#Milano

Azure Foundry Local: l’AI all’Edge che cambia le regole del gioco

Marco Dal Pino

Senior Technical Consultant @Microsoft

#GLOBALAZURE 2026

#Milano

improove

TD SYNNEX

Grazie ai nostri sponsor 🙏

Adaptive cloud

Fragmented IT solutions result in:

Difficult to manage disparate locations, systems, and tools

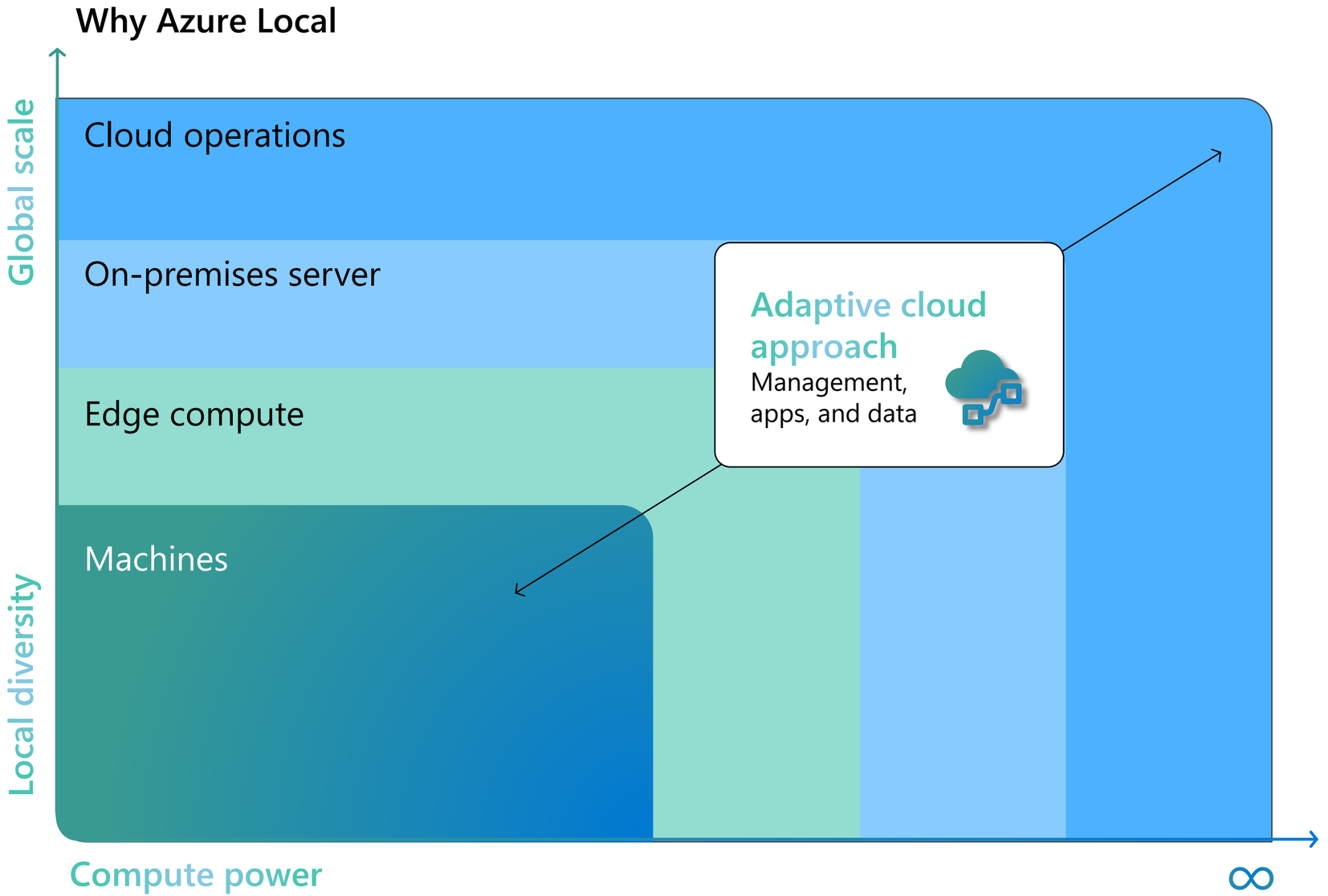
Hard to innovate and adapt

Data stuck in silos

Why is this a problem?

