

The [Interactions API](/gemini-api/docs/interactions-overview) (/gemini-api/docs/interactions-overview) is now generally available. We recommend using this API for access to all the latest features and models.

Batch API

The Gemini API supports batch APIs, which let you process multiple requests in a single call. For more details, see the [Batch API guide](https://ai.google.dev/gemini-api/docs/batch-api) (https://ai.google.dev/gemini-api/docs/batch-api).

Method: models.batchGenerateContent

Enqueues a batch of `models.generateContent` requests for batch processing.

Endpoint

POST `https://generativelanguage.googleapis.com/v1beta/{batch.model=models/*}:batchGenerateContent`

Path parameters

`batch.model` string

Required. The name of the Model to use for generating the completion.
Format: `models/{model}`. It takes the form `models/{model}`.

Request body

The request body contains data with the following structure:

Fields

`batch.name` string

Output only. Identifier. Resource name of the batch.
Format: `batches/{batchId}`.

`batch.displayName` string

Required. The user-defined name of this batch.

batch.inputConfig object ([InputConfig](#) (/api/batch-api#InputConfig))

Required. Input configuration of the instances on which batch processing are performed.

batch.output object ([GenerateContentBatchOutput](#) (/api/batch-api#GenerateContentBatchOutput))

Output only. The output of the batch request.

batch.createTime

string ([Timestamp](#) (<https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp>) format)

Output only. The time at which the batch was created.

Uses RFC 3339, where generated output will always be Z-normalized and use 0, 3, 6 or 9 fractional digits. Offsets other than "Z" are also accepted. Examples: "2014-10-02T15:01:23Z", "2014-10-02T15:01:23.045123456Z" or "2014-10-02T15:01:23+05:30".

batch.endTime

string ([Timestamp](#) (<https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp>) format)

Output only. The time at which the batch processing completed.

Uses RFC 3339, where generated output will always be Z-normalized and use 0, 3, 6 or 9 fractional digits. Offsets other than "Z" are also accepted. Examples: "2014-10-02T15:01:23Z", "2014-10-02T15:01:23.045123456Z" or "2014-10-02T15:01:23+05:30".

batch.updateTime

string ([Timestamp](#) (<https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp>) format)

Output only. The time at which the batch was last updated.

Uses RFC 3339, where generated output will always be Z-normalized and use 0, 3, 6 or 9 fractional digits. Offsets other than "Z" are also accepted. Examples: "2014-10-02T15:01:23Z", "2014-10-02T15:01:23.045123456Z" or "2014-10-02T15:01:23+05:30".

batch.batchStats object ([BatchStats](#) (/api/batch-api#BatchStats))

Output only. Stats about the batch.

batch.state enum ([BatchState](#) (/api/batch-api#v1beta.BatchState))

Output only. The state of the batch.

batch.priority string ([int64](#) (<https://developers.google.com/discovery/v1/type-format>) format)

Optional. The priority of the batch. Batches with a higher priority value will be processed before batches with a lower priority value. Negative values are allowed. Default is 0.

Response body

If successful, the response body contains an instance of [Operation](#) (/api/batch-api#Operation).

Method: models.asyncBatchEmbedContent

Enqueues a batch of `models.embedContent` requests for batch processing. We have a `models.batchEmbedContents` handler in `GenerativeService`, but it was synchronized. So we name this one to be `Async` to avoid confusion.

Endpoint

```
POST https://generativelanguage.googleapis.com/v1beta/{batch.model=models/*}:asyncBatchEmbedContent
```

Path parameters

`batch.model` string

Required. The name of the `Model` to use for generating the completion.
Format: `models/{model}`. It takes the form `models/{model}`.

Request body

The request body contains data with the following structure:

Fields

`batch.name` string

Output only. Identifier. Resource name of the batch.
Format: `batches/{batchId}`.

`batch.displayName` string

Required. The user-defined name of this batch.

`batch.inputConfig` object ([InputEmbedContentConfig](#) (/api/embeddings#InputEmbedContentConfig))

Required. Input configuration of the instances on which batch processing are performed.

batch.output object ([EmbedContentBatchOutput](#) (/api/embeddings#EmbedContentBatchOutput))

Output only. The output of the batch request.

batch.createTime

string ([Timestamp](#) (https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp) format)

Output only. The time at which the batch was created.

