Semantic Caching and Token Rate Limiting with Azure API Management

Jon Butler | Solutions Engineer | Microsoft

Microsoft

What is Azure API Management?

A fully managed service that enables customers to publish, secure, transform, maintain, and monitor APIs.

Key Features:

- Comprehensive API platform for different stakeholders and teams
- Abstract backend architecture diversity and complexity from API consumers
- Securely expose services hosted on and outside of Azure as APIs
- Scaling and performance
- Enable API discovery and consumption by internal and external users

API Management Components

API Gateway

- Routes API calls
- Verifies credentials
- Enforces rate limits
- Transforms requests
- Caches responses
- Emits logs & metrics

Management Plane

- Service configuration
- API schemas import
- Package APIs
- Set up policies
- Analytics insights
- User management

Developer Portal

- API documentation
- Interactive console
- Account creation
- API key management
- Usage analytics
- API definitions

Challenges in Managing Generative AI APIs

Token Cost Management

Track and allocate TPM quotas across multiple apps

Fair Resource Distribution

Prevent single apps from consuming entire quotas

Security & Key Management

Securely distribute API keys across applications

Response Latency

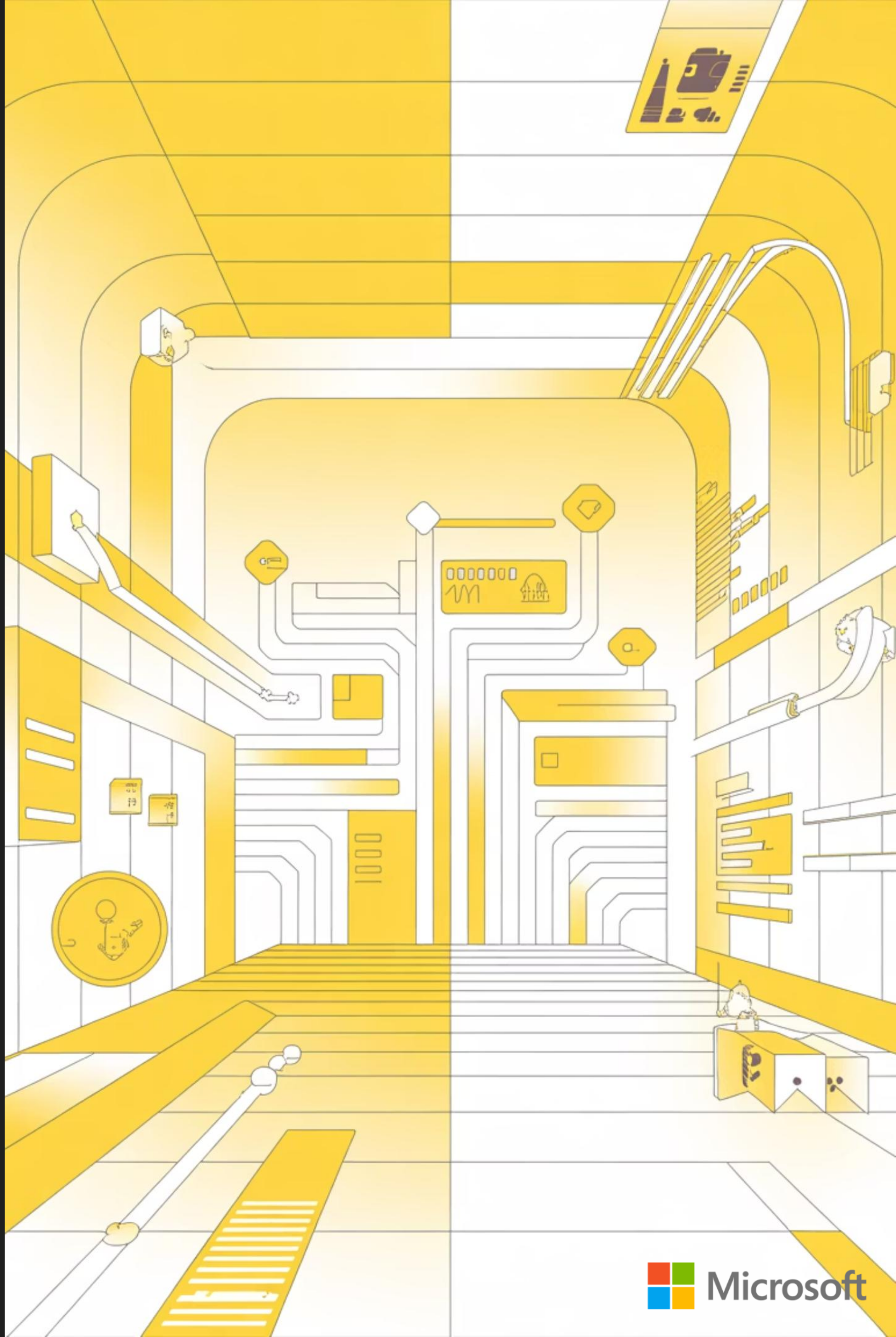
Minimize latency for similar or repeated prompts

