



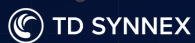


#GlobalAzureTorino



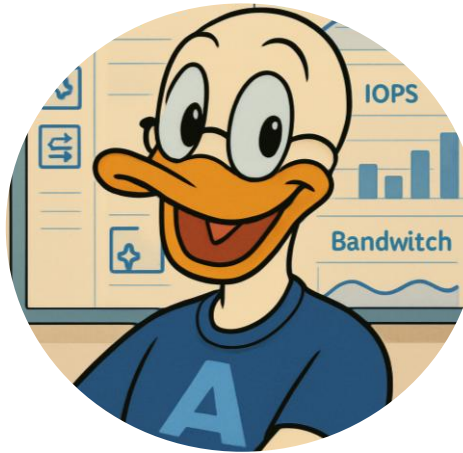
Deploying and Protecting LLM Models at Scale with Azure API Management

Mattia Contessa, Fabio Cannas





Who we are?



Mattia Contessa
Cloud Architect
@Alveo Expertise



Fabio Cannas
Cloud Engineer
@Alveo Expertise

Agenda

- Intro
- Azure API Management (aka APIM) in short
- AI Gateway overview + DEMO
- MCP overview
- Remote MCP Servers & AI Gateway + DEMO

#GlobalAzureTorino

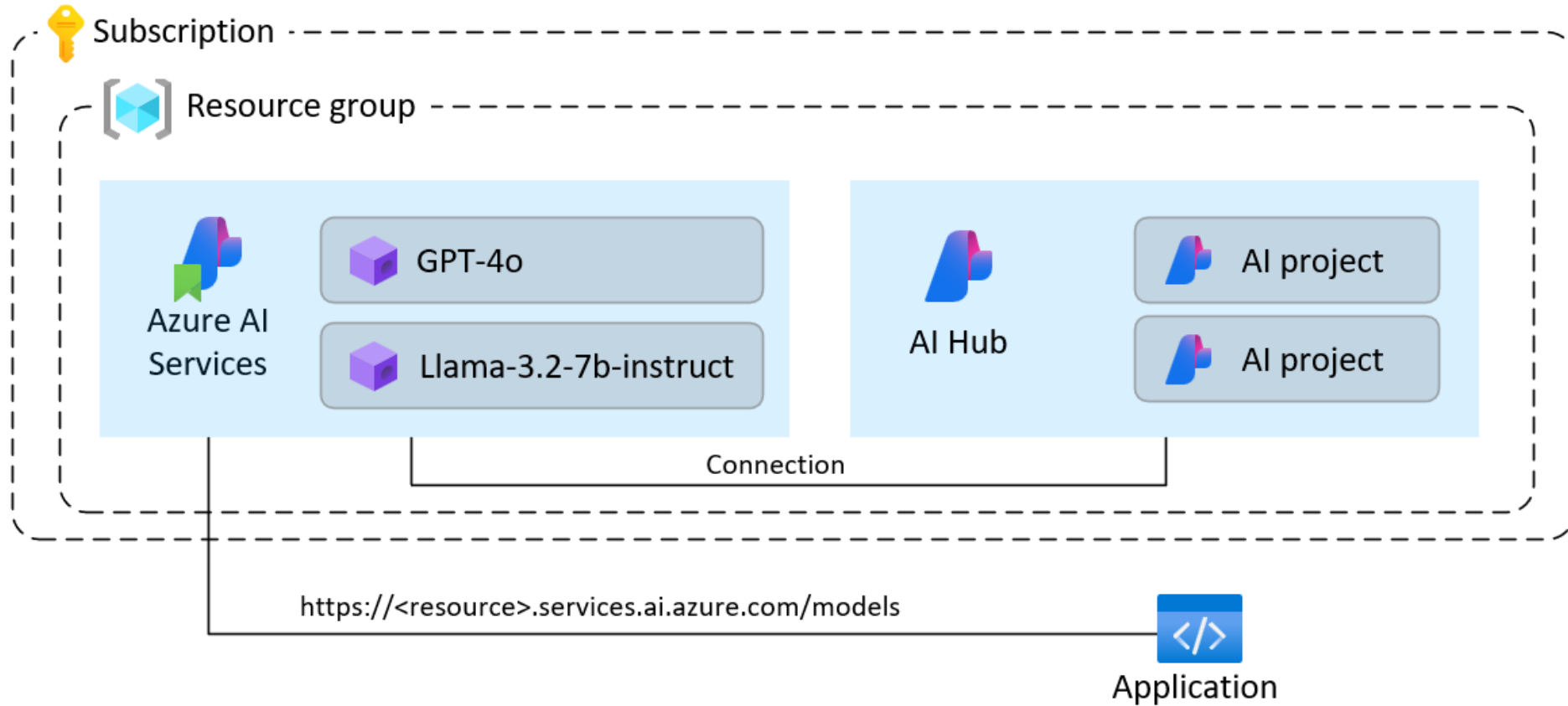


AI Gateway

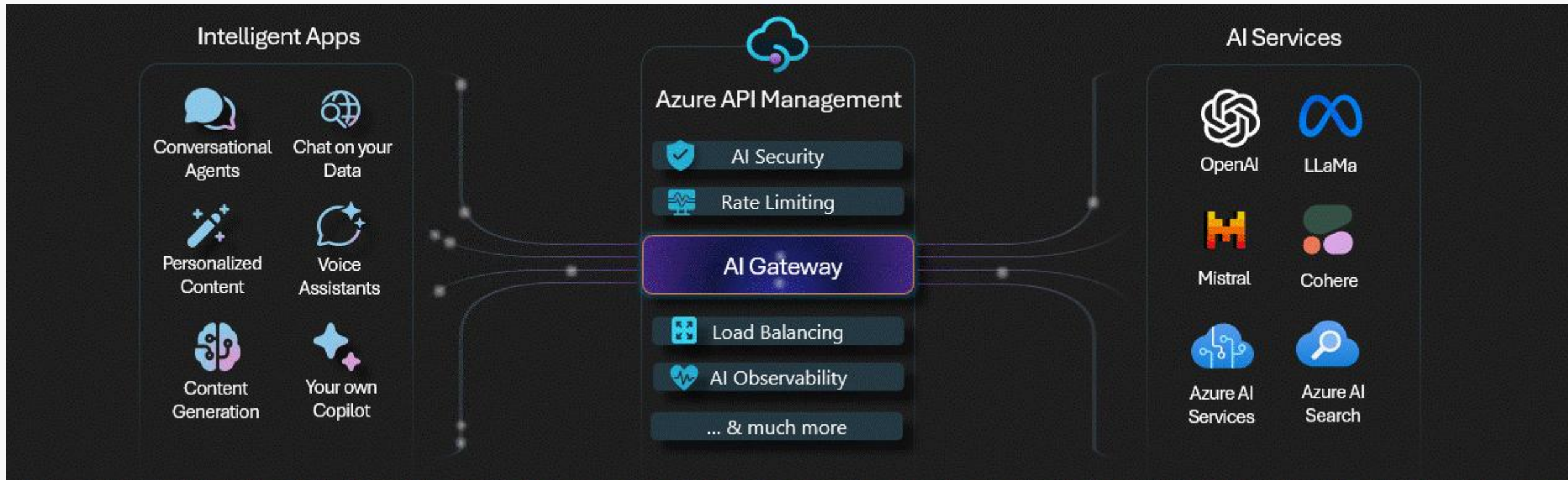
Do we really need an AI Gateway?



