Mastering Azure AI Foundry with Agents

Lino Tadros, RD
Co-Founder & CEO - Tahubu

LIVE! 360
TECH EVENTS WITH PERSPECTIVE

TAHUBU

Lino Tadros

Co-founder & Chief Executive Officer

lino@tahubu.com

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

Distinguished executive leader and renowned technical expert in AI, Data, and Machine Learning. Leads cross-functional architectural teams to award-winning performance by developing strategic roadmaps and powering enterprise-wide projects. Serves as board member and advisor for multiple corporations, delivering strategic guidance on product line developments and business solutions. Industry influencer and mastermind of strategic programs and innovations, leading modernization efforts to alter the global IT landscape as a Microsoft MVP and Microsoft Regional Director for 23 years.

Lino has been in the industry for 33 years and has presented at technology conferences in 54 countries so far, with over a dozen books published and hundreds of publications.

The future of AI

AI is reshaping our world.

Not in years. Not in months. Right now.

AI-driven business transformation

Enrich employee **experiences**

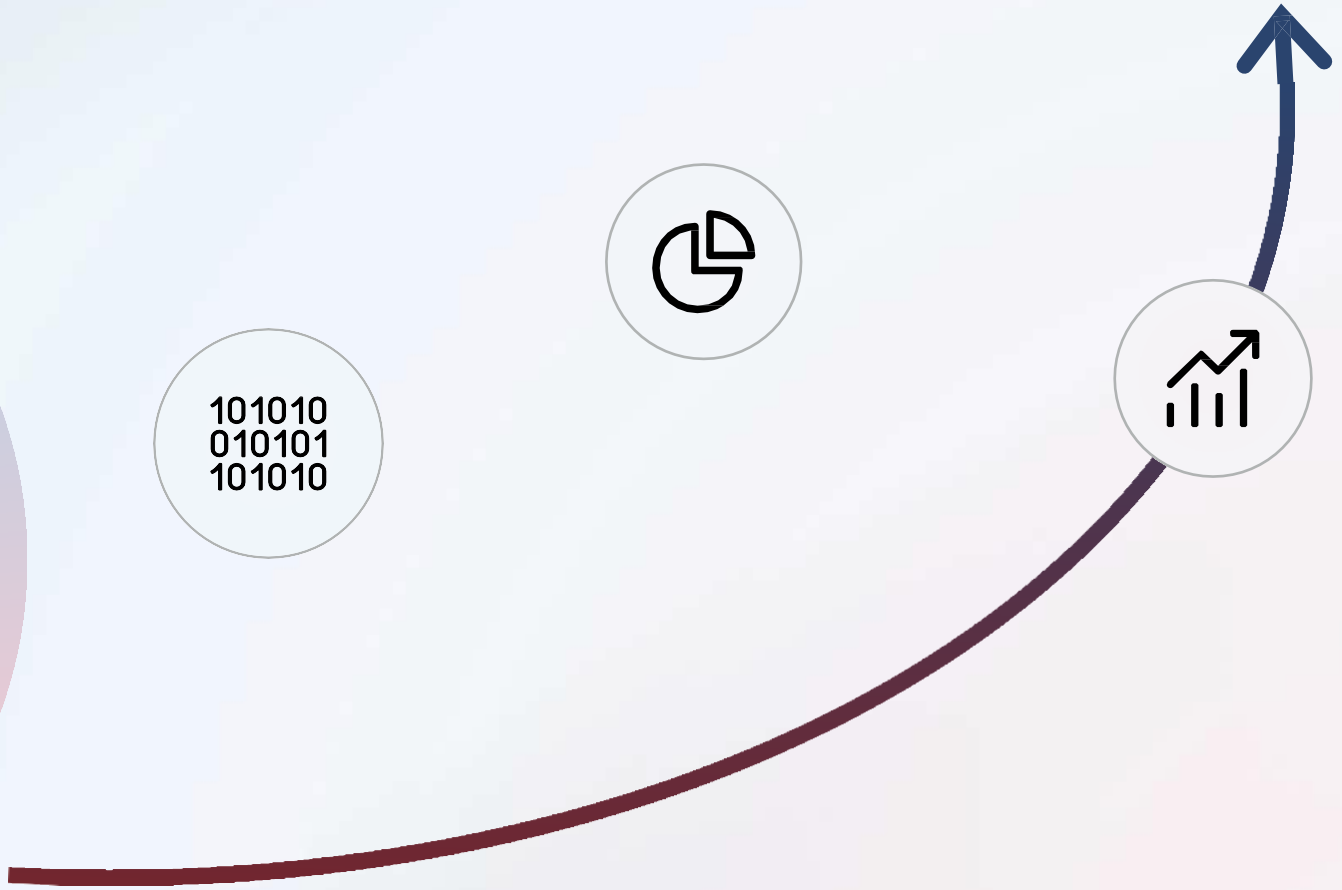
Reinvent customer **engagement**

Reshape business **processes**

Bend the curve on **innovation**

The business case for investing in AI

For every $1 a company invests, the return on investment is $3.7x.



IDC InfoBrief: sponsored by Microsoft, 2024 Business Opportunity of AI, IDC# US52699124, November 2024

Generative AI trends

81%

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

93%

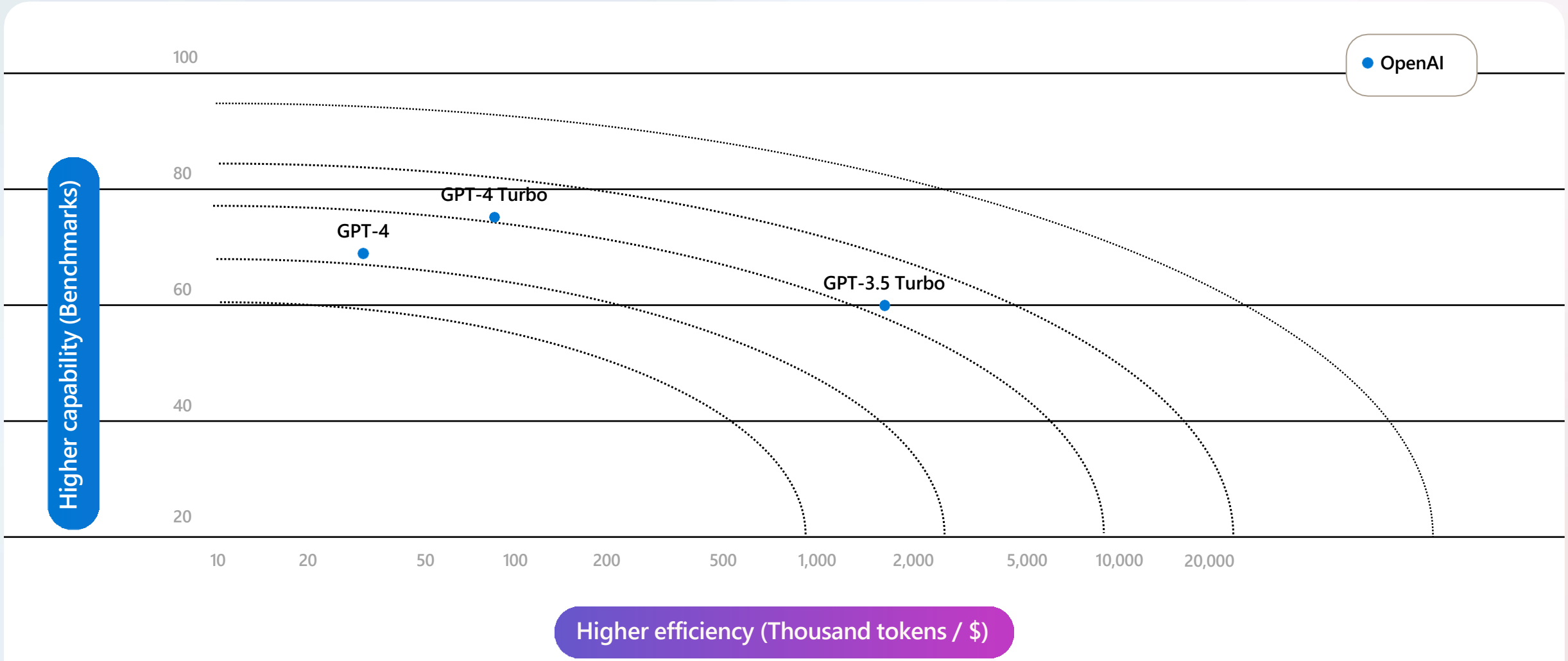
organizations are experimenting with multiple models[2]

70%

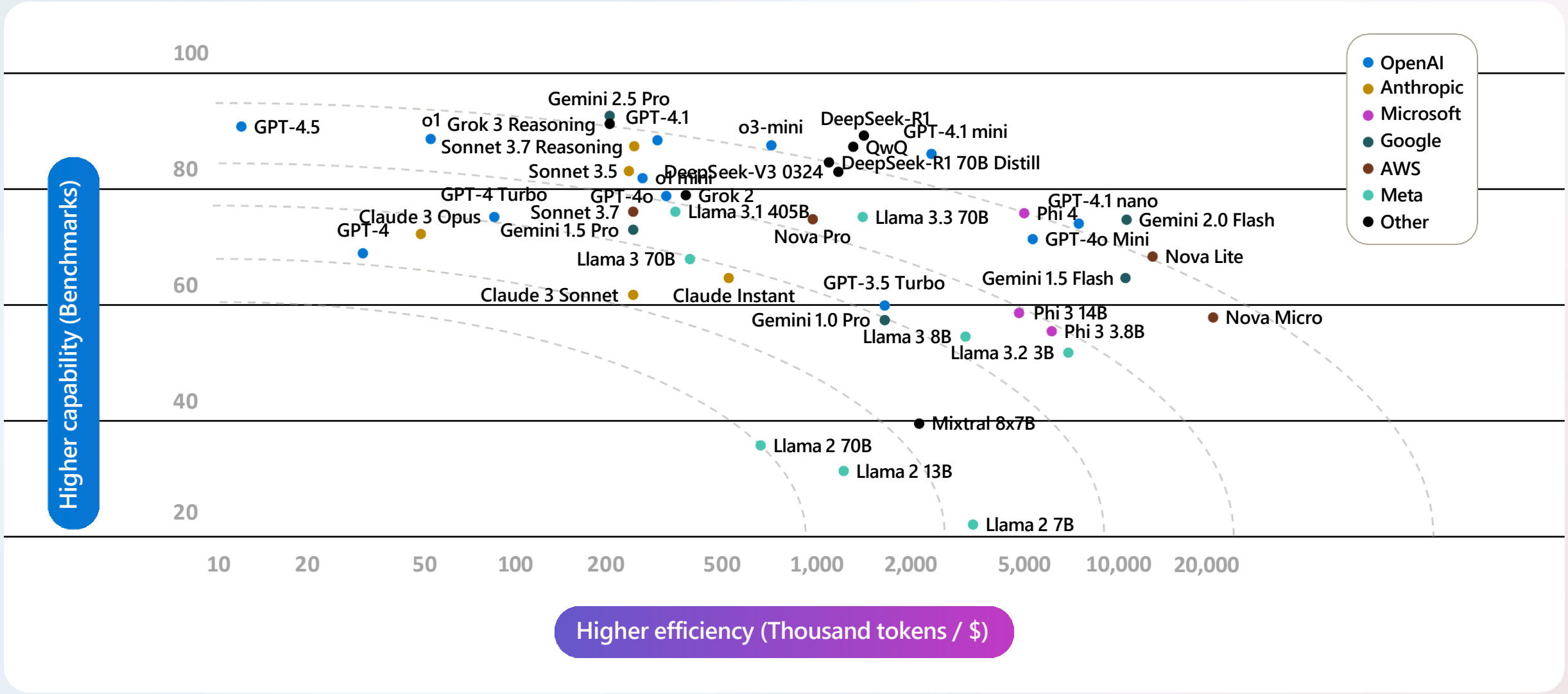
of generative AI experiments have not moved to production[3]

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

From a few foundational models…

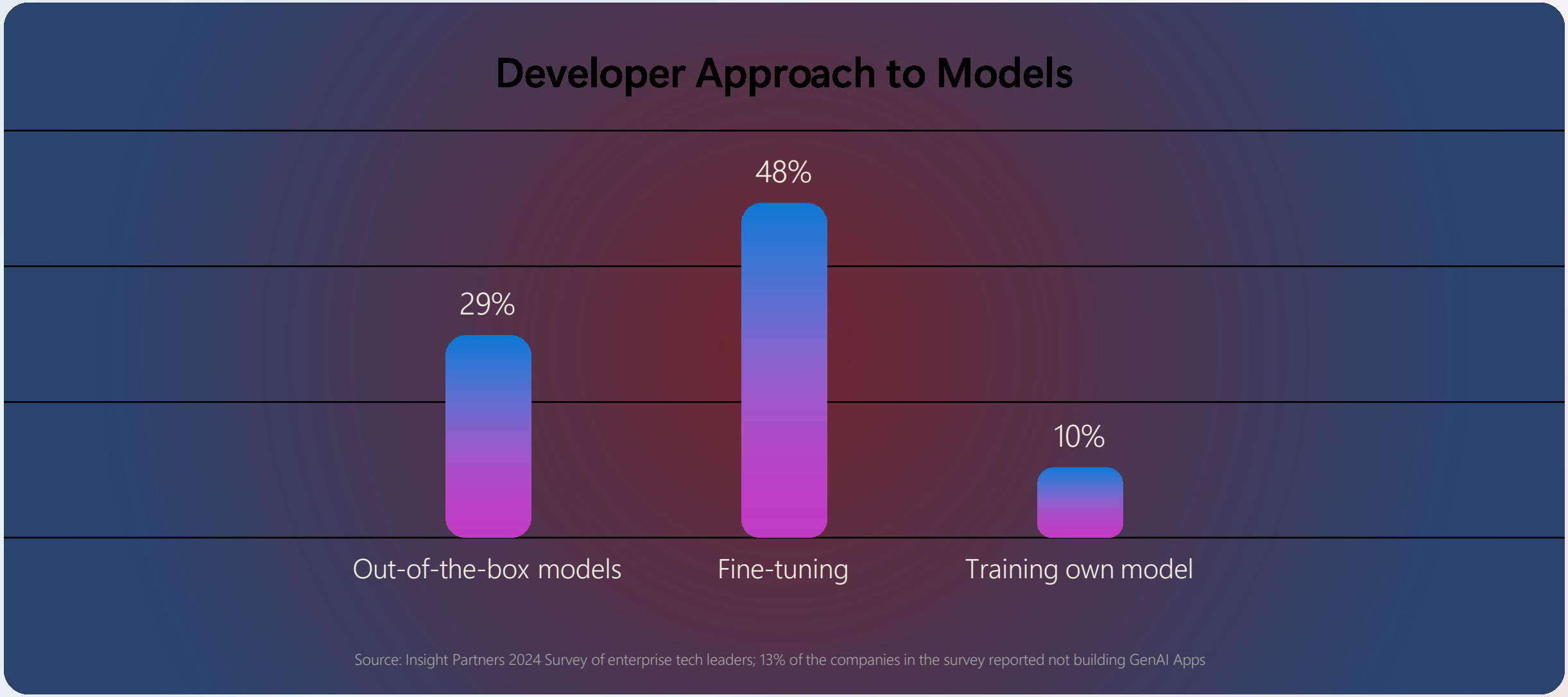


...to explosion of foundational models



Tier-1 | Babel

Fine-tuning is becoming critical



Source: Insight Partners 2024 Survey of enterprise tech leaders; 13% of the companies in the survey reported not building GenAI Apps

Agentic AI is the new frontier

Yesterday

Automation

Completes repetitive tasks

"Inputs new hire information into HR System"

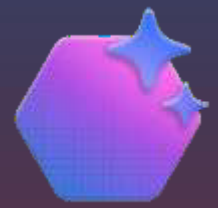
Assistants

Answers & suggests

"What is the dress code in the office?"

