

Azure AI Foundry

The AI app and agent factory

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 the best models for your use case

11

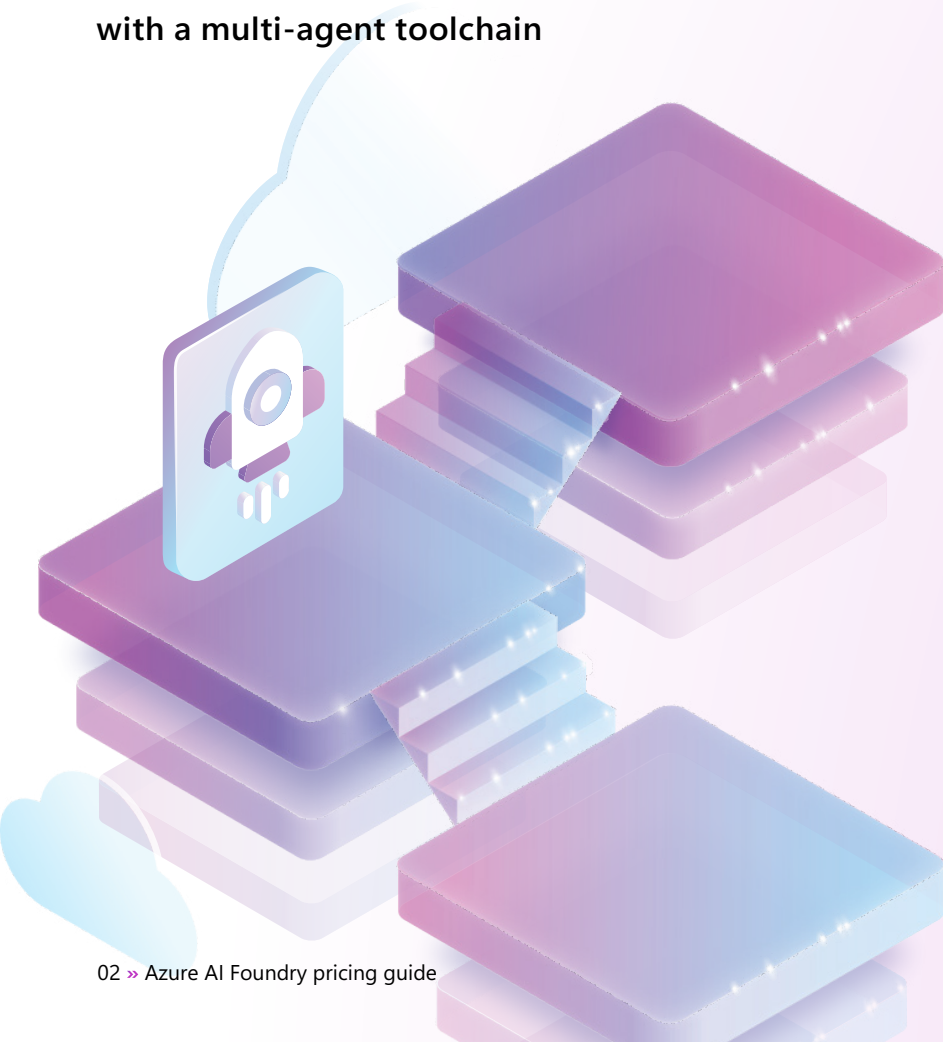
Pricing considerations: Customize with a multi-agent toolchain

13

Pricing considerations: Manage enterprise-ready AI with confidence

16

Start planning and managing costs for Azure AI Foundry



Introduction

This guide is designed to provide Developers and IT Leaders 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 at scale.

Azure AI Foundry is an AI application and agent factory offering models, agents, and tools accessible through a simple portal, unified SDK, and API. These capabilities are seamlessly integrated into developer workspaces like GitHub, Visual Studio, Copilot Studio, and Microsoft Fabric - 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 that allows our customers to 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 means 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 in Azure AI Foundry.



