BY COMMUNITY • FOR COMMUNITY •
#GLOBALAZURE
TORINO

Azure Meetup
TORINO

alveo®
EXPERTISE

TD SYNNEX

#GlobalAzureTorino

Deploying and Protecting LLMs at Scale with Azure API Management

Mattia Contessa, Fabio Cannas

Azure Meetup TORINO

alveo EXPERTISE

TD SYNNEX

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?

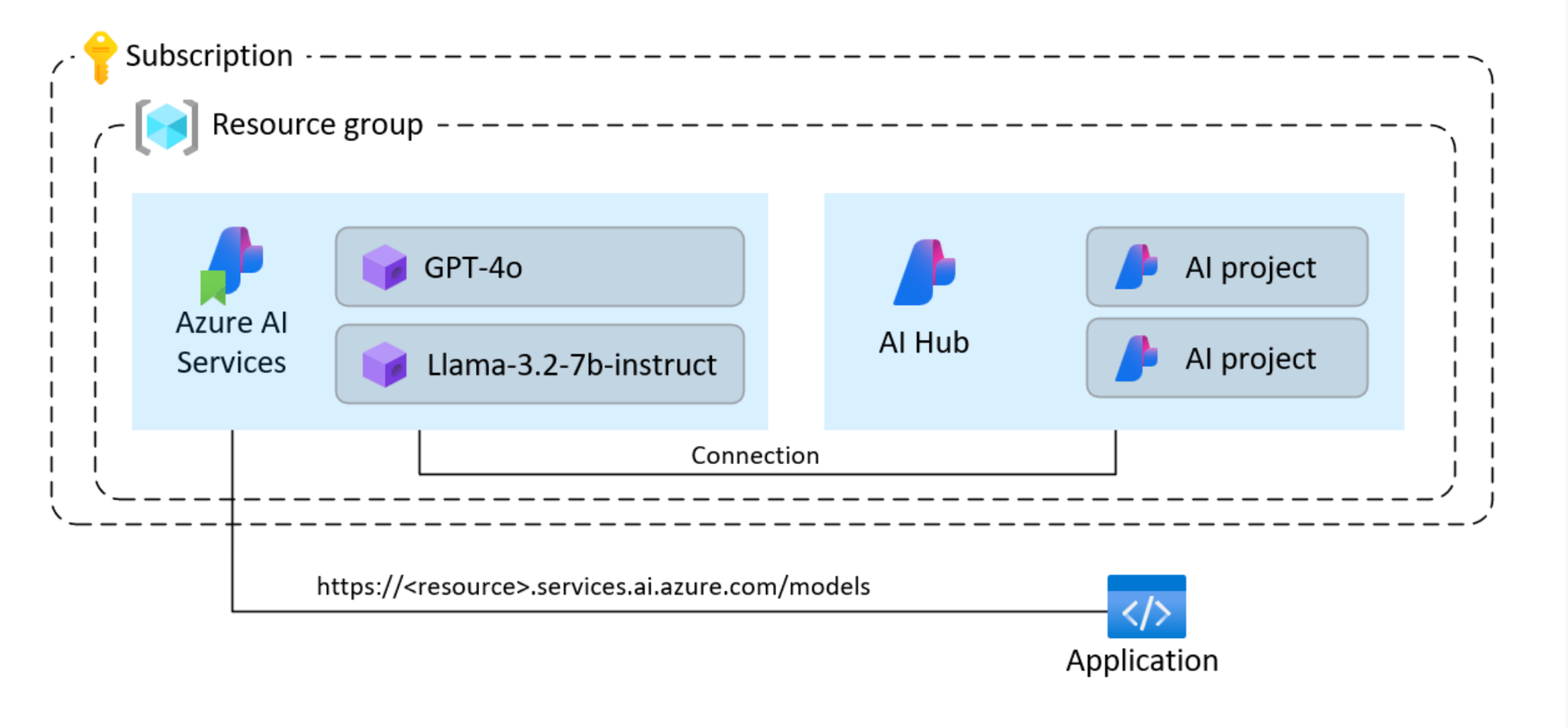
Azure Meetup TORINO

alveo EXPERTISE

TD SYNNEX

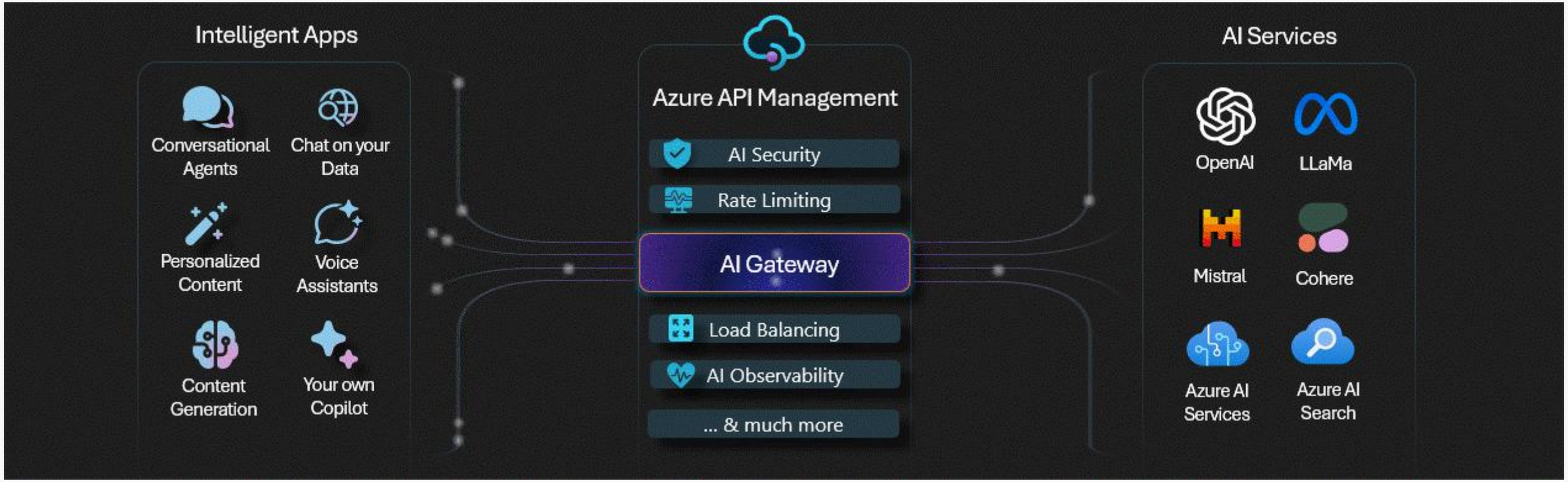
#GlobalAzureTorino

Integrating LLMs with Azure



#GlobalAzureTorino