Today

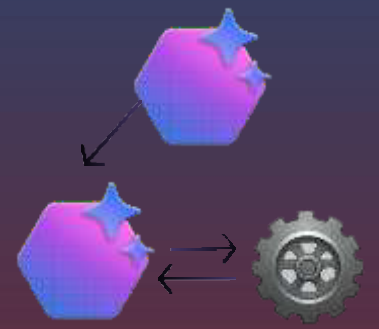
Agents



Plans, reasons & act

"Order a laptop for a new employee"

Agentic Systems

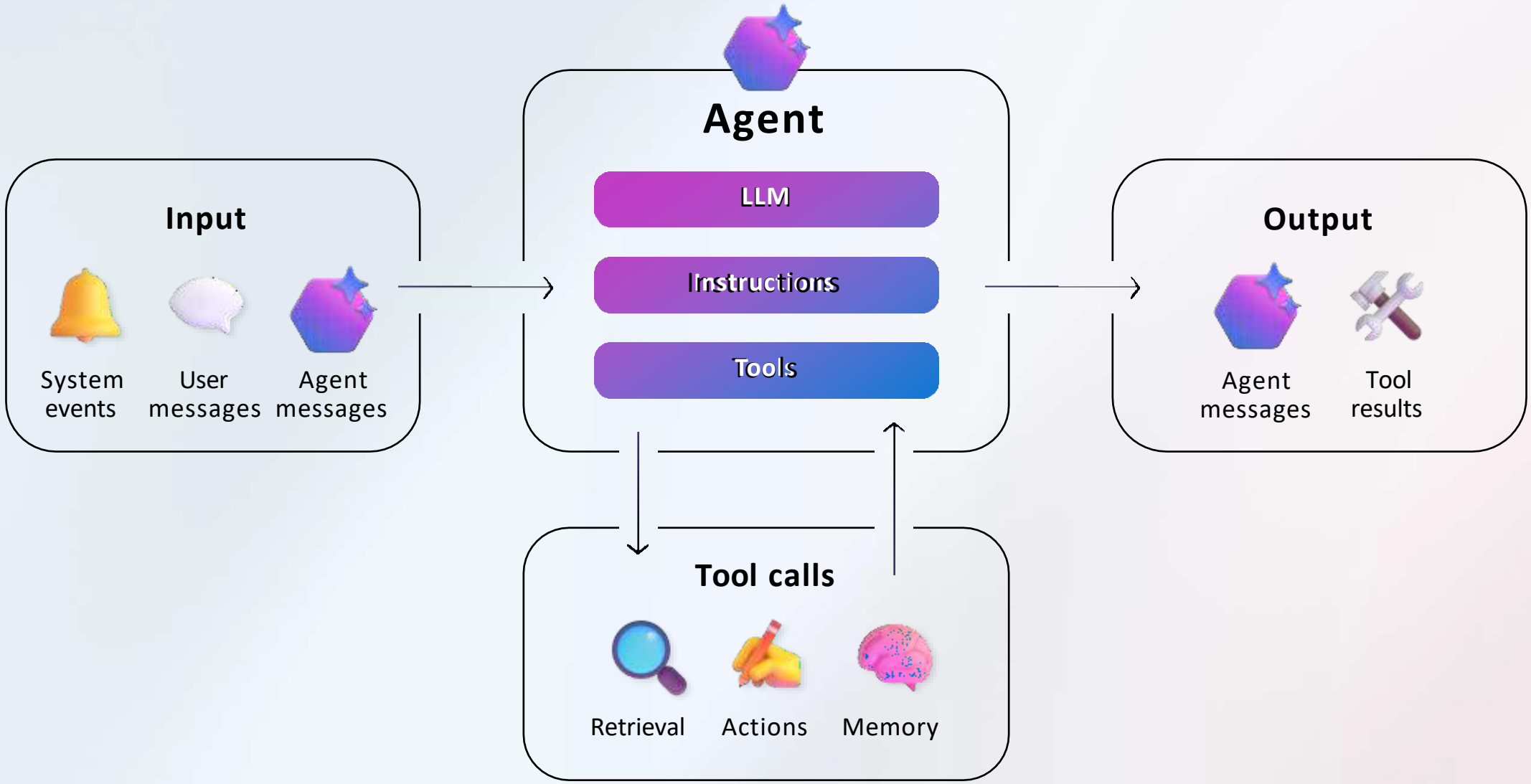


Solves complex tasks

"Onboard 5 employees by Monday"

VALUE

What is an agent?



Source: What Are AI Agents, and How Do They Help Businesses?

Agents are already changing the way we…

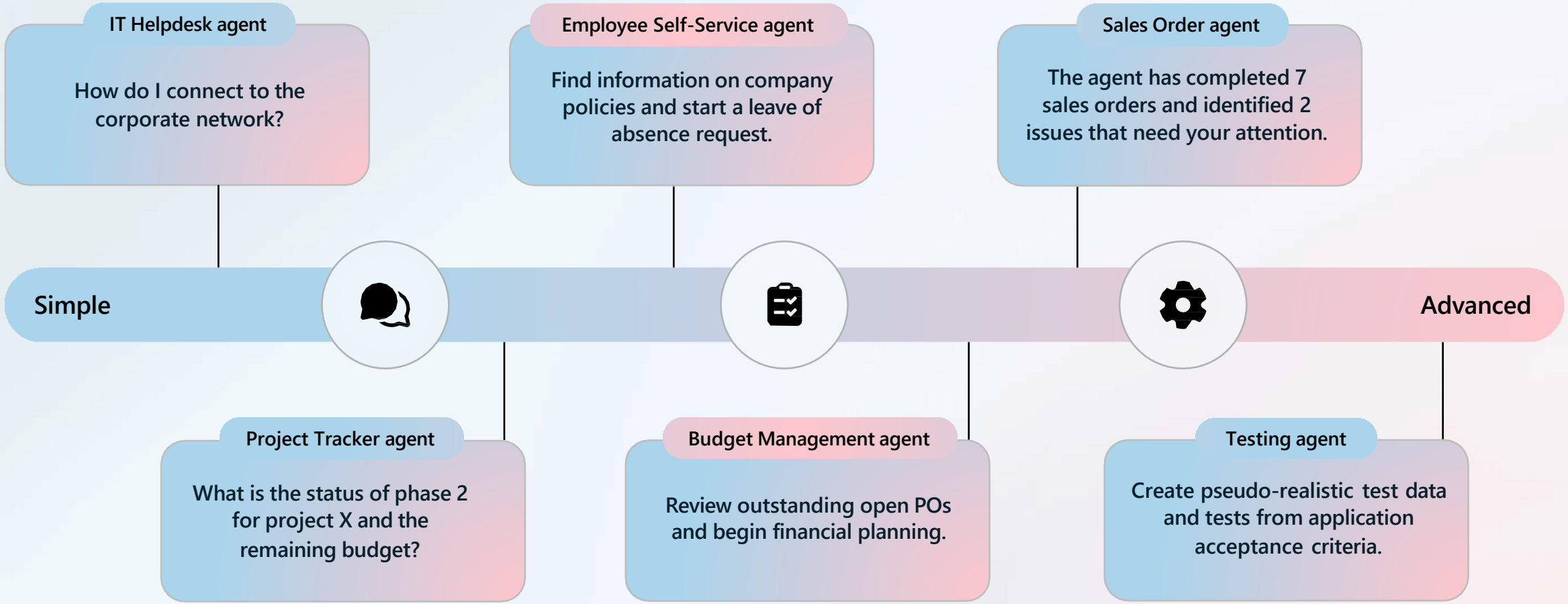
Build and
manage AI

Interact with
software

Gain value
from data

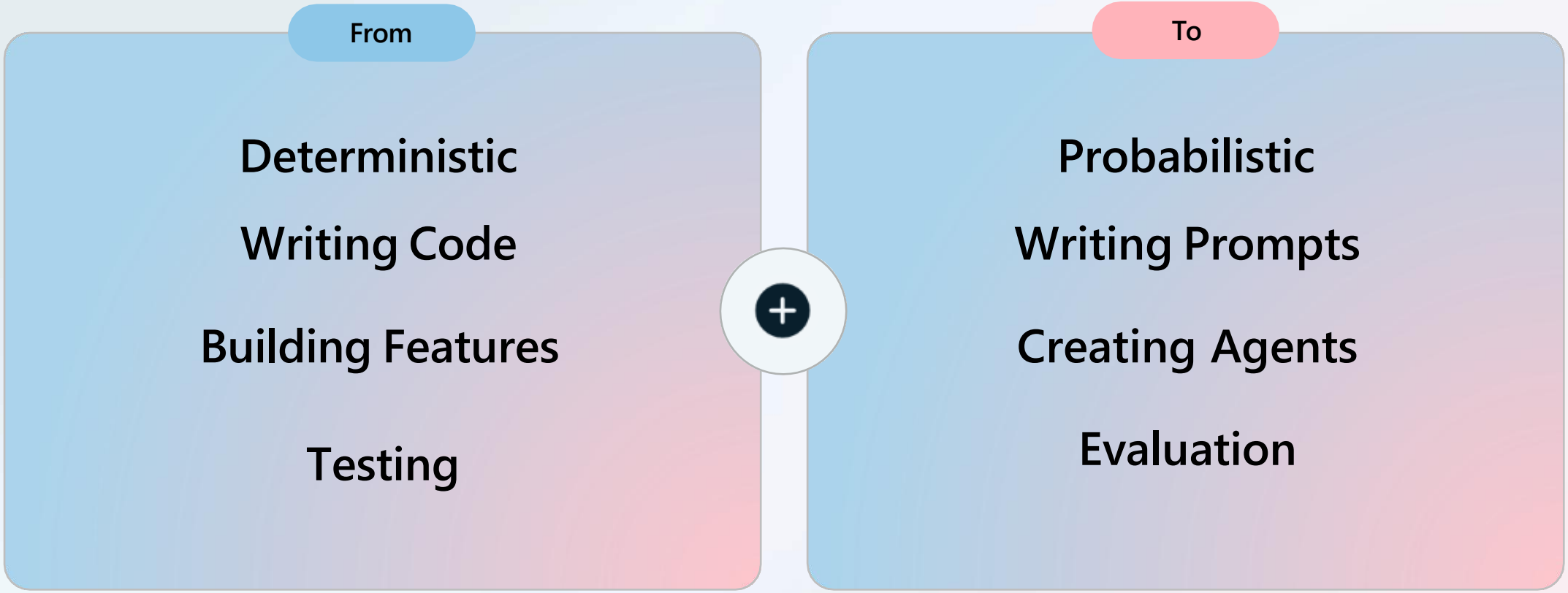
Solve complex
challenges

…with customized solutions…



Who is using Generative AI?

