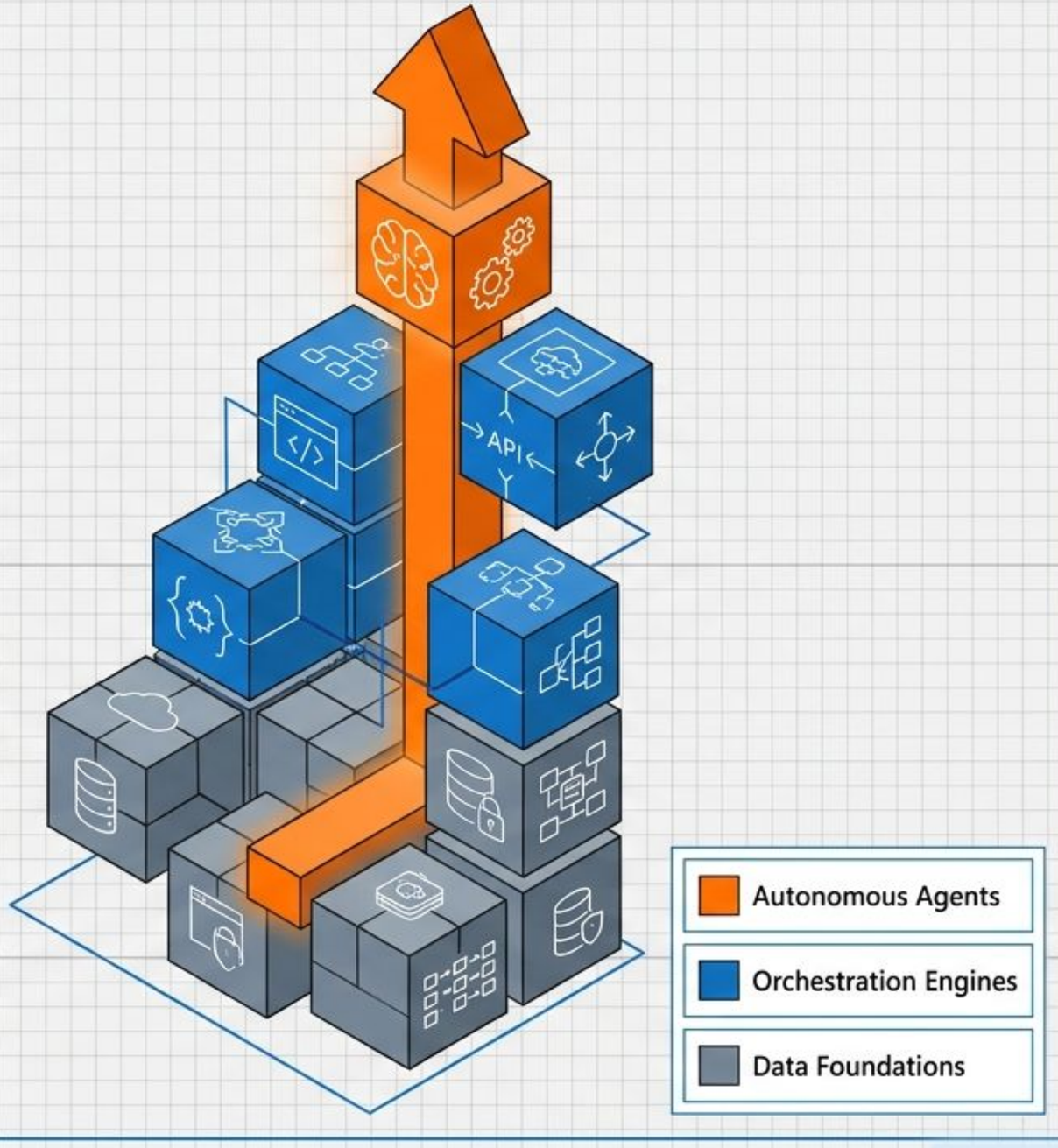
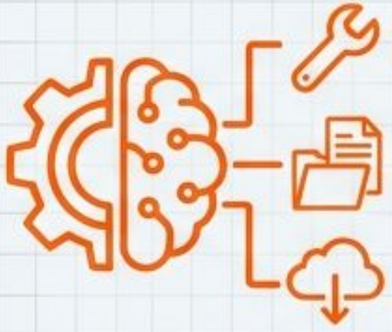

AZURE AI
Building Agentic Applications on the Microsoft Cloud
Microsoft Cloud
Agentic workflow
AzureCloud.pro

# Architecting Autonomous Enterprise Ecosystems with Agentic AI on Azure

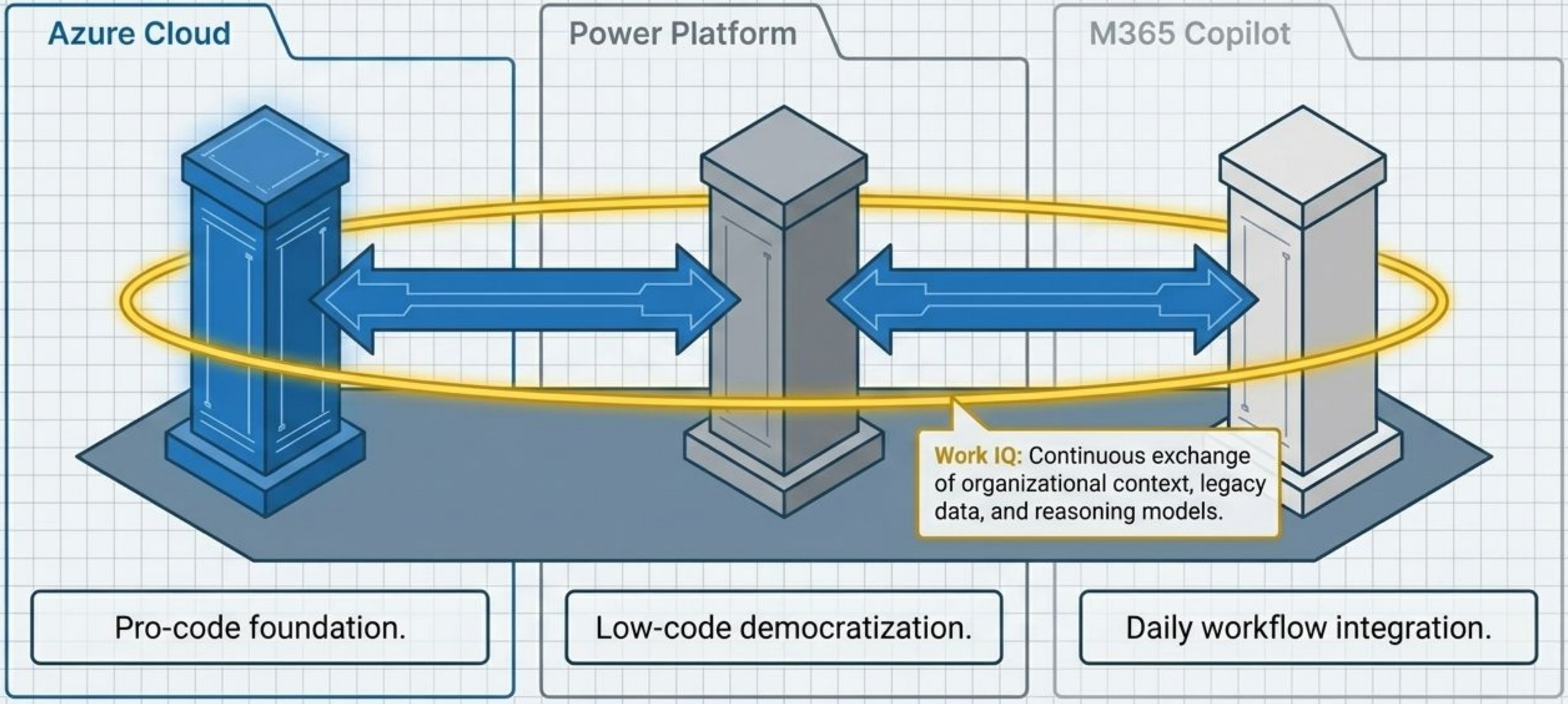
Blueprint for secure, scalable, and self-optimizing enterprise AI infrastructure.



# Transitioning from passive information retrieval to autonomous system action fundamentally changes AI architecture

| | Traditional RAG - The Librarian | Agentic AI - The Staffer |
|---|---|---|
| |  |  |
| Capabilities | Retrieval & Summarization | Task Execution & Tool Invocation |
| Scope | Protecting a static data index | Managing the blast radius of a digital worker |
| Flow | Deterministic outputs | Non-deterministic probabilistic workflows |

# Unifying custom code, low-code, and daily productivity requires a synchronized three-platform control plane



# Combining foundational agentic patterns enables automation that is faster, smarter, and capable of self-correction

## Tool Use

Agents invoke APIs to retrieve data or execute transactions. (e.g., Fujitsu reduced proposal production time by 67% by automating assembly).

