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).

Generative AI on Vertex AI inference API errors

This guide provides a list of errors that you might encounter from using the Model API reference for Generative AI (/vertex-ai/generative-ai/docs/model-reference/overview). The errors follow the error model (/apis/design/errors) of the Google Cloud API, which recommends that we provide guidance on the causes and the solutions specific to the generative AI models.

API errors

This table provides API error codes and descriptions.

HTTP error code	Canonical error code	Cause	Example	Solution
400	**`INVALID_ARGUMENT / FAILED_PRECONDITION`**	Request fails API validation, or you tried to access a model that requires allowlisting or is disallowed by the organization's policy.	Request exceeds the model's input token limit.	Refer to the Model API reference for Generative AI (/vertex-ai/generative-ai/docs/model-reference/overview) for request parameters, token count, and other parameters.
403	**`PERMISSION_DENIED`**	Client doesn't have sufficient permission to call the API.	Service account doesn't have permission to access the Cloud Storage bucket hosting image or video resources.	1. Verify that all necessary APIs are enabled, and the service account has the right permission (/vertex-ai/generative-ai/docs/access-control) to access the selected Vertex AI service. 2. Vertex AI per-product, per-project service account (P4SA) is granted the

HTTP error code	Canonical error code	Cause	Example	Solution
				necessary permission to access resources referenced in the input.
404	`NOT_FOUND`	No valid object is found from the designated URL.	Image file not found in the storage URL.	Check and fix the file location.
429	`RESOURCE_EXHAUSTED`	Depending on the error message, the error could be caused by the following: 1. API quota over the limit. 2. Server overload due to shared server capacity. 3. You've reached the daily limit for requests using `logprobs`.	Gemini API exceeds request per minute limit.	1. Check Vertex AI Generative AI quota limits (/vertex-ai/generative-ai/docs/quotas). If needed, apply for a higher quota. 2. Retry after a few seconds. If the error persists after a prolonged period of time (hours), contact Vertex AI support (/vertex-ai/docs/support/getting-support) .
499	`CANCELLED`	Request is cancelled by the client.		
500	`UNKNOWN / INTERNAL`	Server error due to overload or dependency failure.	Request is throttled, because the service is temporarily overloaded.	Retry after a few seconds. If the error persists after a prolonged period of time (hours), contact Vertex AI support (/vertex-ai/docs/support/getting-support) .
503	`UNAVAILABLE`	Service is temporarily unavailable.	Server isn't responding to the incoming requests.	The unavailable status might be temporary. However, if the error persists, contact Vertex AI support (/vertex-ai/docs/support/getting-support) .
504	`DEADLINE_EXCEEDED`	The client sets a deadline shorter than the server's default deadline (10 minutes), and the request didn't finish within the client-provided deadline.	Consider increasing the client-provided deadline.	

Handle errors

Avoid spikes in traffic. Spikes are sudden and significant increases in the number of requests within a very short period of time. Sometimes, spikes in traffic might cause issues for quota enforcement and might increase the chance of server overloading.

Be careful about retrying an event. We recommend retrying no more than two times. The minimum delay is one second with subsequent requests backing up exponentially.

What's next

- Generative AI on Vertex AI has some limitations. To learn more, see PaLM API limitations (/vertex-ai/generative-ai/docs/learn/responsible-ai#limitations).
- Try a quickstart tutorial using Vertex AI Studio (/vertex-ai/generative-ai/docs/start/quickstarts/quickstart) or the Vertex AI API (/vertex-ai/generative-ai/docs/start/quickstarts/quickstart-multimodal).
- Explore pretrained models in Model Garden (/vertex-ai/generative-ai/docs/model-garden/explore-models).
- Learn about quotas and limits (/vertex-ai/docs/quotas).
- Learn about pricing (/vertex-ai/pricing#generative_ai_models).

Last updated 2025-06-06 UTC.