Microsoft Azure

The Fast Track to Secure AI Apps and Agents

A complete and integrated platform for creating enterprise-grade AI apps and agents with Azure

Contents

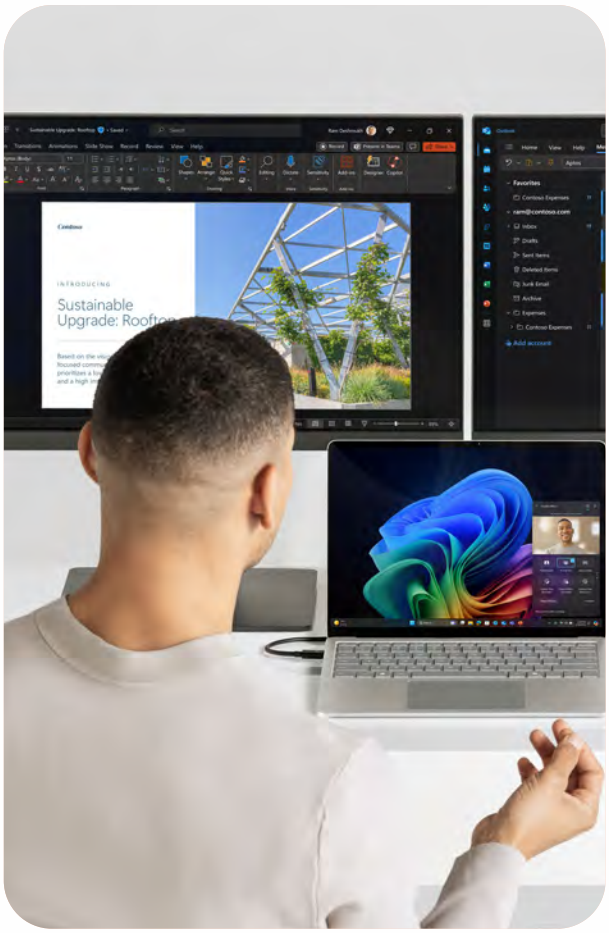
Introduction

Innovating with AI apps and agents

The last few years of application development have focused on modernization, including migrating to the cloud, adopting microservices, and building scalable platforms. But the years ahead will be defined by innovation, specifically in how businesses will create value, move faster, and differentiate themselves in the market with AI apps and agents.

AI development is evolving from being built with the help of AI tools to being fundamentally powered by AI itself. This approach makes AI an integral part of every application, enabling apps to reason over data in real time, adapt to context, and continuously improve through learning. Instead of static experiences, organizations can deliver dynamic, personalized, and predictive interactions that respond instantly to user needs and market changes. The result is not only faster development cycles but also new opportunities to create products and services that were previously impossible.

However, building apps that are truly powered by AI changes how development itself must work. Traditional methods aren't enough to manage software that learns, adapts, and interacts autonomously. This is where **agentic devops** comes in. Agentic devops is a new approach to software development that integrates AI agents across the software lifecycle. They can handle routine tasks, assist with coding, streamline testing, and enforce security practices automatically. Agentic devops goes beyond efficiency, empowering teams to accelerate innovation and bring AI-powered apps into production safely, reliably, and at scale.



Why AI app and agent innovation matters now

- **81% of leaders** expect AI agents to be part of their company's strategy in the next 12–18 months[1]
- **By 2026, 80% of enterprises** will run AI-enabled apps in production[2]

[1]Microsoft 2025 Work Trend Index Annual Report

[2]Gartner Press Release, Gartner Says More Than 80% of Enterprises Will Have used Generative AI APIs or Deployed Generative AI-Enabled Applications by 2026, October 11, 2024, https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026. GARTNER is a registered trademark and service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

What makes AI apps different?

AI applications are cloud-native, built on unified data platforms that can process information in real time, and powered by generative AI that enables agents to go beyond responding to requests—they can take action, adapt, and continuously improve. Whether it's enhancing customer experiences, streamlining business processes, empowering employees, or accelerating innovation, these agents enable apps to operate with a level of autonomy that amplifies business impact.

To fully leverage AI's capabilities, organizations often adopt approaches like agentic devops, where AI agents assist throughout the development lifecycle, helping with tasks such as testing, deployment, and monitoring. This creates a more efficient, adaptive, and intelligence-driven development process.

This approach reduces repetitive work, accelerates innovation, and allows developers to focus on solving meaningful problems. However, achieving this requires a platform and operating model specifically built for AI app development, which must be secure, responsible, scalable, and designed to integrate AI into the software creation and delivery process.

A roadmap for leaders and developers

This guide is designed for technical decision-makers and application developer leaders who want to ensure their AI adoption fuels long-term innovation. To help you embrace AI app and agent development on Azure, this e-book provides a clear roadmap for:

- Activating use cases with real-world customer momentum
- Empowering developers with integrated AI tools
- Creating AI apps and agents on Azure
- Laying the right data and platform foundations
- Ensuring security, compliance, and responsible AI
- Optimizing performance, observability, and scale
- Preparing teams and culture for AI adoption

Use cases for AI apps and agents

Before we talk about the tools that make AI apps and agents possible, it's worth taking a step back to explore what organizations are actually doing with them today. The use cases are expanding quickly, and chances are one of them may be close to what you're already planning—or spark an idea you hadn't considered yet. From transforming customer engagement to reshaping internal processes, AI-powered apps and agents are showing up everywhere, driving productivity, efficiency, and innovation in ways that once felt out of reach.

Reinvent customer engagement

- **Personalize every interaction:** Use AI to deliver engaging, on-brand experiences that boost loyalty and satisfaction.
- **Smarter product support:** Move beyond basic bots—let AI copilots and assistants answer questions and guide customers naturally about your products and services.

Reshape business processes

- **Automate and optimize:** Streamline operations with AI-powered knowledge and process automation to cut costs and improve performance.
- **Simplify policy and document workflows:** Let AI review, align, and automate policy decisions for faster, more accurate outcomes.