## Reflection

Agents evaluate their own outputs and auto-correct errors before taking action, crucial for high-stakes compliance.

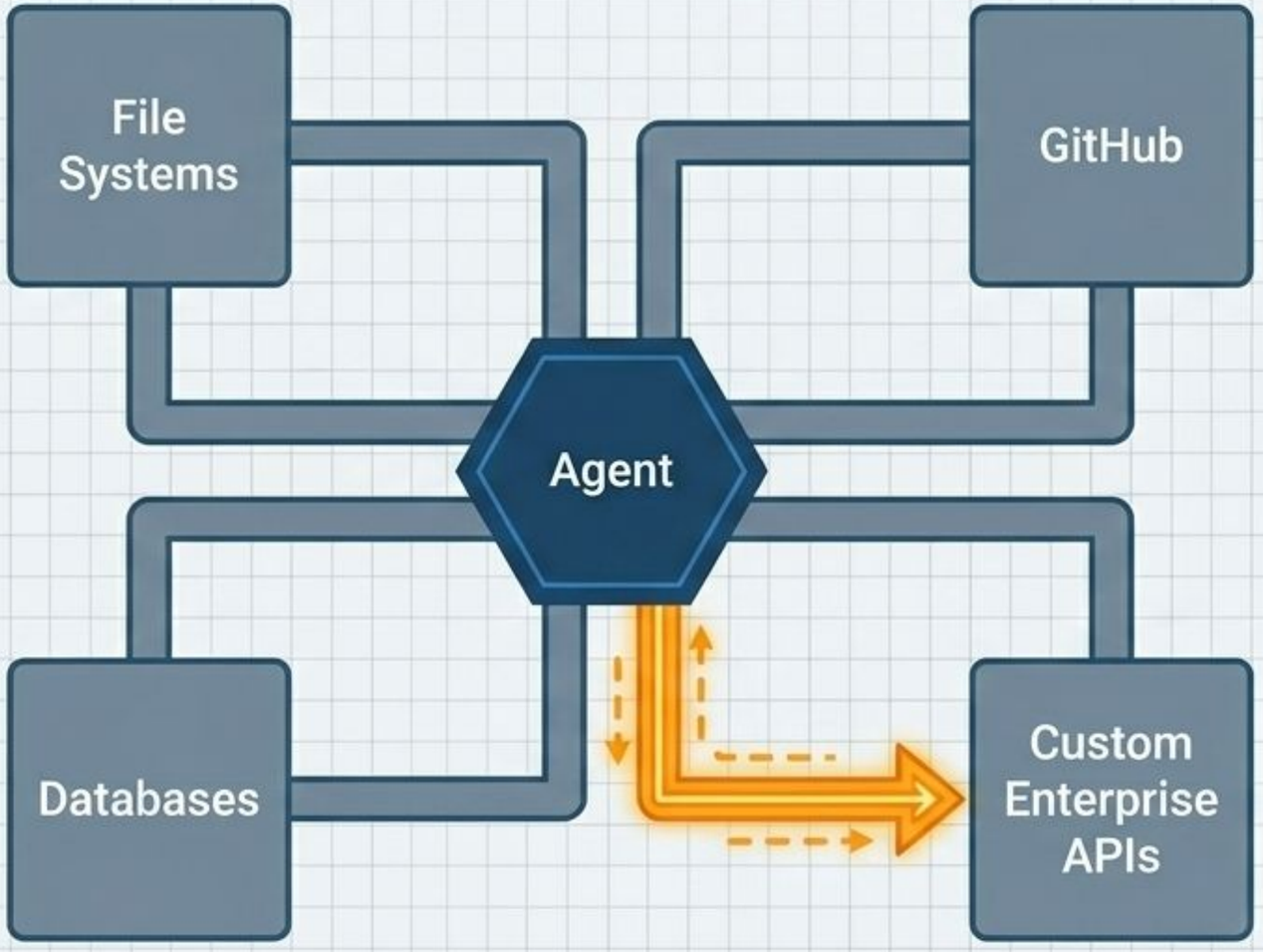
## Planning

Agents dynamically decompose complex tasks into sequential or parallel execution steps.

## Multi-Agent

Specialized agents collaborate under an orchestrator, distributing cognitive load for enterprise-scale workflows.

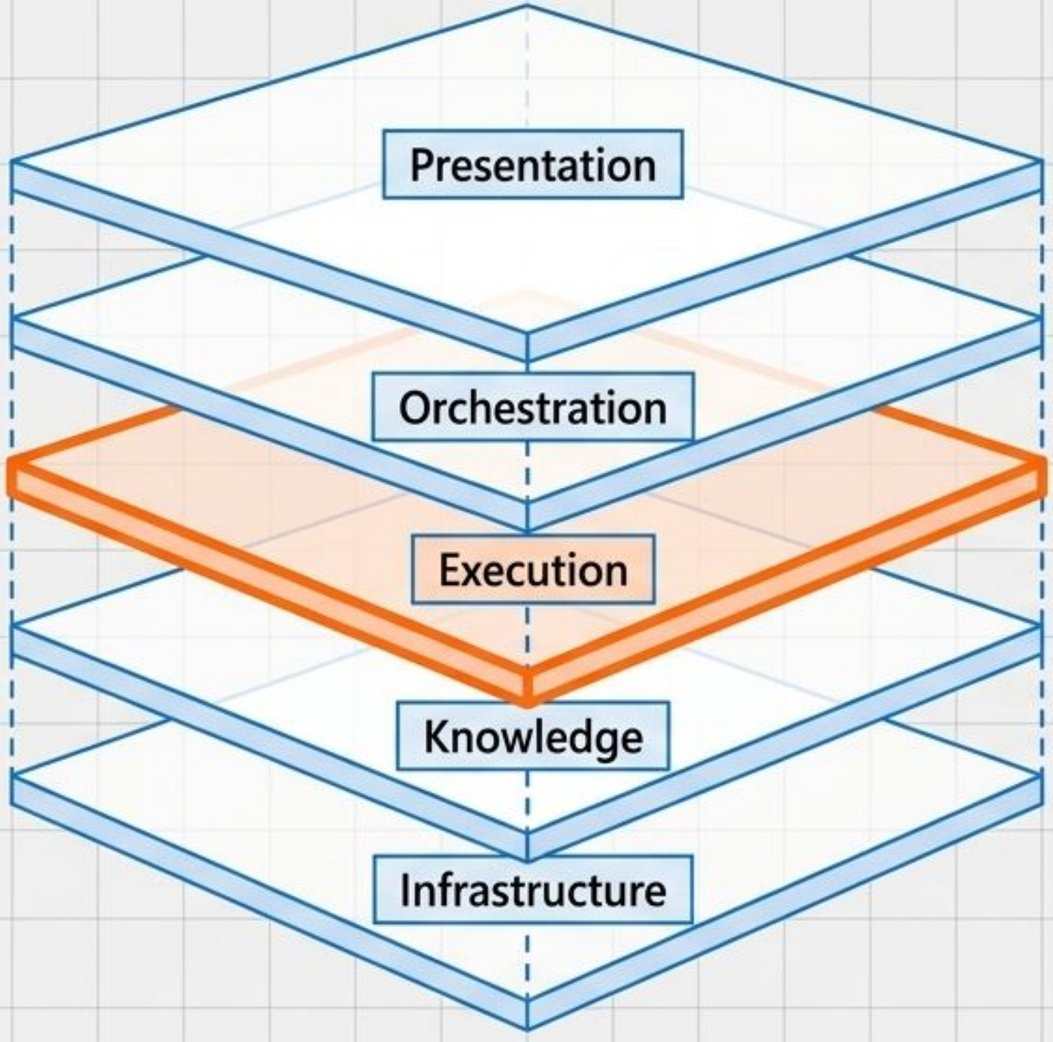
# Adopting the Model Context Protocol (MCP) eliminates manual wiring and standardizes tool discovery at runtime




**Dynamic Tool Discovery:** Tools are self-describing. The agent discovers and invokes tools on the fly without hardcoded pathways.

**Location Agnostic:** Allows tools to be hosted anywhere (on-premises or across clouds) without losing centralized governance.

# Handling the non-determinism of LLMs requires a strictly layered, five-tier modular architecture




**Web UI** protected by **App Gateway & WAF.**

**Azure AI Agent Service** managing reasoning, planning, and task delegation.

Active system modification. **Sandboxed code execution** and external API triggers via MCP.

Short-term state (Cosmos DB) and long-term vector indexing (AI Search).

Scalable compute foundation via AKS or Azure App Service.

