codit
proximus NXT

APIM ❤ OpenAI

exploring the generative AI Gateway pattern

Massimo Crippa

Lead Architect at codit
proximus NXT

https://www.linkedin.com/in/massimocrippa | https://massimocrippa.com/blog | https://github.com/massimoc | https://apimhotrod.cloud

Generative AI

has ignited a remarkable range of possibilities

All industry sectors

are embracing AI advancements

Explore and Understand

run and accelerate experimentation

Unique challenges

could slow down or even completely block adoption

codit
proximus NXT

Strategy

ensure the effective use of AI services

codit
proximus NXT

Azure model catalog - flexible deployment options

Managed compute

- Deployed to managed Azure VMs
- One click deployment
- Pay per GPU billing
- 100+ open models

Serverless APIs and Model as a Service

- Ready to use model APIs
- Pay per token
- Pay per provisioned capacity
- 30+ flagship models

Azure OpenAI Service

GPT-4

GPT-4-Turbo

GPT-3.5-Turbo

GPT-4 for Vision

GPT-4o

Generative Text Models, with varying capabilities and uses

DALL·E 3

Generative Image Model

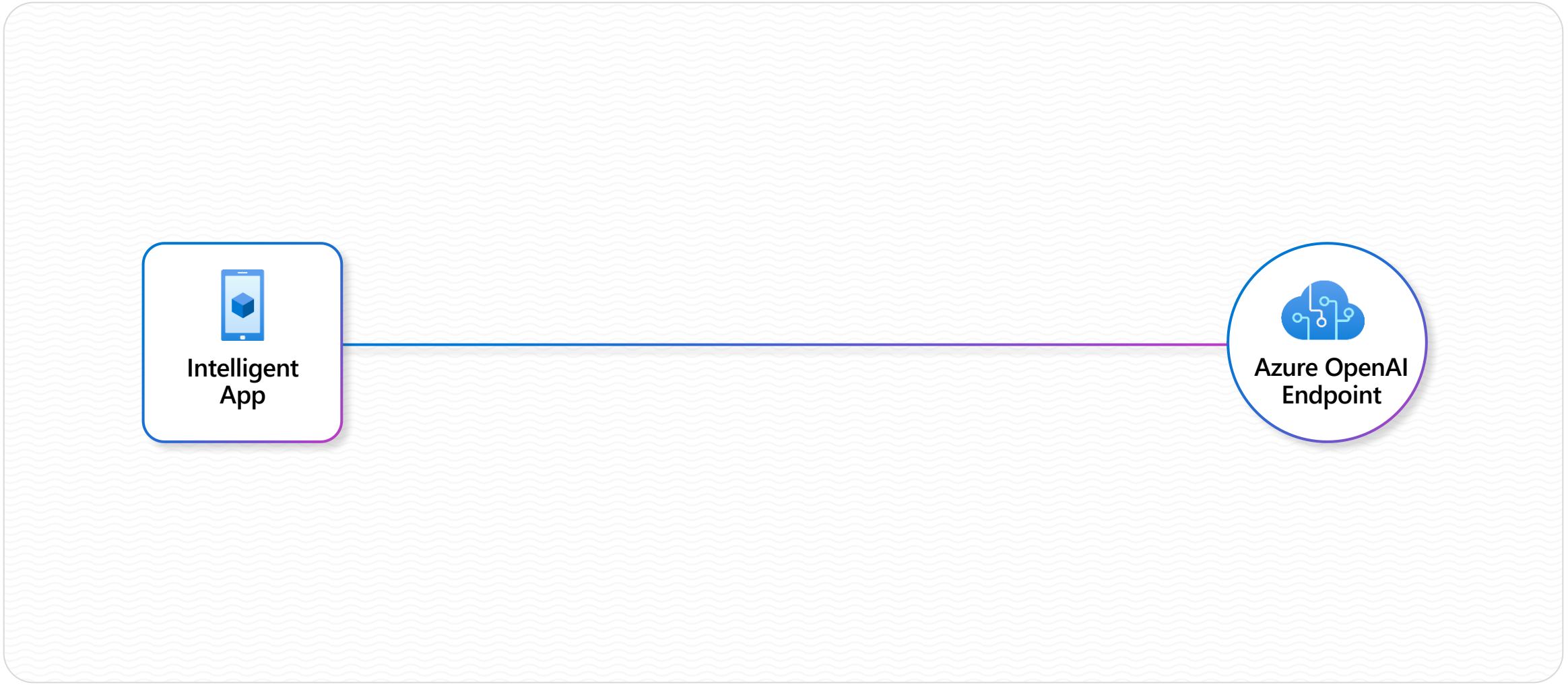
Tokens & Quotas

Pay-as-you-go & Provisioned Throughput Units (PTUs)

