Azure AI Foundry

Monitor model deployments in Azure AI model inference

Private Preview

When you have critical applications and business processes that rely on Azure resources, you need to monitor and get alerts for your system. The Azure Monitor service collects and aggregates metrics and logs from every component of your system, including Azure AI model inference model deployments. You can use this information to view availability, performance, and resilience, and get notifications of issues.

This document explains how you can use metrics and logs to monitor model deployments in Azure AI model inference, available as Private Preview.

Contents

Prerequisites 2

Metrics 2

Viewing metrics 2

Azure AI Foundry portal 3

Metrics explorer 4

Kusto query language (KQL) 6

Other tools 7

Configure diagnostic settings 7

Metrics reference 8

Foundry Models - HTTP Requests 8

Foundry Models - Latency 9

Foundry Modes – Usage 10

Support and feedback 11

Prerequisites

To use monitoring capabilities for model deployments in Azure AI model inference, you need the following:

- Enroll your subscription to the Private Preview.
    - Sign up for the private preview by completing this form. You will need the subscription ID you want to enable.
- An Azure AI services resource. For more information, see Create an Azure AI Services resource.

> If you are using Serverless API Endpoints and you want to take advantage of monitoring capabilities explained in this document, migrate your Serverless API Endpoints to Azure AI model inference.

- At least one model deployment.
- Access to diagnostic information for the resource.

Metrics

Once you have sign up for the private preview, Azure Monitor collects metrics from Azure AI model inference automatically. **No configuration is required**. These metrics are:

- Stored in the Azure Monitor time-series metrics database.
- Lightweight and capable of supporting near real-time alerting.
- Used to track the performance of a resource over time.

Your model deployments will report metrics to Azure Monitor under the category "Models". Other categories, like "Cognitive Services" and "Azure OpenAI" are also available in the resource.

> **Important**
> During Private Preview, Azure OpenAI model deployments are reported on metrics in category **Azure OpenAI**. At Public Preview, all models - including Azure OpenAI models - will report to the category **Foundry Models**.

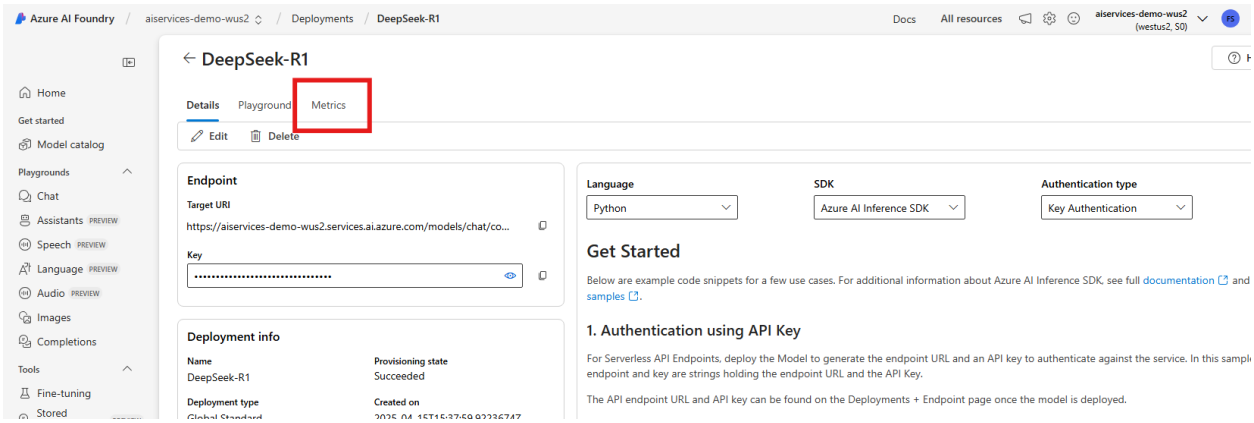
Viewing metrics

Azure Monitor metrics can be queried using multiple tools, including:

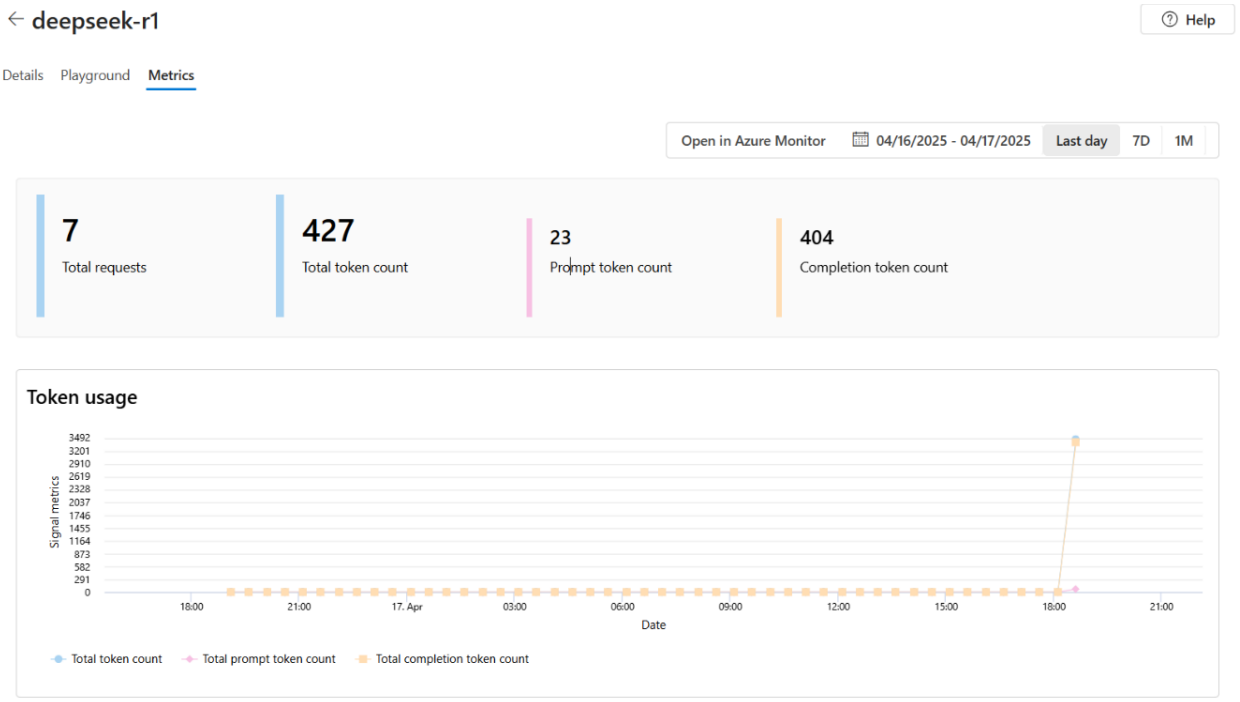
Azure AI Foundry portal

You can view metrics within Azure AI Foundry portal. To view them, follow these steps:

1. Go to Azure AI Foundry portal.
2. Navigate to your model deployment by selecting **Deployments**, and then select the name of the deployment you want to see metrics about.
3. Select the tab **Metrics**.

