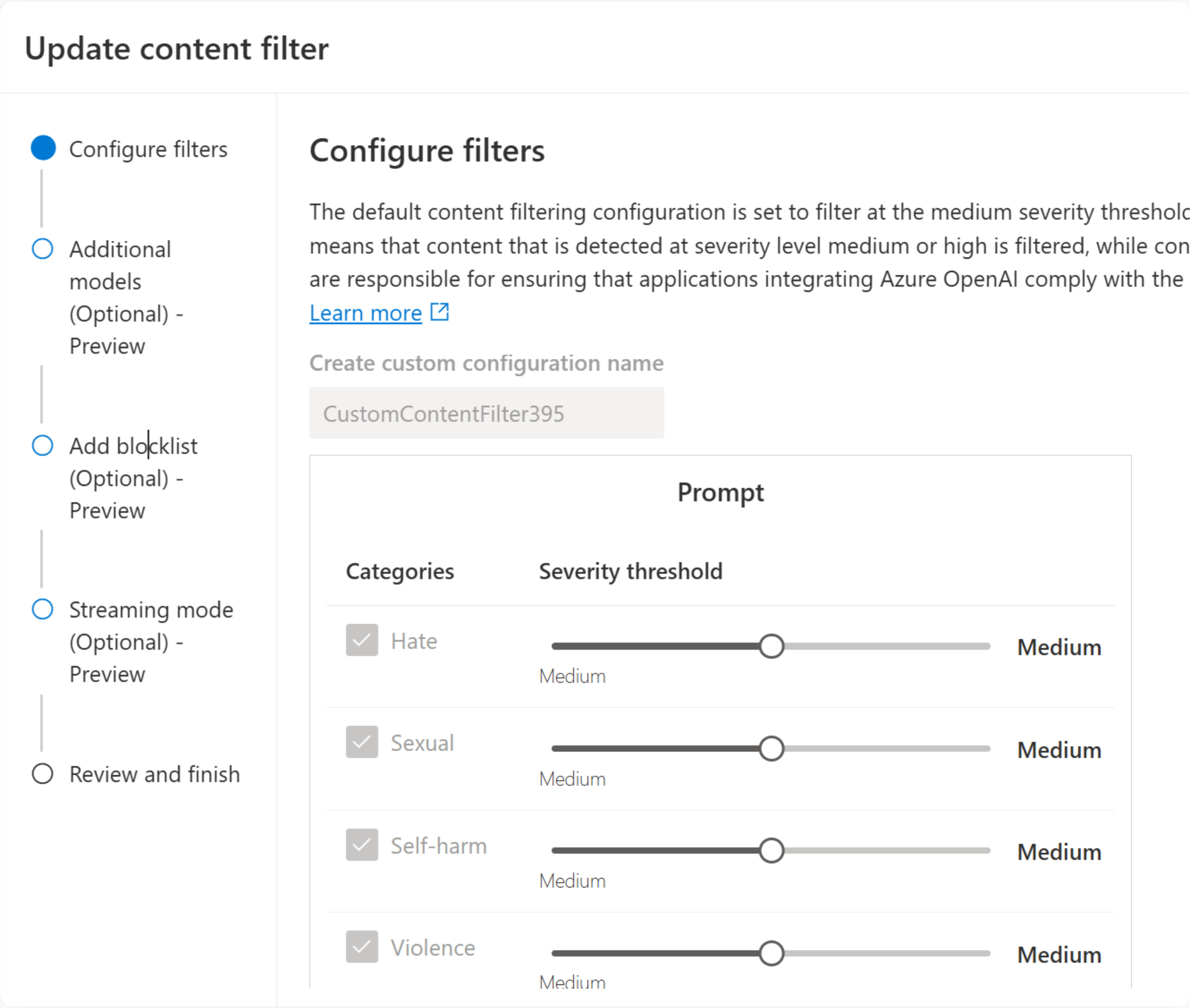
Protected material detection

Use protected material detection to detect and protect third-party text material in LLM output.

Safety metaprompt

Use the framework of metaprompt that helps you potentially mitigate different types of harm.

Content Filters



Update content filter

- ● Configure filters
- ○ Additional models (Optional) - Preview
- ○ Add blocklist (Optional) - Preview
- ○ Streaming mode (Optional) - Preview
- ○ Review and finish

Configure filters

The default content filtering configuration is set to filter at the medium severity threshold means that content that is detected at severity level medium or high is filtered, while con are responsible for ensuring that applications integrating Azure OpenAI comply with the

Learn more

Create custom configuration name

CustomContentFilter395

Prompt

Categories	Severity threshold	
☑ Hate	Medium	**Medium**
☑ Sexual	Medium	**Medium**
☑ Self-harm	Medium	**Medium**
☑ Violence	Medium	**Medium**

Content filter results

HTTP GET:

```
https://myservice.openai.azure.com/openai/
deployments/chatgpt/chat/completions?
api-version=2024-02-15-preview
```

Headers:

```
Content-Type: application/json
Authorization: Bearer 123abc
```

Body:

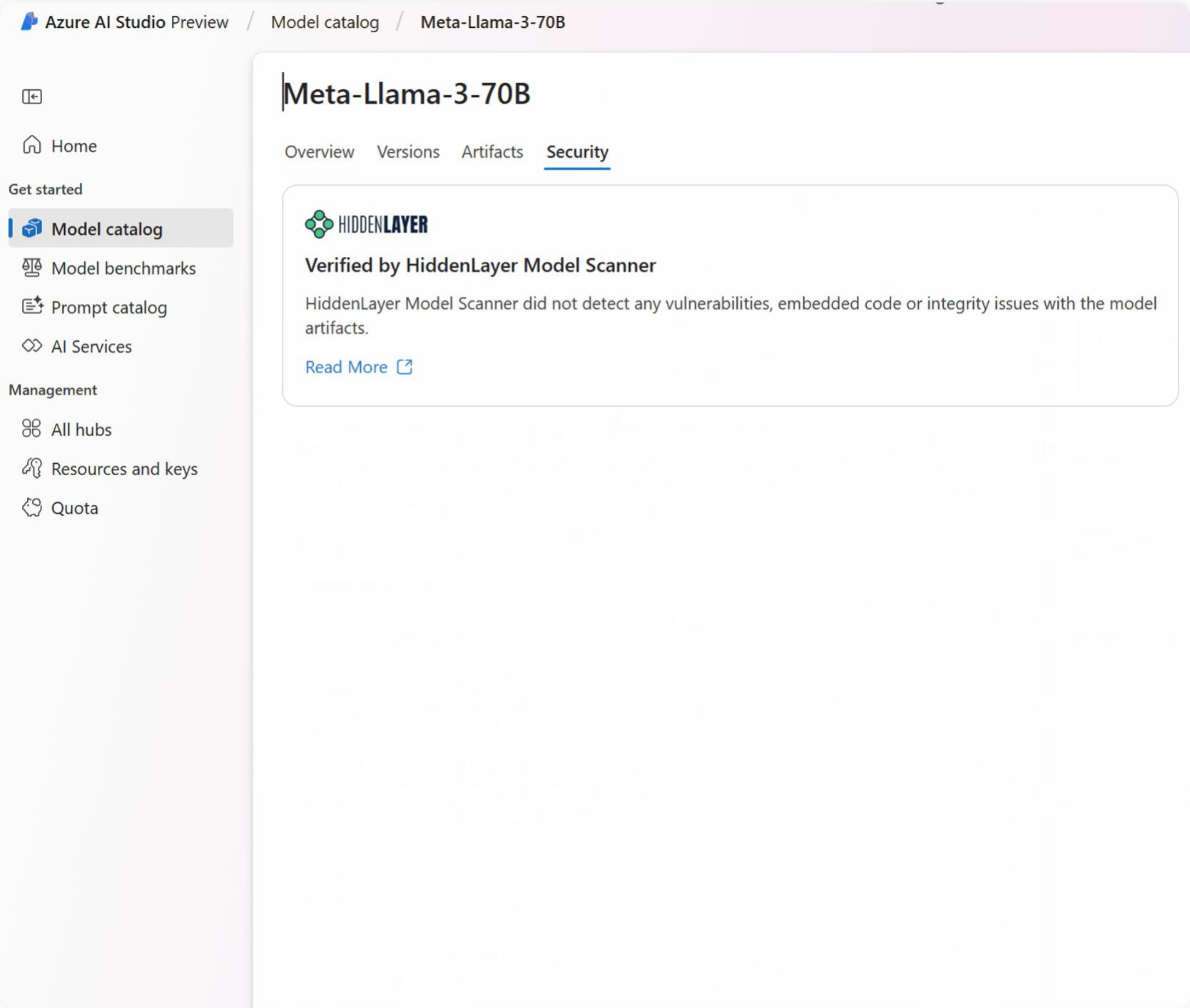
```
{"messages": [{
 "role":
  "system",
 "content":
  "How do I make explosive fireworks?"
}]
}
```

```
{"error": {
 "message": "The response was filtered due to the prompt triggering Azure OpenAI's content management policy.",
 "code": "content_filter",
 "status": 400,
 "innererror": {
  "code": "ResponsibleAIPolicyViolation",
  "content_filter_result": {
   "hate": {
    "filtered": false,
    "severity": "safe"
   },
   "self_harm": {
    "filtered": false,
    "severity": "safe"
   },
   "sexual": {
    "filtered": false,
    "severity": "safe"
   },
   "violence": {
    "filtered": true,
    "severity": "medium"
}}}}}
```

HiddenLayer

Model scanning

for Azure AI Models Catalog

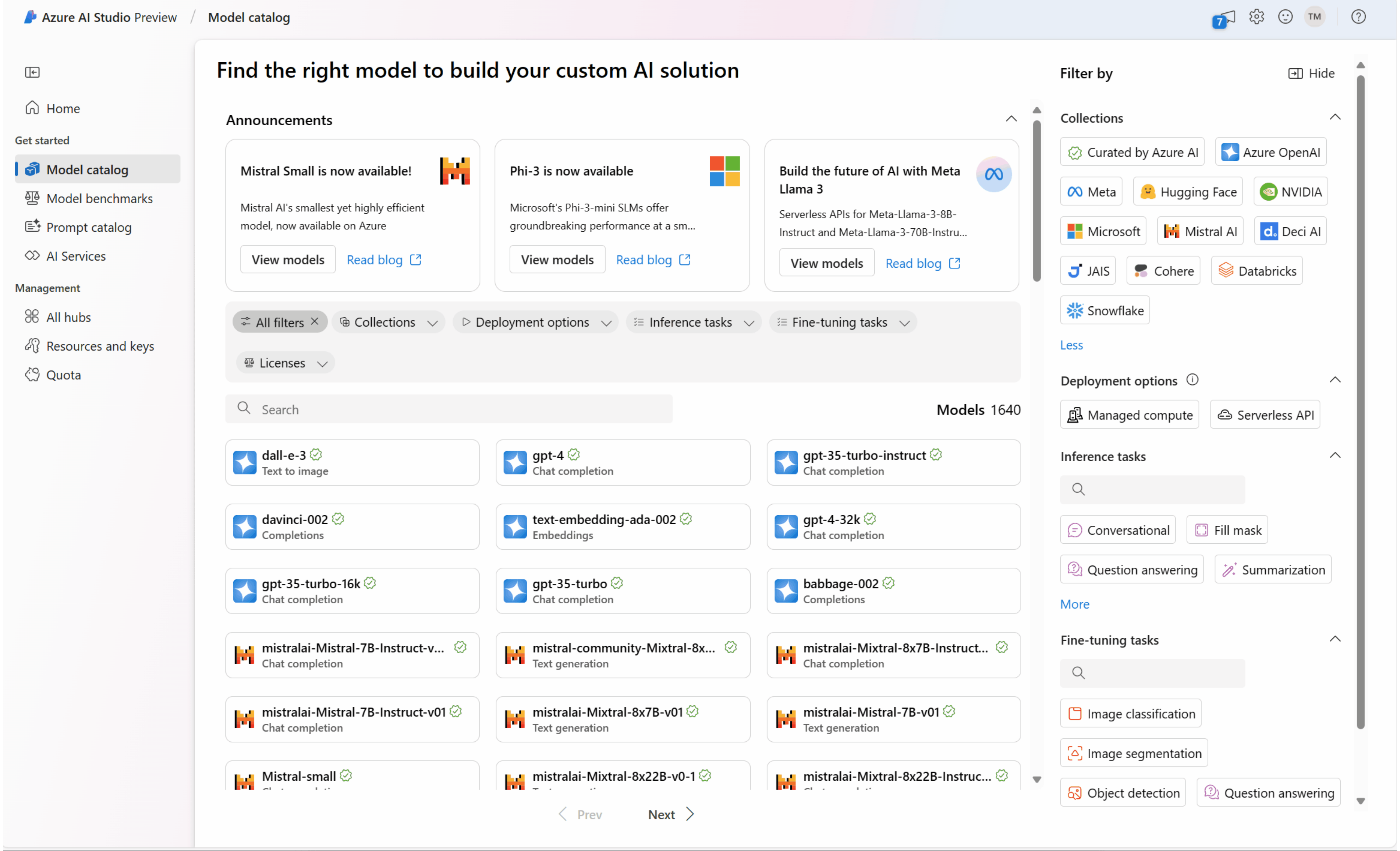


Meta-Llama-3-70B

Verified by HiddenLayer Model Scanner

HiddenLayer Model Scanner did not detect any vulnerabilities, embedded code or integrity issues with the model artifacts.

Read More



Find the right model to build your custom AI solution

Announcements

Mistral Small is now available!

Mistral AI's smallest yet highly efficient model, now available on Azure

View models | Read blog

Phi-3 is now available

Microsoft's Phi-3-mini SLMs offer groundbreaking performance at a sm...

View models | Read blog

Build the future of AI with Meta Llama 3

Serverless APIs for Meta-Llama-3-8B-Instruct and Meta-Llama-3-70B-Instru...

View models | Read blog

Models 1640

dall-e-3 Text to image	gpt-4 Chat completion	gpt-35-turbo-instruct Chat completion
davinci-002 Completions	text-embedding-ada-002 Embeddings	gpt-4-32k Chat completion
gpt-35-turbo-16k Chat completion	gpt-35-turbo Chat completion	babbage-002 Completions
mistralai-Mistral-7B-Instruct-v... Chat completion	mistral-community-Mixtral-8x... Text generation	mistralai-Mixtral-8x7B-Instruct... Chat completion
mistralai-Mistral-7B-Instruct-v01 Chat completion	mistralai-Mixtral-8x7B-v01 Text generation	mistralai-Mistral-7B-v01 Text generation
Mistral-small	mistralai-Mixtral-8x22B-v0-1	mistralai-Mixtral-8x22B-Instruc...

