SME&C
Small Medium Enterprises and Channel

New Innovations in Azure AI Foundry

Guy Gregory
Partner Solution Architect – Microsoft

Microsoft

Microsoft confidential

Azure AI Foundry

Introduction – Guy Gregory

Partner Solution Architect,
UK&I Partner Team (GCPS)

Azure AI Apps and Agents

Focus on Channel Partner

20+ years of Microsoft
partner experience

Guy Gregory
Azure AI Partner Solution Architect

Agenda

01
Introduction

02
Azure AI Foundry overview

03
What's new: Foundry Models

04
Demo – Model router

05
Image, video, and audio models

06
Demo – gpt-audio

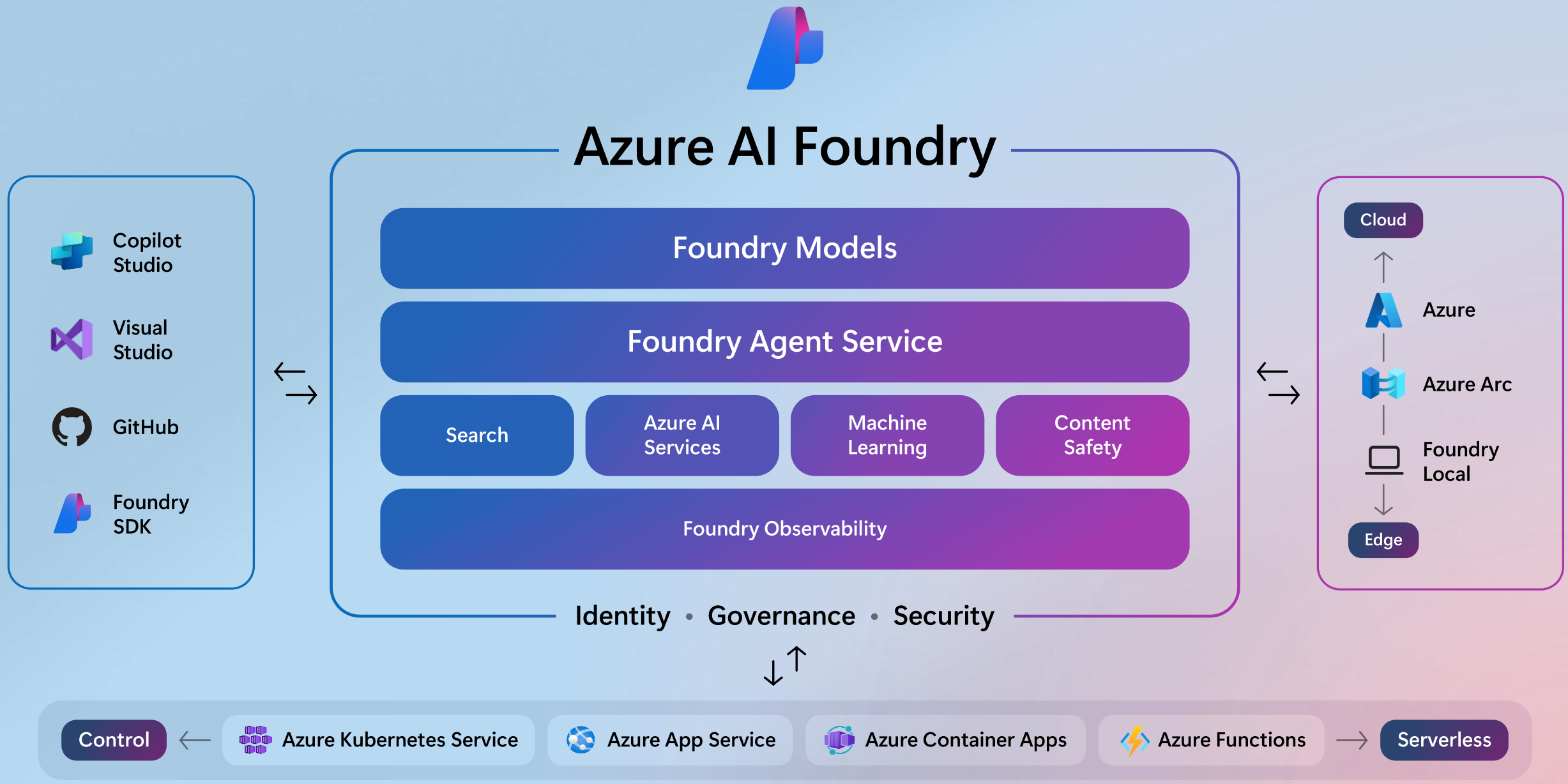
07
What's new – Agents and APIs

08
Demo – Browser Automation

09
Demo – Web Search Preview

10
Further reading

11
Q&A


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

Microsoft

What’s new in Azure AI Foundry Models?