Fragmented IT solutions slow operations, stifle innovation, and isolate data, **making it harder for teams to collaborate, react to change, implement AI and deliver value at scale.**

Consolidate with cloud-native solutions



Outcomes

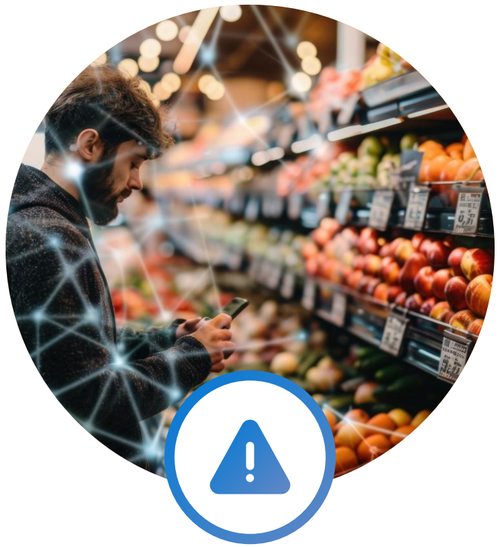
- Lower TCO
- Better margins
- Higher quality
- Faster deployment
- Enhanced efficiency
- Cloud innovation

Some workloads need to stay on-premises

Local AI inferencing

Pipeline leak detection

Personnel safety checks



Mission critical business continuity

Production line operations

Point of sale systems

Near real-time systems

Quality assurance

Manufacturing execution system

Custom sovereignty and regulatory requirements

Highly regulated industries

Defense and intelligence

Azure’s adaptive cloud approach

Cloud services and tools

- Operate with AI-enhanced central **management & security**
- Develop and scale **applications** across boundaries
- Unify **data and AI** across a distributed estate

Global infrastructure

- Innovate on limitless and trusted **infrastructure**

Public cloud — Hybrid cloud — Sovereign Cloud — Multi-cloud — Edge — IoT

Enabled by Azure Arc

Azure’s adaptive cloud approach

Cloud services and tools

- Operate with AI-enhanced central **management & security**
- Develop and scale **applications** across boundaries
- Unify **data and AI** across a distributed estate

Global infrastructure

- Innovate on limitless and trusted **infrastructure**

Public cloud — Hybrid cloud — Sovereign Cloud — Multi-cloud — Edge — IoT

Enabled by Azure Arc

Microsoft Foundry

Agentic AI is the next wave

81%

of leaders expect agents in their company's AI strategy in next 12-18 months[1]

93%

of organizations are experimenting with multiple models[2]