AI Gateway Key Features



Credits: aka.ms/ai-gateway

#GlobalAzureTorino

Azure API Management

Overview

Azure Meetup TORINO

alveo EXPERTISE

TD SYNNEX

#GlobalAzureTorino

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 Meetup TORINO

alveo EXPERTISE

TD SYNNEX

#GlobalAzureTorino

Azure AI model inference

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

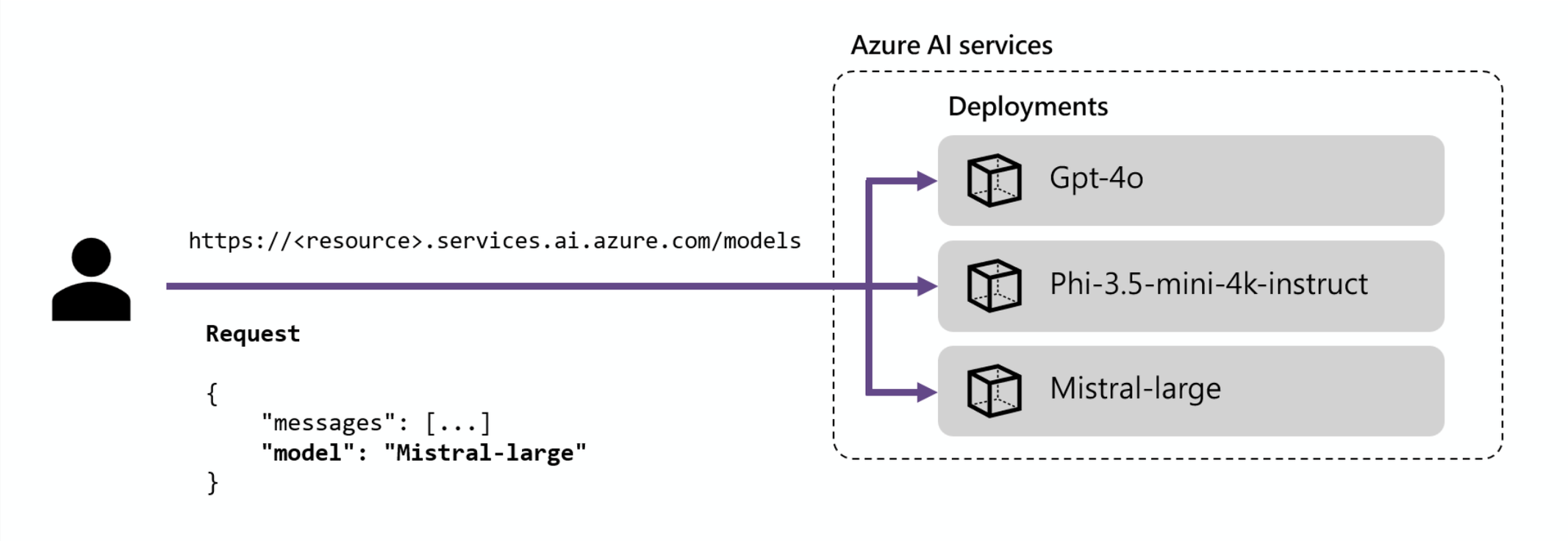
Models are hosted in a Microsoft-managed infrastructure, which enables API-based access to the model provider's model.

