Model Garden

Model Garden

Model Garden

Modalities

Language	39
Vision	79
Tabular	2
Documents	2
Speech	1
Video	4

Tasks

Generation	39
Classification	42
Detection	29
Extraction	8
Recognition	8
Translation	4
Embedding	3
Segmentation	6
Retrieval	1
Tracking	1

Features

Generative AI Studio	7
API available	9

Search models

Suggestions text embedding essay outline BERT

Pre-trained multi-task models that can be further tuned or customised for specific tasks. Models marked with ✦ are available in Generative AI Studio.

✦ Generative AI · Language

PaLM 2 for Text

Fine-tuned to follow natural language instructions and is suitable for a variety of language tasks, such as: classification, extraction, summarization and content...

text-bison

VIEW DETAILS

Foundation · Language

Llama 2

Fine-tune & deploy Meta's Llama 2 models on Vertex AI.

Llama2-7B-001

VIEW DETAILS

Foundation · Language

Embeddings for Text

Text embedding is an important NLP technique that converts textual data into numerical vectors that can be processed by machine learning algorithms, especially large models...

textembedding-gecko@001

VIEW DETAILS

✦ Generative AI · Vision

Imagen for Image Generation

Create or edit studio-grade images at scale via text prompts

google/imagegeneration-002

VIEW DETAILS

SHOW ALL (54)

Fine-tunable models

Models that data scientists can further fine-tune through a custom notebook or pipeline.

Classification · Vision

tfhub/EfficientNetV2

EfficientNet V2 are a family of image classification models, which achieve better parameter efficiency and faster training speed than prior arts.

tensorflow-hub/efficientnetv2

VIEW DETAILS

Classification · Vision

tfvision/vit

The Vision Transformer (ViT) is a transformer-based architecture for image classification.

tfvision/vit-s16

VIEW DETAILS

Detection · Vision

tfvision/SpineNet

SpineNet is an image object detection model generated using Neural Architecture Search.

tfvision/spinenet49

VIEW DETAILS

Detection · Vision

tfvision/YOLO

YOLO algorithm is a one-stage object detection algorithm that can achieve real-time performance on a single GPU.

tfvision/scaled-yolo

VIEW DETAILS

Model Garden ✦ EXPLORE GENERATIVE AI

Search models

Suggestions text embedding essay outline BERT

Modalities

Language	39
Vision	79
Tabular	2
Documents	2
Speech	1
Video	4

Tasks

Generation	39
Classification	42
Detection	29
Extraction	8
Recognition	8
Translation	4
Embedding	3
Segmentation	6
Retrieval	1
Tracking	1

Features

Generative AI Studio	7
API available	9

Pre-tra

✦ Generative AI · Language

PaLM 2 for Text

Fine-tuned to follow natural language instructions and is suitable for a variety of language tasks, such as: classification, extraction, summarization and content...

text-bison

VIEW DETAILS

Foundation · Language

Llama 2

Fine-tune & deploy Meta's Llama 2 models on Vertex AI.

Llama2-7B-001

VIEW DETAILS

Foundation · Language

Embeddings for Text

Text embedding is an important NLP technique that converts textual data into numerical vectors that can be processed by machine learning algorithms, especially large models...

textembedding-gecko@001

VIEW DETAILS

✦ Generative AI · Vision

Imagen for Image Generation

Create or edit studio-grade images at scale via text prompts

google/imagegeneration-002

VIEW DETAILS

SHOW ALL (54)

Fine-tunable models

Models that data scientists can further fine-tune through a custom notebook or pipeline.

Classification · Vision

tfhub/EfficientNetV2

EfficientNet V2 are a family of image classification models, which achieve better parameter efficiency and faster training speed than prior arts.

tensorflow-hub/efficientnetv2

VIEW DETAILS

Classification · Vision

tfvision/vit

The Vision Transformer (ViT) is a transformer-based architecture for image classification.

tfvision/vit-s16

VIEW DETAILS

Detection · Vision

tfvision/SpineNet

SpineNet is an image object detection model generated using Neural Architecture Search.

tfvision/spinenet49

VIEW DETAILS

Detection · Vision

tfvision/YOLO

YOLO algorithm is a one-stage object detection algorithm that can achieve real-time performance on a single GPU.

tfvision/scaled-yolo

VIEW DETAILS

Modalities

Language	39
Vision	79
Tabular	2
Documents	2
Speech	1
Video	4

Tasks

Generation	39
Classification	42
Detection	29
Extraction	8
Recognition	8
Translation	4
Embedding	3
Segmentation	6
Retrieval	1
Tracking	1

MY MODELS

els

t embedding essay outline BERT

models

task models that can be further tuned or customized for specific tasks.

Language

r Text

follow natural language
nd is suitable for a
uage tasks, such as:
extraction,
n and content...
01

S

Foundation Language

PaLM 2 for Chat

Fine-tuned to conduct natural conversation. Use this model to build and customize your own chatbot application.

chat-bison@001

VIEW DETAILS

Foundation Language

Embeddings for text

Text embedding is an important NLP technique that converts textual data into numerical vectors that can be processed by machine learning algorithms, especially large models...

textembedding-gecko@001

VIEW DETAILS

Foundation Speech

Chirp

Chirp is a version of a Universal Speech Model that has over 2B parameters and can transcribe in over 100 languages in a single model.

chirp-rnnt1

VIEW DETAILS

Foundation Vision

Embeddings for Image

Generates vectors based on images, which can be used for downstream tasks like image classification, image search, and so on.

imageembedding-001

VIEW DETAILS

5)

le models

scientists can further fine-tune through a custom notebook or pipeline.

