



Azure OpenAI Provisioned Throughput Unit (PTU)

From dev & test to production

Pietro Brambati

Cloud Solution Architect – Azure Data & Generative AI

Agenda

1. Azure Open AI Quota
2. PTUs – What are they?
3. PTU Calculator
4. PTUs How do they work?



1. Azure Open AI Quota



Quota Limits

Model	Regions	Tokens per minute
gpt-35-turbo	East US, South Central US, West Europe, France Central, UK South	240 K
	North Central US, Australia East, East US 2, Canada East, Japan East, Sweden Central, Switzerland North	300 K
gpt-35-turbo-16k	East US, South Central US, West Europe, France Central, UK South	240 K
	North Central US, Australia East, East US 2, Canada East, Japan East, Sweden Central, Switzerland North	300 K
gpt-35-turbo-instruct	East US, Sweden Central	240 K
gpt-35-turbo (1106)	Australia East, Canada East, France Central, South India, Sweden Central, UK South, West US	120 K
gpt-4	East US, South Central US, France Central	20 K
	North Central US, Australia East, East US 2, Canada East, Japan East, UK South, Sweden Central, Switzerland North	40 K
gpt-4-32k	East US, South Central US, France Central	60 K
	North Central US, Australia East, East US 2, Canada East, Japan East, UK South, Sweden Central, Switzerland North	80 K
gpt-4 (1106-preview) GPT-4 Turbo	Australia East, Canada East, East US 2, France Central, UK South, West US	80 K
	South India, Norway East, Sweden Central	150 K
gpt-4 (vision-preview) GPT-4 Turbo with Vision	Sweden Central, Switzerland North, Australia East, West US	30 K
text-embedding-ada-002	East US, South Central US, West Europe, France Central	240 K
	North Central US, Australia East, East US 2, Canada East, Japan East, UK South, Switzerland North	350 K
Fine-tuning models (babbage-002, davinci-002, gpt-35-turbo-0613)	North Central US, Sweden Central	50 K
all other models	East US, South Central US, West Europe, France Central	120 K

<https://learn.microsoft.com/en-us/azure/ai-services/openai/quotas-limits>

Quota Limits

When a deployment is created, the assigned TPM will directly map to the tokens-per-minute rate limit enforced on its inferencing requests. A **Requests-Per-Minute (RPM)** rate limit will also be enforced whose value is set proportionally to the TPM assignment using the following ratio:

6 RPM per 1000 TPM.

The flexibility to distribute TPM globally within a subscription and region has allowed Azure OpenAI Service to loosen other restrictions:

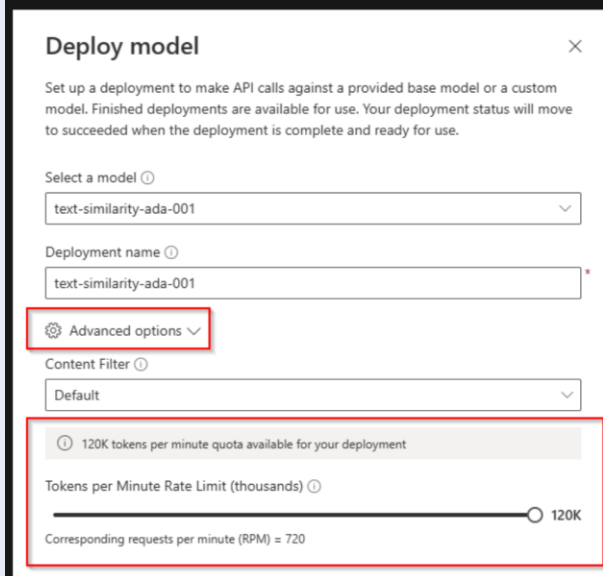
- The maximum resources per region are increased to 30.
- The limit on creating no more than one deployment of the same model in a resource has been removed.

Assign quota

When you create a model deployment, you have the option to assign Tokens-Per-Minute (TPM) to that deployment. TPM can be modified in increments of 1,000, and will map to the TPM and RPM rate limits enforced on your deployment, as discussed above.

To create a new deployment from within the Azure AI Studio under **Management** select **Deployments > Create new deployment**.