...and the way we write code is changing



46%

of code is now written by AI

Boosting developer productivity by

55%

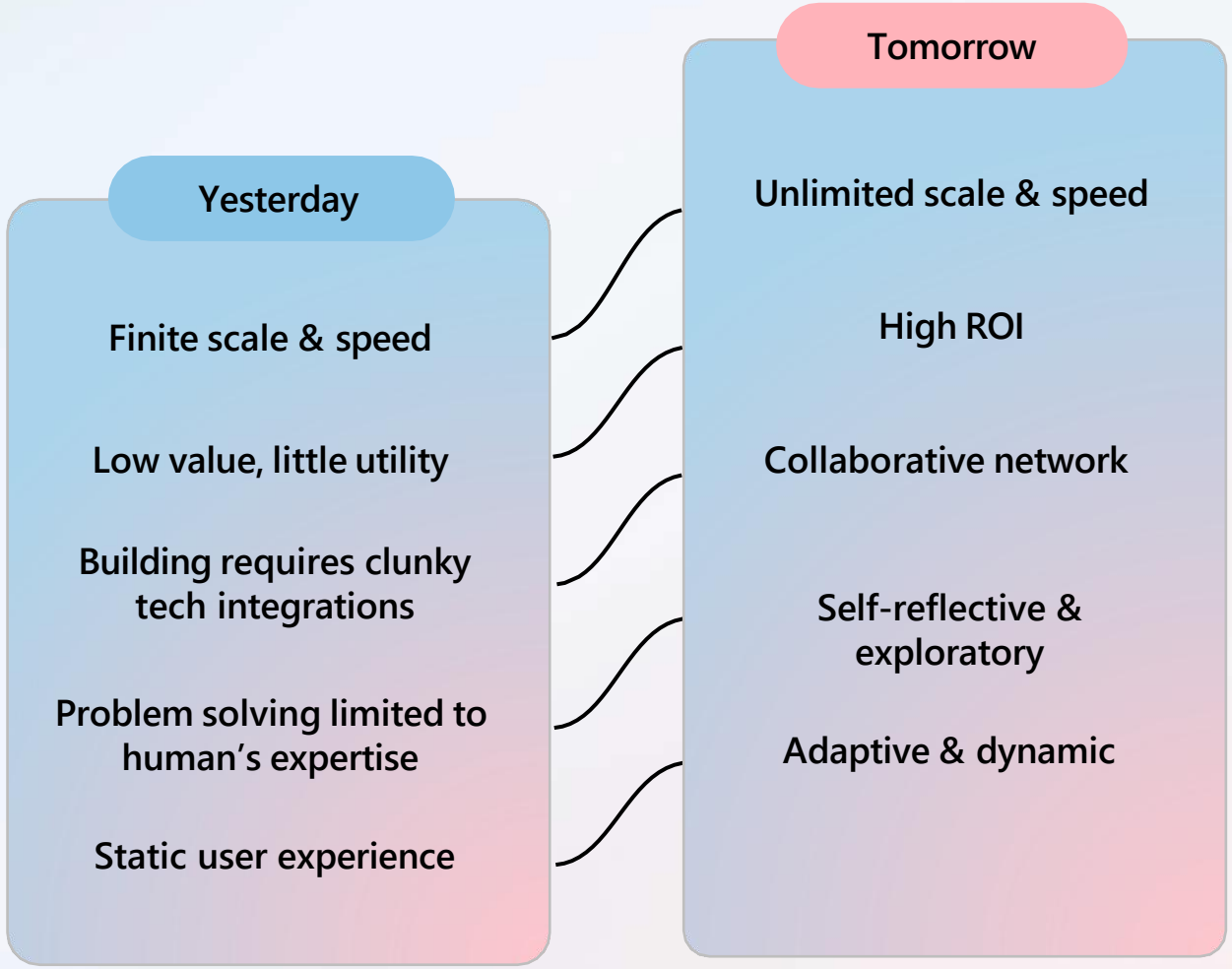
AI tools | Vibe coding | Low code

Source: GitHub Copilot for Business

Every application achieves more with AI and agents

	From		To
Customer Service & Support	Limited business hours Costly routine inquiries Inconsistent quality	→	24/7 availability Routine automation savings Quality standardization
Sales & Marketing	Generic campaigns and content Inefficient lead qualification Reactive sales strategies	→	Personalized engagement Predictive lead management Dynamic, real-time selling
Finance & Business Operations	Manual data entry, errors, delays Post-transaction risk management Resource-intensive auditing	→	Fast and accurate reporting Real-time fraud prevention Audits in less time, less labor
Supply Chain	Inaccurate forecasting, old data Cumbersome logistics Hindered visibility and tracking	→	Accurate, predictive forecasting Intelligent route planning Continuous monitoring, optimizing
Human Resources	Manual, slow time-to-hire Generic training/development Reactive employee engagement	→	Streamlined recruitment Individualized programming Timely intervention, retention

...and they ask more of your applications



70%

or more generative AI experiments never make it to production[2]

Agentic AI development challenges

Model selection & deployment

Complex workflows

Content safety

Observability and governance

Tool sprawl

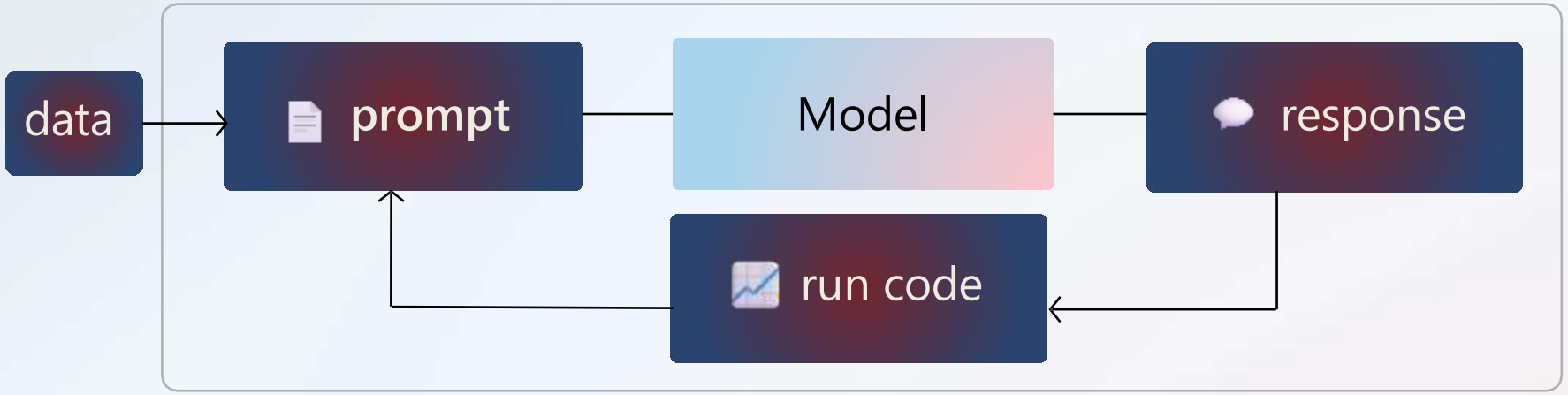
Source: Deloitte, October 2024. GenAI and the future enterprise

De-mystifying AI App Development

At a basic level

Providing prompt to a model

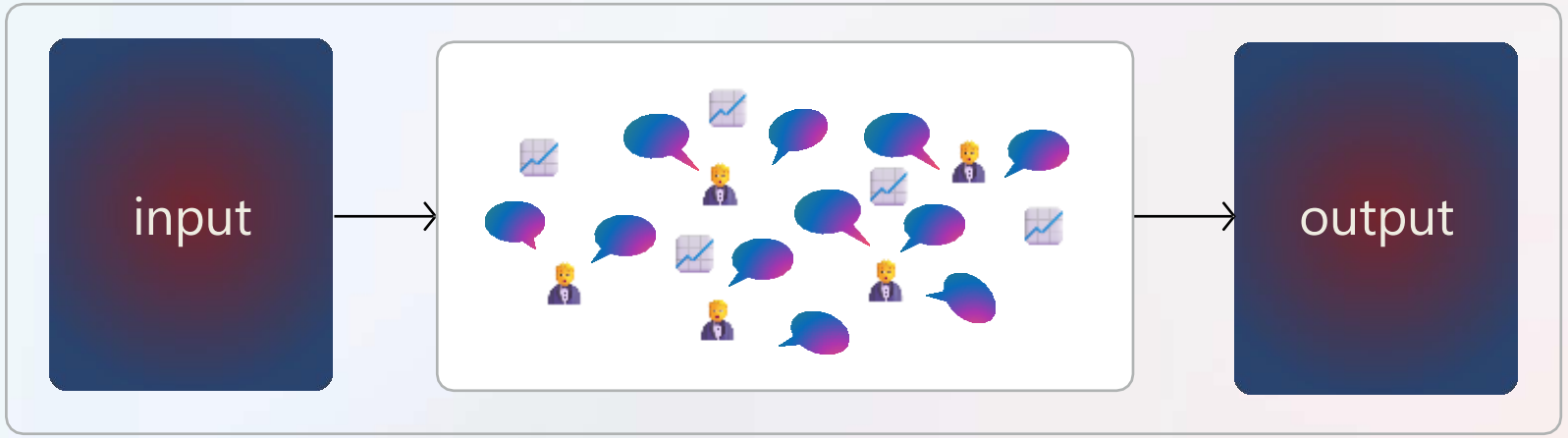
Acting on the response



At scale

Holding many different conversations with models & agents

Extracting information into useful results



Developer needs

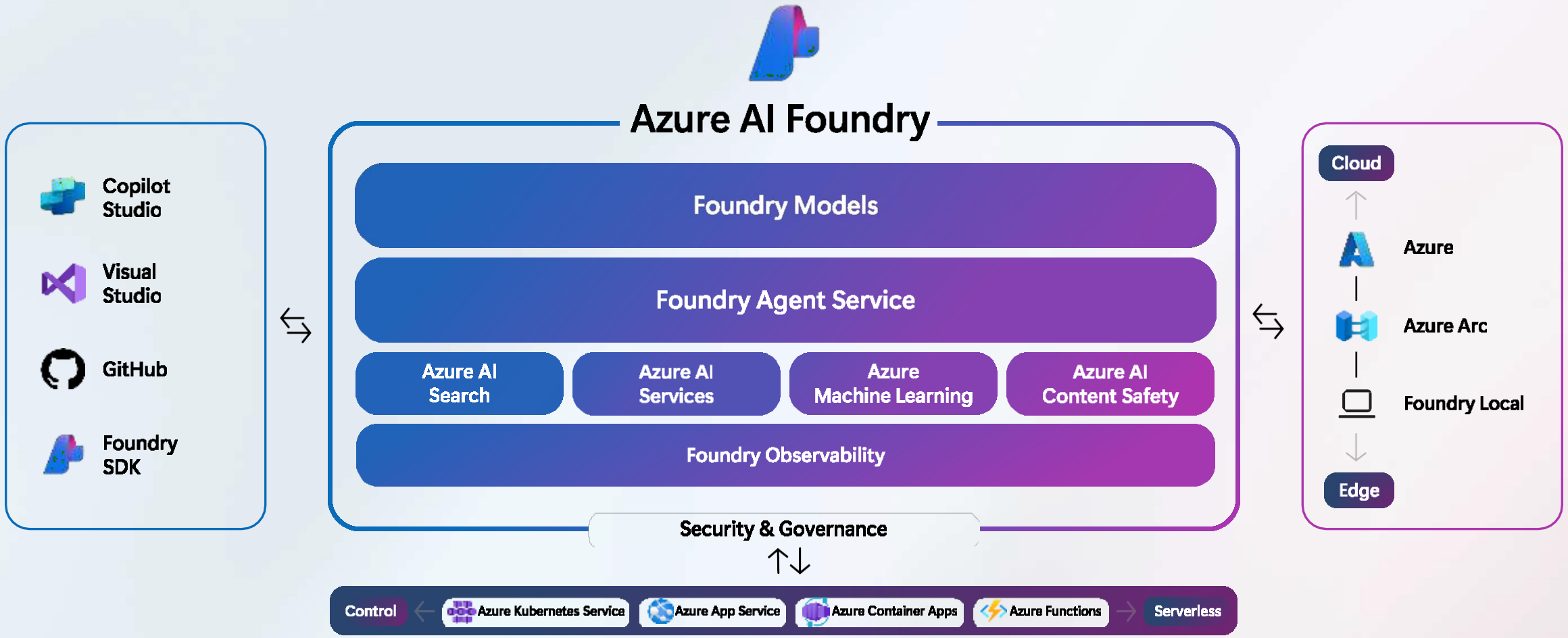
Rapidly prototype, build, and deploy to share concepts with customers

Easy experimentation and evaluation to quickly test prompts, models, and datasets without waiting for provisioning

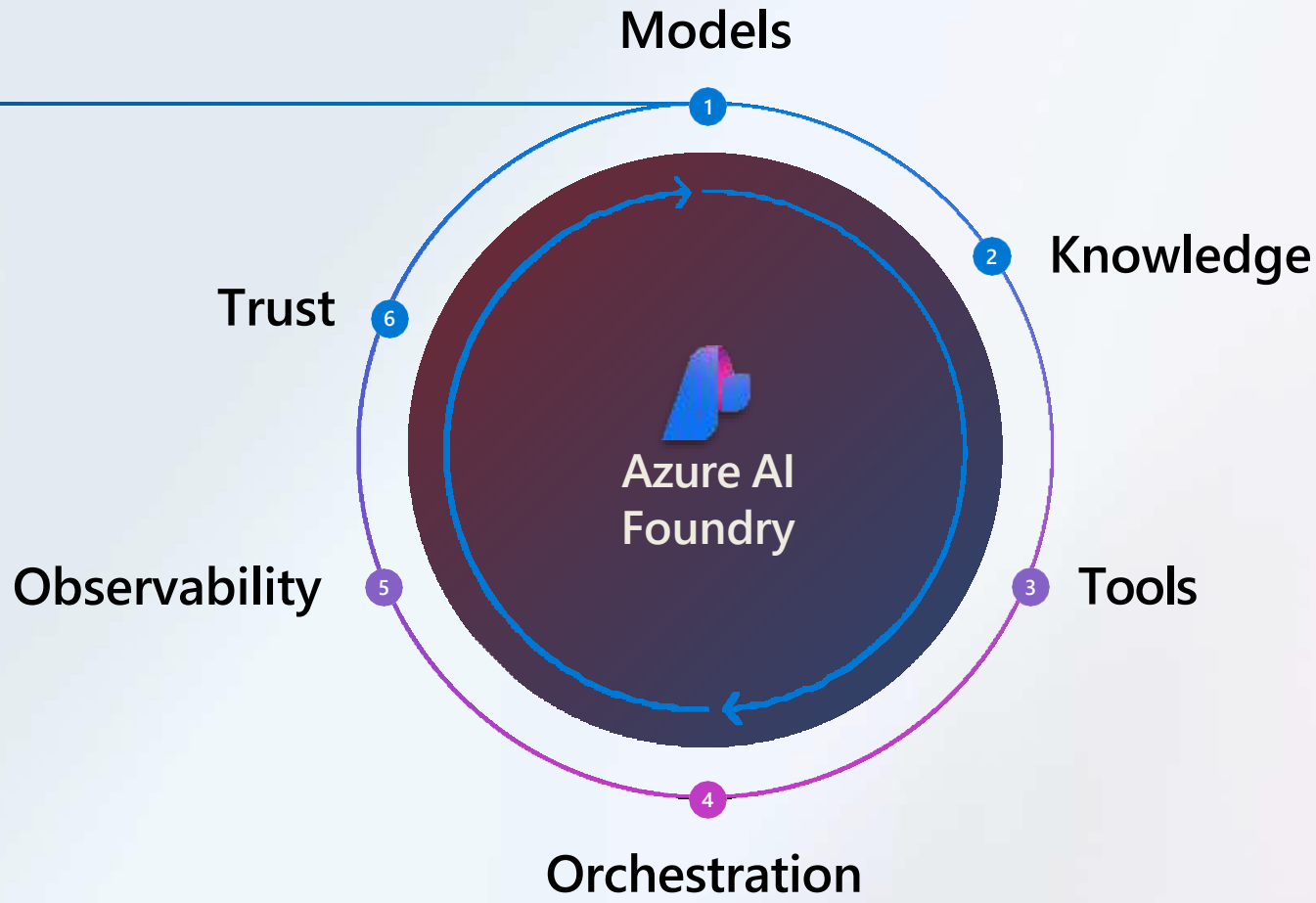
Secure, scalable, production ready AI platform to manage AI applications and agents

