Microsoft Azure

Azure AI Foundry pricing guide

Plan and manage costs for Azure AI Foundry at every stage of the development process

Contents

03
Introduction

04
Pricing considerations: Design with leading AI models

11
Pricing considerations: Customize with access to a comprehensive agent toolchain

14
Pricing considerations: Safeguard with Trustworthy AI

16
Pricing considerations: Manage AI performance in production

18
Start planning and manage costs for Azure AI Foundry

Introduction

This guide is designed to provide Developers and IT Administrators with a clear and engaging overview of the pricing structure for Azure AI Foundry, a trusted and integrated platform for designing, customizing, and managing AI applications and agents.

Azure AI Foundry offers a rich set of AI capabilities and tools accessible through a simple portal, unified SDK, and APIs. These capabilities are seamlessly integrated into developer workspaces like GitHub, Visual Studio, Copilot Studio, and the Azure AI Foundry SDK facilitating secure data integration, model customization, app orchestration, evaluation, and experimentation with enterprise-grade observability, governance, and management in production.

At Microsoft, we prioritize transparent pricing to ensure our customers can make informed decisions. This guide will walk you through the pricing pages, helping you navigate detailed pricing tables and resources. Our pricing models are designed to be scalable and flexible, allowing customers to adjust their usage and costs based on their specific needs. This adaptability is crucial for managing budgets effectively.

Microsoft determines pricing based on several factors:

- **Service-specific pricing:** Different services have distinct pricing models.
- **Deployment level:** Costs are calculated based on the specific deployment and usage of each service.
- **Resource usage:** Compute costs are calculated by hourly usage, and storage costs are incurred based on the amount of data stored. This ensures that customers only pay for the resources they use.
- **Cost management tools:** Tools like the Azure pricing calculator help customers estimate costs before adding resources, allowing for better planning and budgeting.
- **Volume discounts and custom pricing:** Volume discounts and custom pricing options are available to accommodate different customer needs and usage patterns.

General pricing structure

Azure AI 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)**
- **Provisioned Throughput Unit (PTU)**
- **Third-party Model Pricing**
- **Region-specific Pricing**

Keep in mind, pricing can vary by region, so it's important to check the costs for the specific region where you plan to deploy your models and AI solutions.

By the end of this guide, you will have a thorough understanding of Azure AI Foundry's pricing structure, enabling you to make well-informed decisions and optimize your AI projects effectively. Let's dive in and explore the transparent, scalable, and flexible pricing models that Azure AI Foundry has to offer.

Pricing considerations

Design with leading AI models

Azure AI Foundry equips developers with leading AI models for their use case. Developers can choose from the latest cutting-edge foundational, open, task, and industry-specific models. 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 Azure AI Foundry, including the **Model catalog**, which has both Managed compute and Serverless APIs (Model as a Service [MaaS]) offerings in addition to Azure OpenAI Service, and Microsoft's Phi family of models and Azure AI Services' task-based models, **Model benchmarks**, and the **Model Inference API**. 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

Pay-As-You-Go (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, and 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-3-mini 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. Azure AI Foundry offers serverless endpoints, provisioned endpoints, and managed instances for secure and simple deployment.

Model benchmarks in Azure AI 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 MaaP vs. MaaS

Azure AI Foundry provides Model as a Platform (MaaP) and Model as a Service (MaaS) offerings, each catering to different needs.

- **Model as a Platform (MaaP):** This approach allows developers to build, customize, and manage their own models using the tools and capabilities provided by Azure AI Foundry. MaaP is ideal for organizations that require a high degree of control and customization over their AI models.
- **Model as a Service (MaaS):** In contrast, MaaS offers pre-built models that can be easily integrated into applications with minimal customization. This approach is suitable for organizations looking for quick deployment and ease of use without the need for extensive customization.

Which is more expensive?

Generally, MaaP 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.