Enrich employee experiences

- **Boost productivity with Copilot:** Empower your team with AI-driven tools that provide real-time insights and personalized support right in their workflow.
- **Work smarter, not harder:** Automate repetitive tasks and connect your data for deeper insights and faster content creation.

Bend the curve on innovation

- **Accelerate innovation:** Use AI agents in innovation labs to quickly turn ideas into new products, shortening the time from concept to launch.
- **Speed up research and discovery**: Leverage AI to automate routine analysis and bring new solutions to market faster.

Chapter 1

Empowering developers with integrated AI tools

When you're building AI apps, one of the first priorities is enabling teams to move quickly and without unnecessary obstacles. Speed and productivity are critical not only for keeping projects on track, but also for ensuring an efficient path to market.

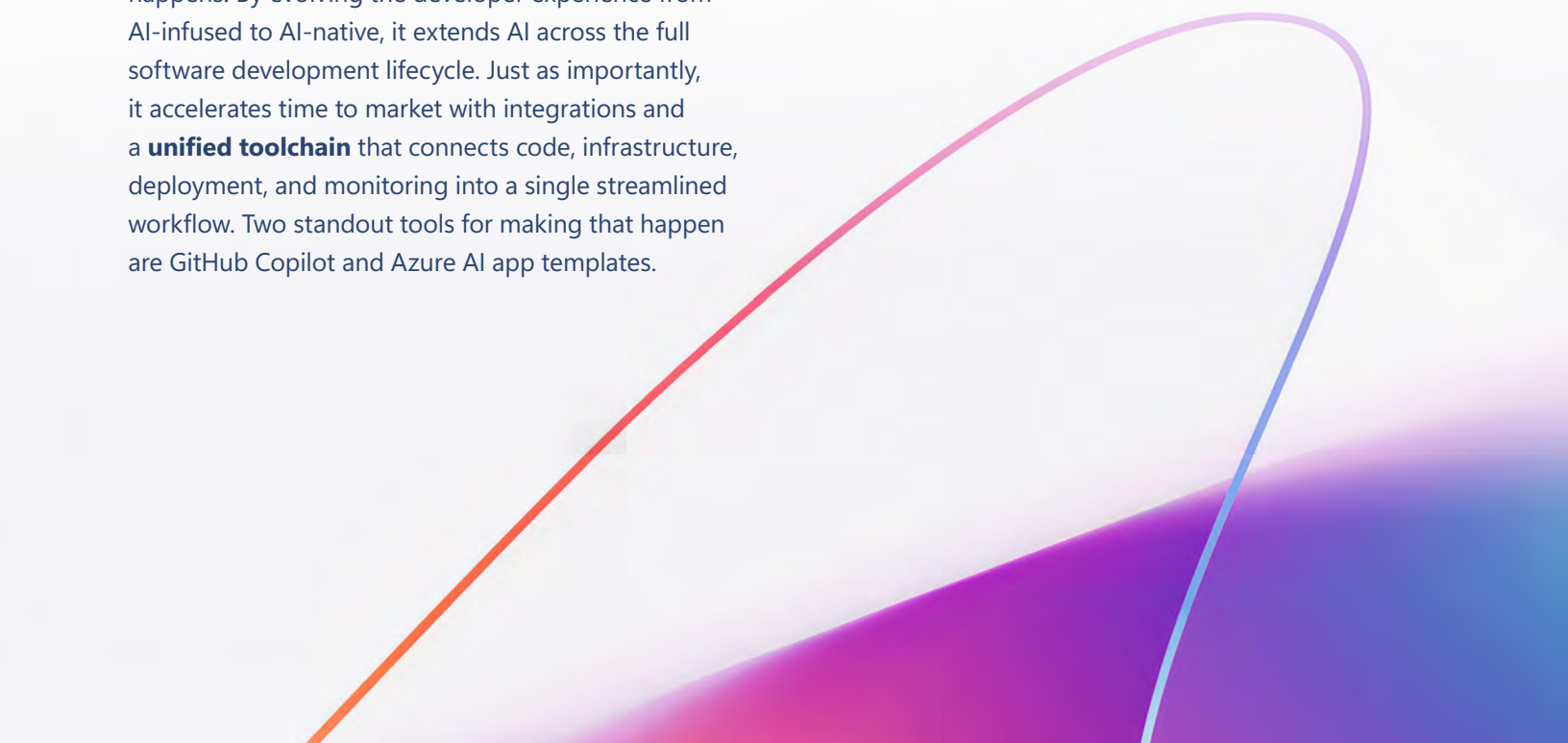
Developers striving to build AI applications often find themselves navigating a maze of complexity, switching between fragmented tools, wiring up infrastructure, configuring services, and stitching together CI/CD pipelines just to get a prototype off the ground. These tasks are essential for modern cloud development. Still, they also have the potential to introduce friction at every step, slowing progress, increasing the risk of errors, and pulling focus away from the real goal: creating impactful AI apps and agents.

Agentic devops transforms how development happens. By evolving the developer experience from AI-infused to AI-native, it extends AI across the full software development lifecycle. Just as importantly, it accelerates time to market with integrations and a **unified toolchain** that connects code, infrastructure, deployment, and monitoring into a single streamlined workflow. Two standout tools for making that happen are GitHub Copilot and Azure AI app templates.

GitHub Copilot: Freeing developers for focused innovation

GitHub Copilot is at the heart of enabling agentic devops. Beyond simple code suggestions, it can review code, identify potential vulnerabilities, and help developers resolve issues before they reach production. GitHub Copilot **agent mode** extends these capabilities to more complex, multi-step tasks: analyzing entire codebases, editing across files, generating and running tests, fixing bugs, and suggesting terminal commands, all from a single prompt and in popular editors like **Microsoft Visual Studio, Visual Studio Code, JetBrains, Eclipse, and Xcode.**

Additionally, the GitHub Copilot **coding agent** turns Copilot from a pair programmer into a more collaborative teammate. You can assign it tasks such as code reviews, test creation, bug fixes, or implementing full specifications. It can also work with other agents on more complex workflows, while maintaining audit logs and branch protections. By handling routine and repetitive tasks, GitHub Copilot enables developers to focus on higher-value work, such as designing new features, enhancing app functionality, and experimenting with AI applications.