The option to set the TPM is under the **Advanced options** drop-down:



Deploy model [X]

Set up a deployment to make API calls against a provided base model or a custom model. Finished deployments are available for use. Your deployment status will move to succeeded when the deployment is complete and ready for use.

Select a model ⓘ
text-similarity-ada-001

Deployment name ⓘ
text-similarity-ada-001 *

⚙️ **Advanced options** ▾

Content Filter ⓘ
Default

📘 120K tokens per minute quota available for your deployment

Tokens per Minute Rate Limit (thousands) ⓘ
120K

Corresponding requests per minute (RPM) = 720

Azure OpenAI in Pay As You Go



Inconsistent
throughput



Hard to get
Quota



Latency
variance

2. PTU – What are they?



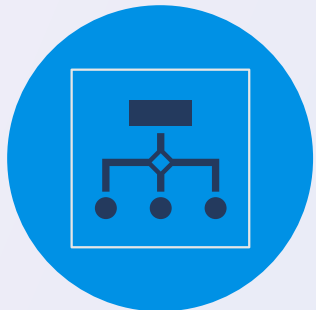
What is Provisioned Throughput?



A new Azure OpenAI Service feature that lets customers **reserve model processing capacity** for running high-volume or latency-sensitive workloads



Reserved processing capacity provides **consistent throughput** for workloads **with consistent characteristics**, such as prompt size, completion size, and number of concurrent API requests



Processing capacity is defined in units called "Provisioned Throughput Units" (PTUs) that are purchased on a monthly commitment



Once purchased, customers use PTUs to create provisioned Azure OpenAI deployments of GPT models during the term of their commitment

Key Concepts

- A Provisioned throughput unit (PTU) gives you an amount of model-processing capacity. The amount of capacity each call takes varies heavily by model, version, prompt size, generation size and call parameters.
- In-product capacity calculator provides the throughput per PTU you will get for a given workload.
 - Requests per minute will scale roughly linearly with PTUs for that workload
 - Throughput (Tokens per minute) varies significantly by both models and workloads

How many PTUs do I require?

To size the number of PTUs you require, we recommend the following steps:

- 1. Understand your throughput requirements.** To size your needs, you will need to know the [prompt input, generation output and expected calls per minute](#).
- 2. Use the Azure OpenAI Capacity Calculator.** Use the calculator to translate the workload to PTUs required for that call shape. <https://oai.azure.com/portal/calculator>
- 3. Validate via our benchmark tool & real-traffic.** Since workloads are not a static shape, it is always best to assess against real-traffic patterns. The exact distribution of your calls may change your PTU requirements. <https://aka.ms/aoai/benchmarking>

3. PTU Calculator

PTU Calculator

The screenshot shows the Azure AI Studio interface. At the top, there's a blue header with the Azure AI logo and 'Azure OpenAI Studio'. Below that, the main header area says 'Azure AI Studio PUBLIC PREVIEW' and 'Presenting the new Azure AI Studio (Preview)'. A sub-header reads 'Build, evaluate, and deploy your AI solutions from end to end.' with a button 'Explore Azure AI Studio'.

The main content area is titled 'Capacity calculator'. It includes a descriptive paragraph: 'This Azure OpenAI calculator enables you to estimate the number of PTUs needed for your workload. The calculator assumes a static prompt and generation size as well as call rate and are provided as an estimation only. Variations on these values will cause changes to the overall throughput per PTU you receive. For more accurate evaluation, run a benchmark test after deploying with a representational workload and monitor the Provisioned-Managed utilization values in the metrics tab.'

The calculator form has the following fields:

- 'Select a model *' dropdown menu with 'gpt-35-turbo' selected.
- 'Model version *' dropdown menu with '1106' selected.
- 'Workload size' slider.
- 'Prompt tokens *' input field with '1000'.
- 'Generation tokens *' input field with '100'.
- 'Peak calls per min *' input field with '10'.
- 'Estimate' slider.
- 'Suggested value' section showing 'PTU estimate' as '50'.

A left-hand navigation menu is visible, containing sections for 'Azure OpenAI', 'Playground' (with sub-items 'Chat', 'Completions', 'DALL-E (Preview)'), and 'Management' (with sub-items 'Deployments', 'Models', 'Data files', 'Quotas', 'Content filters (Preview)').

