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

# Azure OpenAI Service quotas and limits

* Article
* 12/11/2023
* 8 contributors

Feedback

## In this article

1. [Quotas and limits reference](https://learn.microsoft.com/en-us/azure/ai-services/openai/quotas-limits#quotas-and-limits-reference)
2. [Regional quota limits](https://learn.microsoft.com/en-us/azure/ai-services/openai/quotas-limits#regional-quota-limits)
3. [Next steps](https://learn.microsoft.com/en-us/azure/ai-services/openai/quotas-limits#next-steps)

This article contains a quick reference and a detailed description of the quotas and limits for Azure OpenAI in Azure AI services.

## Quotas and limits reference

The following sections provide you with a quick guide to the default quotas and limits that apply to Azure OpenAI:

Expand table

| **Limit Name** | **Limit Value** |
| --- | --- |
| OpenAI resources per region per Azure subscription | 30 |
| Default DALL-E 2 quota limits | 2 concurrent requests |
| Default DALL-E 3 quota limits | 2 capacity units (6 requests per minute) |
| Maximum prompt tokens per request | Varies per model. For more information, see [Azure OpenAI Service models](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models) |
| Max fine-tuned model deployments | 5 |
| Total number of training jobs per resource | 100 |
| Max simultaneous running training jobs per resource | 1 |
| Max training jobs queued | 20 |
| Max Files per resource | 30 |
| Total size of all files per resource | 1 GB |
| Max training job time (job will fail if exceeded) | 720 hours |
| Max training job size (tokens in training file) x (# of epochs) | 2 Billion |
| Max size of all files per upload (Azure OpenAI on your data) | 16 MB |

## Regional quota limits

The default quota for models varies by model and region. Default quota limits are subject to change.

Expand table

|  |  |  |
| --- | --- | --- |
| **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 | 10 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 |

### General best practices to remain within rate limits

To minimize issues related to rate limits, it's a good idea to use the following techniques:

* Implement retry logic in your application.
* Avoid sharp changes in the workload. Increase the workload gradually.
* Test different load increase patterns.
* Increase the quota assigned to your deployment. Move quota from another deployment, if necessary.

### How to request increases to the default quotas and limits

Quota increase requests can be submitted from the [Quotas](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/quota) page of Azure OpenAI Studio. Please note that due to overwhelming demand, quota increase requests are being accepted and will be filled in the order they are received. Priority will be given to customers who generate traffic that consumes the existing quota allocation, and your request may be denied if this condition is not met.

For other rate limits, please [submit a service request](https://learn.microsoft.com/en-us/azure/ai-services/cognitive-services-support-options?context=/azure/ai-services/openai/context/context).

## Next steps

Explore how to [manage quota](https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/quota) for your Azure OpenAI deployments. Learn more about the [underlying models that power Azure OpenAI](https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models).