WHITEPAPER

Microsoft Foundry:

The Platform Approach to AI

Artificial Intelligence is accelerating faster than most enterprises can control, govern, or operationalise it. Organisations require more than access to powerful models. They require a platform that brings structure, governance, and lifecycle management to AI, in the same way that DevOps brought discipline to application development.

This is where **Microsoft Foundry** becomes essential.

What is Microsoft Foundry

Microsoft Foundry (formerly Azure AI Studio) is Microsoft's unified platform for building, deploying, and governing enterprise-grade AI solutions at scale.

Rather than being "just another AI tool", **Foundry** acts as a control layer that sits between powerful AI models and real-world business use cases. It enables organisations to design, test, deploy, monitor, and govern AI agents in a consistent and secure manner.

A useful way to think about **Foundry** is not as the AI engine itself, but as the factory and operating system for AI delivery. While model providers such as OpenAI supply the engines (LLMs), Foundry provides the assembly line, ensuring that AI solutions are:

- **Built using repeatable patterns**
- **Governed by enterprise security and compliance standards**
- **Observable, auditable, and improvable over time**
- **Aligned to real business outcomes rather than isolated experiments**

Foundry brings together model selection, prompt orchestration, agent development, evaluation, deployment, and governance into a single, cohesive platform, reducing the fragmentation that commonly undermines enterprise AI initiatives.

Why a platform approach to AI is now necessary

Early AI adoption often begins with direct API calls to a single model. While this approach can deliver quick wins, it does not scale safely or sustainably.

As AI usage grows, organisations quickly encounter fundamental questions:

? How do we change models without rebuilding applications?
? How do we prevent sensitive data from being exposed or misused?
? How do we measure accuracy, reliability, and cost over time?
? How do we ensure AI behaves consistently across teams and use cases?
? How do we prove compliance, accountability, and responsible use?

Without a central platform, these questions are answered in isolation, if at all.

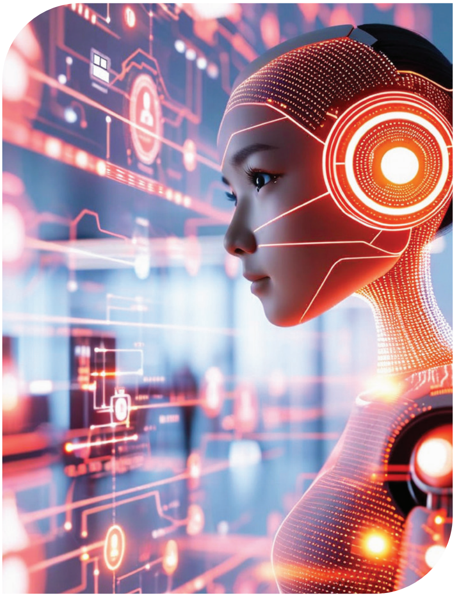
Microsoft Foundry addresses this gap by providing a structured, governed environment in which AI solutions can move from experimentation to production, without sacrificing agility. It enables organisations to embrace AI innovation while maintaining control, trust, and accountability.

From experimentation to Enterprise AI

The true value of **Microsoft Foundry** lies not in individual features, but in its ability to operationalise AI. **Foundry** enables organisations to:

- **Decouple business solutions from specific AI models**
- **Introduce governance and guardrails without slowing innovation**
- **Treat prompts, agents, and evaluations as production assets**
- **Monitor cost, performance, and risk across all AI initiatives**
- **Scale AI adoption confidently across departments and use cases**