MaaS, 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 MaaP and MaaS 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, MaaP might be the better option despite the higher initial costs. However, if you need quick deployment, ease of use, and predictable costs, MaaS could be the more cost-effective choice.

Azure OpenAI Service

Azure OpenAI Service is a first-party service 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 the Azure AI services and includes models like GPT-3.5, GPT-4o, o1, Codex, DALL-E 2, and more.

Pricing model

Azure OpenAI Service offers a flexible and transparent pricing model that caters to various organizational needs, making it an attractive choice for some 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 Service 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 upon demand and ease of management to control costs based on actual usage. Standard is optimized for low-to-medium-volume workloads.
- **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 a 24-hour turnaround time at half the cost of the global standard.

- **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.

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

Deployment types

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

- **Global deployment – Global SKU:** Suitable for organizations with global operations, providing consistent performance and pricing across multiple regions. Great for services needing to be available globally with low latency and where cost savings is a priority.
- **Data Zone deployment – Geographic-based (EU or US):** Ideal for organizations with specific geographic data processing requirements, ensuring compliance with regional regulations. Data Zone deployments load balance cross region within a geographic boundary (EU or US).
- **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.

Microsoft Phi

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 multi-lingual support and is ideal for real-time applications like chatbots and virtual assistants. Using high-quality training data and safety measures, Phi 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 the Azure AI model catalog. 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.

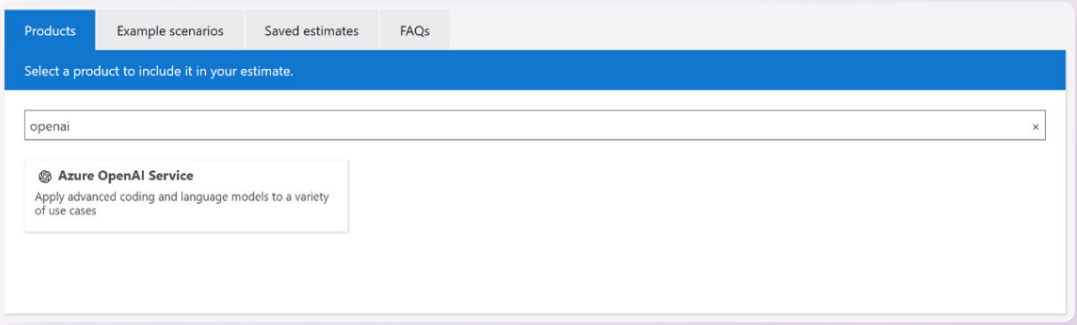
Azure AI service

Azure AI services offer a suite of AI tools, including Speech, Translation, Content Understanding, Language, Document Intelligence, and Vision. These services are designed for tasks like natural language processing, computer vision, speech recognition, multimodal data analysis, machine learning, and conversational AI.

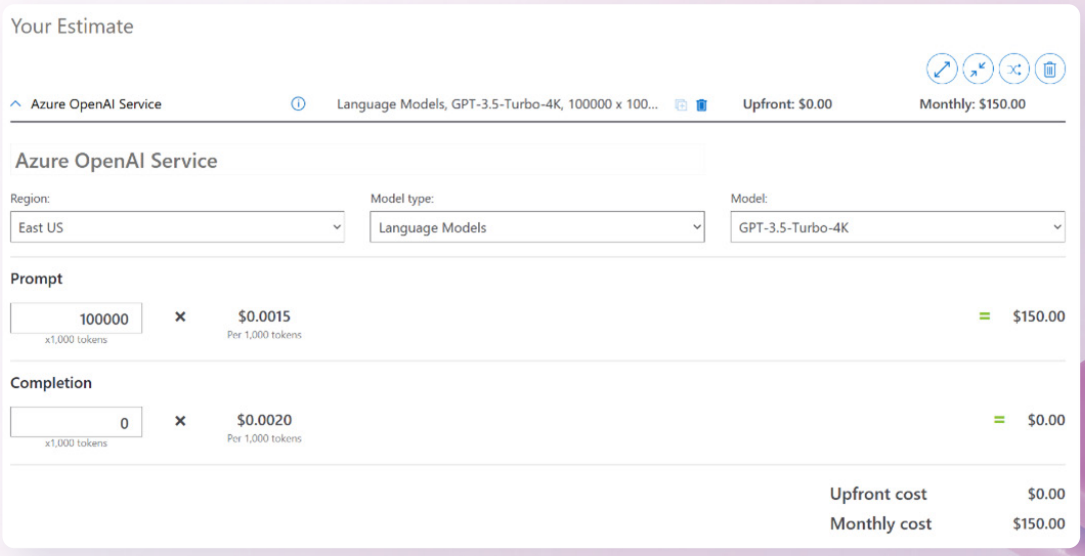
Estimate costs before using Azure AI services