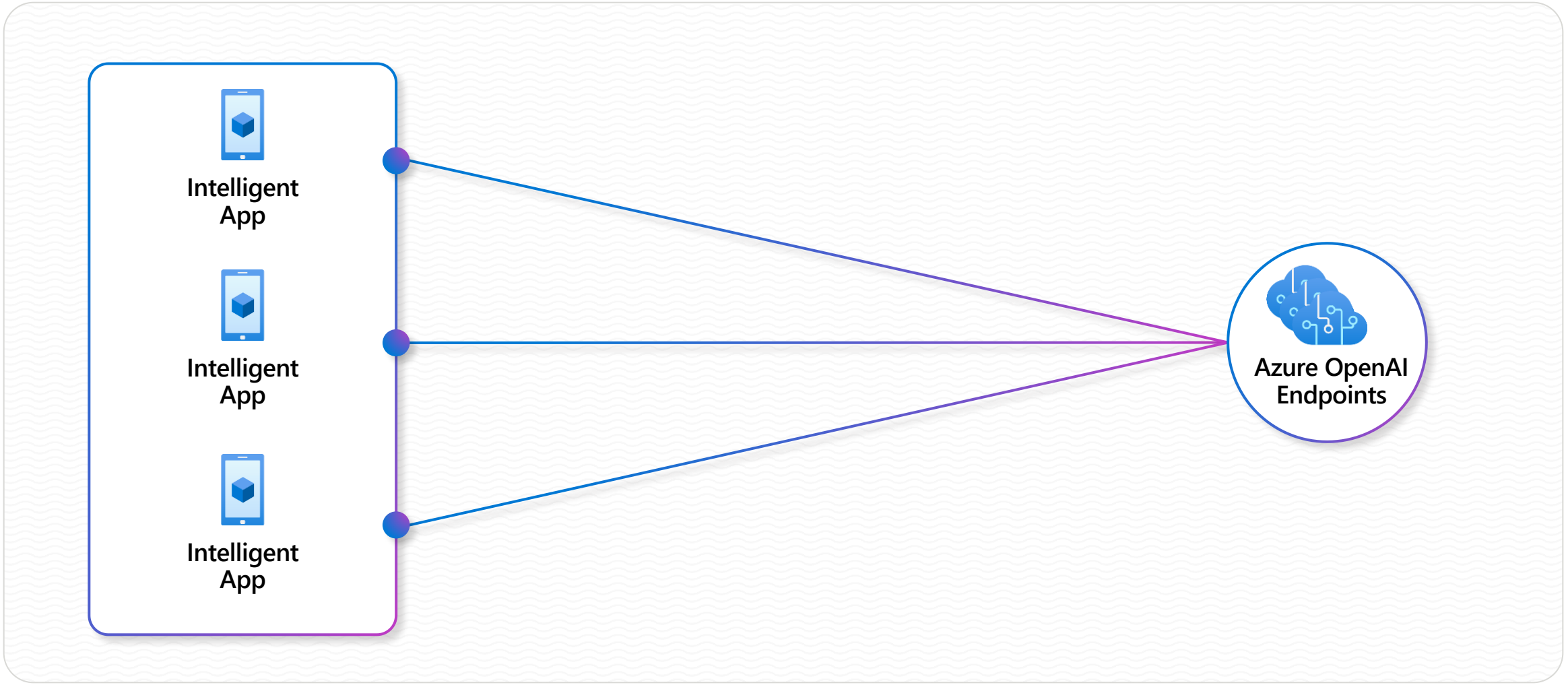
API & SDK

codit
proximus NXT

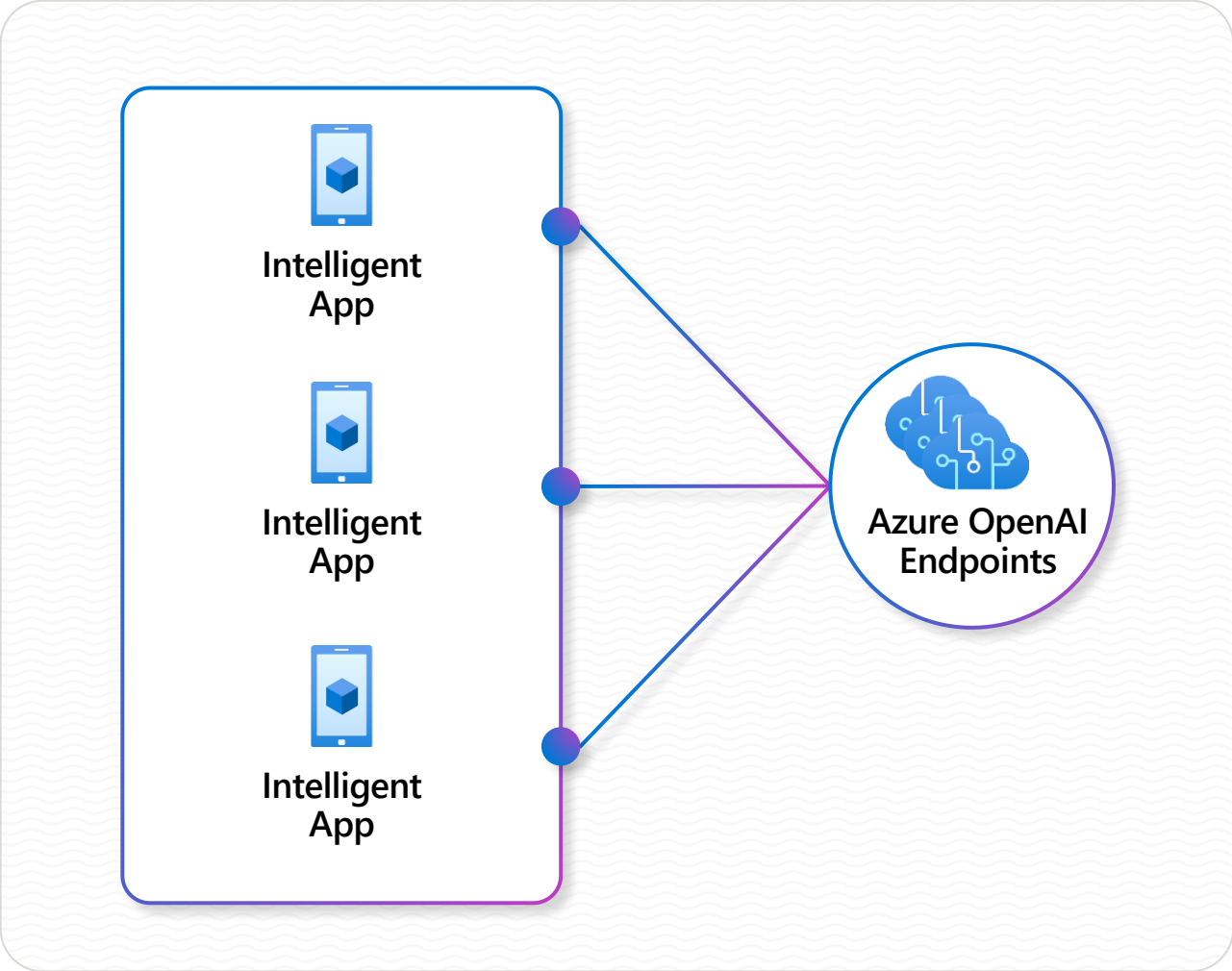
Connecting an Intelligent App with OpenAI



Scaling Up: Multiple Apps, Multiple OpenAI Endpoints



Scaling Up: Multiple Apps, Multiple OpenAI Endpoints

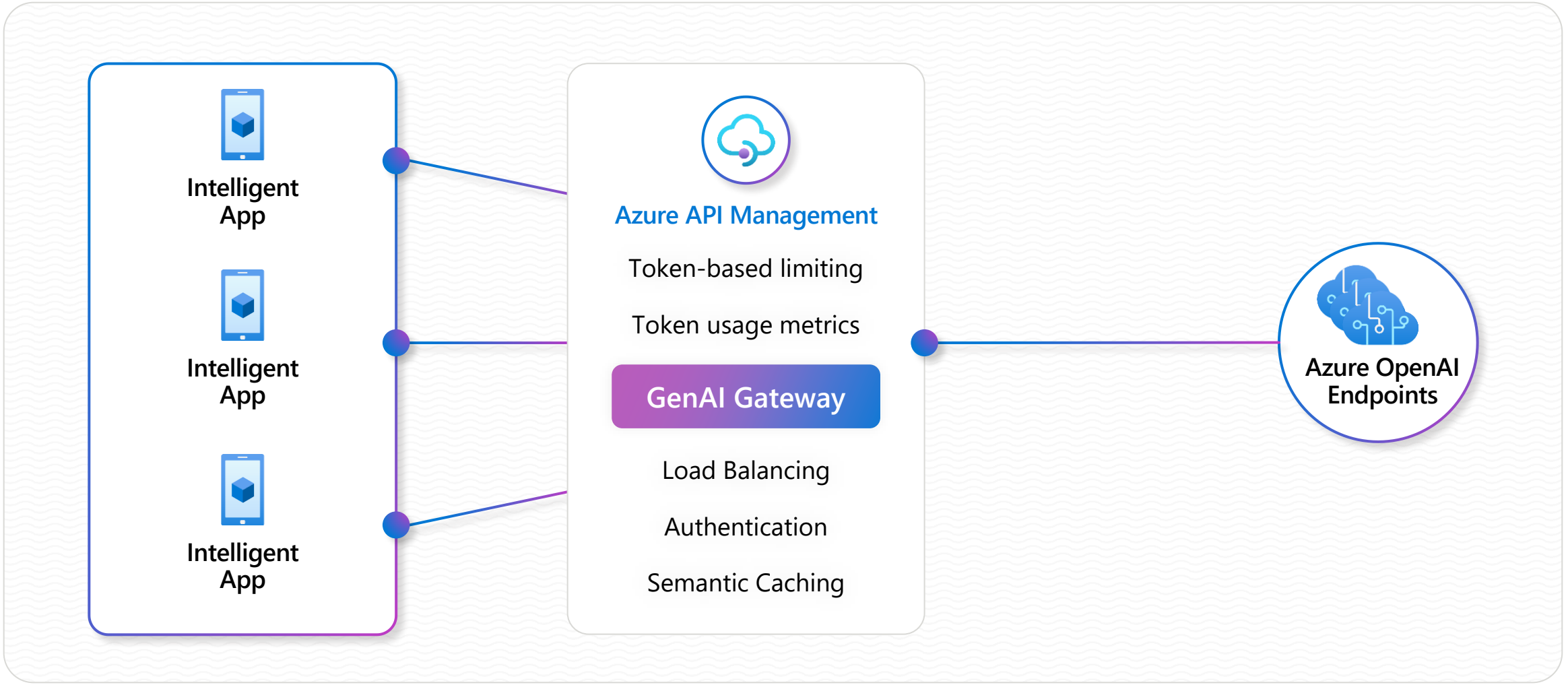


Scaling Challenges

- Track token usage
- Multiple OpenAI endpoints
- Authentication and authorization
- Assign token-based limits

codit
proximus NXT

Delegate security, scalability and observability to the GenAI Gateway

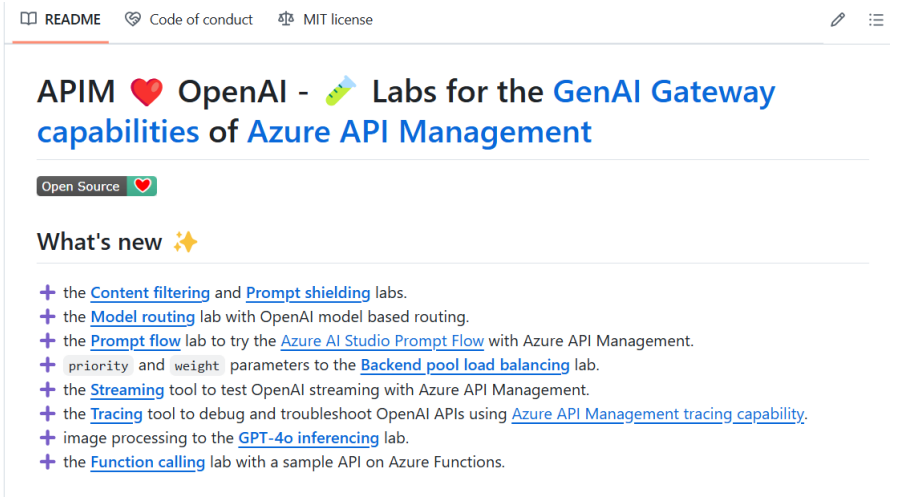


codit
proximus NXT

Explore GenAI Gateway – where to start?

GenAI Gateway Labs
https://aka.ms/apim/genai/labs
(https://aka.ms/ai-gateway)

→ experimentation-driven approach that pave the road to success


README Code of conduct MIT license

APIM ❤️ OpenAI - 🧪 Labs for the GenAI Gateway capabilities of Azure API Management

Open Source ❤️

What's new ✨

- the Content filtering and Prompt shielding labs.
- the Model routing lab with OpenAI model based routing.
- the Prompt flow lab to try the Azure AI Studio Prompt Flow with Azure API Management.
- `priority` and `weight` parameters to the Backend pool load balancing lab.
- the Streaming tool to test OpenAI streaming with Azure API Management.
- the Tracing tool to debug and troubleshoot OpenAI APIs using Azure API Management tracing capability.
- image processing to the GPT-4o inferencing lab.
- the Function calling lab with a sample API on Azure Functions.

codit
proximus NXT

Security

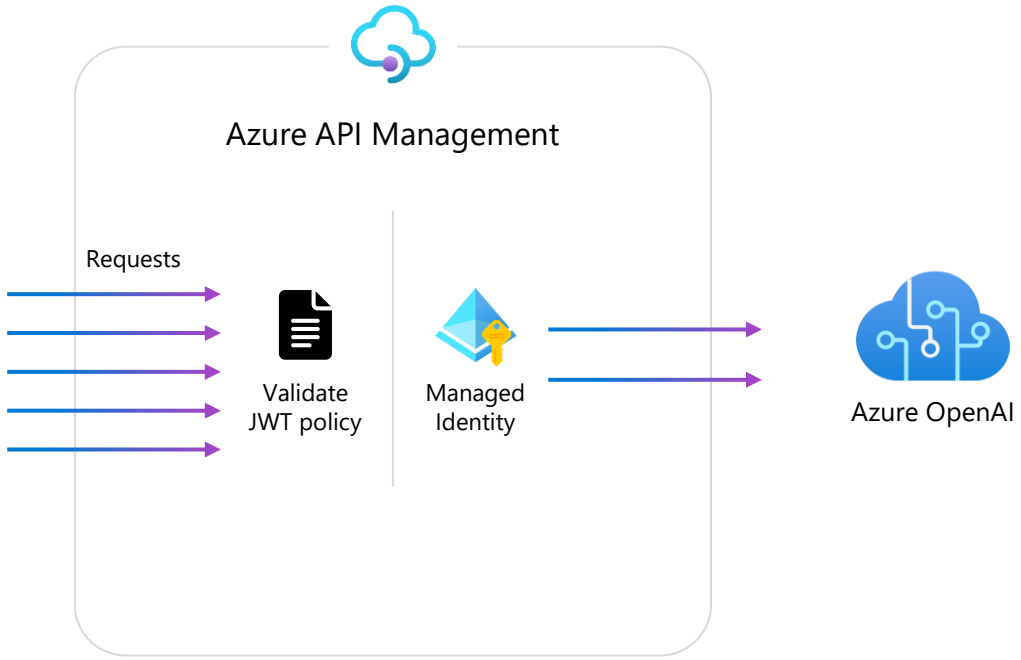
Security

Specialized security policies : OAuth, API key/token validation, schema validation, and threat detection tailored to your API's structure.

