

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).

# Tuning API

Model tuning is a crucial process in adapting Gemini to perform specific tasks with greater precision and accuracy. Model tuning works by providing a model with a training dataset that contains a set of examples of specific downstream tasks.

Use the Gemini tuning API for the following use-cases:

- [Supervised fine tuning](#) (/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning)

## Supported Models:

You can use supervised fine-tuning on the following Gemini models:

- [Vertex AI Model Optimizer](#) (/vertex-ai/generative-ai/docs/model-reference/vertex-ai-model-optimizer) ▲
- [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)

Translation LLM V2 (`translation-11m-002`) is also supported.

## Example syntax

Syntax to tune a model.

**curl**  
(#curl)

```
curl -X POST \  
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \  
  -H "Content-Type: application/json" \  
  
https://TUNING_JOB_REGION-aiplatform.googleapis.com/v1/projects/PROJECT_ID/loca
```

```
-d '{
  "baseModel": "...",
  "supervisedTuningSpec" : {
    ...
    "hyper_parameters": {
      ...
    },
  },
  "tunedModelDisplayName": "",
}'
```

## Parameters list

See [examples](#) (#examples) for implementation details.

### Request body

The request body contains data with the following parameters:

Parameters	
source_model	<div>Optional: <b>string</b></div> <div>Name of the foundation model that's being tuned.</div>
tunedModelDisplayName	<div><b>string</b></div> <div>The display name of the <b>TunedModel</b>. The name can be up to 128 characters long and can consist of any UTF-8 characters.</div>
supervisedTuningSpec	
Parameters	
training_dataset	<div><b>string</b></div> <div>Cloud Storage URI of your training dataset. The dataset must be formatted as a JSONL file. For best results, provide at least 100 to 500 examples. For more information, see <a href="#">About supervised tuning datasets</a>. (/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-about)</div>

validation_dataset	Optional: <b>string</b>  Cloud Storage URI of your validation dataset. Your dataset must be formatted as a JSONL file. A dataset can contain up to 256 examples. If you provide this file, the data is used to generate validation metrics periodically during fine-tuning. For more information, see <a href="#">About supervised tuning datasets</a> (/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-about).
epoch_count	Optional: <b>int</b>  Number of complete passes the model makes over the entire training dataset during training. Vertex AI automatically adjusts the default value to your training dataset size. This value is based on benchmarking results to optimize model output quality.
learning_rate_multiplier	Optional: <b>float</b>  Multiplier for adjusting the default learning rate.
adapter_size	Optional: <b>AdapterSize</b>  Adapter size for tuning.
tuned_model_display_name	Optional: <b>string</b>  Display name of the <b>TunedModel</b> . The name can be up to 128 characters long and can consist of any UTF-8 characters.

**AdapterSize**

Adapter size for tuning job.

Parameters

ADAPTER_SIZE_UNSPECIFIED	Unspecified adapter size.
ADAPTER_SIZE_ONE	Adapter size 1.
ADAPTER_SIZE_FOUR	Adapter size 4.
ADAPTER_SIZE_EIGHT	Adapter size 8.
ADAPTER_SIZE_SIXTEEN	Adapter size 16.

Examples

## Create a supervised tuning Job

You can create a supervised text model tuning job by using the Vertex AI SDK for Python or by sending a POST request.





### Basic use case

The basic use case only sets values for `baseModel` and `training_dataset_uri`. All other parameters use the default values.


**REST**Python (#python)  
(#rest)

To create a model tuning job, send a POST request by using the `tuningJobs.create` (/vertex-ai/docs/reference/rest/v1/projects.locations.tuningJobs/create) method. Note that some of the parameters are not supported by all of the models. Ensure that you only include the applicable parameters for the model that you're tuning.