<https://oai.azure.com/portal/calculator>

PTU Calculator

The screenshot displays the Azure AI Studio interface. At the top, the navigation bar includes 'Azure AI | Azure OpenAI Studio' and the 'Azure AI Studio PUBLIC PREVIEW' logo. A main heading reads 'Presenting the new Azure AI Studio (Preview)' with a subtext 'Build, evaluate, and deploy your AI solutions from end to end.' and a button 'Explore Azure AI Studio'.

The left sidebar contains navigation items: 'Azure OpenAI', 'Playground', 'Chat', 'Completions', 'DALL-E (Preview)', 'Management', 'Deployments', 'Models', 'Data files', 'Quotas', and 'Content filters (Preview)'. The 'Quotas' item is selected.

The main content area is titled 'Quotas' and includes a 'View your quota by subscription and region and track' section with a 'Learn more' link. Below this are dropdowns for 'Subscription' and 'Region' (set to 'SWEDENCENTRA'). There are tabs for 'Standard', 'Provisioned', and 'Other', with 'Provisioned' being the active tab. A 'Manage commitment tiers' link and a 'Capacity calculator' button are visible.

A 'Capacity calculator' modal is open in the foreground. It contains the following text: 'This Azure OpenAI calculator enables you to estimate the number of PTUs needed for your workload. The calculator assumes a static prompt and generation size as well as call rate and are provided as an estimation only. Variations on these values will cause changes to the overall throughput per PTU you receive. For more accurate evaluation, run a benchmark test after deploying with a representational workload and monitor the Provisioned-Managed utilization values in the metrics tab.'

The calculator form includes:

- 'Select a model *' dropdown: gpt-4
- 'Model version *' dropdown: 0613
- 'Workload size' slider: 0
- 'Prompt tokens *' input: 0
- 'Generation tokens *' input: 0
- 'Peak calls per min *' input: 0
- 'Estimate' slider: 0
- 'Suggested value' section: 'PTU estimate' 0

A 'Close' button is located at the bottom right of the modal.

Sizing Examples: GPT4-0613 (8k)

The following are example throughputs for specific workload shapes.