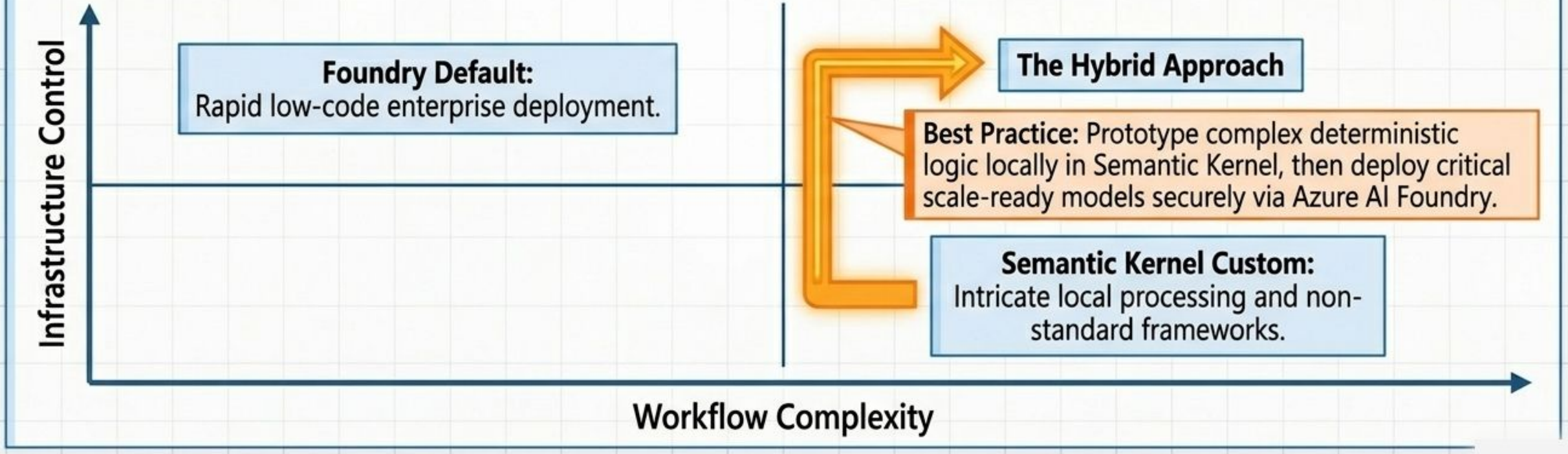
# Strategic deployment requires balancing the managed scale of Azure AI Foundry with the custom flexibility of Semantic Kernel

**Managed Platform (Azure AI Foundry)**

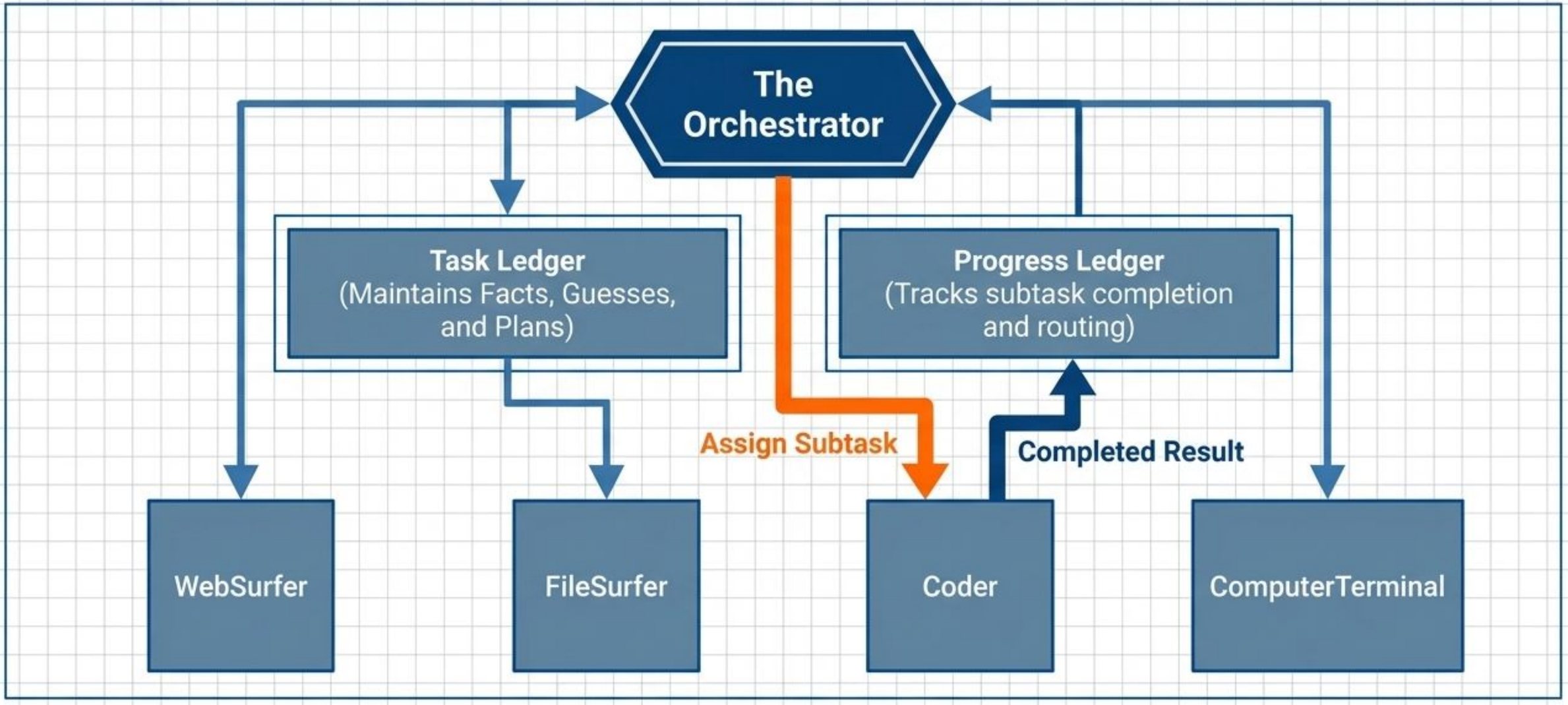
1,400+ built-in connectors

**Code-First Orchestration (Semantic Kernel)**

Deep local execution



**Generalist multi-agent systems** delegate specialized tasks to optimize performance and reduce single-agent cognitive overload

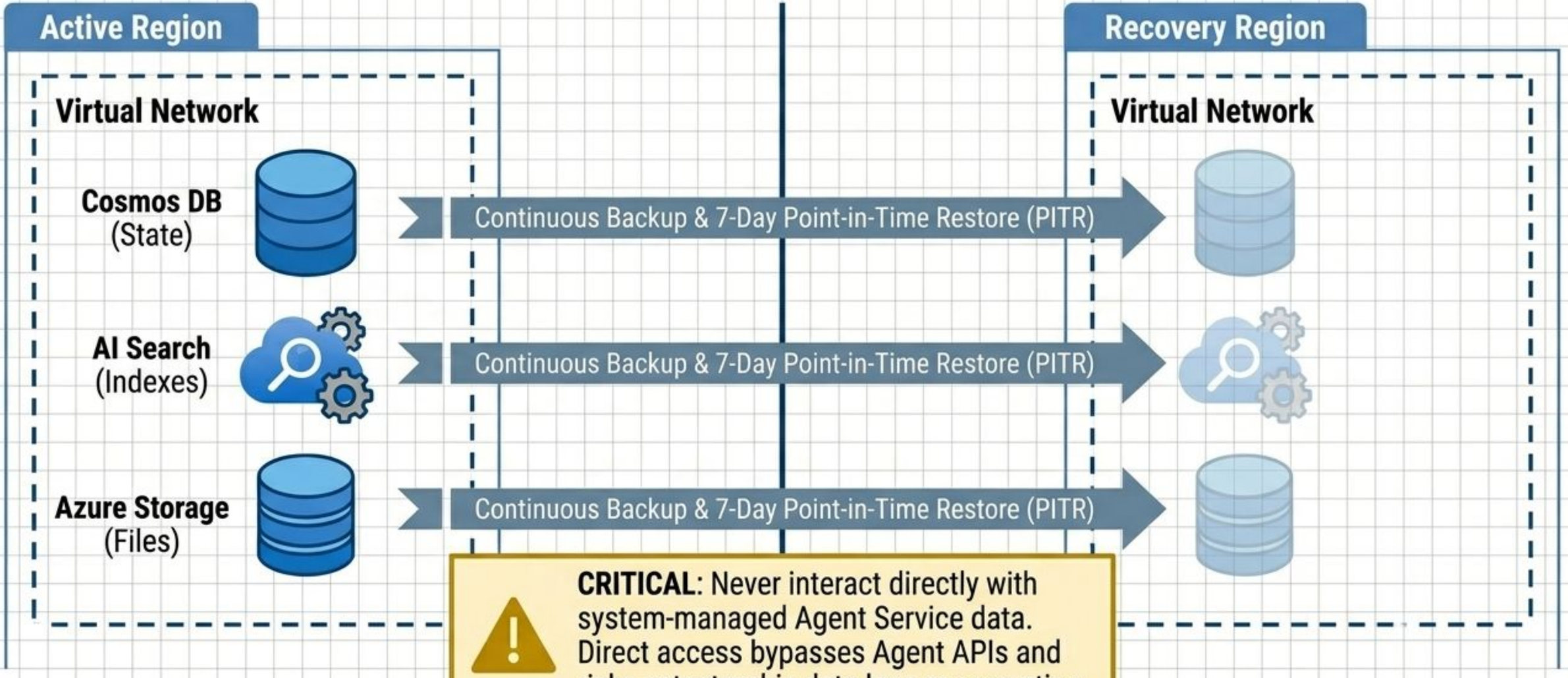


**Integrating agents** into the **daily flow of work** requires choosing the right extensibility mechanism for your data

| Extensibility Type | Core Capability | Underlying Mechanism |
|---|---|---|
| **Graph Connectors** | Knowledge retrieval & semantic summarization | Ingests unstructured data into Microsoft Graph |
| **API Plugins** | Real-time updates, triggers, & transactions | Interacts directly with REST APIs or MCP servers |
| **Declarative Agents** | Tailored, in-context conversational experiences | Combines instructions, knowledge, and actions via Copilot |