Integrating LLMs with Azure



AI Gateway Key Features

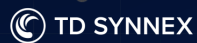


#GlobalAzureTorino



Azure API Management

Overview



What is Azure API Management?

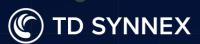
Azure API Management (APIM) is a service that enables organizations to publish, secure, transform, maintain and monitor APIs.

Key features:

- API Gateway
- Security and Rate Limiting
- Analytics
- Developer Portal



DEMO





Azure AI model inference

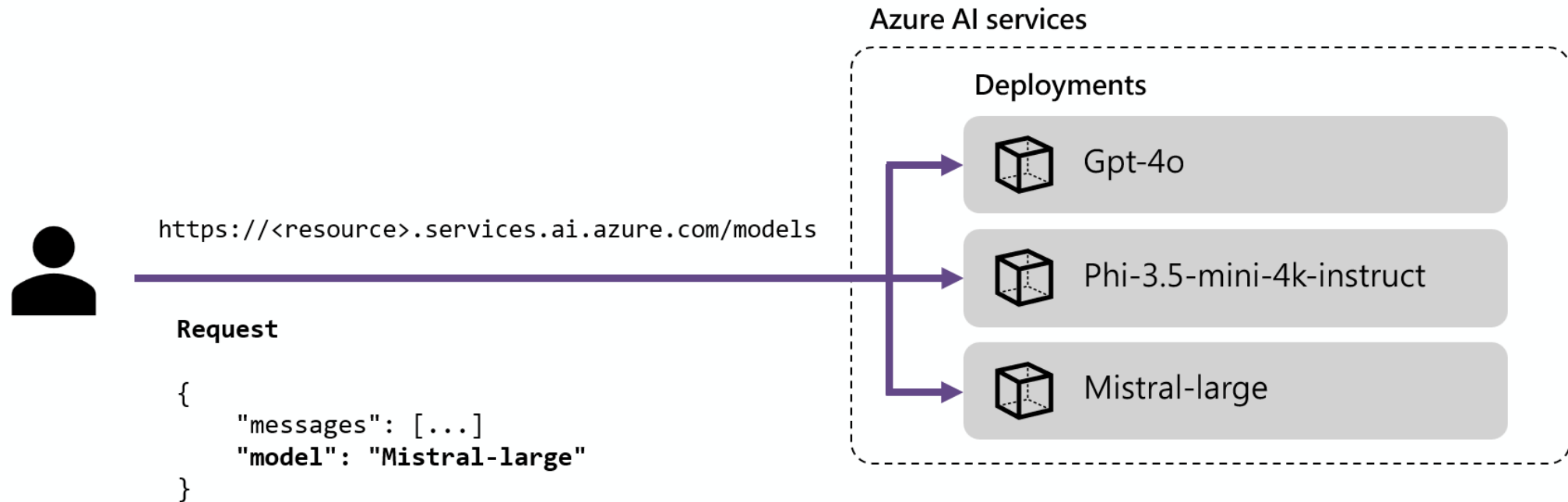
Azure AI model inference provides a way to consume models as APIs **without hosting them on your infrastructure.**