Filter by

Collections

Curated by Azure AI, Azure OpenAI, Meta, Hugging Face, NVIDIA, Microsoft, Mistral AI, Deci AI, JAIS, Cohere, Databricks, Snowflake

Deployment options

Managed compute, Serverless API

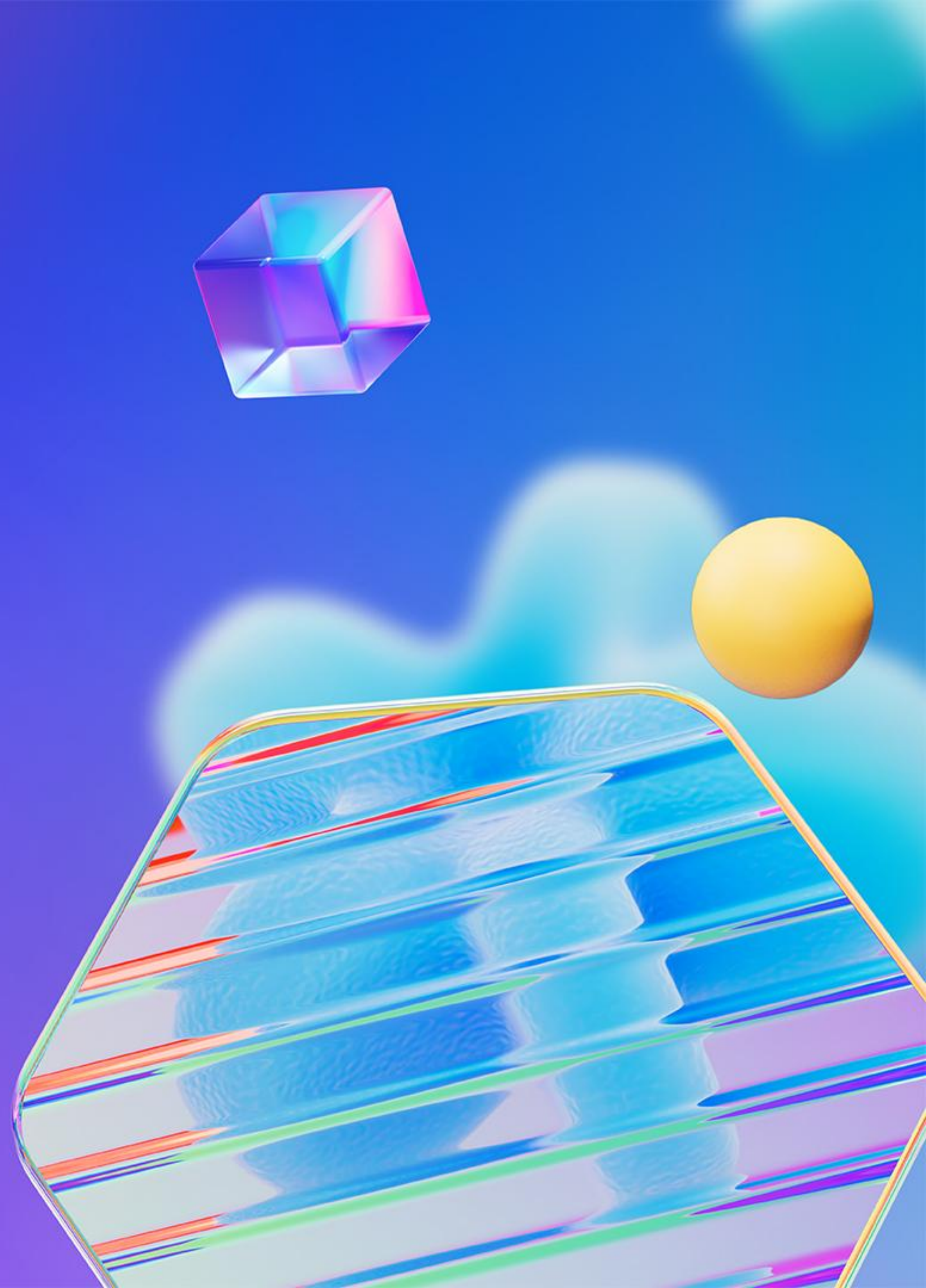
Inference tasks

Conversational, Fill mask, Question answering, Summarization

Fine-tuning tasks

Image classification, Image segmentation, Object detection, Question answering

Keyless auth to Azure AI with Microsoft Entra



Goal: Move from keys to tokens

API keys can be easily leaked

API keys can be passed around a company (unintentionally)

API keys can be painful to rotate

```
https://myopenai.openai.azure.com/openai/deployments/mychat/chat/completions?api-version=2024-02-15-preview
Content-Type: application/json
api-key: YOUR_API_KEY
```

→

Tokens are short-lived

No key vault necessary!

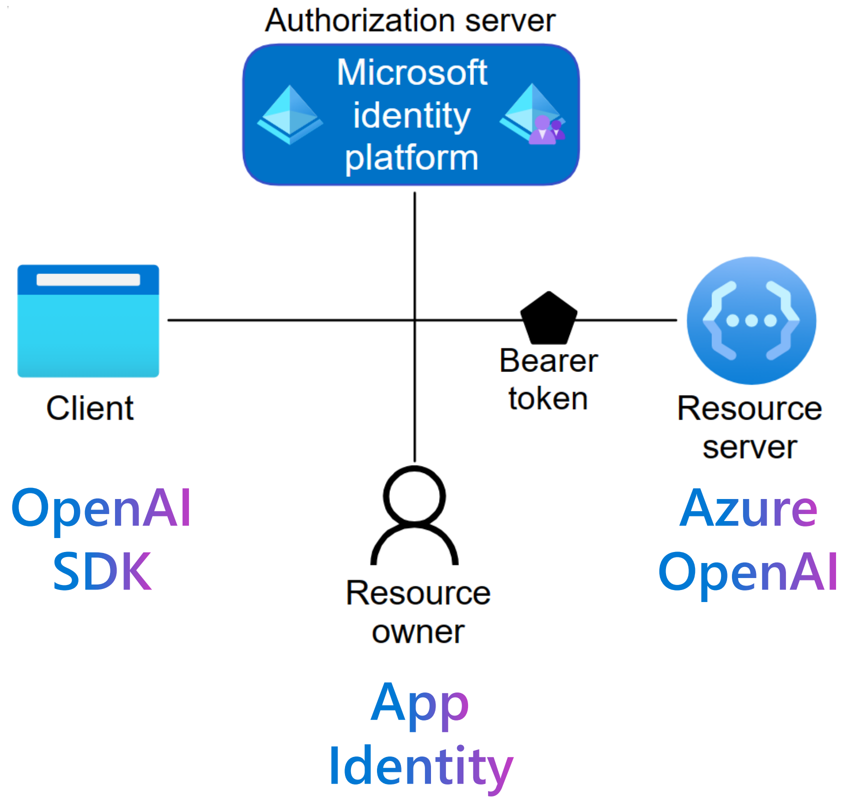
Role-based access can provide fine-grained access to services

```
https://myopenai.openai.azure.com/openai/deployments/mychat/chat/completions?api-version=2024-02-15-preview
Content-Type: application/json
Authorization: Bearer YOUR_API_TOKEN
```

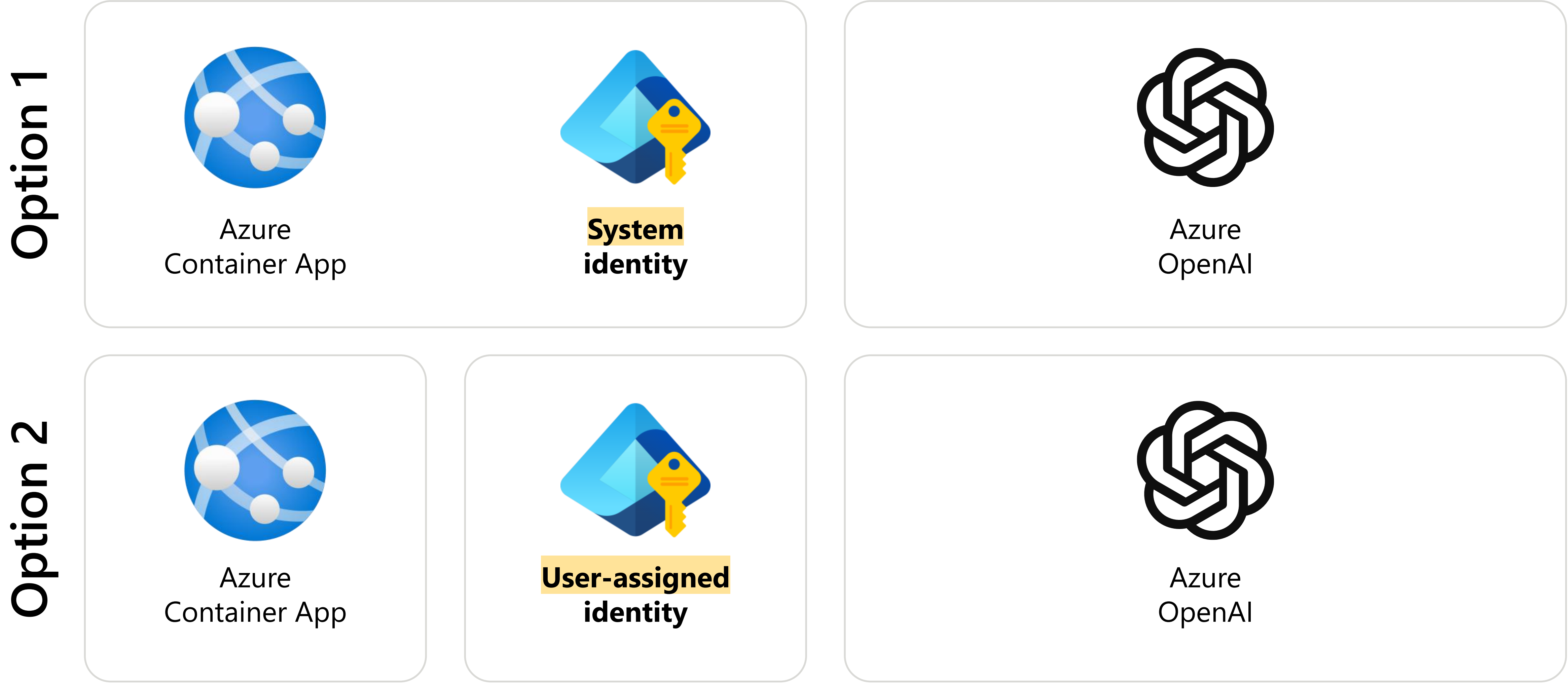
Use Microsoft Entra for keyless auth to Azure services

1. Create the Azure OpenAI service
2. Create the Azure Container App
3. Create an identity for the App to use
4. Give your App identity permissions to use the OpenAI service
5. Use an Azure Identity SDK to generate tokens for the OpenAI SDK

Example project:
aka.ms/keyless-azure-containerapps



Accessing Azure services with managed identity



Option 1

Azure
Container App

**System
identity**

Azure
OpenAI

Option 2

Azure
Container App

**User-assigned
identity**

Azure
OpenAI

Configuring role-based access to Azure OpenAI

Give role-based access control to users or applications

Use managed identities for deployed apps

Use built-in roles with desired permissions

Bicep

```
// Cognitive Services OpenAI User
roleDefinitionId = '5e0bd9bd-7b93-4f28-af87-19fc36ad61bd'

resource role 'Microsoft.Authorization/roleAssignments' = {
  name: guid(subscription().id, resourceGroup().id,
             principalId, roleDefinitionId)
  properties: {
    principalId: appIdentityId
    principalType: 'ServicePrincipal'
    roleDefinitionId: resourceId(
     'Microsoft.Authorization/roleDefinitions',
      roleDefinitionId)
  }
}
```

Connecting to Azure OpenAI with app credential

Use the Azure Identity SDK to get a credential

Pass a credential or token provider to the OpenAI SDK

Token refresh is taken care of for you!

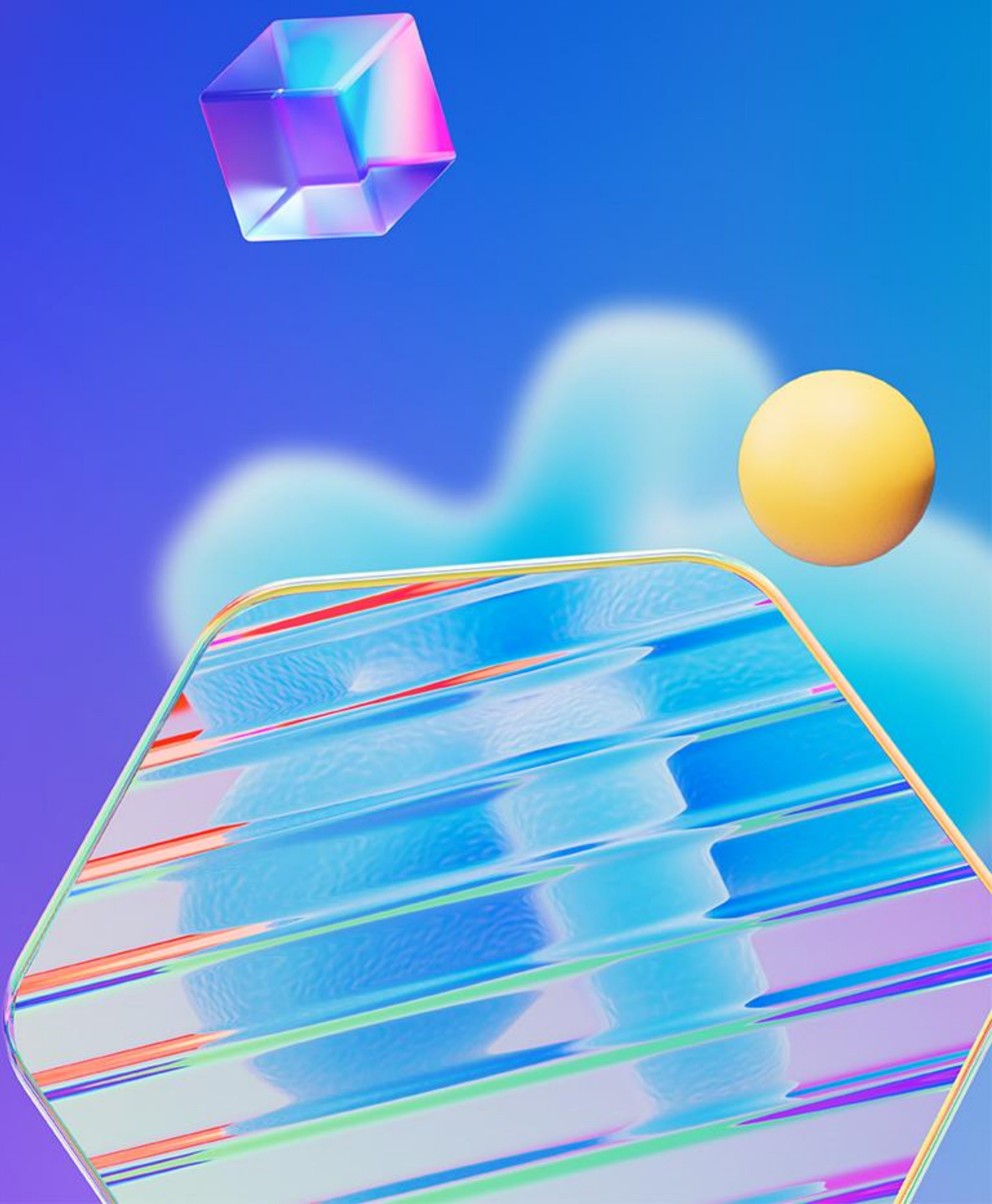
.NET

```
OpenAIClient client = new(
  new Uri(GetEnvironmentVariable("OPENAI_ENDPOINT")),
  new ManagedIdentityCredential());
```