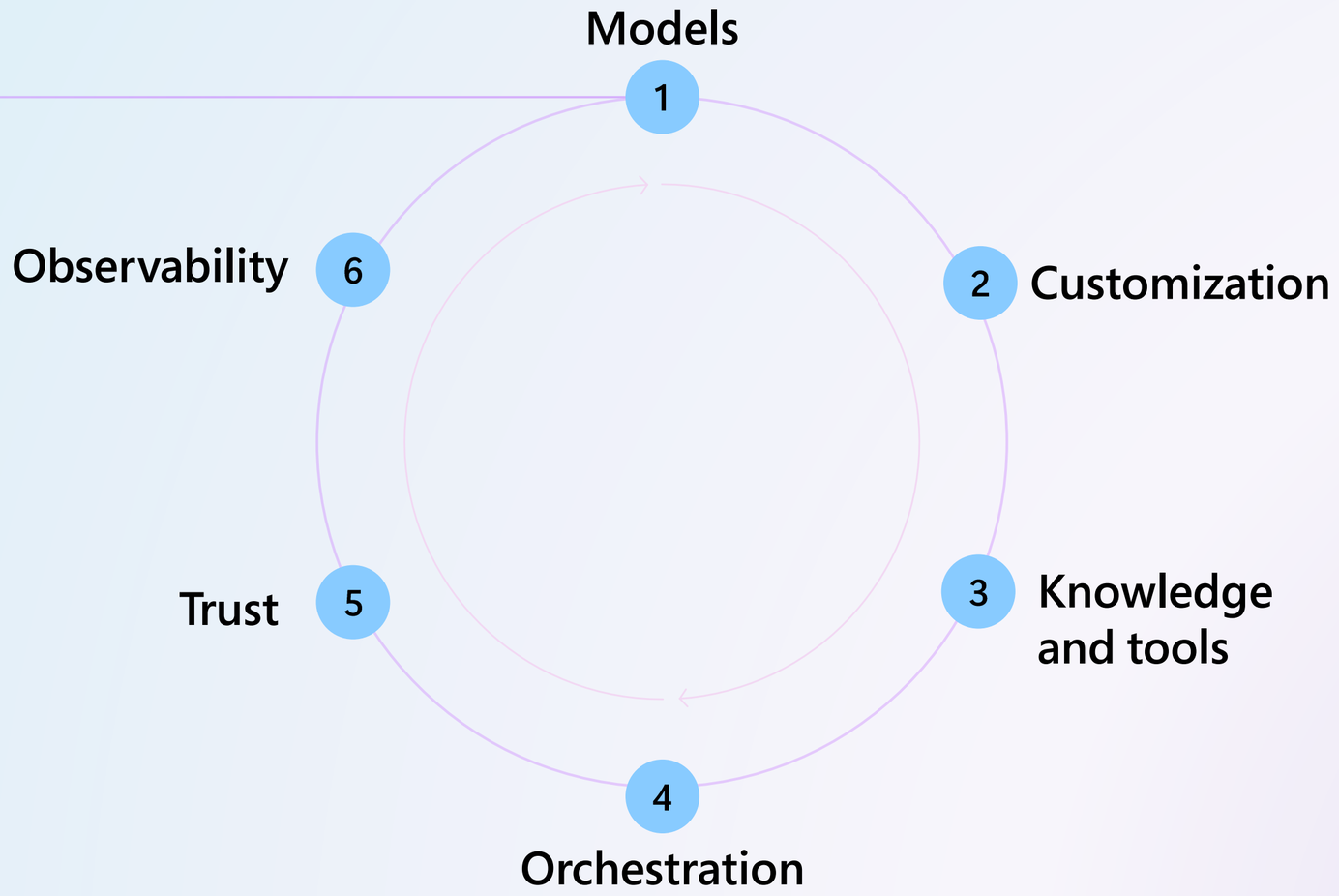
70%

of generative AI experiments have not moved to production[3]

1. Work Trend Index Annual Report
2. 16 Changes to the Way Enterprises Are Building and Buying Generative AI | Andreessen Horowitz
3. GenAI and the future enterprise | Deloitte Insights