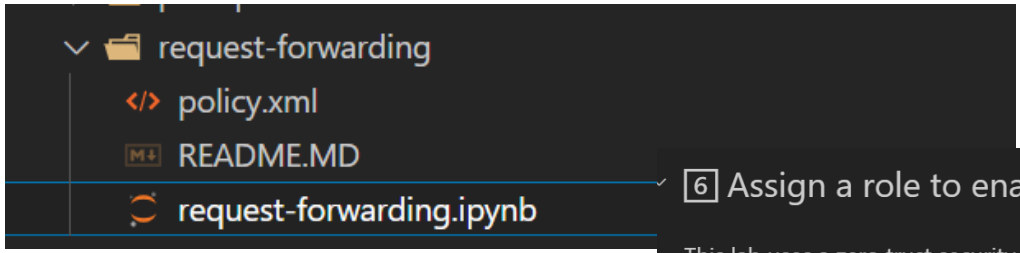
Positive Security Model : explicitly defines what is allowed rather than just relying on detecting abnormalities.

Traffic Control : Rate limits and quotas protect your AI resources from abuse and uncontrolled usage spikes.

Authentication and Authorization



- Configure managed identity authentication
- Validate claims in JWT to manage access to OpenAI endpoints
- Authenticate API consumers using subscription keys



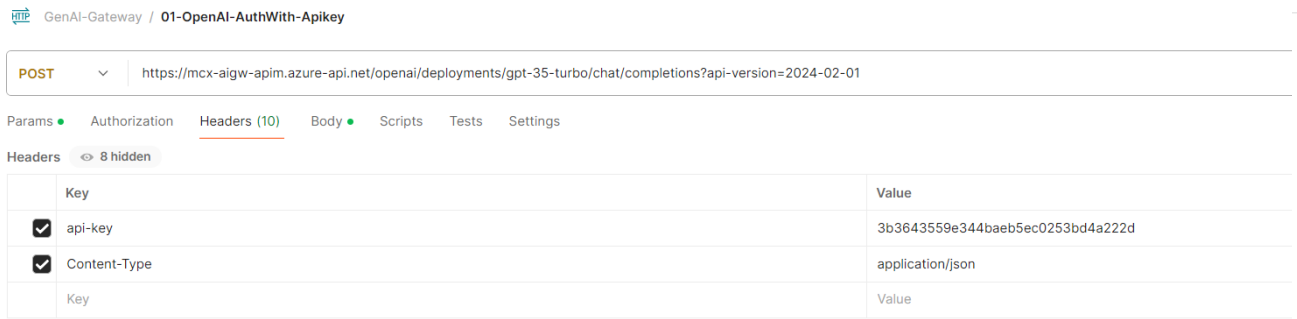
Capacity	Rate limit (tokens per minute)
20K TPM	20,000

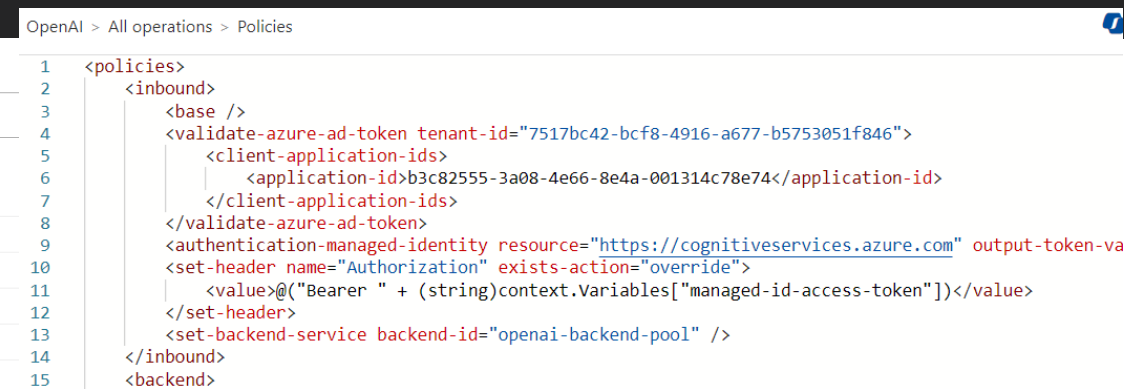
6 Assign a role to enable APIM to access OpenAI API

This lab uses a zero trust security strategy with a key less approach using an Azure Managed Identity. The following script assigns the `Cognitive Services OpenAI User` the OpenAI API.

Generate | + Code | + Markdown

```
if mock_disabled:
    openai_resource_name = openai_resources[0].get("name")
    openai_resource_stdout = ! az cognitiveservices account show --name {openai_resource_name} --resource-group {resource_group}
    openai_resource = json.loads(openai_resource_stdout.n)
    openai_resource_id = openai_resource.get("id")
    role_assignment_stdout = ! az role assignment create --assignee {apim_managed_identity} \
        --role "Cognitive Services OpenAI User" \
        --scope {openai_resource_id}
```

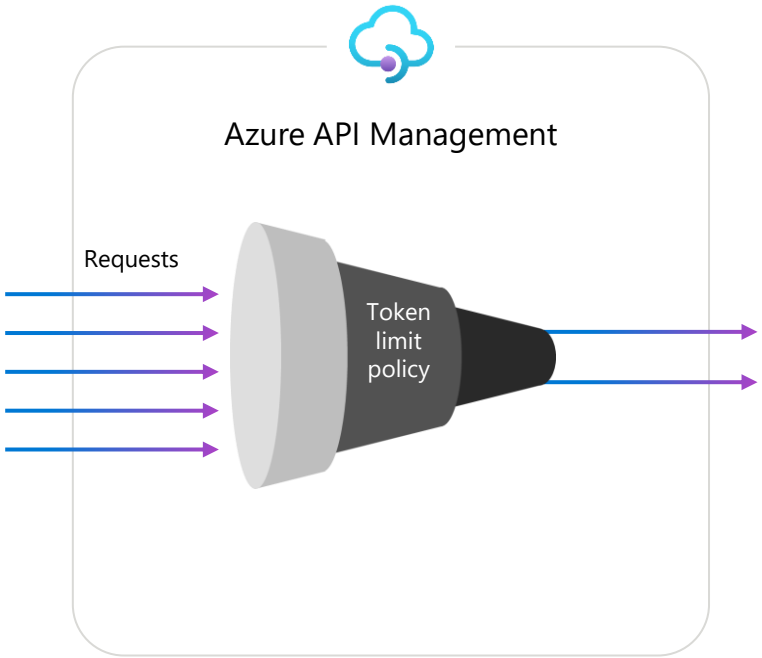




Token count

timestamp	model	region	promptTokens	completionTokens	totalTokens	remainingTokens	remainingRequests
10/18/2024, 10:17:09.534 AM		France Central	32	76	108	19952	18
10/18/2024, 10:17:06.337 AM		Sweden Central	32	31	63	19952	18
10/18/2024, 10:17:01.652 AM		France Central	32	37	69	19968	18
10/18/2024, 10:16:58.415 AM		Sweden Central	32	85	117	19968	19
10/18/2024, 10:16:56.518 AM		France Central	32	19	51	19984	19

[GA] Azure OpenAI Token Limit policy



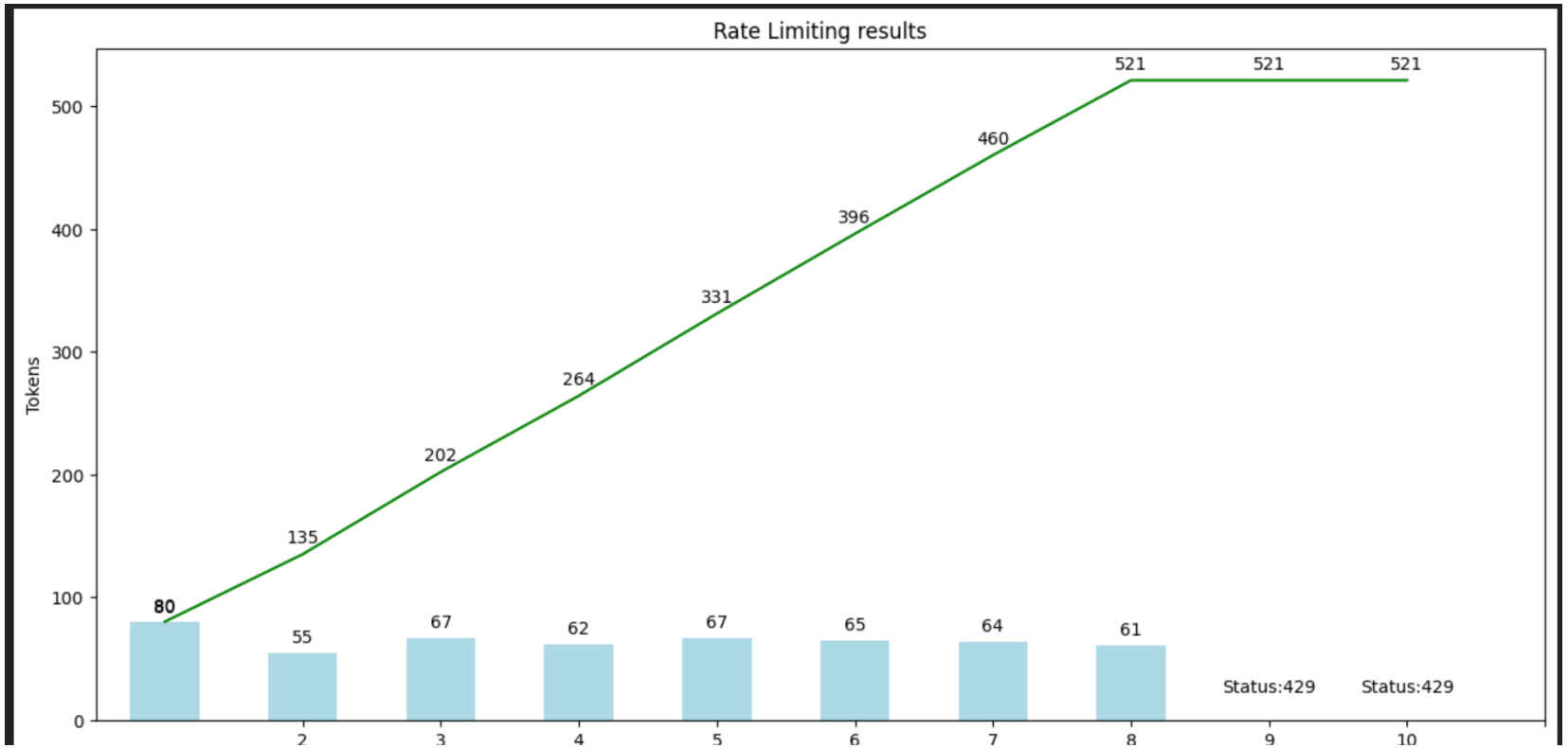
Configure tokens per minute (TPM) limits based on counter keys