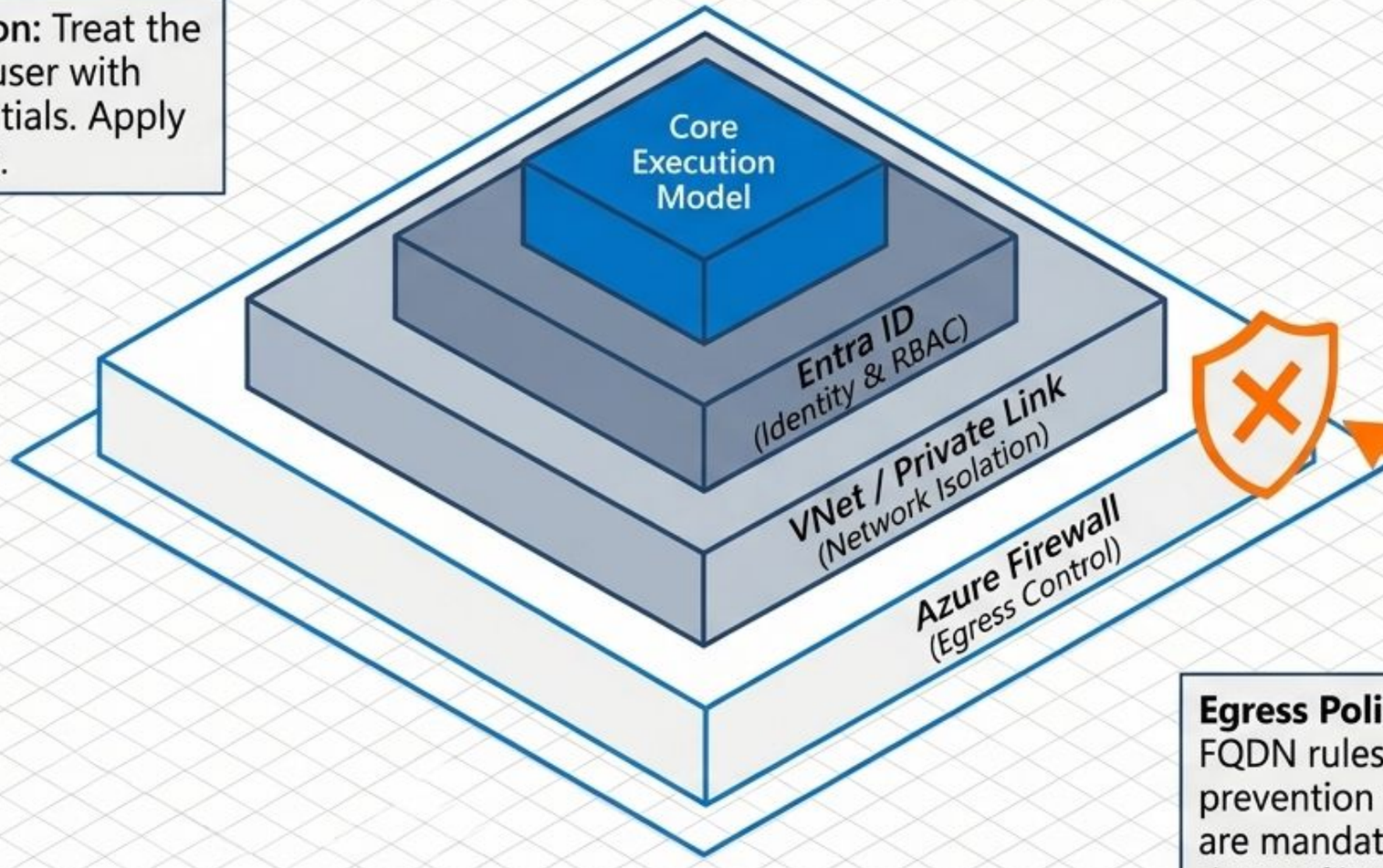
**Critical Distinction:**
API Plugins are the sole mechanism capable of real-time transactional updates and active system modification within the M365 environment.

**Enterprise reliability** requires strict isolation of agent dependencies and asynchronous geo-redundancy



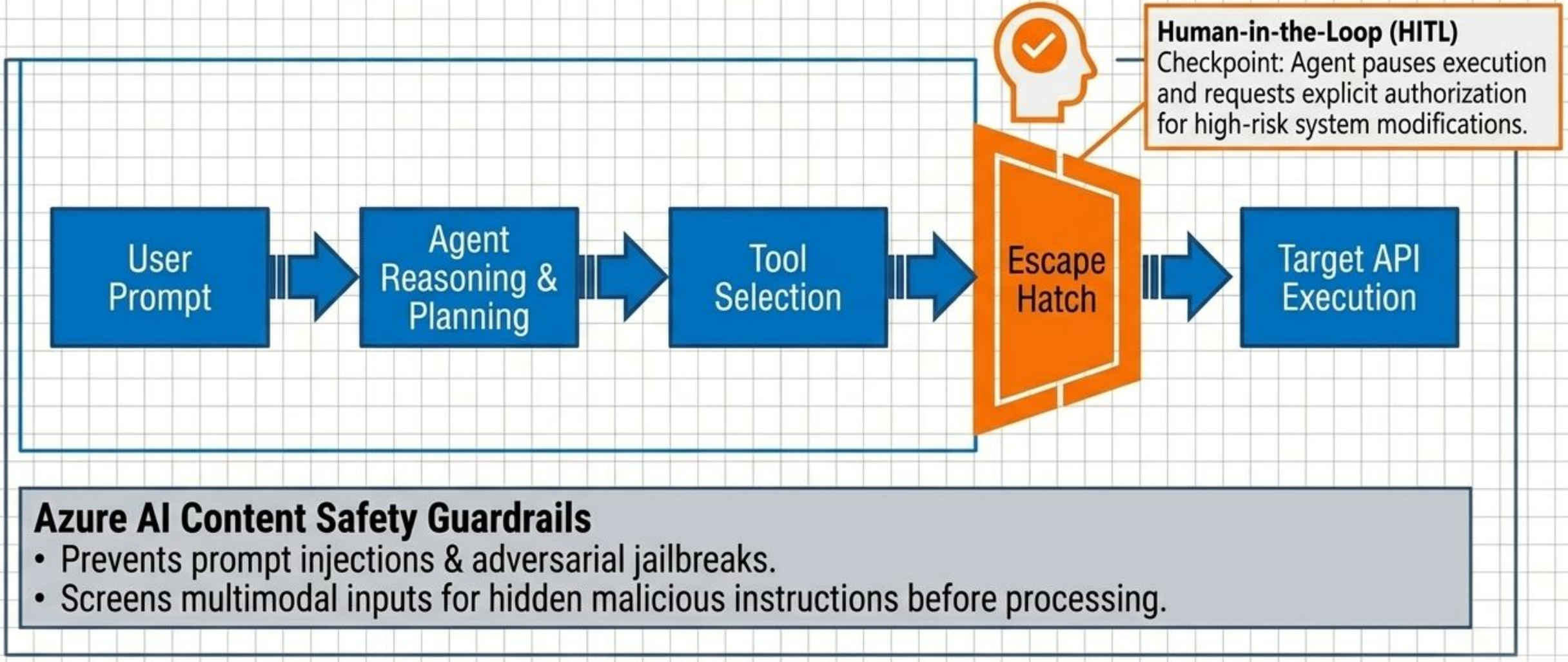
# Autonomous agents require a shift in security from index protection to managing the blast radius of a digital worker

**Zero Trust Assumption:** Treat the agent as an internal user with compromised credentials. Apply least-privilege access.




**Egress Policy Enforcement:** Strict FQDN rules and SNAT exhaustion prevention via multiple public IPs are mandatory to stop unauthorized data exfiltration.