Foundry

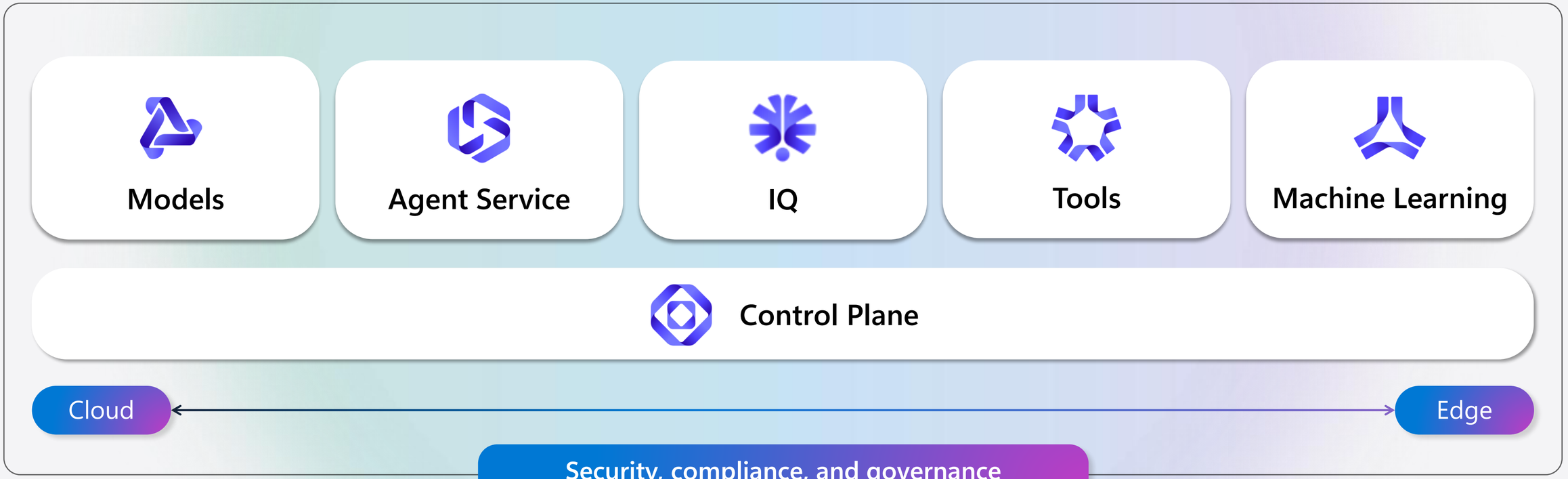
AI Development Fundamentals



Foundry

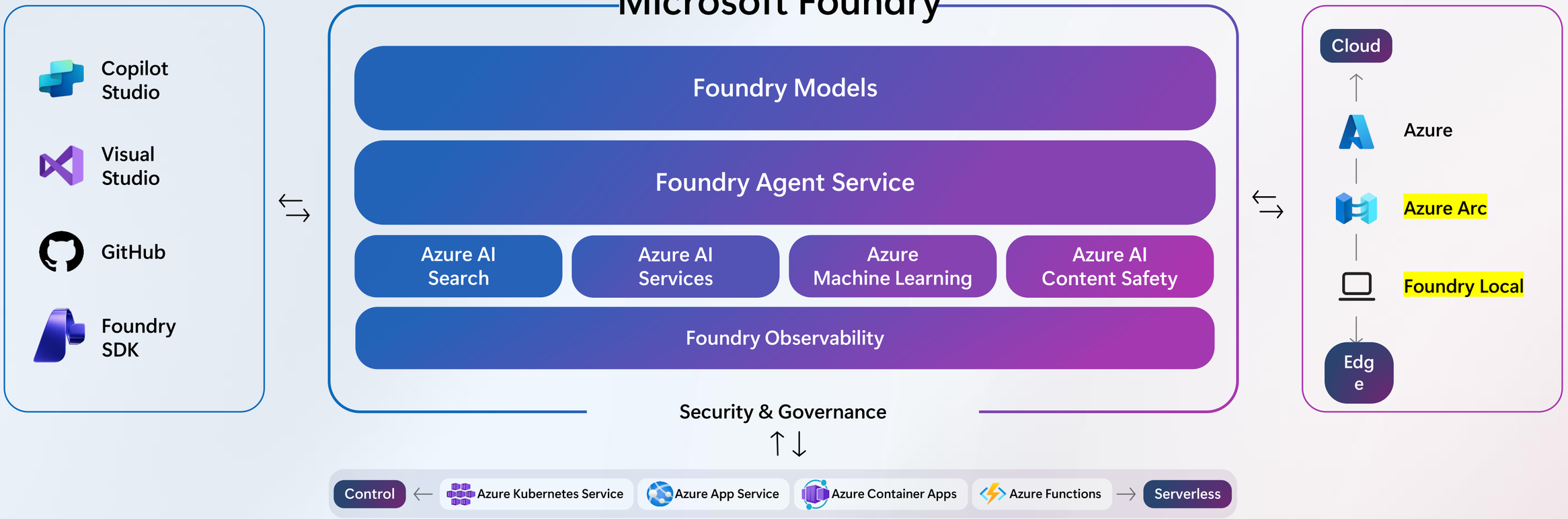
Microsoft Foundry

The AI app and agent factory

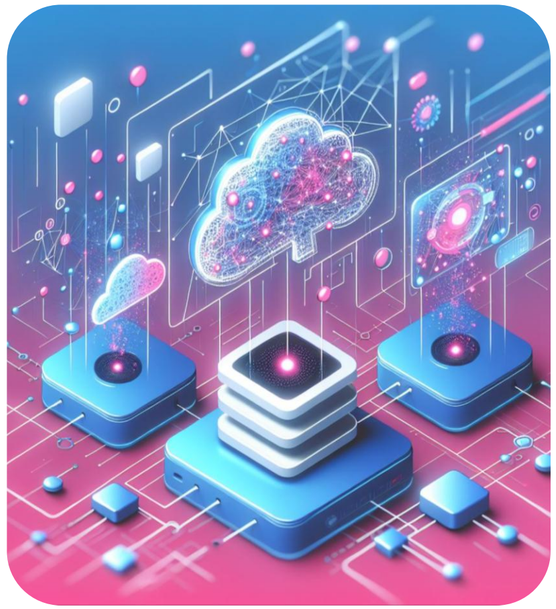


Foundry

Microsoft Foundry



Unlock AI Locally with Foundry Local



Easily build, test, deploy AI Apps using local AI models

Phi SLM

Mistral7B

DeepSeek R1

Meta Llama 3.1 8B

Rich collection of models, tools, agents designed for local AI

High performance on-device inference fully utilizes silicon on Windows 11 and Mac OS

Microsoft Azure

Foundry for Azure Local
Foundry Local for Azure Local
Foundry Local for Azure

Unlock AI innovation while keeping your data where it belongs for adaptive cloud customers

Microsoft Foundry spans cloud to edge

Microsoft Foundry

Frontier models, agents
& fine-tuning hub

Cloud

Foundry for Azure Local enabled by Azure Arc

Edge and On-premise

Edge, hybrid, air-gapped

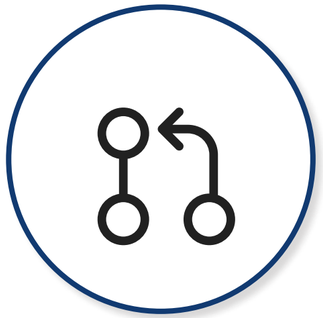
Foundry Local

Windows, MacOS, &
Android (Private Preview)

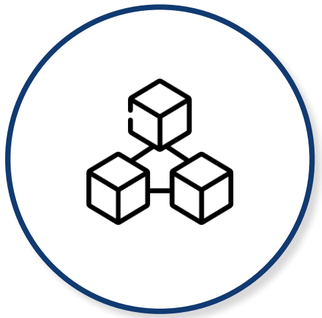
IoT, phones, laptops, desktops

What We're Hearing from Foundry Customers

Asking for clarity on how the various Foundry offerings fit together- positioning must be simplified and aligned



"We want integration with our software distribution pipelines"



"We've standardized on k8s and microservices- your solution should fit into our architecture"

"We will need to bring our own AI model" (25% of users)

"We need one AI management layer across cloud and edge-no silos, no duplication" (78% of users)

"We need deployment across large fleets- multi-node coordination and scale are critical"

"Edge AI agents can deliver high or even transformational impact on daily operation" (88% of users)

Foundry enabled by Azure Arc

cloud and edge AI offering
enabled by Azure Arc for distributed deployments

Out-of-the-box AI for real-world scenarios

Ultra-fast inference with no compromise on model accuracy

Leverage Foundry models, deploy at the edge

Secure and compliant, by keeping data local

Today: Products enabled by Azure Arc

RAG enabled by Arc

(Knowledge)

On-device intelligence with built-in retrieval-augmented generation for secure, real-time insights at the edge.

