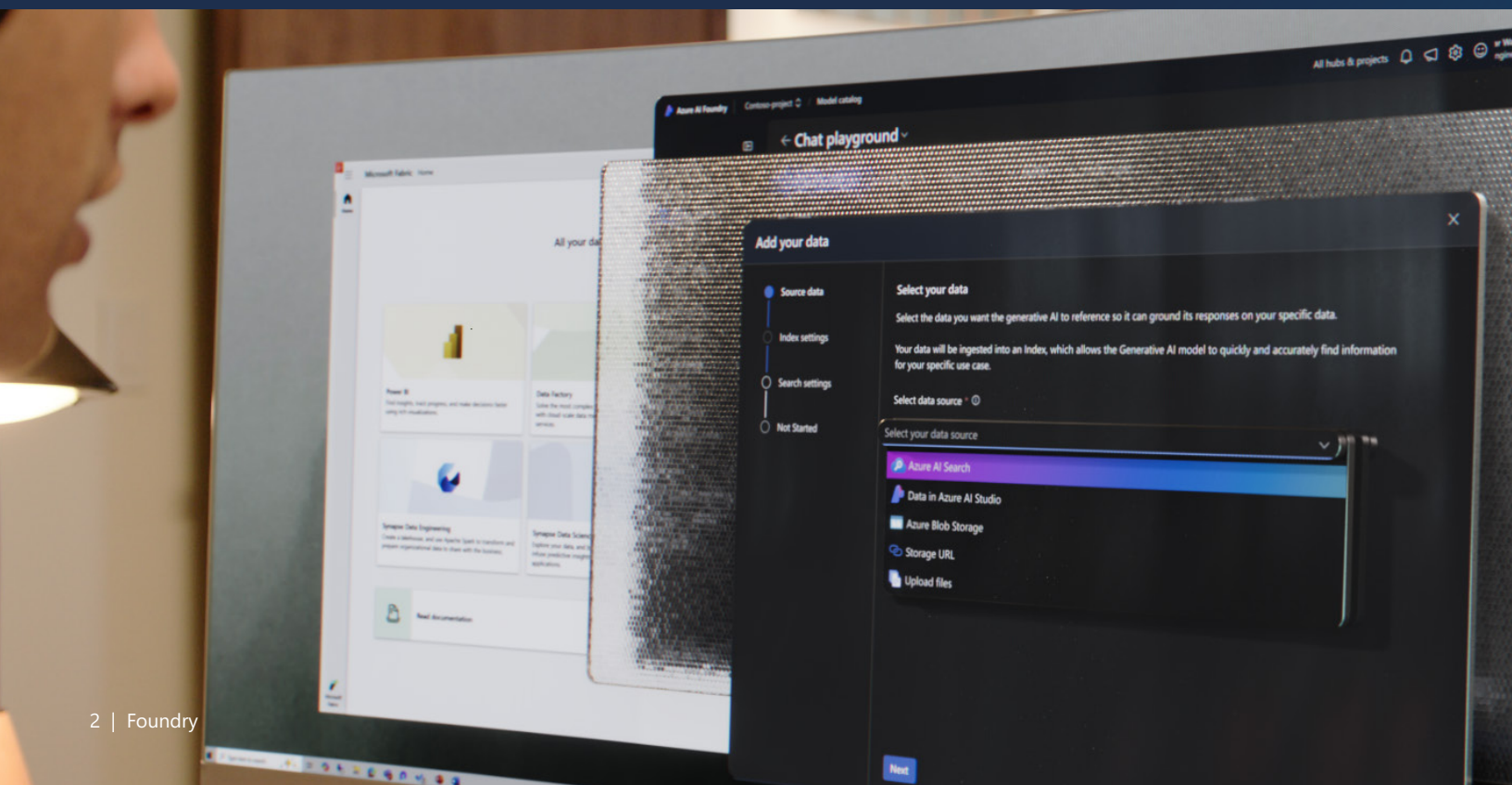
Microsoft Azure

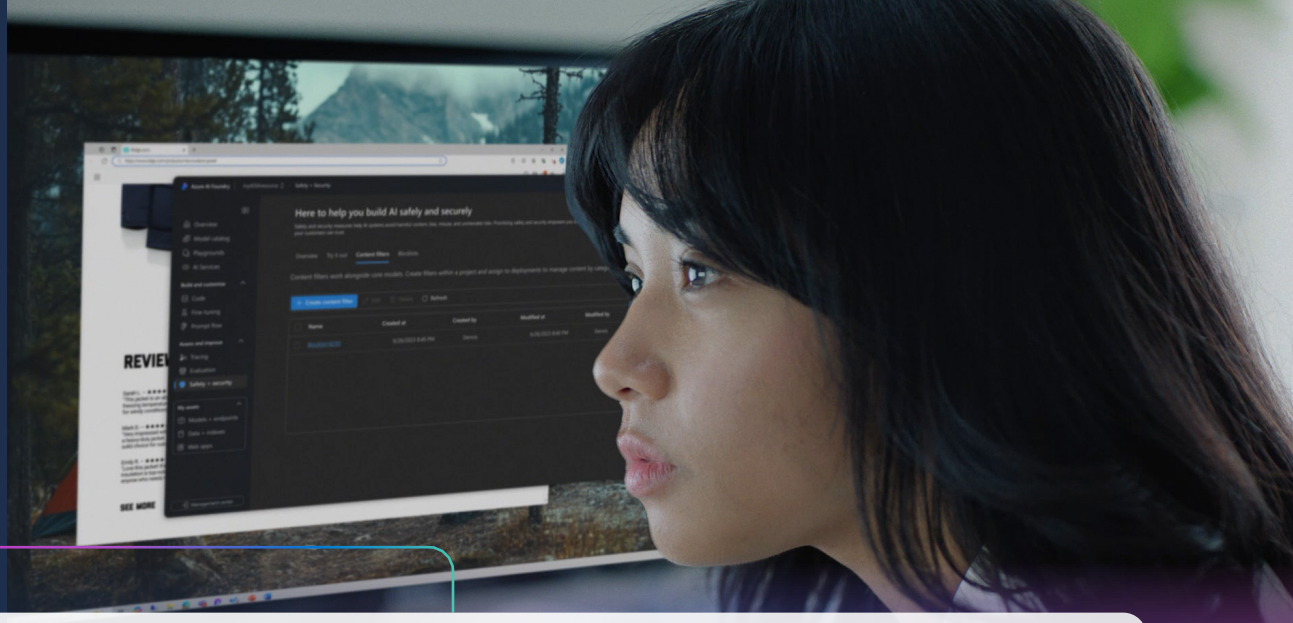
Unlocking the future:

Harnessing the power of Microsoft Foundry across organizations

Design, customize, and manage AI apps and agents with an all-in-one platform

Contents





AI is reshaping the world

Every new generation of applications brings with it a changing set of needs, and just as web, mobile, and cloud technologies have driven the rise of new application platforms, AI is changing how we build, run, govern, and optimize applications.

