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).

CountTokens API

The CountTokens API calculates the number of input tokens before sending a request to the Gemini API.

Use the CountTokens API to prevent requests from exceeding the model context window, and estimate potential costs based on billable characters.

The CountTokens API can use the same `contents` parameter as Gemini API inference requests.

Supported models

- Gemini 2.0 Flash with image generation (/vertex-ai/generative-ai/docs/models/gemini/2-0-flash)
- Vertex AI Model Optimizer (/vertex-ai/generative-ai/docs/model-reference/vertex-ai-model-optimizer)
- 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)

Parameter list

This class consists of two main properties: `role` and `parts`. The `role` property denotes the individual producing the content, while the `parts` property contains multiple elements, each representing a segment of data within a message.

Parameters	
`role`	Optional: `string`

	The identity of the entity that creates the message. Set the string to one of the following: • `user`: This indicates that the message is sent by a real person. For example, a user-generated message. • `model`: This indicates that the message is generated by the model. The `model` value is used to insert messages from the model into the conversation during multi-turn conversations. For non-multi-turn conversations, this field can be left blank or unset.
`parts`	`part` A list of ordered parts that make up a single message. Different parts may have different IANA MIME types (https://www.iana.org/assignments/media-types/media-types.xml).

`Part`

A data type containing media that is part of a multi-part `Content` message.

Parameters

`text`	Optional: `string` A text prompt or code snippet.
`inline_data`	Optional: `Blob` Inline data in raw bytes.
`file_data`	Optional: `FileData` Data stored in a file.

`Blob`

Content blob. If possible this send as text rather than raw bytes.

Parameters

`mime_type`	`string` IANA MIME type (https://www.iana.org/assignments/media-types/media-types.xml) of the data.
`data`	`bytes` Raw bytes.

`FileData`

URI based data.

Parameters

`mime_type`	`string` IANA MIME type (https://www.iana.org/assignments/media-types/media-types.xml) of the data.
`file_uri`	`string` The Cloud Storage URI to the file storing the data.

`system_instruction`

This field is for user provided `system_instructions`. It is the same as `contents` but with a limited support of the content types.

Parameters

`role`	`string` IANA MIME type (https://www.iana.org/assignments/media-types/media-types.xml) of the data. This field is ignored internally.
`parts`	`Part` Text only. Instructions that users want to pass to the model.

`FunctionDeclaration`

A structured representation of a function declaration as defined by the OpenAPI 3.0 specification (https://spec.openapis.org/oas/v3.0.3) that represents a function the model may generate JSON inputs for.

Parameters	
`name`	`string` The name of the function to call.
`description`	Optional: `string` Description and purpose of the function.
`parameters`	Optional: `Schema` Describes the parameters of the function in the OpenAPI JSON Schema Object format: OpenAPI 3.0 specification (https://spec.openapis.org/oas/v3.0.3).
`response`	Optional: `Schema` Describes the output from the function in the OpenAPI JSON Schema Object format: OpenAPI 3.0 specification (https://spec.openapis.org/oas/v3.0.3).

Examples

Get token count from text prompt

This example counts the tokens of a single text prompt:

REST Gen AI SDK for Python... Gen AI SDK for Go... Gen AI SDK for Node.js... Gen AI SDK (#rest)

To get the token count and the number of billable characters for a prompt by using the Vertex AI API, send a `POST` request to the publisher model endpoint.

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

- ***LOCATION***: The region to process the request. Available options include the following:

  Click to expand a partial list of available regions

  - `us-central1`
  - `us-west4`

    - `northamerica-northeast1`
    - `us-east4`
    - `us-west1`
    - `asia-northeast3`
    - `asia-southeast1`
    - `asia-northeast1`
- ***PROJECT_ID***: Your project ID (/resource-manager/docs/creating-managing-projects#identifiers).
- ***MODEL_ID***: The model ID of the multimodal model that you want to use.
- ***ROLE***: The role in a conversation associated with the content. Specifying a role is required even in singleturn use cases. Acceptable values include the following:
    - `USER`: Specifies content that's sent by you.
- ***TEXT***: The text instructions to include in the prompt.

HTTP method and URL:

```
POST https://LOCATION-aiplatform.googleapis.com/v1/projects/PROJECT_ID/loc
```

Request JSON body:

```
{
  "contents": [{
    "role": "ROLE",
    "parts": [{
      "text": "TEXT"
    }]
  }]
}
```

To send your request, choose one of these options:

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/PROJECT_IL
```

You should receive a JSON response similar to the following.

Response

```
{ "totalTokens": 43 }
```

Get token count from media prompt

This example counts the tokens of a prompt that uses various media types.

RESTGen AI SDK for Python... Gen AI SDK for Go... Gen AI SDK for Node.js... Gen AI SDK (#rest)

To get the token count and the number of billable characters for a prompt by using the Vertex AI API, send a POST request to the publisher model endpoint.

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

- *LOCATION*: The region to process the request. Available options include the following:

  Click to expand a partial list of available regions

  - `us-central1`
  - `us-west4`
  - `northamerica-northeast1`
  - `us-east4`
  - `us-west1`
  - `asia-northeast3`
  - `asia-southeast1`
  - `asia-northeast1`

- *PROJECT_ID*: Your project ID (/resource-manager/docs/creating-managing-projects#identifiers).
- *MODEL_ID*: The model ID of the multimodal model that you want to use.
- *ROLE*: The role in a conversation associated with the content. Specifying a role is required even in singleturn use cases. Acceptable values include the following:
  - USER: Specifies content that's sent by you.
- *TEXT*: The text instructions to include in the prompt.
- *FILE_URI*: The URI or URL of the file to include in the prompt. Acceptable values include the following:
  - **Cloud Storage bucket URI:** The object must either be publicly readable or reside in the same Google Cloud project that's sending the request. For `gemini-2.0-flash` and `gemini-2.0-flash-lite`, the size limit is 2 GB.
  - **HTTP URL:** The file URL must be publicly readable. You can specify one video file, one audio file, and up to 10 image files per request. Audio files, video files, and documents can't exceed 15 MB.
  - **YouTube video URL:**The YouTube video must be either owned by the account that you used to sign in to the Google Cloud console or is public. Only one YouTube video URL is supported per request.

When specifying a `fileURI`, you must also specify the media type (`mimeType`) of the file. If VPC Service Controls is enabled, specifying a media file URL for `fileURI` is not supported.

- ***MIME_TYPE***: The media type of the file specified in the `data` or `fileUri` fields. Acceptable values include the following:

  Click to expand MIME types

  - `application/pdf`
  - `audio/mpeg`
  - `audio/mp3`
  - `audio/wav`
  - `image/png`
  - `image/jpeg`
  - `image/webp`
  - `text/plain`
  - `video/mov`
  - `video/mpeg`
  - `video/mp4`
  - `video/mpg`
  - `video/avi`
  - `video/wmv`
  - `video/mpegps`
  - `video/flv`

HTTP method and URL:

```
POST https://LOCATION-aiplatform.googleapis.com/v1/projects/PROJECT_ID/loc
```

Request JSON body:

```
{
  "contents": [{
    "role": "ROLE",
    "parts": [
      {
        "file_data": {
          "file_uri": "FILE_URI",
          "mime_type": "MIME_TYPE"
        }
      },
      {
        "text": "TEXT
      }
    ]
  }]
}
```

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/PROJECT_IL
```

You should receive a JSON response similar to the following.

Response

```
{ "totalTokens": 43 }
```

What's next

- Learn more about the Gemini API (/vertex-ai/generative-ai/docs/model-reference/gemini).

Last updated 2025-06-06 UTC.