Models are hosted in a Microsoft-managed infrastructure, which enables API-based access to the model provider's model.

Models:

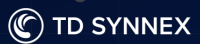
- AI21 Labs
- Azure OpenAI
- Cohere
- Core42
- DeepSeek
- Meta
- Microsoft
- Mistral AI
- NTT Data

Azure AI model inference

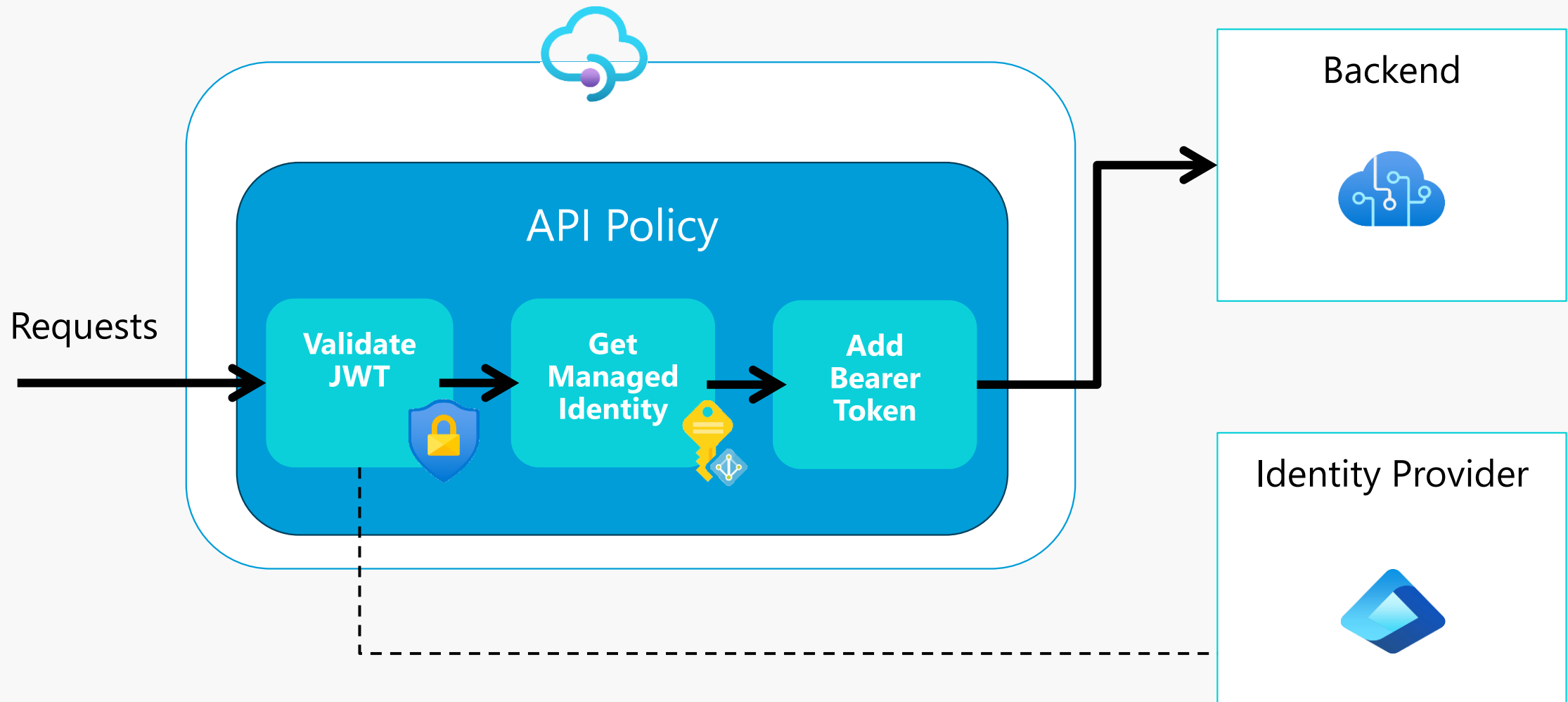




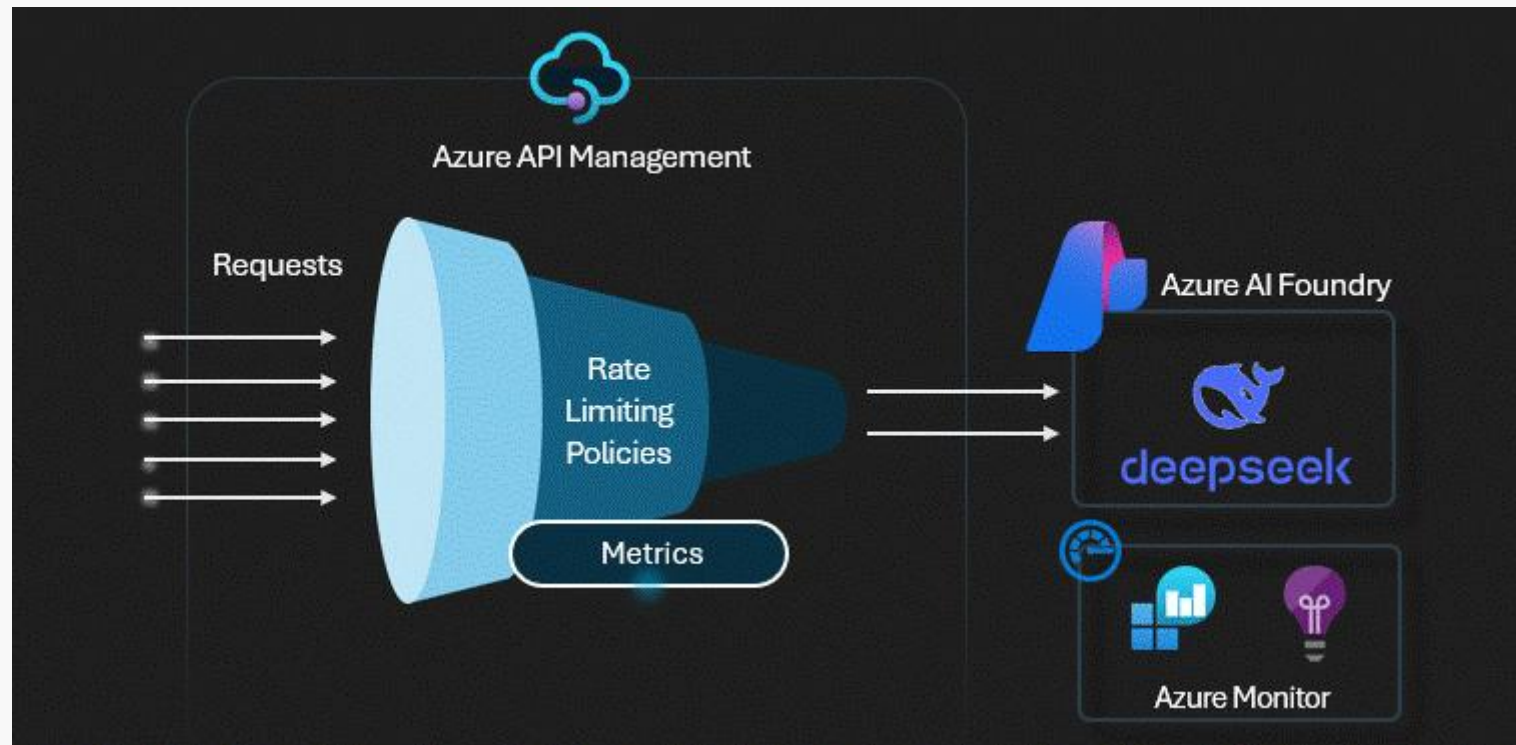