Pricing considerations

Design with the best models for your use case

Azure AI Foundry equips developers with a choice of AI models for their use case. Developers can choose from 11,000+ foundational, open, task, and industry-specific models in **Azure AI Foundry Models**. They can compare and benchmark models 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 **Foundry Models**. 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)

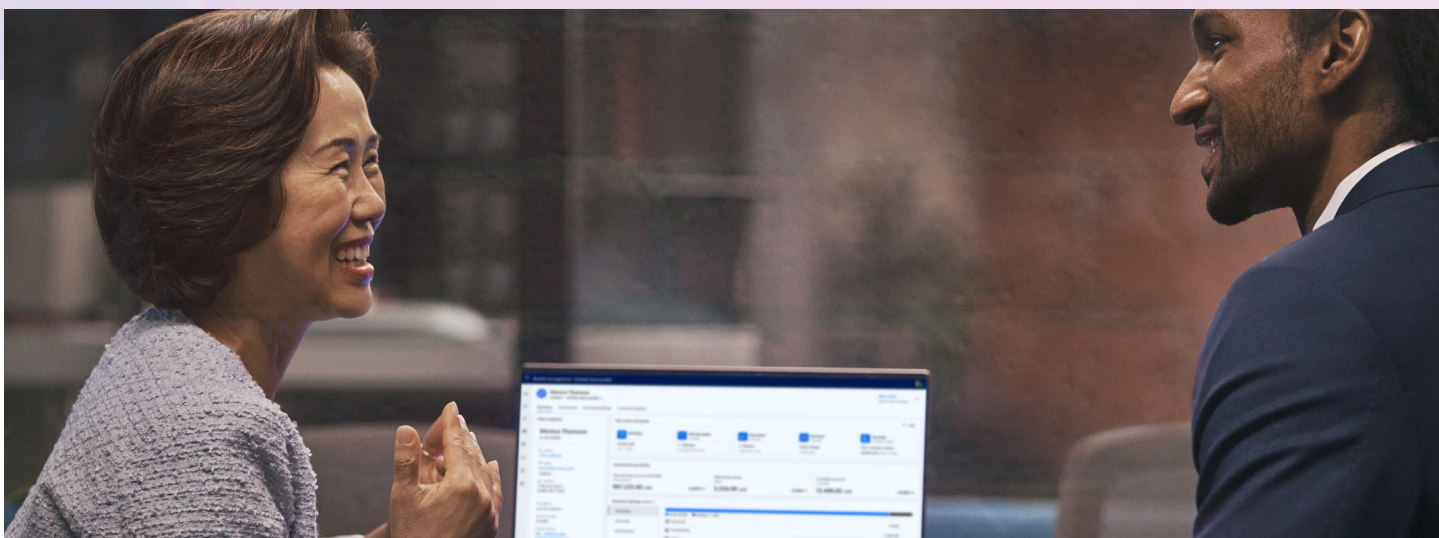
Inference costs: Charges are based on the number of input and output tokens processed during fine-tuning or inferencing. For example, the Phi-3 models have specific costs per 1,000 tokens for both input and output.

Provisioned Throughput Unit (PTU)

Provisioned throughput in Foundry Models allows you to specify the amount of throughput you need for a deployment. This supports having the necessary model processing capacity allocated and ready for use. Throughput is measured in **Provisioned Throughput Units (PTU)**, which is a standardized way of representing the throughput for your deployment. PTU is available as Global, DataZone, or Regional.

Managed Compute offering

Customers bring their own compute quota and they can use third-party models at a surcharge or they can use open models from Hugging Face for free.



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 provide customers 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 in Foundry Models 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 leaderboard and benchmarks in Foundry Models 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 Managed Compute vs. Serverless API

Azure AI Foundry provides flexibility in deployment of models.

- **Managed Compute:** This approach allows developers to use their own Azure infrastructure to build, customize, and manage models using the tools and capabilities provided by Azure AI Foundry. Managed Compute is ideal for organizations that require a high degree of control and customization over their AI models.
- **Serverless API:** In contrast, Serverless API offers pre-built models that can be easily integrated into applications with minimal infrastructure overhead. This approach is suitable for organizations looking for quick deployment and ease of use especially during initial POCs and early prototyping without having to worry about extensive customizations.

Which is more expensive?

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



Azure OpenAI in Foundry Models

Azure OpenAI, part of Azure Foundry models, brings the power of OpenAI's advanced language models – like GPT-4.1 series, the o-series reasoning models, Sora, Image-1, Realtime, Audio, and more – into the secure, scalable, and enterprise-ready environment of Microsoft Azure. It enables developers and organizations to build intelligent applications with capabilities such as reasoning, deep research, code generation, multimodal content generation, natural language understanding, summarization, translation, all while benefiting from Azure's compliance, data privacy, and expansive deployment options. Whether you're fine-tuning models, integrating with your own data, or deploying at scale, Azure OpenAI offers the tools and infrastructure to accelerate innovation responsibly and cost-efficiently.