Foundry Models

OpenAI

Grok

Mistral AI

Meta

black forest labs.

deepseek

https://aka.ms/foundryprojects

Generally Available

GPT-5

Unified reasoning + conversation in one family

Frontier-level code generation & explanation

Developer controls reasoning effort & verbosity

Free-form function calling

Generally Available

GPT-5 APIs

GPT-5 — deeper, faster reasoning for analytics & decisions

GPT-5-chat — natural, multimodal, multi-turn, context-aware

GPT-5-mini — low-latency & cost for real-time copilots/summarization

GPT-5-nano — ultra-low latency for edge, mobile, embedded, only via API

Multimodal reasoning | Agentic workflows | Available via Azure AI Model Router | Enhanced coding and logical reasoning

Generally Available

GPT-5-Codex

- Multimodal coding
- Interactive workflows in a model-guided loop
- Code review and defect detection
- Repository-aware automation
- Seamless experience across developers surfaces

Smarter Development | Faster Workflows | Accelerated Refactors and Migrations | Safer Engineering

Public Preview

Model Router in Azure AI Foundry Models

Choose the right LLM for your diverse use cases

Optimized Routing | High Performance | Optimized Cost

In our tests comparing use of Model Router versus use of GPT-4.1 only, we saw up-to 60% cost savings with similar accuracy

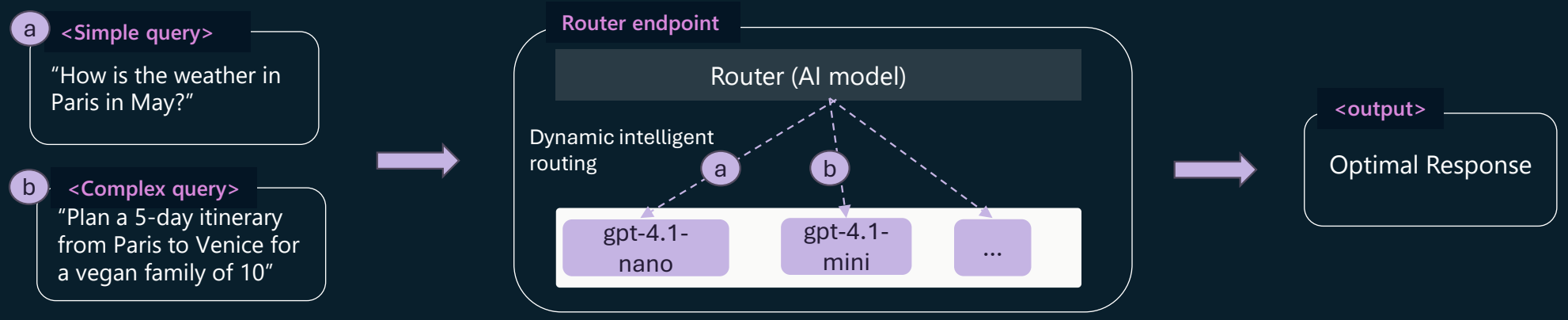
What is a model router?

Challenge

Choosing the right LLM for a task is difficult. Users often stick to a single LLM for a given use case.

Opportunity

Smart routing **optimizes cost** while **maintaining quality**. Dynamically selecting the best model reduces expenses without degrading user experience.



Step 1
Users invoke the endpoint with Chat Completions or Responses API

Step 2
Router assesses the input parameters (like prompt, tool use) and dynamically routes to the optimal model

Step 3
The endpoint returns underlying model's response.

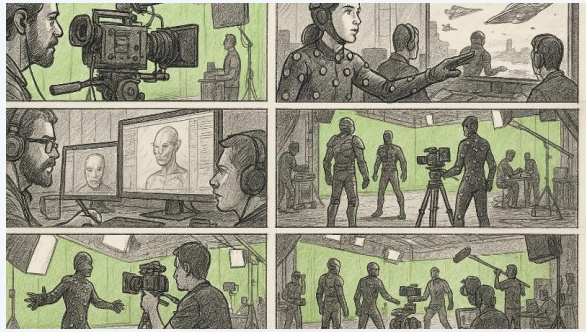
Demo – Model router

Model Pricing

GPT-5	• $1.25 / 1,000,000 input tokens • $10.00 / 1,000,000 output tokens
GPT-5-mini	• $0.25 / 1,000,000 input tokens • $2.00 / 1,000,000 output tokens
GPT-5-nano	• $0.05 / 1,000,000 input tokens • $0.40 / 1,000,000 output tokens
GPT-5-chat	• $1.25 / 1,000,000 input tokens • $10.00 / 1,000,000 output tokens
Model Router	• $0.14 / 1,000,000 tokens