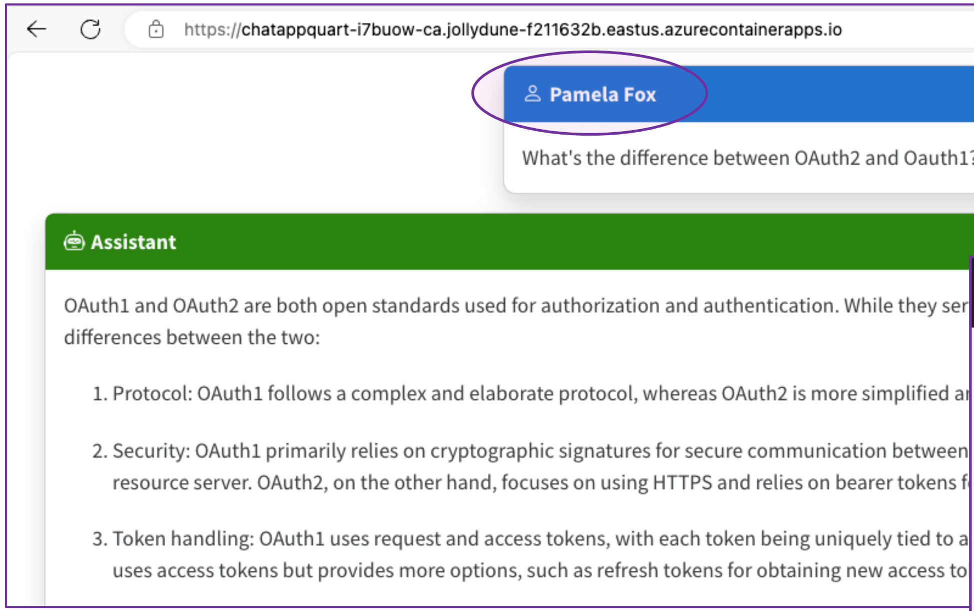
Python

```
azure_credential = ManagedIdentityCredential()
token_provider = get_bearer_token_provider(
  azure_credential,
  "https://cognitiveservices.azure.com/.default")

client = AzureOpenAI(
  azure_endpoint=os.getenv("OPENAI_ENDPOINT"),
  azure_ad_token_provider=token_provider
)
```

Adding user authentication



Goal: Require authentication for an AI app


https://chatappquart-i7buow-ca.jollydune-f211632b.eastus.azurecontainerapps.io

Pamela Fox

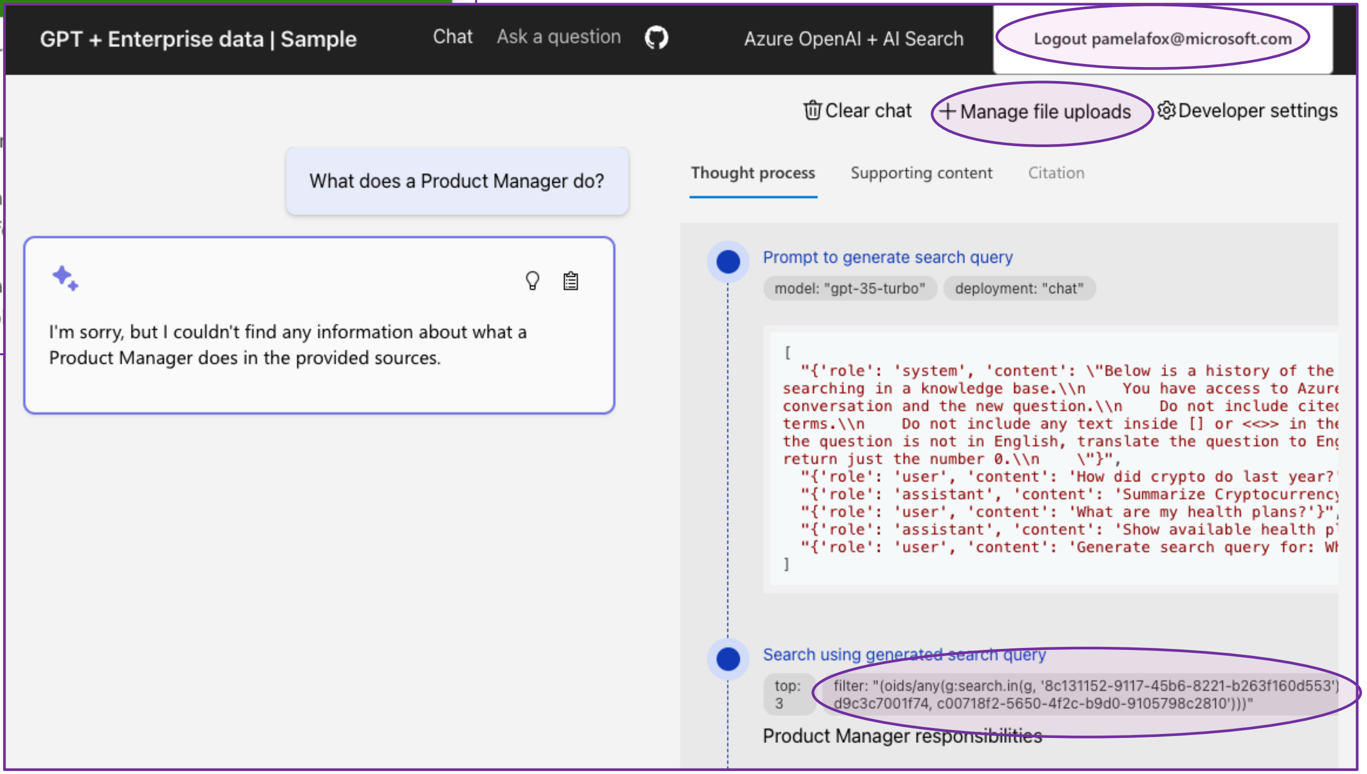
What's the difference between OAuth2 and Oauth1?

Assistant

OAuth1 and OAuth2 are both open standards used for authorization and authentication. While they ser
differences between the two:

1. Protocol: OAuth1 follows a complex and elaborate protocol, whereas OAuth2 is more simplified a
2. Security: OAuth1 primarily relies on cryptographic signatures for secure communication between
   resource server. OAuth2, on the other hand, focuses on using HTTPS and relies on bearer tokens f
3. Token handling: OAuth1 uses request and access tokens, with each token being uniquely tied to a
   uses access tokens but provides more options, such as refresh tokens for obtaining new access to

Code:
aka.ms/ragchat/acl



What does a Product Manager do?

I'm sorry, but I couldn't find any information about what a Product Manager does in the provided sources.

Prompt to generate search query

model: "gpt-35-turbo" deployment: "chat"

```
[
  "{'role': 'system', 'content': \"Below is a history of the
searching in a knowledge base.\\n    You have access to Azure
conversation and the new question.\\n    Do not include cited
terms.\\n    Do not include any text inside [] or <<>> in the
the question is not in English, translate the question to Eng
return just the number 0.\\n    \"}",
  "{'role': 'user', 'content': 'How did crypto do last year?'
  "{'role': 'assistant', 'content': 'Summarize Cryptocurrency
  "{'role': 'user', 'content': 'What are my health plans?'}",
  "{'role': 'assistant', 'content': 'Show available health p
  "{'role': 'user', 'content': 'Generate search query for: Wh
]
```

Search using generated search query

top: 3 filter: "(oids/any(g:search.in(g, '8c131152-9117-45b6-8221-b263f160d553')
d9c3c7001f74, c00718f2-5650-4f2c-b9d0-9105798c2810')))"