To meet this moment, there's been an explosion of tools and complexity across the ecosystem. We count more than 70,000 tools across governance, code, cloud, and edge.[1] It's no wonder why 80% of early POCs fail due to complexity.[2]

But it's not just the tech stack that's hard. The applications themselves are evolving. Research from Andreessen shows that innovators are using on average three models to support their applications. At the same time, agent intelligence is approaching human-level intelligence that learns over time and evolves with customers, according to Carnegie Mellon University's WebArena. Indeed, applications are moving from single models to orchestrated systems that learn and adapt continuously.

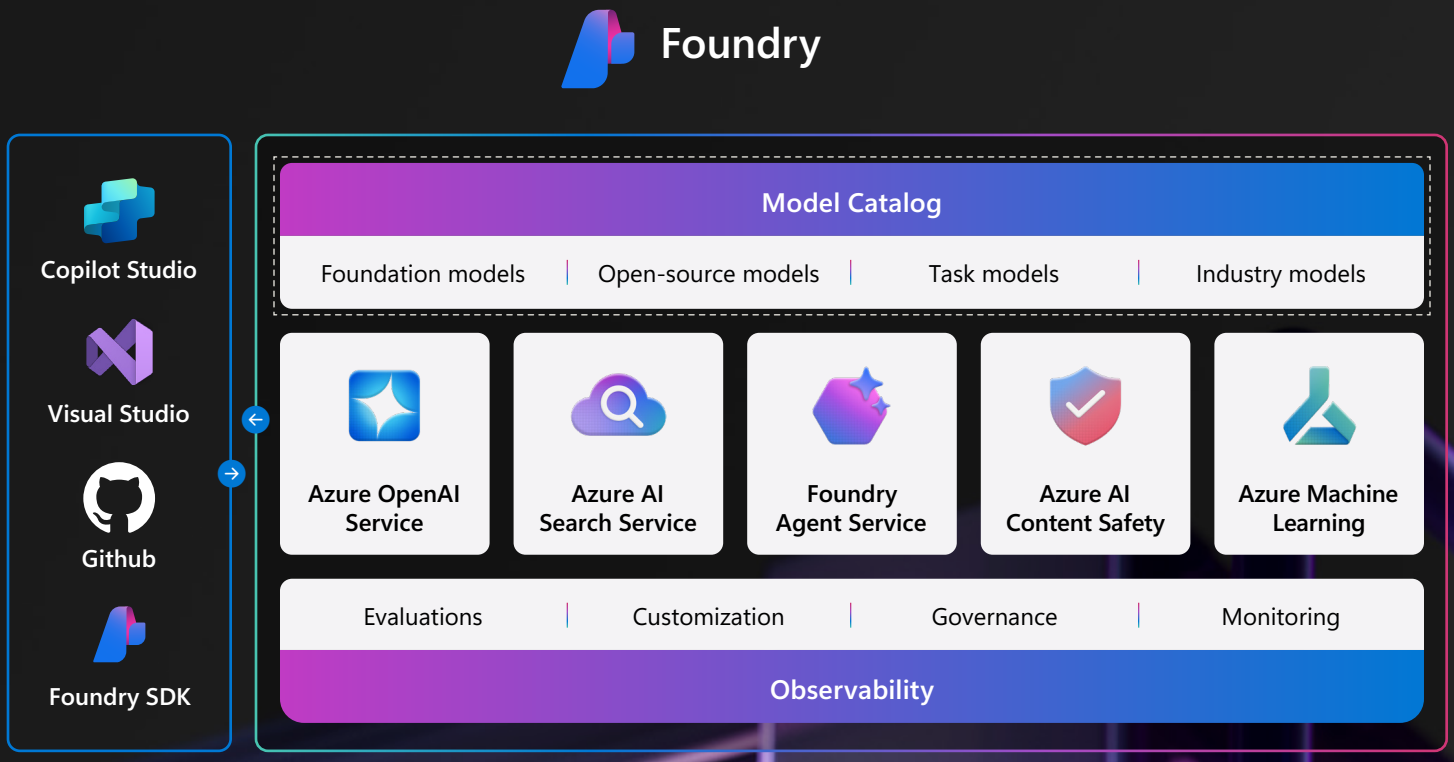
Business leaders are looking to reduce the time and cost of bringing their AI solutions to market while continuing to monitor, measure, and evaluate their performance and ROI.

Introducing Foundry

Where innovators are creating the future

Foundry is a trusted, integrated platform for developers and IT administrators to design, customize, and manage AI applications and agents. It offers a rich set of AI capabilities and tools through a simple portal, unified SDK, and APIs. What sets Foundry apart is its accessibility through the world's most loved developer tools: GitHub, Visual Studio, and Copilot Studio. This integration enables developers to work within their preferred environments, enhancing productivity and collaboration.

Foundry facilitates secure data integration, model customization, app orchestration, evaluation, and experimentation with trustworthy AI tools and principles. It also provides enterprise-grade governance and management, to help ensure AI operations are secure and compliant. By unifying data, models, and operations into a single platform, Foundry enables enterprises, startups, and software development companies to fully harness the potential of AI, driving innovation and operational excellence.



Platform experiences

Foundry SDK

A unified toolchain for AI development makes Foundry capabilities accessible through the world's most loved developer tools.

- Access popular models through a single interface.
- Integrate Foundry capabilities into apps easily.
- Develop faster with a simplified coding experience.
- Code in Python and C#.

Foundry portal

A comprehensive visual user interface that helps developers discover AI models, services, and tools.

IT administrators, operations, and compliance teams can:

- Manage AI applications at scale.
- Provision resources and manage usage across multiple hubs and subscriptions.
- Assign roles and effortlessly manage users.
- Manage quotas and compliance with organizational standards.