The AI app & agent factory

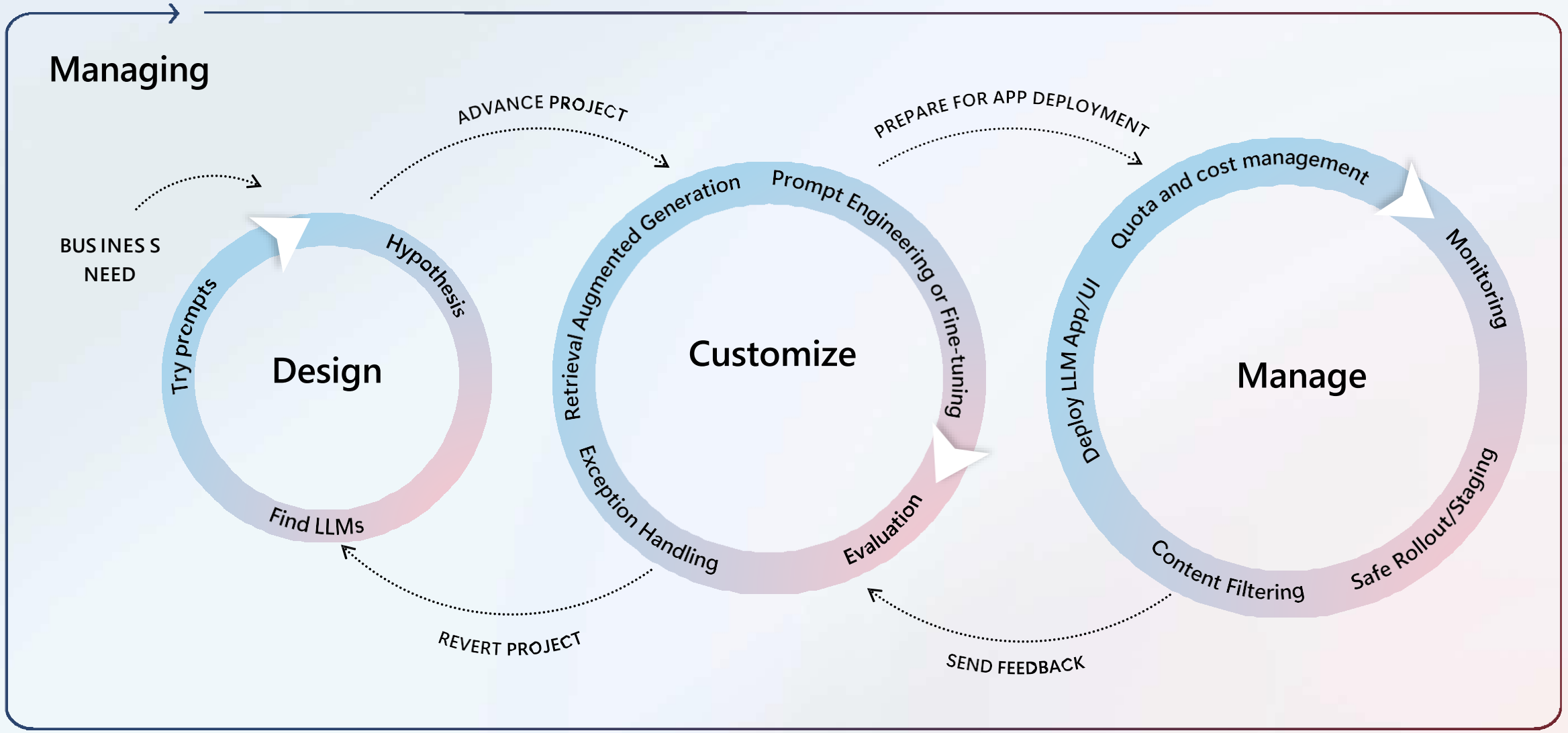
Design, customize, and manage AI apps and agents at scale

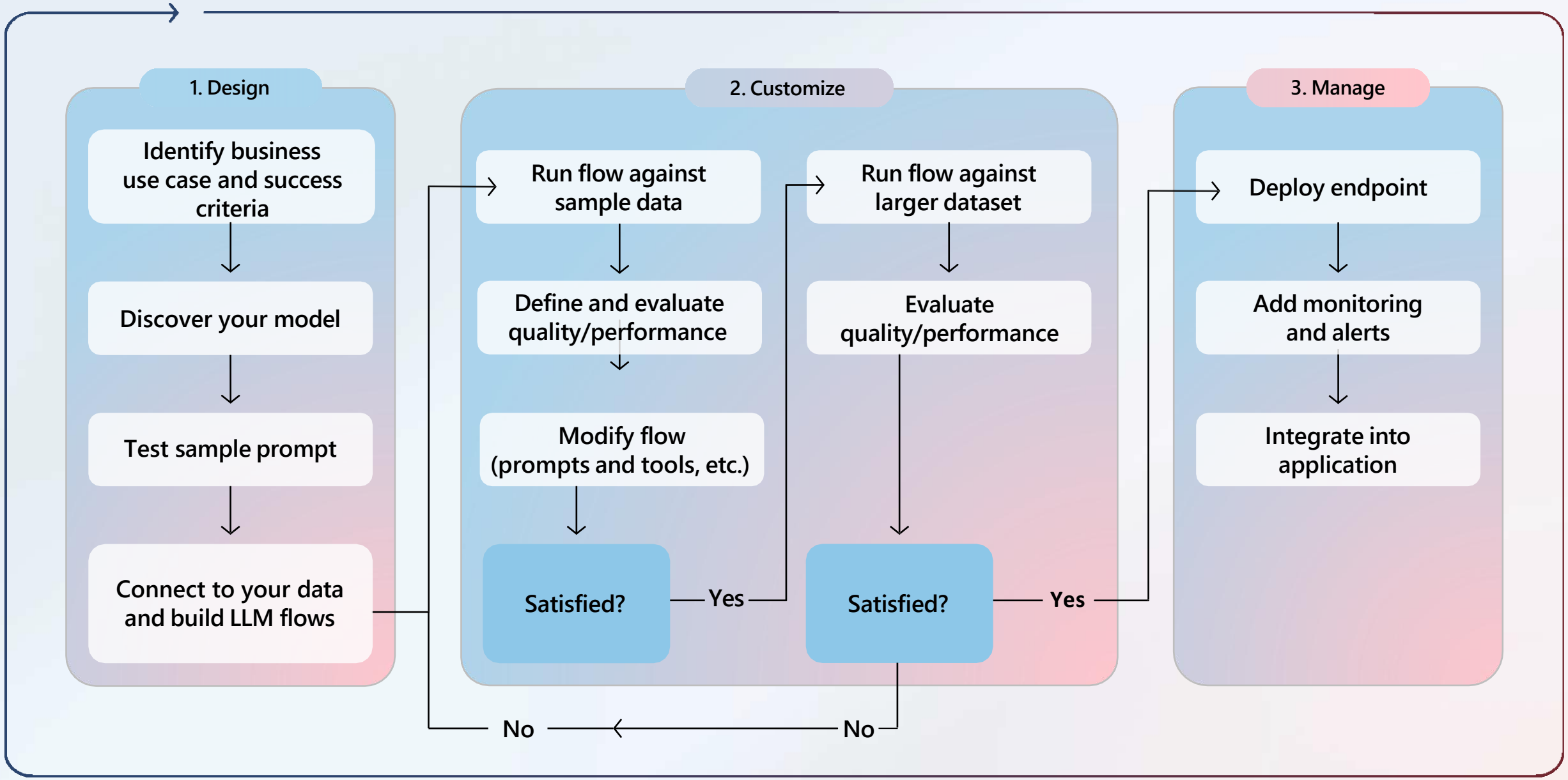

Azure AI Foundry
Copilot Studio
Visual Studio
GitHub
Foundry SDK
Foundry Models
Foundry Agent Service
Azure AI Search
Azure AI Services
Azure Machine Learning
Azure AI Content Safety
Foundry Observability
Security & Governance
Control
Azure Kubernetes Service
Azure App Service
Azure Container Apps
Azure Functions
Serverless
Cloud
Azure
Azure Arc
Foundry Local
Edge

The AI app and agent factory



AI Development Lifecycle




1. Design
Identify business use case and success criteria
Discover your model
Test sample prompt
Connect to your data and build LLM flows
2. Customize
Run flow against sample data
Define and evaluate quality/performance
Modify flow (prompts and tools, etc.)
Satisfied?
Yes
Run flow against larger dataset
Evaluate quality/performance
Satisfied?
Yes
No
No
3. Manage
Deploy endpoint
Add monitoring and alerts
Integrate into application

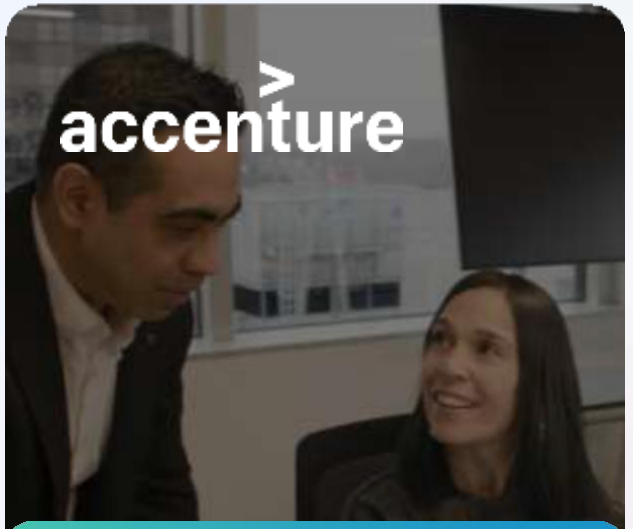
Microsoft runs on Azure AI Foundry

Microsoft Copilot	Microsoft 365 Copilot	Microsoft Security Copilot	GitHub Copilot	DAX Copilot

ROI with Azure AI Foundry

30% average productivity gain

95% reduction in configuration time



50% potential reduction in time to build AI applications

99.97% time reduction for model training

SIEMENS

"We're anticipating more problems being reported and faster resolution time, and using [Azure AI Foundry] allows us to do that."

Joe Bohman,
Executive Vice President,
Siemens Digital Industries Software

Business goal

Siemens wanted to bridge the gap between field and shop floor workers and operations and engineering teams to better drive innovation and efficiency, but more importantly, help companies rapidly address problems as they arise.

Technology solution

Using Microsoft Teams and Azure AI Foundry, Siemens created an **AI-powered collaborative app** to help support their industry leading product lifecycle management (PLM) solution, Teamcenter, and connect people who find problems with those who can fix them.

Impact

The Siemens Teamcenter for Microsoft Teams app promises to close feedback loops faster and help cross-functional teams resolve more issues faster—and together.

[Read full story here]

Key products

Foundry Models

"We really appreciate the one-click deployment of the models in Azure AI [Foundry] and that it makes Azure AI offerings transparent and available to the user."

Shah Muhammad,
Head of AI Innovation,
Sweco

Business goal

Sweco's engineers and architects manage numerous, complex tasks every day. To allow more time for more creative solutions in client projects, the company recognized the need for a supportive solution.

Technology solution

The company chose Microsoft Azure AI Foundry to build its **custom copilot, SwecoGPT, a digital assistant** that automates document creation and analysis, delivers advanced search, and provides language translation.

Impact

Though it was recently deployed, nearly half of Sweco's employees use SwecoGPT and report increased productivity, giving them more time to focus on creativity and helping customers.

[Watch the full story here]

Key products

Foundry Models
Azure AI Services
Azure Machine Learning

70,000+ organizations using Azure AI Foundry today

Design with the best models for your use case

AI templates in

Azure AI Foundry

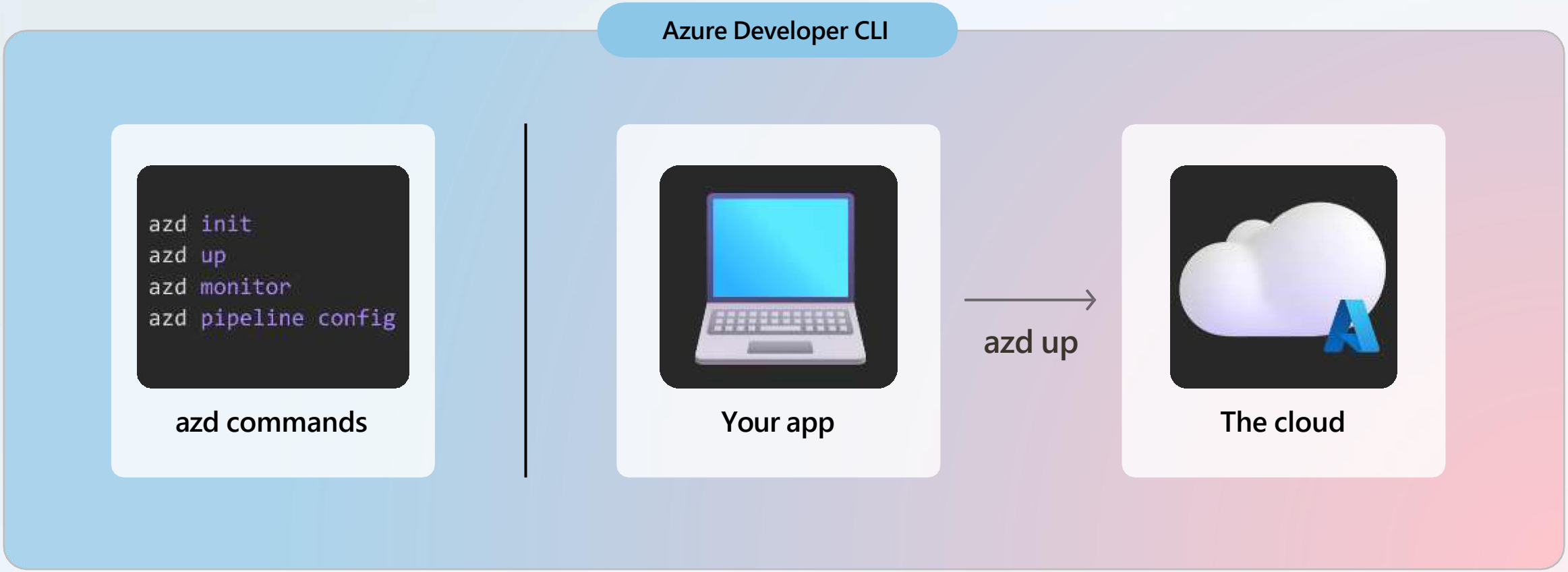
- Unlock insights from conversational data
- Get started with AI chat
- Get started with AI agents
- Create a conversational agent
- Generate documents from your data
- Improve client meetings with agents
- Process multi-modal content
- Modernize your code with agents
- Automate workflows with multi-agents
- Deploy your AI application in production

AI templates in Azure AI Foundry

Get started quickly with customizable AI templates to support a wide variety of use cases.