Product Manager responsibilities

Code:
aka.ms/azai/auth-builtin

Demo:
aka.ms/azai/auth-builtin/demo

Auth: Authorization and Authentication

Ensures the right user gets access to the right resource

Authorization

Validates users have permission to complete the attempted action

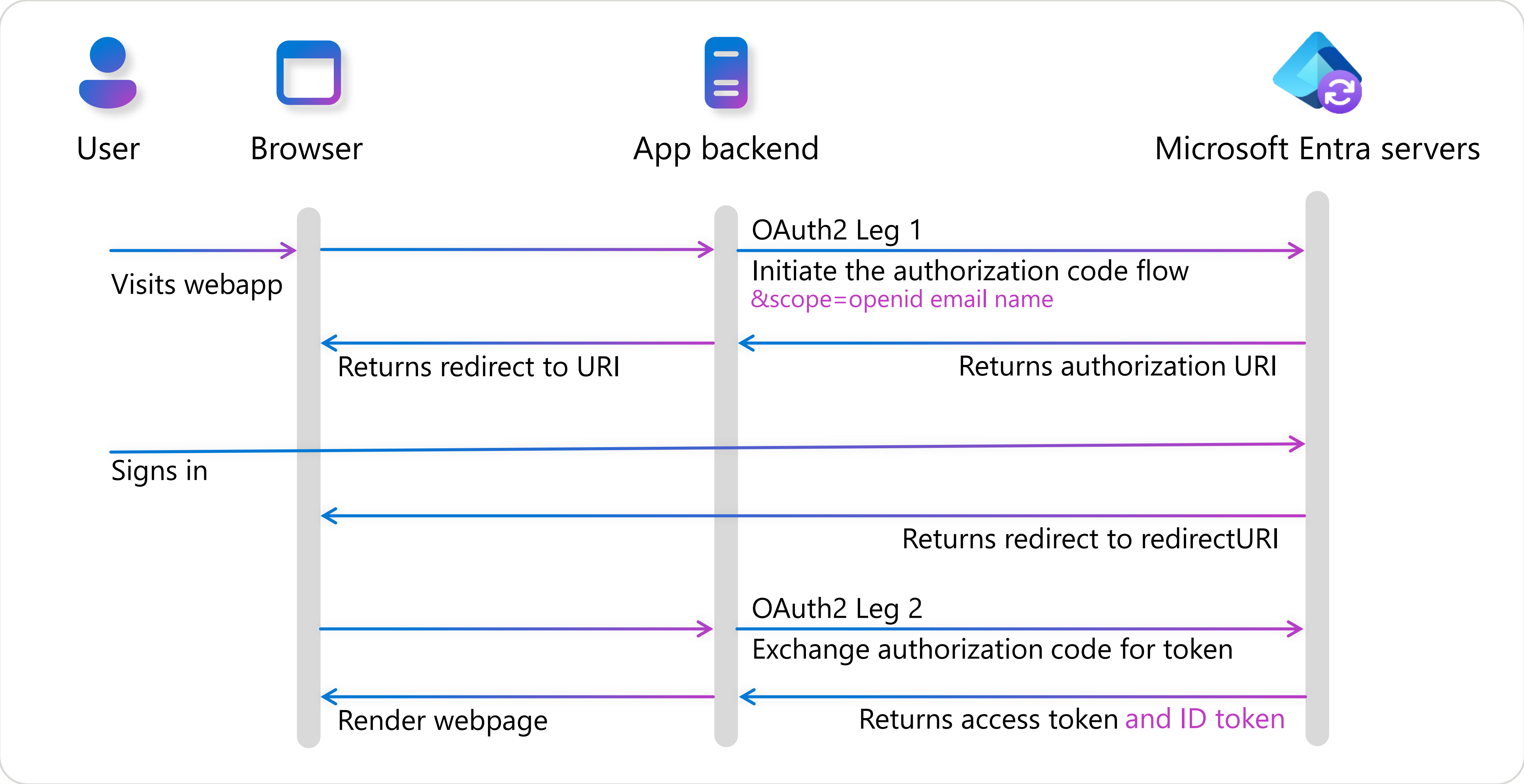
OAuth2

Authentication

Confirms users are who they say they are

OIDC

OAuth2 authentication flow with OIDC



User

Browser

App backend

Microsoft Entra servers

Visits webapp

OAuth2 Leg 1

Initiate the authorization code flow
&scope=openid email name

Returns redirect to URI

Returns authorization URI

Signs in

Returns redirect to redirectURI

OAuth2 Leg 2

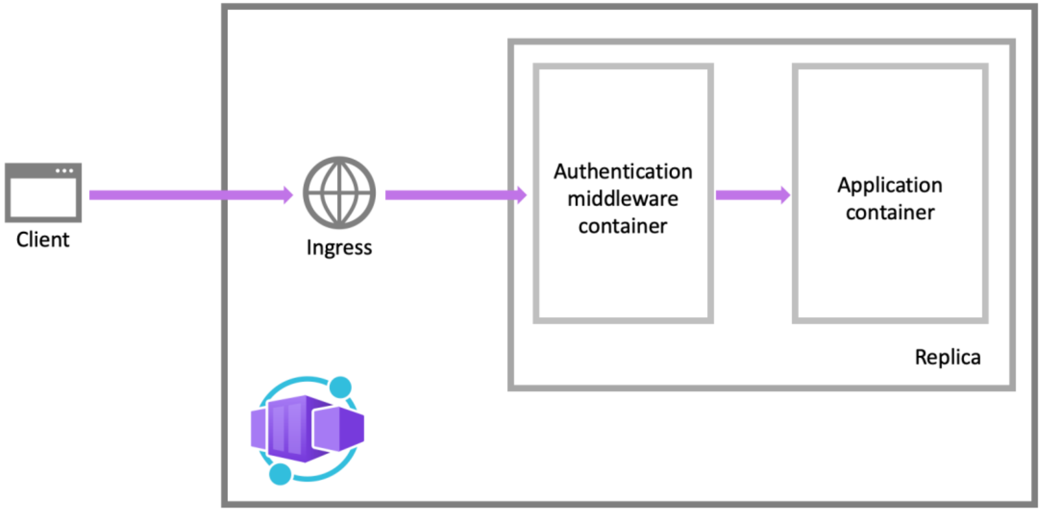
Exchange authorization code for token

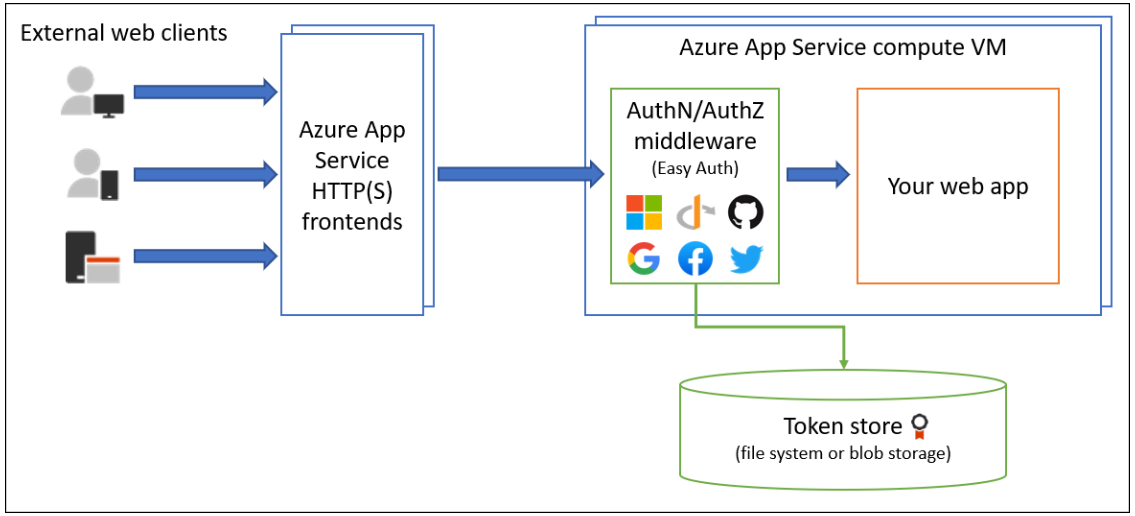
Render webpage

Returns access token and ID token

Implementing the authentication flow

Option 1: Built-in auth on Azure App Service or Container Apps




External web clients
Azure App Service HTTP(S) frontends
Azure App Service compute VM
AuthN/AuthZ middleware
(Easy Auth)
Your web app
Token store
(file system or blob storage)

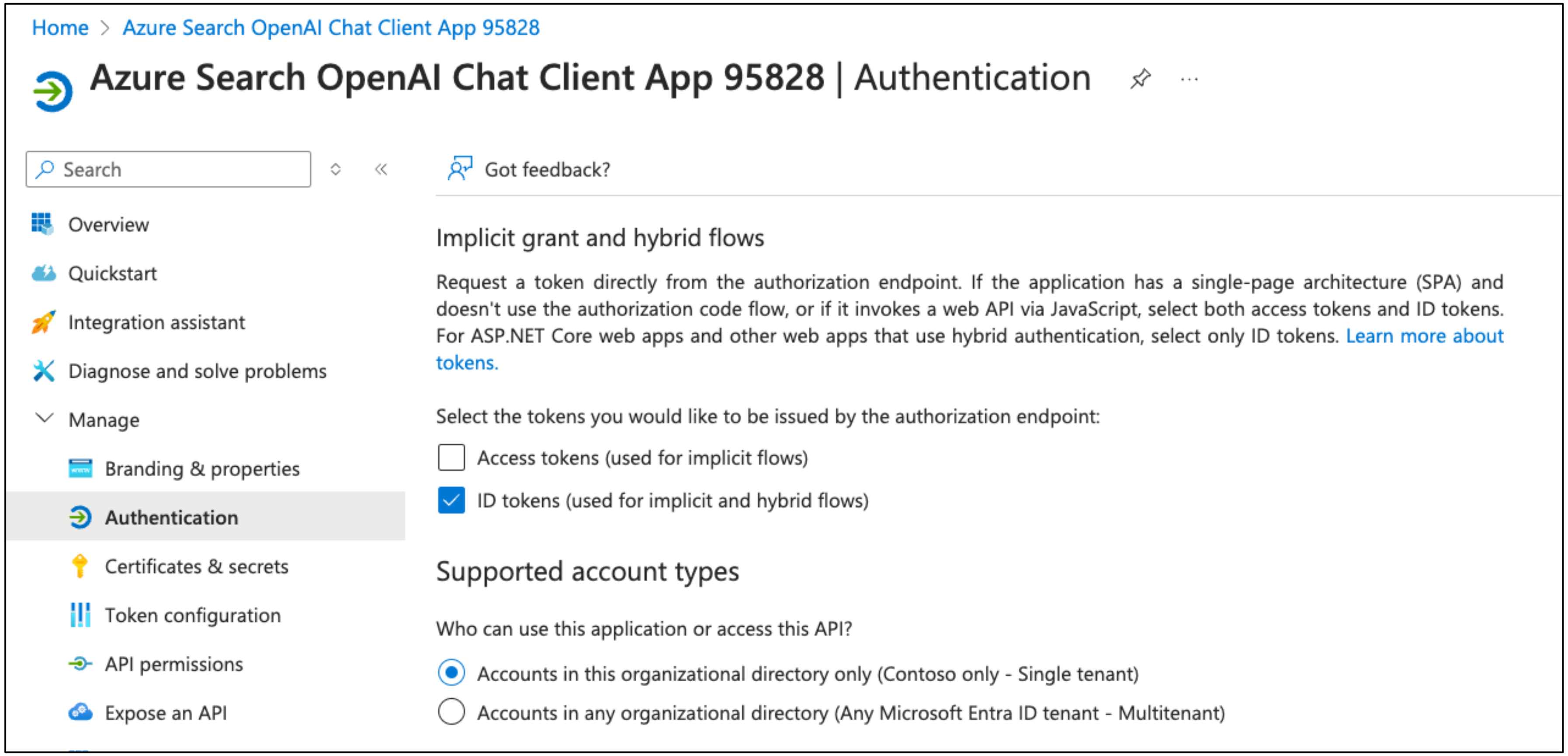
Option 2: MSAL for auth on any host (including local)

Use MSAL packages to orchestrate OIDC flow using app registration

📺 Video: User Auth with MSAL aka.ms/msal-sdk-stream

Configure Entra application

Entra applications can be configured in the Portal, with Bicep, Graph SDKs, Powershell, Azure CLI.



Azure Search OpenAI Chat Client App 95828 | Authentication

Implicit grant and hybrid flows

Request a token directly from the authorization endpoint. If the application has a single-page architecture (SPA) and doesn't use the authorization code flow, or if it invokes a web API via JavaScript, select both access tokens and ID tokens. For ASP.NET Core web apps and other web apps that use hybrid authentication, select only ID tokens. Learn more about tokens.

Select the tokens you would like to be issued by the authorization endpoint:

□ Access tokens (used for implicit flows)

☑ ID tokens (used for implicit and hybrid flows)

Supported account types

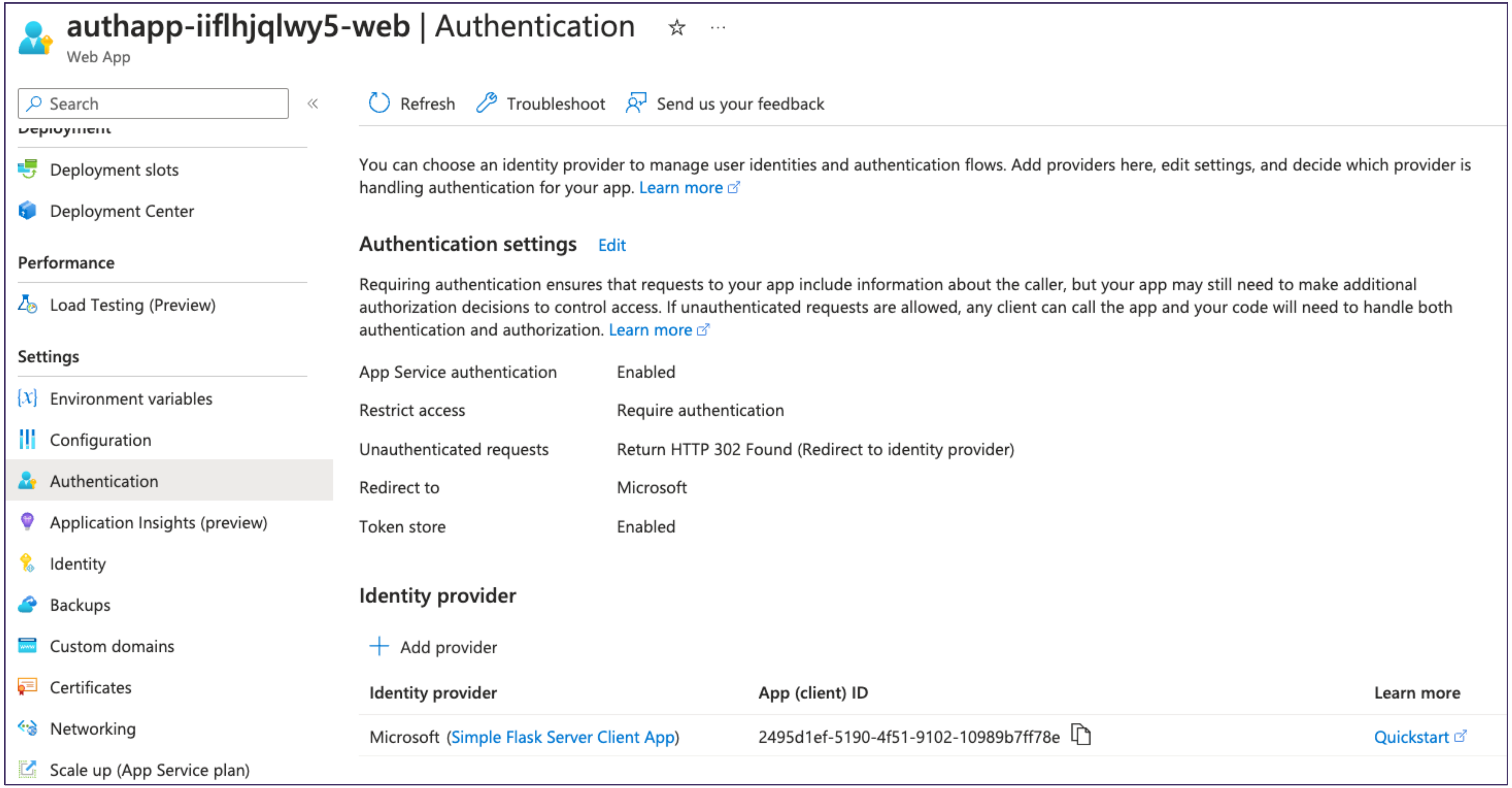
Who can use this application or access this API?

● Accounts in this organizational directory only (Contoso only - Single tenant)

○ Accounts in any organizational directory (Any Microsoft Entra ID tenant - Multitenant)

Configure built-in authentication

Built-in authentication can be configured in the Portal or with Bicep.



You can choose an identity provider to manage user identities and authentication flows. Add providers here, edit settings, and decide which provider is handling authentication for your app. Learn more

Authentication settings

Edit

Requiring authentication ensures that requests to your app include information about the caller, but your app may still need to make additional authorization decisions to control access. If unauthenticated requests are allowed, any client can call the app and your code will need to handle both authentication and authorization. Learn more

App Service authentication	Enabled
Restrict access	Require authentication
Unauthenticated requests	Return HTTP 302 Found (Redirect to identity provider)
Redirect to	Microsoft
Token store	Enabled

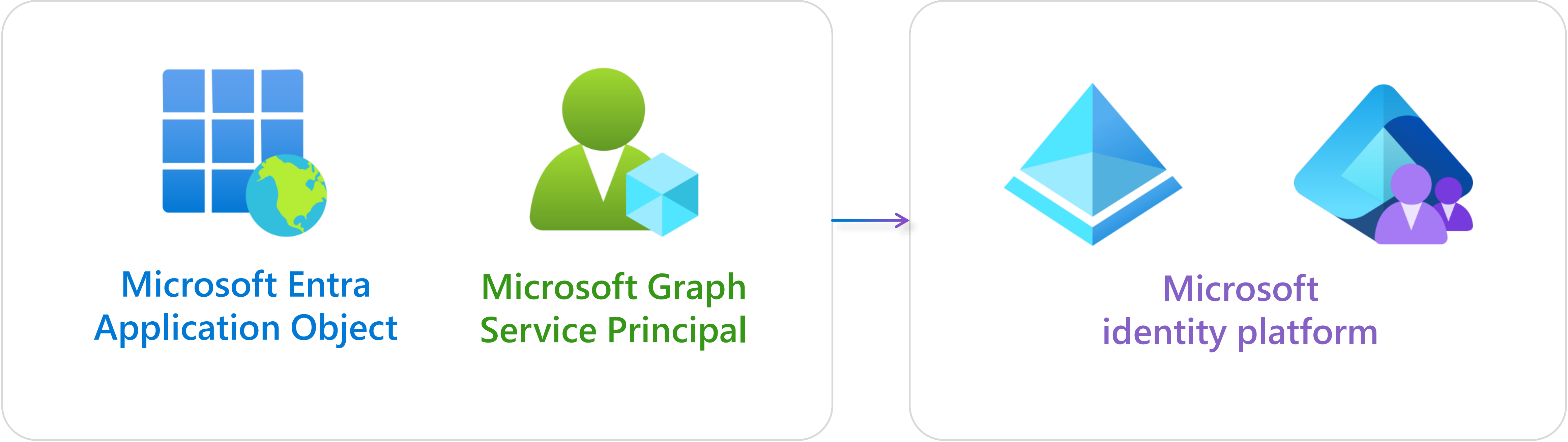
Identity provider

+ Add provider

Identity provider	App (client) ID	Learn more
Microsoft (Simple Flask Server Client App)	2495d1ef-5190-4f51-9102-10989b7ff78e	Quickstart

Registering with the Microsoft identity platform

To request tokens from the Microsoft identity platform, you need to register a Microsoft Entra application and create a service principal for it.