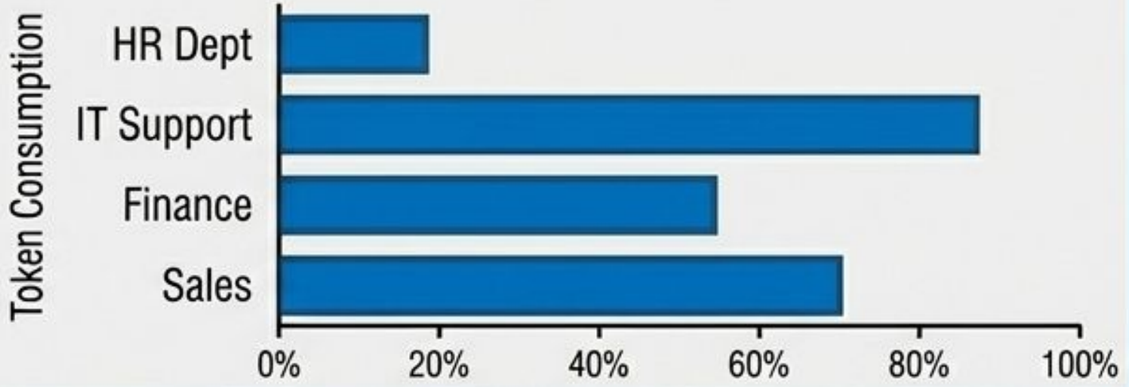
# High-stakes autonomous execution necessitates mandatory human-in-the-loop checkpoints and strict content safety guardrails



# Non-deterministic AI behavior requires continuous telemetry, cost allocation tracking, and automated CI/CD evaluation

## Performance Telemetry


## Financial Governance



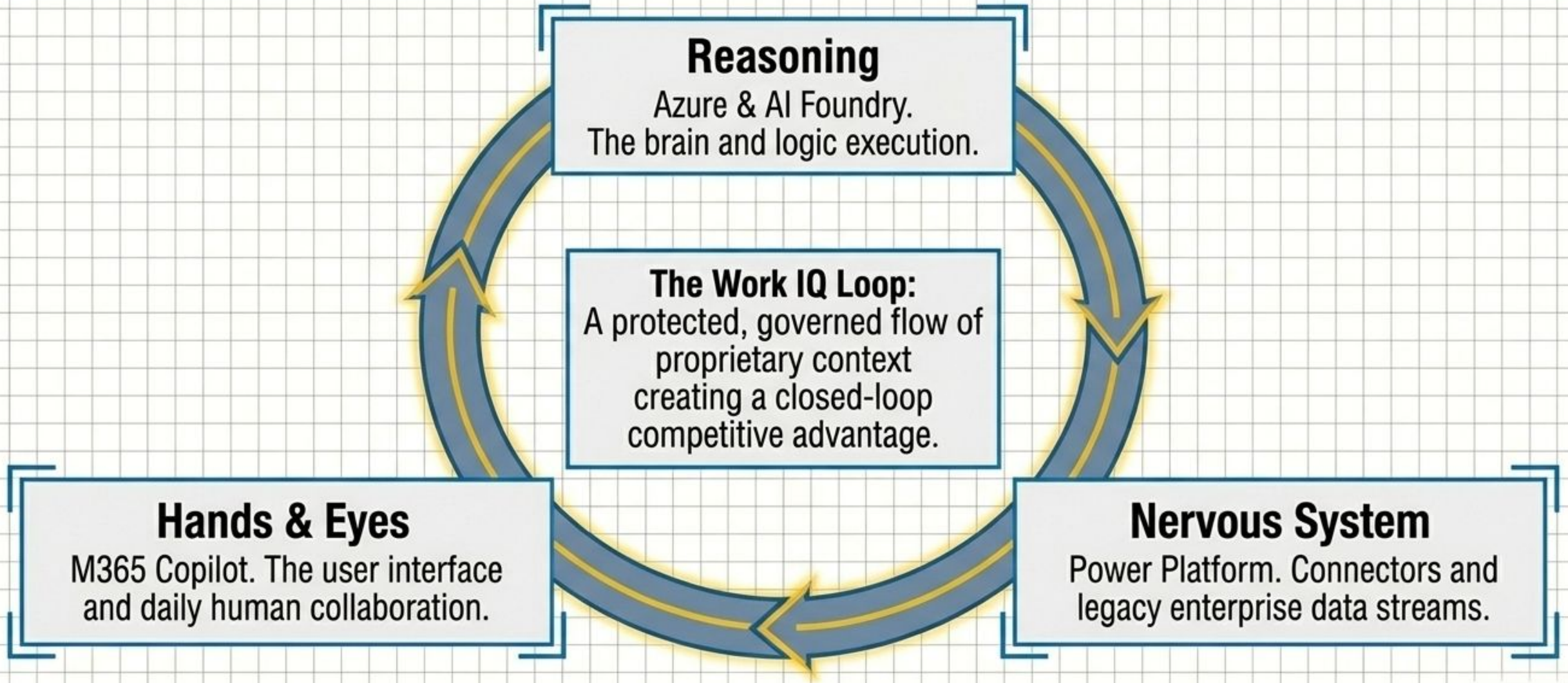

## Deployment Safety


## Evaluating Multi-Step Tasks

$$S_{rate} = \frac{\sum(C_i)}{N}$$

Where $C_i$ represents binary task completion and $N$ represents total invocations.

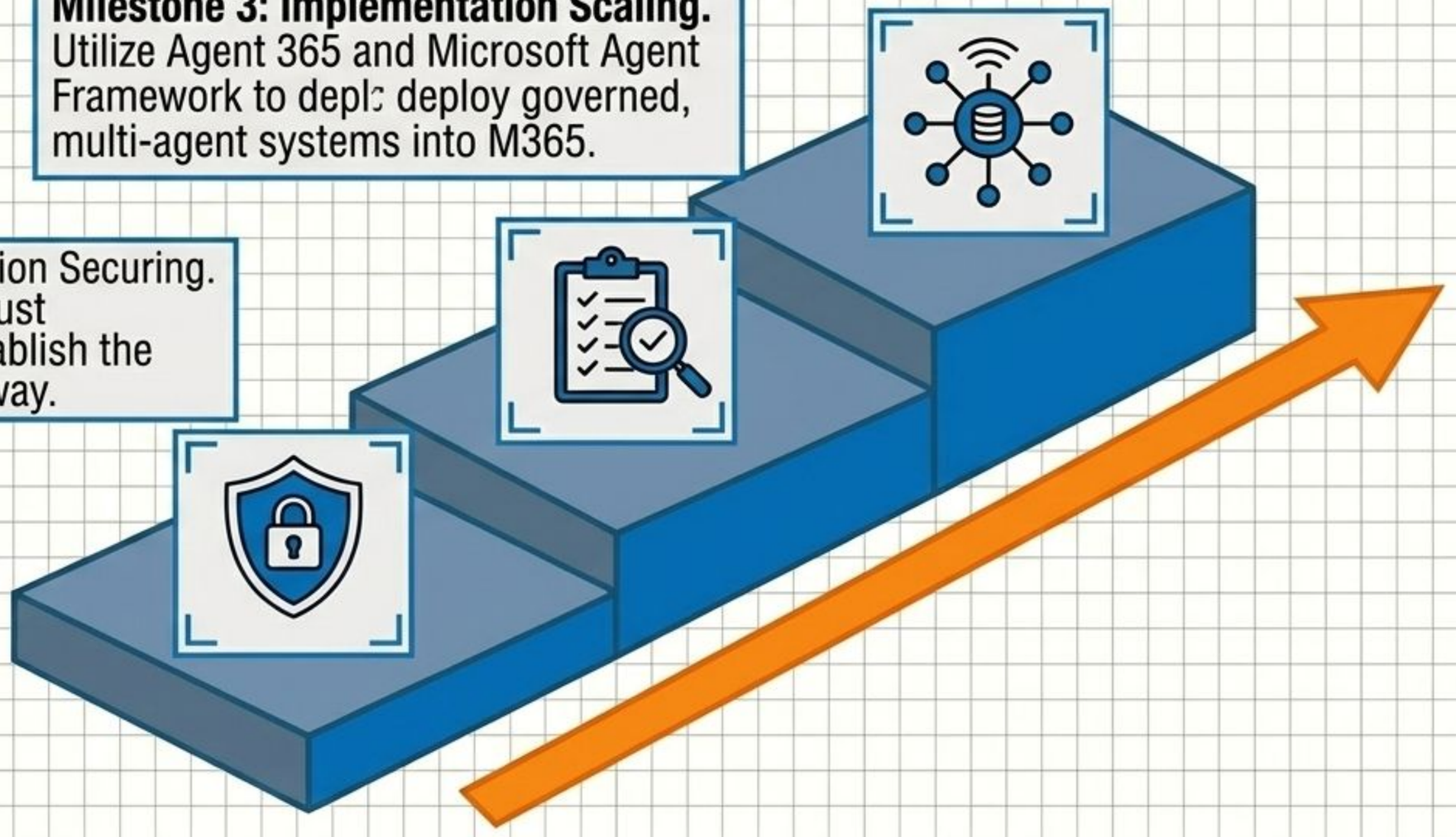
# The ultimate value of Agentic AI is an autonomous, continuously learning value chain tailored to organizational context