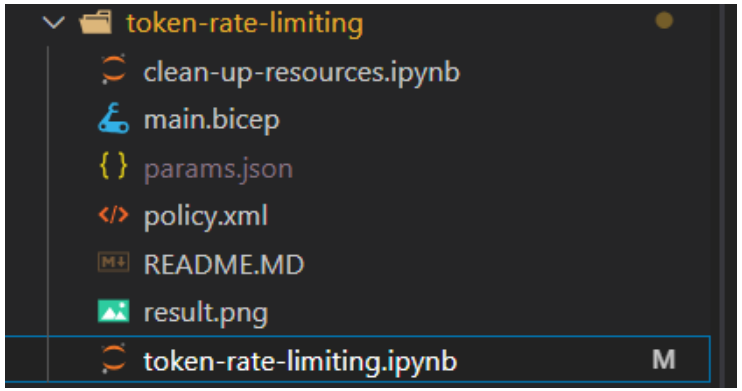
Define policy behavior for throttling

policy.xml

```
<azure-openai-token-limit
        counter-key="@(context.Subscription.Id)"
        tokens-per-minute="1000"
        estimate-prompt-tokens="false" />
```

codit
proximus NXT





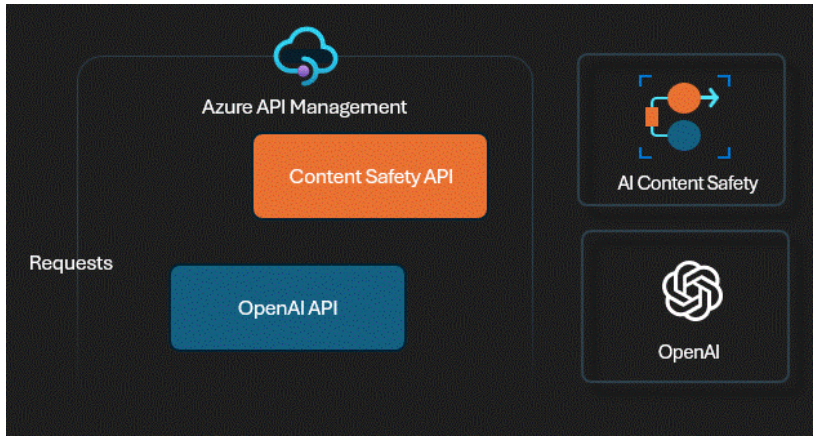
OpenAI > All operations > Policies

```xml
<policies>
    <inbound>
        <base />
        <authentication-managed-identity resource="https://cognitiveservices.azure.com" output-token-variable-name="managed-id-access-token" ignore-error="false" />
        <set-header name="Authorization" exists-action="override">
            <value>@("Bearer " + (string)context.Variables["managed-id-access-token"])</value>
        </set-header>
        <set-backend-service backend-id="openai1" />
        <azure-openai-token-limit counter-key="@(context.Request.IpAddress)" tokens-per-minute="500" estimate-prompt-tokens="false" remaining-tokens-variable-name="remainingTokens" />
    </inbound>
    <backend>
        <base />
    </backend>
    <outbound>
        <base />
    </outbound>
    <on-error>
        <base />
    </on-error>
</policies>
```

Deployment type	Fine-tune	Capacity
Standard		20K TPM

codit
proximus NXT

Prompt guarding, prompt shielding




```
<policies>
    <inbound>
        <!-- Get the request body and store it in a context variable -->
        <base />
        <set-variable name="requestBody" value="@((context.Request.Body?.As<string>(preserveContent: true) ?? ""))" />
        <!-- Define a list of prohibited words or phrases -->
        <set-variable name="prohibitedWords" value="Zeus,Apollo" />
        <!-- Check if the request body contains any prohibited content -->
        <choose>
            <when condition="@{
                var prohibitedWords = context.Variables.GetValueOrDefault<string>("prohibitedWords").Split(',');
                var requestBody = context.Variables.GetValueOrDefault<string>("requestBody");
                return prohibitedWords.Any(word => requestBody.IndexOf(word, StringComparison.OrdinalIgnoreCase) >= 0);
            }">
                <!-- If prohibited content is found, block the request and return an error response -->
                <return-response>
                    <set-status code="400" reason="Bad Request" />
                    <set-body>@{
                            var error = new {
                                error = new {
                                    code = 400,
                                    message = "Bad Request",
                                    description = "Prompt Guard - Ancient Greek mythology not allowed.",
                                    details = new {
                                        errorType = "InvalidInput",
                                        inputPrompt = "Zeus, Apollo and similar"
                                    }
                                }
                            };
                            return Newtonsoft.Json.JsonConvert.SerializeObject(error);
                        }</set-body>
                </return-response>
            </when>
        </choose>
    </inbound>
```

- Check against a list of allowed or denied expressions
- Analyzes LLM inputs and detects User Prompt attacks and Document attacks
- Leverage the APIM policies to build additional Generative AI capabilities.

codit
proximus NXT

Load Balancing

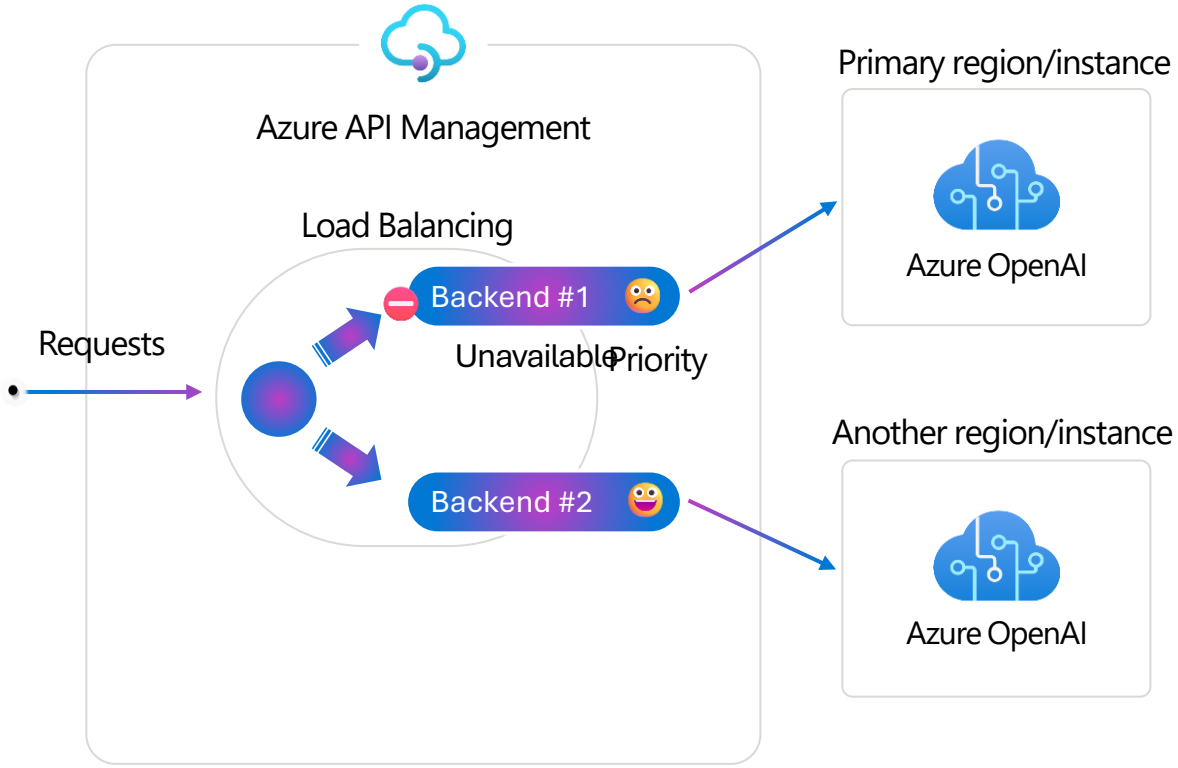
Load Balancing

Intelligent Load Balancing : Distribute traffic intelligently (location, weight, priority, ..) across different AI endpoints deployed in multiple regions.