DEMO



Access Control

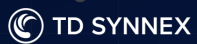


Rate Limiting and Monitoring

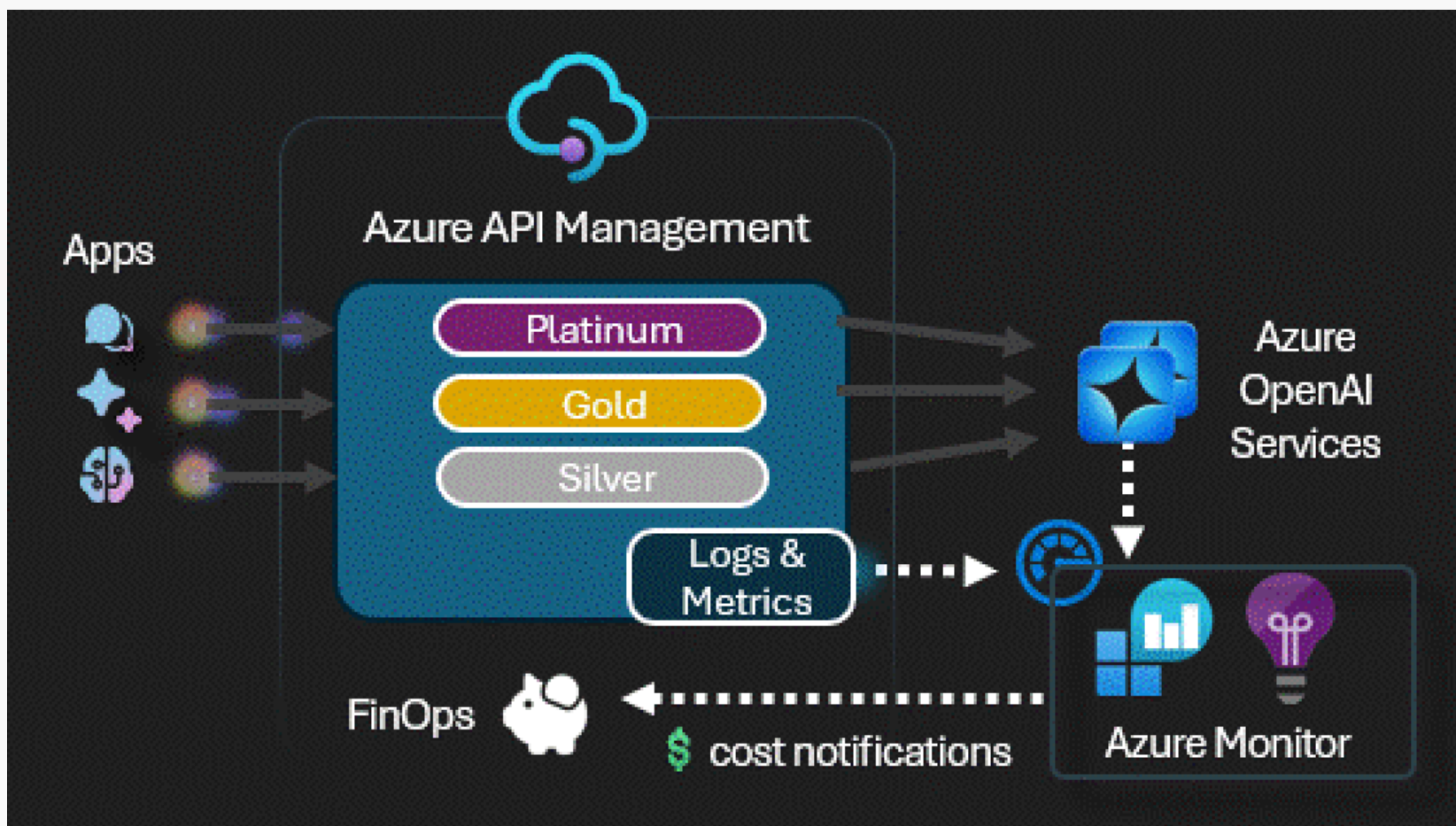




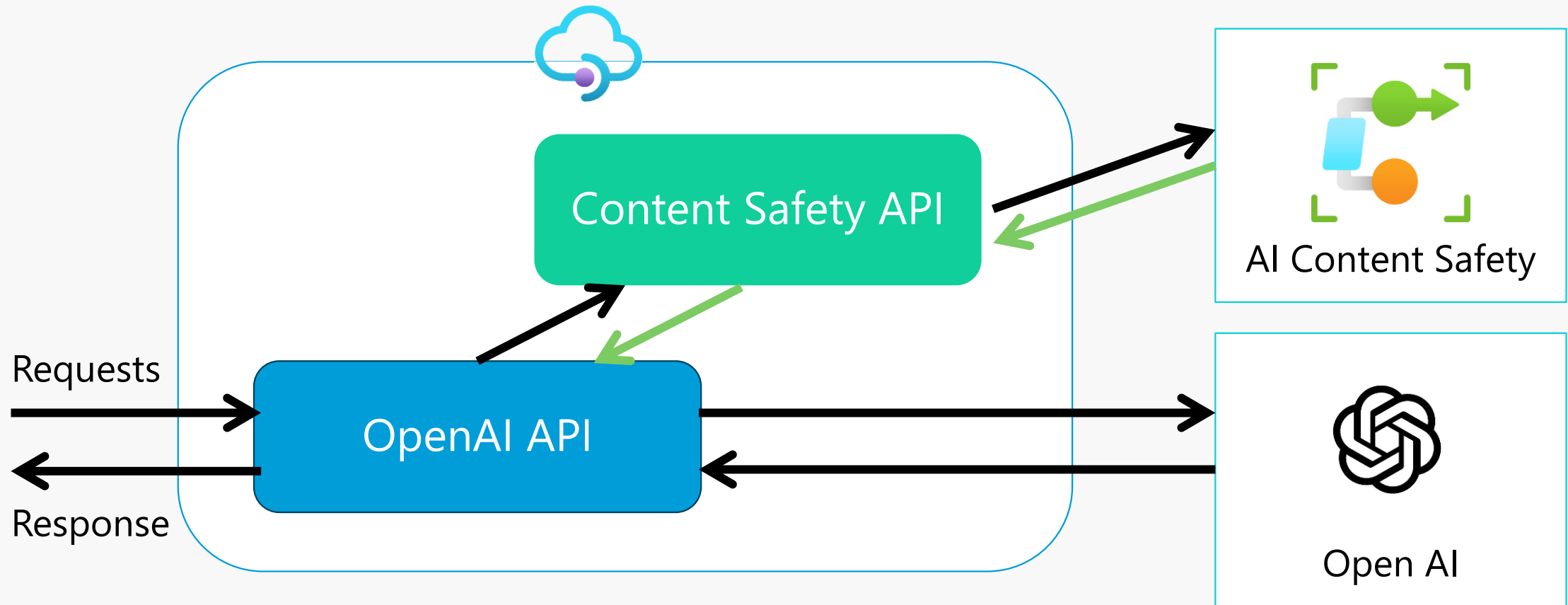
DEMO



FinOps



Manage Content

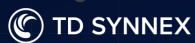


#GlobalAzureTorino

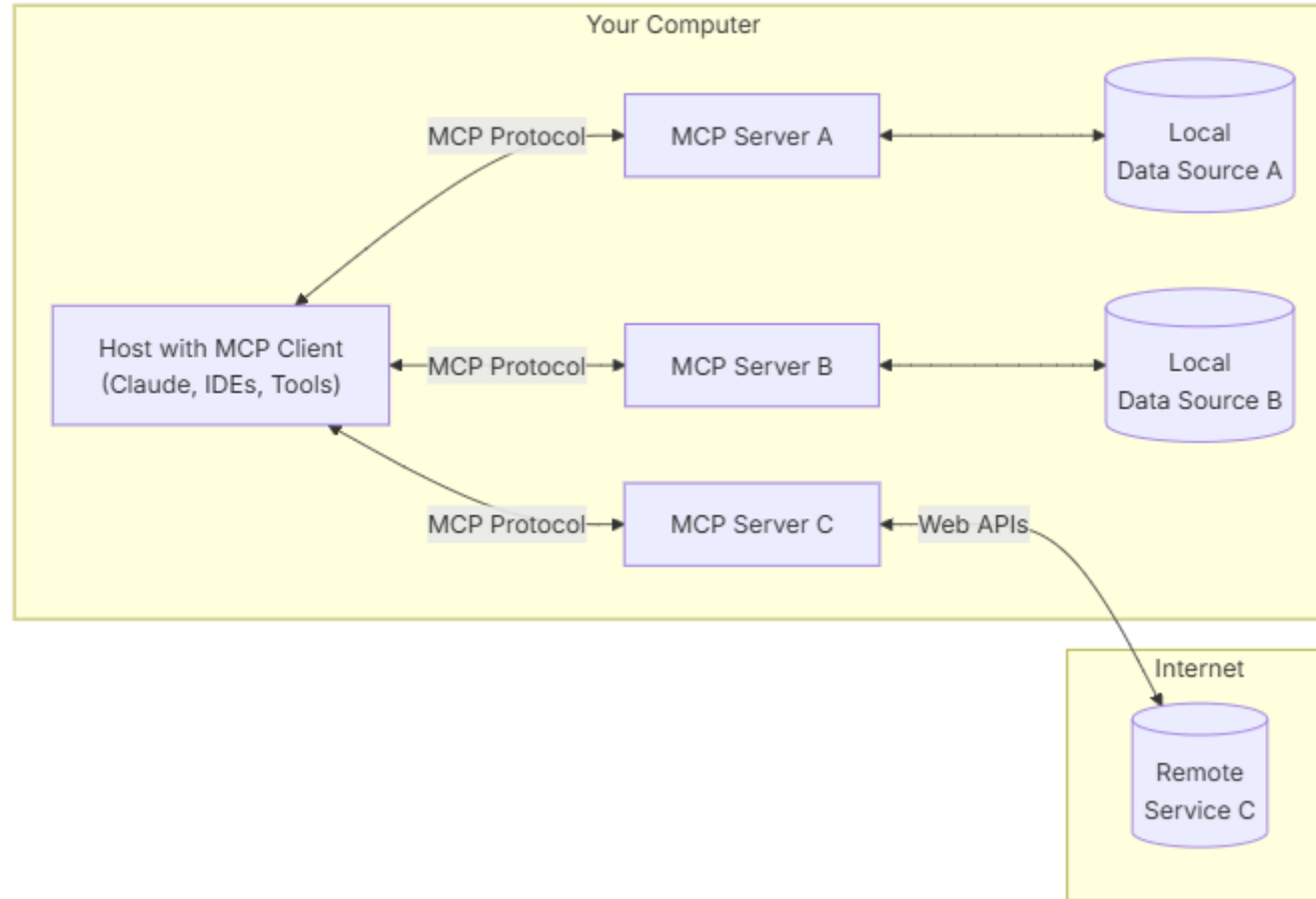


Model Context Protocol

