Microsoft Foundry

A guide to platform capabilities, deployment options, and cost models

Pricing Guide

Explore Microsoft Foundry features and learn more about cost implications at every stage of the development process

Microsoft Azure

Contents

Introduction

Microsoft is committed to transparent pricing to help organizations make informed decisions as they plan and scale their AI investments. Use this guide to gain a deeper understanding of the capabilities and pricing considerations of Microsoft Foundry, your unified platform for building, optimizing, and governing AI innovation at scale. Explore products and learn how different cost models apply to your workloads. Tools like the Azure pricing calculator can also help you estimate costs before adding resources, allowing for better planning and budgeting.
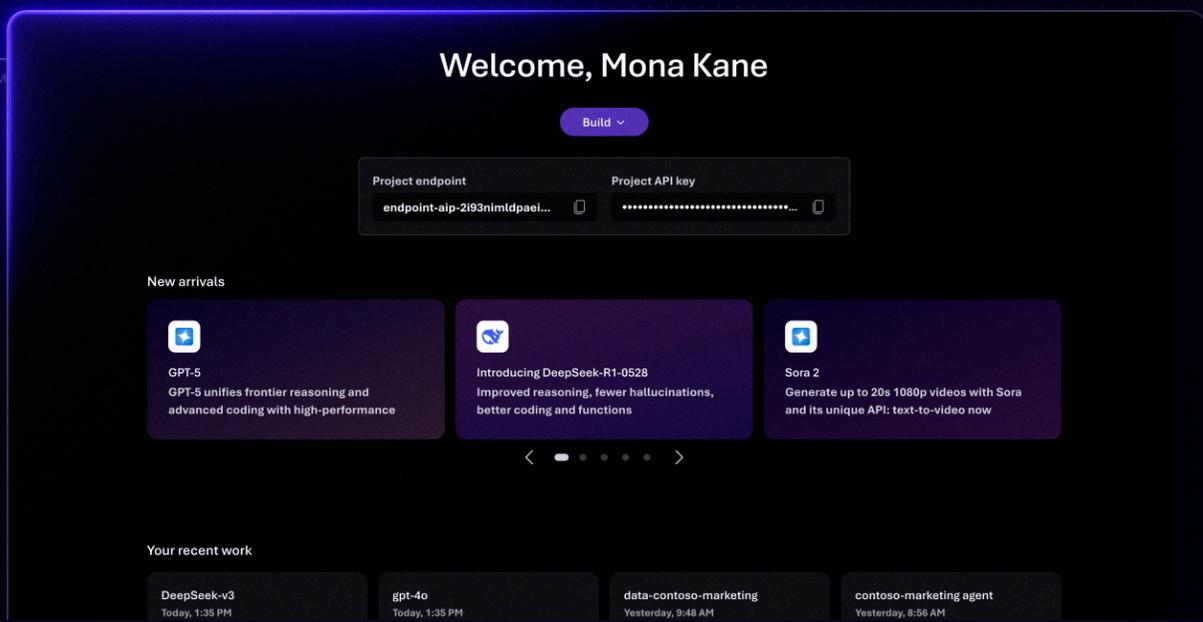
Microsoft determines pricing based on several factors, such as:

- **Deployments:** Costs are calculated based on the specific deployment of each service. For example, running a model through a serverless endpoint incurs pay-as-you-go token charges, while deploying it to a dedicated, real-time or provisioned endpoint adds compute-based costs tied to the underlying infrastructure.
- **Resource usage:** Compute costs are calculated by hourly usage, and storage costs are incurred based on the amount of data stored. This helps ensure that customers only pay for the resources they use.
- **Volume discounts and custom pricing:** Volume discounts and custom pricing options are available to accommodate different customer needs and usage patterns. For example, the Microsoft Agent Pre-Purchase Plan (P3) allows customers to use Microsoft Copilot Studio and Microsoft Foundry through one pool of funds with predictable discounts. Some customers may also have a Microsoft Azure Consumption Commitment (MACC), resulting in discounts for a defined period.