APIM - AI Gateway Features

- Token Rate limiting and quotas
- Semantic caching
- Security and safety
- Observability and governance
- Multi-cloud model management

MCP Servers

- Expose existing REST APIs as MCP servers
- MCP Server Pass Thru (Proxy)



Semantic Caching with Azure API Management

Stores and retrieves LLM responses based on the meaning of prompts, using vector similarity to identify semantically equivalent queries

Benefits:

Reduced Latency

Token Savings

Throughput Increase

What are Vector Embeddings?

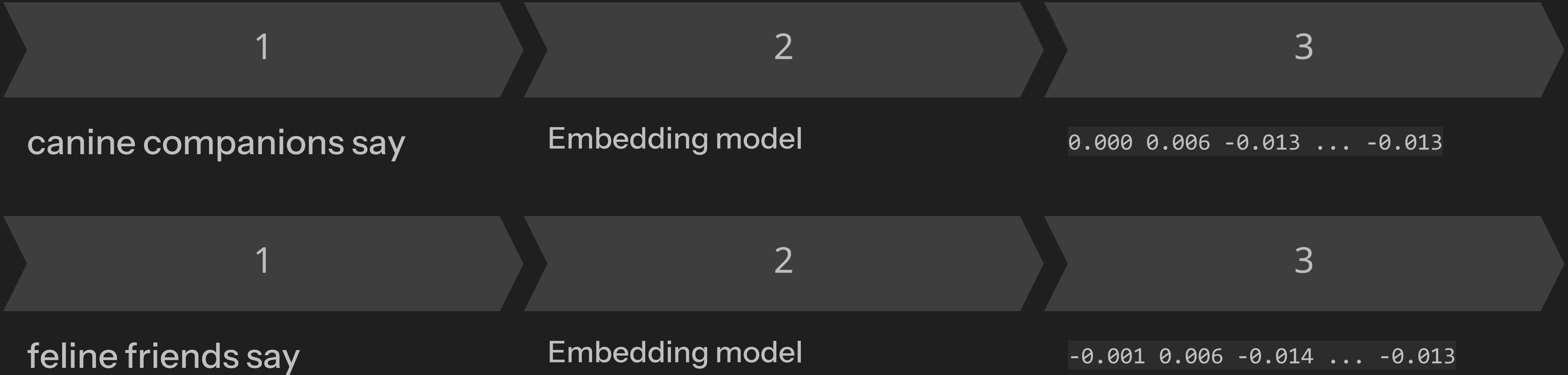
Vector embeddings are a way of representing words, sentences, images, or other data as numerical vectors in a high-dimensional space.

Key Characteristics:

- Numerical representations of data
- Capture semantic meaning
- Similar items cluster together
- Generated by Azure OpenAI text-embedding models

Microsoft

How Vector Embeddings Work



1

canine companions say

2

Embedding model

3

`0.000 0.006 -0.013 ... -0.013`

1

feline friends say

2

Embedding model

3

`-0.001 0.006 -0.014 ... -0.013`

Semantic Similarity: Notice how similar phrases have similar vector values!

What is Semantic Search?

A search approach that uses vector embeddings to understand the meaning behind your words to return results that match what you meant, not just what you typed.

Traditional Keyword Search

Query: "Retrieve all pictures of hot dogs"

✓ Returns a picture of a hot dog on a grill

✗ Returns a picture of puppy that is hot out of breath

✗ Returns a picture of a dog outside on a hot sunny day

Semantic Search

Query: "Retrieve all pictures of hot dogs"

✓ Returns a picture of a hot dog on a grill

✓ Returns a picture a hot dog on a bun

✓ Returns a picture of a hot dogs at a baseball game

Benefits: Intent Understanding ▪ Synonym Recognition ▪ Context Aware

Azure Native Vector Storage Options

Azure Cognitive (AI) Search

A fully managed search service that supports vector embeddings, enabling hybrid search that combines keyword retrieval with semantic