Azure AI app templates: Jumpstart your AI projects

Azure AI app templates provide a fast, reliable way to kickstart your AI projects. They offer pre-built, customizable blueprints for common use cases, so you can focus on refining functionality instead of starting from scratch. Whether you're creating a customer support agent, a document summarizer, a code assistant, or a multimodal search tool, these templates come ready with components like model configurations, orchestration logic, and UI scaffolding. You can edit and deploy them to Azure using VS Code or GitHub Codespaces.

Each template integrates seamlessly with Azure services such as Microsoft Foundry, Azure AI Search, Azure OpenAI, and Azure Container Apps. They support flexible deployment options—from serverless APIs to containerized microservices—and are built with responsible AI principles, including observability, safety filters, and governance tools.

You can also extend these templates with your own data, tools, and workflows, making them ideal for both prototyping and production. By using Azure AI app templates, teams can accelerate development cycles, reduce setup complexity, and maintain consistency across projects.

Learning resources

GitHub Copilot agent mode: Learn how to build apps using autonomous agents that can fix errors, refactor code, and develop new features.
Take the learning module ›

Azure Developer CLI: Follow this step-by-step quick start guide for provisioning and deploying app resources to Azure using a template.
Learn more ›

Chapter 2

Customize, design, and manage AI apps and agents on Azure

Creating AI apps and agents requires models, infrastructure, and workflows to operate as a single system. With a catalog of over 11,000 models, Foundry helps unify development pipelines, enabling teams to move faster from experimentation to deployment with minimal overhead.

Foundry: Your factory for AI apps and agents

Foundry is a flexible, secure platform that helps enterprises bring AI apps and agents to production quickly. It provides a comprehensive catalog of models, agents, and tools so you can unlock your data and create innovative apps using familiar environments like GitHub, Visual Studio, and Copilot Studio.

With Foundry, teams can design, deploy, and manage agents end to end. Whether building with low-code interfaces or SDKs, developers can create agents that reason, act, and collaborate, integrating both proprietary and open-source models through a single endpoint. Meanwhile, built-in monitoring, tracing, and governance ensure enterprise-grade oversight, while prebuilt templates, benchmarked models, and observability tools accelerate time to value.

Perhaps most importantly, Foundry enables multi-agent orchestration, allowing agents to work together across workflows and platforms. By unifying model selection, orchestration, and lifecycle management in one place, it empowers teams to scale innovation with the governance, security, and performance required for enterprise AI.

Foundry developer workbench integrations

Whether you're creating agents in low-code environments or deploying enterprise-grade models, Foundry's integrations support every stage of the AI lifecycle, from design and debugging to collaboration and runtime execution.

- **Microsoft Copilot Studio:** Design and deploy custom copilots with low-code tools, integrating LLMs and business data.
- **Visual Studio:** Develop, debug, and integrate AI agents using rich IDE features and extensions for Foundry.
- **GitHub:** Collaborate on AI projects, manage code, and automate workflows with GitHub Actions and Copilot for developers.

Choose from over 11,000 models

Foundry offers access to more than 11,000 models ready for out-of-the-box use. From proprietary to open-source, these models cover a wide range of AI needs, making it easier to integrate intelligence into apps without provisioning or managing infrastructure. Developers can experiment, customize, and deploy models at scale, supported by a unified environment for training, fine-tuning, and evaluation.
Explore Foundry models ›

Foundry services

Foundry includes a rich set of services to help streamline development, orchestration, and governance across the AI lifecycle:

- **Azure AI Search:** Ground agents with enterprise data using semantic search and vector indexing for Retrieval-Augmented Generation (RAG).
- **Foundry Agent Service:** Operate and orchestrate agents across development and production with secure, scalable runtime integration.
- **Azure Machine Learning:** Train, fine-tune, and deploy custom models with MLOps support for scalable AI workflows.
- **Foundry Tools:** Access prebuilt APIs for vision, speech, language, and decision-making to accelerate app development.

Trustworthy AI in Foundry

From development to deployment, Foundry offers built-in security, observability, and content safety features that help teams build responsibly and operate with confidence:

- **Azure AI Content Safety:** Enhance the safety of generative AI apps with advanced guardrails for responsible AI
- **Foundry Observability:** Optimize and scale AI apps and agents with end-to-end monitoring, tracking, and evaluation

Azure OpenAI in Foundry Models: The intelligence layer

Having access to a diverse set of models is crucial because each model excels at different things. Some are optimized for speed, others for reasoning, some for multimodal input, and others for domain-specific tasks. This flexibility empowers developers to choose the right tool for the job, experiment across modalities, and build solutions that are both performant and cost-effective.

Azure OpenAI in Foundry Models brings OpenAI's most advanced language and multimodal models into Foundry, giving developers a secure, scalable, and governed platform for building generative AI apps and agents. Whether you're looking to summarize complex documents, generate code, build conversational interfaces, or orchestrate multi-step reasoning across enterprise data, you can take advantage of this expansive intelligence layer to rapidly prototype, fine-tune, and deploy AI-powered agents and apps that adapt to real-world tasks with precision and reliability.

Capabilities and benefits

Foundry helps teams transform ideas into AI-powered solutions across a variety of use cases, including copilots, summarization, and personalization. From everyday productivity tools to complex enterprise workflows, it enables developers to build, deploy, and scale apps and agents quickly and efficiently.