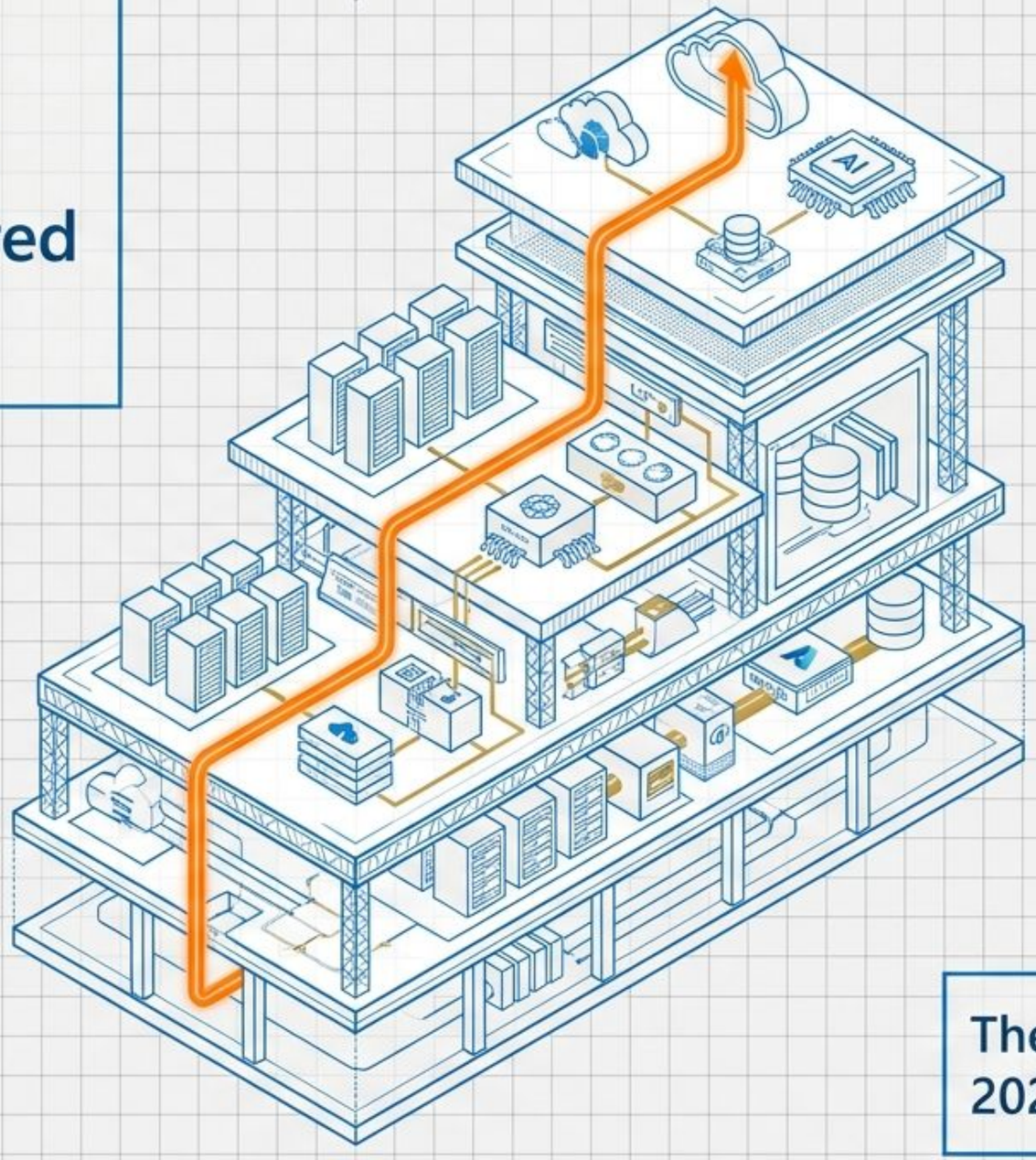
# Balance rapid innovation with structural security by following a phased, three-milestone deployment roadmap

**Milestone 3: Implementation Scaling.** Utilize Agent 365 and Microsoft Agent Framework to deplc deploy governed, multi-agent systems into M365.

Milestone 1: Foundation Securing. Transition to Zero Trust architecture and establish the secure agentic gateway.

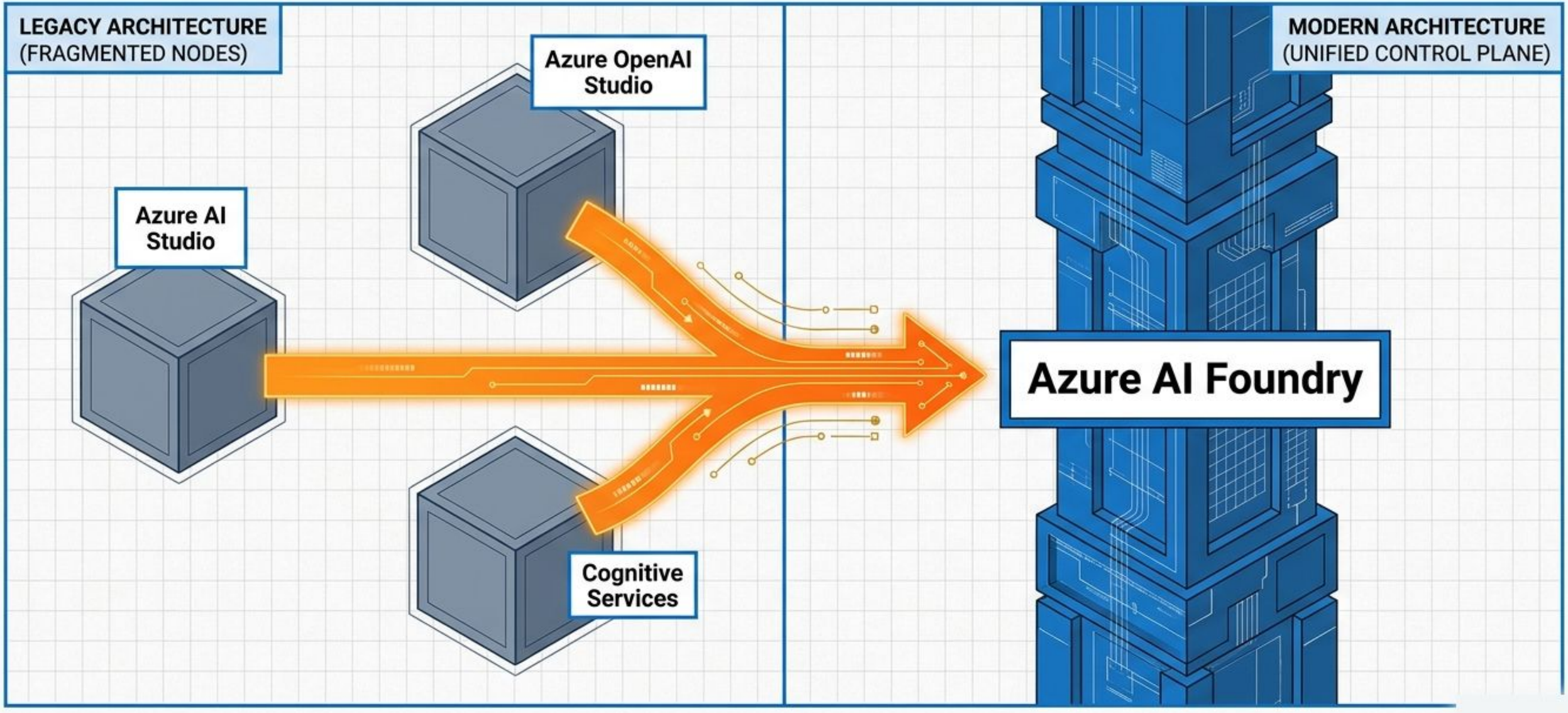


# Unified enterprise intelligence requires a meticulously engineered architecture




The Azure AI 2025/2026 Blueprint

# Azure AI Foundry unifies fragmented development nodes into a singular control plane



# Azure AI Foundry orchestrates the entire AI lifecycle from a centralized hub

## Prompt Engineering

Optimize LLM prompts with visual flow frameworks.

## Responsible AI

Apply built-in content safety and compliance guardrails.

Control Tower

## Model Catalog

Access 11,000+ pre-built and frontier models.

## Agentic Orchestration

Build and route multi-agent enterprise workflows.