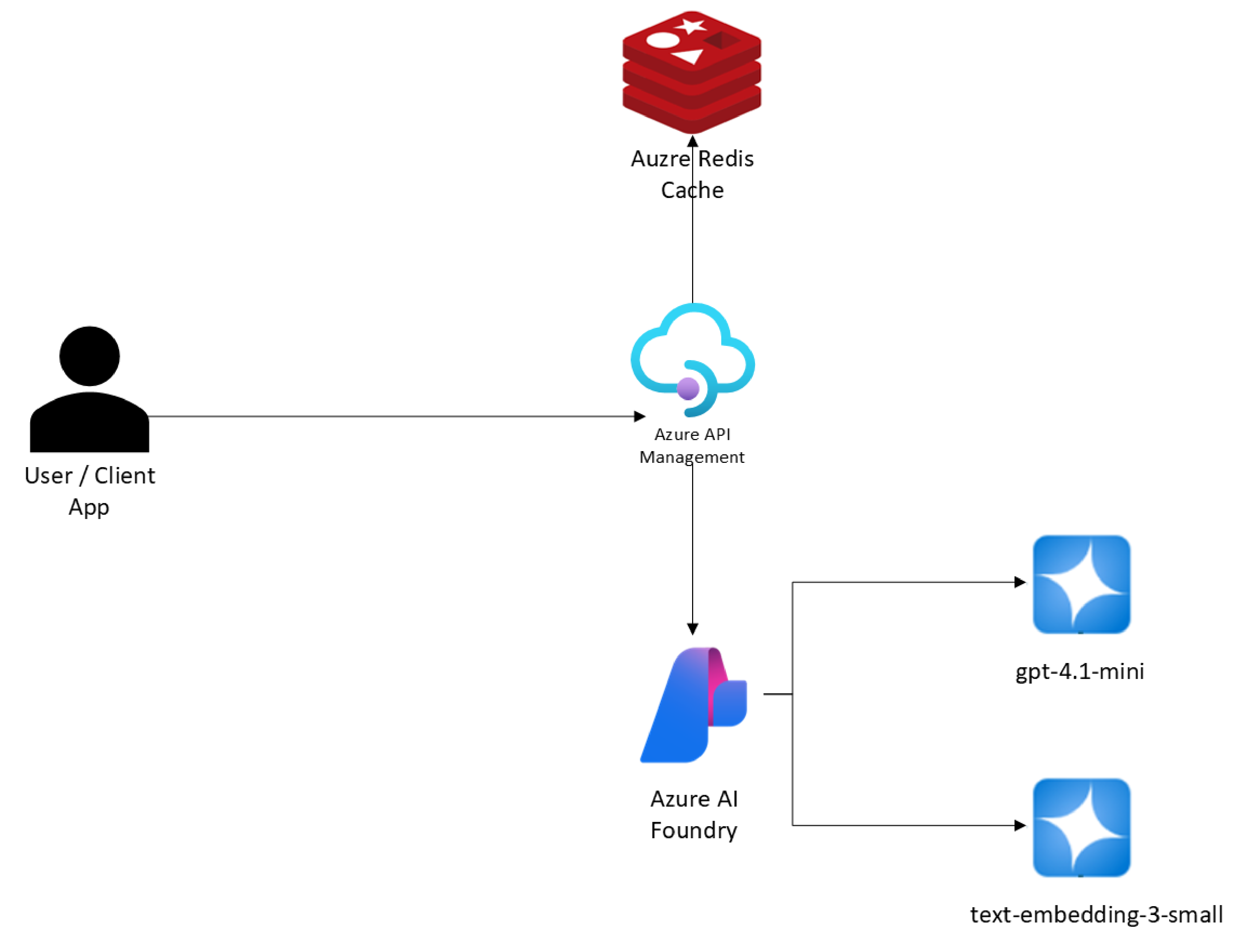
Azure Cosmos DB

A globally distributed NoSQL database capable of storing high-dimensional vector data alongside operational data, making it suitable for apps that need low-latency

Azure Redis Cache

An in-memory cache augmented with vector similarity modules, allowing extremely fast vector comparisons for real-time semantic caching.