- **Access to advanced models:** Streamline access to cutting-edge models—including open-source, proprietary, and Microsoft-hosted models—within a unified development environment.
- **Flexible deployment:** Use serverless APIs, managed compute, or batch processing depending on your workload.
- **Agent integration:** Models can power intelligent agents that automate workflows, generate content, or interact with users. For example, a Copilot agent might generate and review code, while a customer service agent can summarize conversations and suggest responses.
- **Customization and fine-tuning:** Tailor models for domain-specific tasks or workflows using easy-to-use fine-tuning tools.
- **Responsible AI tooling:** Built-in safety features, content filters, and compliance frameworks help ensure the ethical and secure use of AI.

Azure Direct Models

Foundry models include Azure Direct Models, a curated set of models sold and hosted directly by Microsoft. This collection includes all Azure OpenAI models, such as GPT-4.1, GPT-4o (multimodal), GPT-5 series, Codex, and Embeddings—as well as select models from top providers like Meta's Llama 3, Mistral, and Cohere Command R+. These models are optimized for Azure infrastructure and come with enterprise-grade support, SLAs, and responsible AI tools.

Bring it all together with agent orchestration

Orchestration is the coordinated management of multiple AI agents, ensuring they work together seamlessly to complete tasks and achieve desired outcomes. It's a crucial capability for building AI-native apps that can efficiently handle complex workflows. By enabling multiple AI agents to collaborate, share knowledge, and respond dynamically to context, orchestration allows teams to create more sophisticated solutions. Teams can build multi-agent copilots, personalized assistants, or automated summarization systems without introducing added complexity or redundancy.

Azure provides the infrastructure and tools to make this possible. Developers can design and manage agents using the Microsoft Agent Framework, run and manage multi-agent workflows in Foundry with built-in observability, memory, scaling, and enterprise security—or self-host flexibly on Azure Container Apps. Azure AI Search and vector similarity help agents retrieve relevant data or tools, while the Agent2Agent Protocol facilitates communication and memory sharing between agents. Event-driven orchestration is supported via Azure Functions and Logic Apps, enabling flexible, modular, and secure AI solutions.

Orchestrating your agents with Foundry Agent Service

Foundry Agent Service is the central runtime layer that connects the core components of Foundry (models, tools, and frameworks) into a single, production-ready platform. It manages threads, orchestrates tool calls, enforces content safety, and integrates with identity, networking, and observability systems, ensuring agents are secure, scalable, and reliable. By handling infrastructure complexity and building trust and safety by design, the service enables organizations to confidently move AI agents from prototype to production.

The Foundry Agent Service also integrates seamlessly with Azure data services—including Cosmos DB for agentic memory and chat history, Azure SQL Database, and Azure Data Lake—allowing agents to query, store, and reason over structured and unstructured data. The result is a system of agents that work together like a well-orchestrated team, boosting productivity, accelerating development, and delivering smarter, more adaptive applications.

To enable collaboration and task execution across agents, the service works with orchestration frameworks that define how agents communicate, coordinate, and reason together, offering flexible patterns for building scalable, goal-driven AI systems. Here are just a few of the available frameworks:

- **Microsoft Agent Framework:** SDK for building AI agents and copilots with LLMs, plugins, and memory, enabling modular skills and structured, goal-driven workflows.
- **AutoGen:** Framework for multi-agent collaboration where agents communicate to refine outputs, ideal for brainstorming, negotiation, and consensus tasks.

Chapter 3

Laying a modern data foundation for AI

Databases are the beating heart of AI applications. They do more than store data; they fuel intelligence, enable contextual understanding, and drive real-time responsiveness. Choosing the right database is a strategic decision that shapes how your AI apps can learn, adapt, and deliver value.

AI apps and agents depend on massive volumes of structured and unstructured data, fast retrieval of relevant context, and the ability to make decisions in real time. The right database supports the storage of embeddings, powers retrieval-augmented generation (RAG) workflows, enables real-time analytics, and provides secure, scalable, and low-latency access to operational data. Top-tier options like **Azure Cosmos DB**, **Azure SQL Database**, and **Azure Database for PostgreSQL** offer these capabilities and more, making them powerful assets for AI app development.

Choosing the right Azure database for your AI app and agent workloads

Database	Best for	
Azure Cosmos DB	Globally distributed, multimodel apps needing low-latency access across regions	IoT, real-time analytics, personalization, and mobile backends
Azure SQL Database	Open-source relational workloads with advanced SQL and extensibility	Web apps, geospatial apps, and analytics-heavy workloads
Azure Database for PostgreSQL	Enterprise-grade relational apps with deep Microsoft integration	Line-of-business apps, financial systems, and reporting platforms

Key database capabilities for building intelligent, integrated, and trusted AI apps include:

- **Vector search:** Modern AI applications need more than keyword matching. Vector search allows databases to store high-dimensional embeddings, representations of text, images, or other data, and retrieve information based on semantic similarity. This capability powers recommendation engines, chatbots, and summarization tools. Azure Cosmos DB and Azure SQL Database now support native vector indexing, while PostgreSQL offers vector search through extensions like pgvector.
- **Retrieval augmented generation (RAG):** RAG combines large language models with external knowledge stored in databases. By querying vector embeddings for relevant context before generating a response, AI models improve accuracy, reduce hallucinations, and gain domain-specific intelligence. Azure Cosmos DB integrates seamlessly with Foundry Tools for RAG, while Azure SQL and Azure Database for PostgreSQL can connect to embedding models to build flexible pipelines.
- **Real-time analytics:** AI apps thrive on immediacy. Tracking user behavior, system performance, and data trends in milliseconds allows apps to adapt dynamically, whether delivering personalized content, detecting anomalies, or adjusting interfaces. Azure Cosmos DB's analytical store and change feed, Azure SQL's integration with Synapse, and PostgreSQL extensions like TimescaleDB all support this capability.
- **Thread storage:** Azure Cosmos DB for NoSQL lets AI agents securely store and manage threads in your own Azure Cosmos DB account. This enables developers to persist and retrieve multi-turn conversations directly within their resources. Maintaining structured interaction histories improves an agent's contextual understanding, leading to more coherent and relevant responses. It also provides valuable insights into user behavior, agent decisions, and system performance, helping troubleshoot issues and optimize outcomes.