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

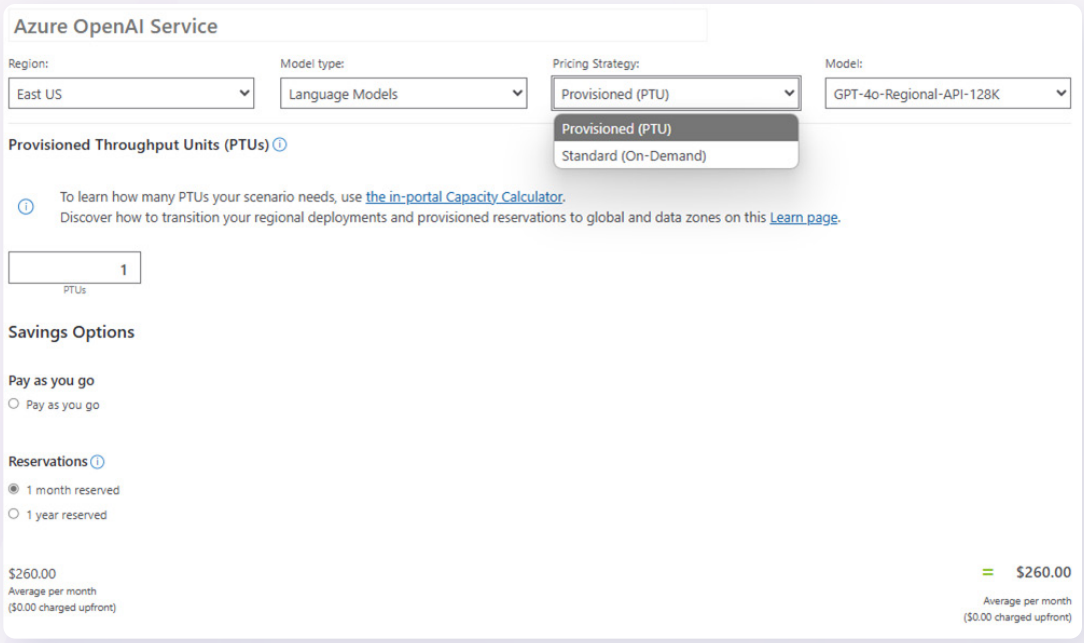
1. Select a product such as Azure OpenAI Service in the Azure pricing calculator.



