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 and zero data retention

Google was the first in the industry to publish an AI/ML Privacy Commitment (https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-unveils-ai-and-ml-privacy-commitment), which outlines our belief that customers should have the highest level of security and control over their data that is stored in the cloud. That commitment extends to Google Cloud's generative AI products. Google ensures that its teams are following these commitments through robust data governance practices, which include reviews of the data that Google Cloud uses in the development of its products. More details about how Google processes data can also be found in Google's Cloud Data Processing Addendum (CDPA) (https://cloud.google.com/terms/data-processing-addendum).

Training restriction

As outlined in Section 17 "Training Restriction" in the Service Terms section of Service Specific Terms (https://cloud.google.com/terms/service-terms), Google won't use your data to train or fine-tune any AI/ML models without your prior permission or instruction. This applies to all managed models on Vertex AI, including GA and pre-GA models.

Customer data retention and achieving zero data retention

Customer data is retained in Vertex AI for Google models for limited periods of time in the following scenarios and conditions. To achieve zero data retention, customers must take specific actions within each of these areas:

- **Data caching for Google models**: By default, Google foundation models cache inputs for Gemini models. This is done to reduce latency and accelerate responses to subsequent prompts from the customer. Cached contents are stored for up to 24 hours in the data center where the request was served. Data caching is enabled or disabled at the Google Cloud project level, and project-level privacy is enforced for cached data. The same cache settings for a Google Cloud project apply to all regions. To achieve zero data retention, you

must disable data caching. See Enabling and disabling data caching (#enabling-disabling-caching).

- **Prompt logging for abuse monitoring for Google models**: As outlined in Section 4.3 "Generative AI Safety and Abuse" of Google Cloud Platform Terms of Service (https://cloud.google.com/terms), Google may log prompts to detect potential abuse and violations of its Acceptable Use Policy (https://cloud.google.com/terms/aup) and Prohibited Use Policy (https://policies.google.com/terms/generative-ai/use-policy) as part of providing generative AI services to customers. Only customers whose use of Google Cloud is governed by the Google Cloud Platform Terms of Service (https://cloud.google.com/terms) and who don't have an Invoiced Cloud Billing account (/billing/docs/concepts#billing_account_types) are subject to prompt logging for abuse monitoring. If you are in scope for prompt logging for abuse monitoring and want zero data retention, you can request an exception for abuse monitoring. See Abuse monitoring (/vertex-ai/generative-ai/docs/learn/abuse-monitoring).
- **Grounding with Google Search**: As outlined in Section 19 "Generative AI Services: Grounding with Google Search" of the Service Specific Terms (https://cloud.google.com/terms/service-terms), Google stores prompts and contextual information that customers may provide, and generated output for thirty (30) days for the purposes of creating grounded results and search suggestions, and this stored information may be used for debugging and testing of systems that support grounding with Google Search. There is no way to disable the storage of this information if you use Grounding with Google Search.
- **Grounding with Google Maps**: As outlined in Section 19 "Generative AI Services: Grounding with Google Maps" of the Service Specific Terms (https://cloud.google.com/terms/service-terms), Google stores prompts and contextual information that customers may provide, and generated output for thirty (30) days for the purposes of creating grounded results, and this stored information may only be used for reliability engineering, such as debugging in case of service issues, of systems that support grounding with Google Maps. There is no way to disable the storage of this information if you use Grounding with Google Maps.
- **Session resumption for Gemini Live API:** This feature is disabled by default. It must be enabled by the user every time they call the API by specifying the field in the API request, and project-level privacy is enforced for cached data. Enabling Session Resumption allows the user to reconnect to a previous session within 24 hours by storing cached data, including text, video, and audio prompt data and model outputs, for up to 24 hours. To achieve zero data retention, do not enable this feature. For more information about this feature, including how to enable it, see Live API (/vertex-ai/generative-ai/docs/live-api#session-resumption).

This applies to all managed models on Vertex AI, including GA and pre-GA models.

Enabling and disabling data caching

You can use the following curl commands to get caching status, disable caching, or re-enable caching. When you disable or re-enable caching, the change applies to all Google Cloud regions. For more information about using Identity and Access Management to grant permissions required to enable or disable caching, see Vertex AI access control with IAM (/vertex-ai/docs/general/access-control). Expand the following sections to learn how to get the current cache setting, to disable caching, and to enable caching.

Get current caching setting

Run the following command to determine if caching is enabled or disabled for a project. To run this command, a user must be granted one of the following roles: `roles/aiplatform.viewer`, `roles/aiplatform.user`, or `roles/aiplatform.admin`.

```
PROJECT_ID=PROJECT_ID
# Setup project_id
$ gcloud config set project PROJECT_ID

# GetCacheConfig
$ curl -X GET -H "Authorization: Bearer $(gcloud auth application-default print-ac

# Response if caching is enabled (caching is enabled by default).
{
  "name": "projects/PROJECT_ID/cacheConfig"
}

# Response if caching is disabled.
{
  "name": "projects/PROJECT_ID/cacheConfig"
  "disableCache": true
}
```

Disable caching

Run the following curl command to disable caching for a Google Cloud project. To run this command, a user must be granted the Vertex AI administrator role, `roles/aiplatform.admin`.

```
PROJECT_ID=PROJECT_ID
# Setup project_id
$ gcloud config set project PROJECT_ID

# Setup project_id.
```

```
$ gcloud config set project ${PROJECT_ID}

# Opt-out of caching.
$ curl -X PATCH -H "Authorization: Bearer $(gcloud auth application-default print-
  "name": "projects/PROJECT_ID /cacheConfig",
  "disableCache": true
}'

# Response.
{
  "name": "projects/PROJECT_ID /locations/us-central1/projects/PROJECT_ID /cache
  "done": true,
  "response": {
    "@type": "type.googleapis.com/google.protobuf.Empty"
  }
}
```

Enable caching

If you disabled caching for a Google Cloud project and want re-enable it, run the following curl command. To run this command, a user must be granted the Vertex AI administrator role, `roles/aiplatform.admin`.

```
PROJECT_ID=PROJECT_ID
LOCATION_ID="us-central1"
# Setup project_id
$ gcloud config set project PROJECT_ID

# Setup project_id.
$ gcloud config set project ${PROJECT_ID}

# Opt in to caching.
$ curl -X PATCH     -H "Authorization: Bearer $(gcloud auth application-default pr
  "name": "projects/PROJECT_ID /cacheConfig",
  "disableCache": false
}'

# Response.
{
  "name": "projects/PROJECT_ID /locations/us-central1/projects/PROJECT_ID /cache
  "done": true,
  "response": {
    "@type": "type.googleapis.com/google.protobuf.Empty"
  }
```

```
}
```

What's next

- Learn about responsible AI best practices and Vertex AI's safety filters (/vertex-ai/generative-ai/docs/learn/responsible-ai).
- Learn about Gemini in Google Cloud data governance (/gemini/docs/discover/data-governance).

Last updated 2025-09-19 UTC.