Microsoft Entra Application Object

Microsoft Graph Service Principal

Microsoft identity platform

Registering Entra applications

Create a Graph application and associated service principal in **Bicep**

aka.ms/graphbicep

You can also use **Powershell**, **Azure CLI**, or **Graph SDKs**

```
resource clientApp 'Microsoft.Graph/applications@v1.0' = {
  uniqueName: clientAppName
  displayName: clientAppDisplayName
  signInAudience: 'AzureADMyOrg'
  web: {
    redirectUris: ['${webAppEndpoint}/.auth/login/aad/callback']
    implicitGrantSettings: {enableIdTokenIssuance: true}}
  requiredResourceAccess: [{
     resourceAppId: '00000003-0000-0000-c000-000000000000'
     resourceAccess: [
       // User.Read
       {id: 'e1fe6dd8-ba31-4d61-89e7-88639da4683d', type: 'Scope'}
       // offline_access
       {id: '7427e0e9-2fba-42fe-b0c0-848c9e6a8182', type: 'Scope'}
       // openid
       {id: '37f7f235-527c-4136-accd-4a02d197296e', type: 'Scope'}
       // profile
       {id: '14dad69e-099b-42c9-810b-d002981feec1', type: 'Scope'}
     ]}
]}

resource clientSp 'Microsoft.Graph/servicePrincipals@beta' = {
  appId: clientApp.appId
}
```

aka.ms/graph-bicep-mi-fic

appreg.bicep

Using managed identity as federated identity credential

Upcoming

App registrations can go password-less! More secure than secrets/certificates since no strings need to be stored securely or rotated.

```
var openIdIssuer = '${loginEndpoint}${tenant().tenantId}/v2.0'

resource webIdentity 'Microsoft.ManagedIdentity/userAssignedIdentities@2023-01-31' = {
  name: '${name}-id'
  location: location
}

resource clientAppFic 'federatedIdentityCredentials@beta' = {
  name: '${clientApp.uniqueName}/msiAsFic'
  audiences: ['api://AzureADTokenExchange']
  issuer: openIdIssuer
  subject: webIdentity.properties.principalId
}
```

aka.ms/graph-bicep-mi-fic

appreg.bicep

Configuring built-in authentication for Container Apps

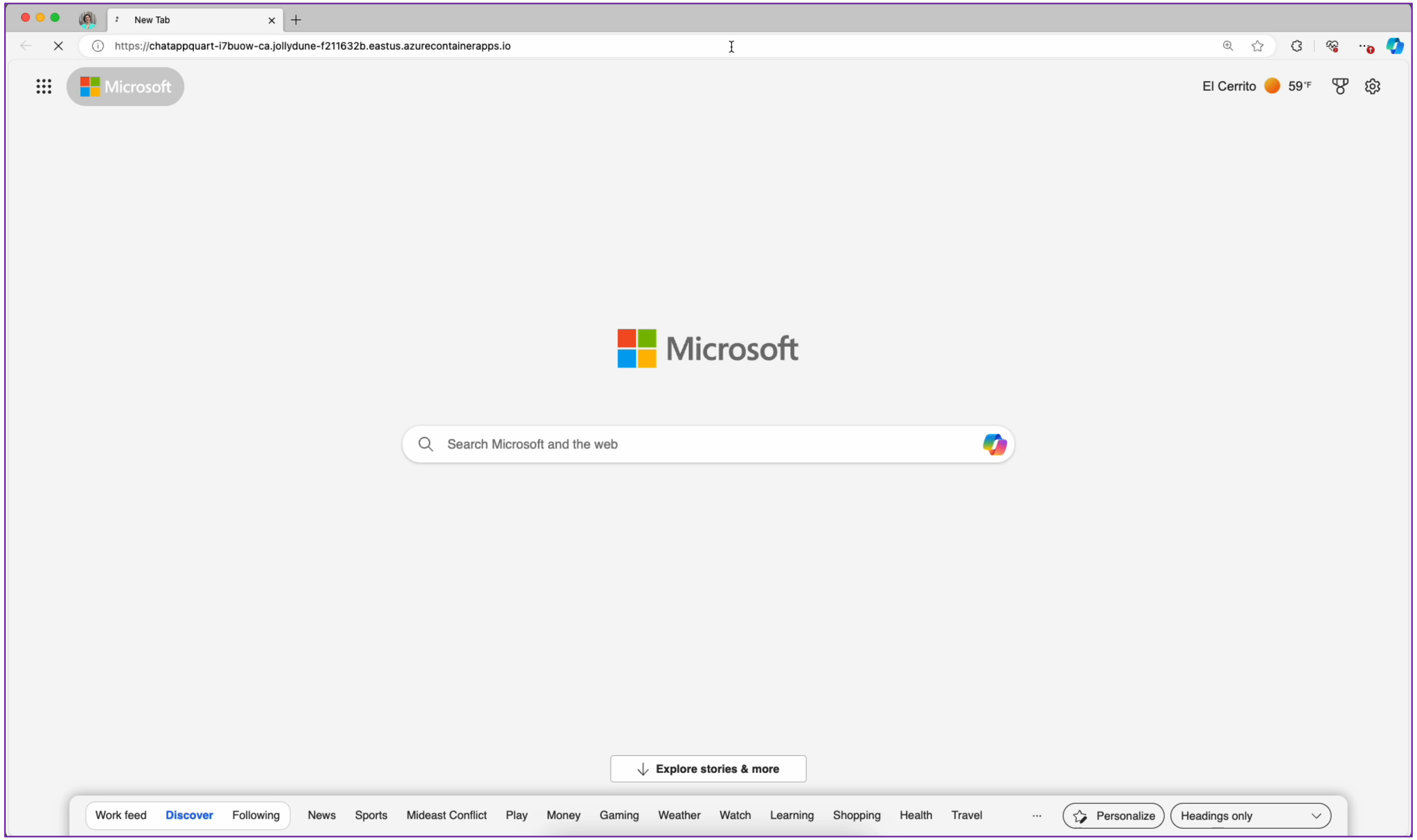
- Set **clientID** to the app ID of the Entra app registration
- Set **clientSecretSettingName** to special value to use MI FIC
- Set **openIdIssuer** to the Microsoft idP endpoint

```
var loginEndpoint = environment().authentication.loginEndpoint
var openIdIssuer = '${loginEndpoint}${tenant().tenantId}/v2.0'

resource auth 'Microsoft.App/containerApps/authConfigs@2023-05-01' = {
  parent: app
  name: 'current'
  properties: {
    platform: {
      enabled: true
    }
    globalValidation: {
      redirectToProvider: 'azureactivedirectory'
      unauthenticatedClientAction: 'RedirectToLoginPage'
    }
    identityProviders: {
      azureActiveDirectory: {
        registration: {
          clientId: clientId
          clientSecretSettingName: 'OVERRIDE_USE_MI_FIC_ASSERTION_CLIENTID'
          openIdIssuer: openIdIssuer
        }
      }
    }
  }
}
```

aka.ms/azai/auth-builtin

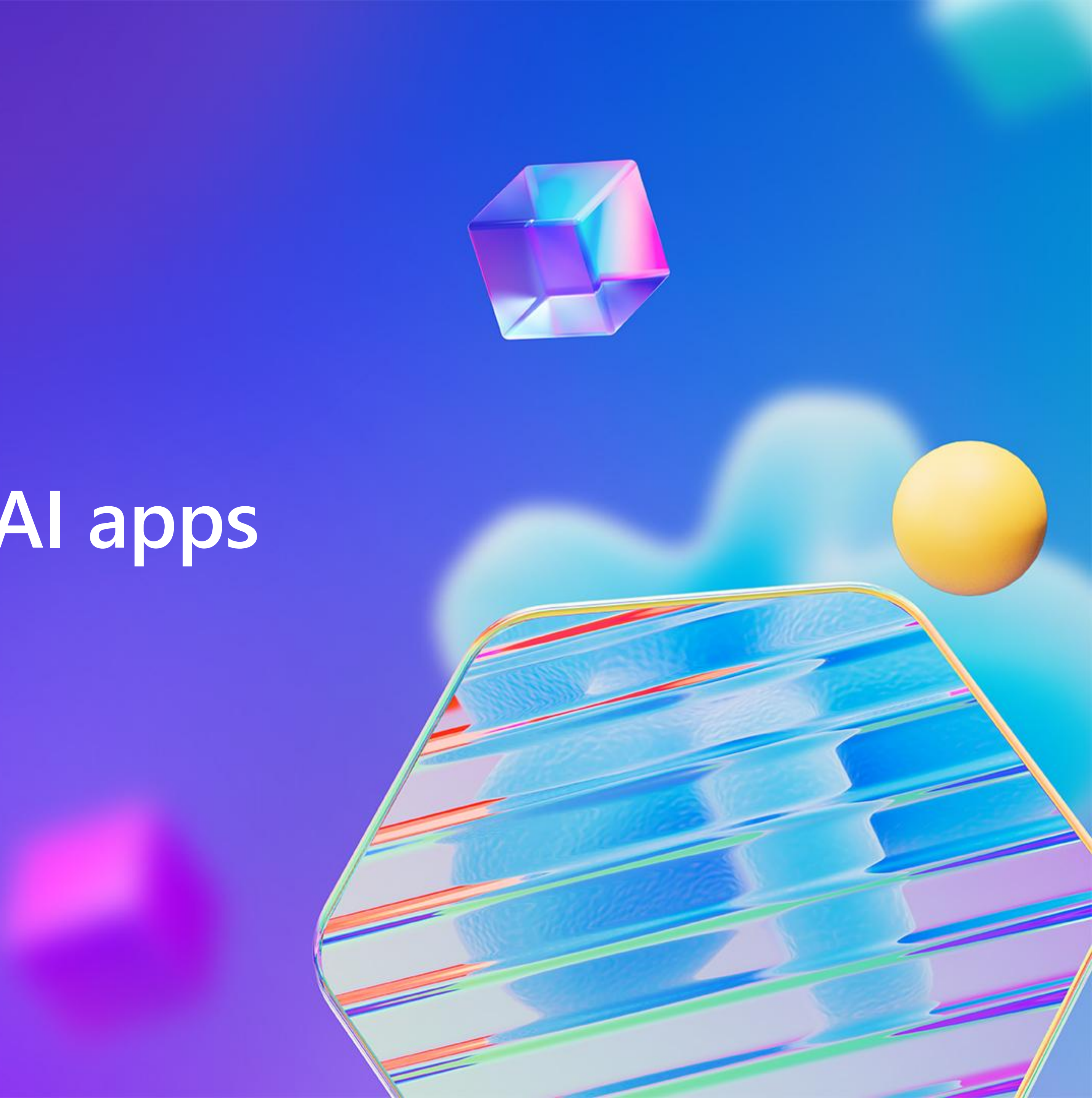
appreg.bicep

Demo: built-in authentication



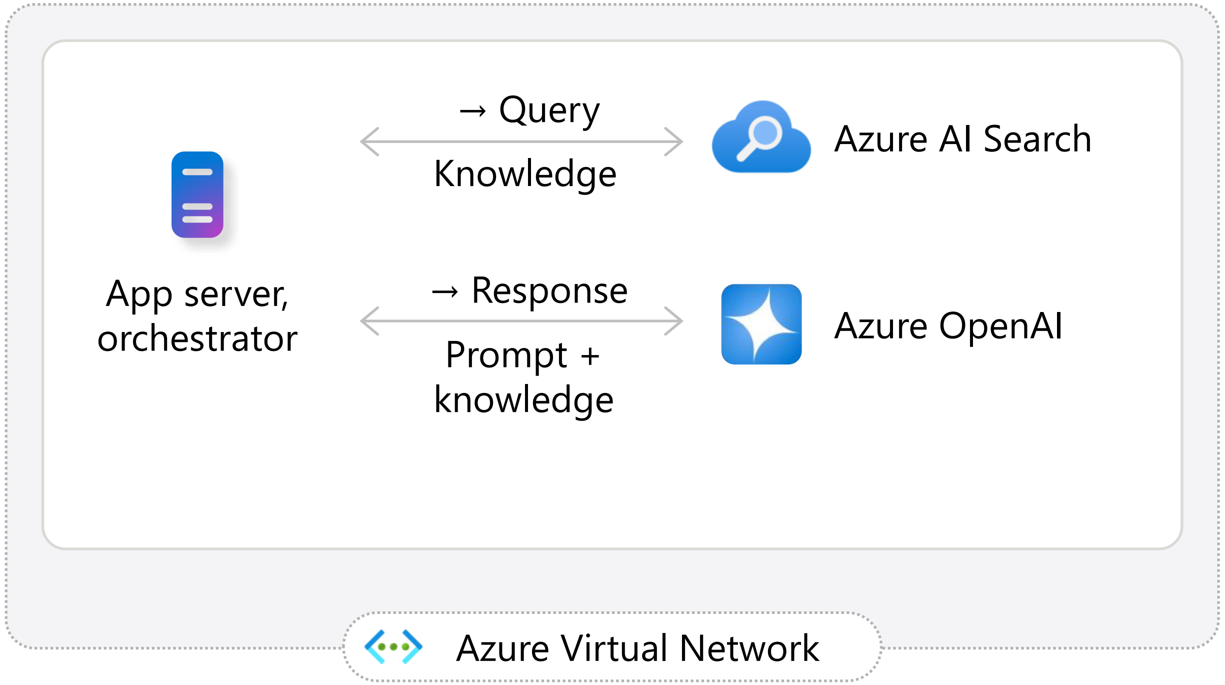
https://chatappquart-i7buow-ca.jollydune-f211632b.eastus.azurecontainerapps.io

Network security for AI apps

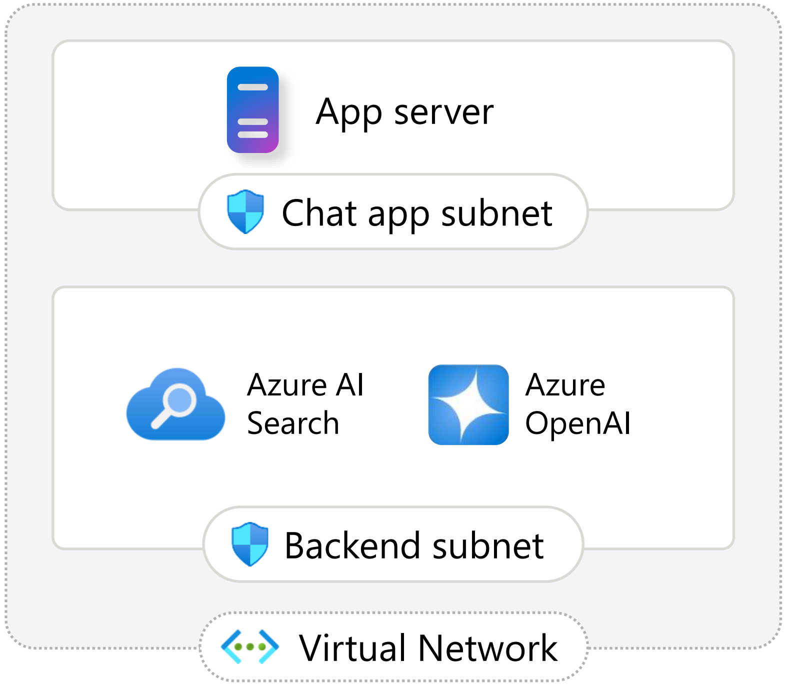


Securely networked architecture (internal facing app)

Move all resources into a virtual network:



Use subnets for further isolation:



Deploy a RAG chat inside a VNet: aka.ms/ragchat/private

VNet configuration in Bicep

Creates a subnet for:

1. App Service app
2. Backend services

Different rules can be applied to each subnet.