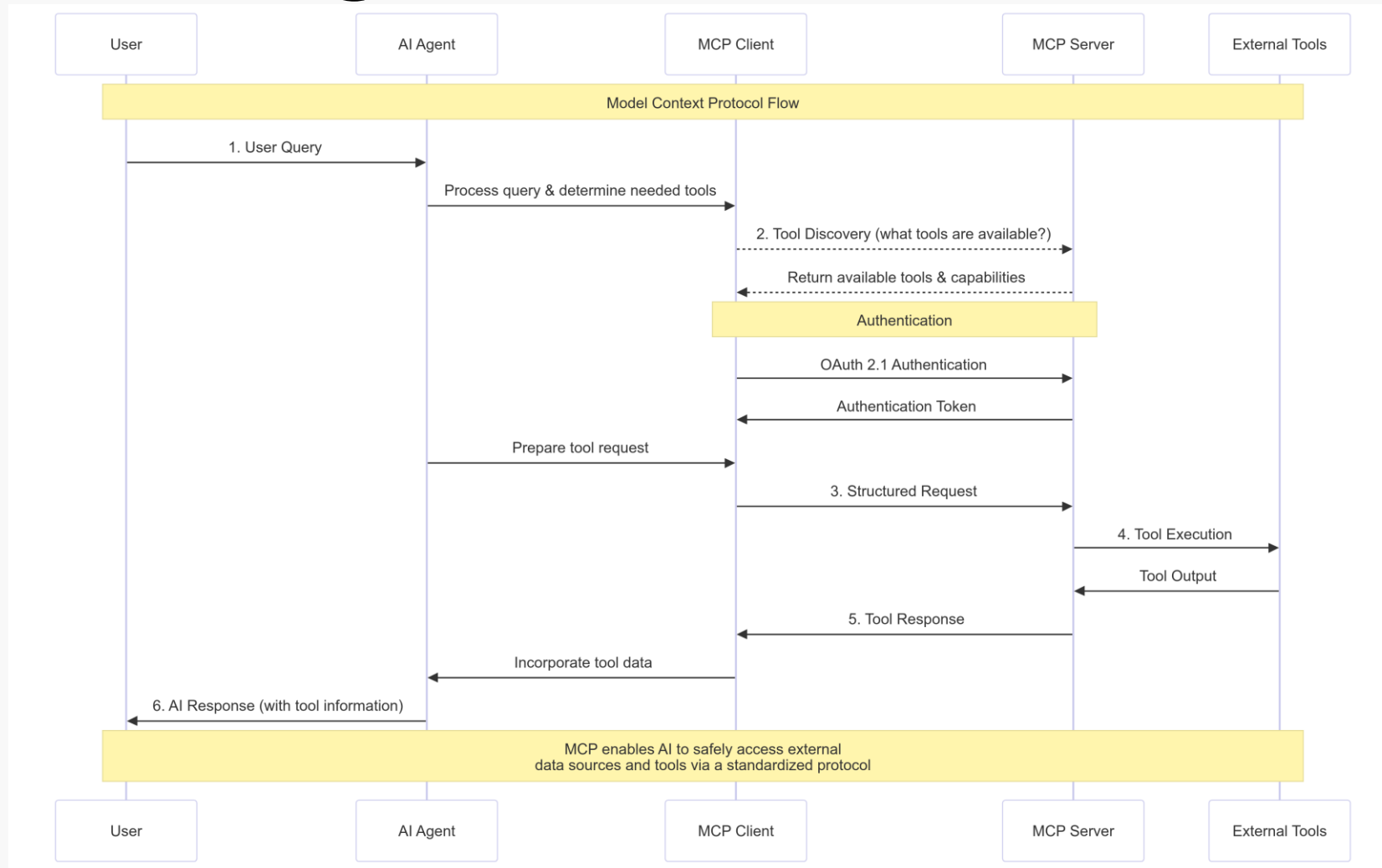
Extending AI models capabilities



MCP Architecture I



MCP Flow Diagram





MCP and AI Gateway Integration

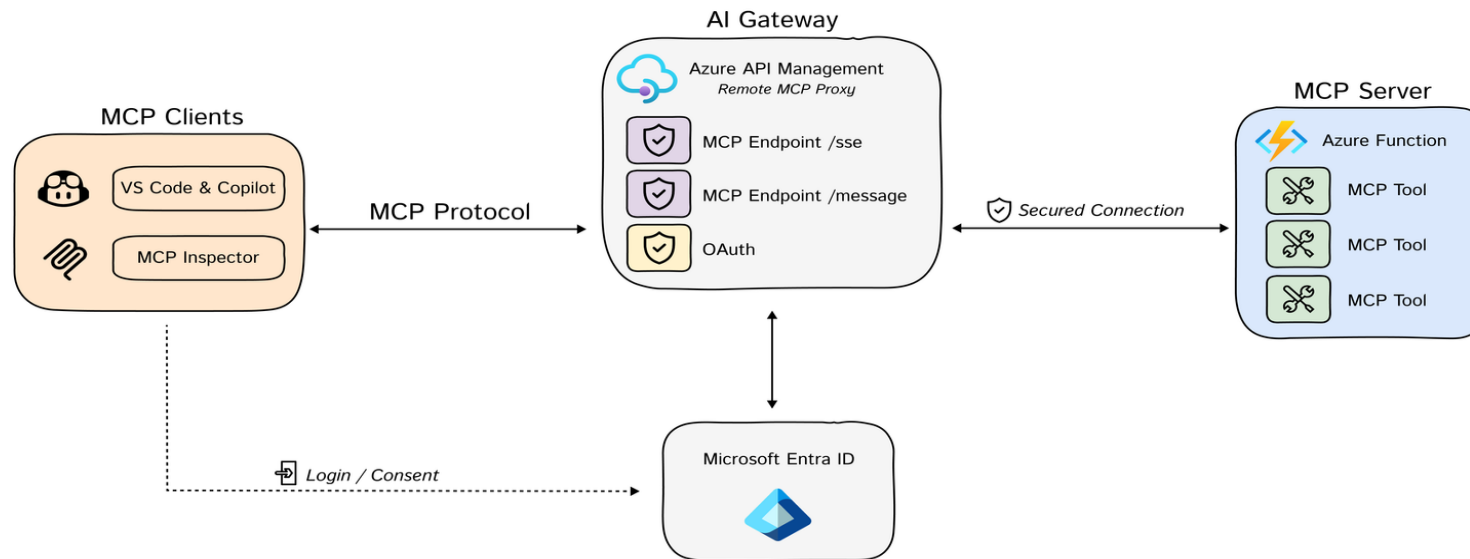
MCP enhances the capabilities of AI Gateways by providing a structured method for AI models to interact with various tools and data sources.

Real-time Data Fetching: AI models can retrieve fresh information from APIs, databases and internal systems

Contextual AI Responses: Enhances AI responses by providing accurate, up-to-date information

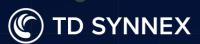
Enterprise-Ready: Secure and scalable for business applications

MCP Server on Function Apps and AI Gateway Integration





DEMO





MCP Servers examples

Azure MCP Server (Public Preview) - [Github repo](#)

Supports the following Azure services:

- Azure Cosmos DB
- Azure Storage
- Azure Monitor (Log Analytics)
- Azure App Configuration
- Azure Resource Group

Can execute Azure CLI and azd commands directly.

Lokka - [Project's site](#)

Supports Microsoft Graph and Azure APIs.

Example prompts:

- "Create a new security group called 'Sales and HR' with a dynamic rule based on the department attribute"
- "What was the most expensive service in Azure last month?"



Links

Azure Samples:

AI Gateway @ <https://github.com/Azure-Samples/AI-Gateway>

Follow us @:

<https://globalazuretorino.welol.it/>

<https://www.meetup.com/it-IT/meetup-microsoft-azure-torino/>