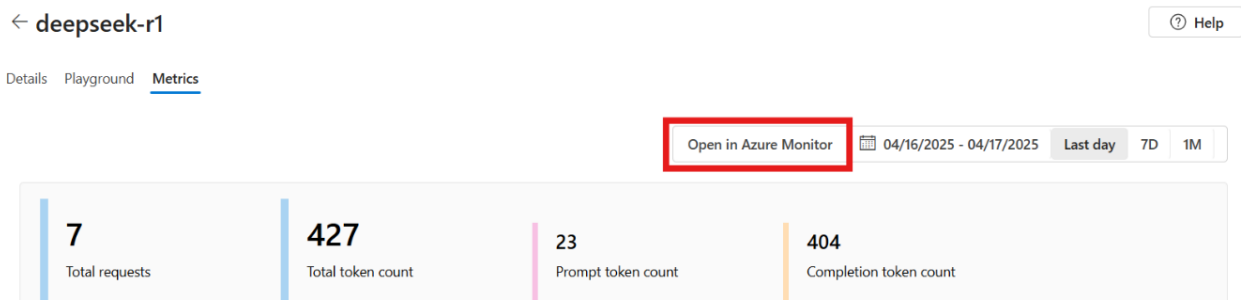


4. If the Metrics tab is missing, ensure you are selecting a model deployment for Azure AI Services resources. **Serverless API Endpoints are not supported**.
5. You can see a high-level view of the most common metrics you may be interested in.



IMPORTANT
During Private Preview, Azure AI Foundry portal may not show the metrics values as the page may not be updated for your subscription. If you see values all zero in the metrics, use **Open in Azure Monitor** option to view the metrics as explained later.

6. To slice, filter, or view model details about the metrics you can **Open in Azure Monitor**, where you have more advanced options.



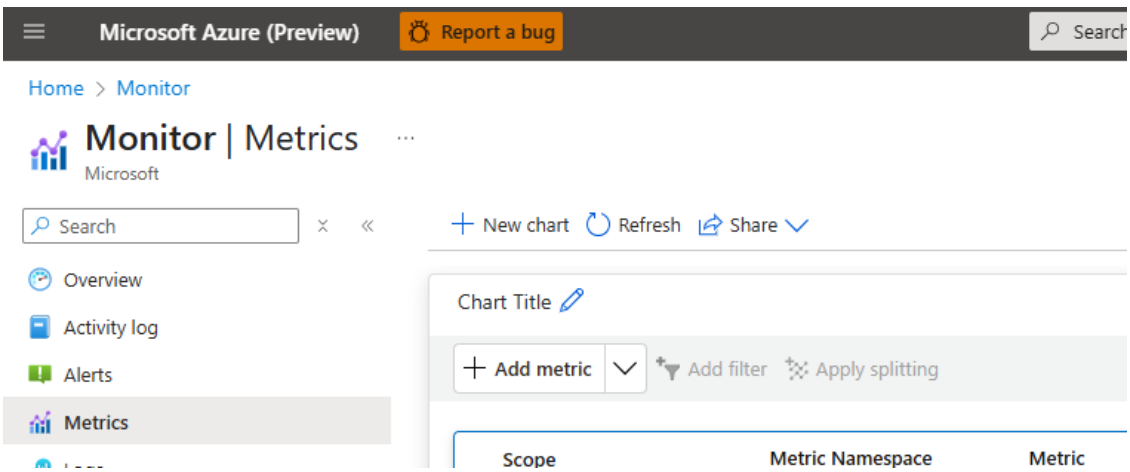
7. Use Metrics explorer to analyze the metrics.

Metrics explorer in Azure Monitor

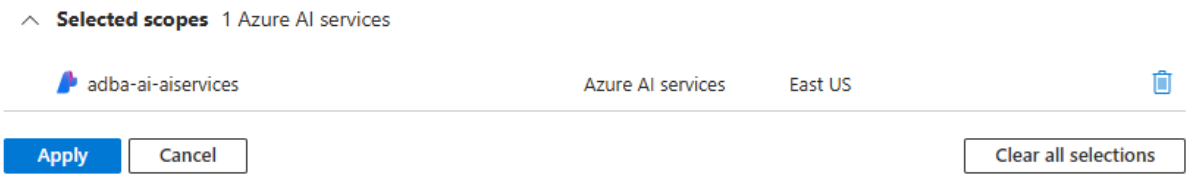
Metrics explorer is a tool in the Azure portal that allows you to view and analyze metrics for Azure resources. For more information, see Analyze metrics with Azure Monitor metrics explorer.