See full Bicep in:

aka.ms/ragchat

infra/network-isolation.bicep

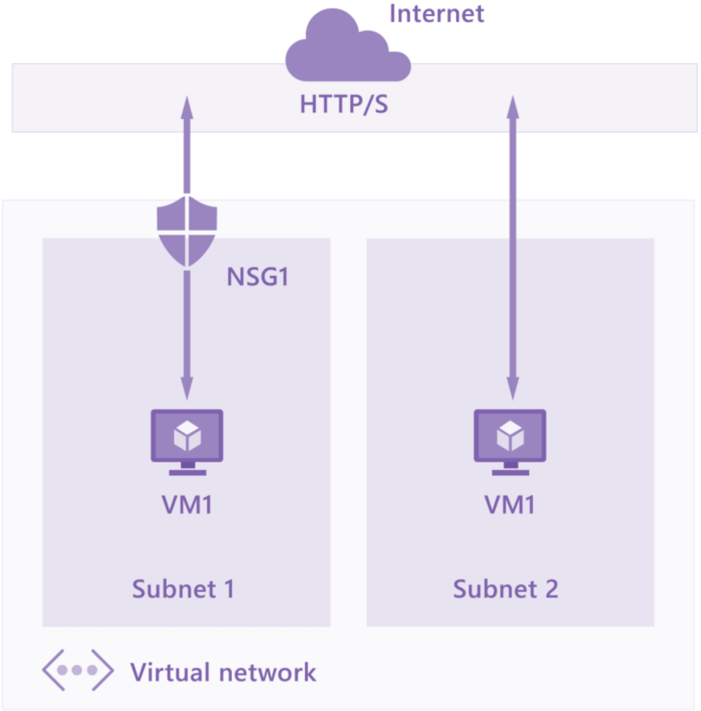
```
module vnet './core/networking/vnet.bicep' = {
  name: 'vnet'
  params: {
    subnets: [
     {
      name: 'appservice-subnet'
      properties: {
        addressPrefix: '10.0.3.0/24'
        privateEndpointNetworkPolicies: 'Enabled'
        privateLinkServiceNetworkPolicies: 'Enabled'
        delegations: [{
          id: appServicePlan.id
          name: appServicePlan.name
          properties: {
            serviceName: 'Microsoft.Web/serverFarms'
        }}]
     }
     {
      name: 'backend-subnet'
      properties: {
        addressPrefix: '10.0.1.0/24'
        privateEndpointNetworkPolicies: 'Enabled'
        privateLinkServiceNetworkPolicies: 'Enabled'
     }
    }
...
```

Azure Network Security Groups (NSG)

Azure network security groups can automatically allow or deny traffic

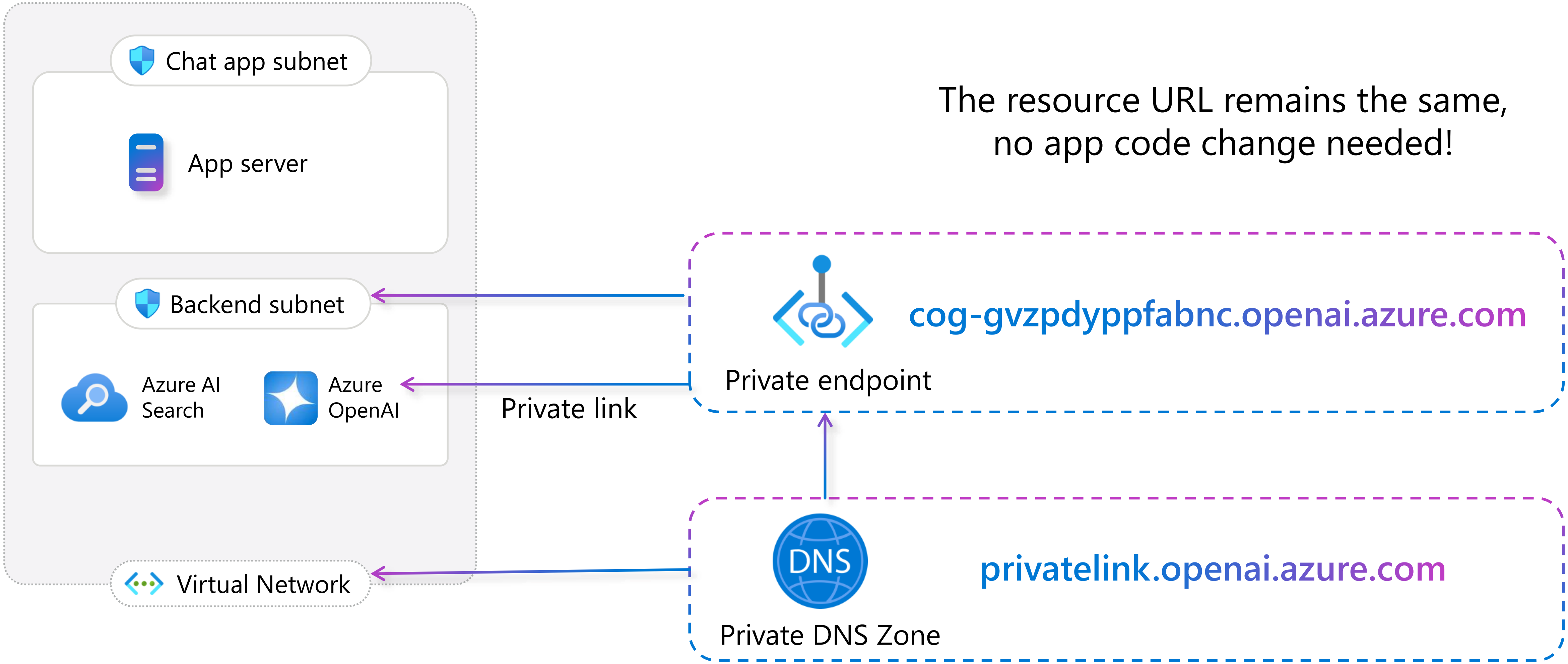
Contains security rules

NSG security rules are evaluated by priority using five information points



Private endpoints and DNS zones

The resource URL remains the same, no app code change needed!



Private endpoints in Bicep

Create private DNS zones and endpoints for:

- Azure Blob Storage
- Azure OpenAI
- Azure AI Search
- Azure App Service

The endpoint for the service remains the same! *No changes to backend code are needed.*

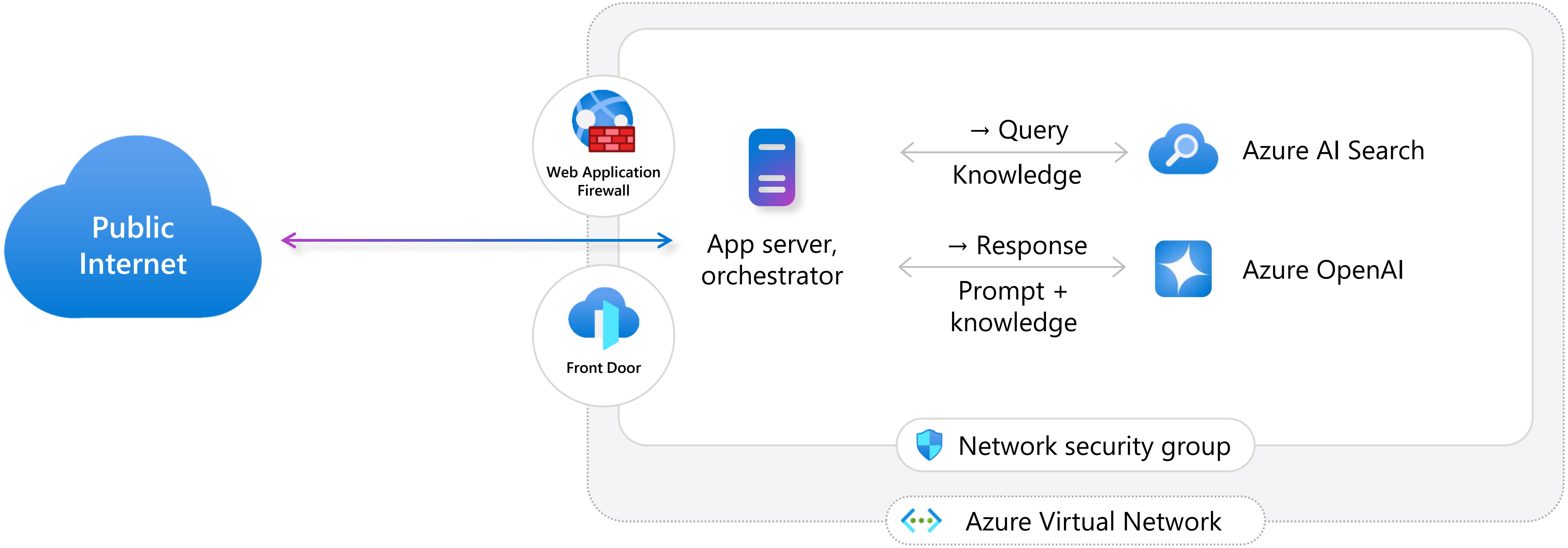
See full Bicep in: aka.ms/ragchat

`infra/network-isolation.bicep`

```
module dnsZones 'private-dns-zone.bicep' =
[for privateEndpointConnection in privateEndpointConnections:
  {
    name: '${privateEndpointConnection.groupId}-dnszone'
    params: {
      dnsZoneName: privateEndpointConnection.dnsZoneName
      tags: tags
      virtualNetworkName: vnetName
}}]

module privateEndpoints 'private-endpoint.bicep' =
[for privateEndpointInfo in flatten(privateEndpointInfo):
  {
    name: '${privateEndpointInfo.name}-privateendpoint'
    params: {
      location: location
      name: '${privateEndpointInfo.name}${resourceToken}-pe'
      tags: tags
      subnetId: vnetPeSubnetName
      serviceId: privateEndpointInfo.resourceId
      groupIds: [ privateEndpointInfo.groupId ]
      dnsZoneId: dnsZones[privateEndpointInfo.dnsIdx].outputs.id
    }
    dependsOn: [ dnsZones ]
}]
```

Securely networked architecture (public app)

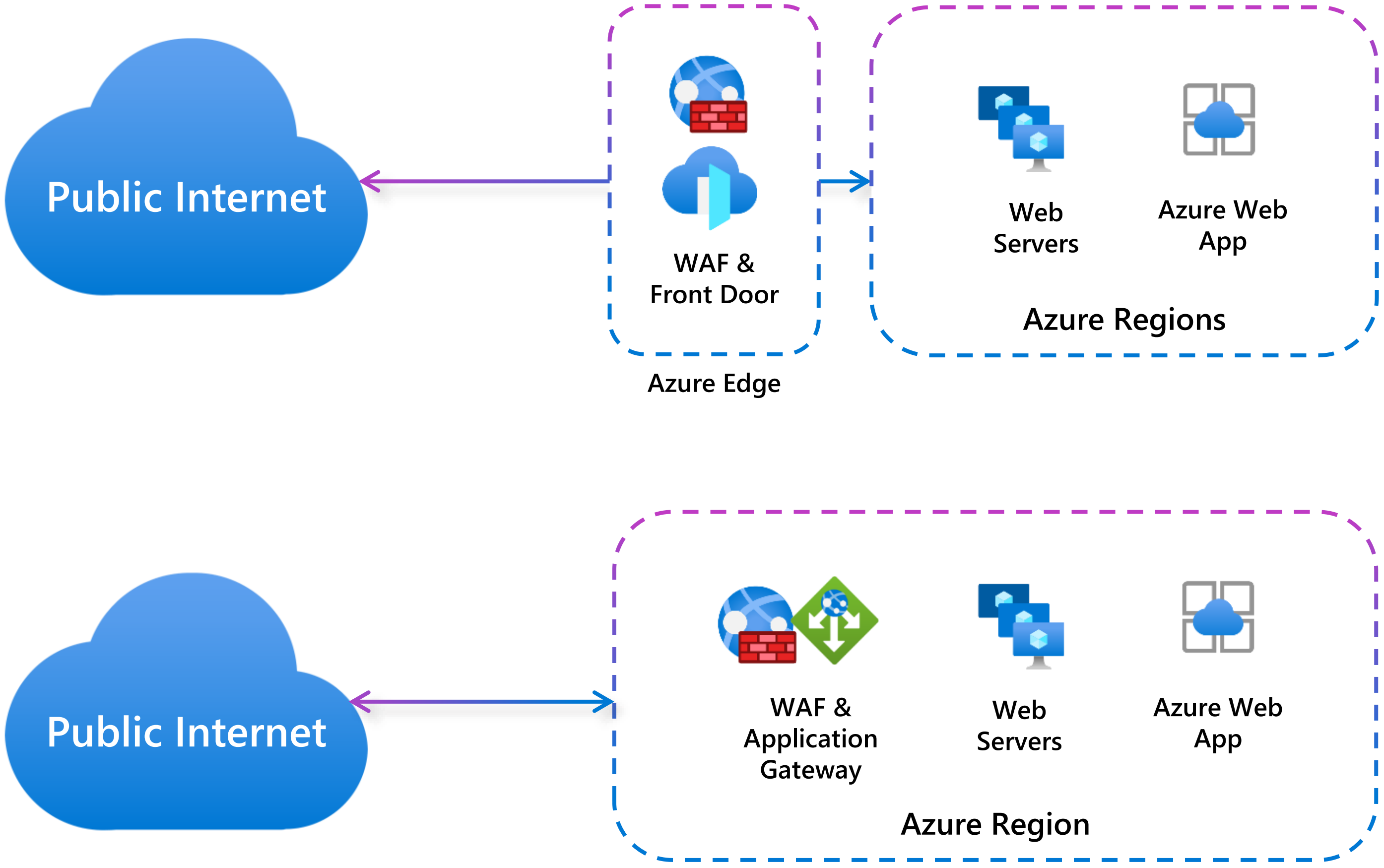
Protect public-facing applications with Azure Web Application Firewall plus Front Door:



**Front Door can be replaced with Application Gateway for a regionally distributed app*

Azure Web Application Firewall (WAF): Front Door or Application Gateway?