General pricing structure

Foundry offers a flexible and transparent pricing structure designed to accommodate various usage patterns and needs. The general pricing structure includes:

- Standard on-demand pay-as-you-go (PAYGO) pricing
- Provisioned Throughput Unit (PTU) pricing
- Third-party model pricing
- Region-specific pricing

Welcome, Mona Kane

Build ∨

Project endpoint	Project API key
endpoint-aip-2i93nimldpaei...	•••••••••••••••••••...

New arrivals

GPT-5	Introducing DeepSeek-R1-0528	Sora 2
GPT-5 unifies frontier reasoning and advanced coding with high-performance	Improved reasoning, fewer hallucinations, better coding and functions	Generate up to 20s 1080p videos with Sora and its unique API: text-to-video now

‹ • · · · · · ›

Your recent work

DeepSeek-v3	gpt-4o	data-contoso-marketing	contoso-marketing agent
Today, 1:35 PM	Today, 1:35 PM	Yesterday, 9:48 AM	Yesterday, 8:56 AM

Accelerate innovation with a complete, integrated, and interoperable AI platform

As you move from experimentation to scaled AI adoption, having clarity on pricing becomes essential to managing cost, optimizing performance, and choosing the right model for each workload. In this chapter, let's explore how Foundry's model catalog, deployment options, and tooling influence total cost.

Foundry Agent Service

Foundry Agent Service enables professional developers to streamline development using seamless hosting and integration of the most popular open-source agent frameworks. Agent Service provides the ability to build, deploy, and scale enterprise-grade AI agents that automate complex business processes. There is no additional charge to use Foundry Agent Service; however, you will incur charges for model token consumption through Foundry Models and separate charges and licenses for tools and knowledge connections, including Azure Logic Apps connectors, Microsoft Fabric, Microsoft SharePoint, Grounding with Bing Search, Foundry IQ (by Azure AI Search), and your own licensed data. Foundry Agent Service capabilities, including hosted agents, multi-agent workflows, and memory, are in preview and are currently free of charge. Pricing details will be shared in the coming months.

Foundry Models

Foundry Models equips developers with one of the widest selections of AI models. Developers can choose from over 11,000 foundational, open, reasoning, multimodal, and industry-specific models, including models from OpenAI and Anthropic. They can compare models using benchmarking features to assess differences in performance and other key parameters. In this section, let's take a look at some of the pricing considerations you should consider as you leverage different tools in Foundry, including a robust model catalog, which has both managed compute and serverless API offerings in addition to Azure direct models, like OpenAI, Anthropic, Mistral, and more. We'll also take a look at Microsoft's Phi family of models, model benchmarks, model routing, and a uniform inference API that eases model swapping. Understanding the pricing implications as you use these tools to begin your next AI development project can help you make informed decisions about model choice, saving costs long term.

Pricing structure

Standard (PAYGO)

Inference costs: Charges are based on the number of tokens processed during fine-tuning or inferencing. For example, the Phi-3 models have specific costs per 1,000 tokens for both input and output.

Provisioned Throughput Unit (PTU)

Some models might be available through subscription plans that can offer more predictable costs and potentially lower rates for high-volume usage.

Third-party models featured in the model catalog

Models from third-party providers are billed through the Azure Marketplace. Their pricing can vary based on the provider's terms.

Key considerations

Models from different providers may be priced differently due to several factors, including the complexity of the models, the computational resources required, and the specific terms set by the providers. These variations ensure that customers have access to a diverse range of models that cater to different needs and budgets. When evaluating different AI models in the model catalog, consider some of the following factors that contribute to overall cost.

Model family

Different models have different pricing based on their capabilities and computational requirements. For instance, smaller models like Phi-4 are more cost-effective for applications with lower computational needs.

Usage patterns

Understanding your usage patterns can help you choose the most cost-effective pricing model. PAYGO is ideal for variable usage, while subscription plans might be better for consistent, high-volume usage.

Region-specific pricing

Prices can vary by region, so it's important to check the costs for the specific region where you plan to deploy the models.

Integration and scalability

Consider the ease of integrating the models into your existing systems and the scalability options available. Foundry offers serverless endpoints, provisioned endpoints, and managed instances for secure and simple deployment.

Model benchmarks in Foundry provide a comprehensive evaluation of models across various metrics, including quality, performance, and cost. By analyzing these benchmarks, you can gain insights into the cost implications of different models, helping you make informed decisions about which models to use based on their cost-effectiveness.