# Select the right platform based on your desired level of architectural control

| Azure AI Foundry | Azure Machine Learning |
| --- | --- |
| **Target User**<br>Developers & Business Leaders | **Target User**<br>Data Scientists & ML Experts |
| **Customization**<br>Pre-built models & prompt tuning | **Customization**<br>Full architecture control & custom training ✓ |
| **Primary Use Case**<br>Rapid prototyping, Agentic apps, RAG ✓ | **Primary Use Case**<br>Deep MLOps, custom ML pipelines, scratch models |

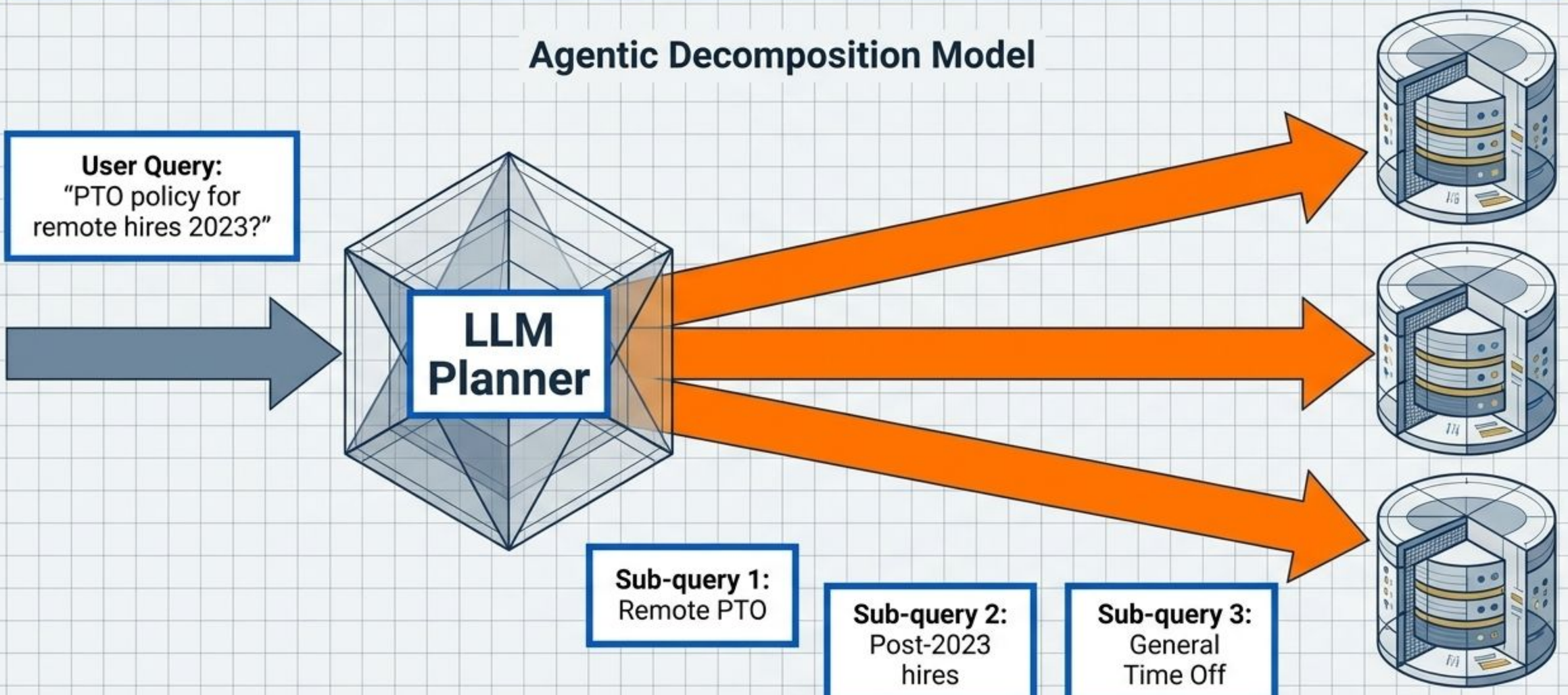
# Foundry IQ and Azure AI Search provide the unified knowledge layer for agentic reasoning


3. Grounded Agentic Output

2. Universal Retrieval Engine

1. Multi-source Data Access

# Agentic retrieval shatters vague queries into precise, parallel execution paths

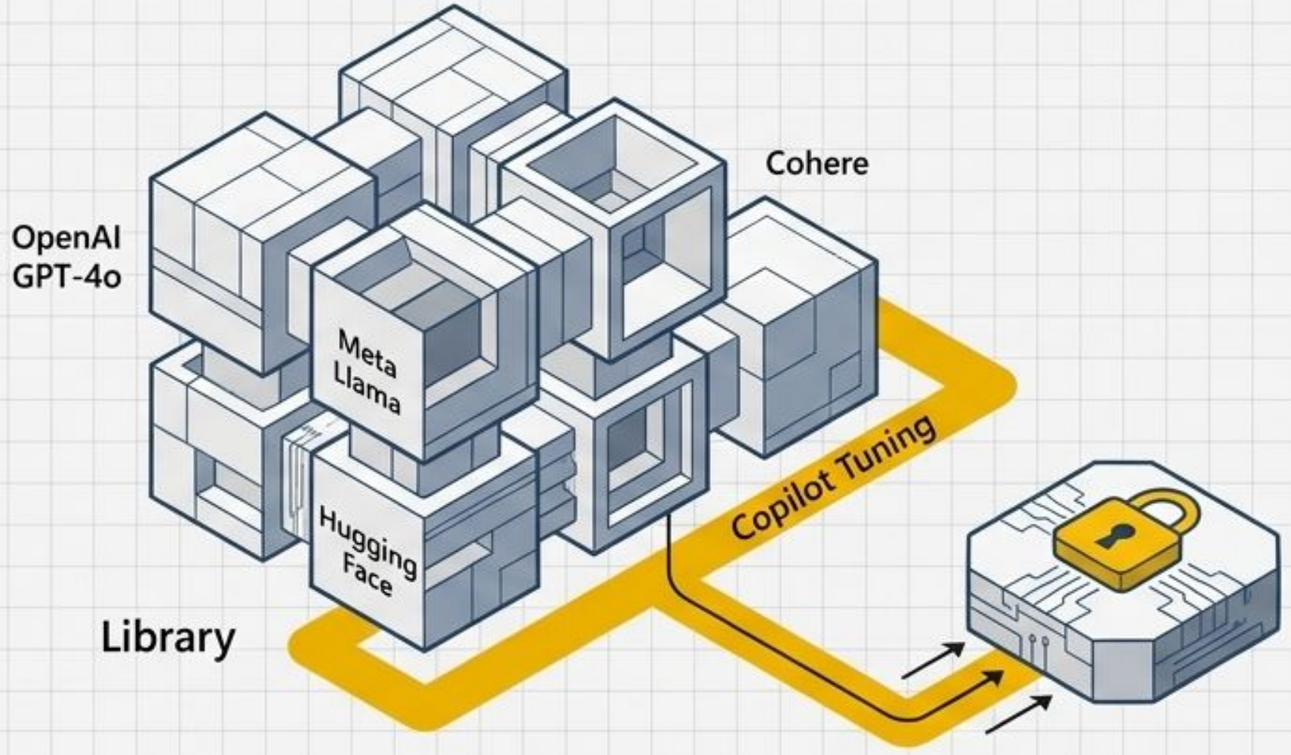


# Agentic RAG supersedes Classic RAG by automating query execution and reasoning

## The RAG Architecture Matrix

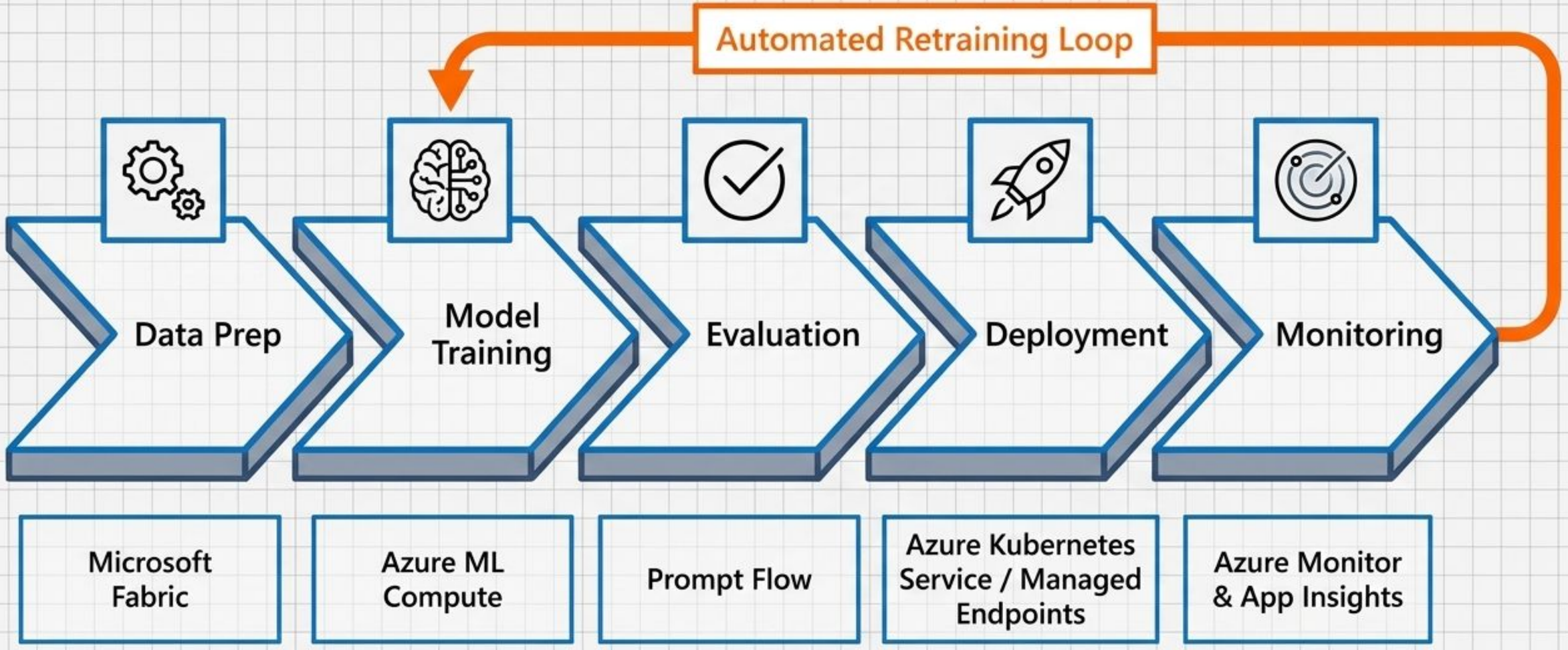
| Classic RAG | Agentic RAG |
| --- | --- |
| Single-shot query execution | ✓ Context-aware query planning |
| Keyword and vector hybrid matches | ✓ Parallel multi-source execution |
| Prone to token waste on large documents | ✓ Structured responses with built-in citations |
| Manual orchestration required | ✓ Automated L2 semantic ranking |