Foundry portal is a central hub where developers can discover, experiment, evaluate, configure, and manage AI resources. It is designed to provide a unified experience for all AI-related activities – except for the actual building and coding of applications. Building happens in Visual Studio, VS Code, or GitHub for pro-code developers, and in Copilot Studio for low-code solutions.

The double-click

Foundry portal gives you the freedom to experiment and innovate

Unified discovery: A single interface consolidates all AI models and services.

Portal environment: Interactive playground for direct model testing and experimentation offers immediate results visualization with performance metrics. Experiment version control enables reproducibility and iteration tracking.

Developer workspace: Foundry offers an integrated development environment for model testing and deployment, including custom data input support with immediate feedback loops. Saved experiments enable continuous refinement.

Setup experience: Guided configuration workflows help to prevent common setup errors. Automated validation helps ensure viable configurations. Clear upgrade paths help scale development to production.

Documentation integration: In-context help and examples appear within technical documentation in addition to API references that include working code samples.

Configuration control: Manage services and environment-specific configurations with granular settings. Track system changes with audit logs.

Team collaboration: Development teams access shared workspaces with role-based access controls on resources for a unified view of team experiments and deployments.

Ecosystem value

Empowering every role with an all-in-one platform

In the modern AI era, success hinges on collaboration between roles and departments. Data scientists are tasked with customizing foundation models to meet business needs. Developers are enlisted to build applications on these tailored models. And enterprise teams must securely scale AI solutions across the organization. Everyone has a part to play, and it helps to have an all-in-one platform that empowers an entire organization – developers, AI engineers, and IT professionals – to design, customize, and manage AI solutions with greater ease and confidence.

Foundry believes in the power of choice and interoperability. It enables collaboration, empowering an organization to harness the full potential of AI.

For enterprises, startups, and software development teams...

This means AI initiatives are secure, compliant, and scalable across the enterprise. It means seamless connection with existing security systems, data warehouses, and operational tools – for end-to-end governance, security, and operational excellence. Run your compliance tools, connect to your data lakes, and maintain your governance practices all while adopting AI capabilities.

For developers...

This means accelerating the journey from idea to code to cloud and edge deployment – turning innovative ideas into fully functional AI applications while streamlining development processes – reducing time to market. This means writing code in your preferred language and running it anywhere. Your existing applications, libraries, and development practices become part of the AI lifecycle without rewriting or restructuring.

For data scientists...

This means rapid experimentation and model customization. Our workbenches are designed to facilitate quick iterations and refinements, allowing scientists to fine-tune their models. Once perfected, these models can flow directly into production systems.

In this ecosystem, innovation flows freely between systems and teams. Run any code. Call your functions. If it can be containerized, it can be part of your AI workflow. Foundry provides smart defaults while preserving your freedom to choose the tools that work best for your needs and within your role.

Platform value for enterprises, startups, and software development teams

Foundry provides a unified platform for AI operations, model builders, and application developers. Our foundation combines production-grade infrastructure with user-friendly interfaces, to help enterprises, startups, and software developers design, customize, and manage AI applications with confidence.

Design with the best models

Discover the best models for your use case with our catalog of foundational, open, task, and industry-specific models – including same-day access to the latest OpenAI models and access to Microsoft's proprietary Phi family of models so developers never lag in AI innovation. And our unified endpoint provides seamless access to all Azure-hosted models through a consistent interface.

Customize with a comprehensive agent toolchain

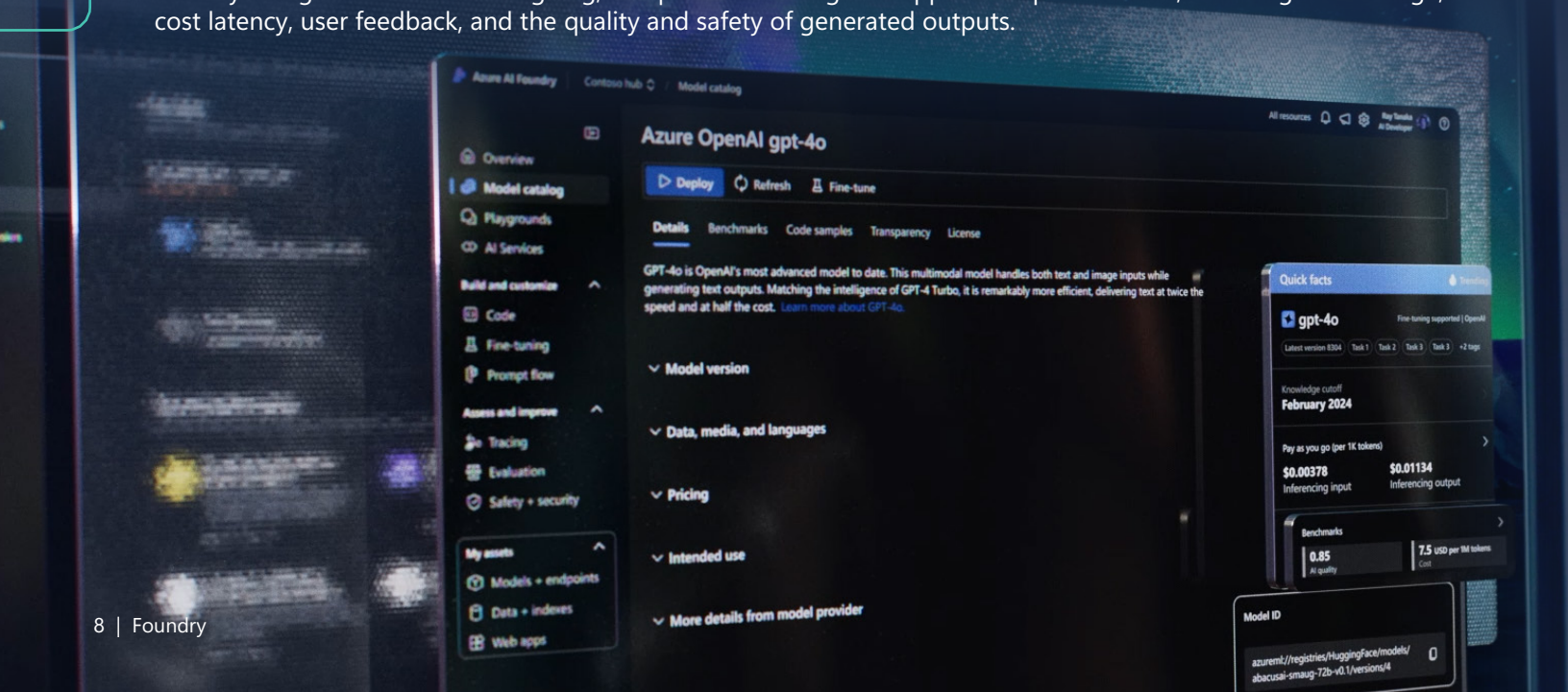
Differentiate your apps and accelerate development with a unified yet modular, flexible, and API-centric platform, including access to a collection of AI Application Templates. Foundry offers comprehensive tools and services for every stage of AI development, including Azure AI Search, Foundry Agent Service, model customization, evaluation, and experimentation. Our built-in data pipelines connect to existing data sources and operational systems, allowing organizations to leverage their existing investments while adopting AI capabilities.

