


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).

Get batch predictions for Gemini

Batch predictions let you send a large number of multimodal prompts in a single batch request.

For more information about the batch workflow and how to format your input data, see [Get batch predictions for Gemini](#) (/vertex-ai/generative-ai/docs/multimodal/batch-prediction-gemini).

Supported models

- [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)

Example syntax

The following example shows how to send a batch prediction API request using the `curl` command. This example is specific to BigQuery storage.

```
curl -X POST \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json" \
  https://$${LOCATION}-aiplatform.googleapis.com/v1/projects/$${PROJECT_ID}/locations/$
  -d '{
    "displayName": "...",
    "model": "publishers/google/models/$${MODEL_ID}",
    "inputConfig": {
      "instancesFormat": "bigquery",
      "bigquerySource": {
        "inputUri" : "..."
      }
    }
  }
```

```
    },
    "outputConfig": {
      "predictionsFormat": "bigquery",
      "bigqueryDestination": {
        "outputUri": "...
      }
    }
  }
}
```

Parameters

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

Body request

Parameters	
displayname	A name you choose for your job.
model	The model to use for batch prediction.
inputConfig	The data format. For Gemini batch prediction, Cloud Storage and BigQuery input sources are supported.
outputConfig	The output configuration which determines model output location. Cloud Storage and BigQuery output locations are supported.

inputConfig

Parameters	
instancesFormat	The prompt input format. Use jsonl for Cloud Storage or bigquery for BigQuery.
gcsSource.uri	The input source URI. This is a Cloud Storage location of the JSONL file in the form gs://bucketname/path/to/file.jsonl .
bigquerySource.inputUri	The input source URI. This is a BigQuery table URI in the form bq://project_id.dataset.table . The region of the input BigQuery dataset must be the same as the Vertex AI batch prediction job.

outputConfig

Parameters	
predictionsFormat	The output format of the prediction. Use bigquery .
gcsDestination.outputUriPrefix	The Cloud Storage bucket and directory location, in the form gs://mybucket/path/to/output .
bigqueryDestination.outputUri	The BigQuery URI of the target output table, in the form bq://project_id.dataset.table . If the table doesn't already exist, then it is created for you. The region of the output BigQuery dataset must be the same as the Vertex AI batch prediction job.

Examples

Request a batch response

Batch requests for multimodal models accept Cloud Storage storage and BigQuery storage sources. To learn more, see the following:

- [Batch request input format details](#)
(/vertex-ai/generative-ai/docs/multimodal/batch-prediction-gemini#prepare_your_inputs)

Depending on the number of input items that you submitted, a batch generation task can take some time to complete.



RESTGen AI SDK for Python...
(#rest)










Node.js (#node.js)Java (#java)Go (#go)

To create a batch prediction job, use the **`projects.locations.batchPredictionJobs.create`** (/vertex-ai/docs/reference/rest/v1/projects.locations.batchPredictionJobs/create) method.







Cloud Storage inputBigQuery input...
(#cloud-storage-input)

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

- **`LOCATION`** : A region that supports Gemini models.
- **`PROJECT_ID`** : Your project ID (/resource-manager/docs/creating-managing-projects#identifiers).

- **MODEL_PATH** : the publisher model name, for example, publishers/google/models/gemini-2.0-flash-001; or the tuned endpoint name, for example, projects/**PROJECT_ID** /locations/**LOCATION** /models/**MODEL_ID** , where **MODEL_ID** is the model ID of the tuned model.
- **INPUT_URI** : The Cloud Storage location of your JSONL batch prediction input such as gs://bucketname/path/to/file.jsonl.
- **OUTPUT_FORMAT** : To output to a BigQuery table, specify bigquery. To output to a Cloud Storage bucket, specify jsonl.
- **DESTINATION** : For BigQuery, specify bigqueryDestination. For Cloud Storage, specify gcsDestination.
- **OUTPUT_URI_FIELD_NAME** : For BigQuery, specify outputUri. For Cloud Storage, specify outputUriPrefix.
- **OUTPUT_URI** : For BigQuery, specify the table location such as bq://myproject.mydataset.output_result. The region of the output BigQuery dataset must be the same as the Vertex AI batch prediction job. For Cloud Storage, specify the bucket and directory location such as gs://mybucket/path/to/output.

Request JSON body:

```
{
  "displayName": "my-cloud-storage-batch-prediction-job",
  "model": "MODEL_PATH INPUT_URI OUTPUT_FORMAT DESTINATION OUTPUT_URI_FIELD_NAME OUTPUT_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://LOCATION -aiplatform.googleapis.com/v1/projects/PROJ
```

You should receive a JSON response similar to the following.

+ Response

```
{
  "name": "projects/PROJECT_ID/locations/LOCATION/batchPredictionJobs/BATCH_PREDICTION_JOB_ID",
  "displayName": "my-cloud-storage-batch-prediction-job",
  "model": "publishers/google/models/gemini-2.0-flash-001",
  "inputConfig": {
    "instancesFormat": "jsonl",
    "gcsSource": {
      "uris": [
        "INPUT_URI"
      ]
    }
  },
  "outputConfig": {
    "predictionsFormat": "OUTPUT_FORMAT",
```

```

    "DESTINATION": {
      "OUTPUT_URI_FIELD_NAME": "OUTPUT_URI"
    },
    "state": "JOB_STATE_PENDING",
    "createTime": "2024-10-16T19:33:59.153782Z",
    "updateTime": "2024-10-16T19:33:59.153782Z",
    "modelVersionId": "1"
  }

```

The response includes a unique identifier for the batch job. You can poll for the status of the batch job using the **BATCH_JOB_ID** until the job state is **JOB_STATE_SUCCEEDED**. For example:

```

curl -X GET \
  -H "Authorization: Bearer $(gcloud auth print-access-token)" \
  -H "Content-Type: application/json" \
  https://us-central1-aiplatform.googleapis.com/v1/projects/PROJECT_ID/locations/

```

Note: The upper limit for concurrent batch jobs is eight per region (/vertex-ai/docs/quotas#concurrent-batch-prediction-request-limits). Custom service accounts and CMEK aren't supported.

Retrieve batch output

When a batch prediction task completes, the output is stored in the Cloud Storage bucket or the BigQuery table that you specified in your request.

What's next

- Learn how to tune a Gemini model in Overview of model tuning for Gemini (/vertex-ai/generative-ai/docs/models/tune-gemini-overview).
- Learn more about how to Get batch predictions for Gemini (/vertex-ai/generative-ai/docs/multimodal/batch-prediction-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.