Public Preview

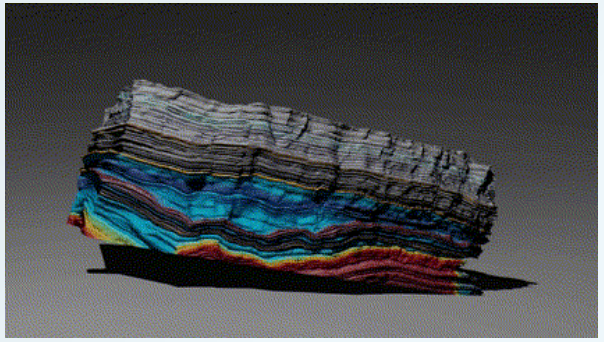
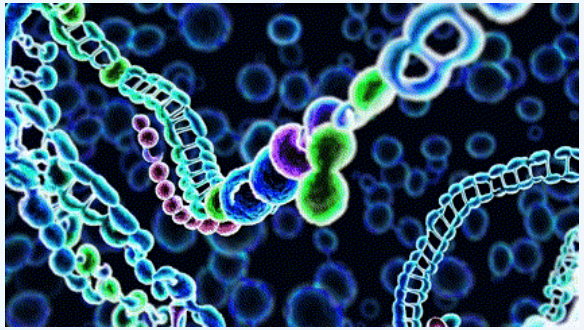
GPT-image-1



Granular Instruction Response | Reliable Text Rendering | Image Input Acceptance & Editing | Professional Quality: 1024px+ resolution

Public Preview

Sora

API Integration | Reliable Text Rendering | Image Input Acceptance (Coming Soon) | High Fidelity 1080p+

Generally Available

gpt-realtime

New, natural voices – two near clear and emotionally expressive voice options (Marin and Cedar).

Instruction Following – Enhanced adherence to tone, pacing, escalation triggers, accents, emotional shifts, and role-based behavior.

Function Calling – Reliable and accurate tool invocation with improved timing and argument accuracy. Supports patterns like answer(question), escalate_to_human(), and finish_session().

Audio Quality – improved alphanumeric reproduction, and strict modality control.

New, natural expressive voices

20% cheaper than gpt-4o-realtime-preview

Multimodal inputs

Conversational behavior

New Release

Generally Available

gpt-audio

Sign up to access in Azure AI Foundry or talk to your representative

https://aka.ms/gpt-audio

Demo – gpt-audio

Microsoft

What’s new in Agents and APIs?

New | Preview

Introducing Voice Live API

Redefining agent communication with advanced, real-time voice

Use the foundation model of your choice

Unparalleled speech recognition and synthesis capabilities

Fully customizable voices and avatars

aka.ms/voice-live

New | Preview

Browser Automation Tool

Build agents capable of performing real-world browser tasks

Automate complex web UIs, including form fills and multi-step processes

Unparalleled speech recognition and synthesis capabilities

Powered by Azure Playwright Workspaces

aka.ms/browserautomation

Demo – Browser Automation

Computer Use tool vs. Browser Automation

Capabilities	Browser Automation	Computer Use
Model support	Most models supported by Foundry Agent Service	Computer-use-preview model
Visualize what's happening	❌	✅
How it understands the screen	Parse the HTML or XML pages into DOM documents	Raw pixel data from screenshots
How it acts	A list of actions provided by the model	Virtual keyword and mouse
Multi-step?	✅	✅
Interfaces	Browser	Computer and browser
Can I bring my own resource	BYO Playwright resource and store keys in connection	No resource required, we recommend using **low-privilege virtual machines**

Generally Available

Responses API

- Combines simplicity of Chat Completions API with tool use capabilities of Assistants API
- Seamless interaction with tools like CUA, code interpreter, function calling, and file search – all in a single API call
- Allows AI systems to retrieve data, process information, and take actions
- Enables AI systems to easily integrate with enterprise workflows

Simplifies AI Agent Development

Orchestrates Multiple AI Tools

Enhances Scalability for Enterprise Use Cases

Offers Flexibility with Azure AI Agent Service

Azure OpenAI API comparison (24 Sep ‘25)