Safeguard with Trustworthy AI

Design apps responsibly and safeguard with configurable filters and controls. Our Trustworthy AI commitments and capabilities, including our Copilot copyright commitment, enable a defense in depth approach with built-in tooling and guidance for responsible AI practices.

Manage AI performance in production

Deploy with continuous monitoring and governance across environments. The platform also supports AI operations with centralized resource provisioning, management, and integration with enterprise systems while maintaining security and governance. Gain ongoing, comprehensive insight of application performance, including token usage, cost latency, user feedback, and the quality and safety of generated outputs.



The double-click

Code-first platform: Foundry prioritizes developer experiences through native IDE integration. Full development and deployment support in Visual Studio, VS Code, and GitHub, with Copilot Studio available for low-code scenarios. Directly discover, experiment, and manage AI resources used in any developer workbench with Foundry portal.

Inclusive innovation: Accessibility features make Foundry easier for everyone to use, validated by developers with disabilities.

Model access: Our unified endpoint provides access to all Azure-hosted models, including those from Azure OpenAI Service and other providers, in a provider-agnostic way.

Advanced services: Task-specific capabilities for content understanding, knowledge retrieval, and action execution are designed for enterprise-scale deployment and integration.

AI app templates: Many pre-built templates accelerate the development of production applications, offering support for core GenAI functionality and use cases including embeddings generation, custom data integration, LLM orchestration, context management, interactive experiences, and more.[3]

Model customization: End-to-end support for model training, fine-tuning, and distillation enables performance optimization for specific use cases and accuracy requirements, including the creation of domain-specific AI solutions.

Service runtime: Foundry Agent Service automates routine tasks with built-in orchestration. The workflow engine handles complex business process automation and is designed for enterprise-scale deployment.

Trustworthy AI: Foundry combines the enterprise security, privacy, and safety capabilities of Azure with modern AI development patterns. Grounded in Microsoft's own best practices and learnings from building trustworthy AI products at scale, Foundry provides built-in frameworks and controls that enable cross-functional teams to proactively assess and manage risks throughout the AI application lifecycle helping align AI outcomes with organizational goals and requirements.

GenAIOps: Production-grade evaluation, monitoring, and logging capabilities support real-time anomaly detection and resource optimization.

Enterprise controls: Data protection, privacy, and governance support diverse regulatory and industry compliance requirements.

Enterprise-grade performance: Foundry leverages Azure's global infrastructure for consistent performance. This includes handling enterprise workloads with guaranteed reliability SLAs and high availability across regions.

The benefits for developers

Foundry accelerates AI development from initial PoC to production. The development experience begins like a highway on-ramp – smooth, predictable, and designed for acceleration. Teams start with their existing GitHub or Entra ID accounts for quick access to development resources. They can leverage the Foundry playgrounds to experiment with AI implementations, while employing integrated tools to evaluate and refine results. Foundry supports every stage of the AI development journey from idea to code to cloud and edge deployment.

Streamline the full lifecycle of development with Foundry

Design

Developers start by discovering and selecting the right tools, models, and documentation inside Foundry. Foundry portal streamlines resource discovery and management.

Customize

Development happens in environments you're already familiar with. Whether you're a pro-code developer using Visual Studio, VS Code, or GitHub, or a low-code developer using Copilot Studio, you can take advantage of your preferred environment with full platform capabilities through native integrations and optimized tooling. Model discovery, experimentation, and deployment tools integrate directly into these development environments, catering to rapid iterations and testing – integrating and scaling with CI/CD pipelines. Integrated AI development tools support rapid prototyping, experimentation, and refinement, including model evaluation dashboards, comparison metrics, and fine-tuning workflows to help developers iterate efficiently while maintaining high performance.

Manage

Post-deployment, developers can monitor, scale, experiment, and optimize their solutions with built-in observability and online experimentation tools. Safety remains a cornerstone of the operation phase, with features like responsible AI assessments, bias detection tools, and automated compliance checks integrated into AI workflows. These tools help ensure deployed AI solutions not only perform well but also adhere to enterprise safety principles.

Adapt with flexible options

Our SDKs and APIs come in two carefully designed flavors:

- The unified Foundry SDK offers a complete toolchain of Foundry APIs including model inferencing and select tooling including search, agent orchestration, evaluation, and tracing.
- Individual APIs enable modular access to Foundry capabilities. For example, the Azure OpenAI Service API provides access to OpenAI's capabilities, enhanced with Azure's enterprise features.

Neither choice is a one-way door – teams can always change approaches as their needs evolve.

The double-click

Discovery experience: The model catalog is accessible directly within Foundry portal, GitHub (partial), VS Code, and Copilot Studio – coming soon. In addition, there is immediate access to model playgrounds and APIs without complex setup.

Infrastructure simplification: Foundry offers automated management of Azure components – subscriptions, resource groups, and application keys. In addition, IT Admins can manage hubs, projects, VMs, GPUs, and key vaults in addition to advanced features for enterprise scenarios.

API excellence: Two robust API approaches support developer needs. The unified SDK includes unified access to models and tooling. The OpenAI-compatible API maintains same-day feature parity while adding enterprise capabilities.

SDK support: The Foundry SDK is available in Python and C#. Individual SDKs offer modular design and comprehensive documentation with easy transition between SDK options without lock-in.