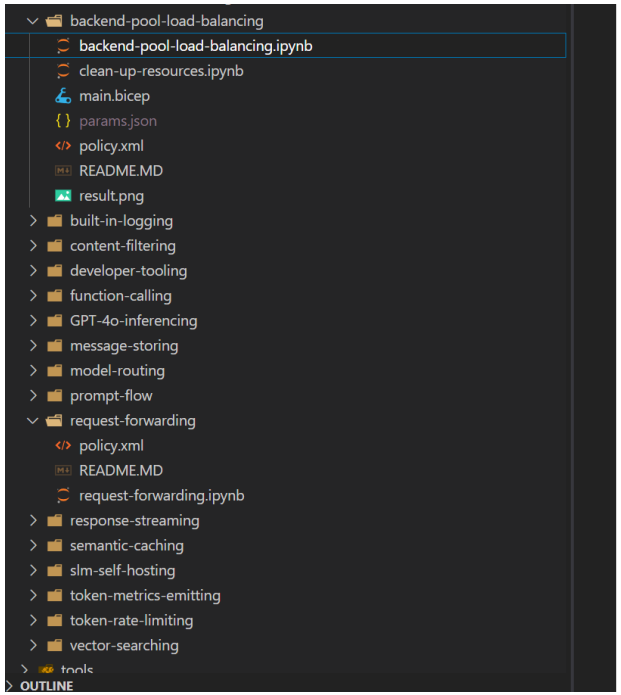
Failover and Resilience : detect a failures and automatically redirect traffic to healthy endpoints, improving your application's reliability.

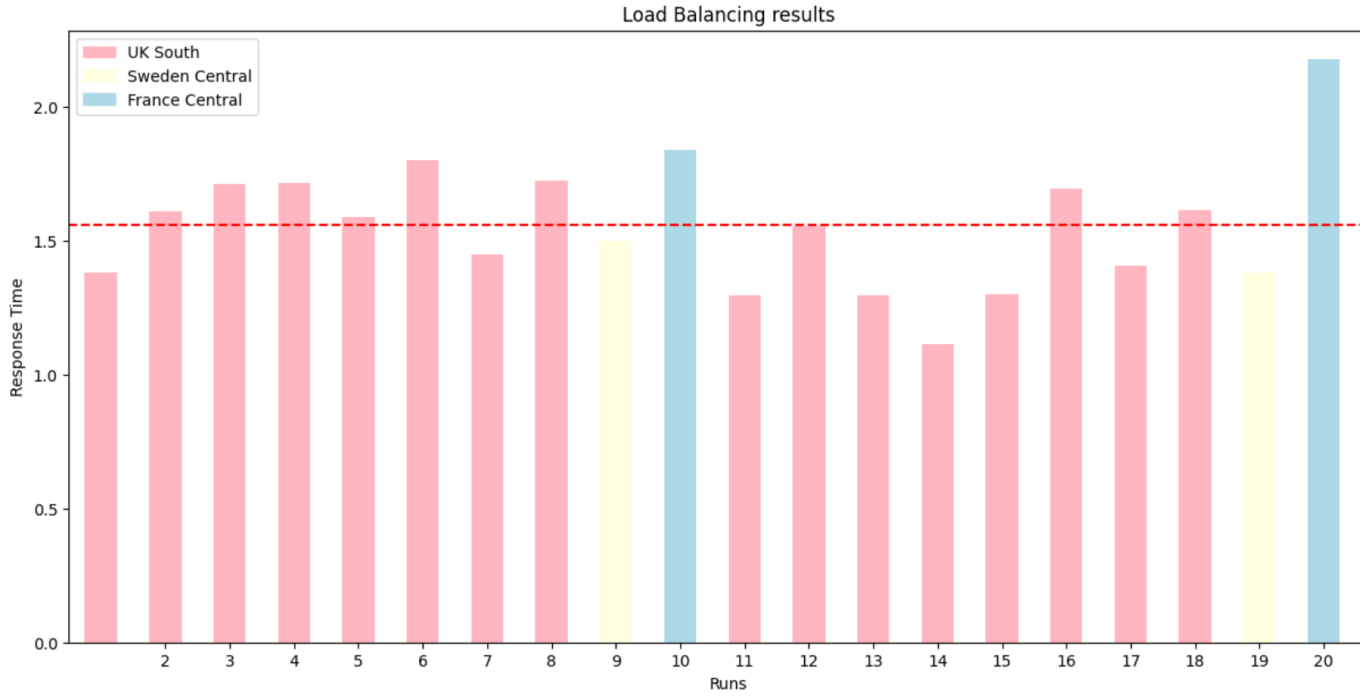
A/B Testing : seamlessly transition across multiple versions of deployed models without disrupting existing applications.

[GA] Load Balancer and Circuit Breaker



- Distribute requests across PTU and Pay-as-you-go instances
- Define load balancing pools to include multiple Azure OpenAI endpoints
- Configure circuit breaker rules forsuccessful failover






```bicep
resource backendOpenAI 'Microsoft.ApiManagement/service/backends@2023-09-01-preview' = [for (config, i) in openAIConfig: if(length(openAIConfig) > 0) {
  name: config.name
  parent: apimService
  properties: {
    description: 'backend description'
    url: '${cognitiveServices[i].properties.endpoint}openai'
    protocol: 'http'
    circuitBreaker: {
      rules: [
        {
          failureCondition: {
            count: 1
            errorReasons: [
              'Server errors'
            ]
            interval: 'PT5M'
            statusCodeRanges: [
              {
                min: 429
                max: 429
              }
            ]
          }
          name: 'openAIBreakerRule'
          tripDuration: 'PT1M'
          acceptRetryAfter: true    // respects the Retry-After header
        }
```

```bicep
resource backendPoolOpenAI 'Microsoft.ApiManagement/service/backends@2023-09-01-preview' = if(length(openAIConfig) > 1) {
  name: openAIBackendPoolName
  parent: apimService
  properties: {
    description: openAIBackendPoolDescription
    type: 'Pool'
//    protocol: 'http'  // the protocol is not needed in the Pool type
//    url: '${cognitiveServices[0].properties.endpoint}/openai'   // the url is not needed in the Pool type
    pool: {
      services: [for (config, i) in openAIConfig: {
          id: '/backends/${backendOpenAI[i].name}'
          priority: config.priority
          weight: config.weight
        }
      ]
    }
  }
}
```

codit
proximus NXT

Observability

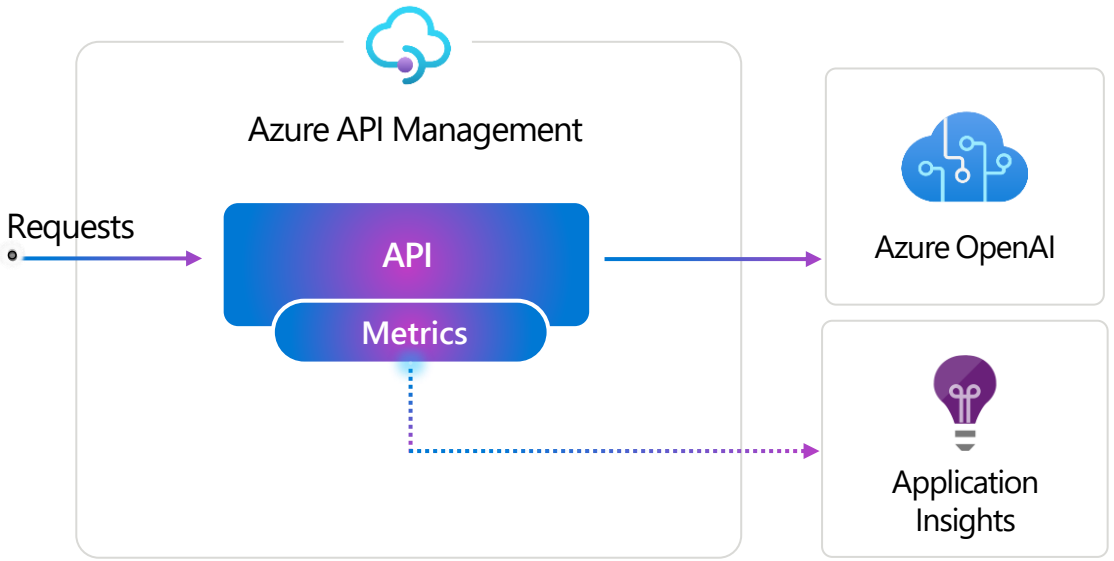
Observability

Visibility into the AI usage patterns and quotas. By proxying AI calls, we gain a unified view of usage, performance, cost, health and more.

Get data data for debugging, troubleshooting, and the metrics necessary to ensure reliability and availability.

End to end traceability for your Intelligent applications.

[GA] Azure OpenAI Token Metric policy



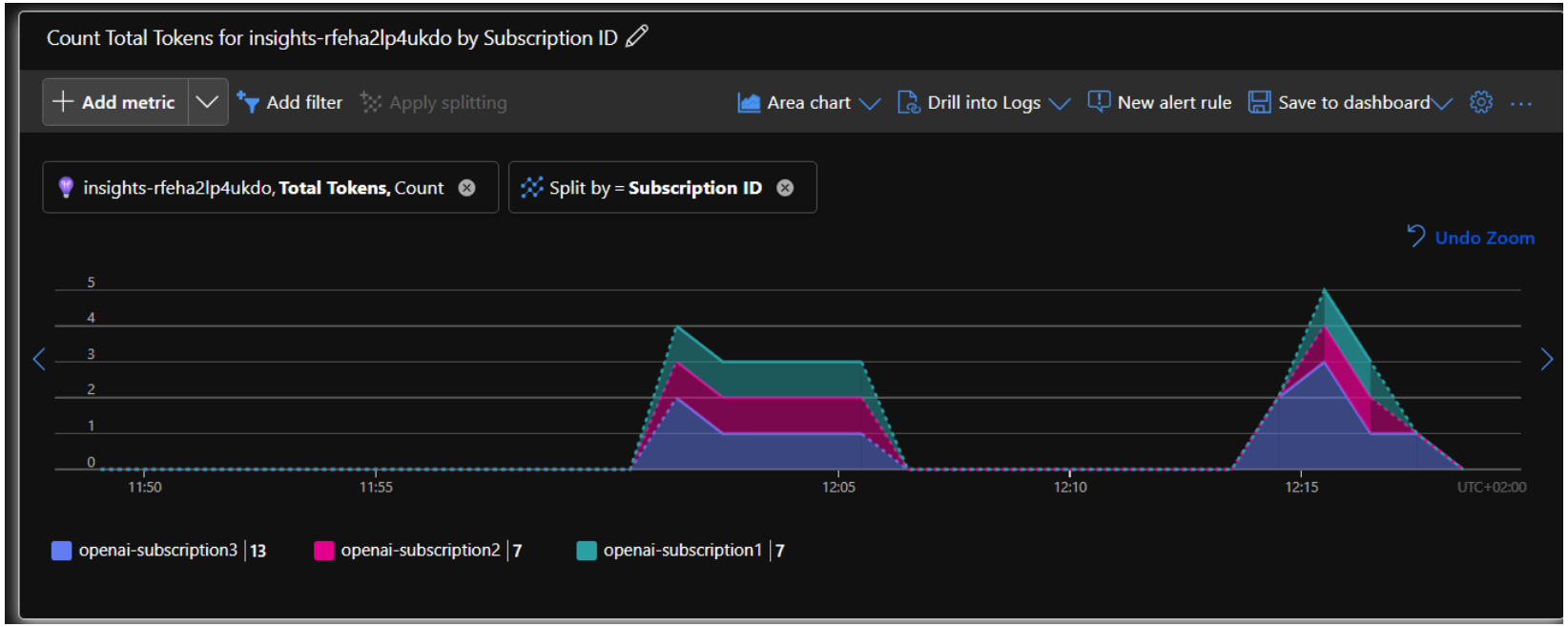
Facilitate accurate cross-charging based on token consumption