Uses RFC 3339, where generated output will always be Z-normalized and use 0, 3, 6 or 9 fractional digits. Offsets other than "Z" are also accepted. Examples: "2014-10-02T15:01:23Z", "2014-10-02T15:01:23.045123456Z" or "2014-10-02T15:01:23+05:30".

batch.endTime

string ([Timestamp](#) (https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp) format)

Output only. The time at which the batch processing completed.

Uses RFC 3339, where generated output will always be Z-normalized and use 0, 3, 6 or 9 fractional digits. Offsets other than "Z" are also accepted. Examples: "2014-10-02T15:01:23Z", "2014-10-02T15:01:23.045123456Z" or "2014-10-02T15:01:23+05:30".

batch.updateTime

string ([Timestamp](#) (https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp) format)

Output only. The time at which the batch was last updated.

Uses RFC 3339, where generated output will always be Z-normalized and use 0, 3, 6 or 9 fractional digits. Offsets other than "Z" are also accepted. Examples: "2014-10-02T15:01:23Z", "2014-10-02T15:01:23.045123456Z" or "2014-10-02T15:01:23+05:30".

batch.batchStats object ([EmbedContentBatchStats](#) (/api/embeddings#EmbedContentBatchStats))

Output only. Stats about the batch.

batch.state enum ([BatchState](#) (/api/batch-api#v1beta.BatchState))

Output only. The state of the batch.

batch.priority string ([int64](#) (https://developers.google.com/discovery/v1/type-format) format)

Optional. The priority of the batch. Batches with a higher priority value will be processed before batches with a lower priority value. Negative values are allowed. Default is 0.

Response body

If successful, the response body contains an instance of [Operation](#) (/api/batch-api#Operation).

Method: batches.get

Gets the latest state of a long-running operation. Clients can use this method to poll the operation result at intervals as recommended by the API service.

Endpoint

```
GET https://generativelanguage.googleapis.com/v1beta/{name=batches/*}
```

Path parameters

name string

The name of the operation resource. It takes the form `batches/{batches}`.

Request body

The request body must be empty.

Response body

If successful, the response body contains an instance of **Operation** (/api/batch-api#Operation).

Method: batches.list

Lists operations that match the specified filter in the request. If the server doesn't support this method, it returns **UNIMPLEMENTED**.

Endpoint

```
GET https://generativelanguage.googleapis.com/v1beta/{name=batches}
```

Path parameters

name string

The name of the operation's parent resource. It takes the form `batches`.

Query parameters

filter string

The standard list filter.

pageSize integer

The standard list page size.

pageToken string

The standard list page token.

returnPartialSuccess boolean

When set to `true`, operations that are reachable are returned as normal, and those that are unreachable are returned in the `ListOperationsResponse.unreachable`

(`/api/rest/v1beta/ListOperationsResponse#FIELDS.unreachable`) field.

This can only be `true` when reading across collections. For example, when `parent` is set to `"projects/example/locations/-"`.

This field is not supported by default and will result in an `UNIMPLEMENTED` error if set unless explicitly documented otherwise in service or product specific documentation.

Request body

The request body must be empty.

Response body

If successful, the response body contains an instance of `ListOperationsResponse`.

Method: `batches.cancel`

Starts asynchronous cancellation on a long-running operation. The server makes a best effort to cancel the operation, but success is not guaranteed. If the server doesn't support this method, it returns `google.rpc.Code.UNIMPLEMENTED`. Clients can use `Operations.GetOperation` or other

methods to check whether the cancellation succeeded or whether the operation completed despite cancellation. On successful cancellation, the operation is not deleted; instead, it becomes an operation with an `Operation.error` (/api/batch-api#Operation.FIELDS.error) value with a `google.rpc.Status.code` (/api/files#FIELDS.code) of 1, corresponding to `Code.CANCELLED`.

Endpoint

```
POST https://generativelanguage.googleapis.com/v1beta/{name=batches/*}:cancel
```

Path parameters

`name` string

The name of the operation resource to be cancelled. It takes the form `batches/{batches}`.