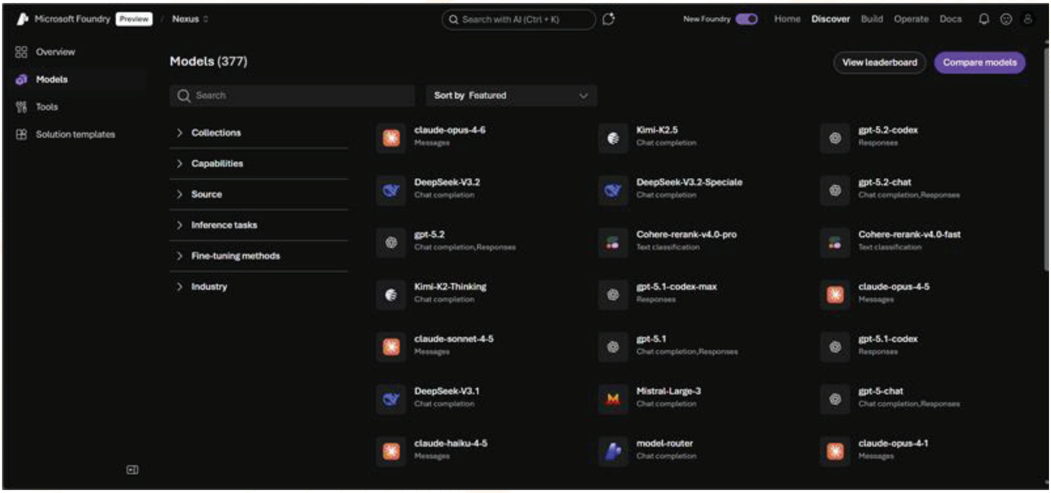
In doing so, **Foundry** shifts AI from an experimental capability to a managed enterprise asset, one that can be trusted, improved, and scaled over time.



Core capabilities

On startup, **Foundry's** UI provides an integrated Model catalogue with access to over 11000 LLM to choose from. From OpenAI GPT to Hangzhou's DeepSeek, AI foundry provides users with the capability to compare models based on:

- Performance
- Cost
- Thoughtput (tokenization)
- Quality
- Safety
- Accuracy
- Hallucination Rate
- Costing
- Response Time
- Context Window
- Scalability
- Data & Security Mindfulness



Models (377)

claude-opus-4-6
Messages

Kimi-K2.5
Chat completion

gpt-5.2-codex
Responses

DeepSeek-V3.2
Chat completion

DeepSeek-V3.2-Speciale
Chat completion

gpt-5.2-chat
Chat completion,Responses

gpt-5.2
Chat completion,Responses

Cohere-rerank-v4.0-pro
Text classification

Cohere-rerank-v4.0-fast
Text classification

Kimi-K2-Thinking
Chat completion

gpt-5.1-codex-max
Responses

claude-opus-4-5
Messages

claude-sonnet-4-5
Messages

gpt-5.1
Chat completion,Responses

gpt-5.1-codex
Responses

DeepSeek-V3.1
Chat completion

Mistral-Large-3
Chat completion

gpt-5-chat
Chat completion,Responses

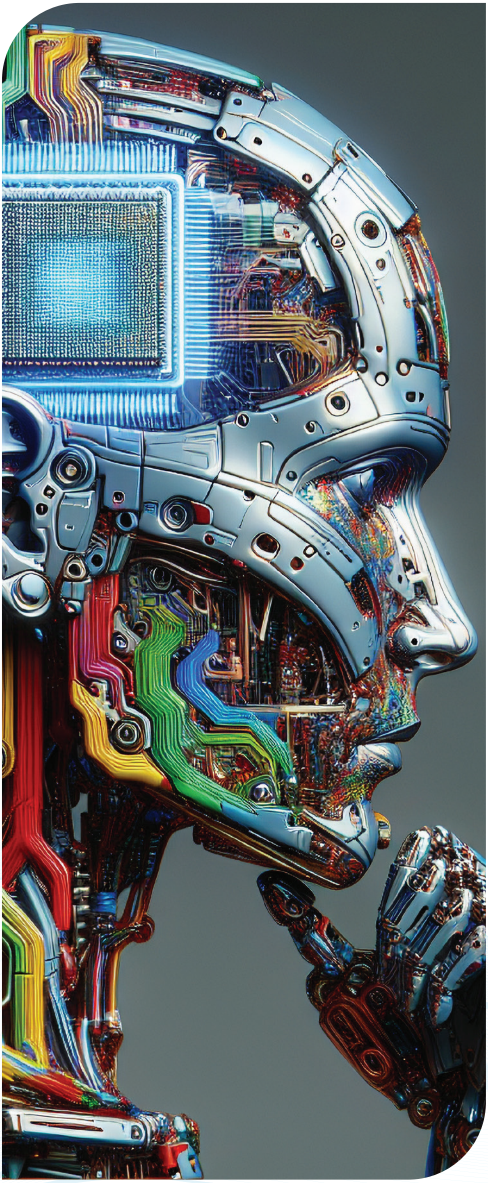
claude-haiku-4-5
Messages

model-router
Chat completion

claude-opus-4-1
Messages

Azure AI agent service:

Azure AI Agent Service is a dedicated platform for developing, managing and deploying Agentic AI solutions. By providing model flexibility, developers can make use of different models (GPT, Claude, Gemini) simultaneously and seamlessly. This goes above and beyond the usual chatbot by empowering agents to think, plan and execute, enabling agents to prompt actions that book a flight, make a reservation, update a database and search for records.



Prompt Flow

Design the logic flow of your agent using Foundry's prompt flow feature, which provides a visual orchestration where engineers can associate data, user prompts, and custom Python code to enhance the user experience of their agents. Think of this as a standard workflow, with the prompt being the trigger, the data being the knowledge base, and the response being the trained output.

Safety & governance

The main driving factor behind every successful AI solution lies in its ability to maintain data security and safety. **Microsoft Foundry's** built-in governance allows for integrating security evaluation tools with platforms like Microsoft Purview and other external tools to define, monitor, and execute AI risks and compliance easily.

Foundry Control Plane:

Foundry Control Plane is an integrated management interface that provides visibility, governance and control for AI Agents, models and tools located in **Foundry.** It aids as a unified area for managing every component of AI solutions from planning to production.

Core Functionalities:

The Control plane combines inventory, observability, compliance and security into a singular user experience. Being Microsoft native, it seamlessly affiliates with Microsoft Security services like Defender, Purview and Entra to provide trust at scale.

Features:

Manage your AI Resources across environments in a single place

- Monitor KPIs like active agents, run completion, compliance metrics, cost effectiveness and unauthorised behaviour.
- Deep links provide evaluation and tracking experiences for swift debugging, diagnosis and improvement.
- Analyse resource health through smart dashboards that visualise live trends and anomalies.

Observe, protect, improve

- Pinpoint alert, evaluation results and traces to identify issues promptly.
- Constantly assess your agent's performance, quality and security.
- Make use of built-in tools such as AI Red Teaming Agent and cluster analysis for an automated risk and error discovery.
- Foundry Agent, with the use of its built-in AI, can advise on improvements from instruction writing to source control.