Insights into hosted and serverless models

Microsoft Foundry provides multiple model offerings catering to different needs:

- **Azure direct models:** This approach gives developers maximum control to build, customize, and manage models using Microsoft Foundry's native tools and capabilities. Direct models support both serverless endpoints and managed compute, making them ideal for organizations that need deeper customization or specialized deployment configurations.
- **Serverless models:** Serverless options – available across direct, partner, and community models – provide pre-built, ready-to-use endpoints that require no infrastructure management. This is ideal for organizations that want rapid integration and low operational overhead without extensive customization.

Which is more expensive?

Generally, direct models can be more expensive due to the need for dedicated infrastructure, customization, and ongoing management. However, it offers greater flexibility and control that can be valuable for organizations with specific needs.

Serverless, on the other hand, can be more cost-effective for organizations looking for quick deployment and ease of use. The pay-as-you-go and subscription models make it easier to predict and manage costs, especially for high-volume usage.

When considering from a pricing perspective, think about your organization's specific needs, usage patterns, and budget. If you require highly customized models and have the resources to manage them, direct might be the better option despite the higher initial costs. However, if you need quick deployment, ease of use, and predictable costs, serverless could be the more cost-effective choice.

Real-time endpoints

In addition to these deployment models, you may also need real-time inference capabilities for certain workloads. Real-time endpoints are essential for deploying AI models that require real-time inference. The pricing for real-time endpoints is based on the compute resources used for running the models. This includes costs for compute instances, storage, and any additional services required for real-time inference. This pricing model allows you to scale your real-time endpoints based on your usage patterns and requirements.

Learn more about pricing for real-time endpoints.

Azure OpenAI in Foundry Models

Azure OpenAI in Foundry Models is a first-party model offering provided by Microsoft that gives developers access to powerful AI models developed by OpenAI with responsible guardrails in place to govern data and robust security measures that are enabled by default. It is part of Foundry Tools and includes models like GPT-4.1, GPT-5.2, GPT-5.1, GPT-5, gpt-oss, Codex mini, o-series, computer-use-preview, and others.

Pricing model

Azure OpenAI offers a flexible and transparent pricing model that caters to various organizational needs, making it an attractive choice for organizations that want to take advantage of the service's enterprise-ready generative AI, built-in data privacy, flexible deployment options, and seamless integration with the Azure ecosystem. Below are several pricing and cost management offers that Azure OpenAI provides:

- **Standard (on-demand):** This pay-as-you-go offer charges for input and output tokens, making it ideal for organizations with variable usage patterns. It provides flexibility to scale up and down on demand and ease of management to control costs based on actual usage. Standard is optimized for low-to-medium-volume workloads.
- **Priority Processing:** This latest addition to Azure OpenAI provides consistent, low-latency performance for a real-time, user-facing application where response speed directly impacts engagement, with pay-as-you-go flexibility and SLA-backed reliability and no commitment required. It's best for latency-sensitive scenarios such as healthcare

transcription, financial services, interactive assistants, and any workload where every millisecond matters.

- **Provisioned (PTUs):** PTUs enable you to allocate throughput with predictable costs, offering monthly and annual reservations to reduce overall spend. It is suitable for organizations with consistent, high-volume usage. You can save on your PTU cost with provisioned reservations. You can commit to paying for a fixed number of PTUs monthly or yearly to receive a discount. Reservations are most beneficial when you have consistent usage for a specific number of PTUs.
- **Batch:** Batch is designed to handle large-scale and high-volume processing tasks efficiently. Batch asynchronously handles requests with separate quotas. Batch also provides cost-effective solutions for large-scale deployments with less stringent time requirements. For example, a global batch deployment provides less than 24-hour turnaround time at half the cost of the global standard.

To learn more about new deployment and cost management solutions for Azure OpenAI, watch this video.