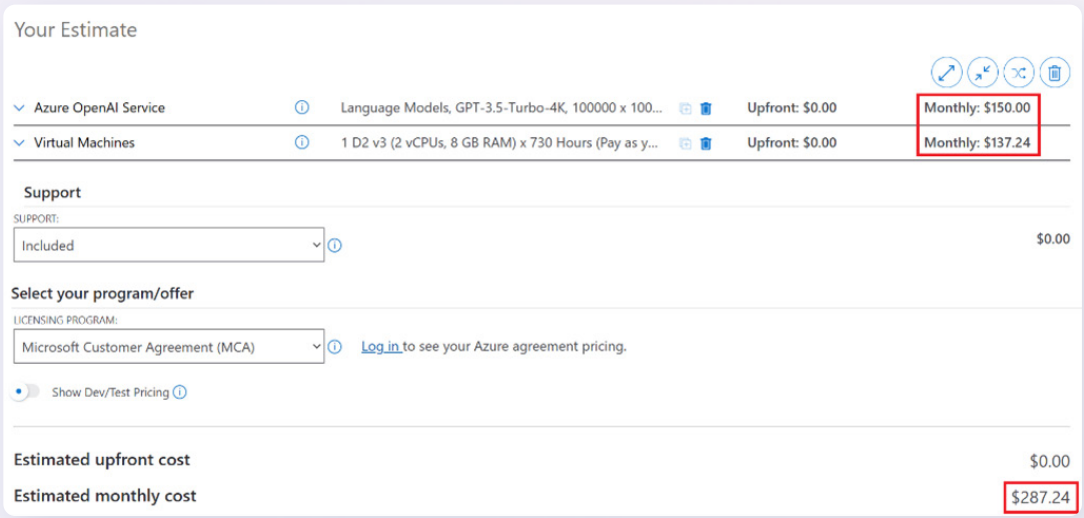
2. Enter the number of units you plan to use. For example, enter the number of tokens for prompts and completions.



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 Azure AI Foundry.

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

Pricing considerations

Customize with access to a comprehensive agent toolchain

Azure AI Foundry offers numerous models and tools to help you evaluate and choose the best fit for your requirements. Additionally, it features a developer toolkit designed to improve your applications and accelerate AI development. This toolkit includes pre-built patterns and core GenAI functionality, developer tools, model customization options, orchestration and evaluation tools, and experimentation environments.

Tools and services you may consider leveraging include **Azure AI Search**, **Azure AI Agent Service**, **Model Customization**, **Prompt Flow**, **OSS Frameworks**, **Evaluations**, **Tracing & Debugging**, **Experimentation**, **GenAIOps**, and **AI Application Templates**, which contribute to the developer experience in Azure AI Foundry. AI Application Templates are distributed across services as these templates use a variety of different services, including 3P services like Pinecone.

In this section, we will explore some of the pricing implications for these various AI tools and services, and how they can be leveraged to differentiate your apps and accelerate development.

Azure AI Search

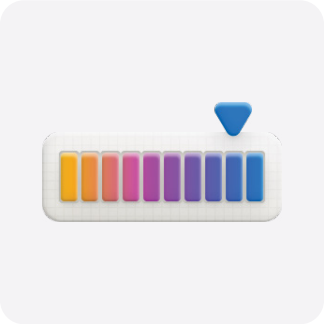
Azure AI Search is a retrieval system built to support GenAI applications at any scale, offering RAG-tuned technology like hybrid search, query rewriting, and reranking. The pricing for Azure AI Search is based on the pricing tier selected, which determines an hourly rate applied to the number of search units allocated to the service. Billing is based on capacity search units (SUs) and the costs of running premium features such as AI enrichment, semantic ranker, and private endpoints.

Learn more about Azure AI Search and choosing a tier.

Learn more about Azure AI Search pricing.

Azure AI Agent Service

Azure AI Agent Service enables professional developers to build, deploy, and scale enterprise-grade AI agents that automate complex business processes. The pricing for Azure AI Agent Service includes charges for consuming model tokens through the Azure OpenAI Service and model catalog. Additionally, you will incur separate charges and licenses for knowledge connections, including Microsoft Fabric, Microsoft SharePoint, Grounding with Bing Search, Azure AI Search, and your own licensed data. Charges for automation services like Azure Logic Apps and Azure Functions also apply.



Fine-tuning

Fine-tuning in Azure AI 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.

Prompt flow

Prompt flow is a tool available in both Azure AI Foundry portal and SDK, designed to streamline the entire development cycle of AI applications. In addition to compute, you will incur separate charges for other Azure services consumed, including but not limited to Azure OpenAI Services, Azure AI Services, Azure Blob Storage, Azure Key Vault, Azure Container Registry, and Azure Application Insights.