Vision

cientNetV2

2 are a family of image
models, which achieve
ter efficiency and
speed than prior arts.

b/efficientnetv2

Classification Vision

tfvision/vit

The Vision Transformer (ViT) is a transformer-based architecture for image classification.

tfvision/vit-s16

Detection Vision

tfvision/SpineNet

SpineNet is an image object detection model generated using Neural Architecture Search.

tfvision/spinenet49

Detection Vision

tfvision/YOLO

YOLO algorithm is a one-stage object detection algorithm that can achieve real-time performance on a single GPU.

tfvision/scaled-yolo

Foundation Language Vision

Stable Diffusion Inpainting

Stable Diffusion Inpainting is a latent diffusion model capable of inpainting images given any text input and a mask image.

runwayml/stable-diffusion-inpainting

Foundation models

Pre-trained multi-task models that can be further tuned or customised for specific tasks. Models marked with ✦ are available in Generati

✦ Generative AI | Language

PaLM 2 for Text

Fine-tuned to follow natural language instructions and is suitable for a variety of language tasks, such as: classification, extraction, summarization and content...

text-bison

VIEW DETAILS

Foundation | Language

Llama 2

Fine-tune & deploy Meta's Llama 2 models on Vertex AI.

Llama2-7B-001

VIEW DETAILS

Foundation | Language

Embeddings for Text

Text embedding is an important NLP technique that converts textual data into numerical vectors that can be processed by machine learning algorithms, especially large models...

textembedding-gecko@001

VIEW DETAILS

SHOW ALL (54)

Modalities

Language	39
Vision	79
Tabular	2
Documents	2
Speech	1
Video	4

Tasks

Generation	39
Classification	42
Detection	29
Extraction	8
Recognition	8
Translation	4
Embedding	3
Segmentation	6
Retrieval	1
Tracking	1

Features

Generative AI Studio	7

tfhub/EfficientNetV2

EfficientNet V2 are a family of image classification models, which achieve better parameter efficiency and faster training speed than prior arts.

tensorflow-hub/efficientnetv2

VIEW DETAILS

tfvision/vit

The Vision Transformer (ViT) is a transformer-based architecture for image classification.

tfvision/vit-s16

VIEW DETAILS

tfvision/SpineNet

SpineNet is an image object detection model generated using Neural Architecture Search.

tfvision/spinenet49

VIEW DETAILS

tfvision/YOLO

YOLO algorithm is a one-stage object detection algorithm that can achieve real-time performance on a single GPU.

tfvision/scaled-yolo

VIEW DETAILS

Model Card

PaLM 2 for Text

Fine-tuned to follow natural language instructions and is suitable for a variety of language tasks, such as: classification, extraction, summarization and content generation.

OPEN PROMPT DESIGN | VIEW API CODE

OVERVIEW | USE CASES | DOCUMENTATION

Overview

text-bison is the name of the PaLM 2 for text large language model that understands and generates language. It's a foundation model that performs well at a variety of natural language tasks such as sentiment analysis, entity extraction, and content creation. The type of content that **text-bison** can create includes document summaries, answers to questions, and labels that classify content.

The PaLM 2 for text is ideal for tasks that can be completed with one API response, without the need for continuous conversation. For text tasks that require back-and-forth interactions, use the PaLM 2 for chat.

Use cases

- **Summarization:** Create a shorter version of a document that incorporates pertinent information from the original text. For example, you might want to summarize a chapter from a textbook. Or, you could create a succinct product description from a long paragraph that describes the product in detail.
- **Question answering:** Provide answers to questions in text. For example, you might automate the creation of a Frequently Asked Questions (FAQ) document from knowledge base content.
- **Classification:** Assign a label to provided text. For example, a label might be applied to text that describes how grammatically correct it is.

Resource ID

text-bison@001

Tags

Task

Classification | Detection | Extraction | Generation | Recognition | Translation

Versions

Resource ID	Release date	Release stage	Description
text-bison	2023-08-29	General Availability	
text-bison-32k	2023-08-29	Public Preview	
text-bison@001	2023-06-07	General Availability	
text-bison@001	2023-05-10	Public Preview	Quality improvements and restage -001 as the first stable base model release

Model Garden

Foundation	Pre-trained APIs	Open Source	Third-Party
Google's state-of-the-art multimodal models	Google's pre-trained, task-specific models	Enterprise-ready open source models	Coming soon!

Foundation Models

Across a variety of model sizes to address use cases

PaLM for Text

Custom language tasks

PaLM for Chat

Multi-turn conversations with session context

Imagen for Text to Image

Create and edit images from simple prompts

Embeddings API for Text and Image

Extract semantic information from unstructured data

Chirp for Speech to Text

Build voice enabled applications

Codey for Code Generation

Improve coding and debugging

PaLM 2

Text

Perform language tasks like summarization, question answering, and classification from natural language prompts.

Chat

Text tasks that flow like natural, multi-turn conversations such as help agents.

Embeddings

Retrieve the vector representation for the semantic meaning of input text.

Codey

Code Generation

Generate functions, tests, and object classes from natural language descriptions.

Code Completion

In-line code generation based on surrounding code.

Code Chat

Multi-turn reasoning, debugging, and explanations of code.

Sec-PaLM2

Limited Access

Text

Analyze security data from multiple sources. Generate rules and queries from natural language prompts.

Imagen

Limited Access

Image Generation

Create images from text prompts.

Embeddings

Vector representation of an image. Also generates text vectors for querying text-to-image and image-to-text.

Chirp

Transcription

Convert speech to text for over 100 languages.