Learning resources

Azure Cosmos DB: Design and build NoSQL generative AI and multi-agent apps. **Learn more ›**

Azure SQL Database: Manage AI-powered APIs, host and manage web apps, and develop data-driven apps with Azure SQL. **Learn more ›**

Azure Database for PostgreSQL: Implement machine learning inferencing with PostgreSQL data. **Learn more ›**

IaaS versus AI-optimized databases

When building AI apps and agents, choosing between infrastructure as a service (IaaS) and AI-optimized databases affects scalability, performance, and development speed.

Aspect	IaaS constraints	AI-optimized database advantages
Setup and configuration	Manual setup of compute, storage, and networking; time-consuming and error-prone	Preconfigured for AI workloads; minimal setup required
AI capabilities	Limited support for embeddings, vector search, or semantic indexing	Native support for RAG, embeddings, multimodal data, and semantic search
Operational overhead	Requires manual scaling, patching, and monitoring	Built-in scalability, observability, and governance
Performance	Potential latency and bottlenecks for real-time AI tasks	Optimized for low-latency, high-throughput AI inference and retrieval
Flexibility	Full control over infrastructure and architecture	Purpose-built for AI workloads; may limit general-purpose querying
Cost efficiency	Cost-effective for static workloads, but can be expensive for AI optimization	High-performance retrieval and processing optimized for AI use cases
Security and compliance	Requires custom implementation of security and compliance controls	Integrated responsible AI tooling with enterprise-grade security features
Developer experience	Slower iteration due to manual infrastructure management	Faster prototyping and deployment with AI APIs and tools, while letting teams work in familiar programming languages

Chapter 4

Integrated platform for cloud-native development

Developing AI apps in the cloud can quickly become complicated without a unified platform. Fragmented services, inconsistent environments, and operational overhead can slow development, introduce errors, and distract teams from building value. Developers may face multiple interfaces for compute, storage, networking, and AI, while manually scaling workloads across cloud and edge or maintaining security and compliance.

Azure provides a set of integrated cloud-native services that streamline this complexity, making it easier to deploy, manage, and scale intelligent AI applications. Key services include Azure Kubernetes Service (AKS), Azure App Service, and Azure Container Apps, with each one addressing different challenges in the AI development lifecycle.

Azure API Management plays a crucial role in integrating these services and exposing their functionality to agents, users, or other applications. It enables teams to create secure, scalable, and well-documented APIs that act as the connective tissue between services and agents. With built-in support for versioning, throttling, authentication, and analytics, Azure API Management ensures that AI agents can interact with tools and data sources reliably, whether through the Model Context Protocol or in Agent-to-Agent (A2A) workflows.

By combining these services, Azure empowers developers to build modular, agentic systems that are not only powerful and intelligent but also governed, observable, and enterprise-ready.

Azure App Service

Best for: Web-based AI agents, APIs, and user-facing applications

Azure App Service streamlines deployment by providing uniform runtime environments across dev, test, and production. Developers can push code directly from GitHub or Visual Studio, with automatic scaling and no infrastructure management. Integrated AI services, such as Azure OpenAI and Azure Machine Learning endpoints, enable seamless addition of AI functionality. Security is built in, with support for VNet integration, managed identities, and Key Vault, while Application Insights provides monitoring and diagnostics.

This makes Azure App Service well-suited for conversational agents, recommendation engines, or real-time dashboards that require consistent performance and secure connectivity. For instance, picture a customer support chatbot that handles order queries, suggests personalized products, and escalates issues seamlessly across web and mobile channels.

Azure Container Apps

Best for: Lightweight, event-driven AI microservices and edge-ready agents

Azure Container Apps offer serverless containers that scale automatically based on traffic or events, including scale-to-zero for cost efficiency. Event-driven triggers enable reactive AI workflows, while GPU support provides pay-per-second compute for AI inferencing. Container Apps abstract away Kubernetes complexity while still running on AKS, and they support secure networking and managed identities for edge deployments.

These features make Container Apps ideal for reactive AI services, such as image classification triggered by uploads or autonomous agents responding to sensor data. Think of an IoT-driven predictive maintenance agent that responds to sensor data, predicts equipment failures, and triggers automated alerts to technicians.

Azure Kubernetes Service (AKS)

Best for: Complex, distributed AI workloads with custom orchestration needs

Azure Kubernetes Service (AKS) centralizes container management, simplifying deployment, scaling, and monitoring of AI microservices, model training jobs, and inference endpoints. It supports hybrid deployments with Azure Arc, allowing models to run at the edge while maintaining centralized control. Built-in integrations with Foundry Tools, databases, and monitoring tools like Prometheus and Grafana provide end-to-end connectivity, while GitOps workflows and CI/CD pipelines make devops continuous and predictable.

AKS is ideal for scenarios requiring full control over AI infrastructure, such as multi-agent systems or real-time inferencing pipelines. For example, imagine a real-time fraud detection system where multiple AI agents monitor transactions, cross-reference user behavior, and flag anomalies across regions.

Azure Kubernetes Service Automatic

With automated provisioning, scaling, and updates, **Azure Kubernetes Service (AKS) Automatic** is a streamlined mode of AKS that makes Kubernetes faster and easier to use, giving developers and teams production-ready clusters without the complexity of manual setup.

Learning resources

AKS: Learn how to build and deploy apps that use OpenAI. **Learn more** ›

Azure App Service: Discover how to create apps using Azure OpenAI and OpenAI. **Take the tutorial** ›

Azure Container Apps: Follow the step-by-step quick start guide to deploy your first container app. **Learn more** ›

The growing importance of cloud-to-edge scalability

Cloud-to-edge capability ensures AI apps stay fast, responsive, secure, and scalable. Without it, performance, cost, and reliability all take a hit.

Without cloud-to-edge deployment:

- AI apps are tied to centralized servers, creating latency for real-time tasks.
- Workloads can bottleneck under high traffic, slowing performance.
- Edge devices can't run AI models locally, limiting responsiveness for IoT, industrial, or mobile applications.
- Data transfer costs and security risks increase as everything must travel to the cloud.