Template	Popular Customer Use Cases
Unlock insights from conversational data	• Mining call center conversation insights • Call center operations improvement • Transcription
Get started with AI chat	• Employee chatbot for HR / benefits • Professional Services assistance (Legal/Tax/Audit) • Analytics and reporting • Contact center agent assistance
Get started with AI agents	
Build your conversational agents	• Customer service • Conversation summarization • Translation
Generate documents from your data	• Client meeting prep • Document generation • AI-enabled product experiences • Product discovery/shopping assistant
Improve client meetings with agents	
Multi-modal content processing	• Clinician-patient visit documentation • Claims/Contract/Invoice processing • ID verification
Modernize your code with agents	• Code conversion • Code validation • Documentation
Multi-agent workflow automation	• Employee On-Boarding and self-service • Travel Booking and Expense Management • Supply Chain Planning
Deploy your AI application in production	Taking proof of concept applications to production

Local dev environment to the cloud in a single step



Difficulty selecting models and services

New models launching and evolving

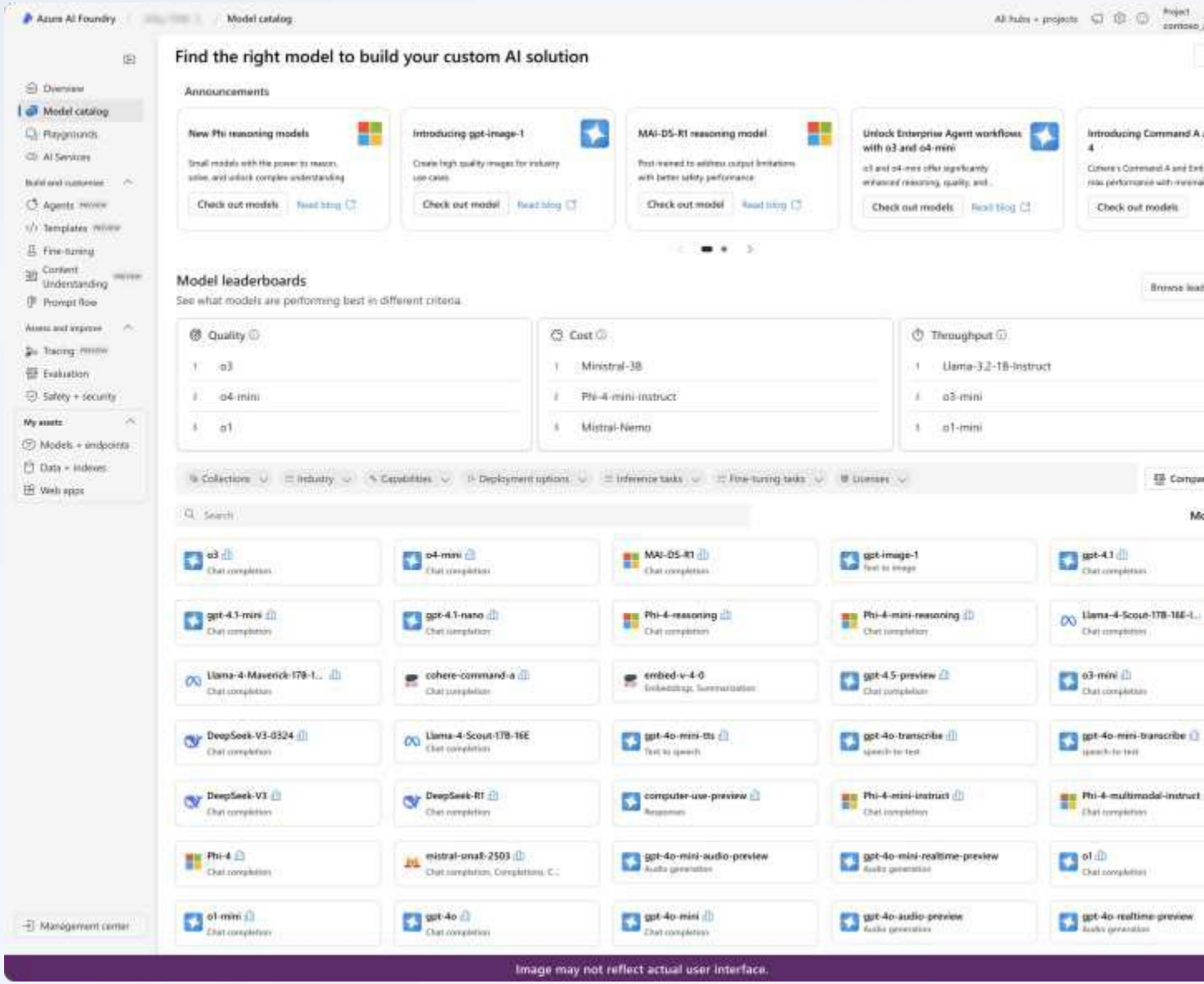
Need to run AI solutions everywhere from cloud to edge

AI innovation requires model flexibility

Azure AI Foundry Models

Design with the best models for your use case

- Discover models for your use case
- Simplify model selection and upgrades
- Run anywhere - cloud, local devices and on-premises



Azure AI Foundry Models

Offering 11,000+ frontier and open models

Built-in safety

Deployment Types:
Standard
Provisioned Throughput
Batch (50%off)

Deployment Locations:
Global
Data Zone (US and EU)
28 Regions

Sold by:
Microsoft

Hosted and sold by Microsoft

- OpenAI Model Family *(available day 1)*
- DeepSeek latest
- Mistral AI latest *(coming soon)*
- Meta Llama latest *(coming soon)*
- xAI Grok *(coming soon)*
- Black Forest Labs *(coming soon)*
- Microsoft Phi Family