# Deploy over 11,000 frontier models within secure, enterprise-grade boundaries




| Azure OpenAI Models | Open-Source Catalog | Copilot Tuning |
| --- | --- | --- |
| Managed, dedicated instances of frontier OpenAI reasoning models. | Serverless APIs for open-weights models, eliminating infrastructure overhead. | Secure customization of base models using proprietary enterprise data. |

# MLOps automates the transition from experimental models to production-grade deployments



# Microsoft Fabric bridges disparate data silos into a singular, AI-ready data foundation

### Native Cosmos DB Support

Millisecond reads for real-time agentic reasoning.

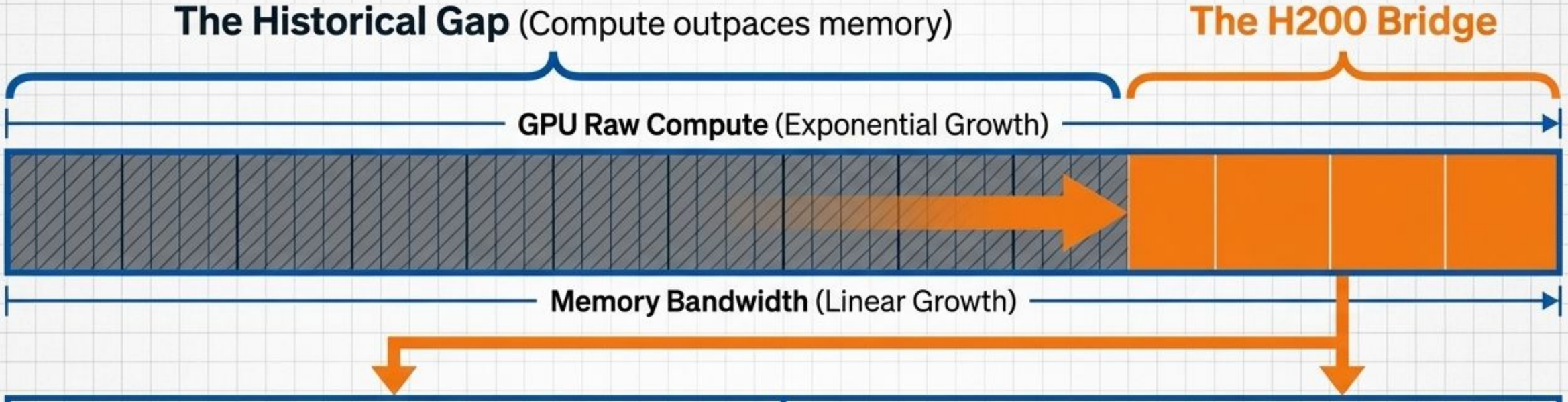
### Real-Time Telemetry

Direct ingestion of operational sensor streams.

### Cross-Domain Interoperability

One unified SaaS environment for AI data integration.

# The ND H200 v5 shatters historical memory bottlenecks for massive generative AI workloads




| 141 GB HBM | 4.8 TB/s Memory Bandwidth |
| --- | --- |
| 76% increase in High Bandwidth Memory allows larger models to fit on a single node. | 43% increase over previous generations, eliminating parameter access latency. |

# Built-in guardrails ensure safety, compliance, and data residency at every layer

### Azure AI Content Safety

Automated filters to detect and block harmful AI-generated content.

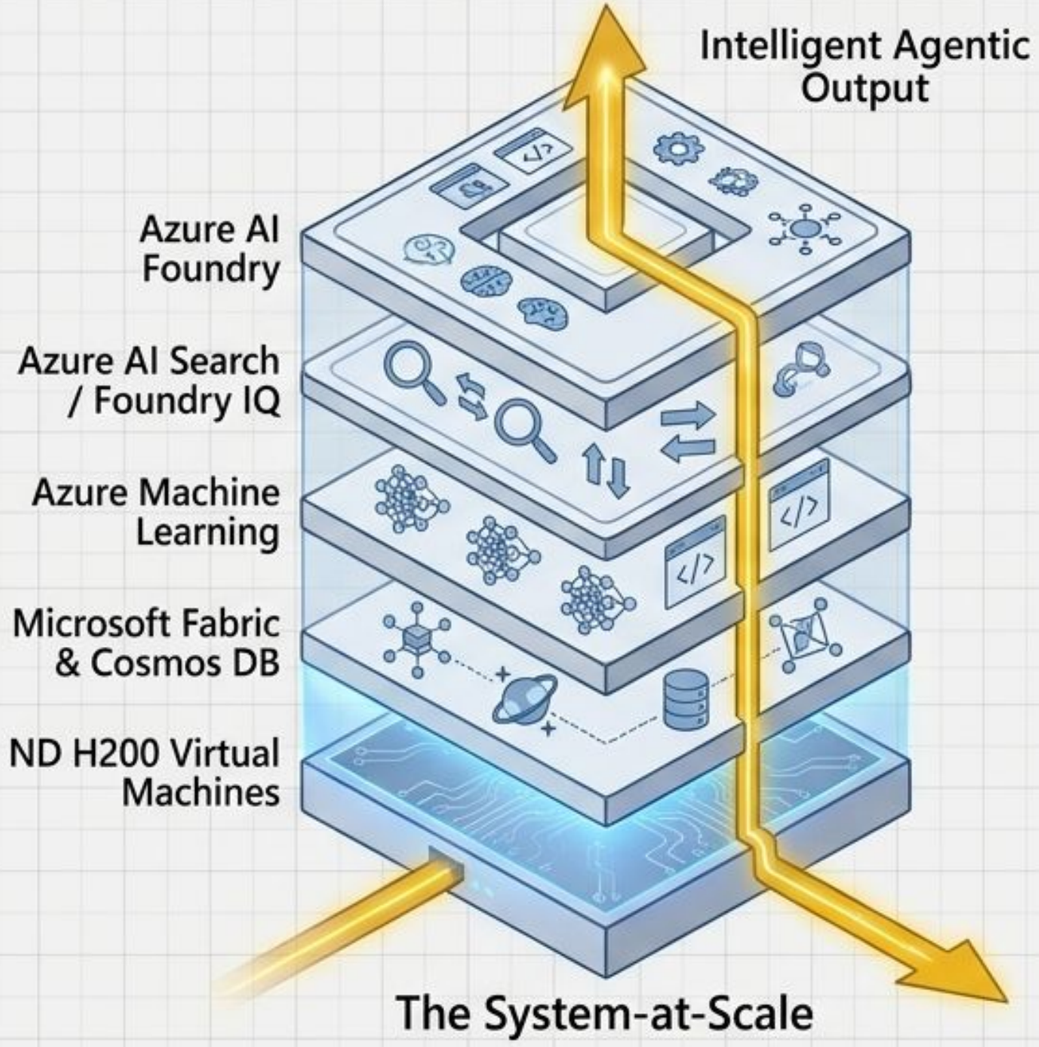
### Entra ID Integration

Granular, role-based access control protecting sensitive model endpoints.

### Network Isolation

Private VNet endpoints ensuring absolute enterprise data residency.

# The Azure AI ecosystem operates as a singular, vertically integrated intelligence machine




From silicon to agentic reasoning, **Azure eliminates integration friction**, allowing enterprises to **focus entirely on application logic** rather than **infrastructure glue**.