Request body

The request body must be empty.

Response body

If successful, the response body is an empty JSON object.

Method: batches.delete

Deletes a long-running operation. This method indicates that the client is no longer interested in the operation result. It does not cancel the operation. If the server doesn't support this method, it returns `google.rpc.Code.UNIMPLEMENTED`.

Endpoint

```
DELETE https://generativelanguage.googleapis.com/v1beta/{name=batches/*}
```

Path parameters

name string

The name of the operation resource to be deleted. It takes the form `batches/{batches}`.

Request body

The request body must be empty.

Response body

If successful, the response body is an empty JSON object.

Method: batches.updateGenerateContentBatch

Updates a batch of GenerateContent requests for batch processing.

Endpoint

```
PATCH https://generativelanguage.googleapis.com/v1beta/{generateContentBatch.name=batches/*}:updateGenerateContentBatch
```

PATCH `https://generativelanguage.googleapis.com/v1beta/{generateContentBatch.name=batches/*}:updateGenerateContentBatch`

Path parameters

generateContentBatch.name string

Output only. Identifier. Resource name of the batch.

Format: `batches/{batchId}`. It takes the form `batches/{batches}`.

Query parameters

updateMask

string ([FieldMask](https://protobuf.dev/reference/protobuf/google.protobuf/#field-mask) (`https://protobuf.dev/reference/protobuf/google.protobuf/#field-mask`) format)

Optional. The list of fields to update.

This is a comma-separated list of fully qualified names of fields. Example: `"user.displayName,photo"`.

Request body

The request body contains an instance of [GenerateContentBatch](#) (/api/batch-api#v1beta.GenerateContentBatch).

Fields

model string

Required. The name of the **Model** to use for generating the completion.
Format: `models/{model}`.

displayName string

Required. The user-defined name of this batch.

inputConfig object ([InputConfig](#) (/api/batch-api#InputConfig))

Required. Input configuration of the instances on which batch processing are performed.

priority string ([int64](#) (<https://developers.google.com/discovery/v1/type-format>) format)

Optional. The priority of the batch. Batches with a higher priority value will be processed before batches with a lower priority value. Negative values are allowed. Default is 0.

Response body

If successful, the response body contains an instance of [GenerateContentBatch](#) (/api/batch-api#v1beta.GenerateContentBatch).

GenerateContentRequest

Request to generate a completion from the model.

Fields

model string

Required. The name of the **Model** to use for generating the completion.
Format: `models/{model}`.

contents[] object ([Content](#) (/api/caching#Content))

Required. The content of the current conversation with the model.

For single-turn queries, this is a single instance. For multi-turn queries like [chat](https://ai.google.dev/gemini-api/docs/text-generation#chat) (<https://ai.google.dev/gemini-api/docs/text-generation#chat>), this is a repeated field that contains the conversation history and the latest request.

tools[] object ([Tool](#) (/api/caching#Tool))

Optional. A list of **Tools** the **Model** may use to generate the next response.

A **Tool** is a piece of code that enables the system to interact with external systems to perform an action, or set of actions, outside of knowledge and scope of the **Model**. Supported **Tools** are **Function** and **codeExecution**. Refer to the [Function calling](https://ai.google.dev/gemini-api/docs/function-calling) (<https://ai.google.dev/gemini-api/docs/function-calling>) and the [Code execution](https://ai.google.dev/gemini-api/docs/code-execution) (<https://ai.google.dev/gemini-api/docs/code-execution>) guides to learn more.

toolConfig object ([ToolConfig](#) (/api/caching#ToolConfig))

Optional. Tool configuration for any **Tool** specified in the request. Refer to the [Function calling guide](https://ai.google.dev/gemini-api/docs/function-calling#function_calling_mode) (https://ai.google.dev/gemini-api/docs/function-calling#function_calling_mode) for a usage example.

safetySettings[] object ([SafetySetting](#) (/api/generate-content#v1beta.SafetySetting))

Optional. A list of unique **SafetySetting** instances for blocking unsafe content.