Deployment types

Azure OpenAI also offers various deployment types for Standard, Batch, and Provisioned offers, enabling greater flexibility and control of pricing and performance:

- **Global deployment – Global distributions:** Suitable for organizations with global operations, providing consistent performance and pricing across multiple regions. Great for services that need to be available globally with low latency and where cost savings is a priority.
- **Data Zones deployment – Geographic-based (EU and US):** Data Zones deployments load balance across regions within a geographic boundary (EU and US). Ideal for organizations with specific geographic data processing requirements, ensuring compliance with regional regulations.
- **Regional deployment – Local region (up to 28 regions):** Offers localized performance and pricing, optimizing costs and performance for specific regions. Best suited for applications required to meet data residency compliance with low latency. Regional deployments are useful for applications requiring localized data processing and storage.

Learn more about deployment and cost management solutions for Azure OpenAI Service.

Explore pricing details for Azure OpenAI Service.

Save costs with Microsoft Azure OpenAI Service Provisioned Reservations.
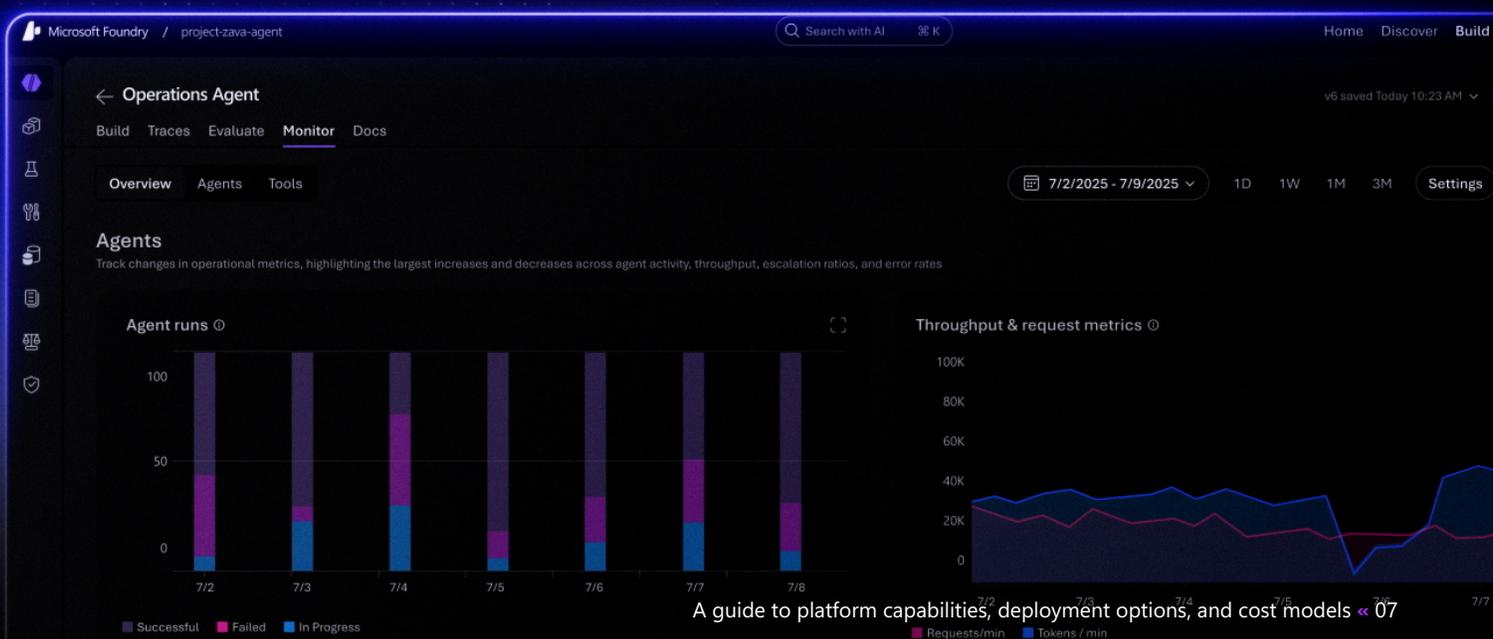
Phi open model family

Phi, Microsoft's latest Small Language Model (SLM), offers efficient performance for commercial and research tasks. It supports various functions with low latency, reduced costs, and offline use, ensuring privacy. Phi excels in mathematical reasoning, code generation, advanced reasoning, summarization, long document QA, and information retrieval. Customizable and deployable, it integrates smoothly into existing systems with multilingual support and is ideal for real-time applications like chatbots and virtual assistants. Using high-quality training data and safety measures, Phi open model family ensures accurate, reliable outputs adaptable to diverse business needs.