Before using any of the request data, make the following replacements:

- **PROJECT\_ID** : Your project ID (/resource-manager/docs/creating-managing-projects#identifiers).
- **TUNING\_JOB\_REGION** : The region (/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-region-settings) where the tuning job runs. This is also the default region for where the tuned model is uploaded.
- **BASE\_MODEL** : Name of the foundation model to tune.
- **TRAINING\_DATASET\_URI** : Cloud Storage URI of your training dataset. The dataset must be formatted as a JSONL file. For best results, provide at least 100 to 500 examples. For more information, see About supervised tuning datasets (/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-about).

HTTP method and URL:

POST `https://TUNING_JOB_REGION -aiplatform.googleapis.com/v1/projects/PROJECT`

Request JSON body:

```
{
  "baseModel": "BASE_MODEL ✎",
  "supervisedTuningSpec": {
    "training_dataset_uri": "TRAINING_DATASET_URI ✎"
  },
}
```

To send your request, choose one of these options:

**curlPowerShell** (#powershell)  
(#curl)

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **gcloud init** (/sdk/gcloud/reference/init) or **gcloud auth login** (/sdk/gcloud/reference/auth/login) , or by using **Cloud Shell** (/shell/docs), which automatically logs you into the **gcloud** CLI . You can check the currently active account by running **gcloud auth list** (/sdk/gcloud/reference/auth/list).

Save the request body in a file named `request.json`, and execute the following command:

```
curl -X POST \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json; charset=utf-8" \
  -d @request.json \
  "https://TUNING_JOB_REGION ✎ -aiplatform.googleapis.com/v1/projects/
```

You should receive a JSON response similar to the following.

### + Response

```
{
  "name": "projects/PROJECT_ID/locations/TUNING_JOB_REGION/tuningJobs/TUNING_JO
  "createTime": CREATE_TIME,
```

```

"updateTime": UPDATE_TIME,
"status": "STATUS",
"supervisedTuningSpec": {
  "training_dataset_uri": "TRAINING_DATASET_URI",
  "validation_dataset_uri": "VALIDATION_DATASET_URI",
  "hyper_parameters": {
    "epoch_count": EPOCH_COUNT,
    "learning_rate_multiplier": LEARNING_RATE_MULTIPLIER
  },
},
"tunedModelDisplayName": "TUNED_MODEL_DISPLAYNAME"
}

```






## Advanced use case








The advance use case expands upon the basic use case, but also sets values for optional `hyper_parameters`, such as `epoch_count`, `learning_rate_multiplier` and `adapter_size`.

### RESTPython (#python) (#rest)


To create a model tuning job, send a POST request by using the [tuningJobs.create](#) (/vertex-ai/docs/reference/rest/v1/projects.locations.tuningJobs/create) method. Note that some of the parameters are not supported by all of the models. Ensure that you only include the applicable parameters for the model that you're tuning.

Before using any of the request data, make the following replacements:

- **PROJECT\_ID** : Your project ID  
(/resource-manager/docs/creating-managing-projects#identifiers).
- **TUNING\_JOB\_REGION** : The region  
(/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-region-settings) where the tuning job runs. This is also the default region for where the tuned model is uploaded.
- **BASE\_MODEL** : Name of the foundation model to tune.
- **TRAINING\_DATASET\_URI** : Cloud Storage URI of your training dataset. The dataset must be formatted as a JSONL file. For best results, provide at least 100 to 500 examples. For more information, see [About supervised tuning datasets](#) (/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-about).
- **VALIDATION\_DATASET\_URI**  Optional: The Cloud Storage URI of your validation dataset file.