Pricing model

Azure OpenAI 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 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 yearly 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 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, supporting 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

Microsoft Phi in Azure Foundry Models offers a family of Small Language Models (SLMs) with efficient performance for commercial and research tasks. It supports various functions with low latency, reduced costs, and offline use, and designed to help protect 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 helps to provide 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.

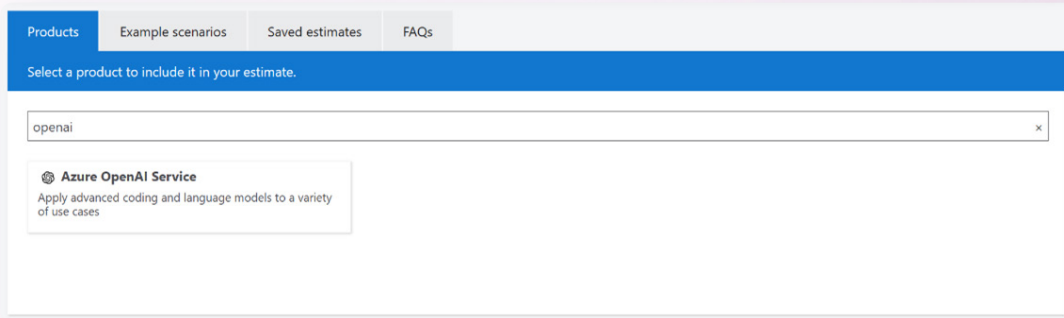
By understanding these pricing considerations, you can confidently start using the Foundry Models that are right for your AI solutions. Determine the most important components of your AI model and assess how those parameters and requirements impact overall cost.



Estimate costs before using Azure AI services

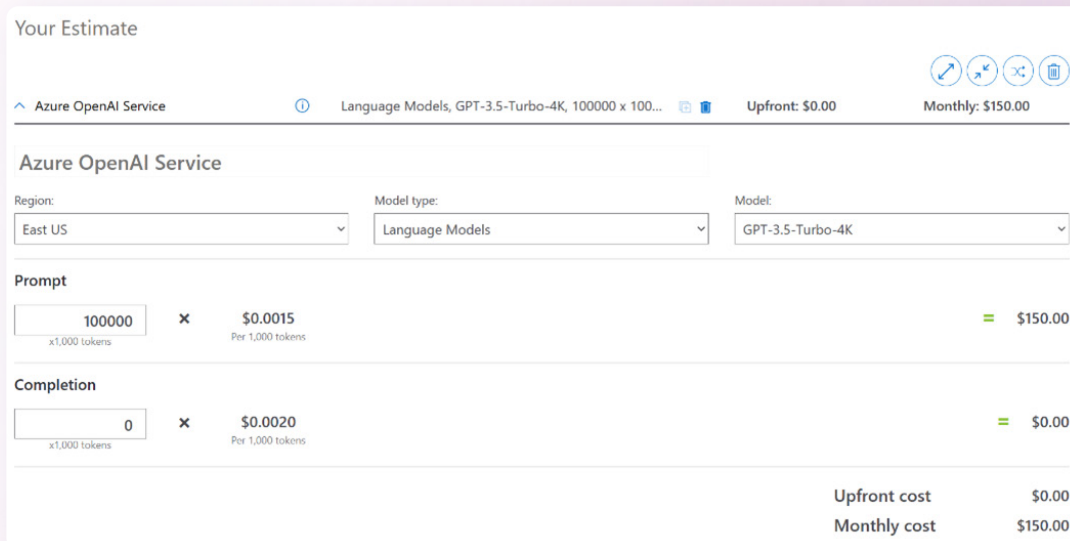
We recommend that you use the [Azure pricing calculator](#) to estimate costs before you add Azure AI services. To do that, follow the steps below:

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



The screenshot shows the 'Products' tab of the Azure pricing calculator. A search bar contains the text 'openai'. Below the search bar, a result card for 'Azure OpenAI Service' is displayed, with the description 'Apply advanced coding and language models to a variety of use cases'.