Govern and administer with Guardrails

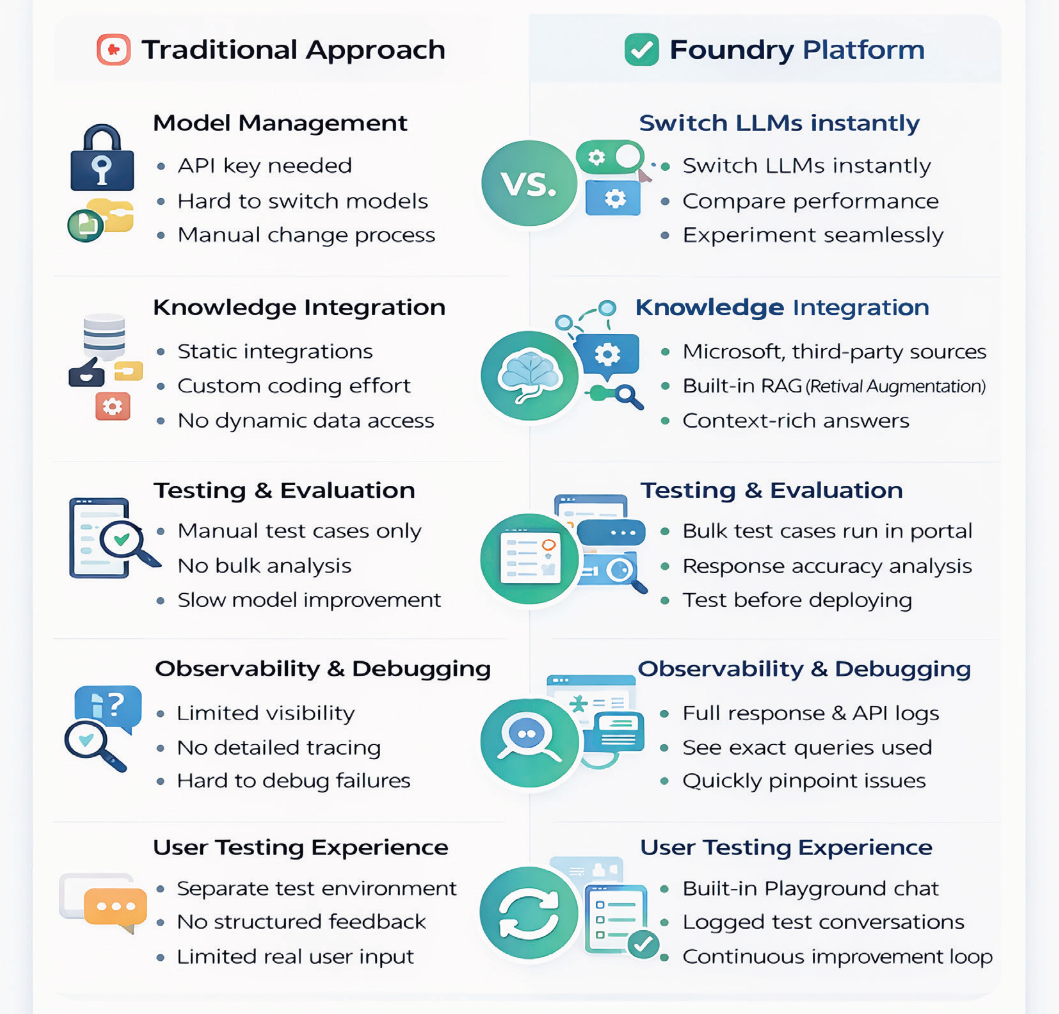
- Clearly describe guardrail policies for safety, compliance and value.
- Apply bulk remediation to rapidly repair non-compliant settings and loopholes across your fleet.

Secure Agents

- Create scheduled security scans to monitor and report performance for rolling agent testing.
- Control Plane consolidates alerts from Microsoft Defender and Purview into a dashboard for analysis, providing a one-stop shop for monitoring.

Why use Foundry over other AI development services?

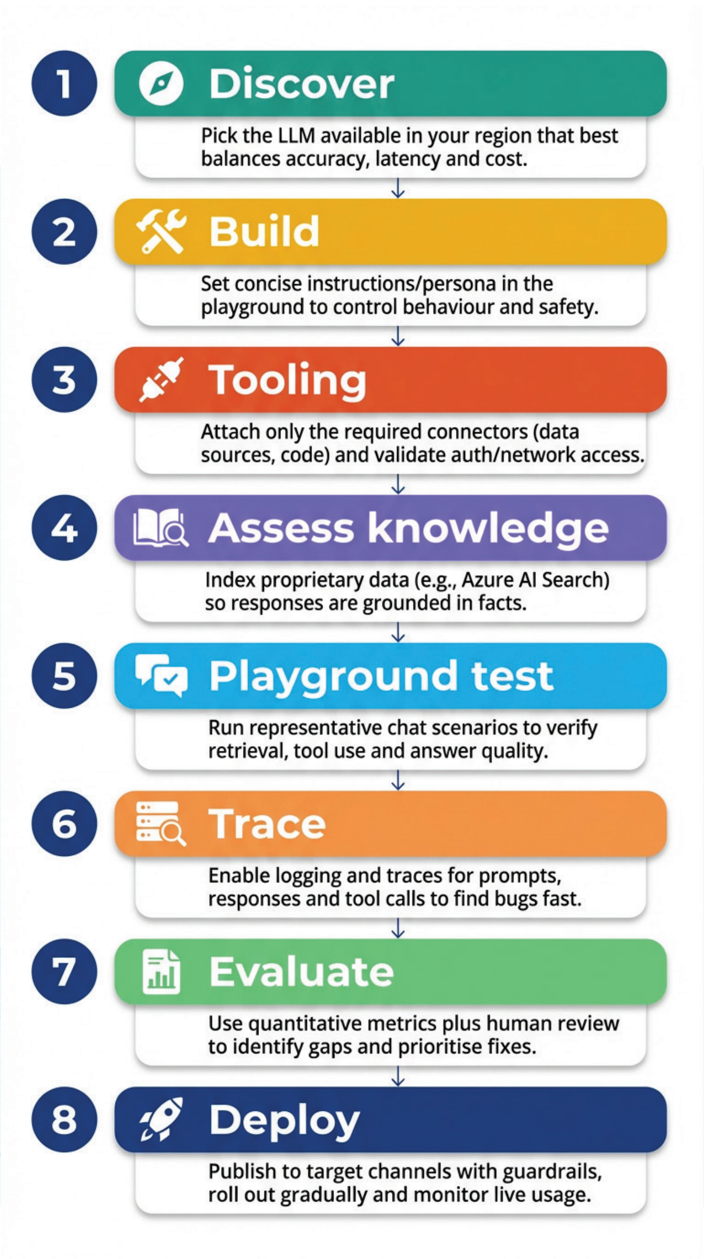
From a non-technical perspective, if you wanted to use an LLM, GPT5, for example, in your agent, you would have to make use of an API key generated by Azure OpenAI. This leaves engineers with little functionality to change LLMs and connect to knowledge bases dynamically. In contrast, **Microsoft Foundry** offers enhanced features as shown in the table below:



Traditional Approach	Foundry Platform
Model Management • API key needed • Hard to switch models • Manual change process	**Switch LLMs instantly** • Switch LLMs instantly • Compare performance • Experiment seamlessly
Knowledge Integration • Static integrations • Custom coding effort • No dynamic data access	**Knowledge Integration** • Microsoft, third-party sources • Built-in RAG (Retival Augmentation) • Context-rich answers
Testing & Evaluation • Manual test cases only No bulk analysis • Slow model improvement	**Testing & Evaluation** • Bulk test cases run in portal • Response accuracy analysis • Test before deploying
Observability & Debugging • Limited visibility • No detailed tracing • Hard to debug failures	**Observability & Debugging** • Full response & API logs • See exact queries used • Quickly pinpoint issues
User Testing Experience • Separate test environment • No structured feedback • Limited real user input	**User Testing Experience** • Built-in Playground chat • Logged test conversations • Continuous improvement loop

The Microsoft Foundry experience

Developing AI solutions in **Microsoft Foundry** requires the following process:



1 **Discover**

Pick the LLM available in your region that best balances accuracy, latency and cost.

2 **Build**

Set concise instructions/persona in the playground to control behaviour and safety.

3 **Tooling**

Attach only the required connectors (data sources, code) and validate auth/network access.

4 **Assess knowledge**

Index proprietary data (e.g., Azure AI Search) so responses are grounded in facts.

5 **Playground test**

Run representative chat scenarios to verify retrieval, tool use and answer quality.

6 **Trace**

Enable logging and traces for prompts, responses and tool calls to find bugs fast.

7 **Evaluate**

Use quantitative metrics plus human review to identify gaps and prioritise fixes.

8 **Deploy**

Publish to target channels with guardrails, roll out gradually and monitor live usage.

Responsible AI

Microsoft Foundry adheres to planning, developing and deploying reliable AI Agents that meet all security, monitoring and governance requirements, along with complete control and checkpoints at each stage. This is enforced by Microsoft Responsible AI Standard.

The following 6 core values are outlined within Microsoft's Standard:

- ✓ **Fairness**
- ✓ **Reliability & Safety**
- ✓ **Privacy & Security**
- ✓ **Transparency**
- ✓ **Accountability**
- ✓ **Inclusiveness**

As emphasised above, responsible AI is a key driving factor in delivering excellence in AI agents. To ensure this is met, Microsoft has applied a Limited Access policy to aid in the responsible deployment of Azure services.

Craig Andrew
Head of Business Development

Scan here

At **Keyrus,** we are experts within the Microsoft ecosystem and understand the fundamental components required to introduce your data into a suitable AI solution that follows industry best practices and Microsoft Foundry security measures with the appropriate guardrails.

Contact us today

www.keyrus.com/za