Other Foundry Models

- Hugging Face
- NVIDIA NIMS
- Cohere

Specialized & Industry models

Deployment Type:
Standard
Managed Compute

Deployment Locations:
Global
7 Regions

Sold by:
Partner

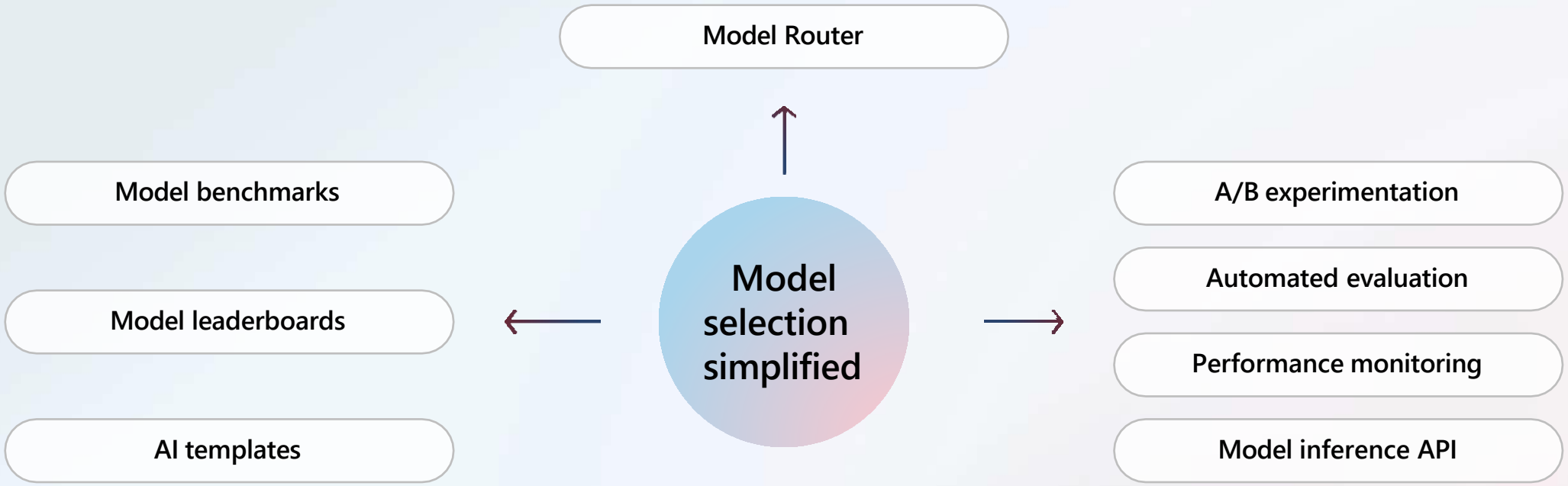
Unified access

Scalable deployment

Enterprise-ready

Model selection simplified with Azure AI Foundry

Use the best model for your use case

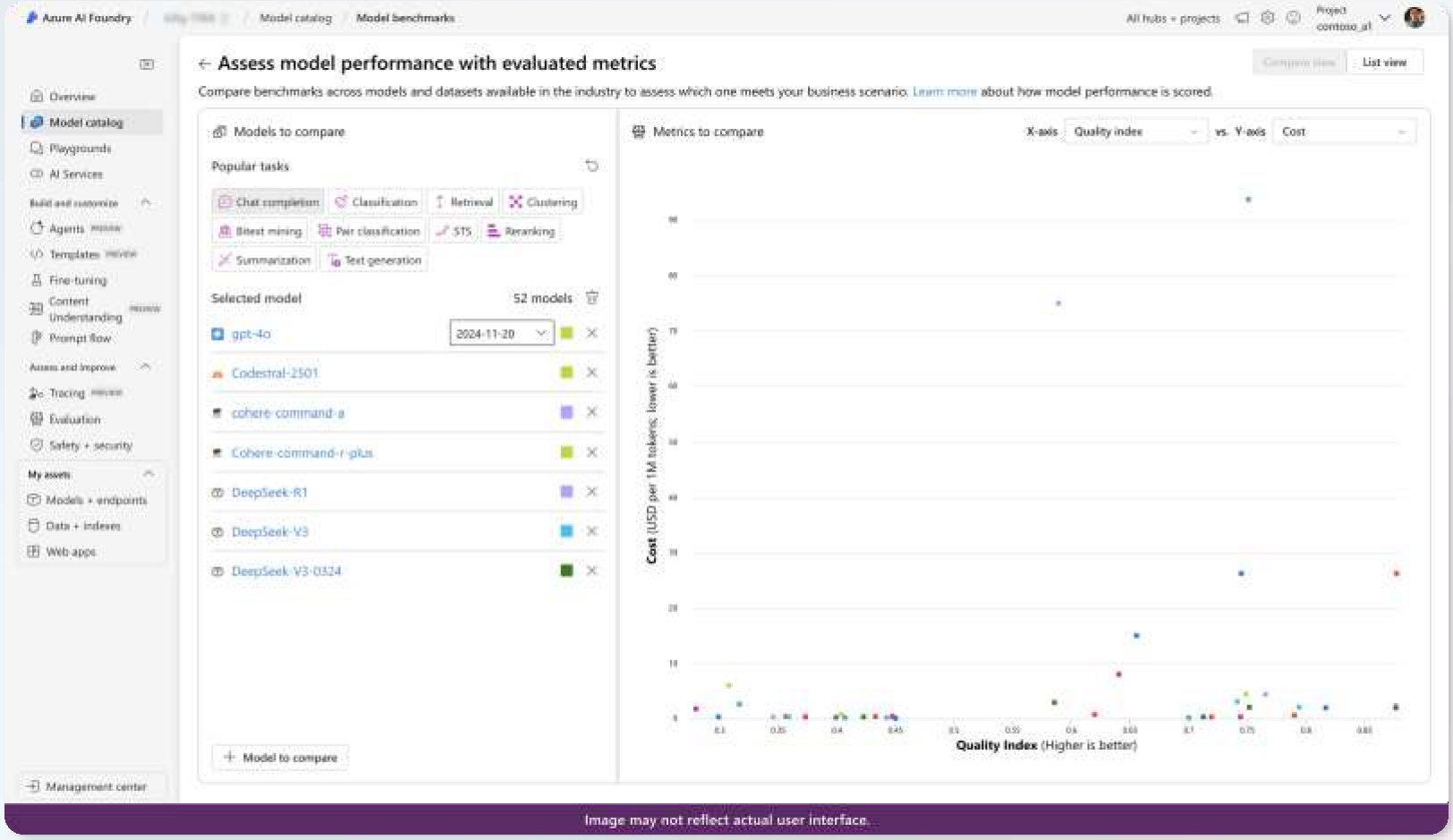


Select the best model

Continue using the best model

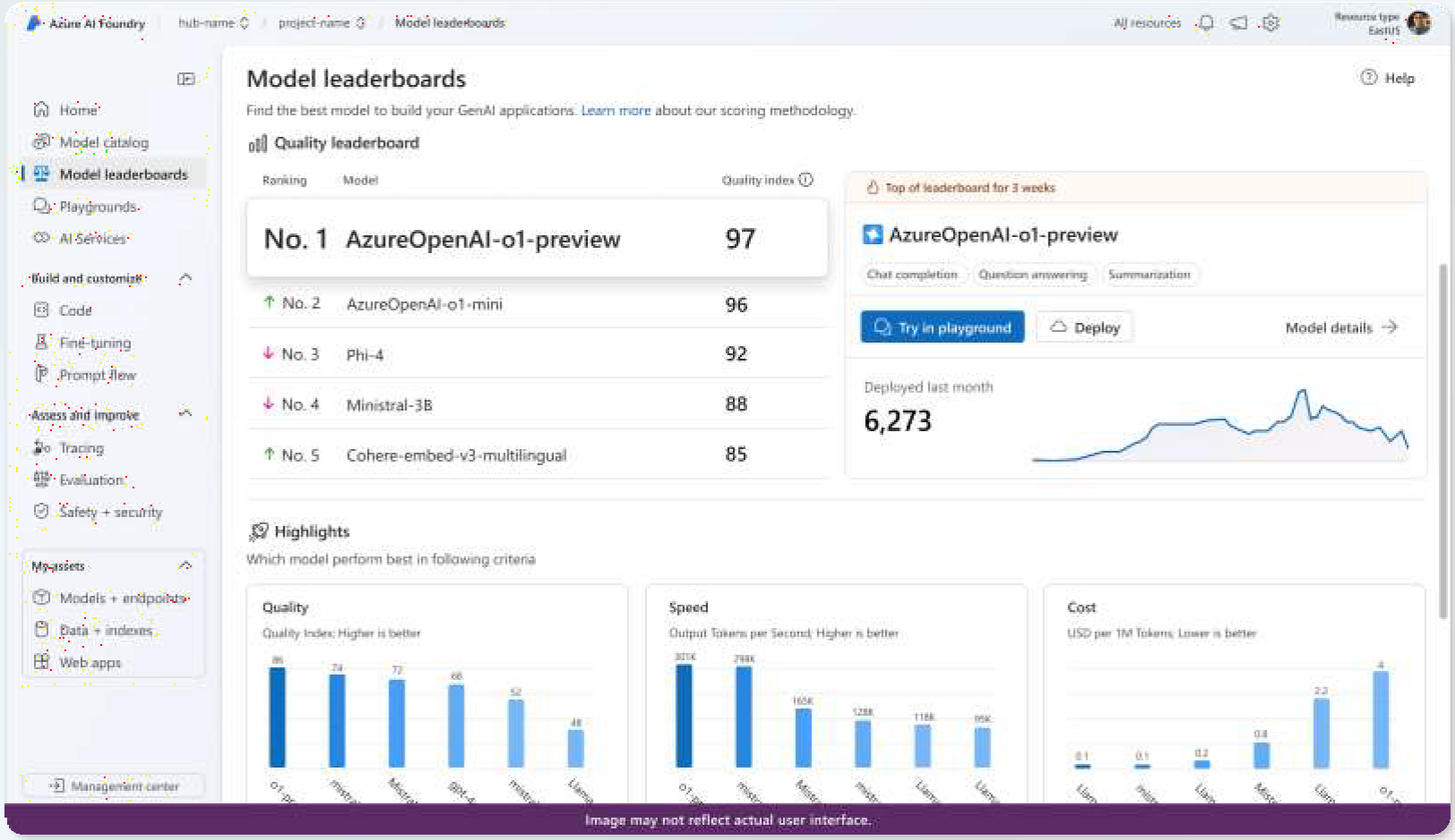
Benchmarking in Foundry Models

Simplify model selection. Review and compare the performance of AI models.



Leaderboard in Foundry Models

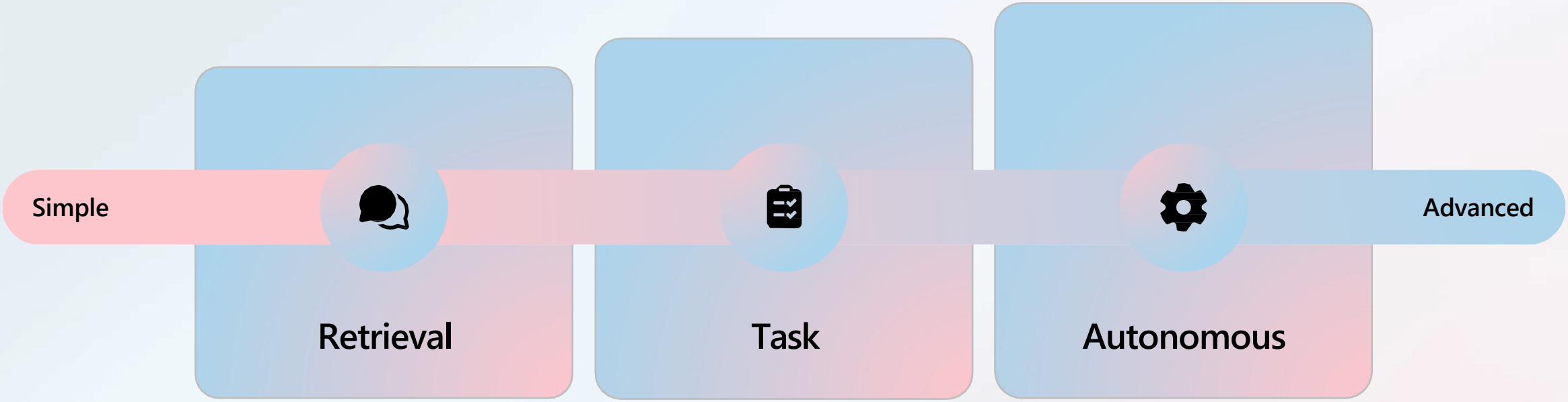
Identify top-performing AI models across different categories and tasks.



AI agents

What are agents?

Agents are programs that use AI to automate and execute business processes, working alongside or on behalf of a person, team or organization



Agents vary complexity and capabilities depending on your need

Freedom to create agents your way

Platform Integrations, ease-of-use, and development speed

IaaS
Infrastructure as a service

Code handling every detail

Azure AI Infrastructure
Bring your own frameworks

PaaS
Platform-as-a-service

Code with managed services

Azure AI Foundry
Agent Service

SaaS
Software as a service

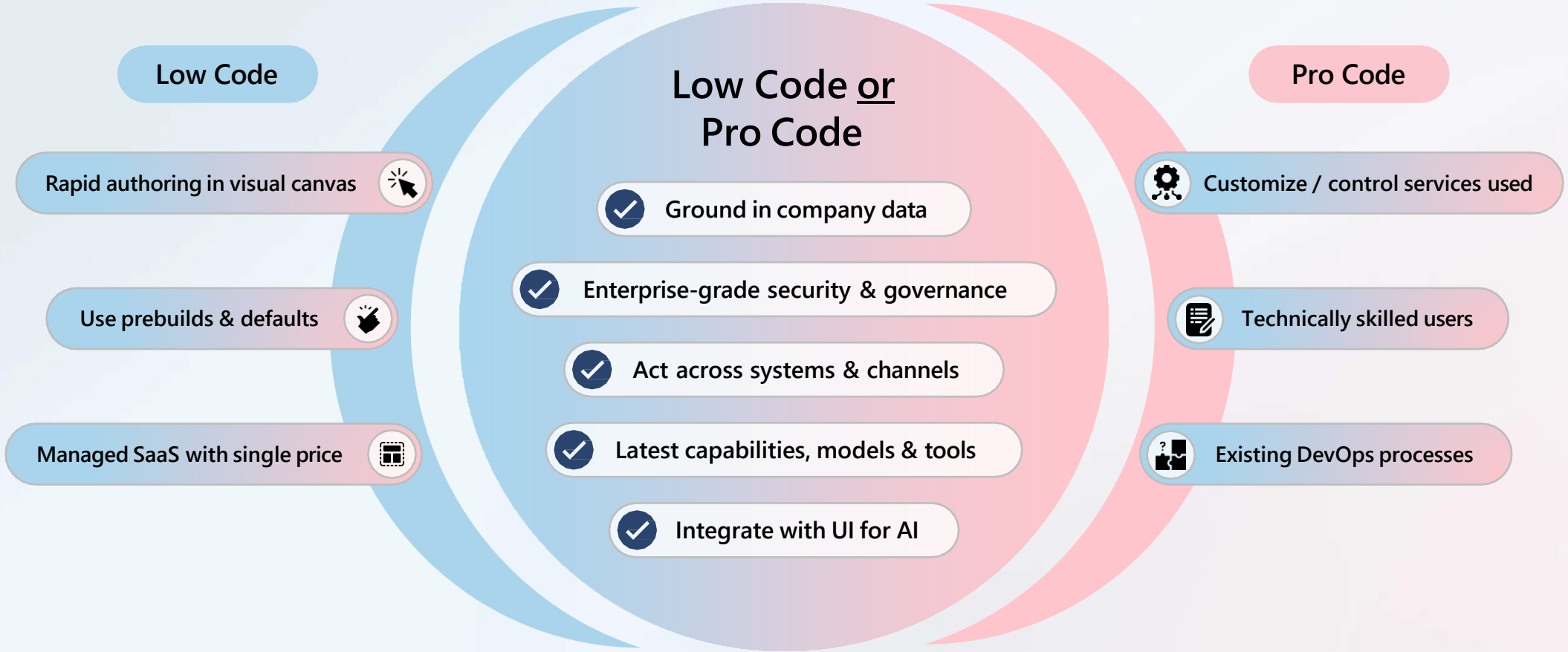
Drag and drop UI building

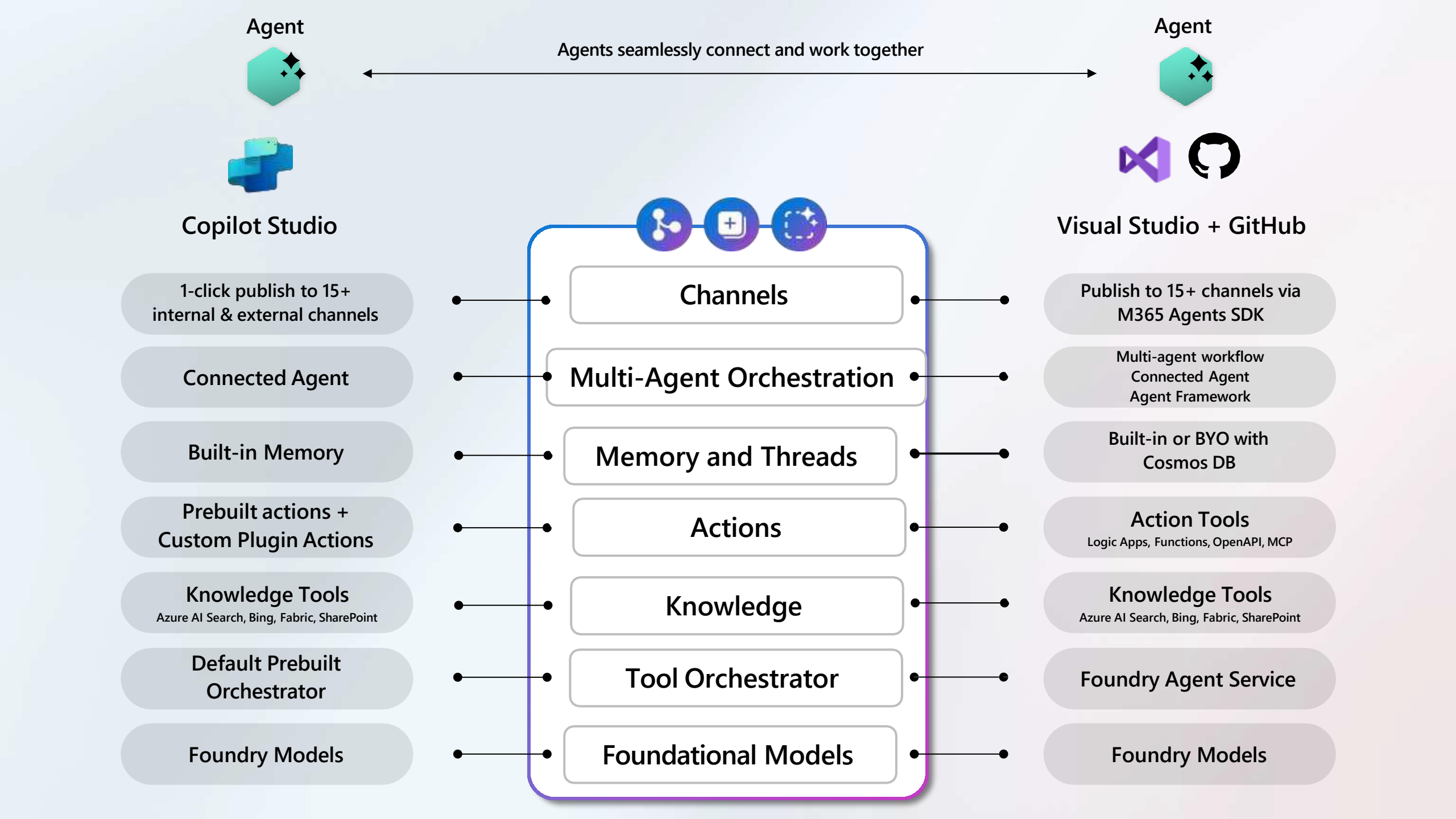
Copilot Studio
Instant agent runtime

Control, visibility, and customization

Choose the tool that best fits your needs

Most agents can be built using either of these differentiated tools – powered by Azure AI Foundry




Agent
Agents seamlessly connect and work together
Agent
Copilot Studio
Visual Studio + GitHub
1-click publish to 15+ internal & external channels
Channels
Publish to 15+ channels via M365 Agents SDK
Connected Agent
Multi-Agent Orchestration
Multi-agent workflow
Connected Agent
Agent Framework
Built-in Memory
Memory and Threads
Built-in or BYO with Cosmos DB
Prebuilt actions + Custom Plugin Actions
Actions
Action Tools
Logic Apps, Functions, OpenAPI, MCP
Knowledge Tools
Azure AI Search, Bing, Fabric, SharePoint
Knowledge
Knowledge Tools
Azure AI Search, Bing, Fabric, SharePoint
Default Prebuilt Orchestrator
Tool Orchestrator
Foundry Agent Service
Foundry Models
Foundational Models
Foundry Models