App services integration: Foundry surfaces a full suite of services for model evaluation, fine-tuning, and operations. Native integration within the Azure platform offers Azure App Service, Azure Kubernetes, Azure Functions, and Microsoft CosmosDB in addition to pre-built components to accelerate GenAI app development.

Tool interoperability: Foundry offers modular integration with leading third-party tools to help preserve existing workflows. Cross-compatibility between all Foundry services supports flexible architecture with options for customizing to optimize for specific use cases.

Developer resources: Comprehensive documentation with inline examples and code samples supports rapid development. Production-ready templates and learning materials and an active internal and external developer community provide ongoing guidance.

Enterprise readiness: Foundry supports complex deployment scenarios with advanced governance and security controls available when needed. Clear upgrade paths from development to production environments are provided.

Explore comprehensive resources

In addition to this flexibility, documentation and learning resources support the developer experience. Every API and feature include integrated documentation. Code samples are available and usable. Interactive tutorials guide you through common scenarios in your IDE, while contextual help provides relevant examples and best practices. Developers also receive real-time guidance on safety and operational practices for responsible and effective deployment.

Benefit from prepackaged, managed services

Although LLMs offer significant built-in capabilities, integrating them into typical enterprise use cases often requires substantial effort. The challenges include – but are not limited to – processing messy documents, managing knowledge Retrieval (RAG), structuring output, and calling external APIs. There are also emerging deployment scenarios like edge, air-gapped, on-device, and latency-sensitive environments where LLMs are not ideal. For these tasks, Foundry offers a family of prepackaged, managed extensions – to help developers become immediately productive for the most common enterprise use cases.


Fine-tune gpt-4o-mini
Basic information
Training data
Validation data
Optional
Task parameters
Optional
Review
Task parameters
These parameters will impact both the performance and training time of your job. Default values will be programmatically defined. Learn more about task parameters.
Number of epochs
Default
Custom
Batch size
Default
Custom
Learning rate multiplier
Default
Custom
Back
Next
Fine-tune model
SAMSUNG

The benefits for data scientists

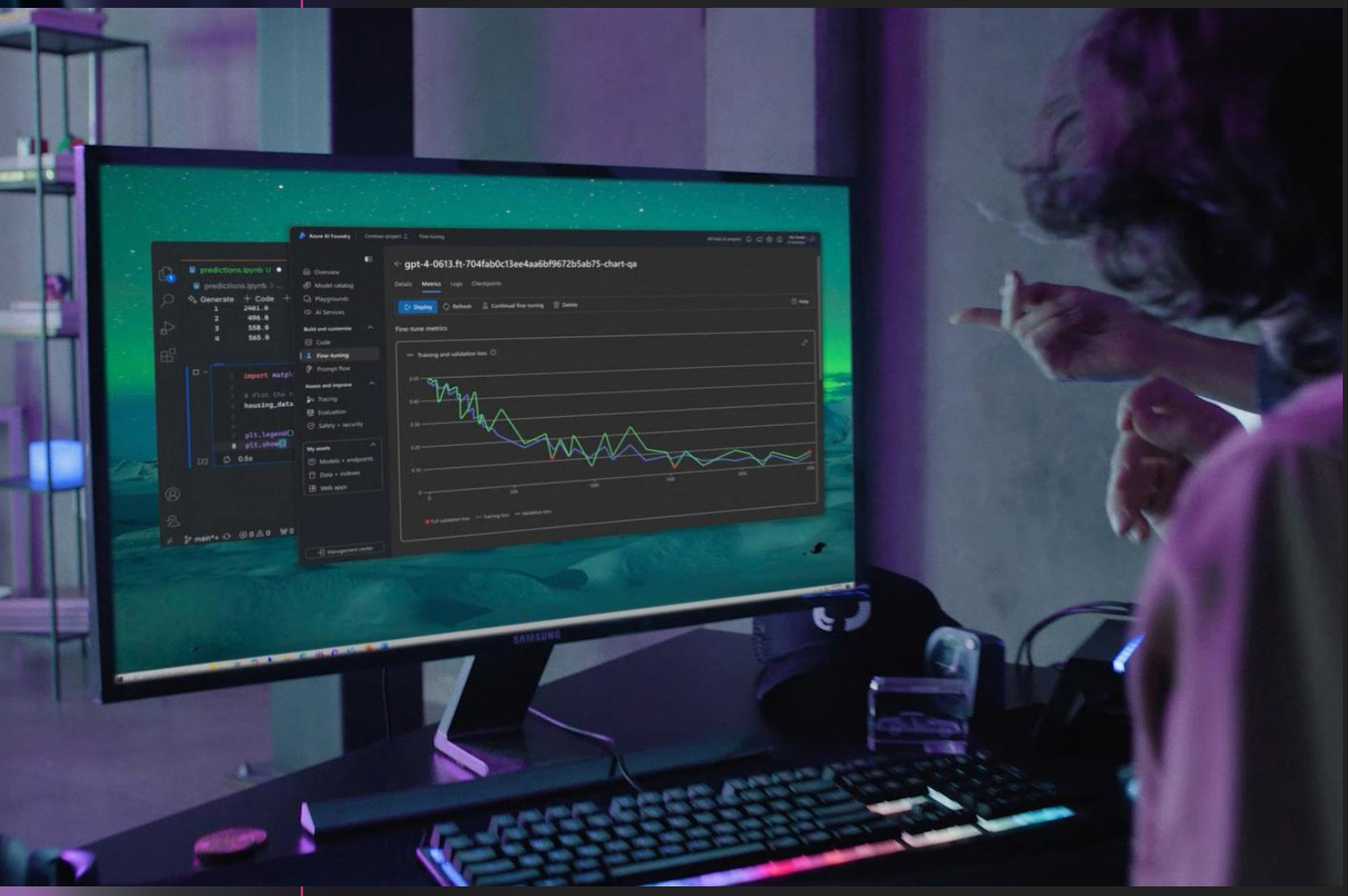
Traditional machine learning (ML) workflows are evolving into foundation model experimentation and customization.

Accelerate time to value

Data scientists can start experimenting within minutes, accessing state-of-the-art models and evaluation frameworks through secure, governed workflows. Experimentation with foundation models, including fine-tuning and prompt engineering, helps advance techniques like model distillation. Foundry also empowers teams to collaborate on model development, share insights, and leverage their organization's data assets to create customized solutions with measurable value.