- **EPOCH\_COUNT**  Optional: The number of complete passes the model makes over the entire training dataset during training. Leave it unset to use the pre-populated recommended value.  
(/vertex-ai/generative-ai/docs/model-reference/tuning#supervisedtuningspec)
- **ADAPTER\_SIZE**  Optional: The Adapter size (/vertex-ai/generative-ai/docs/model-reference/tuning#adaptersize) to use for the tuning job. The adapter size influences the number of trainable parameters for the tuning job. A larger adapter size implies that the model can learn more complex tasks, but it requires a larger training dataset and longer training times.
- **LEARNING\_RATE\_MULTIPLIER**  Optional: A multiplier to apply to the recommended learning rate. Leave it unset to use the recommended value.  
(/vertex-ai/generative-ai/docs/model-reference/tuning#supervisedtuningspec)
- **EXPORT\_LAST\_CHECKPOINT\_ONLY**  Optional: Set to `true` to use only the latest checkpoint.
- **TUNED\_MODEL\_DISPLAYNAME**  Optional: A display name for the tuned model. If not set, a random name is generated.
- **KMS\_KEY\_NAME**  Optional: The Cloud KMS resource identifier of the customer-managed encryption key used to protect a resource. The key has the format: `projects/my-project/locations/my-region/keyRings/my-kr/cryptoKeys/my-key`. The key needs to be in the same region as where the compute resource is created. For more information, see Customer-managed encryption keys (CMEK).  
(https://cloud.google.com/vertex-ai/docs/general/cmek).
- **SERVICE\_ACCOUNT**  Optional: The service account that the tuningJob workload runs as. If not specified, the Vertex AI Secure Fine-Tuning Service Agent in the project is used. See Tuning Service Agent (https://cloud.google.com/iam/docs/service-agents#vertex-ai-secure-fine-tuning-service-account). If you plan to use a customer-managed Service Account, you must grant the `roles/aiplatform.tuningServiceAgent` role to the service account. Also grant the `vertex-ai-service-account` permission to the Tuning Service Agent.

HTTP method and URL:

POST `https://TUNING_JOB_REGION  -aiplatform.googleapis.com/v1/projects/PROJECT`

Request JSON body:

```
{
  "baseModel": "BASE_MODEL ",
  "supervisedTuningSpec": {
    "trainingDatasetUri": "TRAINING_DATASET_URI ",
    "validationDatasetUri": "VALIDATION_DATASET_URI ",
    "hyperParameters": {
      "epochCount": "EPOCH_COUNT ",
      "adapterSize": "ADAPTER_SIZE ",
      "learningRateMultiplier": "LEARNING_RATE_MULTIPLIER "
    },
    "export_last_checkpoint_only": EXPORT_LAST_CHECKPOINT_ONLY ,
  },
  "tunedModelDisplayName": "TUNED_MODEL_DISPLAYNAME ",
  "encryptionSpec": {
    "kmsKeyName": "KMS_KEY_NAME "
  },
  "serviceAccount": "SERVICE_ACCOUNT "
}
```

To send your request, choose one of these options:

**curlPowerShell** (#powershell)  
(#curl)

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **gcloud init** (/sdk/gcloud/reference/init) or **gcloud auth login** (/sdk/gcloud/reference/auth/login) , or by using **Cloud Shell** (/shell/docs), which automatically logs you into the **gcloud** CLI . You can check the currently active account by running **gcloud auth list** (/sdk/gcloud/reference/auth/list).

Save the request body in a file named **request.json**, and execute the following command:

```
curl -X POST \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json; charset=utf-8" \
  -d @request.json \
  "https://TUNING_JOB_REGION -aiplatform.googleapis.com/v1/projects/
```



You should receive a JSON response similar to the following.

### + Response

```
{
  "name": "projects/PROJECT_ID/locations/TUNING_JOB_REGION/tuningJobs/TUNING_JOB_ID",
  "createTime": CREATE_TIME,
  "updateTime": UPDATE_TIME,
  "status": "STATUS",
  "supervisedTuningSpec": {
    "trainingDatasetUri": "TRAINING_DATASET_URI",
    "validationDatasetUri": "VALIDATION_DATASET_URI",
    "hyperParameters": {
      "epochCount": EPOCH_COUNT,
      "adapterSize": "ADAPTER_SIZE",
      "learningRateMultiplier": LEARNING_RATE_MULTIPLIER
    },
  },
  "tunedModelDisplayName": "TUNED_MODEL_DISPLAYNAME",
  "encryptionSpec": {
    "kmsKeyName": "KMS_KEY_NAME"
  },
  "serviceAccount": "SERVICE_ACCOUNT"
}
```



## List tuning Jobs

You can view a list of tuning jobs in your current project by using the Vertex AI SDK for Python or by sending a GET request.