Collect token usage data

policy.xml

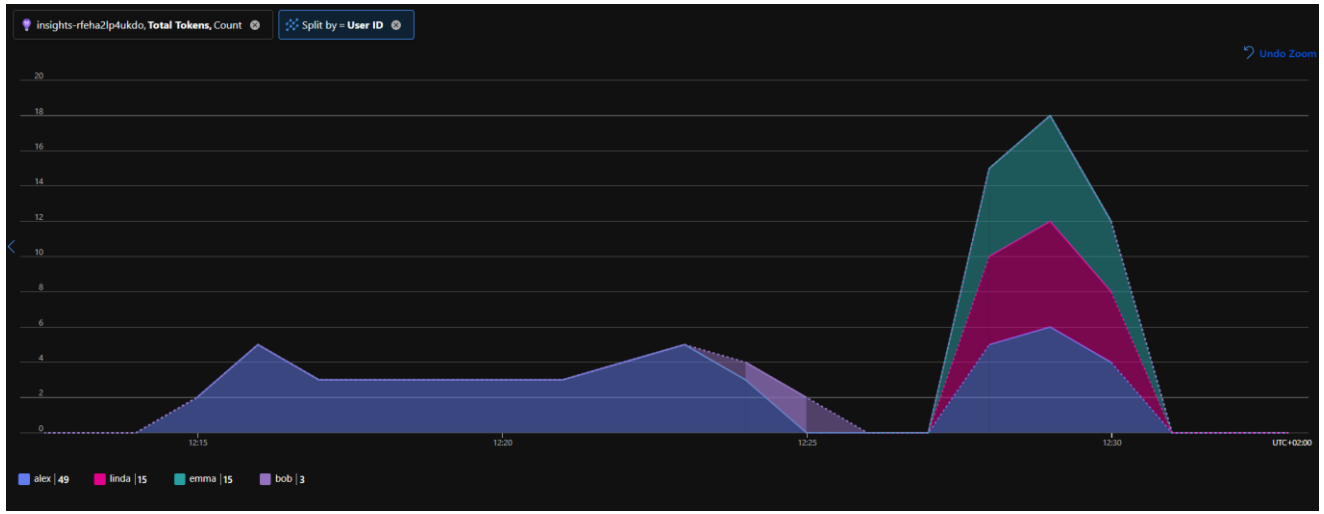
```
<azure-openai-emit-token-metric
namespace="AzureOpenAI">
        <dimension name="User ID" />
        <dimension name="Subscription ID"
/>
</azure-openai-emit-token-metric>
```

codit
proximus NXT

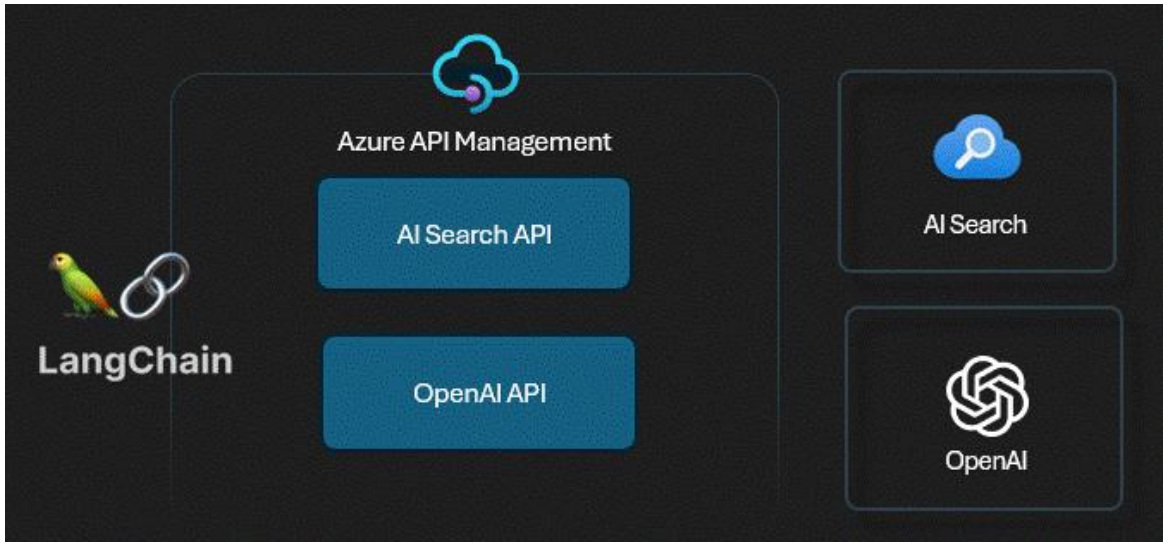


OpenAI > All operations > Policies

```xml
<policies>
    <inbound>
        <base />
        <authentication-managed-identity resource="https://cognitiveservices.azure.com" output-token-variable-name="managed-id-access-token" ignore-error="false" />
        <set-header name="Authorization" exists-action="override">
            <value>@("Bearer " + (string)context.Variables["managed-id-access-token"])</value>
        </set-header>
        <set-backend-service backend-id="openai1" />
        <azure-openai-emit-token-metric namespace="openai">
            <dimension name="Subscription ID" value="@(context.Subscription.Id)" />
            <dimension name="Client IP" value="@(context.Request.IpAddress)" />
            <dimension name="API ID" value="@(context.Api.Id)" />
            <dimension name="User ID" value="@(context.Request.Headers.GetValueOrDefault("x-user-id", "N/A"))" />
        </azure-openai-emit-token-metric>
    </inbound>
    <backend>
        <base />
    </backend>
    <outbound>
```



Explore the RAG pattern : built-in logging

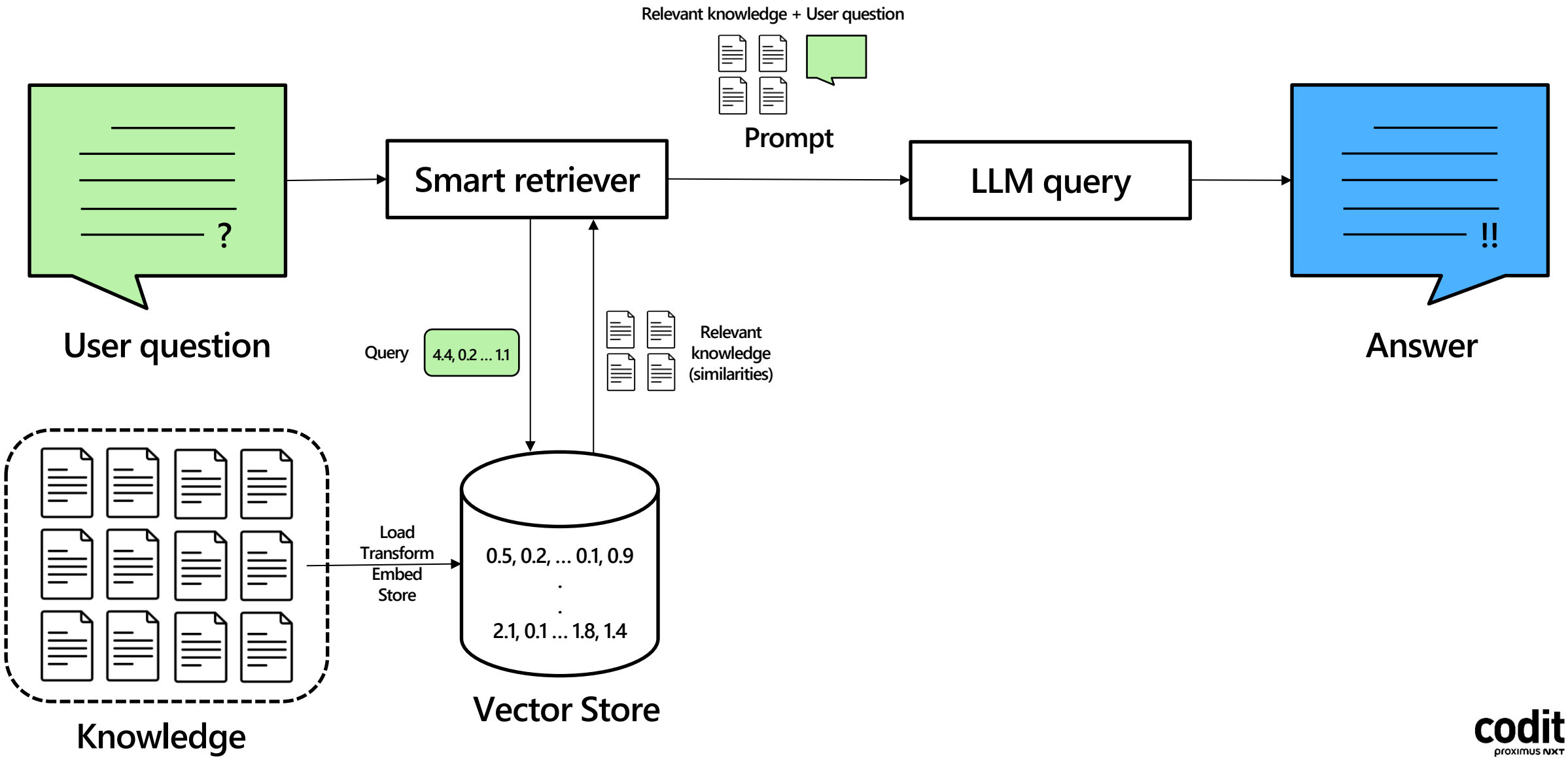


The requests are logged into Application Insights and metrics available in Azure Monitor.

