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.streamGenerateContent

Generate content with multimodal inputs in express mode with streaming support.

Endpoint

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

Path parameters

`model` `string`

Required. The fully qualified name of the publisher model to use.
Publisher model format: `publishers/google/models/*`

Request body

The request body contains data with the following structure:

Fields

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

Required. The content of the current conversation with the model.
For single-turn queries, this is a single instance. For multi-turn queries, this is a repeated field that contains conversation history + latest request.

`cachedContent` `string`

Optional. The name of the cached content used as context to serve the prediction. Note: only used in explicit caching, where users can have control over caching (e.g. what content to cache) and enjoy guaranteed cost savings. Format: `projects/{project}/locations/{location}/cachedContents/{cachedContent}` `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.

`toolConfig`

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

Optional. Tool config. This config is shared for all tools provided in the request.

`labels` `map (key: string, value: string)`

Optional. The labels with user-defined metadata for the request. It is used for billing and reporting only.
label keys and values can be no longer than 63 characters (Unicode codepoints) and can only contain lowercase letters, numeric characters, underscores, and dashes. International characters are allowed. label values are optional. label keys must start with a letter.

`safetySettings[]`

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

Optional. Per request settings for blocking unsafe content. Enforced on GenerateContentResponse.candidates.

`generationConfig`

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

Optional. Generation config.

`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.

Response body

If successful, the response body contains a stream of `GenerateContentResponse` (/vertex-ai/generative-ai/docs/reference/rest/v1beta1/GenerateContentResponse) instances.

Last updated 2024-12-17 UTC.