Starting April 29, 2025, Gemini 1.5 Pro and Gemini 1.5 Flash models are not available in projects that have no prior usage of these models, including new projects. For details, see Model versions and lifecycle (/vertex-ai/generative-ai/docs/learn/model-versions#legacy-stable).

Generative AI on Vertex AI quotas and system limits

Release Notes

This page introduces two ways to consume generative AI services, provides a list of quotas by region and model, and shows you how to view and edit your quotas in the Google Cloud console.

Overview

There are two ways to consume generative AI services. You can choose *pay-as-you-go (PayGo)*, or you can pay in advance using *Provisioned Throughput*.

If you're using PayGo, your usage of generative AI features is subject to one of the following quota systems, depending on which model you're using:

- Models earlier than Gemini 2.0 use a standard quota system for each generative AI model to help ensure fairness and to reduce spikes in resource use and availability. Quotas apply to Generative AI on Vertex AI requests for a given Google Cloud project and supported region.
- Newer models use Dynamic shared quota (DSQ) (/vertex-ai/generative-ai/docs/dynamic-shared-quota), which dynamically distributes available PayGo capacity among all customers for a specific model and region, removing the need to set quotas and to submit quota increase requests. **There are no quotas with DSQ**.

To help ensure high availability for your application and to get predictable service levels for your production workloads, see Provisioned Throughput (/vertex-ai/generative-ai/docs/provisioned-throughput).

Quota system by model

The following models support Dynamic shared quota (DSQ) (/vertex-ai/generative-ai/docs/dynamic-shared-quota):

- Gemini 2.0 Flash with Live API (/vertex-ai/generative-ai/docs/models/gemini/2-0-flash)
- Gemini 2.0 Flash with image generation (/vertex-ai/generative-ai/docs/models/gemini/2-0-flash)
- Gemini 2.5 Pro (/vertex-ai/generative-ai/docs/models/gemini/2-5-pro)
- Gemini 2.5 Flash (/vertex-ai/generative-ai/docs/models/gemini/2-5-flash)
- Gemini 2.0 Flash (/vertex-ai/generative-ai/docs/models/gemini/2-0-flash)
- Gemini 2.0 Flash-Lite (/vertex-ai/generative-ai/docs/models/gemini/2-0-flash-lite)

The following legacy Gemini models support DSQ:

- Gemini 1.5 Pro
- Gemini 1.5 Flash

Non-Gemini and earlier Gemini models use the standard quota system. For more information, see Vertex AI quotas and limits (/vertex-ai/docs/quotas).

Tuned model quotas

Tuned model inference shares the same quota as the base model. There is no separate quota for tuned model inference.

Text embedding limits

Each text embedding model request can have up to 250 input texts (generating 1 embedding per input text) and 20,000 tokens per request. Only the first 2,048 tokens in each input text are used to compute the embeddings. For `gemini-embedding-001`, each request can only include a single input text. The quota for this model (/vertex-ai/docs/quotas#model-region-quotas) is listed under the name `gemini-embedding`.

Embed content input tokens per minute per base model

Unlike previous embedding models which were primarily limited by RPM quotas, the quota for the Gemini Embedding model limits the number of tokens that can be sent per minute per project.

Quota	Value
Embed content input tokens per minute	200000

Vertex AI Agent Engine limits

The following limits apply to Vertex AI Agent Engine (/vertex-ai/generative-ai/docs/agent-engine/overview) for a given project in each region.

Description	Limit
Create/Delete/Update Vertex AI Agent Engine per minute	10
Create/Delete/Update Vertex AI Agent Engine Sessions per minute	100
Query/StreamQuery Vertex AI Agent Engine per minute	60
Append event to Vertex AI Agent Engine Sessions per minute	100
Maximum number of Vertex AI Agent Engine resources	100

Batch prediction

The quotas and limits for batch prediction jobs are the same across all regions.

Concurrent batch prediction job limits

The following table lists the limits for the number of concurrent batch prediction jobs:

Limit	Value
Concurrent batch prediction requests, per region, for Gemini models	8

If the number of tasks submitted exceeds the allocated limit, the tasks are placed in a queue and processed when the limit capacity becomes available.