Securely developing AI agents requires 4 primary considerations

Knowledge

providing agents with the right context

Actions

Giving them the tools to complete their tasks

Security

Provide only access to what they should

Evaluation & Control

Check they complete their tasks correctly

Azure AI Foundry Agent Service

Securely customize, orchestrate, and deploy AI agents

Models

Model choice and flexibility with Foundry Models

Azure OpenAI

o1, o3-mini, GPT-5, 4.1, 4o, etc

Azure Direct Models

- Llama 3.1-405B-Instruct
- Mistral Large
- Cohere-Command-R-Plus

Tools

Richest set of enterprise connectivity

Knowledge

Actions

Logic Apps*Azure functions OpenAPI

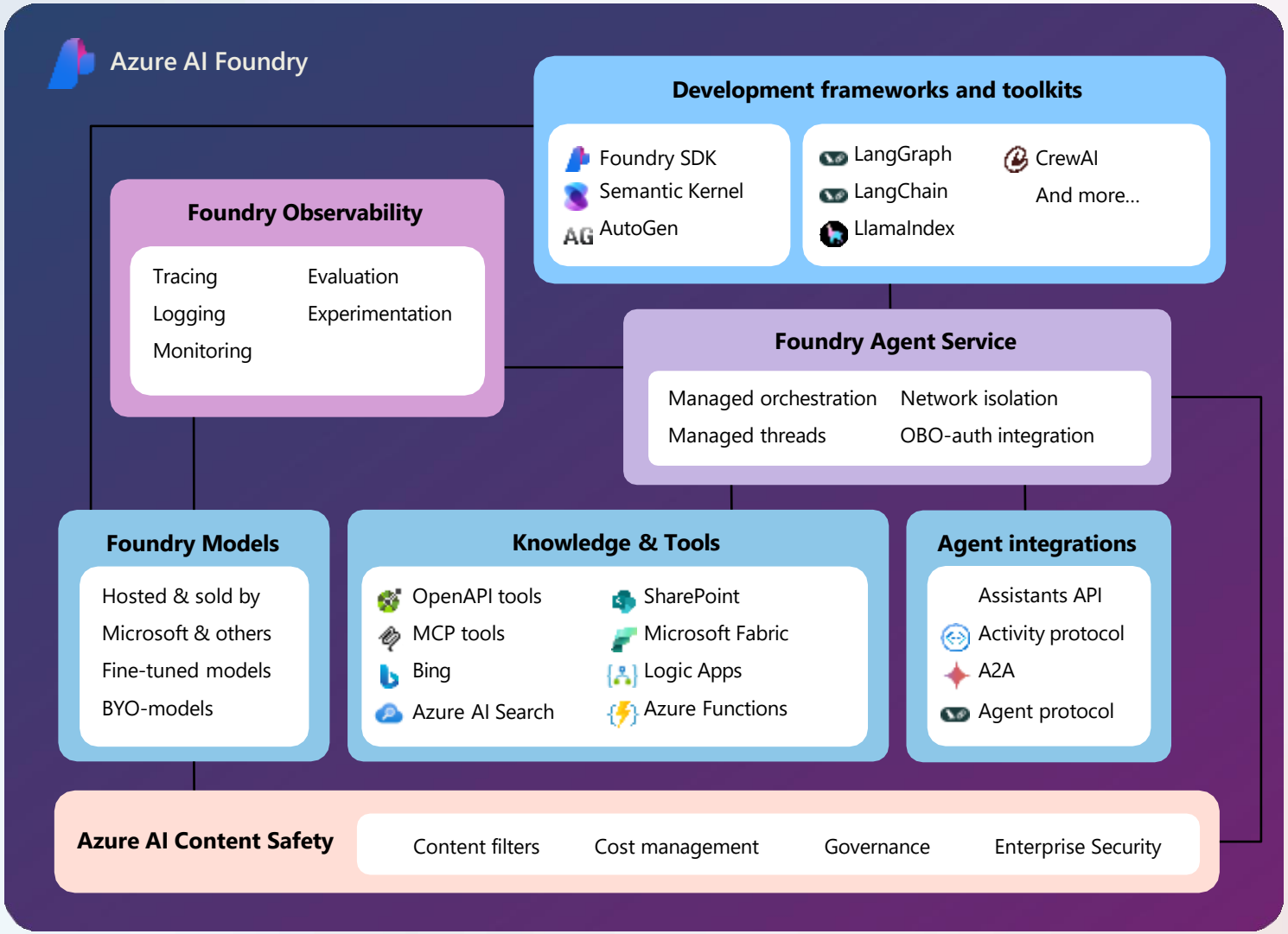
Trust

Customer control over data, networking, and security

- BYO-file storage
- BYO-search index
- BYO-virtual network
- BYO-thread storage
- OBO authentication
- Content filtering

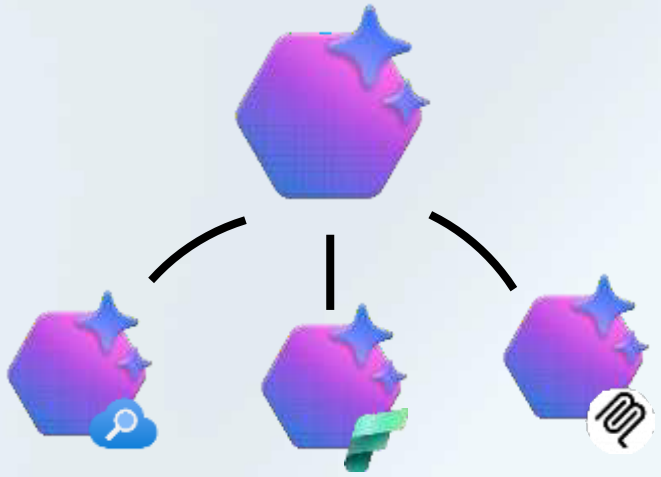
Agents in Azure AI Foundry

Combine the best models, services, and tools in Azure AI Foundry into reusable, testable agentic components.



Multi-agent orchestration in Foundry Agent Service

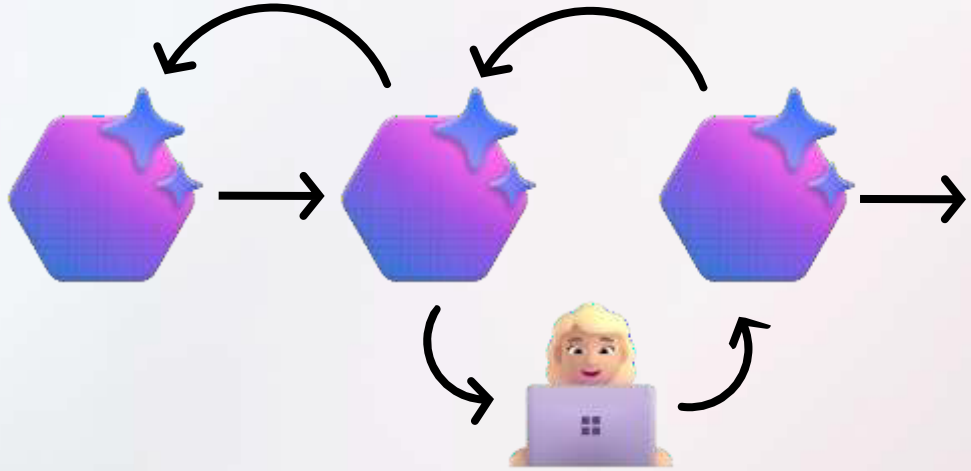
Connected agents



Give one agent the abilities of another.

Multi-agent workflows

Powered by Semantic Kernel

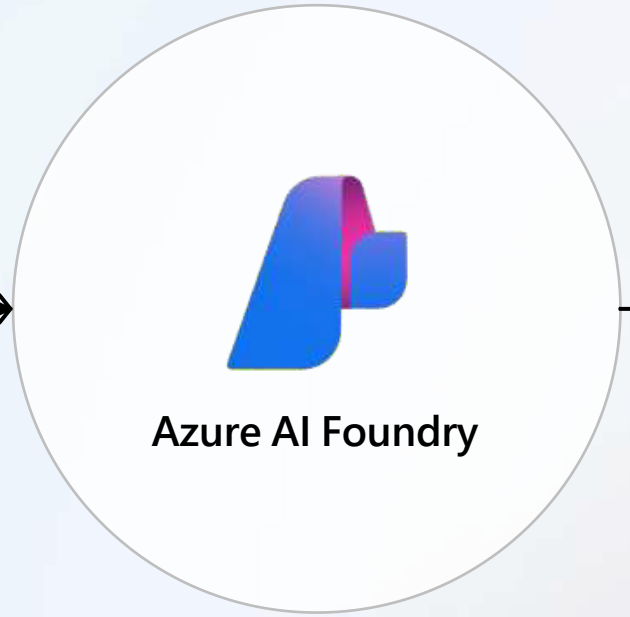


Declaratively orchestrate multiple agents.

Foundry Agent Service will work with the top agent protocols

Import external agents

- Assistants API
- Responses API*
- Activity protocol*
- A2A*
- Agent protocol✧
- MCP*



Expose your agent externally

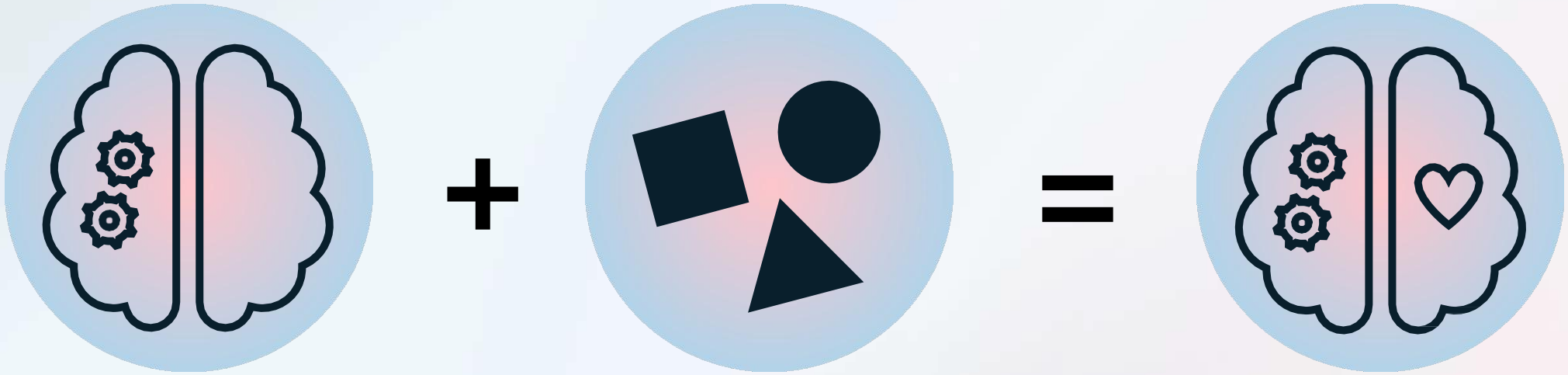
- Assistants API
- Responses API✧
- Activity protocol✧
- A2A
- Agent protocol✧
- MCP*

* Available with Semantic Kernel ✧ Coming soon

Fine-tuning

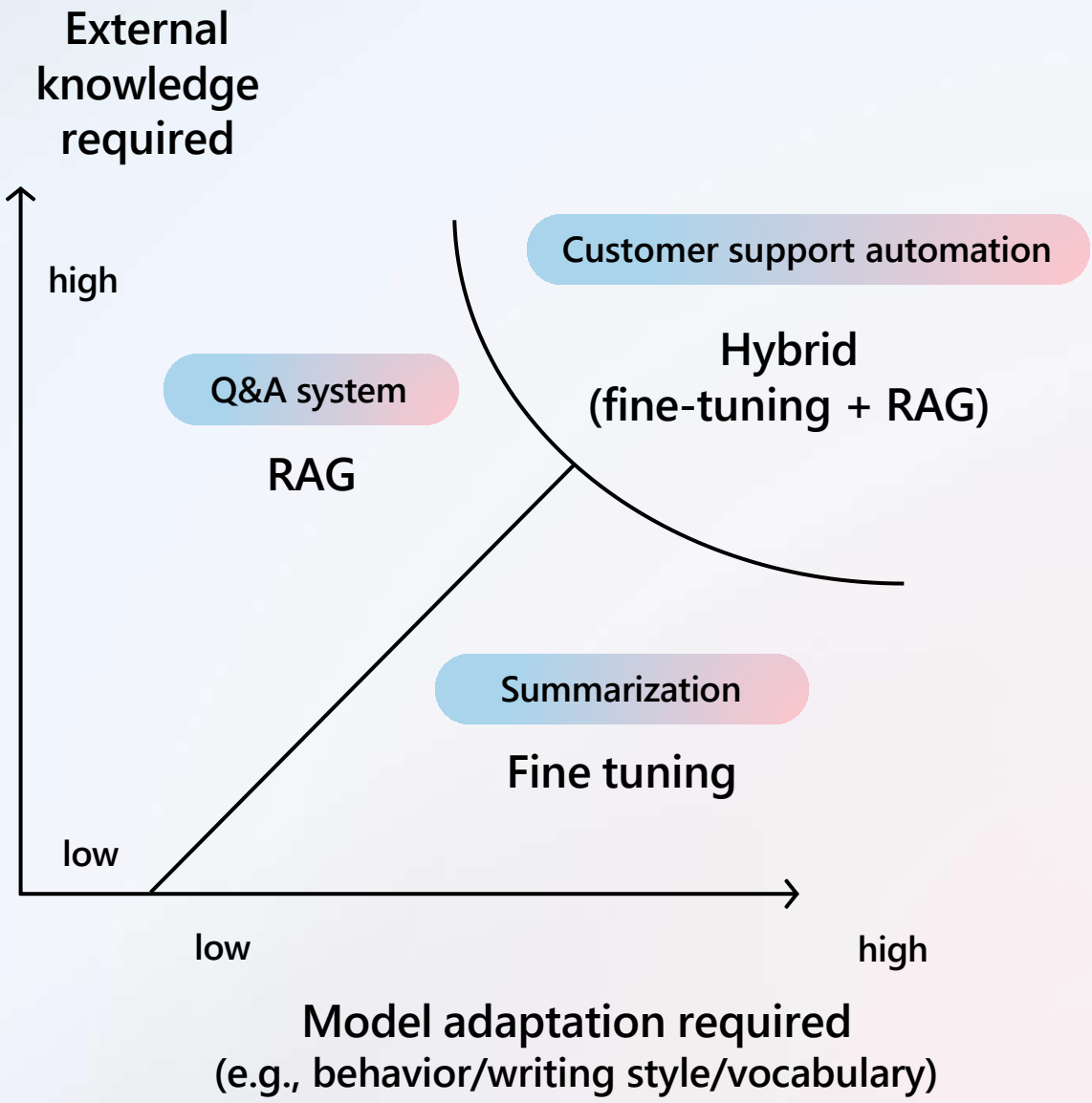
What is fine-tuning?