Models:

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

#GlobalAzureTorino

Azure AI model inference



Azure AI model inference documentation | Microsoft Learn

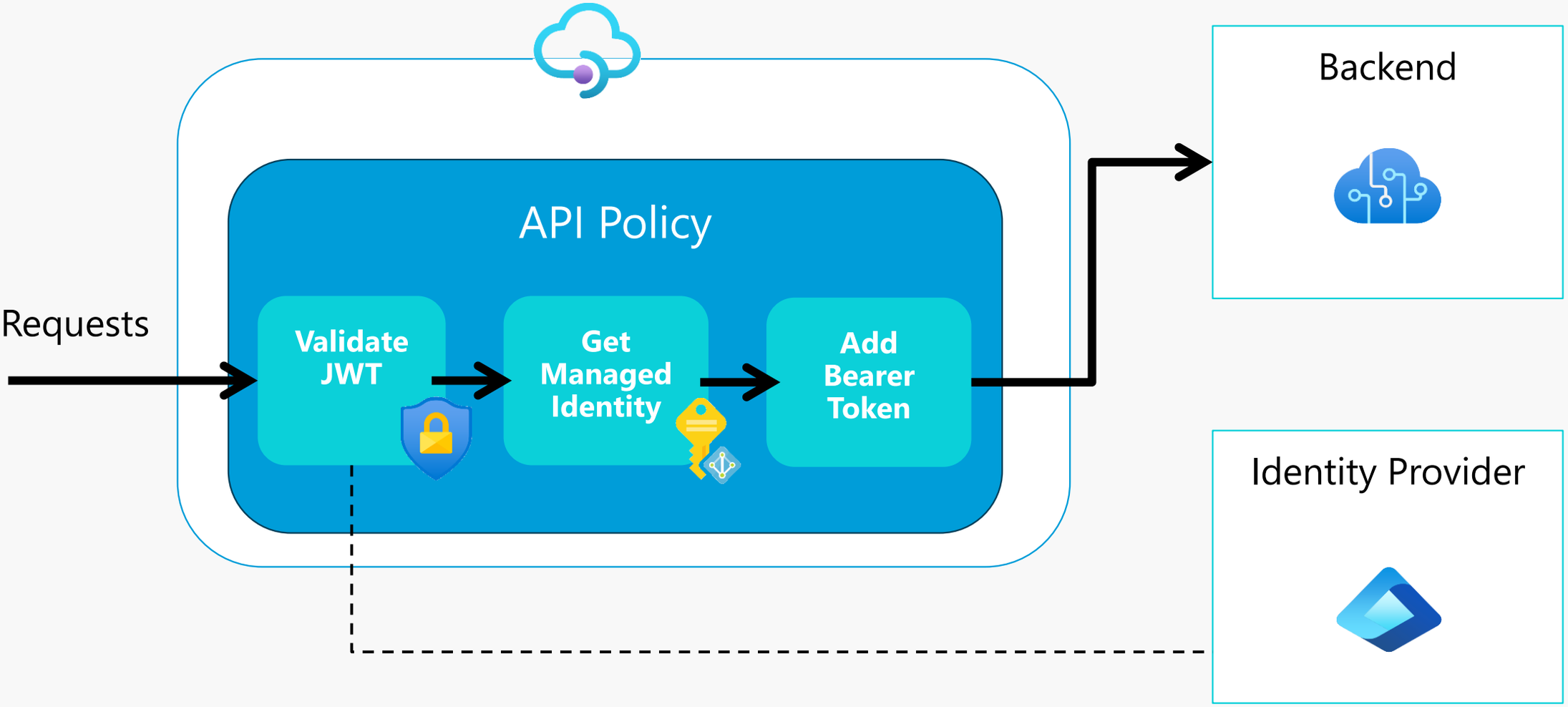
DEMO

Azure Meetup TORINO