Preview

Video Indexer enabled by Arc

(Vision)

Video intelligence solution for cloud and edge - powered by AI agents and LLMs for video analysis of live and recorded content

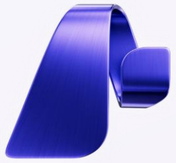
Preview features

Azure Machine Learning enabled by Arc

(Custom Models)

Experiment tracking, fine tuning, and monitoring, while maintaining local control over compute.

General Available



Foundry Local catalog enabled by Arc

(Model Deployment)

Orchestration of AI model catalog and agents at the edge, enabling inferencing and deployment as scale

Coming soon!

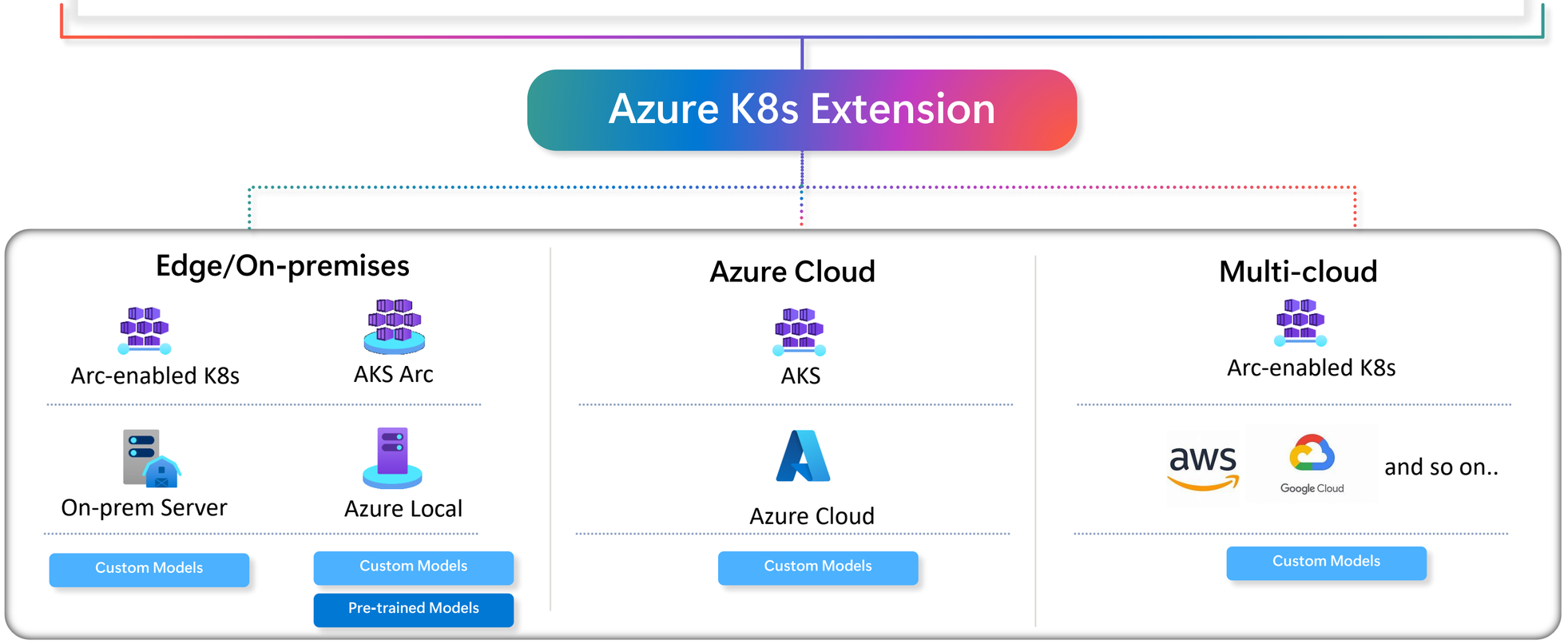
Product names are subjected to update

Kubernetes as a compute target

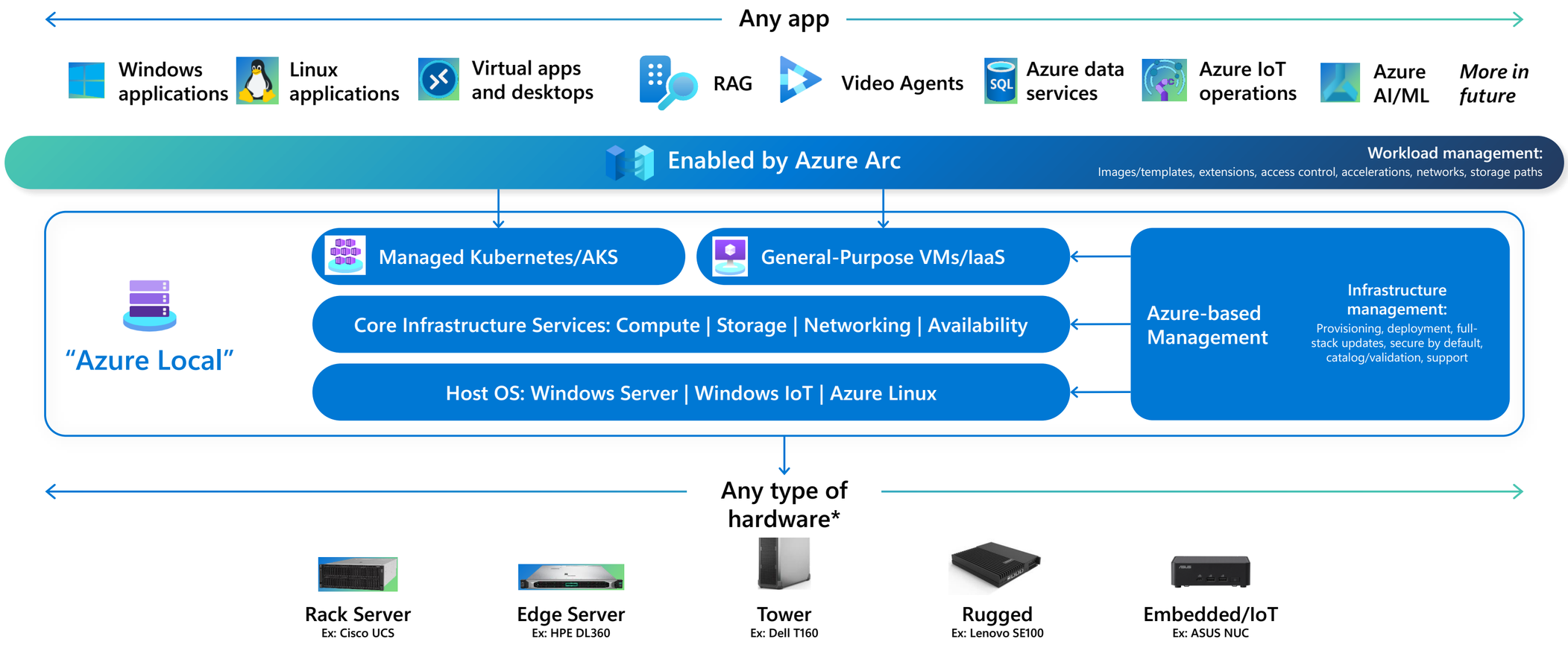
Foundry models and agents

Option 1: Build on Foundry, deploy anywhere

Option 2: Build at the edge (API only), deploy at the edge



Validated on Azure Local, supporting other infrastructure



*Must meet minimum requirements per operating system and solution-level pass validation

Local AI use-cases in every industry

Regulated Customers

- Real-time threat detection
- Secure video analytics for restricted environments.
- Fraud detection models deployed in secure enclaves.
- Local chat experience with local data sources (such as SharePoint, Exchange and others)

Manufacturing

- Real-time defect detection
- Predictive maintenance
- Employee safety monitoring

Retail

- Real-time loss prevention
- Streamline store operations
- Customer safety monitoring