Enables tracking request/response details and token usage with the provided notebook

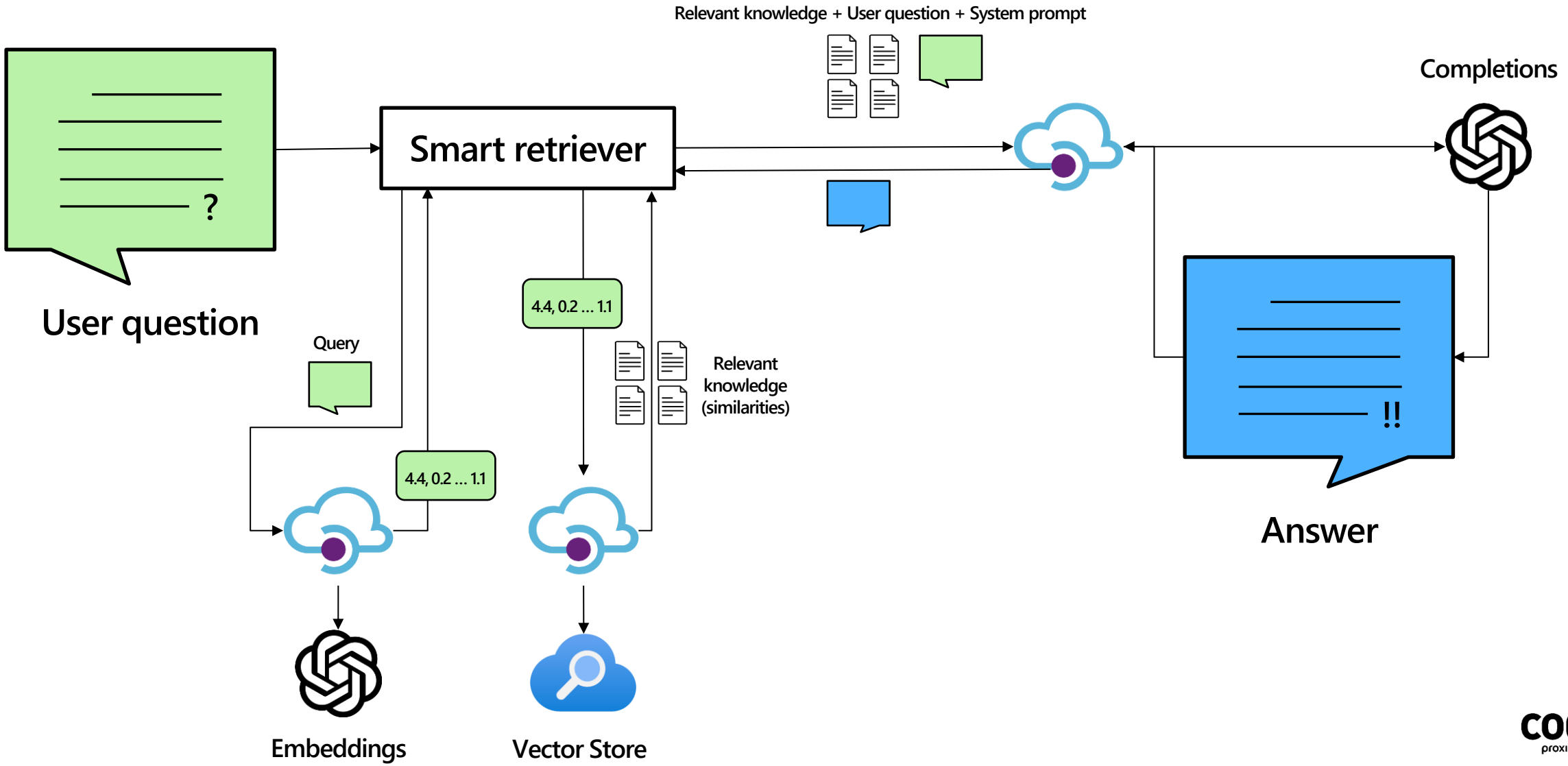
Enables the creation of Azure dashboards for a single pane of glass monitoring approach.

Explore the RAG pattern



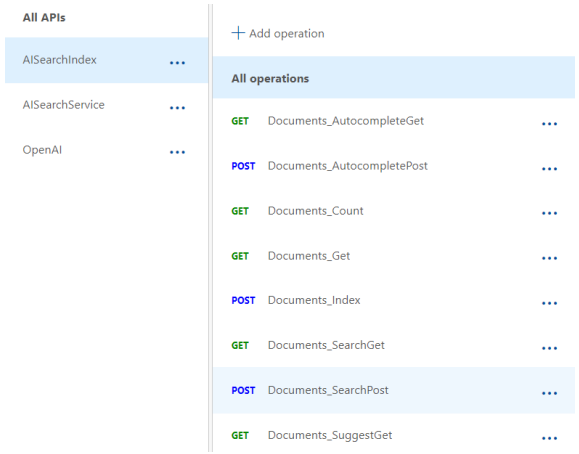
codit
proximus NXT

AI Gateway : RAG pattern



codit
proximus NXT

Explore the RAG pattern : built-in logging



Timeline | Geography | APIs | **Operations** | Products | Subscriptions | Users | Requests

API	Operation	Operation ID	Successful Requests	Failed Requests
openai	embeddings_create	/apis/openai/operations/embeddings_create	4	0
openai	ChatCompletions_Create	/apis/openai/operations/ChatCompletions_Create	4	0
searchindex	Documents_SearchPost	/apis/searchindex/operations/Documents_SearchPost	4	0

Search results | Learn more | Copy link | Feedback | Leave preview

End-to-end transaction
Operation ID: 3e96b480-6d42-4e29-8d33-e41ed7f8cf91

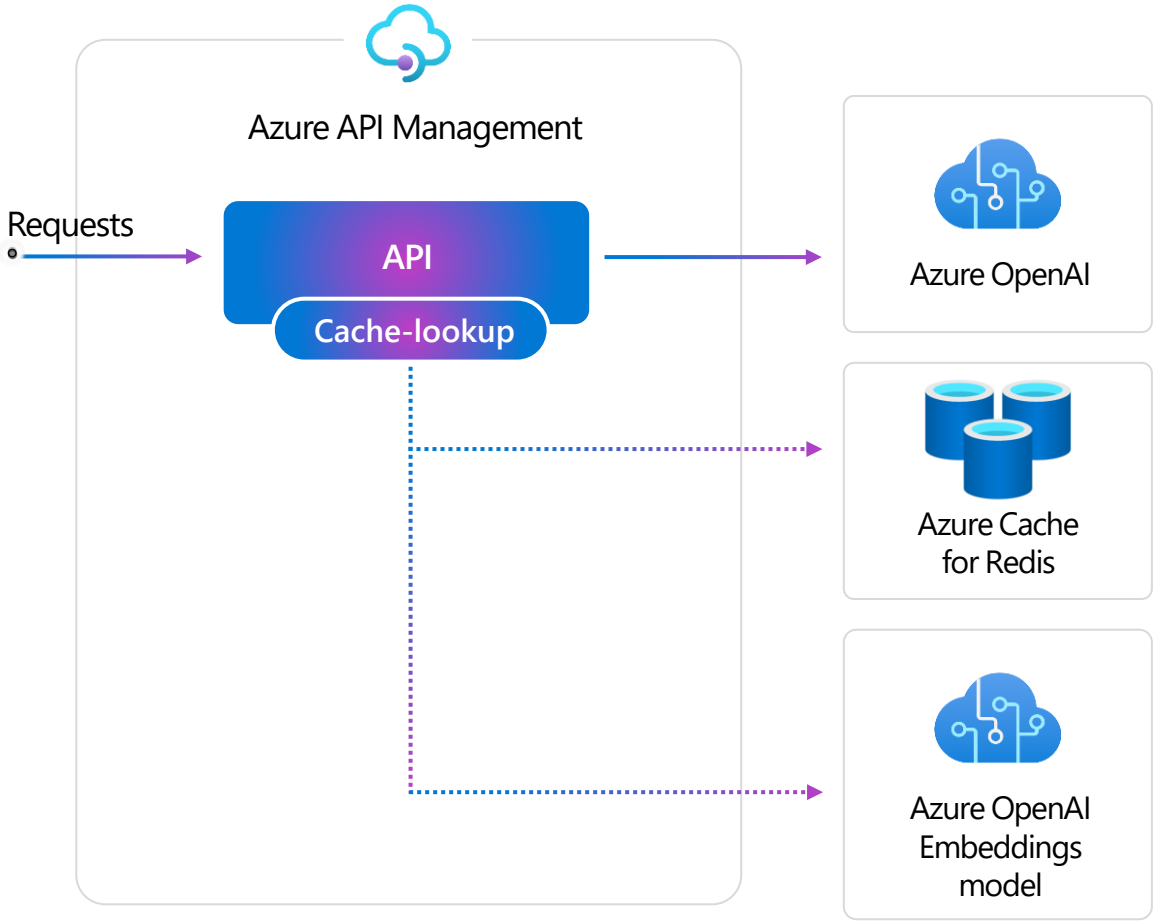
= Request (incoming) = Dependency (outgoing)