To use Azure Monitor, follow these steps:

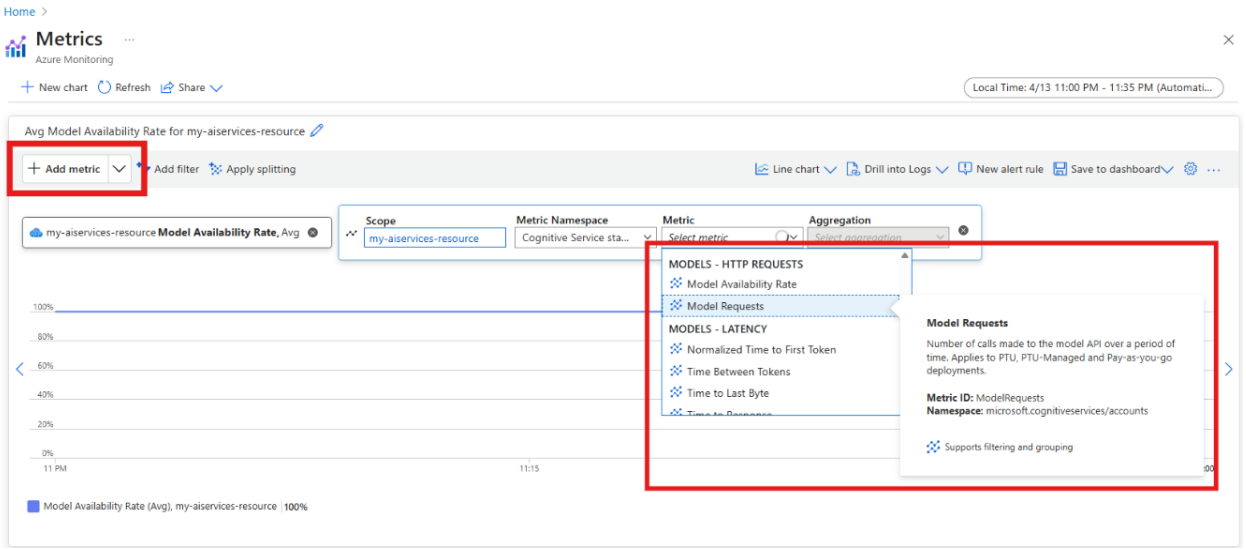
1. Go to Azure portal.
2. On the search box type and select **Monitor**.
3. Select **Metrics** in the left navigation bar.



4. On **Select scope**, select the resources you want to monitor. You can select either one resource or select a resource group or subscription. If that's the case, ensure you select **Resource types** as **Azure AI Services**.

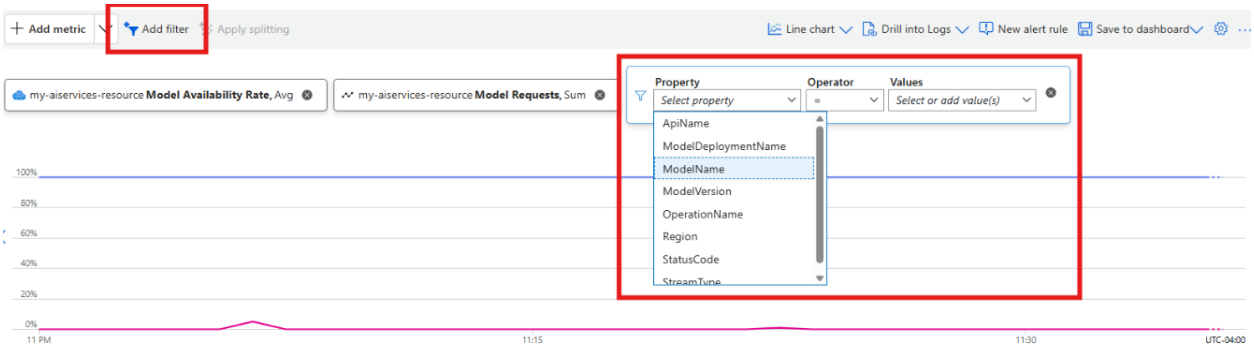


5. The metric explorer shows. Select the **metrics** that you want to explore. Select any of the metrics that start with the text "**Foundry Models** – ". The following example shows the number of requests made for the model deployments in the resource.