This will be enforced on the **GenerateContentRequest.contents** and **GenerateContentResponse.candidates**. There should not be more than one setting for each **SafetyCategory** type. The API will block any contents and responses that fail to meet the thresholds set by these settings. This list overrides the default settings for each **SafetyCategory** specified in the **safetySettings**. If there is no **SafetySetting** for a given **SafetyCategory** provided in the list, the API will use the default safety setting for that category. Harm categories **HARM_CATEGORY_HATE_SPEECH**, **HARM_CATEGORY_SEXUALLY_EXPLICIT**, **HARM_CATEGORY_DANGEROUS_CONTENT**, **HARM_CATEGORY_HARASSMENT**, **HARM_CATEGORY_CIVIC_INTEGRITY** are supported. Refer to the [guide](https://ai.google.dev/gemini-api/docs/safety-settings) (<https://ai.google.dev/gemini-api/docs/safety-settings>) for detailed information on available safety settings. Also refer to the [Safety guidance](https://ai.google.dev/gemini-api/docs/safety-guidance) (<https://ai.google.dev/gemini-api/docs/safety-guidance>) to learn how to incorporate safety considerations in your AI applications.

systemInstruction object ([Content](#) (/api/caching#Content))

Optional. Developer set [system instruction\(s\)](https://ai.google.dev/gemini-api/docs/system-instructions) (<https://ai.google.dev/gemini-api/docs/system-instructions>). Currently, text only.

generationConfig object ([GenerationConfig](#) (/api/generate-content#v1beta.GenerationConfig))

Optional. Configuration options for model generation and outputs.

cachedContent string

Optional. The name of the content [cached](https://ai.google.dev/gemini-api/docs/caching) (<https://ai.google.dev/gemini-api/docs/caching>) to use as context to serve the prediction. Format: **cachedContents/{cachedContent}**

serviceTier enum ([ServiceTier](#) (/api/generate-content#v1beta.ServiceTier))

Optional. The service tier of the request.

store boolean

Optional. Configures the logging behavior for a given request. If set, it takes precedence over the project-level logging config.

JSON representation

```
{
  "model": string,
  "contents": [
    {
      object (Content (/api/caching#Content))
    }
  ],
  "tools": [
    {
      object (Tool (/api/caching#Tool))
    }
  ],
  "toolConfig": {
    object (ToolConfig (/api/caching#ToolConfig))
  },
  "safetySettings": [
    {
      object (SafetySetting (/api/generate-content#v1beta.SafetySetting))
    }
  ],
  "systemInstruction": {
    object (Content (/api/caching#Content))
  },
  "generationConfig": {
    object (GenerationConfig (/api/generate-content#v1beta.GenerationConfig))
  },
  "cachedContent": string,
  "serviceTier": enum (ServiceTier (/api/generate-content#v1beta.ServiceTier)),
  "store": boolean
}
```

Method: batches.updateEmbedContentBatch

Updates a batch of EmbedContent requests for batch processing.

Endpoint

PATCH `https://generativelanguage.googleapis.com/v1beta/{embedContentBatch.name=batches/*}:updateEmbedContentBatch`

PATCH `https://generativelanguage.googleapis.com/v1beta/{embedContentBatch.name=batches/*}:updateEmbedContentBatch`

Path parameters

embedContentBatch.name string

Output only. Identifier. Resource name of the batch.

Format: `batches/{batchId}`. It takes the form `batches/{batches}`.

Query parameters

updateMask

string ([FieldMask](https://protobuf.dev/reference/protobuf/google.protobuf/#field-mask) (<https://protobuf.dev/reference/protobuf/google.protobuf/#field-mask>) format)

Optional. The list of fields to update.

This is a comma-separated list of fully qualified names of fields. Example: `"user.displayName,photo"`.

Request body

The request body contains an instance of [EmbedContentBatch](#) (`/api/embeddings#v1beta.EmbedContentBatch`).

Fields

model string

Required. The name of the **Model** to use for generating the completion.