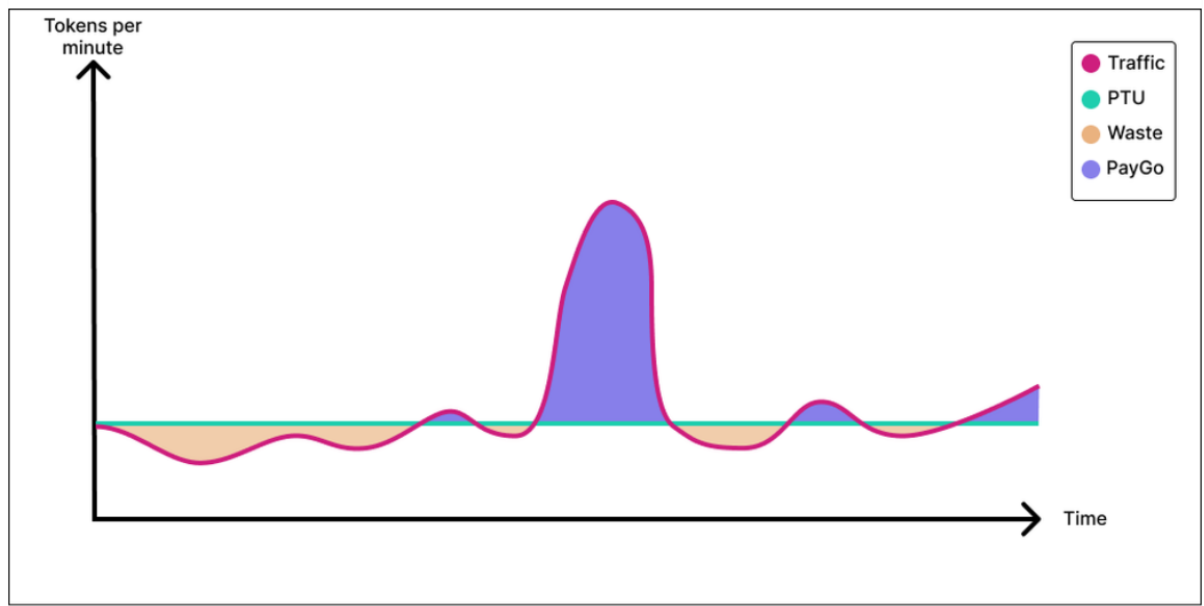
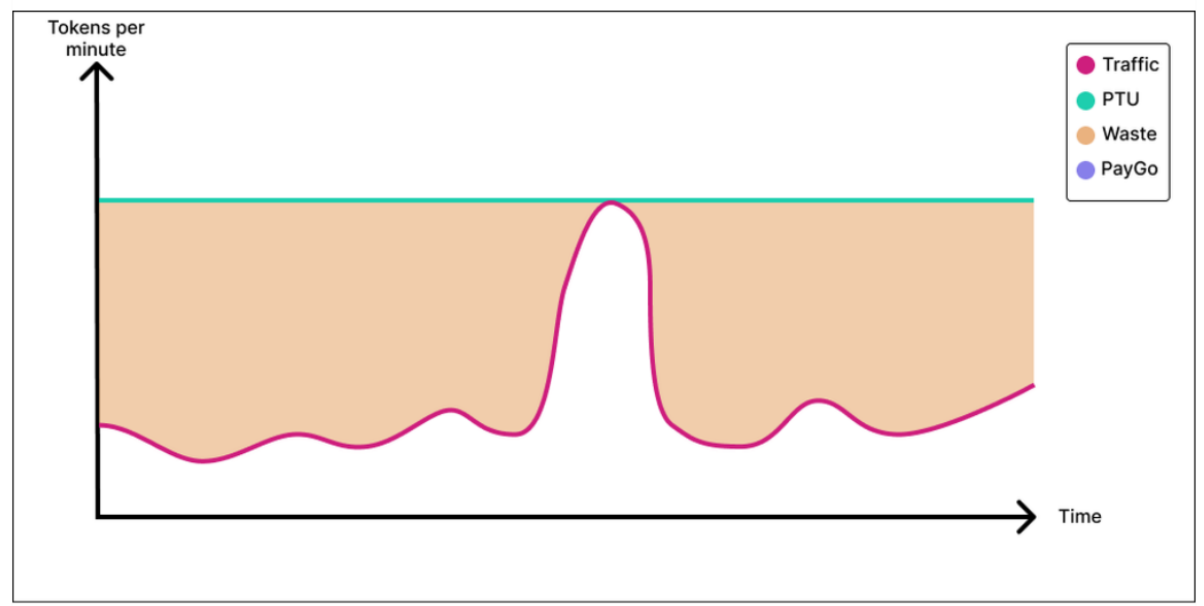
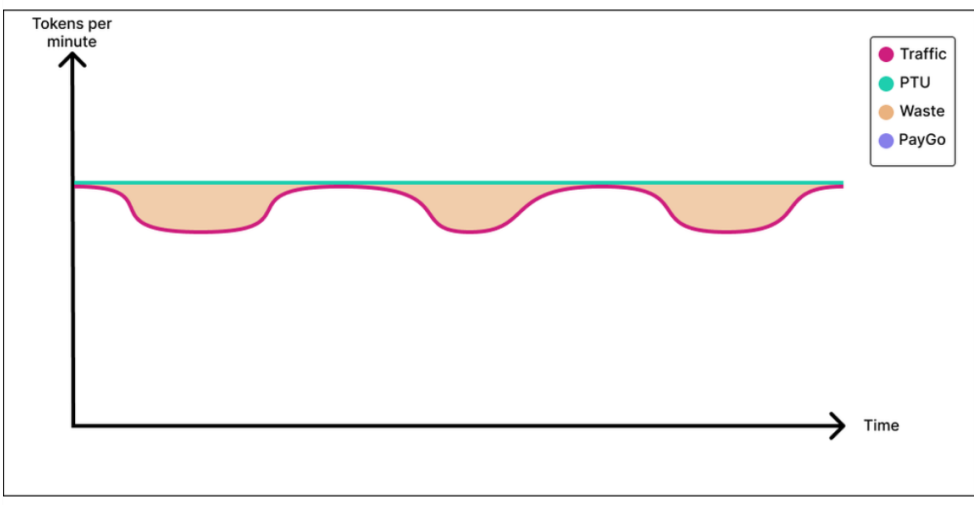
Please use the capacity calculator for an estimate of your specific workload:

<https://oai.azure.com/portal/calculator>

Prompt Size (tokens)	Generation Size (tokens)	Requests Per Minute	Tokens Per minute	PTUs required Deploy (actual)
500	200	60	42,000	200
1000	200	48	54,000	200
1500	50	60	93,000	200
1500	150	40	66,000	200
2000	300	25	57,500	200
3000	200	25	80,000	200
3000	1500	6	27,000	200

** These examples are based on customer scenarios and do not represent the full range of throughput. Always refer to the capacity calculator for the most accurate values

Sizing for PTU

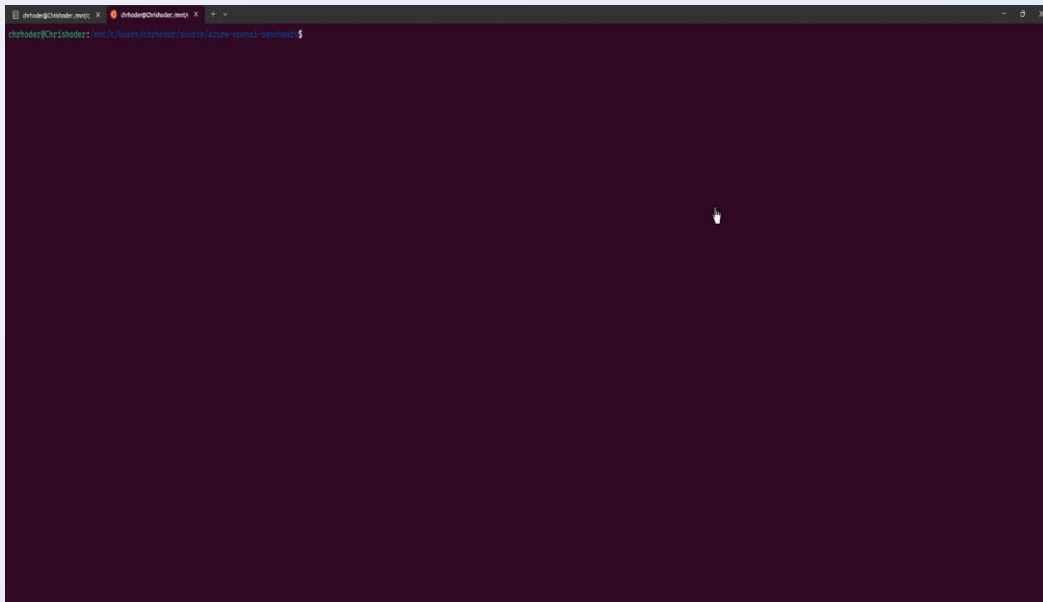


[Right-size your PTU deployment and save big \(microsoft.com\)](https://microsoft.com)

Managing Provisioned Deployments

Performance Assessment

Azure OpenAI provides a Python toolset for assessing deployment performance across the range of prompt/generation token sizes and RPM



Production Monitoring

Deployment metrics are built into Azure Monitor, including utilization, latency, token and request counts



Deployment

via Studio:



via Azure CLI:

```
az cognitiveservices account deployment create \  
  --name <myResourceName> \  
  --resource-group <myResourceGroupName> \  
  --deployment-name MyModel \  
  --model-name GPT-4 \  
  --model-version "0613" \  
  --model-format OpenAI \  
  --sku-capacity "100" \  
  --sku-name "Provisioned Managed"
```

Deploy model

Set up a deployment to make API calls against a provided base model or a custom model. Finished deployments are available for use. Your deployment status will move to succeeded when the deployment is complete and ready for use.

Select a model ⓘ
gpt-4

Model version ⓘ
0613 *

Deployment name ⓘ
cwhpm123 *

⚙️ Advanced options ▾

Content Filter ⓘ
Microsoft.Nil

Deployment type
Provisioned-Managed *

ⓘ 462 Provisioned Throughput Units available for deployment

Provisioned throughput units (PTU) ⓘ
300

Create Cancel

Provisioned Throughput Purchase and Reservation Model

- PTUs are purchased as a monthly commitment
- Committed PTUs are reserved for your use – They are there when you need them
- Billing is up-front for an entire month, starting on the day of purchase
- PTUs can be added to a commitment mid-month, but cannot be reduced
- If a commitment is not renewed, deployed PTUs will be billed a per-hour overage