Capabilities	Chat Completions API (Azure OpenAI)	Assistants API (Azure OpenAI)	Azure AI Foundry Agent Service	Responses API (Azure OpenAI)
Text generation	✅	✅	✅	✅
Audio	✅	❌	✅	Not yet available
Vision	✅	✅	✅	✅
Structured Outputs	✅	✅	✅	✅
Function calling	✅	✅	✅	✅
Web search	❌	❌	✅	Not yet available
File search	❌	✅	✅	✅
Computer use (CUA)	❌	❌	✅	✅
Browser use	❌	❌	✅	via CUA or MCP
MCP tool	❌	❌	✅	✅
Code interpreter	❌	✅	✅	✅
Reasoning models	✅	✅	✅	✅
Entra ID support	✅	✅	✅	✅

NB Some features/tools may be in preview

Demo – Web Search Preview

Semantic Kernel

AI and Agent Orchestration

AI Services

- Azure OpenAI
- Foundry Catalog
- OpenAI
- Hugging Face
- NVIDIA
- DeepSeek
- Google Gemini
- AWS Bedrock

Local models

- Ollama
- ONNX
- LM Studio
- Phi SLM Model Family

Memory Services

- Azure AI Search
- Azure CosmosDB
- Elasticsearch
- MongoDB
- Pinecone
- Qdrant
- Redis

.NET

Python

Java

Agent Services

- Azure AI Agent Service
- Bedrock Agents
- AutoGen
- Crew AI

Plugins

- OpenAPI
- MCP
- A2A
- Logic Apps

Prompts

- YAML
- Semantic Kernel

Filters and telemetry

- OpenTelemetry
- Azure Monitor
- Aspire Dashboard
- Azure AI Content Safety

Converging the two popular OSS frameworks

Microsoft Agent Framework (coming soon)

Semantic Kernel

Full SDK designed to build AI agents with ease, excellent for single agents and can be extended for multi-agents with integrations to AutoGen

AutoGen

Powerful multi-agent research framework with prebuild conversation orchestration patterns for handling complex agent systems

Microsoft

How do I learn more?

FY26 Skilling Hub

https://www.skilling-hub.com/

CLOUD CHAMPION

Upcoming Webinars

2 OCTOBER 2025

Latest Innovations in Azure AI Foundry

See More

2 OCTOBER 2025

Security GTM Programs and Investments Overview for SMB

See More

6 OCTOBER 2025

Cloud Ascent - Microsoft's data-driven customer targeting programme

See More

https://www.cloudchampion.co.uk/

AI for Partners (AI4P)

Last updated: 11th Sep 2025
https://aka.ms/AI4P

Access leading AI models and Azure credit via your MAICPP membership

No membership, or Network Member

Azure Free

New to Azure? $200 free credit to use on Azure services within 30 days

Spending protection—credit card won't be charged

GitHub Models

Unlock limited model access with GitHub Copilot Free

Learn how to integrate GitHub Models into GitHub Actions for "Continuous AI"

Easy upgrade to Azure AI Foundry Models

https://gh.io/models

Partner Launch Benefits, or Partner Success Benefits

Partner Launch Benefits

Activate $700/year Azure Bulk Credit to help launch your business

Partner Success Core Benefits

Access $2,400/year Azure Bulk Credit to build your practice

Partner Success Expanded Benefits

Unlock $4,000/year Azure Bulk Credit to expand and grow success

https://aka.ms/AI4P/Success

Solutions Partner Designation (Azure)

Solutions Partner Benefits

Add $6,000/year Azure Bulk Credit with the Data & AI Solutions Partner designation

25 x Visual Studio Subscriptions, each with $150 Azure credit

Stack multiple benefits

Add more Azure Bulk Credit with additional Solutions Partner designations

Access to programs and incentives

Access to CSP Incentives and Azure Accelerate (SMB Track) offer

https://aka.ms/AI4P/SolutionsPartner

Specialized Partner, or Azure Expert MSP

Specialization/AEMSP Benefits

Add $12,000/year Azure Bulk Credit per Azure specialization

10 x Visual Studio Subscriptions, each with $150 Azure credit, per Azure specialization

Stack multiple benefits

Max cap amount: 5 specializations or expert programs

Access to programs and incentives

Access to CSP Incentives and additional Azure Accelerate offers

https://aka.ms/AI4P/SpecializedPartner

Download the Microsoft AI Cloud Partner Program Benefit Guide: https://aka.ms/solutionspartner.benefits

35,000+ Members!

Join

Azure AI Foundry Developer Community

Let's Build the Future of AI Together

- Discord
- GitHub Discussions
- Community Events
- Open-Source Courses
- Code Samples

Join the Discord: https://aka.ms/AzureAI/Discord

Microsoft

Questions?

Download these slides and demos:

https://aka.ms/FoundryInnovation

© Copyright Microsoft Corporation. All rights reserved.