Other foundation models

Image Generation

- Stable Diffusion
- Stable Diffusion Inpainting
- Stable Diffusion LoRA
- InstructPix2Pix
- ControlNet
- FaceStylizer

Image Processing

- PaLI zero-shot
- BLIP 2
- CLIP
- OWL-ViT
- ViT GPT2
- LayoutLM for VQA
- BiomedCLIP
- ImageBind

Natural Language

- BERT
- T5-FLAN
- NLLB (Translation)
- Mistral-7B
- RoBERTa-large (PEFT)
- Palmyra Med

Fine-tunable models

Foundation

- PaLM (via Gen AI Studio)
- Stable Diffusion Inpainting
- ControlNet

Image Classification

- EfficientNetV2
- ResNet
- BEiT
- DeiT
- MobileNet
- ... many more!

Object Detection

- SpineNet
- YOLO
- Faster R-CNN
- Mask R-CNN
- RetinaNet

Task-specific solutions: Vision

Detection

- Occupancy
- Person/Vehicle
- PPE
- Objects
- Person blur
- Watermark
- Faces

Recognition

- Product
- Tag

Extraction

- Text
- Form Parser
- Document AI OCR

Classification

- Moderation

Task-specific solutions: Language

Classification

- Entity analysis
- Sentiment analysis
- Entity sentiment analysis
- Text Moderation

Translation

- Text Translation

Extraction

- Syntax analysis

Task-specific solutions: Tabular

Classification / Regression

- AutoML E2E
- TabNet

Sometimes, you don't need a model.
You just need a prompt.

Model Garden

01 Introduction to Model Garden

02 Available Model Types

03 **Working with Gen AI Studio**

04 Working with Model Registry

05 Intro: Course Use Cases

Generative AI Studio

PaLM 2 for Text

Fine-tuned to follow natural language instructions and is suitable for a variety of language tasks, such as: classification, extraction, summarization and content generation.

OPEN PROMPT DESIGN | VIEW API CODE

OVERVIEW | USE CASES | DOCUMENTATION

Overview

text-bison is the name of the PaLM 2 for text large language model that understands and generates language. It's a foundation model that performs well at a variety of natural language tasks such as sentiment analysis, entity extraction, and content creation. The type of content that **text-bison** can create includes document summaries, answers to questions, and labels that classify content.

The PaLM 2 for text is ideal for tasks that can be completed with one API response, without the need for continuous conversation. For text tasks that require back-and-forth interactions, use the PaLM 2 for chat.

Use cases

- **Summarization:** Create a shorter version of a document that incorporates pertinent information from the original text. For example, you might want to summarize a chapter from a textbook. Or, you could create a succinct product description from a long paragraph that describes the product in detail.
- **Question answering:** Provide answers to questions in text. For example, you might automate the creation

Codey for Code Generation

Generates code based on natural language input. Good for writing functions, classes, unit tests, and more.

OPEN PROMPT DESIGN | VIEW API CODE

OVERVIEW | USE CASES | DOCUMENTATION

Overview

code-bison is the name of the model that supports code generation. It's a foundation model that generates code based on a natural language description. The type of content that **code-bison** can create includes functions, web pages, and unit tests. **code-bison** is supported by the code generation Codey API. The Codey APIs are in the PaLM API family.

The Codey API for code generation is ideal for tasks that can be completed with one API response, without the need for continuous conversation. For code tasks that require back-and-forth interactions, use the Codey API for code chat.

Use cases

Some common used cases for code generation are:

- **Unit tests:** Use the prompt to request a unit test for a function.
- **Write a function:** Pass a problem to the model to get a function that solves that problem.
- **Create a class:** Use a prompt to describe the purpose of a class and have code that defines the class returned.

Untitled prompt

FREE-FORM | STRUCTURED | SAVE | VIEW CODE

Prompt

Write a prompt and then click Submit

Response

Markdown

The model will generate a response after you click Submit

Model may display inaccurate or offensive information that doesn't represent Google's view. Not all languages are supported. Learn more.

Region: us-central1 (Iowa)

Model: text-bison (latest)

Temperature: 0 — 1 — 0.2

Advanced

Token limit: 1 — 2048 — 1024

Top-k: 1 — 40 — 40

Top-p: 0 — 1 — 0.8

Max. responses: 1 — 8 — 1

Add stop sequence

Press Enter after each sequence

Streaming responses

Print responses as they're generated

Safety filter threshold: Block few

SUBMIT RESET PARAMETERS

Untitled prompt

FREEFORM STRUCTURED SAVE VIEW CODE

Prompt

Write a prompt and then click Submit

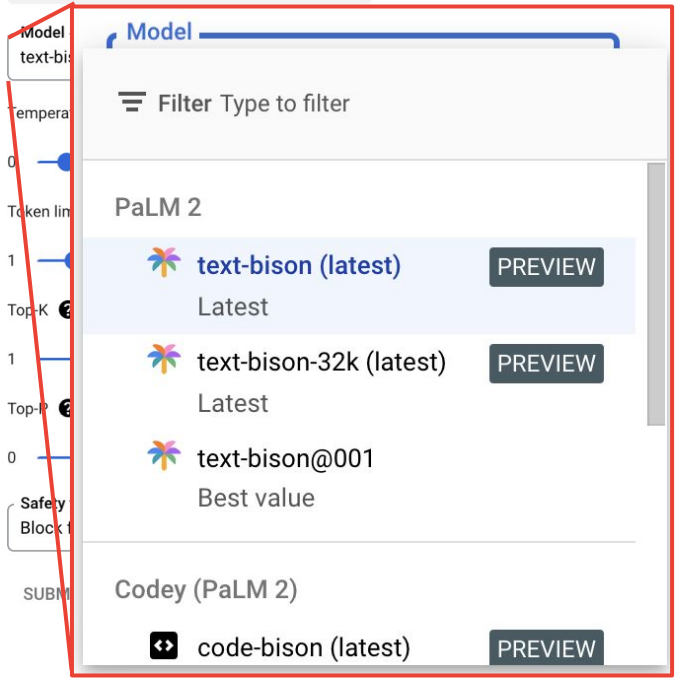
Response

Markdown

The model will generate a response after you click Submit

Model may display inaccurate or offensive information that doesn't represent Google's view. Google models currently only support English.

