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).

Method: publishers.models.countTokens

Perform a token count in express mode.

Endpoint

```
POST https://aiplatform.googleapis.com/v1beta1/{model}:countTokens
```

Path parameters

`model` `string`

Required. The name of the model requested to perform token counting. Format: `/publishers/google/models/*`

Request body

The request body contains data with the following structure:

Fields

`model` `string`

Optional. The name of the publisher model requested to serve the prediction. Format: `publishers/google/models/*`

`instances[]` `value` (`Value` (https://protobuf.dev/reference/protobuf/google.protobuf/#value) `format`)

Optional. The instances that are the input to token counting call. Schema is identical to the prediction schema of the underlying model.

`contents[]` `object` (`Content` (/vertex-ai/generative-ai/docs/reference/rest/v1beta1/Content))

Optional. Input content.

`tools[]`

`object (Tool (/vertex-ai/generative-ai/docs/reference/rest/v1beta1/projects.locations.cachedContents#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.

`systemInstruction` `object (Content (/vertex-ai/generative-ai/docs/reference/rest/v1beta1/Content))`

Optional. The user provided system instructions for the model. Note: only text should be used in parts and content in each part will be in a separate paragraph.

`generationConfig`

`object (GenerationConfig (/vertex-ai/generative-ai/docs/reference/rest/v1beta1/GenerationConfig))`

Optional. Generation config that the model will use to generate the response.

Response body

If successful, the response body contains an instance of `CountTokensResponse` (/vertex-ai/generative-ai/docs/reference/rest/v1beta1/CountTokensResponse).

Last updated 2024-12-17 UTC.