alveo EXPERTISE

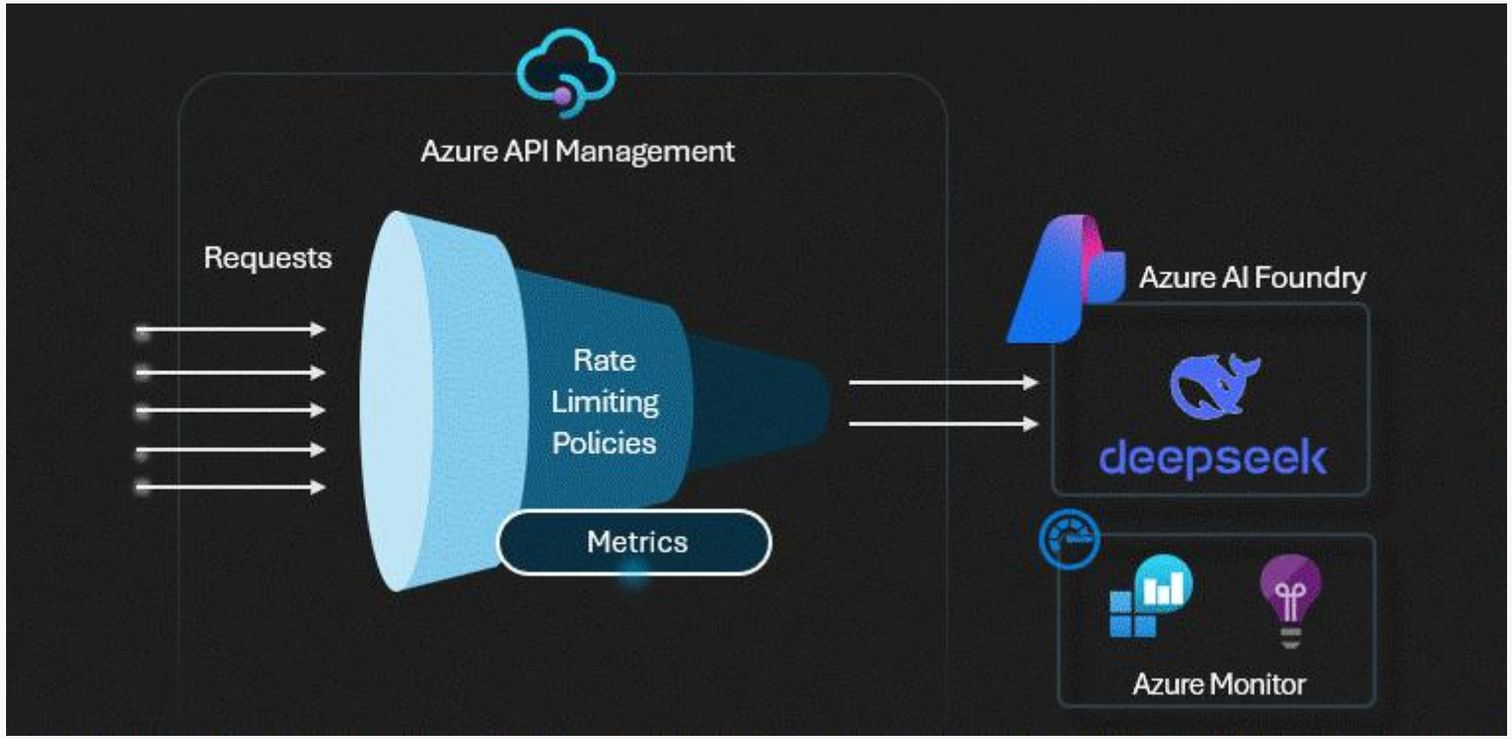
TD SYNNEX

Access Control



#GlobalAzureTorino

Rate Limiting and Monitoring



Credits: AI-Gateway/labs/ai-foundry-deepseek

DEMO

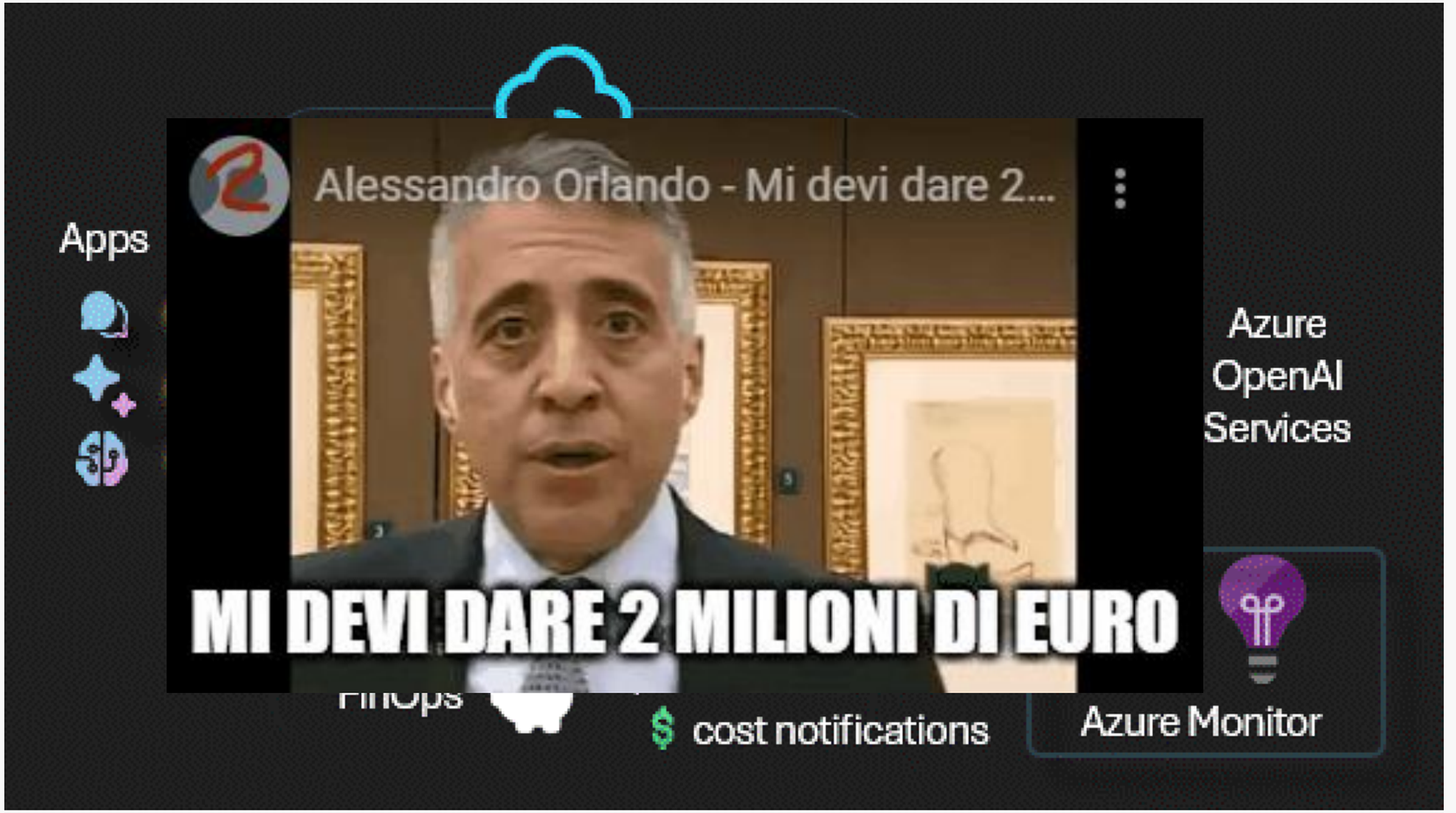