### RESTPython (#python) (#rest)

To create a model tuning job, send a POST request by using the [tuningJobs.create](#) (/vertex-ai/docs/reference/rest/v1/projects.locations.tuningJobs/create) method. Note that some of the parameters are not supported by all of the models. Ensure that you only include the applicable parameters for the model that you're tuning.

Before using any of the request data, make the following replacements:

- **PROJECT\_ID** : Your project ID  
(/resource-manager/docs/creating-managing-projects#identifiers).
- **TUNING\_JOB\_REGION** : The region  
(/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-region-settings) where the tuning job runs. This is also the default region for where the tuned model is uploaded.

HTTP method and URL:

GET [https://TUNING\\_JOB\\_REGION-aiplatform.googleapis.com/v1/projects/PROJECT\\_ID](https://TUNING_JOB_REGION-aiplatform.googleapis.com/v1/projects/PROJECT_ID).

To send your request, choose one of these options:

**curlPowerShell** (#powershell)  
(#curl)

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **gcloud init** (/sdk/gcloud/reference/init) or **gcloud auth login** (/sdk/gcloud/reference/auth/login) , or by using **Cloud Shell** (/shell/docs), which automatically logs you into the **gcloud** CLI . You can check the currently active account by running **gcloud auth list** (/sdk/gcloud/reference/auth/list).

Execute the following command:

```
curl -X GET \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  "https://TUNING_JOB_REGION-aiplatform.googleapis.com/v1/projects/PROJECT_ID"
```

You should receive a JSON response similar to the following.

#### Response

```
{
  "tuning_jobs": [
```

```

    TUNING_JOB_1, TUNING_JOB_2, ...
  ]
}

```




## Get details of a tuning job

You can get the details of a tuning job by using the Vertex AI SDK for Python or by sending a GET request.

**REST**Python (#python)  
(#rest)

To view a list of model tuning jobs, send a GET request by using the [tuningJobs.get](#) (/vertex-ai/docs/reference/rest/v1/projects.locations.tuningJobs/get) method and specify the **TuningJob\_ID**.

Before using any of the request data, make the following replacements:

- **PROJECT\_ID** : Your [project ID](#) (/resource-manager/docs/creating-managing-projects#identifiers).
- **TUNING\_JOB\_REGION** : The [region](#) (/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-region-settings) where the tuning job runs. This is also the default region for where the tuned model is uploaded.
- **TUNING\_JOB\_ID** : The ID of the tuning job.

HTTP method and URL:

GET [https://TUNING\\_JOB\\_REGION-aiplatform.googleapis.com/v1/projects/PROJECT\\_](https://TUNING_JOB_REGION-aiplatform.googleapis.com/v1/projects/PROJECT_)

To send your request, choose one of these options:

**curl**PowerShell (#powershell)  
(#curl)

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **gcloud init** (/sdk/gcloud/reference/init) or **gcloud auth login** (/sdk/gcloud/reference/auth/login) , or by using [Cloud Shell](#) (/shell/docs), which automatically

logs you into the **gcloud** CLI . You can check the currently active account by running **gcloud auth list** (/sdk/gcloud/reference/auth/list).

Execute the following command:

```
curl -X GET \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  "https://TUNING_JOB_REGION-aiplatform.googleapis.com/v1/projects/P
```

You should receive a JSON response similar to the following.