We want your feedback.



REPORT INAPPROPRIATE RESPONSES

Content processed through Vertex AI is assessed against a list of safety attributes. Confidence scores for these attributes are returned in API responses but are not visible in the UI.

Untitled prompt

FREEFORM STRUCTURED SAVE VIEW CODE

Prompt

Write a prompt and then click Submit

Response Markdown

The model will generate a response after you click Submit

Model may display inaccurate or offensive information that doesn't represent Google's view. Google models currently only support English.

Temperature
0 1 0.2

Advanced

Token limit
1 2048 1024

Top-k
1 40 40

Top-p
0 1 0.8

Max. responses
1 8 1

Add stop sequence
Press Enter after each sequence

Streaming responses
Print responses as they're generated

Safety filter threshold
Block few

Untitled prompt

FREE-FORM | STRUCTURED | SAVE | VIEW CODE

Context

Write your context here...

Examples

IMPORT | ADD INPUT COLUMN

INPUT	Add an input prefix	OUTPUT	Add an output prefix
Write an input to add a new row		Write an output to add a new row	

Test

Batch testing | ADD TO EXAMPLES

INPUT	Add an input prefix	OUTPUT	Add an output prefix	Markdown
Write an input		Model responses will appear here		

Model may display inaccurate or offensive information that doesn't represent Google's view. Not all languages are supported. Learn more.

Region: us-central1 (Iowa)

Model: text-bison (latest)

Temperature: 0 — 1 — 0.2

Advanced

Token limit: 1 — 2048 — 1024

Top-k: 1 — 40 — 40

Top-p: 0 — 1 — 0.8

Max. responses: 1 — 8 — 1

Add stop sequence

Press Enter after each sequence

Streaming responses

Print responses as they're generated

Safety filter threshold: Block few

Untitled prompt

FREEFORM | STRUCTURED | SAVE | VIEW CODE

Context

You are a machine learning practitioner training a junior staff member. Provide answers to questions that include practical use cases to help the new employee learn.

Examples | ADD INPUT COLUMN

Test | ADD TO EXAMPLES

INPUT Add an input prefix

OUTPUT Add an output prefix | Markdown

Why would I choose precision versus recall as my evaluation metric?

Precision and recall are two important metrics for evaluating the performance of a binary classifier. Precision is the fraction of positive predictions that are actually positive, while recall is the fraction of actual positives that are predicted as positive.

Precision is important when the cost of false positives is high. For example, if you are using a classifier to identify fraudulent transactions, you would want to have a high precision so that you don't accidentally block legitimate transactions.

Recall is important when the cost of false negatives is high. For example, if you are using a classifier to identify cancer, you would want to have a high recall so that you don't miss any cases of cancer.

The choice of which metric to use depends on the specific application. In general, you should choose the metric that is most important for your particular application.

Model: text-bison (latest)

Temperature: 0 — 1 — 0.2

Advanced

Token limit: 1 — 2048 — 1024

Top-k: 1 — 40 — 40

Top-p: 0 — 1 — 0.8

Max. responses: 1 — 8 — 1

Add stop sequence

Press Enter after each sequence

Streaming responses

Print responses as they're generated

Safety filter threshold: Block few

SUBMIT | RESET PARAMETERS

Untitled prompt

FREEFORM | STRUCTURED | SAVE | VIEW CODE

Context

You are a machine learning practitioner training a junior staff member. Provide answers to questions that include practical use cases to help the new employee learn.

Examples ADD INPUT COLUMN

Test ADD TO EXAMPLES

INPUT	Add an input prefix	OUTPUT	Add an output prefix	Markdown
Why would I choose precision versus recall as my evaluation metric?		Precision and recall are two important metrics for evaluating the performance of a binary classifier. Precision is the fraction of positive predictions that are actually positive, while recall is the fraction of actual positives that are predicted as positive. Precision is important when the cost of false positives is high. For example, if you are using a classifier to identify fraudulent transactions, you would want to have a high precision so that you don't accidentally block legitimate transactions. Recall is important when the cost of false negatives is high. For example, if you are using a classifier to identify cancer, you would want to have a high recall so that you don't miss any cases of cancer. The choice of which metric to use depends on the specific application. In general, you should choose the metric that is most important for your particular application.	

Model
text-bison (latest)

Temperature
0 — 1 — 0.2

Advanced

Token limit
1 — 2048 — 1024

Top-k
1 — 40 — 40

Top-p
0 — 1 — 0.8

Max. responses
1 — 8 — 1

Add stop sequence
Press Enter after each sequence

Streaming responses
Print responses as they're generated

Safety filter threshold
Block few

SUBMIT RESET PARAMETERS

Untitled prompt

FREEFORM | STRUCTURED | SAVE | VIEW CODE

Context

You are a machine learning practitioner training a junior staff member. Provide answers to questions that include practical use cases to help the new employee learn.