Concurrent batch prediction job quotas

The following table lists the quotas for the number of concurrent batch prediction jobs, which don't apply to Gemini models:

Quota	Value
`aiplatform.googleapis.com/textembedding_gecko_concurrent_batch_prediction_jobs`	4

If the number of tasks submitted exceeds the allocated quota, the tasks are placed in a queue and processed when the quota capacity becomes available.

View and edit the quotas in the Google Cloud console

To view and edit the quotas in the Google Cloud console, do the following:

1. Go to the **Quotas and System Limits** page.

   Go to Quotas and System Limits (https://console.cloud.google.com/iam-admin/quotas)

2. To adjust the quota, copy and paste the property `aiplatform.googleapis.com/generate_content_requests_per_minute_per_project_per_base_model` in the **Filter**. Press **Enter**.
3. Click the three dots at the end of the row, and select **Edit quota**.
4. Enter a new quota value in the pane, and click **Submit request**.

Vertex AI RAG Engine

The VPC-SC security control (/vertex-ai/generative-ai/docs/security-controls) is supported by RAG Engine. Data residency, CMEK, and AXT security controls aren't supported.

For each service to perform retrieval-augmented generation (RAG) using RAG Engine, the following quotas apply, with the quota measured as requests per minute (RPM).

Service	Quota	Metric
RAG Engine data management APIs	60 RPM	`VertexRagDataService requests per minute per region`

`RetrievalContexts` API	1,500 RPM	`VertexRagService retrieve requests per minute per region`
`base_model: textembedding-gecko`	1,500 RPM	`Online prediction requests per base model per minute per region per base_model` An additional filter for you to specify is `base_model: textembedding-gecko`

The following limits apply:

Service	Limit	Metric
Concurrent `ImportRagFiles` requests	3 RPM	`VertexRagService concurrent import requests per region`
Maximum number of files per `ImportRagFiles` request	10,000	`VertexRagService import rag files requests per region`

For more rate limits and quotas, see Generative AI on Vertex AI rate limits (/vertex-ai/generative-ai/docs/quotas).

Gen AI evaluation service

The Gen AI evaluation service uses `gemini-2.0-flash` as a default judge model for model-based metrics. A single evaluation request for a model-based metric might result in multiple underlying requests to the Gen AI evaluation service. Each model's quota is calculated on a per-project basis, which means that any requests directed to `gemini-2.0-flash` for model inference and model-based evaluation contribute to the quota. Quotas for the Gen AI evaluation service and the underlying judge model are shown in the following table:

Request quota	Default quota
Gen AI evaluation service requests per minute	1,000 requests per project per region
Online prediction requests per minute for `base_model: gemini-2.0-flash`	See Quotas by region and model. (/vertex-ai/generative-ai/docs/learn/models#quotas_by_region_and_model)

If you receive an error related to quotas while using the Gen AI evaluation service, you might need to file a quota increase request. See View and manage quotas (/docs/quotas/view-manage) for more information.

Limit	Value
Gen AI evaluation service request timeout	60 seconds

When you use the Gen AI evaluation service for the first time in a new project, you might experience an initial setup delay up to two minutes. If your first request fails, wait a few minutes and then retry. Subsequent evaluation requests typically complete within 60 seconds.

The maximum input and output tokens for model-based metrics depend on the model used as the judge model. See Google models (/vertex-ai/generative-ai/docs/learn/models) for a list of models.

Vertex AI Pipelines quotas

Each tuning job uses Vertex AI Pipelines. For more information, see Vertex AI Pipelines quotas and limits (/vertex-ai/docs/quotas#vertex-ai-pipelines).

What's next

- To learn more about dynamic shared quota, see Dynamic shared quota (/vertex-ai/generative-ai/docs/dsq).
- To learn about quotas and limits for Vertex AI, see Vertex AI quotas and limits (/vertex-ai/docs/quotas).
- To learn more about Google Cloud quotas and limits, see Understand quota values and system limits (/docs/quotas/understand-limits).

Last updated 2025-06-06 UTC.