Azure Meetup TORINO

alveo EXPERTISE

TD SYNNEX

#GlobalAzureTorino

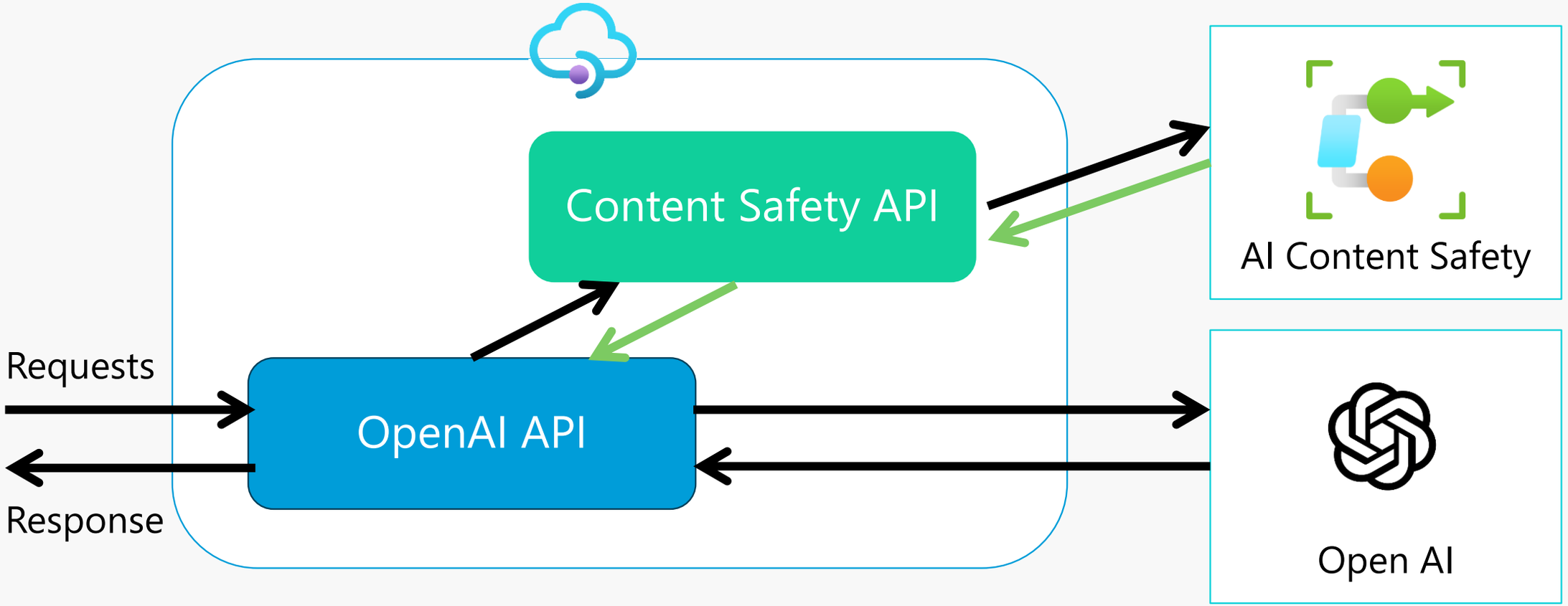
FinOps



Credits: AI-Gateway/labs/finops-framework

#GlobalAzureTorino

Manage Content



#GlobalAzureTorino