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



The screenshot shows the 'Your Estimate' page for Azure OpenAI Service. The configuration is as follows:

- Service: Azure OpenAI Service
- Model type: Language Models
- Model: GPT-3.5-Turbo-4K
- Region: East US

The cost breakdown is shown in the following table:

Category	Units	Unit Price	Cost
Prompt	100000 x1,000 tokens	\$0.0015 Per 1,000 tokens	\$150.00
Completion	0 x1,000 tokens	\$0.0020 Per 1,000 tokens	\$0.00
Upfront cost			\$0.00
Monthly cost			\$150.00

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

Azure OpenAI Service

Region:

Model type:

Pricing Strategy:

Model:

East US

Language Models

Provisioned (PTU)

GPT-4o-Regional-API-128K

Provisioned Throughput Units (PTUs)

To learn how many PTUs your scenario needs, use the [in-portal Capacity Calculator](#).
Discover how to transition your regional deployments and provisioned reservations to global and data zones on this [Learn page](#).

1

PTUs

Savings Options

Pay as you go

Pay as you go

Reservations

1 month reserved

1 year reserved

\$260.00

Average per month
(\$0.00 charged upfront)

\$260.00

Average per month
(\$0.00 charged upfront)

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.

Your Estimate

Azure OpenAI Service

Language Models, GPT-3.5-Turbo-4K, 100000 x 100...

Upfront: \$0.00

Monthly: \$150.00

Virtual Machines

1 D2 v3 (2 vCPUs, 8 GB RAM) x 730 Hours (Pay as y...

Upfront: \$0.00

Monthly: \$137.24

Support

SUPPORT:

Included

\$0.00

Select your program/offer

LICENSING PROGRAM:

Microsoft Customer Agreement (MCA)

Log in to see your Azure agreement pricing.

Show Dev/Test Pricing

Estimated upfront cost

\$0.00

Estimated monthly cost

\$287.24

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 may impact overall cost.



Pricing considerations

Customize with a multi-agent toolchain

Azure AI Foundry offers numerous agents and tools to help your development teams orchestrate agentic workflows and customize models for your use case with fine-tuning and distillation.

Tools and services you may consider leveraging include **Azure AI Search**, **Azure AI Foundry Agent Service**, Azure AI Services, and Azure Machine Learning, which contribute to the developer experience in Azure AI Foundry.

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 retrieval technology like hybrid search, query rewriting, and reranking. The pricing for Azure AI Search is based on Search Units (SUs) and the pricing tier you select. SUs are a combination of replicas and partitions accessible to your tier. Premium features come at an additional cost and follow their own pricing structure.

[Learn more](#) about Azure AI Search and choosing a tier.

[Learn more](#) about Azure AI Search pricing.



Azure AI Foundry Agent Service

Foundry Agent Service enables professional developers to securely customize, deploy, and scale AI agents that automate complex business processes. The pricing for Foundry Agent Service includes charges for consuming Foundry Model tokens. 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 action tool services like code interpreter, Azure Logic Apps, and Azure Functions also apply.

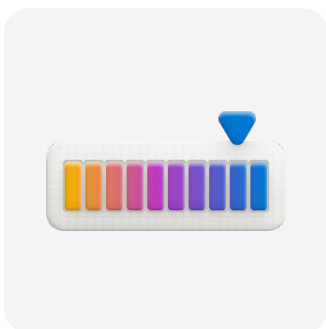


Azure AI Services

[Azure AI Services](#) is a suite of AI services and prebuilt agents, 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, and conversational AI.

We recommend that you use the [Azure pricing calculator](#) to estimate costs before you add Azure AI Services.

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 project, resources for other Azure services are also created. That means some costs can accrue.



Model training and customization

Fine-tuning in Azure AI Foundry involves customizing pre-trained models to better suit specific use cases. Whether you're optimizing for accuracy, speed, tone, or efficiency, Azure AI provides the tools you need to adapt frontier models to your real-world scenarios. Azure AI Foundry provides [fine-tuning capabilities](#) for specific Azure OpenAI (AOAI) models and third-party models. For AOAI models, pricing includes three components: training, hosting, and inference. In contrast, fine-tuning third-party models is priced based on the underlying Azure infrastructure used, including compute, storage, and deployment resources.

As pricing models and service availability may evolve, always refer to the official **Azure pricing page** for Azure OpenAI models and, to see the pricing for individual models, search for the model in [Azure Marketplace](#).