Unleash innovation

Foundry provides immediate insights into model behavior with performance monitoring and automated drift detection. Successful experiments go straight to production. This integration of development and operations means scientists can focus on innovation instead of scaling infrastructure. Data scientists can deliver business impact in hours rather than weeks while inheriting Azure's comprehensive security and governance framework.



Customer commitments

Foundry is built for running AI applications in regulated environments. While we prioritize simplicity in developer experiences, we never compromise on enterprise requirements. Security, compliance, and operational excellence form our foundation.

Secure by design

Microsoft implements advanced security measures to protect sensitive data and AI models in Foundry portal, adhering to international regulations, including HIPAA BAA, ISO, HITRUST, and more. Our consistent security model means enterprises can focus on innovation rather than infrastructure management. Azure AI Content Safety is enabled by default for Azure OpenAI Service through our comprehensive toolkit, including prompt shield and groundedness detection.

Comprehensive data protection

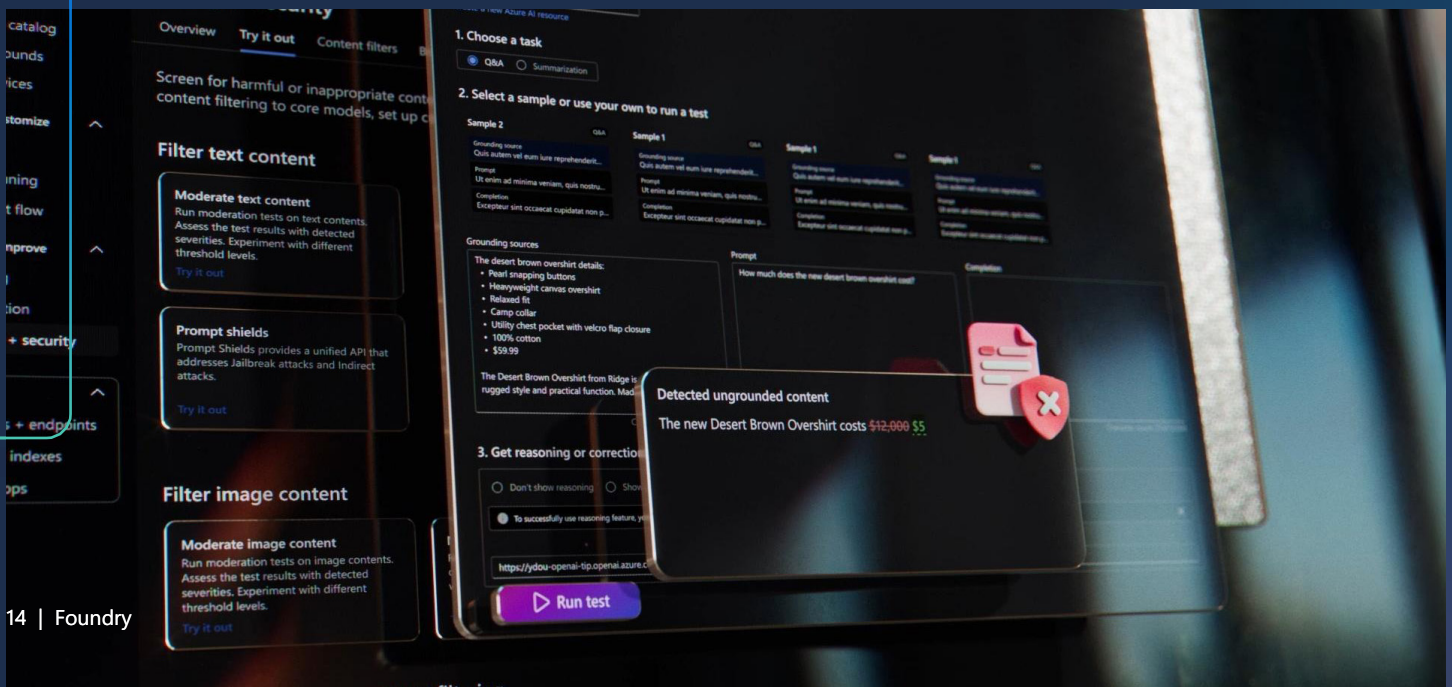
Data residency options span global, data zone, and regional deployments, optimizing for throughput and availability while maintaining comprehensive sovereignty requirements. All customer subscription data uses Microsoft-managed encryption by default, while Microsoft-managed resources employ customer-managed keys.

Zero-trust philosophy

Access control implements zero-trust principles through Entra ID integration to help prevent identity attacks while enforcing least-privilege principles. Network isolation through Private Link keeps traffic within Microsoft's backbone network, enabling secure connections to on-premises resources through ExpressRoute, VPN tunnels, and peered virtual networks. See the security baseline documentation for more details.

Simplified management

Resource management and cost optimization tools in Foundry portal and Azure Portal provide transparency and control. Our management tools enable efficient discovery, deployment, real-time monitoring, and management of AI resources. Organizations can set budgets, track usage, and optimize costs across their entire AI portfolio to optimize their AI estate. Built-in tools automatically identify opportunities to save while budget controls prevent unexpected spending.



The double-click

Security foundation: Benefit from enterprise-grade protection across services. Simplify policy enforcement with a consistent security model while enabling innovation and benefit from advanced threat protection for data and models.

Content safety: Enable industry-leading safety tools by default. Use prompt shield and groundedness detection. Customize severity levels and asynchronous modes for performance optimization, and access enterprise-specific content filtering.

Data controls: Explore flexible residency options across global, data zone, and regional deploments. Encrypt all customer data in subscription by default with Microsoft-managed keys or opt for customer-managed keys.