Customizing a pre-trained foundation model with additional training on curated data of a specific task or domain for enhanced performance, new skills, or improved accuracy.



Generic LLM/SLM + Curated Data Set = Fine tuned LLM/SLM

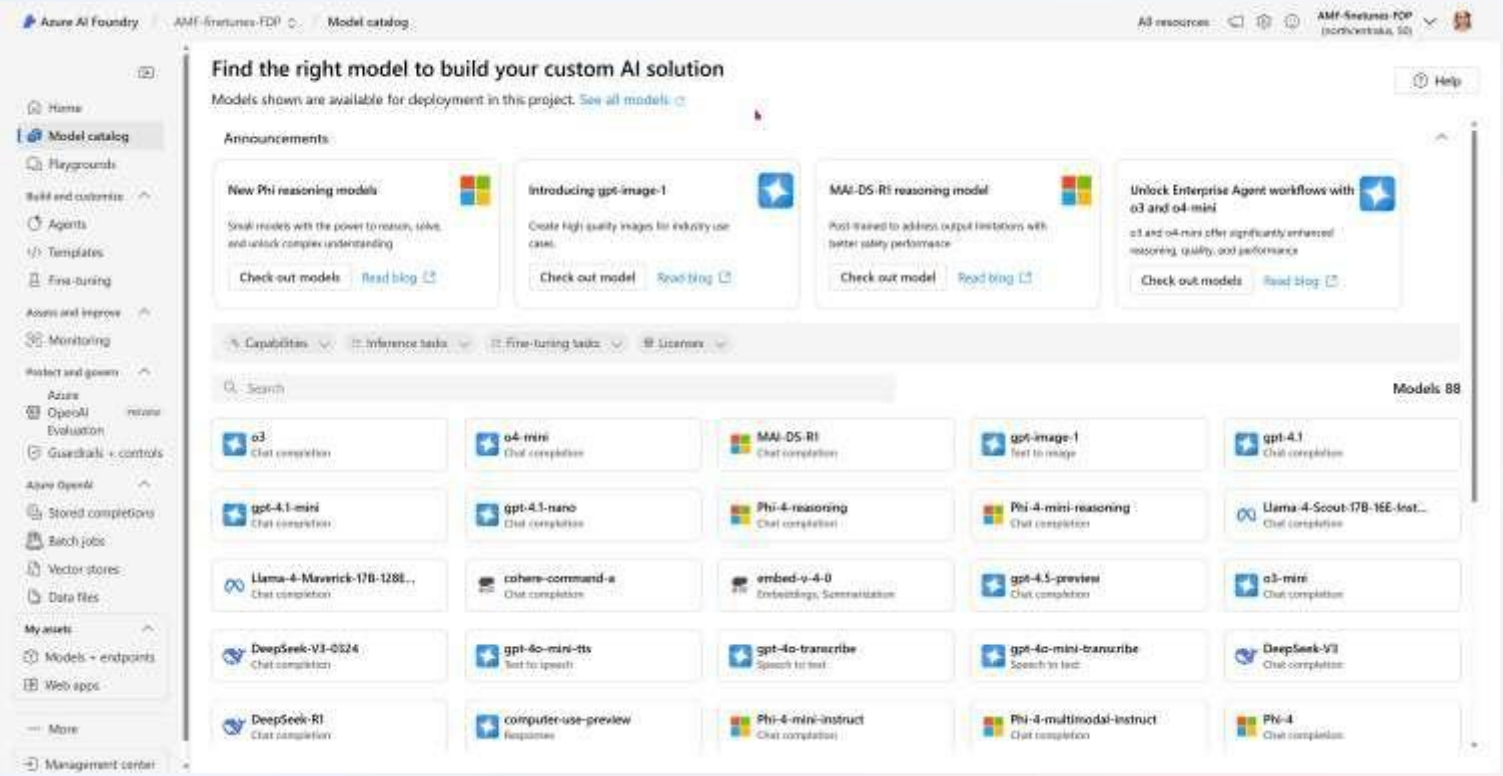
When to use RAG vs. fine-tuning?



Fine-tuning in Azure AI Foundry

Adapt Foundry Models for a specific task or on specific datasets to support business needs.

- Enhance performance and accuracy for specific use cases
- Reduce consumption costs and improve AI efficiency
- Lower latency
- Reduce bias and avoid hallucinations



Manage enterprise-ready AI with confidence

Understand performance trade-offs for agentic applications

New attacks emerging and bad actors also using generative AI

Privacy, security, and compliance complexity

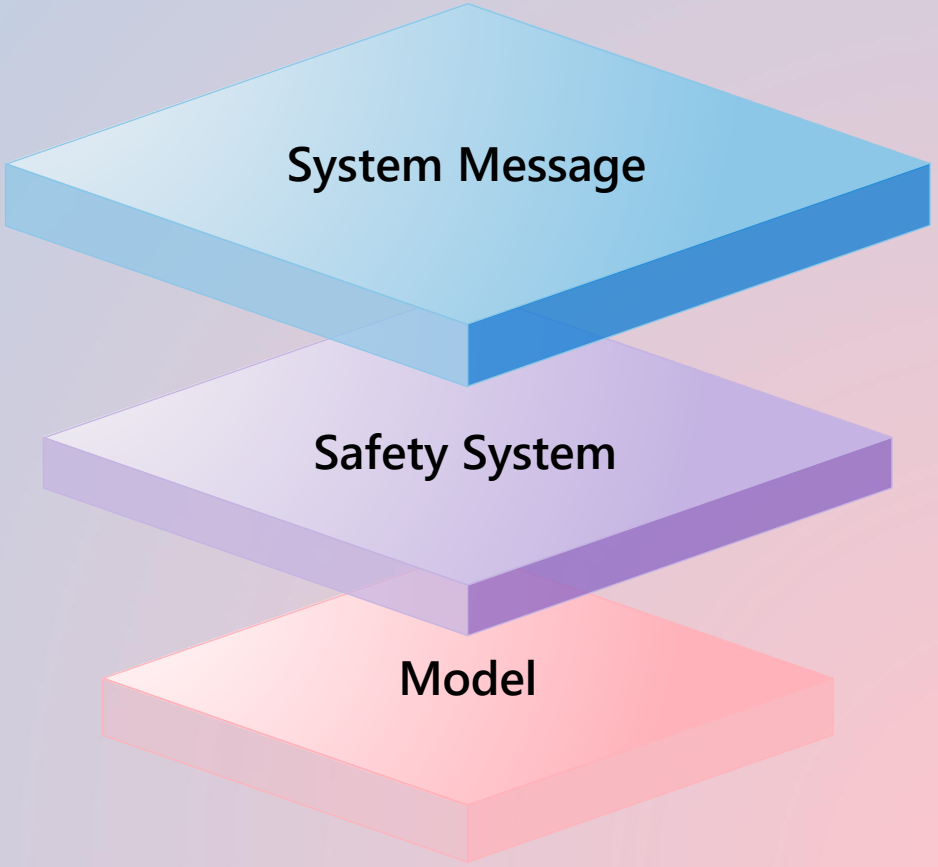
Generative AI introduces new risk and needs to be managed with care

Mitigation layers in Azure AI Foundry

The **system message layer** provides hidden instructions to your model with every user prompt, so you can guide the model's behavior and data retrieval to generate higher quality responses by default.

The **safety system layer** acts as a shield or firewall around your foundation model, by detecting and filtering/blocking risky prompts before they reach your model and risky responses before they reach your end user, for high quality experiences that stay on-brand.

The **model layer** is your foundational model, including any fine-tuning performed by you or the model developer to improve performance and safety.

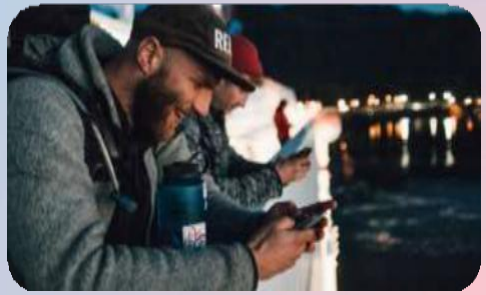


Azure AI Content Safety

Keep your AI content safe with tailored protection, real-time threat detection, and customizable controls.

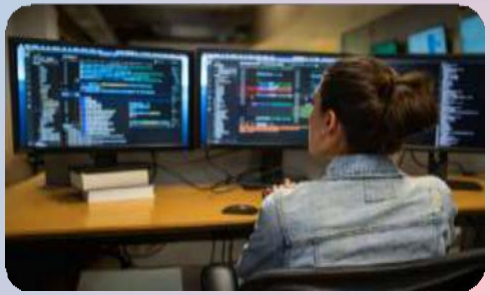
Multimodal filtering

Scan text, images and multi-media to identify, block and monitor harmful content



Customized systems

Create blocklists and custom categories to block entire topics, not just specific words



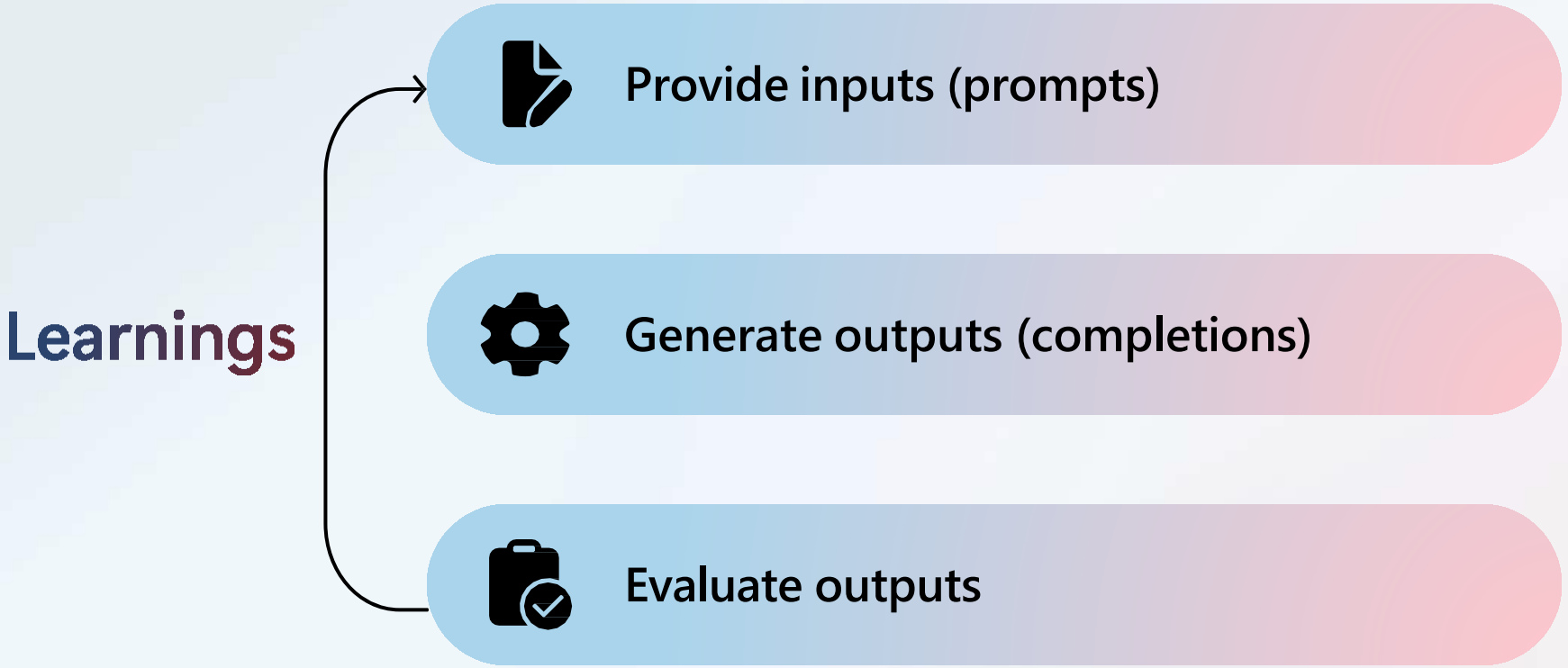
Prompt shielding

Identify and mitigate prompts that could expose you to prompt injection attacks

Protected materials

Avoid outputting known or owned text content with protected materials detection

Evaluations in a nutshell



Evaluations

Quality

Groundedness

Coherence

Fluency

Relevance

Retrieval Score

Similarity

NLP Metrics (e.g., F1 Score)

Risk & Safety

Jailbreak Defect

Hate and Unfairness

Sexual

Violence

Self-Harm

Protected Material

Agents

Intent Resolution

Tool Call Accuracy

Task Adherence

Response Completeness

Risk & Safety

Ungrounded Attributes

Code Vulnerability

Evaluation best practices

Multiply your datasets and metrics

Evaluate systematically using GenAIOps

Review results as a team, with multiple points of view

Getting started

Accelerate AI deployment with Azure AI Foundry

ai.azure.com

Define scope & requirements

Prepare data

Build, train, and evaluate

Review and approve app for production

Deploy and monitor

Govern and monitor continuously

TAHUBU

Thank You!

Lino Tadros

Co-founder & Chief Executive Officer

lino@tahubu.com

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