With cloud-to-edge scalability:

- AI models run close to where data is generated, enabling real-time decisions.
- Workloads scale automatically across cloud and edge, avoiding bottlenecks.
- Edge devices handle sensitive or high-volume data locally, improving privacy and reducing bandwidth costs.
- Distributed deployment supports mission-critical applications across various industries, including manufacturing, healthcare, retail, and logistics.

Chapter 5

Secure, governed, and responsible AI

Security is one of the top concerns in AI app and agent development, as risk leaders consistently identify sensitive data exposure as their greatest concern. And without robust governance frameworks, multi-agent environments can quickly become difficult to manage. On top of that, many businesses fear being locked into outdated models, which limits flexibility and innovation.

This evolving threat landscape means you need more than just tools to build AI apps; you need a secure foundation. Protecting your data, AI models, and applications end-to-end is critical to reducing business risk and fostering trust. Multilayered security offers this protection by layering defenses across infrastructure, identity, governance, and runtime operations. If one layer is compromised, others remain in place to prevent further damage.

Azure provides always-on guardrails and clear policy commitments to help ensure the safety, security, and privacy of AI apps and agents. Deep integration across security, compliance, and identity solutions helps you address risks at every stage of development and deployment, building trust into every layer of your AI stack.

Microsoft Defender for Cloud: Code-to-cloud protection

Microsoft Defender for Cloud is a cloud-native application protection platform (CNAPP) that helps protect AI apps and agents across the entire development lifecycle. It provides visibility into your security posture and detects vulnerabilities before they become threats, allowing you to identify misconfigurations, secure workloads, and monitor compliance benchmarks mapped to industry standards.

Defender for Cloud also helps protect AI-specific scenarios by detecting issues such as prompt injection, policy violations, and runtime threats, with the ability to block active attacks. With workload protection across virtual machines, containers, databases, and storage, plus agentless vulnerability scanning, you can keep your AI environment secure from code to runtime

Defender for Cloud capabilities:

- **AI threat detection:** Detect prompt injection, policy violations, and runtime threats, with the ability to block active attacks.
- **Security posture monitoring:** Visualize and improve security posture proactively.
- **Regulatory compliance:** Get compliance benchmarks mapped to industry standards.
- **Fraud detection:** Use AI and machine learning to catch fraud throughout your app infrastructure.
- **Attack-path analysis:** Discover and prioritize critical risks and contextual threat analysis.
- **Workload protection:** Protect workloads from malware across virtual machines, containers, databases, and storage.
- **Vulnerability scanning:** Use an agentless or agent-based approach to scanning for vulnerabilities.
- **DevOps posture:** visibility Unify visibility into devops inventory across multicloud and multiple-pipeline environments.
- **Infrastructure-as-code security**: Secure configurations throughout the development lifecycle.
- **Code security guidance:** Speed up remediation of critical issues in code.

Microsoft Purview: Data security and governance for AI

Microsoft Purview is a comprehensive data governance and compliance solution that helps organizations manage, protect, and get more value from their data. Whether your data lives on-premises, in the cloud, or across hybrid environments, Purview helps organizations secure and govern the data that powers AI models, from training and fine-tuning to real-time inferencing and grounding.

With Purview, you can:

- Protect sensitive data during runtime to prevent exposure through model inputs and outputs.
- Monitor and detect risky user behavior, with actionable recommendations to reduce insider threats.
- Apply governance policies that classify, label, and manage AI data with precision and transparency.
- Ensure regulatory compliance through built-in audit trails, retention rules, and communication oversight.

By embedding security and governance into every stage of the AI lifecycle, Purview empowers teams to innovate confidently, minimizing risk, preventing data leakage, and maintaining a trusted and compliant data estate.

Microsoft Entra ID: Secure identity for AI apps and agents

Managing identities for AI apps and agents is just as important as managing human users. With **Microsoft Entra ID**, you can securely integrate identities into environments like AKS clusters, while managed identities in App Service eliminate secrets from code and connection strings.

For AI agents, Entra Agent ID enables scalable identity management with least-privileged access, scoped tokens, and complete discovery and auditing across multiple tenants. Together, these capabilities ensure your AI agents operate securely, with controlled and auditable access at every stage.

Azure AI Content Safety: Real-time safeguards

AI apps and agents must be designed to avoid producing harmful or unsafe content. **Azure AI Content Safety** helps detect and filter unsafe prompts and outputs in real time, protecting against issues such as prompt injection, harmful user-generated content, and biased responses. By embedding safety checks directly into your AI workflows, you can deploy responsible AI solutions that protect users, safeguard data, and uphold your brand.

GitHub Advanced Security: DevSecOps for AI development

Secure AI starts with secure code. **GitHub Advanced Security for Azure devops** provides end-to-end protection with tools such as secret scanning, dependency review, and code scanning. These tools help surface vulnerabilities early in the development cycle, allowing teams to remediate faster and ship secure applications with confidence. Integrated into CI/CD pipelines, GitHub Advanced Security helps reduce risks by embedding security checks early in the development process.

Compliance: Meeting regulatory and policy requirements

Compliance is central to responsible AI. With Microsoft's compliance portfolio, you can align your AI solutions with global and industry-specific regulations, while also benefiting from enterprise-grade auditing and monitoring. Built-in tools, such as Azure Policy and Compliance Blueprints, simplify the creation of compliant environments. Additionally, integrations with partners extend governance to encompass fairness and regulatory tracking.