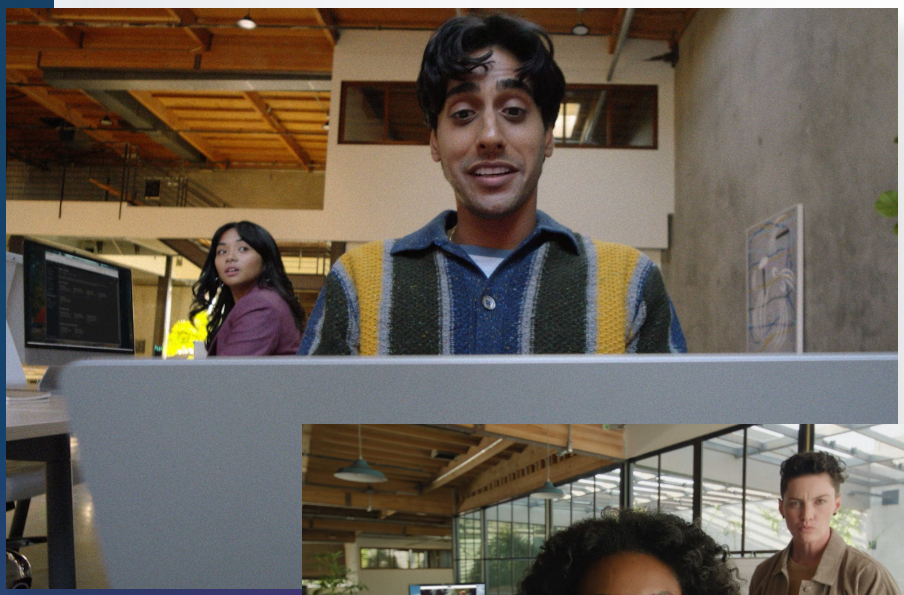
Identity management: Implement zero-trust architecture through Microsoft Entra ID integration. Use Managed Identities to eliminate hard-coded credentials. Enforce least-privilege principles with granular access controls across resources.

Network isolation: Achieve complete public internet isolation via Azure Virtual Networks and Private Link. Contain traffic within Microsoft's backbone. Secure on-premises connectivity through ExpressRoute, VPN tunnels, and peered networks.

Resource optimization: Deploy multiple models with standard or provisioned pricing. Monitor utilization and performance in real time. Prevent resource bottlenecks and cost overruns with automated thresholds.

Unified management: Manage all AI services and resources from a centralized console. Use a single interface for deployment, monitoring, and management. Gain enhanced visibility into complex AI environments. Enable secure solution development with IT Admin controls.

Compliance framework: Achieve comprehensive regulatory compliance across industries. Utilize built-in controls for data governance and audit requirements. Automate compliance reporting and documentation.



Cost governance: Optimize AI investments with flexible pricing models. Analyze and forecast usage with detailed analytics. Manage budgets and allocations. Integrate with Azure Cost Management.

Operational scale: Support reliable enterprise availability with Azure performance SLAs. Deliver consistent service with a global infrastructure backbone. Automate scaling and redundancy across regions.

Service quality commitment

Foundry supports the continuous availability of an organization's mission-critical applications.

High availability

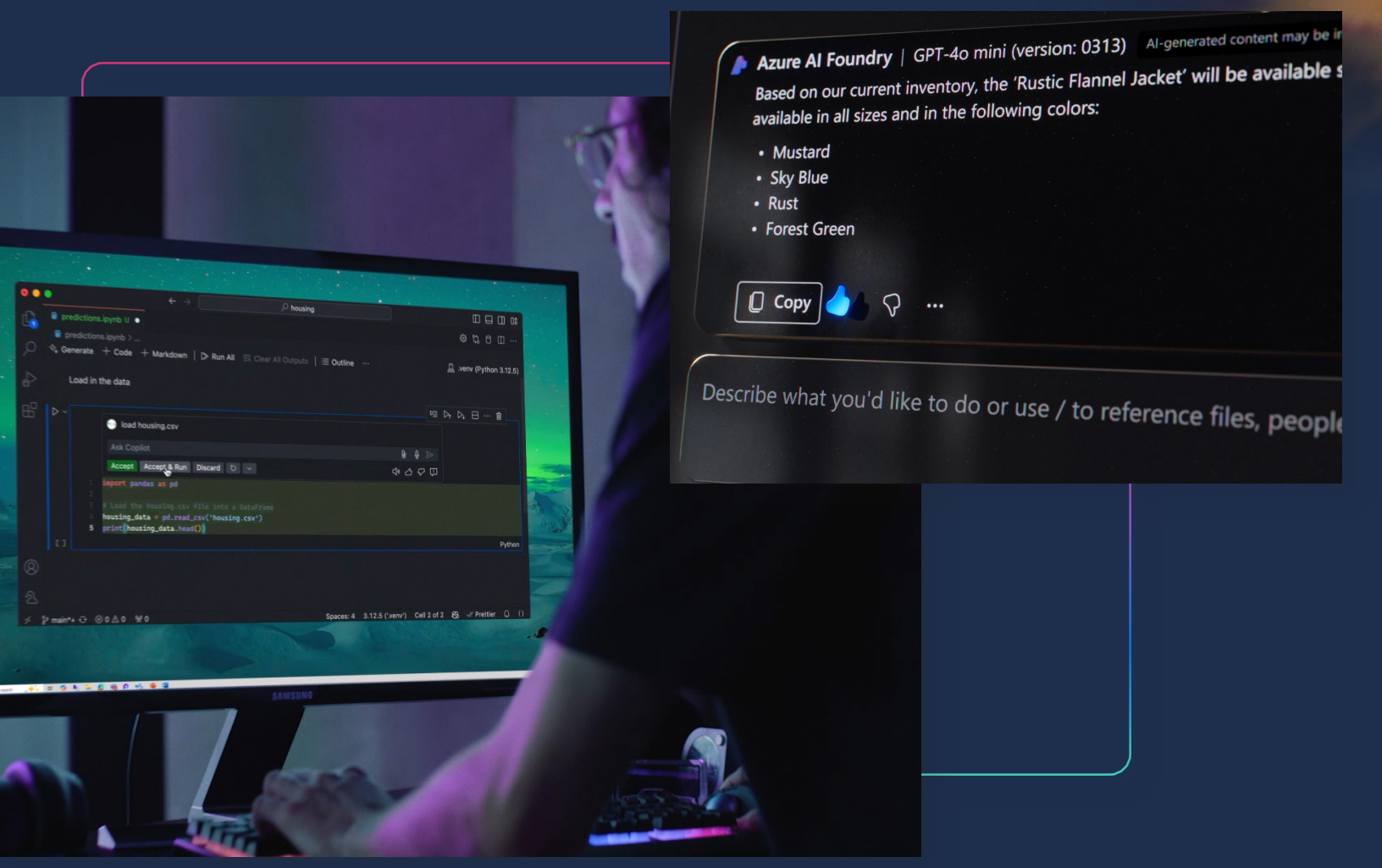
Microsoft guarantees 99.9% uptime with SLAs for availability and performance, including <100 ms response times and low-token generation latency.[4] Our throughput guarantees ensure consistent performance even under peak loads.

