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).

Vertex AI in express mode REST API reference

Preview

This feature is subject to the "Pre-GA Offerings Terms" in the General Service Terms section of the Service Specific Terms (/terms/service-terms#1). Pre-GA features are available "as is" and might have limited support. For more information, see the launch stage descriptions (/products#product-launch-stages).

Vertex AI in express mode lets you try a subset of Vertex AI features by using only an express mode API key. This page shows you the REST resources available for Vertex AI in express mode.

Unlike the standard REST resource endpoints on Google Cloud, endpoints that are available when using Vertex AI in express mode use the global endpoint `aiplatform.googleapis.com` and don't include `projects` or `locations`. For example, the following shows the difference between standard and express mode endpoints for the datasets resource:

Standard Vertex AI endpoint format: `https://{location}-aiplatform.googleapis.com/v1/projects/{project}/locations/{location}/{model}:generateContent`

Endpoint format for Vertex AI in express mode: `https://aiplatform.googleapis.com/v1/{model}:generateContent`

REST Resource: v1.publishers.models

(/vertex-ai/generative-ai/docs/reference/express-mode/rest/v1/publishers.models)

Methods

`countTokens` (/vertex-ai/generative-ai/docs/reference/express-mode/rest/v1/publishers.models/countTokens)	`POST /v1/{endpoint}:countTokens` Perform a token counting.
`generateContent` (/vertex-ai/generative-ai/docs/reference/express-mode/rest/v1/publishers.models/generateContent)	`POST /v1/{model}:generateContent` Generate content with multimodal inputs.
`streamGenerateContent` (/vertex-ai/generative-ai/docs/reference/express-mode/rest/v1/publishers.models/streamGenerateContent)	`POST /v1/{model}:streamGenerateContent` Generate content with multimodal inputs with streaming support.

REST Resource: v1beta1.publishers.models

(/vertex-ai/generative-ai/docs/reference/express-mode/rest/v1beta1/publishers.models)

Methods

`countTokens` (/vertex-ai/generative-ai/docs/reference/express-mode/rest/v1beta1/publishers.models/countTokens)	`POST /v1beta1/{endpoint}:countTokens` Perform a token counting.
`generateContent` (/vertex-ai/generative-ai/docs/reference/express-mode/rest/v1beta1/publishers.models/generateContent)	`POST /v1beta1/{model}:generateContent` Generate content with multimodal inputs.
`streamGenerateContent` (/vertex-ai/generative-ai/docs/reference/express-mode/rest/v1beta1/publishers.models/streamGenerateContent)	`POST /v1beta1/{model}:streamGenerateContent` Generate content with multimodal inputs with streaming support.

Last updated 2025-06-05 UTC.