Pricing model

Phi models are available with pay-as-you-go billing via inference APIs in Foundry Models. The pricing varies based on the model and context length, and charges are based on the number of tokens processed during inference.

Deployment types

Phi models can be deployed in two different ways to allow users flexibility and ease as you integrate a new model into your AI ecosystem:

- **Serverless API endpoints:** Phi models can be deployed to serverless API endpoints with pay-as-you-go billing. This allows you to consume models as an API without hosting them on your subscription, while maintaining enterprise security and compliance.
- **Self-hosted managed compute:** For more control, Phi models can also be deployed to a self-hosted managed inference solution, which allows you to customize all the details about how the model is served. This requires enough quota in your subscription to employ.
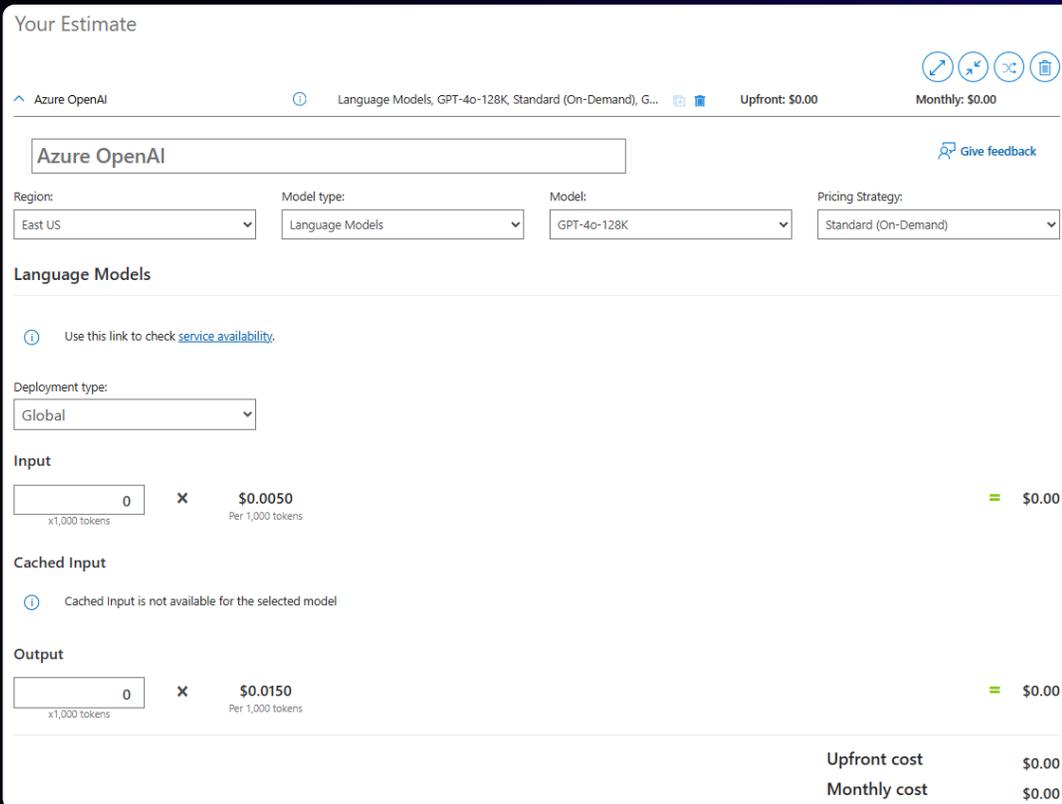
Estimate costs before using Azure AI services

We recommend that you use the Azure pricing calculator to estimate costs before you deploy Models.
To do that, follow the steps below:

1. Select a product such as Azure OpenAI Service in the Azure pricing calculator.
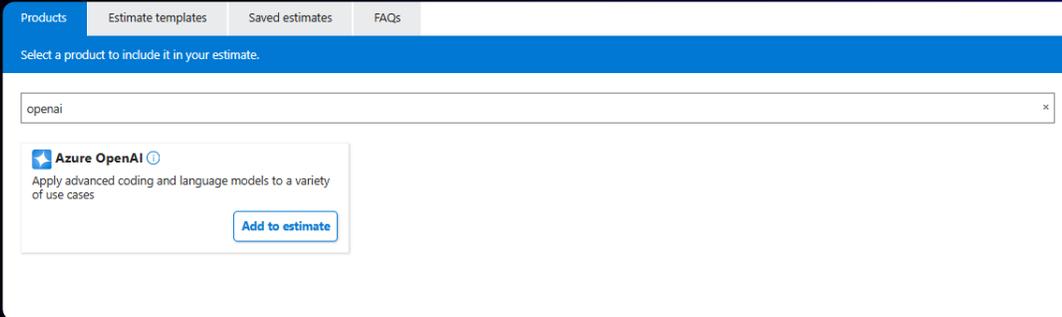
2. Enter the number of units you plan to use. For example, enter
the number of tokens for prompts and completions.

3. You can also enter the number of PTUs you plan to use and compare cost with PAYGO vs. reservations.

4. You can select more than one product to estimate costs for multiple products.
   For example, select Virtual Machines to add potential costs for compute resources.

As you add new resources to your project, return to this calculator and add the newly added resource to update your cost estimates. Note that when you create resources for a hub, resources for other Azure services are also created. That means some costs can accrue with Foundry.

By understanding these pricing considerations, you can confidently use the tools available in Foundry to aid in selecting and deploying the model that's right for your specific use case. Determine what the most important components of your AI model are and assess how those parameters and requirements impact overall cost.

Drive business results with action-oriented and context-aware agents

Microsoft Foundry enables organizations to drive business results by building action-oriented, context-aware agents that automate processes, integrate with enterprise data, and deliver measurable impact. The platform's comprehensive toolkit – including retrieval systems, agent services, model customization, prompt engineering, and evaluation tools – empowers teams to rapidly develop, deploy, and optimize agents tailored to their unique business needs.

When planning for costs, it's important to understand how each tool and service contributes to agent capabilities and business outcomes. Pricing reflects not only model usage and customization, but also the integration of agents with enterprise data sources, automation workflows, and continuous evaluation for quality and safety. The following sections outline the pricing implications for key Foundry features that support agent development and deployment:

- Foundry IQ (by Azure AI Search): Empower agents to understand business context.
- Foundry Tools: Enable AI to act on your business systems using pre-built and custom MCP tools.
- Microsoft Agent Factory: Streamline AI integration with your everyday productivity tools.
- Model, app, and agent customization: Train and upgrade models with ease.

By understanding these pricing considerations, organizations can make informed decisions that maximize the value and impact of their AI agents while optimizing costs.

Foundry IQ by Azure AI Search

Foundry IQ is a unified knowledge layer for agents designed to improve response performance and agent ROI with the next generation of RAG. It is an agent's single endpoint for knowledge, delivering better context with automated source routing and advanced agentic retrieval, all while respecting user and document permissions.

Customers must have an Azure AI Search resource to use Foundry IQ knowledge bases. Knowledge bases incur an additional charge from service and tier pricing, based on retrieval reasoning effort settings and tokens used when using agentic retrieval.

Learn more about agentic retrieval in Azure AI Search.

Foundry Tools

Foundry Tools bring together Microsoft's first-party AI capabilities to help developers automate workflows and build intelligent applications quickly and responsibly. When building AI applications, explore the following Foundry Tools.

Service	Description
Speech	Provide speech to text, text to speech, translation, and speaker recognition.
Translator	Use AI-powered translation technology to translate more than 100 in-use, at-risk, and endangered languages and dialects.
Language	Build apps with industry-leading natural language understanding capabilities.
Content Understanding	Analyze and comprehend various media content types.
Document Intelligence	Turn documents into intelligent data-driven solutions.
Vision	Analyze content in images and videos.
Azure AI Search	Bring AI-powered cloud search to your mobile and web apps.
Content Safety	An AI service that detects unwanted content.
Custom Vision	Customize image recognition for your business.
Immersive Reader	Help users read and comprehend text.

To extend these workflows, Foundry also provides connectors through the Model Context Protocol (MCP), giving you access to over 1,400 business systems, including SAP, Salesforce, and Dynamics 365. Microsoft-built connectors are included at no additional charge within Foundry. If you choose to connect to a third-party system, the connector itself is free to use, but you will be billed separately by the third-party provider for any underlying service or API usage required by that system.