Roadmap

Sovereign AI

Sovereign Public Cloud

Data stays in Europe, under European law

Data Guardian: operations and access controlled by Europeans

Sovereign controls for policy enforcement

Applies to existing Europe cloud datacenter regions with no migration

Sovereign Private Cloud

Azure Local + Microsoft 365 Local: integrated cloud and productivity

Hybrid or disconnected at your location

Validated architecture and partner ecosystem

Virtualization services

National Partner Clouds

For government and critical infrastructure criteria

Government approved local operator independent from Microsoft

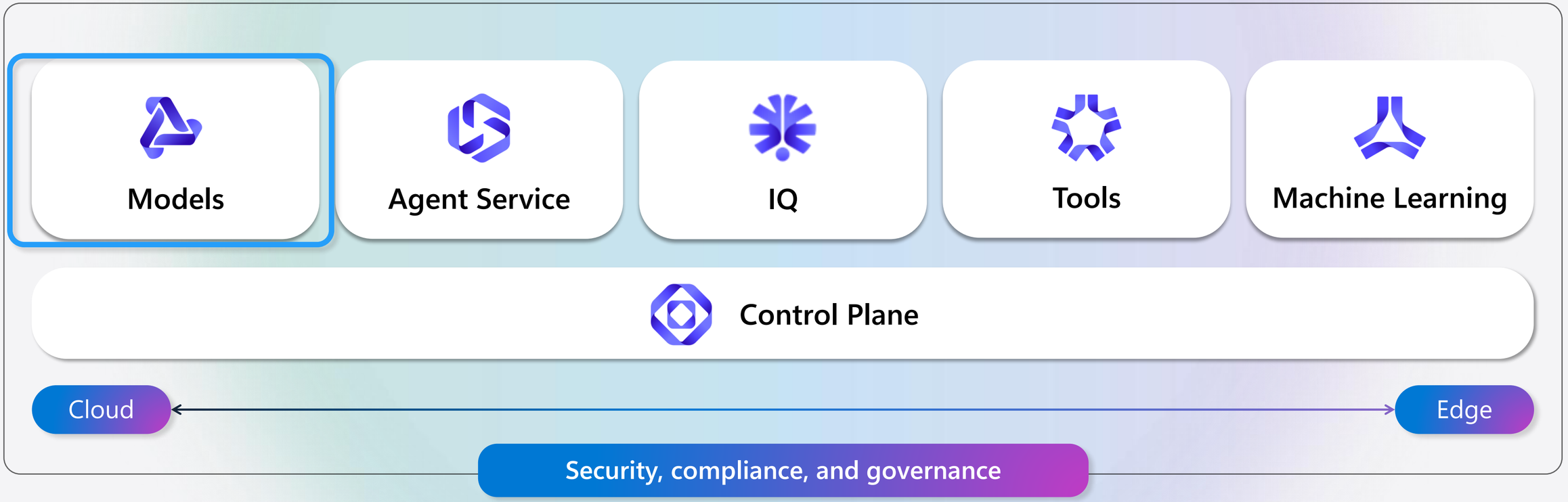
Clouds in Germany (Delos Cloud) and France (Bleu) with local ownership and isolated infrastructure

Microsoft

Foundry Model catalog and inference

Microsoft Foundry

The AI app and agent factory



Every model.
Every use case.
One platform.

Avoid model lock-in:

- Flexible model selection
- Curated for organizations
- Explore, compare, evaluate and swap models with ease

Foundry Models

11,000+ Frontier, specialized and open models

Models direct from Azure

Azure-hosted, Azure-sold. API Licensed and supported directly by Microsoft

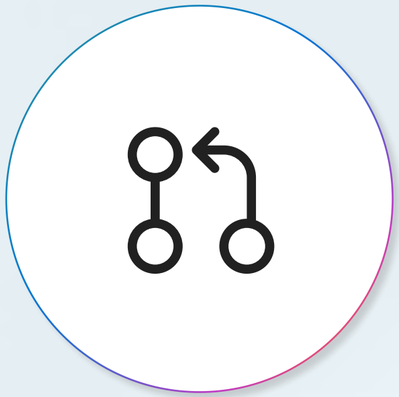
Azure OpenAI · Azure DeepSeek · Azure Mistral AI · Azure Grok · Azure Black Forest Labs · Azure Llama

Models from Partners & Community

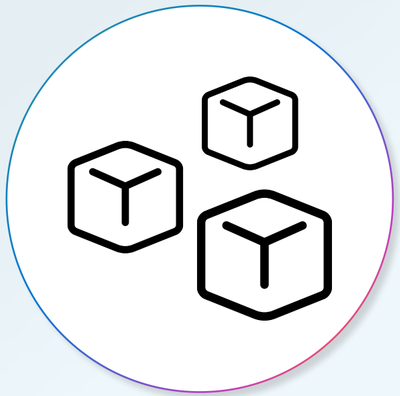
Azure-hosted, partner-sold or open-source, licensing/support from partner or community

Cohere · Hugging Face · NVIDIA · Nixtla · DatabrickS · SIGHT MACHINE · saifr · Rockwell Automation · Paige

What we have heard from Foundry customers using Kubernetes



Integrate with my software distribution pipelines (CI/CD)