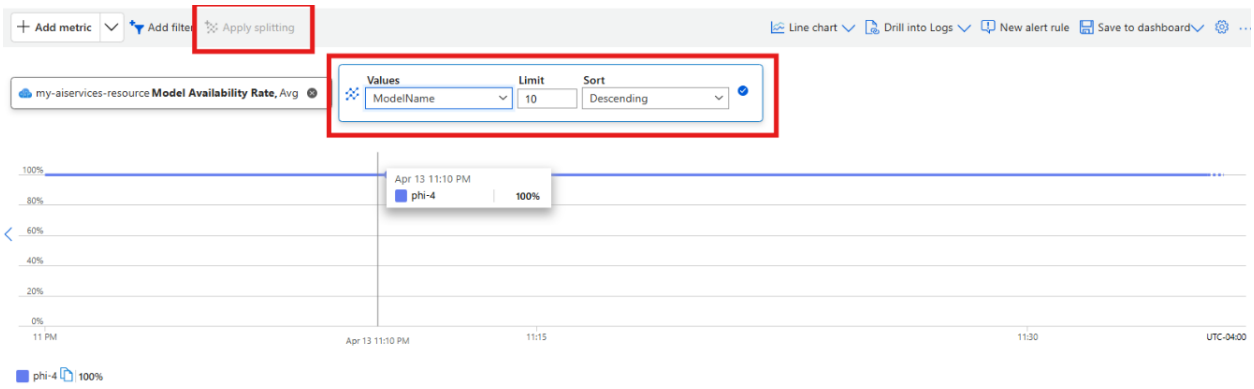


6. You can add as many metrics as needed to either the same chart or to a new chart.

7. If you need, you can filter metrics with any of the available dimensions of it.



8. It is useful to break down specific metrics by some of the dimensions. The following example shows how to break down the number of requests made to the resource by model by using the option **Add splitting**:



9. You can save your dashboards at any time to avoid having to configure them each time.

Kusto query language (KQL)

If you **configured diagnostic settings (read bellow)** to send metrics to Log Analytics, you can use the Azure portal to query and analyze log data using Kusto query language (KQL).

To query metrics, follow these steps:

1. Ensure you have configure diagnostic settings.
2. Go to Azure portal.
3. Locate the Azure AI Services resource you want to query.
4. In the left navigation bar, navigate to **Monitoring** > **Logs**.
5. Select the Log Analytics workspace that you configured with diagnostics.
6. From the Log Analytics workspace page, under Overview on the left pane, select Logs. The Azure portal displays a Queries window with sample queries and suggestions by default. You can close this window.
7. To examine the Azure Metrics, use the table `AzureMetrics` for your resource, and run the following query:

```
AzureMetrics
| take 100
| project TimeGenerated, MetricName, Total, Count, Maximum, Minimum, Average,
TimeGrain, UnitName
```

When you select **Monitoring** > **Logs** in the menu for your resource, Log Analytics opens with the query scope set to the current resource. The visible log queries include data from that specific resource only. To run a query that includes data from other resources or data from other Azure services, select **Logs** from the **Azure Monitor** menu in the Azure portal. For more information, see **Log query scope and time range in Azure Monitor Log Analytics** for details.