4. PTUs How do they work?

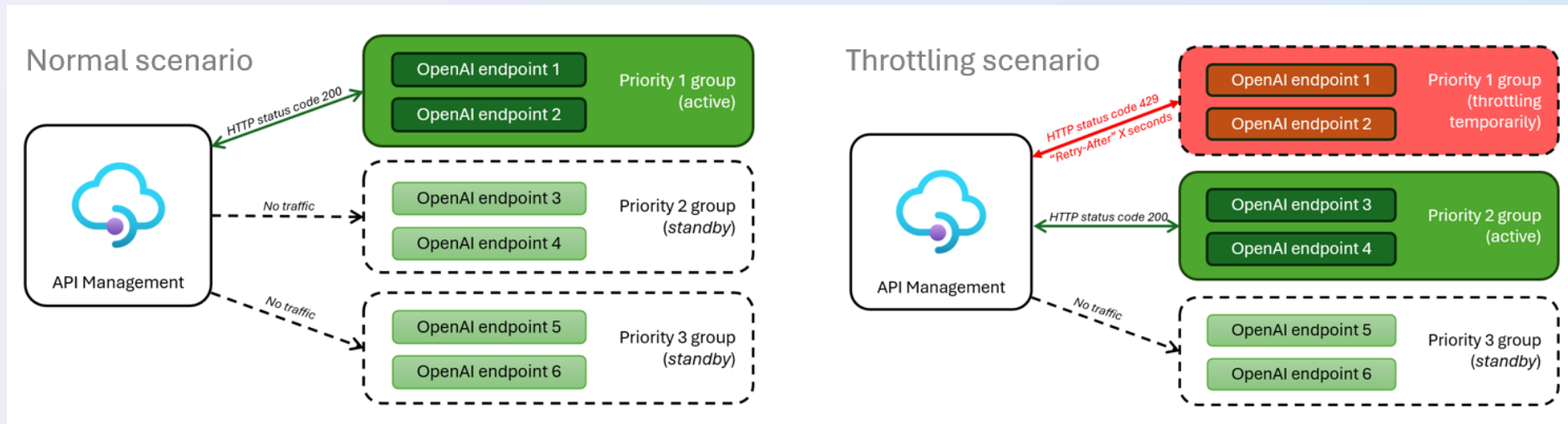
How does the service decide when to send a 429?

We use a variation of the leaky bucket algorithm to maintain utilization below 100% while allowing some burstiness in the traffic. The high-level logic is as follows:

1. Each customer has a set amount of capacity they can utilize on a deployment
2. When a request is made:
 1. When the current utilization is above 100%, the service **returns a 429 code** with the **retry-after-ms header** set to the time until utilization is below 100%
 2. Otherwise, the **service estimates the incremental change to utilization required to serve the request by combining prompt tokens and the specified max_tokens in the call.**
3. When a request finishes, we now know the actual compute cost for the call. To ensure an accurate accounting, we correct the utilization using the following logic:
 1. If the actual > estimated, then the difference is added to the deployment's utilization b. If the actual < estimated, then the difference is subtracted.
4. The overall utilization is decremented down at a continuous rate based on the number of PTUs deployed.

<https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/provisioned-throughput#how-does-the-service-decide-when-to-send-a-429>

Load Balancing



Links

[Azure OpenAI Service Provisioned Throughput Units \(PTU\) onboarding - Azure AI services | Microsoft Learn](#)
[Quickstart - Get started using Provisioned Deployments with Azure OpenAI Service - Azure OpenAI Service |](#)
[Azure OpenAI Service provisioned throughput - Azure AI services | Microsoft Learn](#)
[Azure OpenAI Service performance & latency - Azure OpenAI | Microsoft Learn](#)
[Smarter Azure Open AI Usage - Microsoft Community Hub](#)
[Azure/apim-aoai-smart-loadbalancing:](#)
<https://oai.azure.com/portal/calculator>
<https://aka.ms/aoai/benchmarking>
[Azure OpenAI Service quotas and limits - Azure AI services | Microsoft Learn](#)
[Manage Azure OpenAI Service quota - Azure AI services | Microsoft Learn](#)
[Right-size your PTU deployment and save big \(microsoft.com\)](#)

Grazie per l'ascolto