Learning resources

Microsoft Defender for Cloud:
Discover how to enable multicloud protection. **Learn more ›**

Microsoft Entra ID:
Watch the video to understand key concepts, including managing users, groups, and service principals. **Watch now ›**

Microsoft Purview:
Follow this guide to securing and managing your data estate. **Read the e-book ›**

GitHub Advanced Security:
Take the learning path to explore features like secret scanning, code scanning, and dependency management. **Learn more ›**

Azure AI Content Safety:
Take this learning module to learn how to build generative AI guardrails in Foundry. **Learn more ›**

Chapter 6

Performance, observability, and optimization

Creating AI apps and agents successfully depends on making sure they perform consistently, scale effectively, and remain reliable over time. Performance, observability, and optimization are the cornerstones of long-term success. Without them, teams risk slow response times, escalating costs, and blind spots that make it hard to diagnose issues or improve model behavior. With the right visibility and insights, however, organizations can fine-tune their AI systems, unlock efficiencies, and ensure the solutions they deliver are both impactful and sustainable.

Foundry provides this foundation with Foundry Observability by integrating tightly with Application Insights. Together, these tools provide teams with a comprehensive view of their AI ecosystem, encompassing agents, models, and the infrastructure that supports them.

Foundry Observability

Responsible AI relies on visibility and accountability. Foundry Observability provides teams with detailed visibility into how their AI apps and agents perform and behave. It includes specialized evaluators that review outputs for coherence, fluency, factual accuracy, and ethical considerations, helping ensure generative AI systems operate reliably throughout their lifecycle. The AI **red teaming agent** plays a critical role in proactively stress-testing agents for vulnerabilities, bias, and misuse scenarios, simulating complex adversarial attacks against your AI system to identify weaknesses in model responses prior to real-world deployment.

Telemetry and Application Insights

For deeper diagnostics, **Application Insights** brings a more detailed level of observability. It captures telemetry, including traces, exceptions, and user interactions, helping teams understand how AI agents behave once they're deployed. This data can also be used for continuous evaluation of output quality and safety, providing ongoing feedback on model behavior. Teams can run advanced queries, drill down into telemetry, and uncover root causes of performance issues or unexpected results.

A complete view of your AI systems

When combined, these capabilities provide a 360-degree view of your AI solutions. You gain visibility into operational health, user impact, performance efficiency, and financial sustainability, all within one integrated environment. With these insights, teams can iterate more quickly, deploy smarter solutions, and maintain control as their AI solutions evolve.

Learning resources

Foundry Observability: Discover how to bring continuous visibility across your entire AI app lifecycle.
Read the blog ›

Application Insights: See how to enable application performance monitoring (APM) for live web applications.
Learn more ›

Chapter 7

Skilling and cultural readiness for AI adoption

Technology alone doesn't guarantee successful AI adoption. For many organizations, the real challenge isn't the tools themselves but the difficulty of integrating them into existing workflows and ensuring they align with organizational values. Without addressing these skill gaps, even the most advanced AI solutions may struggle to gain traction.

Closing these gaps requires a dual focus: equipping teams with the right technical skills and fostering a culture that embraces change and responsible innovation.

Microsoft Learn, certifications, and change management

Microsoft Learn is filled with resources to help teams build the skills needed for AI transformation. Its guided learning paths cover everything from prompt engineering and agent design to working with Azure OpenAI and other advanced capabilities. With role-specific tracks for AI engineers, data scientists, and solution architects, your teams can progress at a pace and depth that fit their responsibilities.

Pairing technical skilling with structured change management is equally important. Approaches such as stakeholder mapping, pilot programs, and continuous feedback loops ensure that AI solutions aren't only implemented but also fully embraced across the organization.

Responsible AI practices and team enablement

As teams skill up, they must also adopt responsible AI practices to guide how technology is built and applied. Microsoft's Responsible AI framework emphasizes embedding ethical principles into every stage of development. This includes proactive steps such as bias detection and mitigation, maintaining human oversight in decision-making processes, and providing transparent documentation of a model's intended use and limitations.

True team enablement requires cross-functional collaboration among engineering, compliance, legal, and product teams to define acceptable use policies, monitor the real-world impact of AI systems, and establish feedback mechanisms for ongoing improvement. When responsible AI is integrated into culture—not just code—your organization can create an environment where teams innovate with confidence, align with values, and deliver AI solutions that are both powerful and trustworthy.

Azure Essentials and the AI Center of Excellence

Azure Essentials is a comprehensive learning and enablement platform designed to help teams build AI apps and agents with confidence and clarity. It offers structured guidance, hands-on resources, and best practices across the entire AI development lifecycle.

At the heart of Azure Essentials is the **AI Center of Excellence**, a built-in framework that helps you align business goals, technical capabilities, and team skills. By fostering cross-functional collaboration and providing actionable playbooks, the AI CoE empowers enterprises to scale AI responsibly and accelerate innovation across use cases.

Learning resources

AI skills development: Discover AI skills-building resources based on your role. **Start learning ›**

Responsible AI: Find out how to adopt responsible AI practices to secure your innovation strategy. **Take the learning module ›**

Credentials: Take courses and test your knowledge to earn certifications and applied skills. **Browse credentials ›**

Build AI apps and agents: Explore resources designed to ease your learning journey into the world of building AI apps and agents. **Learn more ›**

AI Center of Excellence: Discover how to scale AI across your organization by establishing a robust Center of Excellence (CoE) with expert support from Microsoft. **Read the e-book ›**

GitHub Copilot: Automate development workflows with AI-powered agents, GitHub Actions, and Azure devops for intelligent CI/CD and self-healing systems. **Learn more ›**

Foundry: Build and integrate agents into applications seamlessly and learn advanced model fine-tuning techniques. **Learn more ›**

Models: Find and deploy models, work with benchmarking tools, and create multimodal applications. **Learn more ›**

Trustworthy AI: Design, govern, and observe AI apps and agents with security, safety, and observability capabilities. **Learn more ›**

Business impact: Initiate your organization's AI strategy, assess infrastructure readiness, and review AI use cases. **Learn more ›**

Chapter 8

Customer momentum and use case activation

AI apps and agents are gaining momentum as organizations unlock tangible business value. According to IDC, generative AI has contributed to 10% of revenue growth and 12% of cost savings over the past year.[3] Businesses are adopting a range of AI tools (including agents, AI coding assistants, text generation, and image/video generation) to transform workflows, modernize applications, automate IT operations, and create new solutions. Popular use cases span customer service, code generation, meeting summaries, synthetic data for model training, drug discovery, and design prototyping.