## + Response

```
{
  "name": "projects/PROJECT_ID/locations/TUNING_JOB_REGION/tuningJobs/TUNING_JOB_ID",
  "tunedModelDisplayName": "TUNED_MODEL_DISPLAYNAME",
  "createTime": CREATE_TIME,
  "endTime": END_TIME,
  "tunedModel": {
    "model": "projects/PROJECT_ID/locations/TUNING_JOB_REGION/models/MODEL_ID",
    "endpoint": "projects/PROJECT_ID/locations/TUNING_JOB_REGION/endpoints/ENDPOINT_ID"
  },
  "experiment": "projects/PROJECT_ID/locations/TUNING_JOB_REGION/metadataStores/METADATA_STORE_ID",
  "tuning_data_statistics": {
    "supervisedTuningDataStats": {
      "tuningDatasetExampleCount": "TUNING_DATASET_EXAMPLE_COUNT",
      "totalTuningCharacterCount": "TOTAL_TUNING_CHARACTER_COUNT",
      "tuningStepCount": "TUNING_STEP_COUNT"
    }
  },
  "status": "STATUS",
  "supervisedTuningSpec": {
    "trainingDatasetUri": "TRAINING_DATASET_URI",
    "validationDatasetUri": "VALIDATION_DATASET_URI",
    "hyperParameters": {
      "epochCount": EPOCH_COUNT,
      "learningRateMultiplier": LEARNING_RATE_MULTIPLIER
    }
  }
}
```

```
}
}
```




## Cancel a tuning job

You can cancel a tuning job by using the Vertex AI SDK for Python or by sending a POST request.

### RESTPython (#python) (#rest)

To view a list of model tuning jobs, send a GET request by using the [tuningJobs.cancel](#) (/vertex-ai/docs/reference/rest/v1/projects.locations.tuningJobs/cancel) method and specify the **TuningJob\_ID**.

Before using any of the request data, make the following replacements:

- **PROJECT\_ID** : Your project ID (/resource-manager/docs/creating-managing-projects#identifiers).
- **TUNING\_JOB\_REGION** : The region (/vertex-ai/generative-ai/docs/models/gemini-supervised-tuning-region-settings) where the tuning job runs. This is also the default region for where the tuned model is uploaded.
- **TUNING\_JOB\_ID** : The ID of the tuning job.

HTTP method and URL:

POST [https://TUNING\\_JOB\\_REGION-aiplatform.googleapis.com/v1/projects/PROJECT](https://<u>TUNING_JOB_REGION</u>-aiplatform.googleapis.com/v1/projects/<u>PROJECT</u>)

To send your request, choose one of these options:

### curlPowerShell (#powershell) (#curl)

★ **Note:** The following command assumes that you have logged in to the **gcloud** CLI with your user account by running **gcloud init** (/sdk/gcloud/reference/init) or **gcloud auth login** (/sdk/gcloud/reference/auth/login) , or by using Cloud Shell (/shell/docs), which automatically logs you into the **gcloud** CLI . You can check the currently active account by running **gcloud auth list** (/sdk/gcloud/reference/auth/list).

Execute the following command:

```
curl -X POST \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json; charset=utf-8" \
  -d "" \
  "https://TUNING_JOB_REGION -aiplatform.googleapis.com/v1/projects/
```

You should receive a JSON response similar to the following.

#### Response

```
{}
```

## What's next

For detailed documentation, see the following:

- [Supervised Tuning Job](#)  
(/vertex-ai/generative-ai/docs/models/gemini-use-supervised-tuning#create\_a\_text\_model\_supervised\_tuning\_job)
- [Gemini API](#) (/vertex-ai/generative-ai/docs/model-reference/gemini)

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/) (https://creativecommons.org/licenses/by/4.0/), and code samples are licensed under the [Apache 2.0 License](https://www.apache.org/licenses/LICENSE-2.0) (https://www.apache.org/licenses/LICENSE-2.0). For details, see the [Google Developers Site Policies](https://developers.google.com/site-policies) (https://developers.google.com/site-policies). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2025-06-06 UTC.