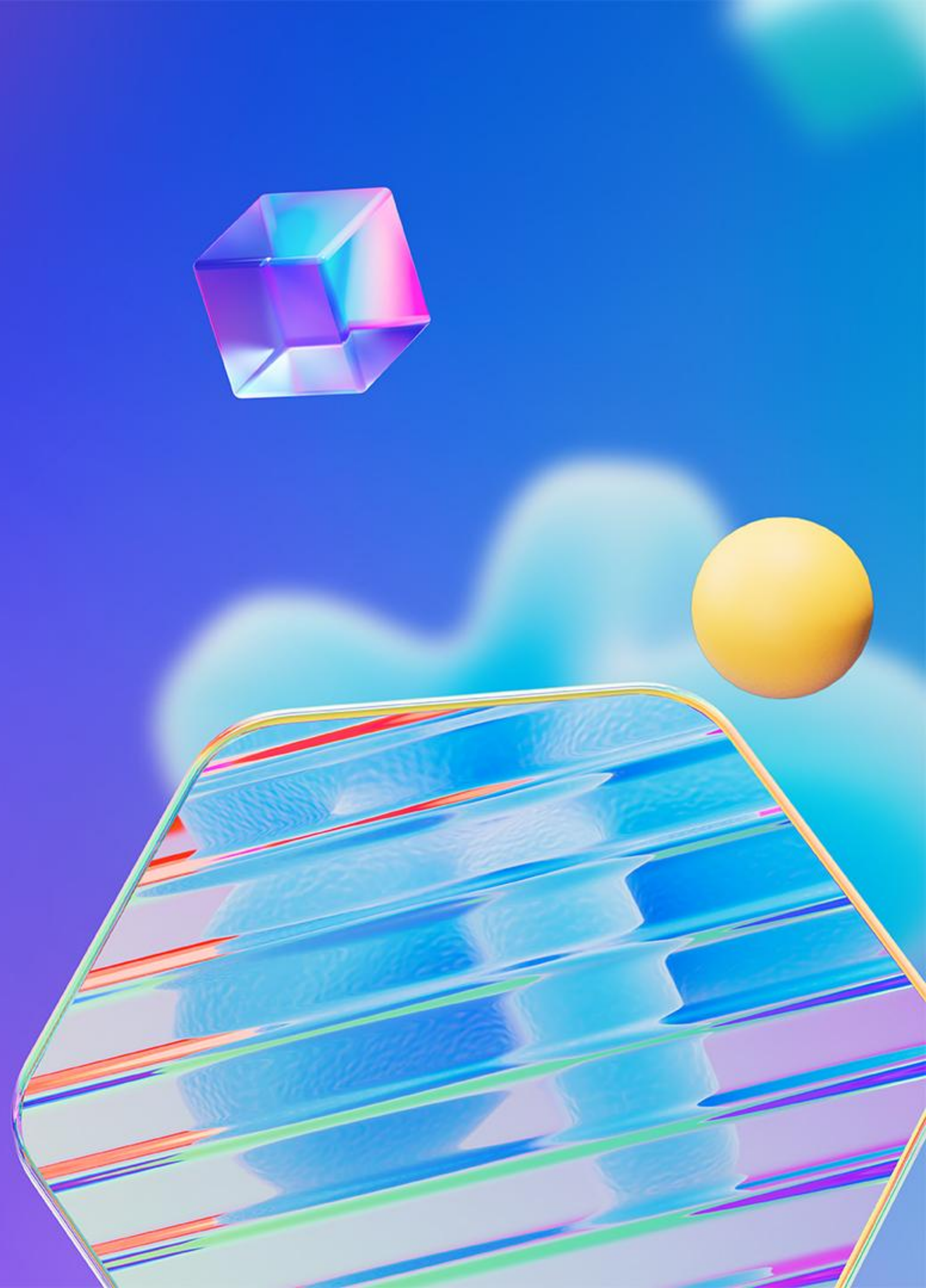
1. Scalable, highly available, Low latency service provided at network edge
2. Easy setup with managed ruleset (OWASP TOP 10) and custom rules
3. Bot protection using threat intelligence-based filtering (preview)
4. Global insights
5. Built-in DDoS protection
6. Azure Front Door provides built-in CDN capabilities
7. Cost efficient:
   Pay as you go



Azure Front Door with WAF in Bicep

```
resource profile 'Microsoft.Cdn/profiles' = {
  name: 'frontdoor-profile'
  location: 'global'
  sku: {
   name: 'Standard_AzureFrontDoor'
  }
}

resource policy
'Microsoft.Network/frontDoorWebApplicationFirewallPolicies' = {
  name: 'waf-policy'
  location: 'global'
  sku: {
    name: 'Standard_AzureFrontDoor'
  }
  properties: {
    policySettings: {
    enabledState: 'Enabled'
    mode: 'Prevention'
  }
 }
}
```

Continuous security for AI



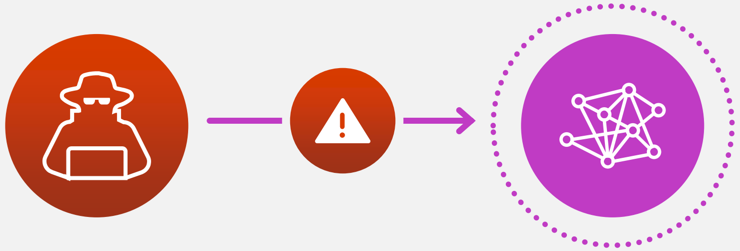
Protect AI apps from code to runtime

Start secure

AI security posture management (AI-SPM)

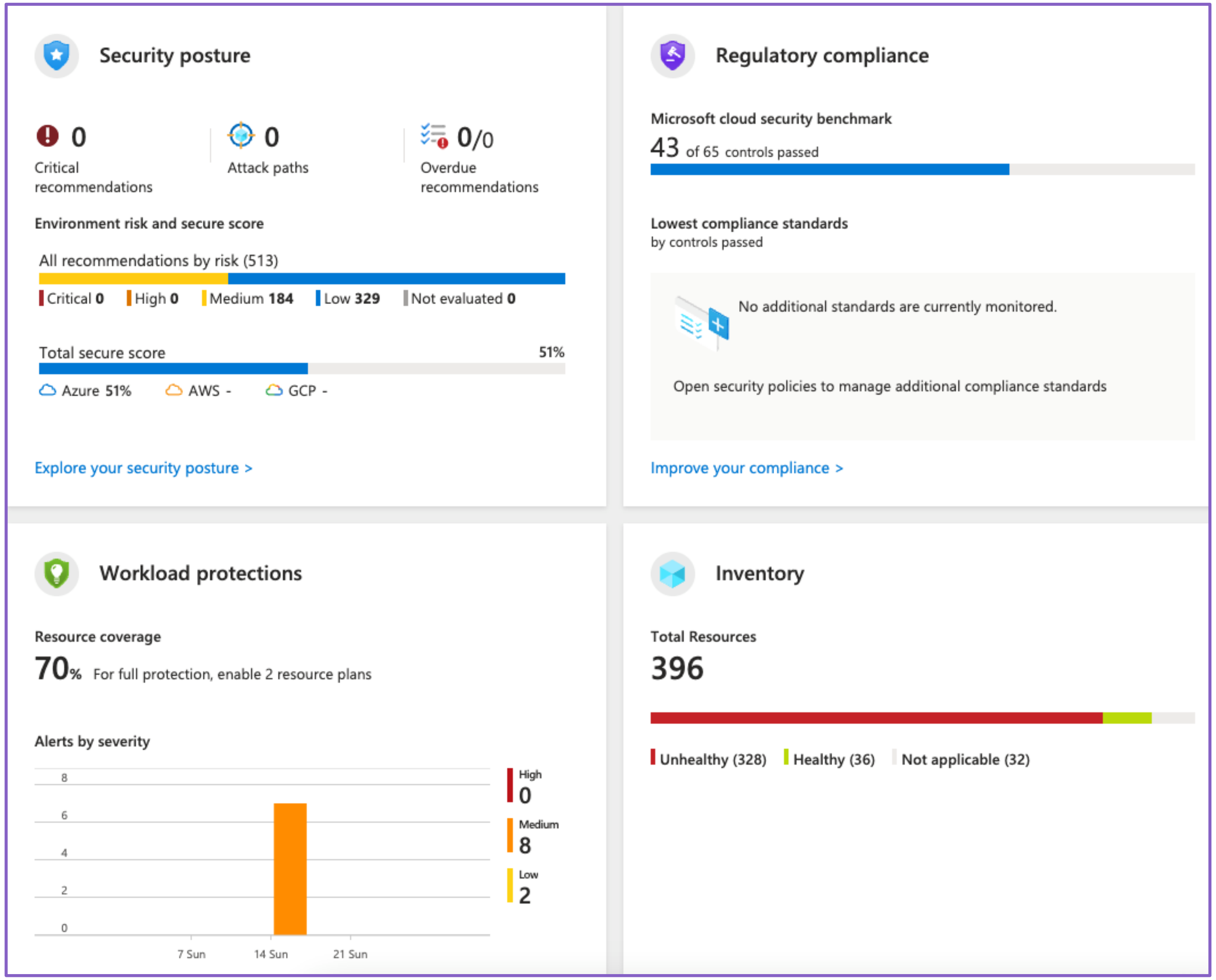
Stay secure

Threat protection for AI workloads



Microsoft Defender for Cloud

Defender for Cloud



Security posture

0 Critical recommendations

0 Attack paths

0/0 Overdue recommendations

Environment risk and secure score

All recommendations by risk (513)

Critical 0 | High 0 | Medium 184 | Low 329 | Not evaluated 0

Total secure score 51%

Azure 51% | AWS - | GCP -

Explore your security posture >

Regulatory compliance

Microsoft cloud security benchmark

43 of 65 controls passed

Lowest compliance standards by controls passed

No additional standards are currently monitored.

Open security policies to manage additional compliance standards

Improve your compliance >

Workload protections

Resource coverage

70% For full protection, enable 2 resource plans

Alerts by severity

High 0

Medium 8

Low 2

Inventory

Total Resources

396

Unhealthy (328) | Healthy (36) | Not applicable (32)

Security alerts: Detects DDOS, suspicious logins, etc.

Security posture: Audits Azure resources and their settings

Workload protections: Scans for known vulnerabilities in SQL, container images, etc.

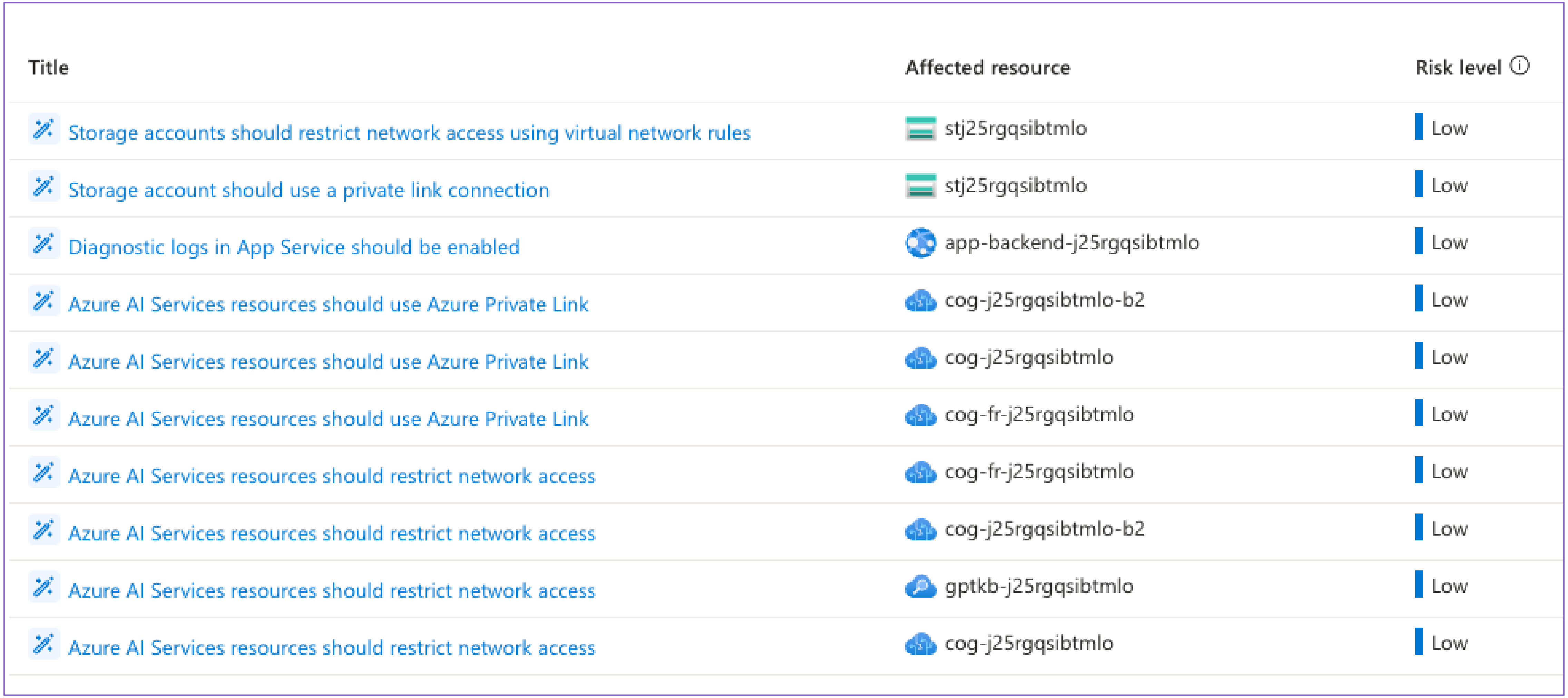
Data security: Scans stored data for PII and sensitive data

Regulatory compliance: Ensure compliance with benchmarks.

aka.ms/enable-defender

DfC recommendations: RAG *without* VNet

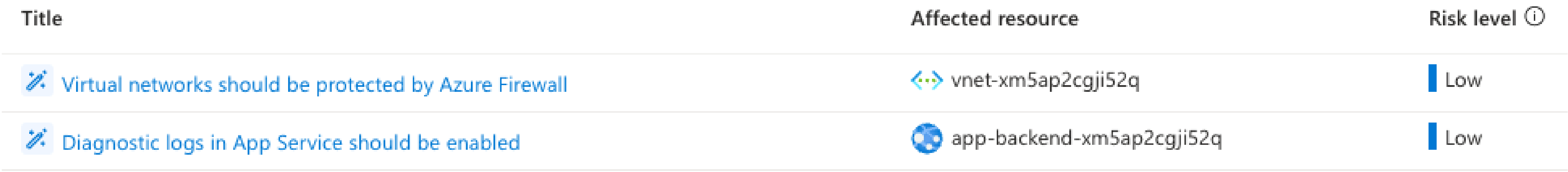
10 recommendations for azure-search-openai-demo, non-private deployment:



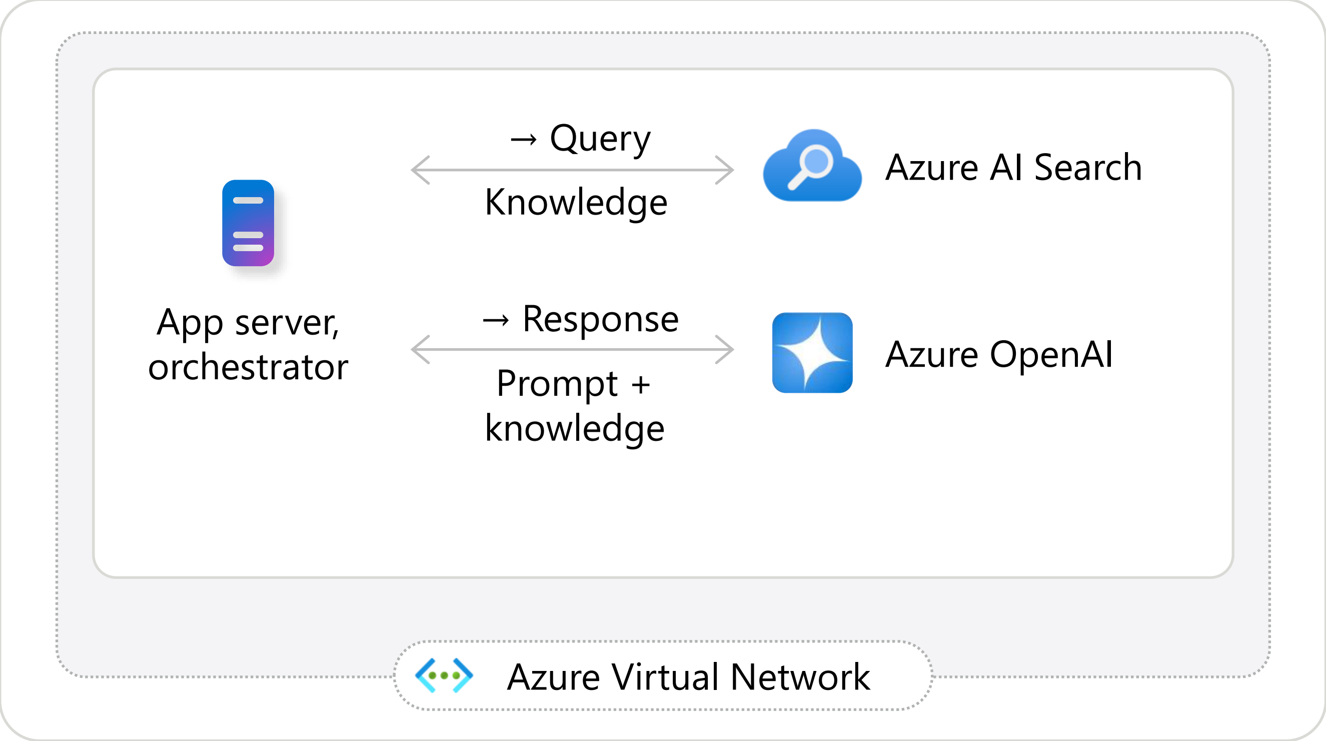
Title	Affected resource	Risk level ⓘ
Storage accounts should restrict network access using virtual network rules	stj25rgqsibtmlo	Low
Storage account should use a private link connection	stj25rgqsibtmlo	Low
Diagnostic logs in App Service should be enabled	app-backend-j25rgqsibtmlo	Low
Azure AI Services resources should use Azure Private Link	cog-j25rgqsibtmlo-b2	Low
Azure AI Services resources should use Azure Private Link	cog-j25rgqsibtmlo	Low
Azure AI Services resources should use Azure Private Link	cog-fr-j25rgqsibtmlo	Low
Azure AI Services resources should restrict network access	cog-fr-j25rgqsibtmlo	Low
Azure AI Services resources should restrict network access	cog-j25rgqsibtmlo-b2	Low
Azure AI Services resources should restrict network access	gptkb-j25rgqsibtmlo	Low
Azure AI Services resources should restrict network access	cog-j25rgqsibtmlo	Low

DfC recommendations: RAG *with* VNet

2 recommendations for azure-search-openai-demo, private deployment:

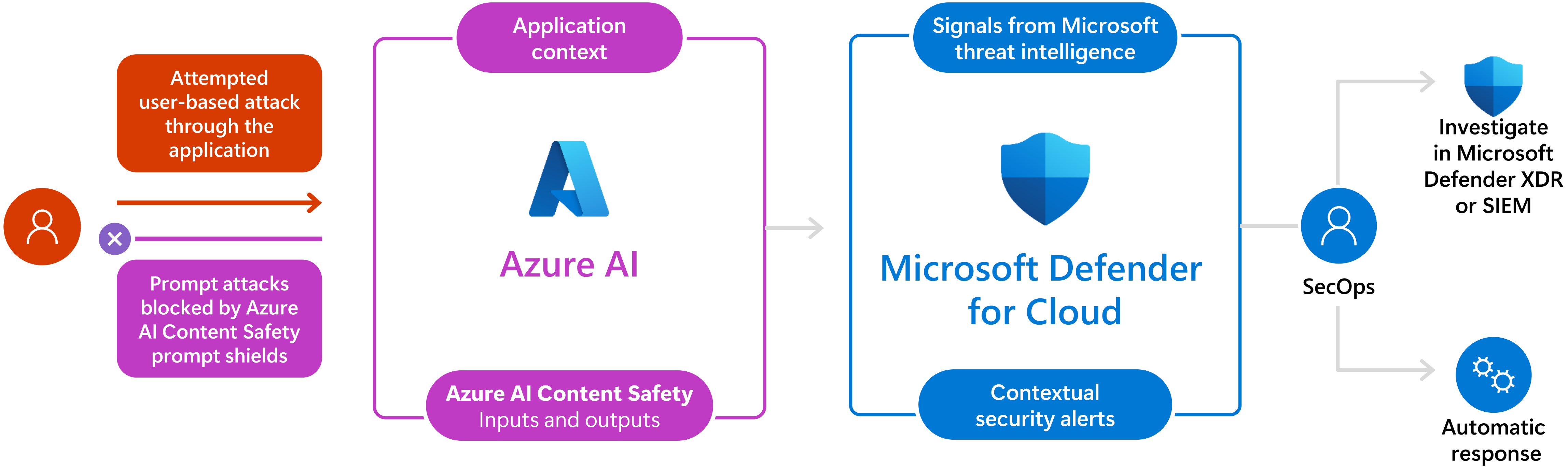


Title	Affected resource	Risk level ⓘ
Virtual networks should be protected by Azure Firewall	vnet-xm5ap2cgji52q	Low
Diagnostic logs in App Service should be enabled	app-backend-xm5ap2cgji52q	Low



Threat protection for AI workloads

Microsoft Defender for Cloud + Azure AI Content Safety



Attempted user-based attack through the application

Prompt attacks blocked by Azure AI Content Safety prompt shields

Application context

Azure AI

Azure AI Content Safety
Inputs and outputs

Signals from Microsoft threat intelligence

Microsoft Defender for Cloud

Contextual security alerts

SecOps

Investigate in Microsoft Defender XDR or SIEM

Automatic response

https://learn.microsoft.com/azure/defender-for-cloud/ai-onboarding

Key: Security teams Developers

GitHub actions for security recommendations

Use ps-rule action on your Bicep to auto-scan for security issues

github.com/microsoft/ps-rule

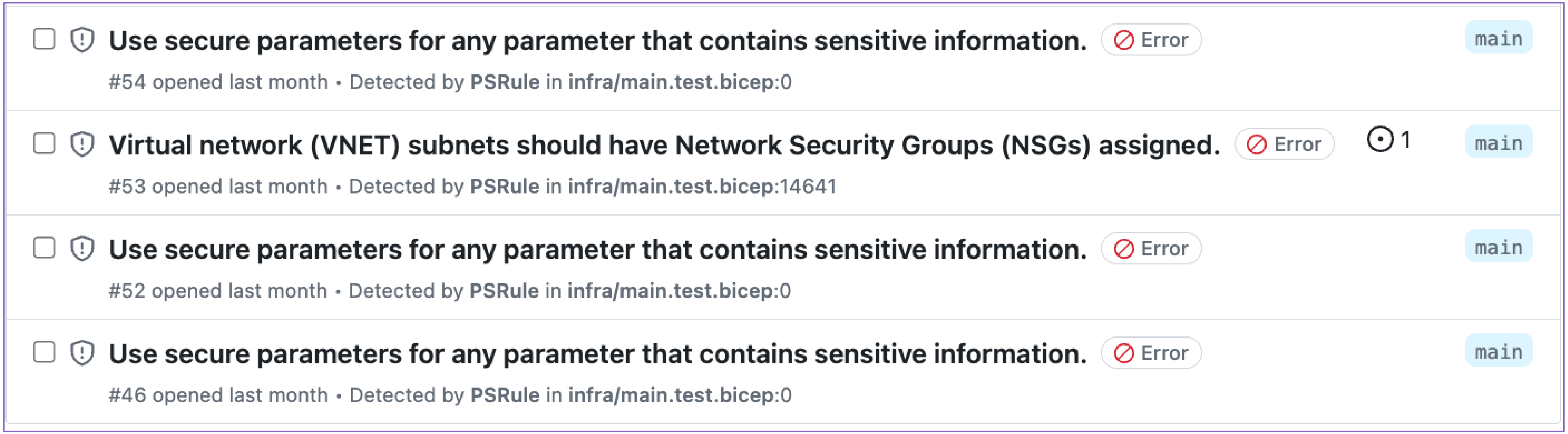
```yaml
- name: Run PSRule analysis
uses: microsoft/ps-rule@v2.9.0
with:
  modules: PSRule.Rules.Azure
  baseline: Azure.Pillar.Security
  inputPath: infra/*.test.bicep
  outputFormat: Sarif
  outputPath: reports/ps-rule-results.sarif
  summary: true
continue-on-error: true
env:
  PSRULE_CONFIGURATION_AZURE_BICEP_FILE_EXPANSION: 'true'
  PSRULE_CONFIGURATION_AZURE_BICEP_FILE_EXPANSION_TIMEOUT: '30'

- name: Upload results to security tab
uses: github/codeql-action/upload-sarif@v3
with:
  sarif_file: reports/ps-rule-results.sarif
```

Blog post:

Securing Azure deployments with PSRule

aka.ms/blog-psrule

GitHub actions for security recommendations: Results



□ Use secure parameters for any parameter that contains sensitive information. Error main

#54 opened last month • Detected by **PSRule** in **infra/main.test.bicep**:0

□ Virtual network (VNET) subnets should have Network Security Groups (NSGs) assigned. Error 1 main

#53 opened last month • Detected by **PSRule** in **infra/main.test.bicep**:14641

□ Use secure parameters for any parameter that contains sensitive information. Error main

#52 opened last month • Detected by **PSRule** in **infra/main.test.bicep**:0

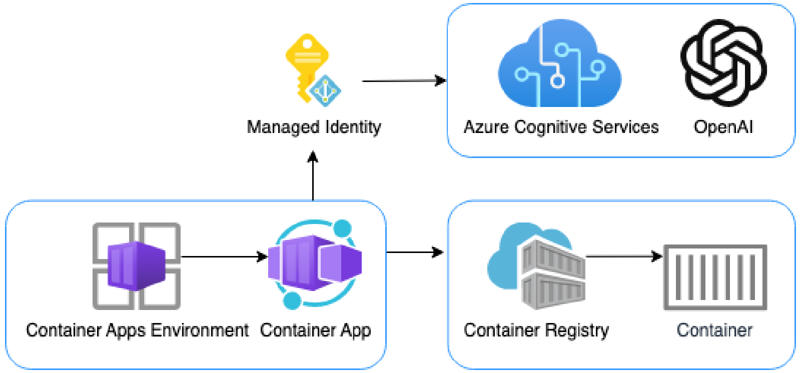
□ Use secure parameters for any parameter that contains sensitive information. Error main

#46 opened last month • Detected by **PSRule** in **infra/main.test.bicep**:0

https://github.com/Azure-Samples/azure-search-openai-demo/actions/runs/9378324878

Wrap up

Get started with our samples

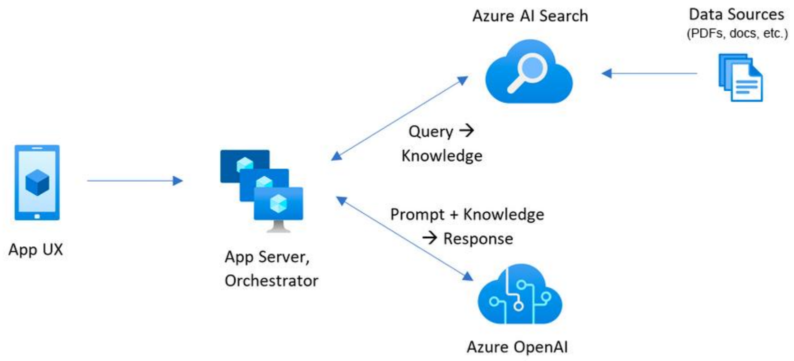


aka.ms/azai/chat

Azure OpenAI + Entra + Container Apps Built-in Auth

aka.ms/azai/chat/identity

Azure OpenAI + Entra + MSAL + Identity package



aka.ms/ragchat

Azure OpenAI + AI Search
+ Entra + MSAL + App Service Built-in Auth
+ VNet + Private Endpoints

Learn more about securing your AI application

Microsoft Entra developer center - aka.ms/dev/ms-entra

Get started with Defender for Cloud - aka.ms/enable-defender

Python Risk Identification Tool for generative AI – aka.ms/pyrit

Azure Well Architected Framework – aka.ms/wellarchitectedframework

Azure AI Content Safety – aka.ms/aicontentsafety

Tune in to our AI security webinar series

Copilot L33T Sp34k is a webinar series where we interview industry experts about how to use AI securely and how organizations should use AI, like Microsoft Copilot for Security, to enhance their security.

aka.ms/copilotl33tsp34k

Feedback

Your feedback is valuable.

Please submit your thoughts about today's experiences at aitour.microsoft.com/en-US/sessions/BRK420-FR

...or use the QR code.

Scan QR code to respond

Microsoft