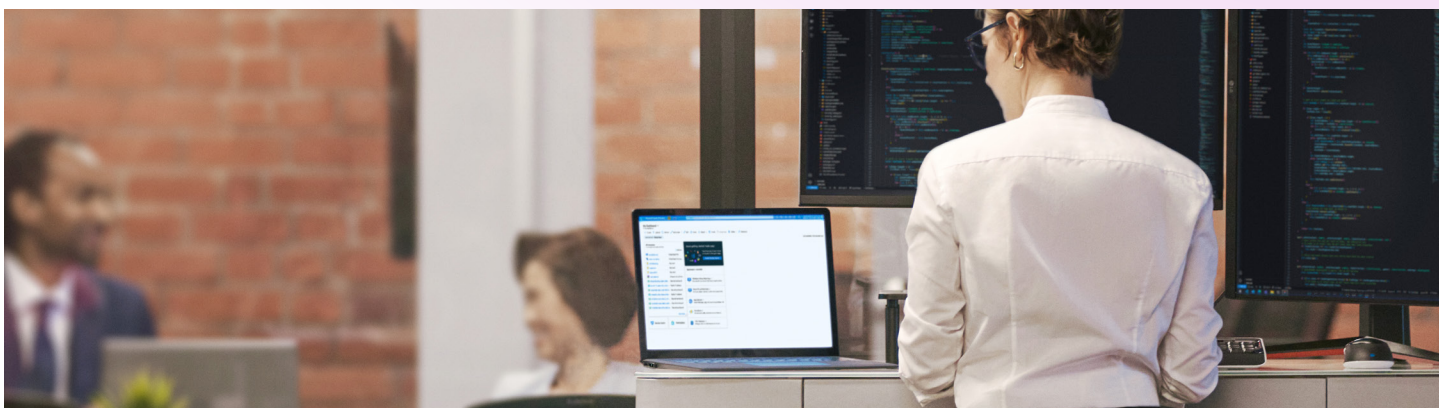


Pricing considerations

Manage enterprise-ready AI with confidence

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 endeavor to maintain the highest standards for 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. Azure AI Foundry offers a comprehensive set of safeguards aligned with Microsoft's Trustworthy AI principles. These safeguards span the entire AI development lifecycle – from model selection to deployment and monitoring – and are designed to ensure security, fairness, transparency, and accountability. 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.

But what happens in production once you've deployed your model, AI app, or agent? 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.





Azure AI Content Safety

Azure AI Content Safety is a critical component for detecting and mitigating risks and harms in AI applications, and it uses advanced language and vision models to detect offensive or inappropriate content in text, images, and multimodal content. The pricing for Azure AI Content Safety includes charges for using Text and Image APIs. While this service can incur additional costs, developers can help maintain content safety and appropriateness by applying custom content filters tailored to your requirements for enhanced safety and reassurance.

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.

[Learn more](#) about Azure AI Content Safety pricing.



Azure AI Foundry Observability

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.

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 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.

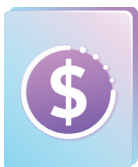
	Risk and safety evaluations	Performance and quality evaluations		Custom evaluations	
Metrics	Violence, hateful and unfair, sexual, self-harm, direct prompt injection attacks, indirect prompt injection attacks, and protected material	Groundedness, retrieval, relevance, coherence, fluency, and similarity	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	<ul style="list-style-type: none"> \$20/1M input tokens \$60/1M output tokens 	Azure OpenAI Service pricing	AzureML compute pricing	Azure OpenAI Service pricing	AzureML compute pricing

Monitoring and observability are critical for tracking model and application performance in production. Foundry Observability 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 is designed so 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.

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.



Start planning and managing costs for Azure AI Foundry

From designing with leading AI models to customizing with a comprehensive agent toolchain, and managing AI performance across the development lifecycle and into production, Azure AI Foundry equips developers and IT administrators with the tools they need to help them 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 AI Foundry 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, so 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 allow you to handle increased workloads without significant cost increases.
- **Utilize pre-built patterns and templates:** Take advantage of pre-built AI app templates and code 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.



Learn more about Azure AI Foundry product pricing and how the platform can benefit your organization.

Accelerate innovation today

[Learn more](#)

[Estimate costs](#)

No purchase necessary. Results may vary. Actual savings depend on individual circumstances. Based on internal data. Your results may differ. Terms and conditions apply. See website for details. Limited time offer. Subject to availability. We respect your privacy. Read our privacy policy. For informational purposes only. Not a guarantee of future performance. This advertisement is for illustrative purposes only.