Model Context Protocol

Extending AI models capabilities

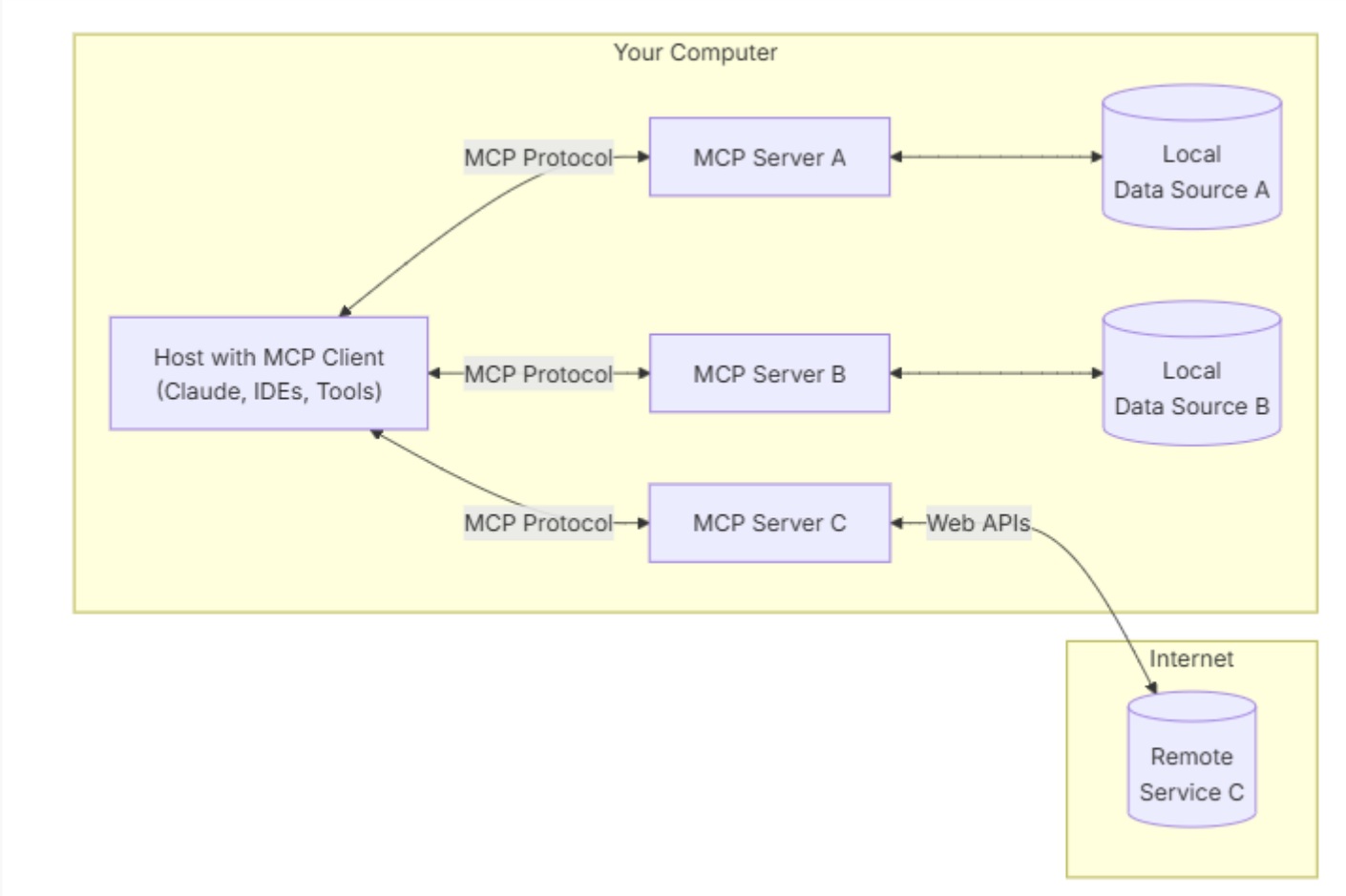
Azure Meetup TORINO

alveo EXPERTISE

TD SYNNEX

#GlobalAzureTorino

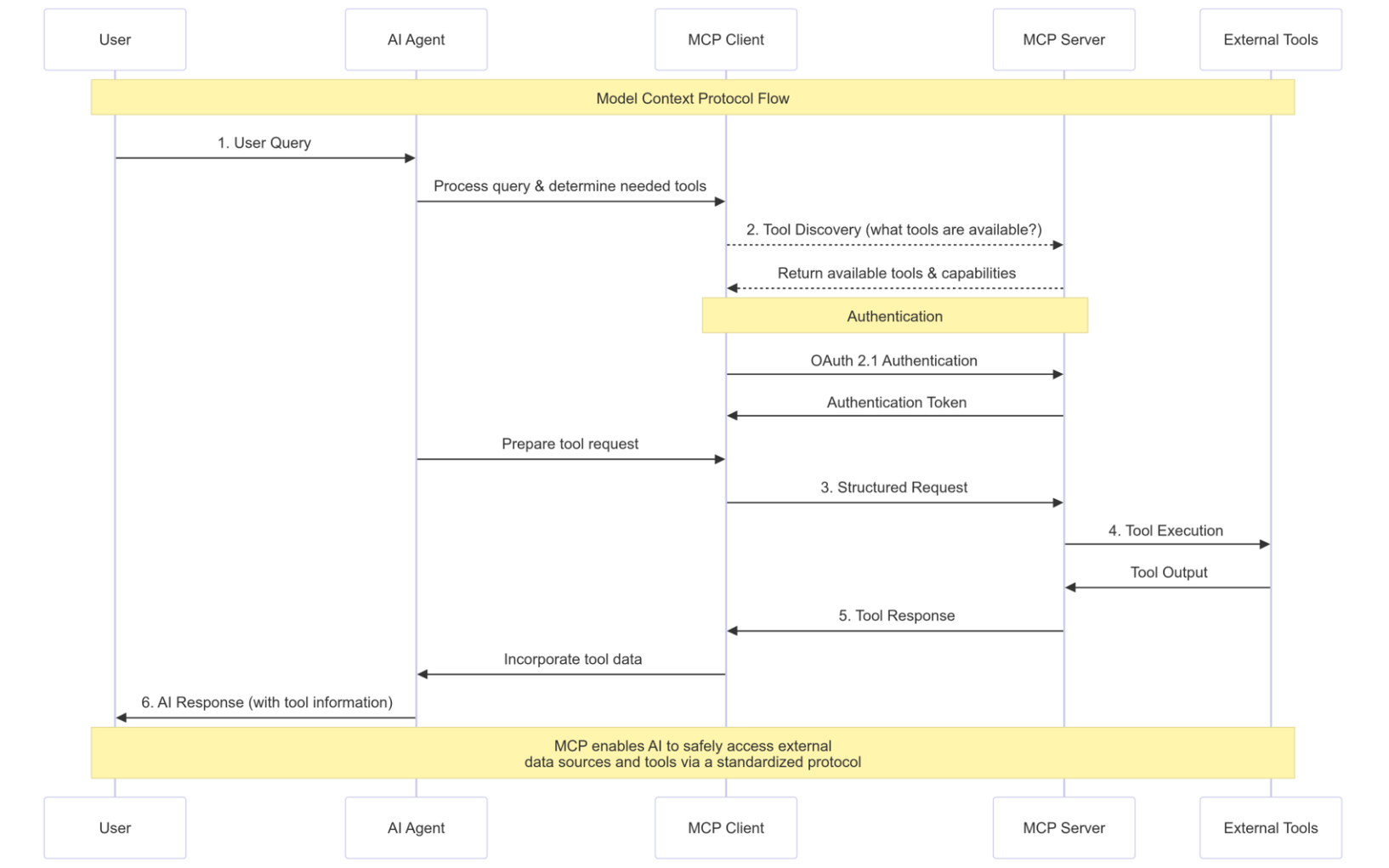
MCP Architecture



Credits: [ModelContextProtocol/introduction/General-Architecture]

#GlobalAzureTorino

MCP Flow Diagram



Credits: [Developer guide to MCP]

#GlobalAzureTorino

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.