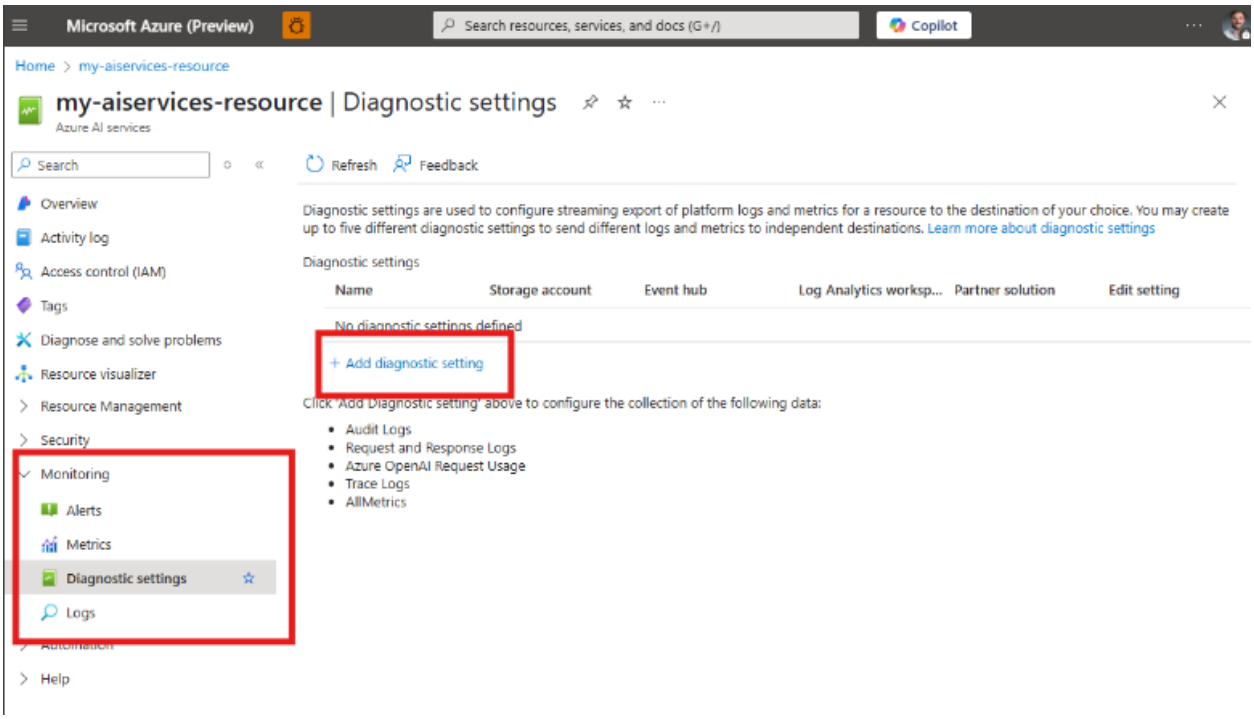
Other tools

Tools that allow more complex visualization include:

- Workbooks, customizable reports that you can create in the Azure portal. Workbooks can include text, metrics, and log queries.
- Grafana, an open platform tool that excels in operational dashboards. You can use Grafana to create dashboards that include data from multiple sources other than Azure Monitor.
- Power BI, a business analytics service that provides interactive visualizations across various data sources. You can configure Power BI to automatically import log data from Azure Monitor to take advantage of these visualizations.

Configure diagnostic settings

All of the metrics are exportable with diagnostic settings in Azure Monitor. To analyze logs and metrics data with Azure Monitor Log Analytics queries, you need to configure diagnostic settings for your Azure AI Services resource. You need to perform this operation on each resource.



There's a cost for collecting data in a Log Analytics workspace, so only collect the categories you require for each service. The data volume for resource logs varies significantly between services.

Metrics reference

The following table describes the metrics available. Metrics related to PTU may not be available from the start of the Private Preview. In case of questions, please reach out using the contact information at the end of the document.