Learn more about extending agents with MCP.

Learn more about Foundry Tools.

Model, app, and agent customization

Train, fine-tune, distill, and automatically upgrade models with minimal coding to optimize performance and costs. Fine-tuning in Foundry involves customizing pre-trained models to better suit specific use cases. The pricing for fine-tuning varies based on the model and the number of tokens processed during the fine-tuning process. For Model as a Service (MaaS), fine-tuning costs are also based on the number of tokens processed and the hosting of the fine-tuned model.

Explore pricing details for fine-tuning for Model as a Service.

Explore prompt flow in the Microsoft Foundry portal.

Learn about observability in generative AI.

Explore the Microsoft Foundry pricing page.

Govern the AI lifecycle with organization-wide observability and controls

Foundry Control Plane is comprised of Content Safety, Observability, and security offerings to help teams evaluate, monitor, and optimize AI apps and agents in real time.

Content Safety in Foundry Control Plane

Simulate, detect, and mitigate harmful content and security threats at runtime, including prompt attacks, hallucinations, and bias, while seeing real-time alerts and recommendations from Microsoft Security. Content Safety in Foundry Control Plane supports real-time detection and mitigation of harmful content across agents and apps. This service is a critical component for monitoring and managing content safety in AI applications, and it uses advanced language and vision models to detect offensive or inappropriate content in text and images. The pricing for Content Safety includes charges for using Text and Image APIs and the Studio experience. While this service can incur additional costs, developers can ensure that content remains safe and appropriate by applying custom content filters tailored to your requirements for enhanced safety and reassurance.

Learn more about Content Safety pricing.

Safety system messages

Microsoft Foundry provides system message templates within the chat playground at no cost to support effective prompt engineering. These templates provide explicit instructions to a generative AI model that can help mitigate risks and guide the model toward more trustworthy interactions with users.

Learn more about safety system messages.

Integrations with Microsoft Security

To provide enterprise-grade security and governance, Foundry integrates with Microsoft Security tools and services such as Azure Policy, Entra ID, Azure Key Vault, App Gateway, Microsoft Defender, and Azure Backup. These will incur separate charges.

Learn more about the Foundry security baseline.
Learn more about Microsoft Security for AI.

Observability in Foundry Control Plane

Monitoring and observability are critical for tracking model and application performance in production. Azure Monitor Application Insights provides advanced application performance monitoring, including tracking token usage, the quality of generated outputs, and other operational metrics. The compute costs for monitoring in the Foundry portal are calculated based on hourly usage, and the logged inference data is stored in Azure Blob Storage, which incurs storage costs. Azure Monitor logs are billed through the Log Analytics workspace, with pricing options based on data ingestion and retention. This pricing model ensures that you only pay for the resources you use, making it cost-effective for monitoring and observability.

Learn more about Observability in Foundry Control Plane.

Learn more about Azure Monitor logs pricing options.

Manual evaluation

Manual evaluations in Foundry portal enable users to score a small set of generated outputs using human feedback and preferences. This feedback can support rapid iteration on big and small changes to various application components, such as the base model, model parameters, content filters, and system message. Users are charged for model inferencing (i.e., generating the outputs for manual evaluation).

Learn more about manual evaluation.

Automated evaluation

Automated evaluations enable developers to systematically assess the quality and safety of model and application outputs at scale, supporting data-driven decisions around model selection, application design, and risk mitigation. Automated evaluations are accessible via the Foundry SDK and portal, where users can choose to run an evaluation ad hoc or schedule online evaluations as part of their continuous monitoring strategy. The costs associated with automated evaluations depend on the evaluation metrics (built-in metrics or custom metrics) and evaluator type used as shown in the following table.