We use microservices architecture patterns

I take advantage of Kubernetes building blocks and tooling

I will need to bring my own AI model

We need one control plane to rule them all

Expand Foundry Local

Windows AI Foundry integrates

Azure Arc uses

Foundry Local

Dev Experience

SDK | CLI | Containers | UX

Asset Management

Models | Agents | MCP Servers | Knowledge

Runtime

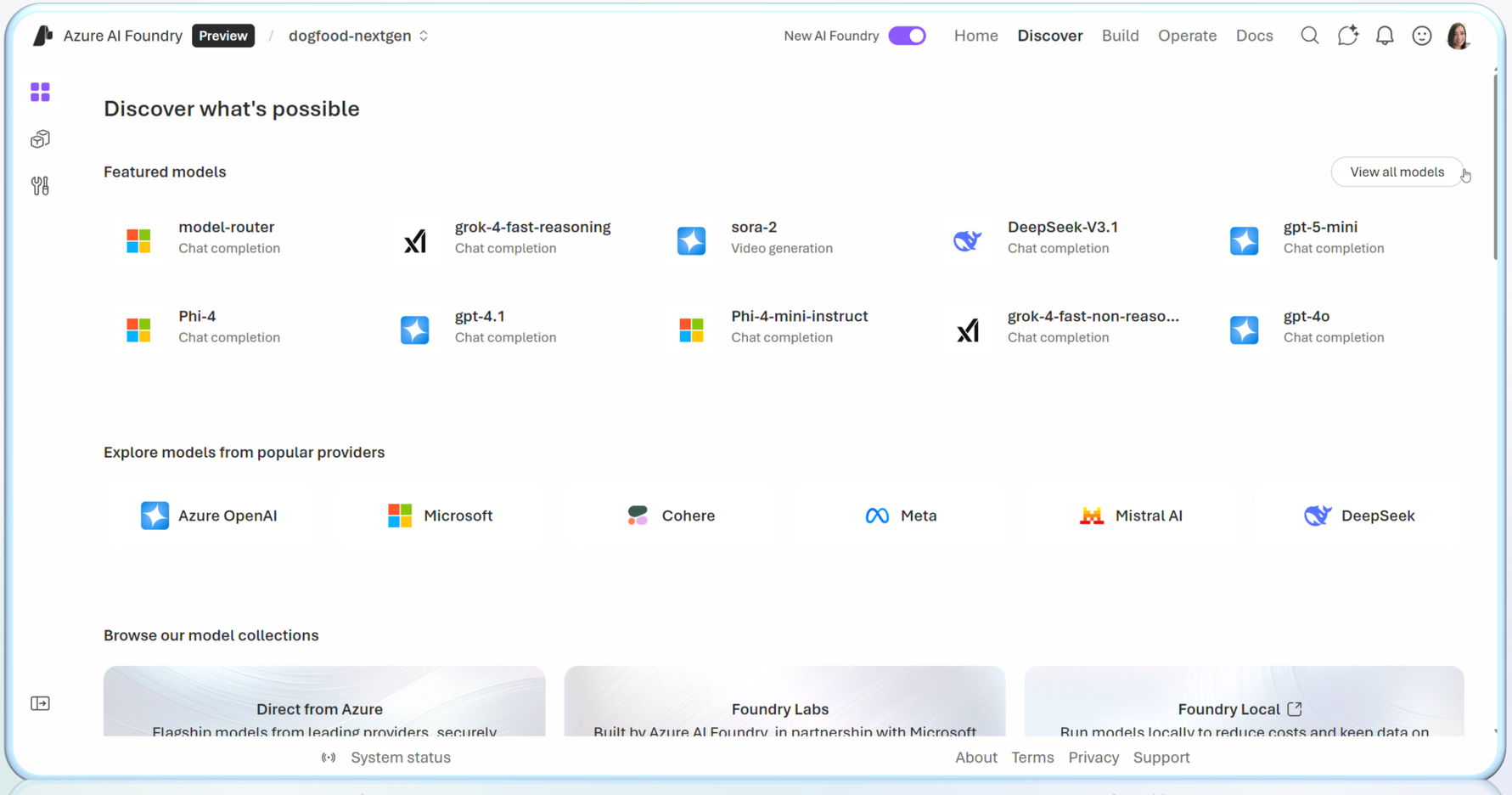
ORT | Other RT

Value-Added Services

Observability
Security
Private model Distro
CI/CD
Compute Vision
....

Android | iOS | MacOS | Windows | Linux

Silicon



Call to action

Explore our models

https://aka.ms/foundrylocal

Sign up for Android Private Preview

https://aka.ms/androidprp

Integrate Foundry Local into your app (Windows, macOS)

https://aka.ms/foundrylocal

Try Foundry Local on Kubernetes

https://aka.ms/FL-K8s-Preview-Signup

Samples for each language are provided in the

Foundry Local GitHub Repository

Foundry Local

What's next:

Foundry Local powered by Azure Local:
brings models and agentic AI — including RAG, chat — to *customer-owned distributed infrastructure*. This is in preview now, with more coming soon.

Expanded model catalog:
More models across more domains, with community contributions

Real-time audio transcription:
Transcribe in real-time from a microphone. Ideal for live captioning scenarios.

Enhanced hardware support:
Broader NPU and GPU coverage as the *silicon landscape evolves*

Enhanced shared cache:
Enhancements to allow *models to be shared* between applications.

- 30+ years in IT (Developer, Architect, Senior Consultant, PM, Trainer)
- Speaker, Community addicted
- IoT Influencer
- Microsoft Certified Trainer

Marco Dal Pino
Senior Technical Consultant
Microsoft

https://www.linkedin.com/in/marcodalpino

https://about.me/marcodalpino

https://twitter.com/marcodalpino

info@contoso.blog

https://www.twitch.tv/dpcons
https://www.twitch.tv/techchat

Microsoft MVP Most Valuable Professional Reconnect

DotNet Toscana

Feedback

Azure Foundry Local: l’AI all’Edge che cambia le regole del gioco

#GLOBALAZURE
2026
#Milano

Slide e video:
https://www.globalazuremilano.it