Examples ADD INPUT COLUMN

Test ADD TO EXAMPLES

INPUT	OUTPUT
Add an input prefix	Add an output prefix
Why would I choose precision versus recall as my evaluation metric?	Precision and recall are two important metrics for evaluating the performance of a binary classifier. Precision is the fraction of positive predictions that are actually positive, while recall is the fraction of actual positives that are predicted as positive. Precision is important when the cost of false positives is high. For example, if you are using a classifier to identify fraudulent transactions, you would want to have a high precision so that you don't accidentally block legitimate transactions. Recall is important when the cost of false negatives is high. For example, if you are using a classifier to identify cancer, you would want to have a high recall so that you don't miss any cases of cancer. The choice of which metric to use depends on the specific application. In general, you should choose the metric that is most important for your particular application.

Markdown

Model
text-bison (latest)

Temperature
0 — 1 — 0.2

Advanced

Token limit
1 — 2048 — 1024

Top-k
1 — 40 — 40

Top-p
0 — 1 — 0.8

Max. responses
1 — 8 — 1

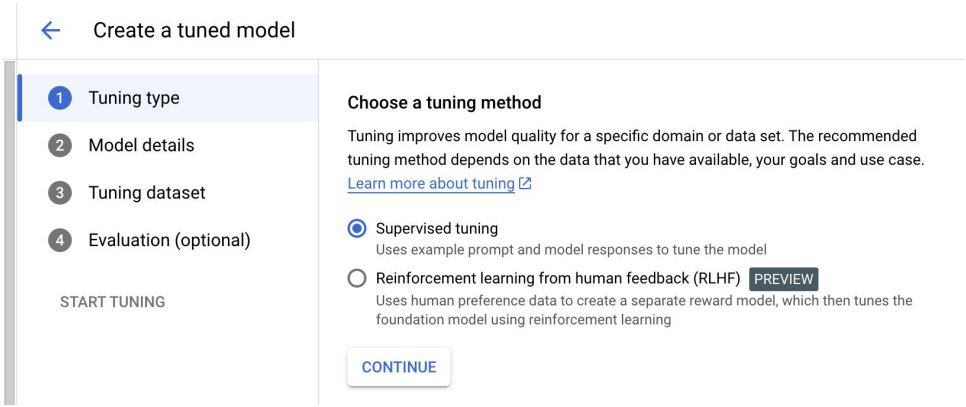
Add stop sequence
Press Enter after each sequence

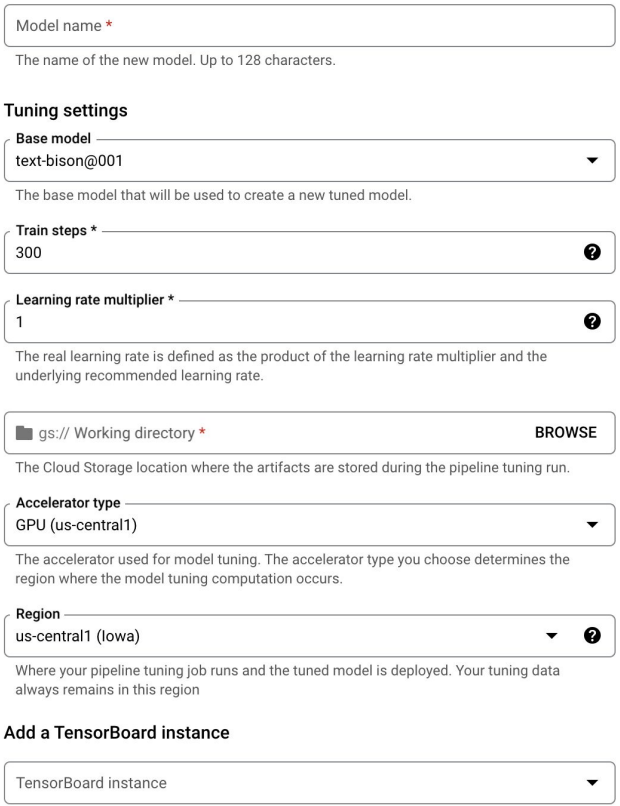
Streaming responses
Print responses as they're generated