	Risk and safety evaluations	Performance and quality evaluations		Custom evaluations	
Metrics	Violence, hateful and unfair, sexual, self-harm, direct prompt injection attacks, indirect prompt injection attacks, and protected material	Groundedness, retrieval, relevance, coherence, fluency, and similarity	F1 Score, BLEU, GLEU, METEOR, ROUGE	Custom (prompt-based)	Custom (code-based)
Evaluator type	AI-assisted (using LLM as a judge)	AI-assisted (using LLM as a judge)	Natural language processing (NLP)	AI-assisted (using LLM as a judge)	Natural language processing (NLP) or custom code
Model used to perform evaluations	Microsoft-hosted GPT model (content filters turned off)	Customer's Azure OpenAI GPT deployment	N/A	Customer's Azure OpenAI GPT deployment	N/A
Pricing	• $20/1M input tokens • $60/1M output tokens	Azure OpenAI Service pricing	AzureML compute pricing	Azure OpenAI Service pricing	AzureML compute pricing

Learn more about automated evaluations in Foundry.
Explore evaluation and monitoring metrics for generative AI.
Explore pricing details for automated evaluations.

CI/CD with GitHub Actions

GitHub Actions allows you to discover, create, and share actions to perform any job you'd like, including CI/CD, and combine actions in a completely customized workflow. The pricing for GitHub Actions is based on usage, with free and paid plans available. The free plan includes a certain amount of free minutes and storage, while the paid plans offer additional minutes and storage at a cost. This flexible pricing model allows you to scale your CI/CD workflows based on your needs.

Learn more about GitHub Actions pricing.

CI/CD with Azure DevOps

Azure DevOps supports a collaborative culture and set of processes that bring together developers, project managers, and contributors to develop software. The pricing for Azure DevOps is based on the number of users and the services used. Azure DevOps offers a standard tier with basic features, and paid plans that provide additional features and capabilities. This pricing model allows organizations to choose the most suitable plan for their needs and adjust it as their business requirements change.

Learn more about Azure DevOps pricing.

Start planning and managing costs for Foundry

From designing with leading AI models to customizing with a comprehensive agent toolchain, safeguarding with trustworthy AI, and managing AI performance in production, Foundry equips developers and IT administrators with the tools they need to succeed. As a trusted, integrated platform, Foundry offers a rich set of AI capabilities and tools through a simple portal, unified SDK, and APIs, facilitating secure data integration, model customization, and enterprise-grade governance to accelerate the path to production.

Key considerations for budgeting

When selecting tools and services for your Foundry project, consider the following factors to balance meeting business requirements with optimizing costs:

- **Optimize usage patterns:** Analyze your usage patterns to choose the most cost-effective pricing model. For example, if your usage is variable, the pay-as-you-go (PAYGO) model might be more suitable. For consistent, high-volume usage, consider subscription plans or provisioned throughput units (PTUs) to reduce overall costs. Save even more on PTUs with Azure OpenAI Service provisioned reservations.
- **Leverage built-in tools:** Utilize built-in tools like the Azure pricing calculator to estimate costs before adding resources. This helps with better planning and budgeting, ensuring that you only pay for what you need.
- **Evaluate model performance:** Use model benchmarks and automated evaluations to assess the performance and cost-effectiveness of different models. This allows you to choose models that offer the best balance of performance and cost, optimizing your budget while meeting your business requirements.
- **Consider regional pricing:** Prices can vary by region, so it's important to check the costs for the specific region where you plan to deploy the models. This can help you optimize costs by selecting regions with more favorable pricing.
- **Integrate efficiently:** Consider the ease of integrating the models into your existing systems and the scalability options available. Efficient integration can reduce development time and costs, while scalable solutions ensure that you can handle increased workloads without significant cost increases.
- **Utilize pre-built patterns and templates:** Take advantage of pre-built patterns, app templates, and prompt samples to accelerate development and reduce costs. These resources provide ready-to-use solutions that can be customized to meet your specific needs, saving time and effort.
- **Monitor and optimize:** Continuously monitor the performance and costs of your AI applications using tracing and debugging tools. Identifying performance bottlenecks and unexpected errors early can help you optimize resource usage and reduce costs.

FINANCIAL PERFORMANCE DASHBOARD

RETURN ON ASSETS | WORKING CAPITAL RATIO | TOTAL ASSETS

16% | 3.4:1

RETURN ON EQUITY | DEBT EQUITY RATIO

31% | 0.4:1

To learn more about Foundry pricing and how it can benefit your organization, visit the Foundry pricing page.

Accelerate innovation today

Learn more

Estimate costs

Microsoft Azure