Across industries, organizations are using AI to gain a competitive advantage, accelerate productivity, enhance outcomes, and drive innovation at scale. From AI-powered chat analysis to coding assistants and industrial data automation, real-world implementations showcase how AI can drive measurable results.

Explore the stories below to see how leading companies are using Azure AI, data, and apps platforms to reimagine operations and deliver impact.

[3]IDC Infographic, sponsored by Microsoft and NVIDIA, GenAI Unleashed: Accelerating Growth, Efficiency, and Innovation Across the Enterprise, #US53347425, June 2025.

Customer success: Accelerating AI Innovation with Azure

Azure enables teams to move from modernization to innovation, building AI apps that deliver real-world impact. Here are just a few examples of how organizations are putting these capabilities into action.

Coca-Cola reaches millions with immersive campaign built on Azure

Coca-Cola had an ambitious campaign idea: create a "Create Real Magic" app that replicates Santa in digital 3D and gives him the ability to have real-time, memory-sharing conversations with consumers. Using Azure Speech in Foundry Tools, Coca-Cola was able to assimilate vast amounts of data from global markets into the Santa campaign. Foundry played a pivotal role in managing and deploying AI models at scale, while Azure Speech supported effortless and natural multilingual conversations. Serverless features in Azure Functions, such as triggers and bindings, seamlessly handled Coca Cola's events and data processing across **Azure Cosmos DB**, **Azure Service Bus**, and **Azure Blob Storage** using a preferred format. This approach enhanced productivity with fewer lines of code. The "Create Real Magic" campaign generated impressive results, with more than one million users interacting with Santa across 43 markets in just three weeks.

Impact:

- **60 days** to launch
- **1M+** consumer engagements
- **26 languages** available

Read the full story ›

Carvana builds efficient, seamless customer experiences with agentic AI

Leading online automotive retailer Carvana delivered high-quality customer support with Sebastian, an AI agent. Sebastian has deep insight and action-taking capabilities to help guide customers through every step of the buying and selling journey. To make both Sebastian and Carvana's Customer Advocate team even more efficient and effective in working with customers, Carvana created CARE (Conversation Analysis Review Engine), an AI-powered platform that pulls insight from every customer touchpoint to keep getting better.

CARE uses **Azure Speech** within **Foundry** for quick and accurate transcription, and CARE data is stored in **Azure Cosmos DB** for scalability, security, and uptime. Almost all of the company's systems run on **Microsoft Azure Kubernetes Service** (AKS) to scale and reliably provide services and simplified management. Meanwhile, the development team works with **GitHub Copilot** to automate repetitive tasks and enable faster, more efficient code writing. With GitHub Copilot handling routine coding tasks, engineers can focus on more complex and creative problems, and Interoperability with the Microsoft ecosystem makes workflows seamless.

Impact:

- **45% decrease** in calls per sale in 2 years
- **100% visibility** into customer interactions

Read the full story ›

Hexagon brings AI-powered speed and scale to industrial data workflows

Hexagon's transformation of engineering data management into SDx2 showcases AI-driven scale and speed. Using Foundry, the company now extracts and contextualizes complex engineering documents in under an hour—a process that used to take days—while delivering real-time analytics and actionable insights to industrial clients. Behind the scenes, **Azure Kubernetes Service** (AKS) helps ensure microservices scale on demand, which has helped Hexagon in getting solutions to market faster. To support large-scale customers, Hexagon uses Azure SQL Database Hyperscale tier for storing tenants' engineering data in elastic pools, which enables automatic scaling without manual oversight.

Impact:

- **Document processing** cut from 2–3 days to under 1 hour
- **Improved data quality** and digital twin accuracy
- **Millions** in productivity gains and operational savings

Read the full story ›

OpenAI scales AI innovation with Azure Database for PostgreSQL

OpenAI relies on Azure Database for PostgreSQL Flexible Server to power its mission-critical AI workloads, including ChatGPT. By leveraging PostgreSQL for both read- and write-heavy operations, the company has optimized performance at a global scale, offloading reads to replicas, using connection pooling with PgBouncer, and implementing schema governance to maintain resilience. These strategies ensure low-latency responses for millions of users worldwide. Managed PostgreSQL service provides automatic scaling, high availability, and global consistency, enabling OpenAI to focus on delivering cutting-edge AI experiences without the burden of infrastructure management.

Impact:

- **Global-scale performance**
- **Operational resilience**
- **Accelerated AI innovation**

Read the full story ›

Getting started with Azure AI apps and agents

Organizations that want to harness AI effectively need more than isolated tools. They need an integrated platform that enables secure, scalable, and AI apps from end to end. Azure provides exactly that: a comprehensive ecosystem with integrated tools for AI, applications, integration services, data, and developers. With Azure, technical and developer decision makers can streamline cloud-native development, accelerate developer velocity, and take advantage of advanced AI platforms and agent management.

Azure also delivers the scalability and infrastructure required for enterprise growth, along with the security, safety, and governance needed to meet the highest standards. The result is an environment where organizations can not only integrate AI seamlessly but also manage costs efficiently and optimize devops processes without compromise. Most importantly, Azure gives enterprises the confidence to adopt responsible AI, allowing them to move beyond experimentation and unlock transformative value with trust at its core.

Next steps

Jumpstart your journey by exploring models, features, and templates.

Try Foundry ›

Browse a wide collection of videos all about using GitHub Copilot to ramp up your productivity.

Watch the GitHub Copilot Series ›

For developers: Learn how to develop AI apps and build AI agents.

Designing and implementing a Microsoft Azure AI Solution ›

Accelerate your AI skills with curated training and resources for individuals and organizations, from building AI agents with Foundry to mastering Microsoft Copilot.

AI learning hub ›