Demo Architecture – Semantic Caching



Semantic Caching Workflow

Semantic Caching

Demo

Azure APIM - Token Rate Limiting & Quotas

Manage and enforce token limits per API consumer to prevent abuse and ensure fair distribution

Key Features:

✓ Tokens-per-minute (TPM) limits

✓ Hourly, daily, weekly, monthly, yearly quotas

✓ Per subscription/IP/custom key

Benefits:

- Prevents token quota exhaustion by individual applications
- Enables fair resource distribution across teams and applications
- Supports chargeback scenarios with per-subscription tracking
- Prevents unnecessary backend calls

Azure APIM - Token Usage Metrics & Monitoring

Send token consumption metrics to Application Insights for monitoring and chargeback

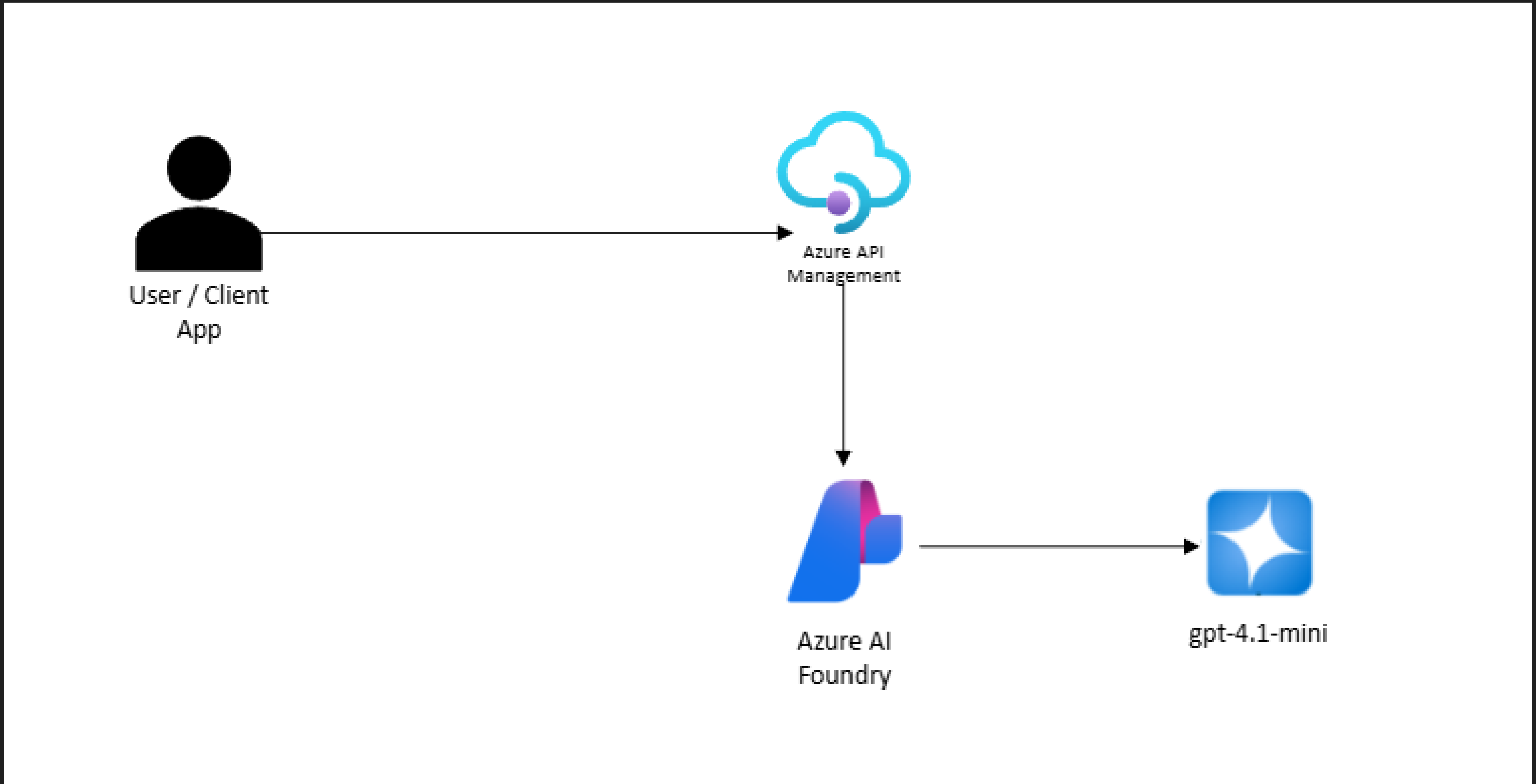
Metrics Captured:

- Prompt tokens consumed (input)
- Completion tokens generated (output)
- Total token usage
- Model utilization

Use Cases:

- Chargeback
- Capacity Planning
- Usage Analysis
- Cost Optimization

Demo Architecture – Token Limiting



Token Rate Limiting

Demo

Questions?

Microsoft

Thank You!

Jon Butler | Solutions Engineer | Microsoft

Learn More:

•APIM Docs • APIM Policies • APIM AI Gateway Docs• APIM AI Gateway GitHub

Microsoft