Format: `models/{model}`.

displayName string

Required. The user-defined name of this batch.

inputConfig object ([InputEmbedContentConfig](#) (`/api/embeddings#InputEmbedContentConfig`))

Required. Input configuration of the instances on which batch processing are performed.

priority string ([int64](https://developers.google.com/discovery/v1/type-format) (https://developers.google.com/discovery/v1/type-format) format)

Optional. The priority of the batch. Batches with a higher priority value will be processed before batches with a lower priority value. Negative values are allowed. Default is 0.

Response body

If successful, the response body contains an instance of [EmbedContentBatch](#) (/api/embeddings#v1beta.EmbedContentBatch).

EmbedContentRequest

Request containing the Content for the model to embed.

Fields

model string

Required. The model's resource name. This serves as an ID for the Model to use. This name should match a model name returned by the `ListModels` method.
Format: `models/{model}`

content object ([Content](#) (/api/caching#Content))

Required. The content to embed. Only the `parts.text` fields will be counted.

taskType

(deprecated) enum ([TaskType](#) (/api/embeddings#v1beta.TaskType))

This item is deprecated!

Optional. Deprecated: Please use `EmbedContentConfig.task_type` instead. Optional task type for which the embeddings will be used. Not supported on earlier models (`models/embedding-001`).

title

(deprecated) string

This item is deprecated!

Optional. Deprecate: Please use EmbedContentConfig.title instead. An optional title for the text. Only applicable when TaskType is RETRIEVAL_DOCUMENT.

Note: Specifying a title for RETRIEVAL_DOCUMENT provides better quality embeddings for retrieval.

outputDimensionality
(**deprecate**) integer

This item is deprecate!

Optional. Deprecate: Please use EmbedContentConfig.output_dimensionality instead. Optional reduced dimension for the output embedding. If set, excessive values in the output embedding are truncated from the end. Supported by newer models since 2024 only. You cannot set this value if using the earlier model (models/embedding-001).

embedContentConfig object ([EmbedContentConfig](#) (/api/embeddings#v1beta.EmbedContentConfig))

Optional. Configuration for the EmbedContent request.

JSON representation

```
{
  "model": string,
  "content": {
    object (Content (/api/caching#Content))
  },
  "taskType": enum (TaskType (/api/embeddings#v1beta.TaskType)),
  "title": string,
  "outputDimensionality": integer,
  "embedContentConfig": {
    object (EmbedContentConfig (/api/embeddings#v1beta.EmbedContentConfig))
  }
}
```

REST Resource: batches

Resource: Operation

This resource represents a long-running operation that is the result of a network API call.

Fields

name string

The server-assigned name, which is only unique within the same service that originally returns it. If you use the default HTTP mapping, the **name** should be a resource name ending with `operations/{unique_id}`.

metadata object

Service-specific metadata associated with the operation. It typically contains progress information and common metadata such as create time. Some services might not provide such metadata. Any method that returns a long-running operation should document the metadata type, if any.

An object containing fields of an arbitrary type. An additional field `"@type"` contains a URI identifying the type.

Example: { `"id": 1234`, `"@type": "types.example.com/standard/id"` }.

done boolean

If the value is `false`, it means the operation is still in progress. If `true`, the operation is completed, and either `error` or `response` is available.

result Union type

The operation result, which can be either an `error` or a valid `response`. If `done == false`, neither `error` nor `response` is set. If `done == true`, exactly one of `error` or `response` can be set. Some services might not provide the result. **result** can be only one of the following:

error object ([Status](#) (/api/files#v1beta.Status))

The error result of the operation in case of failure or cancellation.

response object

The normal, successful response of the operation. If the original method returns no data on success, such as `Delete`, the response is `google.protobuf.Empty`. If the original method is standard `Get/Create/Update`, the response should be the resource. For other methods, the response should have the type `XxxResponse`, where `Xxx` is the original method name. For example, if the original method name is `TakeSnapshot()`, the inferred response type is `TakeSnapshotResponse`.

An object containing fields of an arbitrary type. An additional field `"@type"` contains a URI identifying the type. Example: { `"id": 1234`, `"@type": "types.example.com/standard/id"` }.

JSON representation

```
{
  "name": string,
  "metadata": {
    "@type": string,
    field1: ...,
    ...
  }
}
```

JSON representation

```
{,
  "done": boolean,

  // result
  "error": {
    object (Status (/api/files#v1beta.Status))
  },
  "response": {
    "@type": string,
    field1: ...,
    ...
  }
  // Union type
}
```

GenerateContentBatch

A resource representing a batch of `GenerateContent` requests.