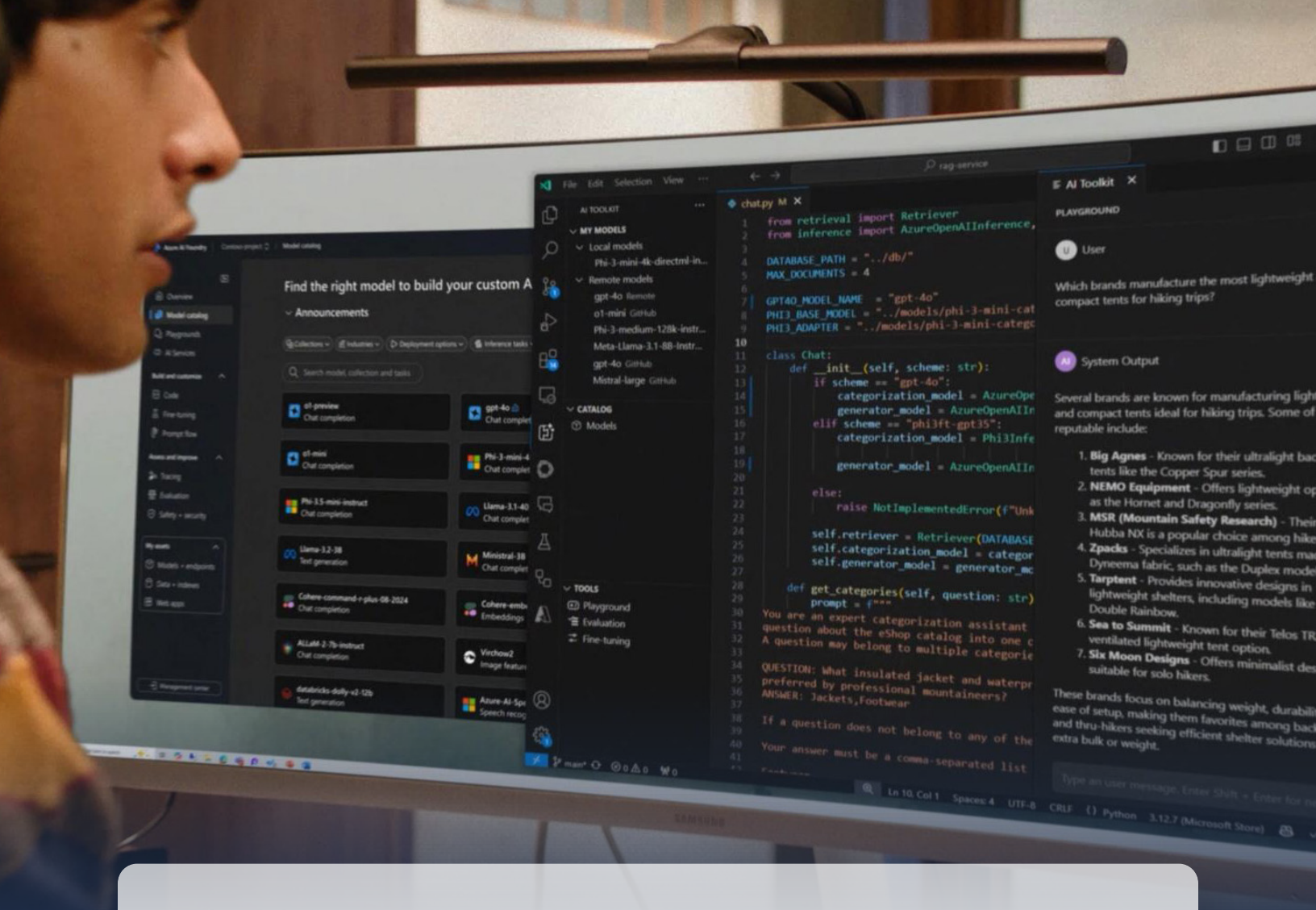
Support for business continuity

Our global infrastructure implements redundant systems and automatic failover and supports business continuity through hardware failures and regional disruptions. Real-time monitoring tools provide insights into service performance and health for proactive issue detection and resolution.

Dynamic scaling

Capacity management automatically adjusts resources based on workload demands, preventing over-provisioning and performance bottlenecks. This dynamic scaling, combined with comprehensive monitoring and alerting, helps teams achieve optimal performance without manual intervention.





The double-click

Service availability: Redundant systems and automated failover support continuous service access. Multi-region deployment options help maintain business continuity. Active architecture reduces the impact of hardware failures and regional disruptions.

Performance SLAs: We guarantee 99.9% uptime across all services.[1] Clear remediation paths and support commitments help address potential service degradation.

Dynamic scaling: Intelligent capacity management adjusts to workload demands in real time. Automated resource allocation prevents performance bottlenecks. Built-in optimizations balance performance and cost efficiency. Global load balancing enables consistent response times.

Operational excellence: Proactive platform maintenance and updates minimize disruption.

Recovery systems: Automated failover mechanisms ensure service continuity. Multi-region redundancy with instant failback capabilities maintains data consistency during recovery events.

Global reliability: Azure's worldwide infrastructure delivers consistent performance. Region-specific availability zones increase resilience. Geographic distribution of workloads improves reliability.

Support framework: We provide 24/7 enterprise support with defined response SLAs. Dedicated technical resources handle critical issues. Regular service health updates and maintenance notifications keep you informed.

Our commitment to empower every role and department to create the future of AI

In embracing the full potential of Foundry, teams are not only harnessing the power of advanced language models and robust security, they are driving a future where AI innovation and responsible AI practices go hand in hand. Foundry is designed to empower every department and role within an organization with collaborative, flexible, integrated tools to create the future of AI.

Ready to start innovating?

Design, customize, and manage AI applications and agents with Foundry.

Discover Foundry

Learn how to accelerate GenAI model selection, evaluation, and multimodal integration.

Learn

Explore in-depth technical documentation, guides, and articles to scale your projects.

Explore

[1] Full Steam Ahead: The 2024 MAD (Machine Learning, AI & Data) Landscape | Matt Turck
[2] The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed: Avoiding the Anti-Patterns of AI | RAND
[3] Build AI applications with pre-made templates | Microsoft
[4] Microsoft Online Subscription Agreement | Microsoft

Microsoft Azure