EVENT	RES.	DURATION
apim-rpyjkv2axnuju.azure-api.net POST /openai/deployments/text-embedding-ada-002/embeddings	200	300.0 ms
BACKEND openai1-rpyjkv2axnuju.openai.azure.com POST //openai/deployments/text-embedding-ada-002/embeddings	200	298 ms
apim-rpyjkv2axnuju.azure-api.net POST /searchindex/indexes('movies')/docs/search.post.search	200	168.2 ms
BACKEND search-rpyjkv2axnuju.search.windows.net POST /indexes('movies')/docs/search.post.search	200	166 ms
apim-rpyjkv2axnuju.azure-api.net POST /openai/deployments/gpt-35-turbo/chat/completions	200	3.5 s
BACKEND openai1-rpyjkv2axnuju.openai.azure.com POST //openai/deployments/gpt-35-turbo/chat/completions	200	3.5 s

codit
proximus NXT

... many more

[Preview] Azure OpenAI Semantic Caching policy



Configure semantic caching for all API consumers

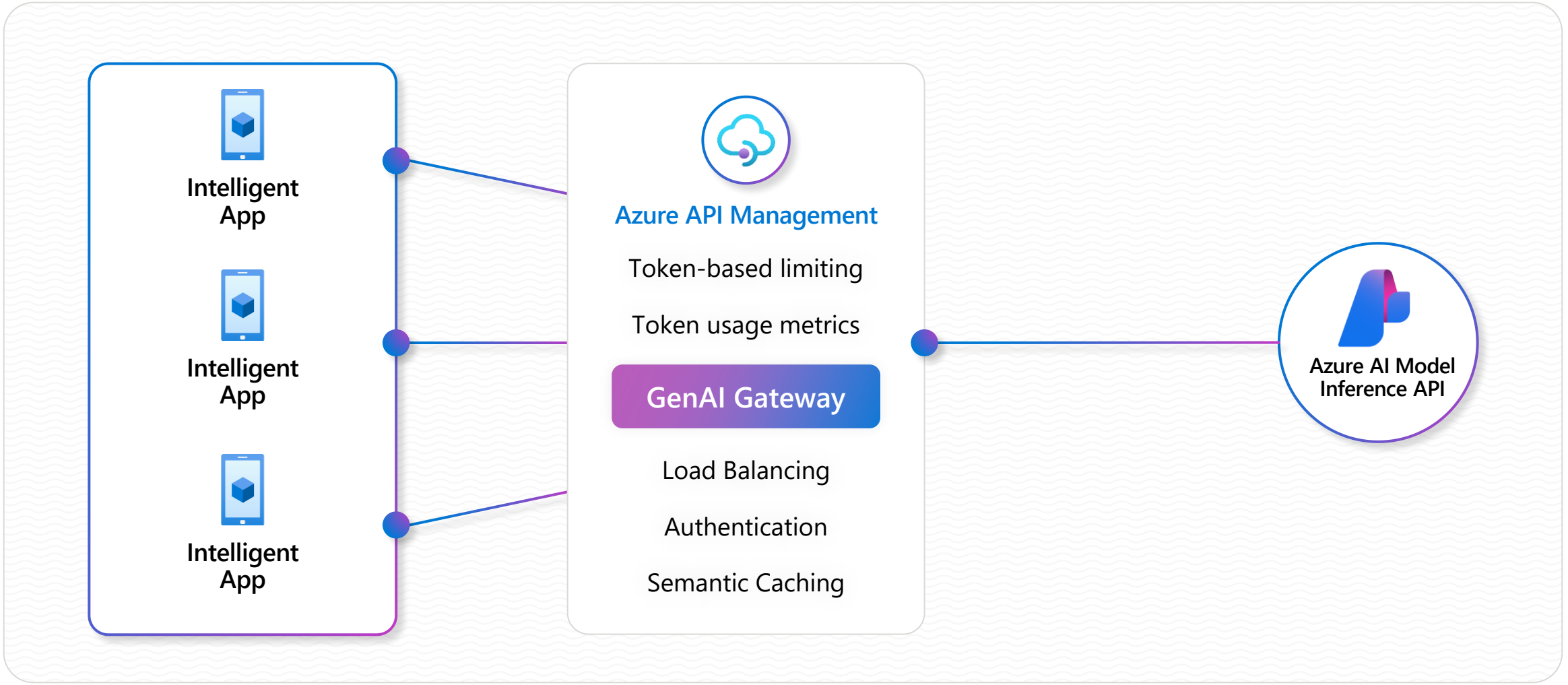
Define similarity score threshold for caching

policy.xml

```
<azure-openai-semantic-cache-lookup
  score-threshold="0.05"
  embeddings-backend-id="azure-openai-backend">
  <vary-by>@(context.Subscription.Id)"</vary-by>
</azure-openai-semantic-cache-lookup>
```

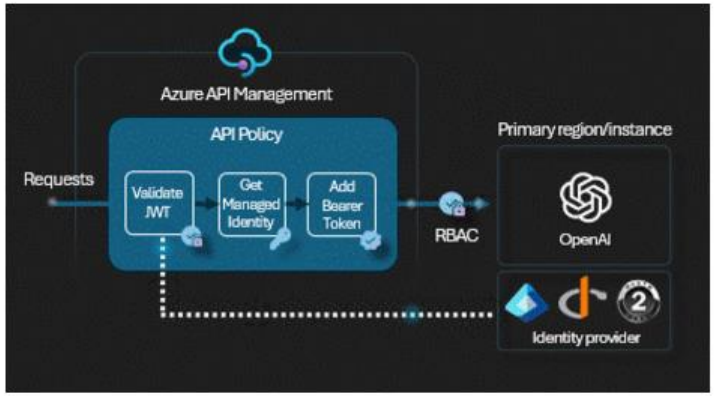
codit
proximus NXT

[Preview] Azure AI Model Inference API Support



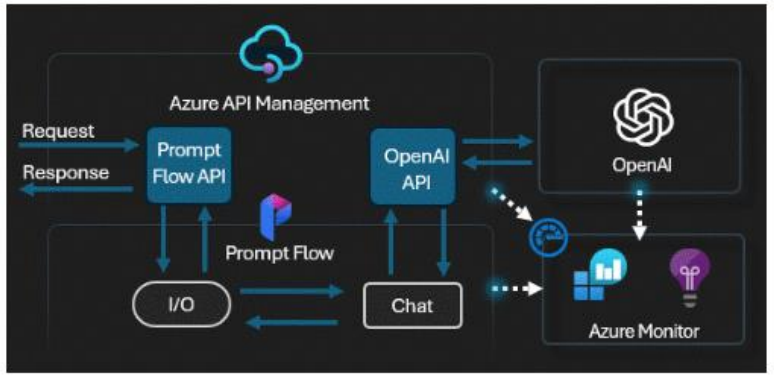
… many more

Access controlling



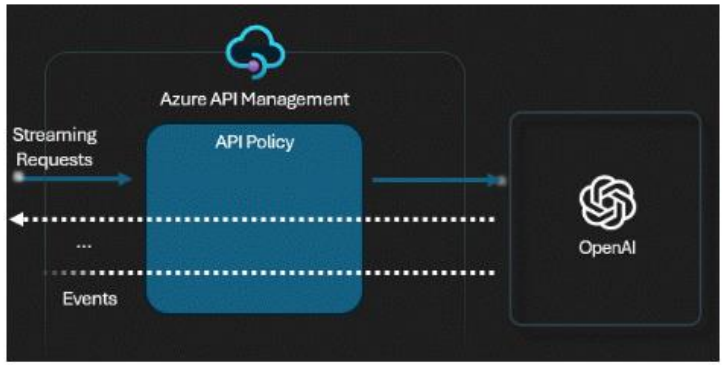
Playground to try the OAuth 2.0 authorization feature using identity provider to enable more fine-grained access to OpenAPI APIs by particular users or client.

Prompt flow



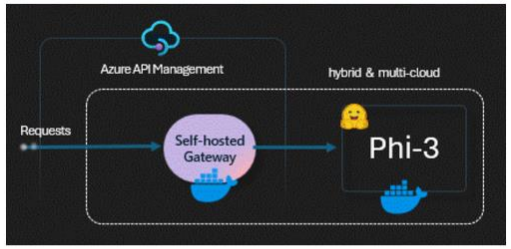
Playground to try the Azure AI Studio Prompt Flow with Azure API Management.

Response streaming



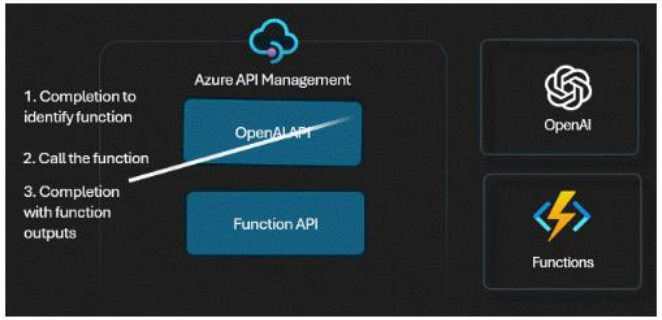
Playground to try response streaming with Azure API Management and Azure OpenAI endpoints to explore the advantages and shortcomings associated with streaming.

SLM self-hosting (phy-3)



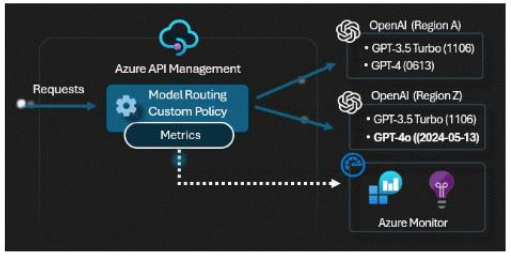
Playground to try the self-hosted phy-3 Small Language Model (SLM) trough the Azure API Management self-hosted gateway with OpenAI API compatibility.

Function calling



Playground to try the OpenAI function calling feature with an Azure Functions API that is also managed by Azure API Management.

Model Routing



Playground to try routing to a backend based on Azure OpenAI model and version.

codit
proximus NXT

Get Started with Azure API Management and OpenAI

Documentation
https://aka.ms/apim/openai-docs

AI Hub Gateway Accelerator
https://aka.ms/ai-hub-gateway

GenAI Gateway Labs
https://aka.ms/apim/genai/labs

GenAI Gateway Accelerator
https://aka.ms/apim-genai-lza

codit
proximus NXT


codit
proximus NXT
Thank you