Foundry Models - HTTP Requests

Metric name	Dimensions	Unit
ModelAvailabilityRate Availability percentage with the following calculation: (Total Calls - Server Errors)/Total Calls. Server Errors include any HTTP responses >=500	Region ModelDeploymentName ModelName ModelVersion	%
ModelRequests Number of calls made to the inference	ApiName OperationName Region	Count

API over a period of time. Applies to PTU, PTU-Managed and Pay-as-you-go deployments.	StreamType ModelDeploymentName ModelName ModelVersion StatusCode	

Foundry Models - Latency

Metric name	Dimensions	Unit
TimeToResponse Recommended latency (responsiveness) measure for streaming requests. Applies to PTU and PTU-managed deployments. Calculated as time taken for the first response to appear after a user sends a prompt, as measured by the API gateway. This number increases as the prompt size increases and/or cache hit size reduces. Note: this metric is an approximation as measured latency is heavily dependent on multiple factors, including concurrent calls and overall workload pattern. In addition, it does not account for any client-side latency that may exist between your client and the API endpoint. Please refer to your own logging for optimal latency tracking	ApiName OperationName Region StreamType ModelDeploymentName ModelName ModelVersion StatusCode	Miliseconds
NormalizedTimeBetweenTokens For streaming requests; model token generation rate, measured in milliseconds. Applies to PTU and PTU-managed deployments.	ApiName OperationName Region StreamType ModelDeploymentName ModelName ModelVersion	Miliseconds
TimeToLastByte For streaming and non-streaming requests; time it takes for last byte of response data to be received after request is made by model. Applies to	ApiName OperationName Region StreamType ModelDeploymentName ModelName ModelVersion	Miliseconds

PTU, PTU-managed, and Pay-as-you-go deployments		
NormalizedTimeToFirstToken For streaming and non-streaming requests; time it takes for first byte of response data to be received after request is made by model, normalized by token. Applies to PTU, PTU-managed, and Pay-as-you-go deployments.	ApiName OperationName Region StreamType ModelDeploymentName ModelName ModelVersion	Miliseconds
TokensPerSecond Enumerates the generation speed for a given model response. The total tokens generated is divided by the time to generate the tokens, in seconds. Applies to PTU and PTU-managed deployments.	ApiName OperationName Region StreamType ModelDeploymentName ModelName ModelVersion	Count

Foundry Modes – Usage

Metric name	Dimensions	Unit
InputTokens Number of prompt tokens processed (input) on a model. Applies to PTU, PTU-managed and Pay-as-you-go deployments.	ApiName Region ModelDeploymentName ModelName ModelVersion	Count
OutputTokens Number of tokens generated (output) from an OpenAI model. Applies to PTU, PTU-managed and Pay-as-you-go deployments.	ApiName Region ModelDeploymentName ModelName ModelVersion	Count
TotalTokens Number of inference tokens processed on a model. Calculated as prompt tokens (input) plus generated tokens (output). Applies to PTU, PTU-managed and Pay-as- you-go deployments.	ApiName Region ModelDeploymentName ModelName ModelVersion	Count
TokensCacheMatchRate	Region ModelDeploymentName	

Percentage of prompt tokens that hit the cache. Applies to PTU and PTU-managed deployments.	ModelName ModelVersion	
ProvisionedUtilization Utilization % for a provisoned-managed deployment, calculated as (PTUs consumed / PTUs deployed) x 100. When utilization is greater than or equal to 100%, calls are throttled and error code 429 returned.	Region ModelDeploymentName ModelName ModelVersion	Percent
ProvisionedConsumedTokens Total tokens minus cached tokens over a period of time. Applies to PTU and PTU-managed deployments.	Region ModelDeploymentName ModelName ModelVersion	Count
AudioOutputTokens Number of audio prompt tokens generated (output) on an OpenAI model. Applies to PTU-managed model deployments.	ApiName Region ModelDeploymentName ModelName ModelVersion	
AudioInputTokens Number of audio prompt tokens processed (input) on an OpenAI model. Applies to PTU-managed model deployments.	ApiName Region ModelDeploymentName ModelName ModelVersion	

Support and feedback

If you need help, support, or want to provide feedback please contact Facundo Santiago at fasantia@microsoft.com