Learn more about prompt flow.

Manual evaluation

Manual evaluations in Azure AI 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 Azure AI SDK and Azure AI Foundry 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 table below.

	Risk and safety evaluations	Performance and quality evaluations		Custom evaluations	
Metrics	Harmful content, jailbreaks, protected material, etc. View all metrics	Groundedness, coherence, relevance, etc View all metrics	F1 score, BLEU, GLEU, etc. View all metrics	Custom (prompt-based) View example	Custom (code-based) View all metrics
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 Azure AI Foundry.

Explore evaluation and monitoring metrics for generative AI.

Explore pricing details for automated evaluations.

Pricing considerations:

Safeguard with Trustworthy AI

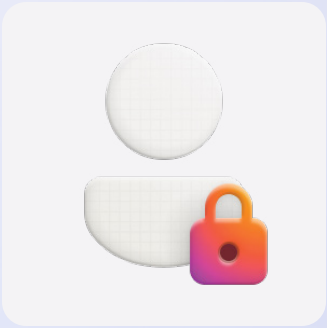
Azure AI Foundry is built for running AI applications in regulated environments, which is why we take security and data privacy very seriously. While we prioritize simplicity in developer experiences, we never compromise on enterprise requirements. Security, compliance, and operational excellence form our foundation.

Azure AI Foundry provides enterprise security, privacy, and safety features by default, putting our Trustworthy AI commitments into practice with real-world capabilities. These capabilities help you design apps responsibly using industry-leading tools and controls, such as prompt shields to detect and block prompt injection attacks on your application, groundedness detection, and correction to address hallucinations in real time, and PII detection and masking to remove personal identifiable information (PII) from inputs or outputs. We also equip organizations with system message templates to help guide models' behavior toward more trustworthy outputs. Finally, Azure AI Foundry is built to integrate with the best of Azure and Microsoft Security, such as Azure Policy, Entra ID, and Microsoft Defender, to protect and govern AI workloads from code to cloud.

Azure AI Content Safety pricing

One additional service that is available through Azure AI Foundry is **Azure AI Content Safety**. 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 Azure AI 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, applying custom content filters tailored to your requirements for enhanced safety and reassurance.

Learn more about Azure AI Content Safety pricing.



Safety system messages

Azure AI 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, Azure AI 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 Azure AI Foundry security baseline.
Learn more about Microsoft Security for AI.

Pricing considerations:

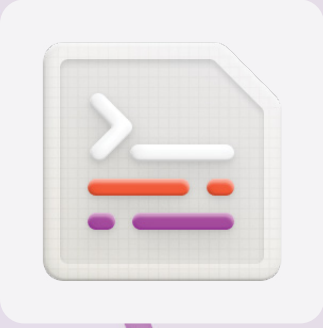
Manage AI performance in production

What happens in production once you've deployed your AI model? In this section, we will break down additional pricing considerations for some of the key tools and services available through Azure AI Foundry. These resources equip developers with what they need to deploy AI applications, including continuous monitoring and governance across environments. Understanding the pricing structure for these tools will help you make informed decisions and optimize costs at this stage of the development process.

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.



Azure DevOps Integrations

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.

Real-time endpoints

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.

Monitoring and observability

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 Azure AI 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 Azure Monitor logs pricing options.

Start planning and manage costs for Azure AI Foundry

From designing with leading AI models to customizing with a comprehensive agent toolchain, safeguarding with trustworthy AI, and managing AI performance in production, Azure AI Foundry equips developers and IT administrators with the tools they need to succeed. As a trusted, integrated platform, Azure AI 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 Azure AI 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
16%
3.4:1
RETURN ON EQUITY
DEBT-EQUITY RATIO
31%
0.4:1

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

Accelerate innovation today

Learn more

Estimate costs

Microsoft Azure