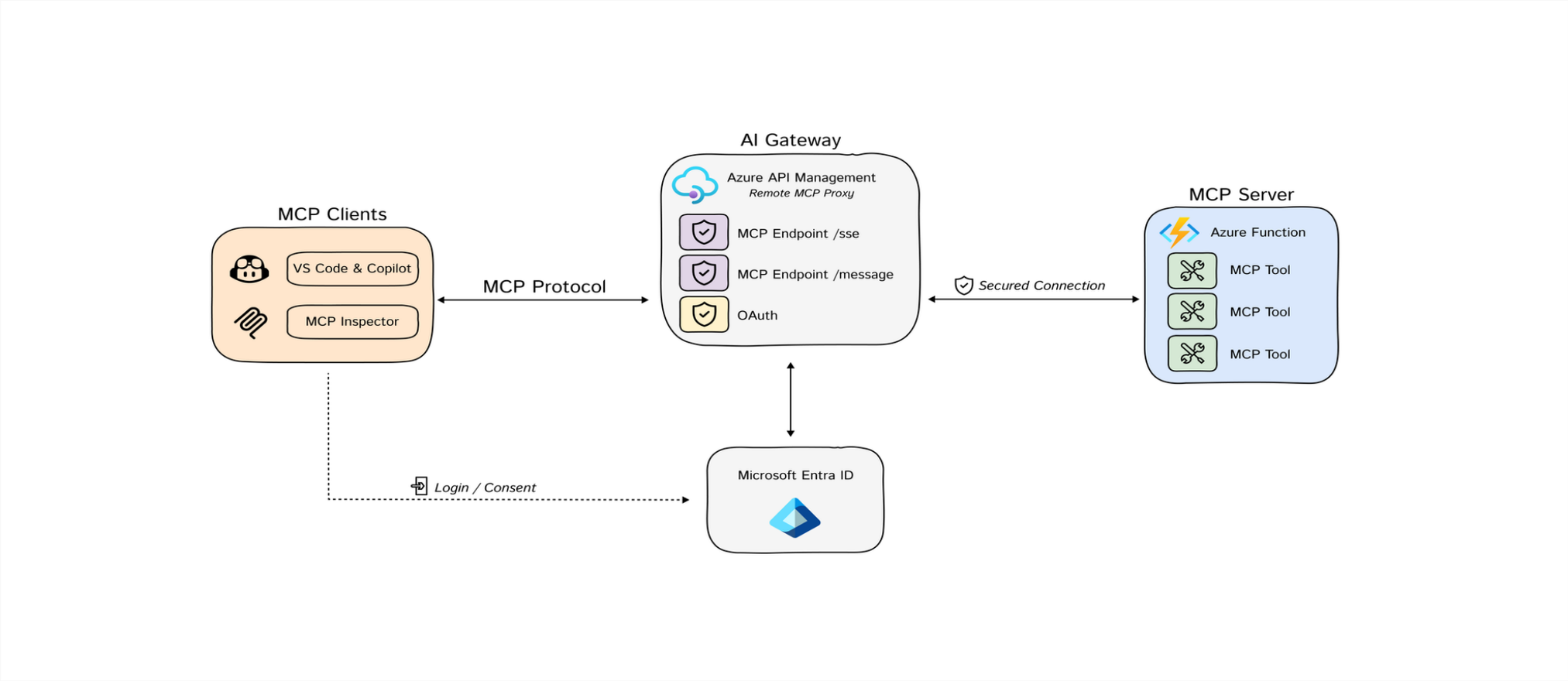
Real-time Data Fetching: AI models can retrieve fresh information from APIs, databases and internal systems

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

Enterprise-Ready: Secure and scalable for business applications

#GlobalAzureTorino

MCP Server on Function Apps and AI Gateway Integration



Credits: [Azure API Management Your Auth Gateway For MCP Servers]

DEMO

Azure Meetup TORINO

alveo EXPERTISE

TD SYNNEX

#GlobalAzureTorino

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?"

#GlobalAzureTorino

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/

BY COMMUNITY • FOR COMMUNITY •
#GLOBALAZURE
TORINO