Safety filter threshold
Block few

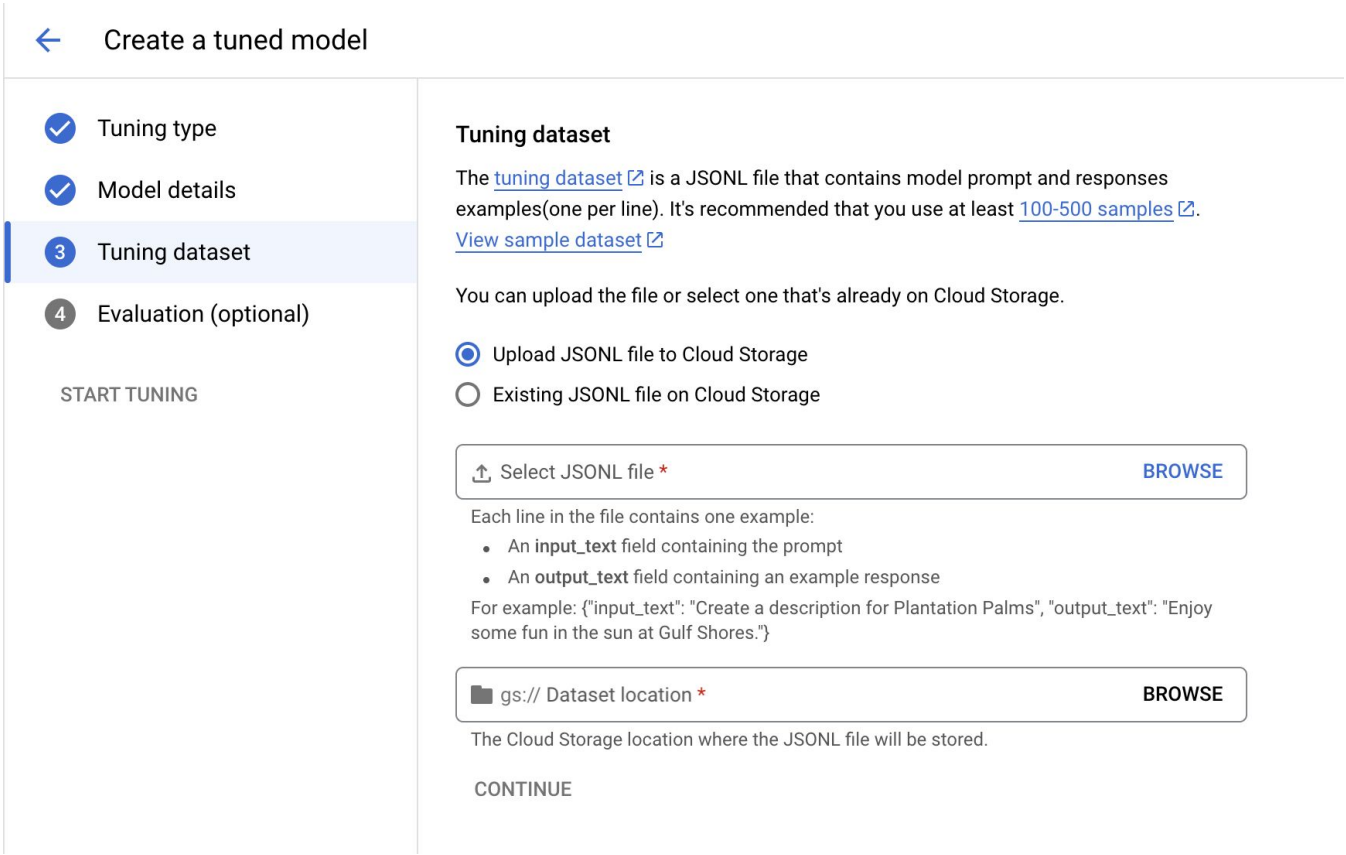
SUBMIT RESET PARAMETERS

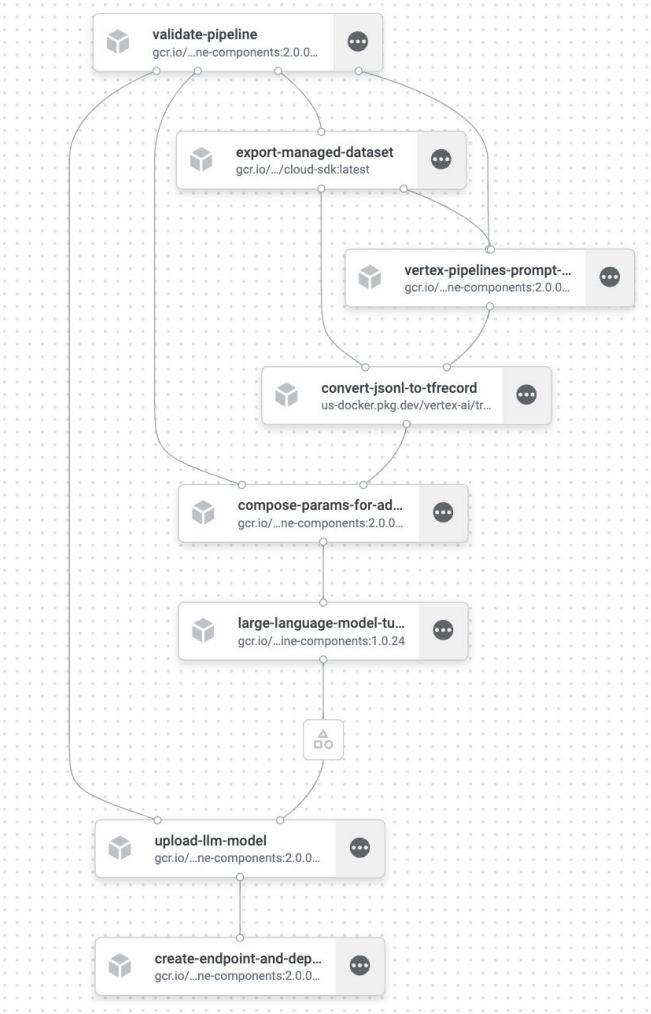
Tune a Model





Tune a Model





Model Garden

01 Introduction to Model Garden

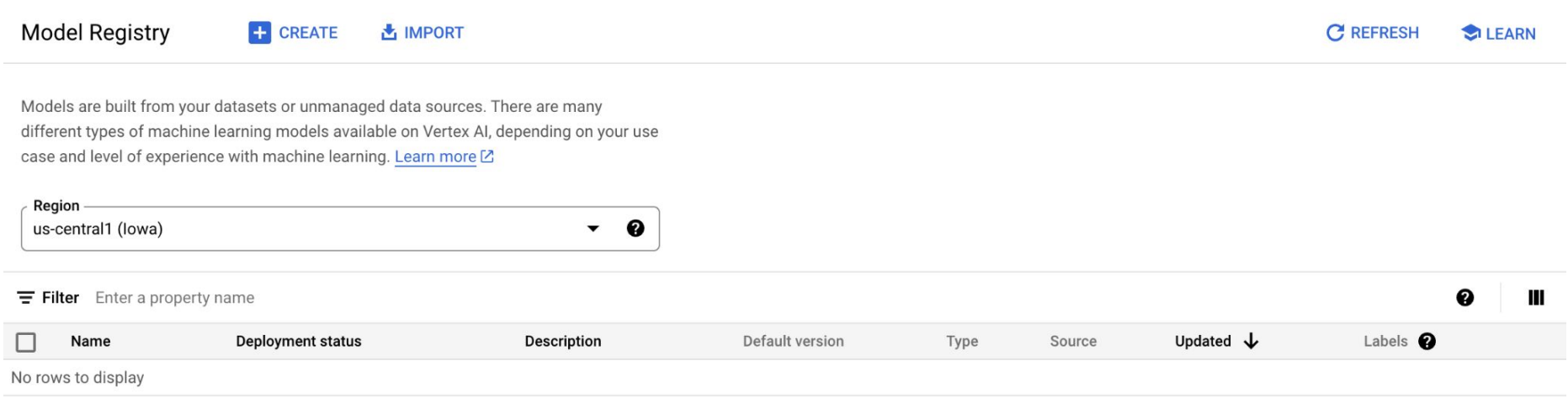
02 Available Model Types

03 Working with Gen AI Studio

04 **Working with Model Registry**

05 Intro: Course Use Cases

Model Registry



Model Registry

CREATE

IMP

Models are built from your datasets or unmanaged data s
different types of machine learning models available on V
case and level of experience with machine learning. Learn

Region
us-central1 (Iowa)

Filter Enter a property name

Name	Deployment status
No rows to display	

Train new model

1. Training method
2. Model details
3. Training container
4. Hyperparameters (optional)
5. Compute and pricing
6. Prediction container (optional)

START TRAINING CANCEL

Dataset *
No managed dataset

Annotation set
-

Objective
Custom

Please refer to the pricing guide for more details (and available deployment options) for each method.

AutoML options are only available when you train with a managed dataset.

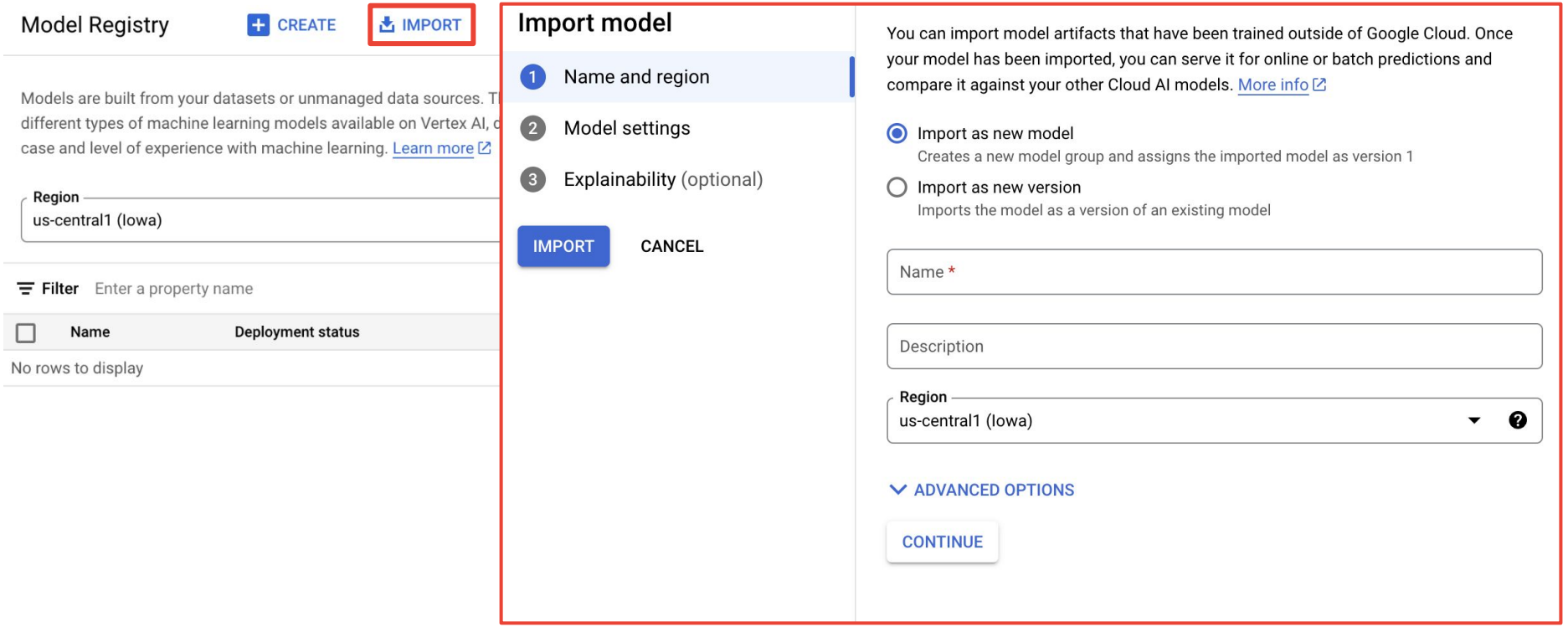
Model training method

AutoML
Train high-quality models with minimal effort and machine learning expertise. Just specify how long you want to train. Learn more

AutoML Edge
Train a model that can be exported for on-prem/on-device use. Typically has lower accuracy. Learn more

Custom training (advanced)
Run your TensorFlow, scikit-learn, and XGBoost training applications in the cloud. Train with one of Google Cloud's pre-built containers or use your own. Learn more

CONTINUE


Model Registry
CREATE
IMPORT
Models are built from your datasets or unmanaged data sources. T
different types of machine learning models available on Vertex AI, c
case and level of experience with machine learning. Learn more
Region
us-central1 (Iowa)
Filter
Enter a property name
Name
Deployment status
No rows to display
Import model
1 Name and region
2 Model settings
3 Explainability (optional)
IMPORT
CANCEL
You can import model artifacts that have been trained outside of Google Cloud. Once your model has been imported, you can serve it for online or batch predictions and compare it against your other Cloud AI models. More info
Import as new model
Creates a new model group and assigns the imported model as version 1
Import as new version
Imports the model as a version of an existing model
Name *
Description
Region
us-central1 (Iowa)
ADVANCED OPTIONS
CONTINUE

Model Card

Stable Diffusion v1-5

Latent text-to-image diffusion model capable of generating photo-realistic images given a text input.