Fields

model string

Required. The name of the `Model` to use for generating the completion.
Format: `models/{model}`.

name string

Output only. Identifier. Resource name of the batch.
Format: `batches/{batchId}`.

displayName string

Required. The user-defined name of this batch.

inputConfig object ([InputConfig](#) (/api/batch-api#InputConfig))

Required. Input configuration of the instances on which batch processing are performed.

output object ([GenerateContentBatchOutput](#) (/api/batch-api#GenerateContentBatchOutput))

Output only. The output of the batch request.

createTime

string ([Timestamp](https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp) (https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp) format)

Output only. The time at which the batch was created.

Uses RFC 3339, where generated output will always be Z-normalized and use 0, 3, 6 or 9 fractional digits. Offsets other than "Z" are also accepted. Examples: "2014-10-02T15:01:23Z", "2014-10-02T15:01:23.045123456Z" or "2014-10-02T15:01:23+05:30".

endTime

string ([Timestamp](https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp) (https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp) format)

Output only. The time at which the batch processing completed.

Uses RFC 3339, where generated output will always be Z-normalized and use 0, 3, 6 or 9 fractional digits. Offsets other than "Z" are also accepted. Examples: "2014-10-02T15:01:23Z", "2014-10-02T15:01:23.045123456Z" or "2014-10-02T15:01:23+05:30".

updateTime

string ([Timestamp](https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp) (https://protobuf.dev/reference/protobuf/google.protobuf/#timestamp) format)

Output only. The time at which the batch was last updated.

Uses RFC 3339, where generated output will always be Z-normalized and use 0, 3, 6 or 9 fractional digits. Offsets other than "Z" are also accepted. Examples: "2014-10-02T15:01:23Z", "2014-10-02T15:01:23.045123456Z" or "2014-10-02T15:01:23+05:30".

batchStats object ([BatchStats](#) (/api/batch-api#BatchStats))

Output only. Stats about the batch.

state enum ([BatchState](#) (/api/batch-api#v1beta.BatchState))

Output only. The state of the batch.

priority string ([int64](https://developers.google.com/discovery/v1/type-format) (https://developers.google.com/discovery/v1/type-format) format)

Optional. The priority of the batch. Batches with a higher priority value will be processed before batches with a lower priority value. Negative values are allowed. Default is 0.

JSON representation

```
{
  "model": string,
  "name": string,
  "displayName": string,
  "inputConfig": {
    object (InputConfig (/api/batch-api#InputConfig))
  },
  "output": {
```

JSON representation

```
    object (GenerateContentBatchOutput (/api/batch-api#GenerateContentBatchOutput))
  },
  "createTime": string,
  "endTime": string,
  "updateTime": string,
  "batchStats": {
    object (BatchStats (/api/batch-api#BatchStats))
  },
  "state": enum (BatchState (/api/batch-api#v1beta.BatchState)),
  "priority": string
}
```

InputConfig

Configures the input to the batch request.

Fields

Required. The source of the input. `source` can be only one of the following:

fileName string

The name of the File containing the input requests.

source Union type

requests object (InlinedRequests (/api/batch-api#InlinedRequests))

The requests to be processed in the batch.

JSON representation

```
{
  // source
  "fileName": string,
  "requests": {
    object (InlinedRequests (/api/batch-api#InlinedRequests))
  }
  // Union type
}
```

InlinedRequests

The requests to be processed in the batch if provided as part of the batch creation request.

Fields

requests[] object ([InlinedRequest](#) (/api/batch-api#InlinedRequest))

Required. The requests to be processed in the batch.

JSON representation

```
{
  "requests": [
    {
      object (InlinedRequest (/api/batch-api#InlinedRequest))
    }
  ]
}
```

InlinedRequest

The request to be processed in the batch.

Fields

request object ([GenerateContentRequest](#) (/api/batch-api#GenerateContentRequest))

Required. The request to be processed in the batch.

metadata object ([Struct](#) (https://protobuf.dev/reference/protobuf/google.protobuf/#struct) format)

Optional. The metadata to be associated with the request.

JSON representation

```
{
  "request": {
    object (GenerateContentRequest (/api/batch-api#GenerateContentRequest))
  },
  "metadata": {
    object
  }
}
```

JSON representation

```
}  
}
```

GenerateContentBatchOutput

The output of a batch request. This is returned in the `BatchGenerateContentResponse` or the `GenerateContentBatch.output` field.

Fields

output Union type

The output of the batch request. `output` can be only one of the following:

responsesFile string

Output only. The file ID of the file containing the responses. The file will be a JSONL file with a single response per line. The responses will be `GenerateContentResponse` messages formatted as JSON. The responses will be written in the same order as the input requests.

inlinedResponses object ([InlinedResponses](#) (/api/batch-api#InlinedResponses))

Output only. The responses to the requests in the batch. Returned when the batch was built using inlined requests. The responses will be in the same order as the input requests.

JSON representation

```
{  
  
  // output  
  "responsesFile": string,  
  "inlinedResponses": {  
    object (InlinedResponses (/api/batch-api#InlinedResponses))  
  }  
  // Union type  
}
```

InlinedResponses

The responses to the requests in the batch.

Fields

inlinedResponses[] object ([InlinedResponse](#) (/api/batch-api#InlinedResponse))

Output only. The responses to the requests in the batch.

JSON representation

```
{
  "inlinedResponses": [
    {
      object (InlinedResponse (/api/batch-api#InlinedResponse))
    }
  ]
}
```

InlinedResponse

The response to a single request in the batch.

Fields

metadata object ([Struct](#) (<https://protobuf.dev/reference/protobuf/google.protobuf/#struct>) format)

Output only. The metadata associated with the request.

output Union type

The output of the request. **output** can be only one of the following:

error object ([Status](#) (/api/files#v1beta.Status))

Output only. The error encountered while processing the request.

response

object ([GenerateContentResponse](#) (/api/generate-content#v1beta.GenerateContentResponse))

Output only. The response to the request.

JSON representation

```
{
  "metadata": {
    object
  },

  // output
  "error": {
    object (Status (/api/files#v1beta.Status))
  },
  "response": {
    object (GenerateContentResponse (/api/generate-content#v1beta.GenerateContentResponse))
  }
  // Union type
}
```

BatchStats

Stats about the batch.

Fields

requestCount string ([int64](#) (<https://developers.google.com/discovery/v1/type-format>) format)

Output only. The number of requests in the batch.

successfulRequestCount

string ([int64](#) (<https://developers.google.com/discovery/v1/type-format>) format)

Output only. The number of requests that were successfully processed.

failedRequestCount string ([int64](#) (<https://developers.google.com/discovery/v1/type-format>) format)

Output only. The number of requests that failed to be processed.

pendingRequestCount

string ([int64](#) (<https://developers.google.com/discovery/v1/type-format>) format)

Output only. The number of requests that are still pending processing.

JSON representation

```
{
  "requestCount": string,
  "successfulRequestCount": string,
  "failedRequestCount": string,
  "pendingRequestCount": string
}
```

BatchState

The state of the batch.

Enums

BATCH_STATE_UNSPECIFIED	The batch state is unspecified.
-------------------------	---------------------------------

BATCH_STATE_PENDING	The service is preparing to run the batch.
---------------------	--

BATCH_STATE_RUNNING	The batch is in progress.
---------------------	---------------------------

BATCH_STATE_SUCCEEDED	The batch completed successfully.
-----------------------	-----------------------------------

BATCH_STATE_FAILED	The batch failed.
--------------------	-------------------

BATCH_STATE_CANCELLED	The batch has been cancelled.
-----------------------	-------------------------------

BATCH_STATE_EXPIRED	The batch has expired.
---------------------	------------------------

ServiceTier

Service tier of the request.

Enums

unspecified	Default service tier, which is standard.
-------------	--

standard	Standard service tier.
----------	------------------------

Enums

flex	Flex service tier.
-------------	--------------------

priority	Priority service tier.
-----------------	------------------------

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 2026-05-20 UTC.