DEPLOY OPEN NOTEBOOK

OVERVIEW USE CASES DOCUMENTATION PRICING

Overview

The **Stable Diffusion** model was first published in the paper "High-Resolution Image Synthesis with Latent Diffusion Models" by Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer in CVPR 2022. To enable diffusion models training on limited computational resources while retaining their quality and flexibility, the authors applied them in the latent space of powerful pretrained autoencoders. Training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. The latent diffusion models achieved a new state of the art for image inpainting and highly competitive performance on various tasks, including unconditional image generation, semantic scene synthesis, and super-resolution, while significantly reducing computational requirements compared to pixel-based diffusion models.

Supported by Stability AI and LAION, the 1st version of Stable Diffusion was trained on 512x512 images from a subset of the LAION-5B database. The model uses a frozen CLIP ViT-L/14 text encoder to condition the model on text prompts. With its 860M UNet and 123M text encoder, the model is relatively lightweight and runs on a GPU with at least 10GB VRAM.

The Stable Diffusion V1.5 was trained by runwayml which published the modeling implementation in their Github repository runwayml/stable-diffusion.

Resource ID

runwayml/stable-diffusion-v1-5

Tags

Task

Generation

Stable Diffusion v1-5

Latent text-to-image diffusion model capable of g
images given a text input.

DEPLOY OPEN NOTEBOOK

OVERVIEW USE CASES DOCUMENTATION P

Overview

The **Stable Diffusion** model was first published in the pape
Models" by Robin Rombach, Andreas Blattmann, Dominik L
enable diffusion models training on limited computational
authors applied them in the latent space of powerful pretra
representation allows for the first time to reach a near-optir
preservation, greatly boosting visual fidelity. The latent diffu
inpainting and highly competitive performance on various t
scene synthesis, and super-resolution, while significantly re
based diffusion models.

Supported by Stability AI and LAION, the 1st version of Stab
of the LAION-5B database. The model uses a frozen CLIP V
prompts. With its 860M UNet and 123M text encoder, the m
least 10GB VRAM.

The Stable Diffusion V1.5 was trained by runwayml which p
repository runwayml/stable-diffusion.

Save model

To deploy this model, you will need to save a copy to Model Registry in your Cloud project.

Resource ID
runwayml/stable-diffusion-v1-5 (Latest model)

- (•) Save as new model
  Creates a new model group and assigns this model as version 1
- ( ) Save as new version
  Saves the model as a version of an existing model

Model name *
stable_diffusion

Some locations have been restricted due to a policy set by your organization. Learn more about restricting locations.

Region *
us-central1 (Iowa)

Advanced Options

SAVE CANCEL

Stable Diffusion v1-5

Latent text-to-image diffusion model capable o
images given a text input.

DEPLOY | OPEN NOTEBOOK

OVERVIEW | USE CASES | DOCUMENTATION

Overview

The **Stable Diffusion** model was first published in the pap
Models" by Robin Rombach, Andreas Blattmann, Dominik
enable diffusion models training on limited computationa
authors applied them in the latent space of powerful pret
representation allows for the first time to reach a near-op
preservation, greatly boosting visual fidelity. The latent di
inpainting and highly competitive performance on variou
scene synthesis, and super-resolution, while significantly
based diffusion models.

Supported by Stability AI and LAION, the 1st version of S
of the LAION-5B database. The model uses a frozen CLIP
prompts. With its 860M UNet and 123M text encoder, the
least 10GB VRAM.

The Stable Diffusion V1.5 was trained by runwayml which
repository runwayml/stable-diffusion.

Deploy to endpoint

1. Define your endpoint
2. Model settings
3. Model monitoring

DEPLOY CANCEL

Create new endpoint | Add to existing endpoint

Endpoint name *

Location

Some locations have been restricted due to a policy set by your organization. Learn more about restricting locations.

Region
us-central1 (Iowa)

Access

Determines how your endpoint can be accessed. By default, endpoints are available for prediction serving through a REST API. Endpoint access can't be changed after the endpoint is created.

- Standard
  Makes the endpoint available for prediction serving through a REST API. AutoML and custom-trained models can be added to standard endpoints.
- Private
  Create a private connection to this endpoint using a VPC network and private services access. Only custom-trained and tabular models can be added to private endpoints. Learn more

ADVANCED OPTIONS

CONTINUE

Model versions

stable_diffusion EDIT DETAILS

Model description
–

Region
us-central1 (Iowa)

Model labels
–

Versions

Filter Enter a property name

Version ID	Alias	Status	Description	Endpoints	Created	Labels
1	default	Ready	–	–	Jun 29, 2023, 7:48:55 AM	

Model Garden

01 Introduction to Model Garden

02 Available Model Types

03 Working with Gen AI Studio

04 Working with Model Registry

05 **Intro: Course Use Cases**

Course Use Cases

03

Task-Specific Models

Problem focus: text classification

Vertex AI provides multiple ways to solve classification tasks. Task-specific models are an easy way to get started.

04

Foundation Models

Problem focus: language understanding

The foundation models provide a lot of power just as they are. We can use them to solve many tasks.

05

Fine-tunable models

Problem focus: Productionalization

Vertex AI provides us the tools to tune models for our own use cases. We're going to see how to formalize that into a pipeline.

Lab: Exploring Model Garden