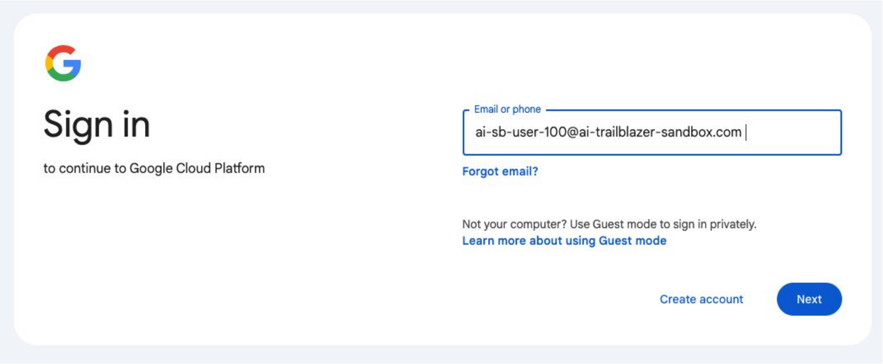
LAB GUIDE

AI Labs - Prudential with Google

Labs Access: Accessing Google Cloud Console

Step 1: Navigate to Google Cloud Console.
https://console.cloud.google.com/



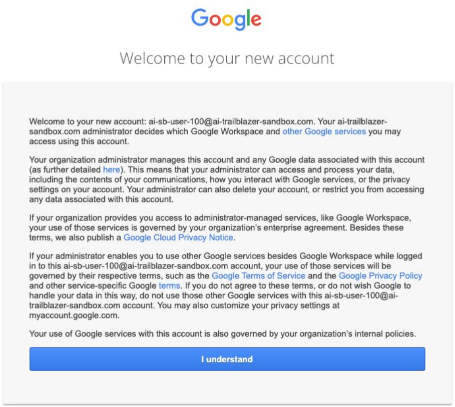
Sign in

to continue to Google Cloud Platform

Email or phone: ai-sb-user-100@ai-trailblazer-sandbox.com

Not your computer? Use Guest mode to sign in privately.

Step 2: For the first login, you have to click "I Understand"



Welcome to your new account

Welcome to your new account: ai-sb-user-100@ai-trailblazer-sandbox.com. Your ai-trailblazer-sandbox.com administrator decides which Google Workspace and other Google services you may access using this account.

Your organization administrator manages this account and any Google data associated with this account (as further detailed here). This means that your administrator can access and process your data, including the contents of your communications, how you interact with Google services, or the privacy settings on your account. Your administrator can also delete your account, or restrict you from accessing any data associated with this account.

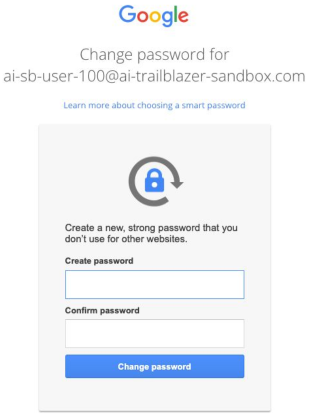
If your organization provides you access to administrator-managed services, like Google Workspace, your use of those services is governed by your organization's enterprise agreement. Besides these terms, we also publish a Google Cloud Privacy Notice.

If your administrator enables you to use other Google services besides Google Workspace while logged in to this ai-sb-user-100@ai-trailblazer-sandbox.com account, your use of those services will be governed by their respective terms, such as the Google Terms of Service and the Google Privacy Policy and other service-specific Google terms. If you do not agree to these terms, or do not wish Google to handle your data in this way, do not use those other Google services with this ai-sb-user-100@ai-trailblazer-sandbox.com account. You may also customize your privacy settings at myaccount.google.com.

Your use of Google services with this account is also governed by your organization's internal policies.

Labs Access: Accessing Google Cloud Console

Step 3: You will be prompted to change password. Please enter a new password.



Google

Change password for
ai-sb-user-100@ai-trailblazer-sandbox.com

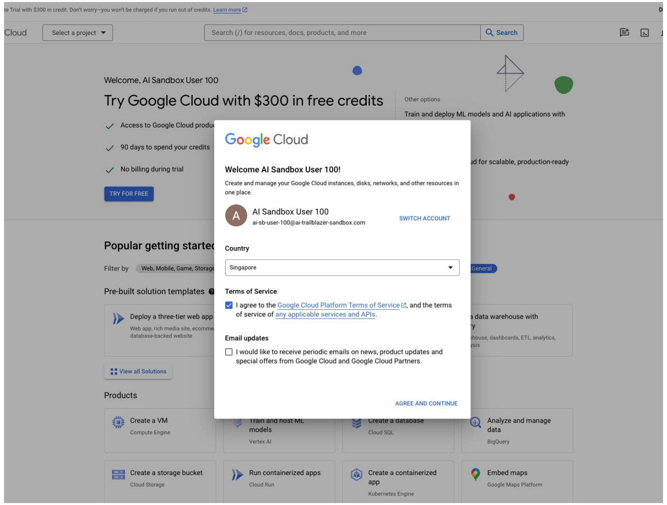
Learn more about choosing a smart password

Create a new, strong password that you don't use for other websites.

Create password

Confirm password

Step 4: Check the checkbox for “Terms of Service” and then proceed to click on “Agree and Continue”.



ee Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more
Cloud
Select a project
Search (/) for resources, docs, products, and more
Search
Welcome, AI Sandbox User 100
Try Google Cloud with $300 in free credits
Access to Google Cloud produ
90 days to spend your credits
No billing during trial
TRY FOR FREE
Other options
Train and deploy ML models and AI applications with
ud for scalable, production-ready
Popular getting startec
Filter by
Web, Mobile, Game, Storage
General
Pre-built solution templates
Deploy a three-tier web app
Web app, rich media site, ecomme database-backed website
a data warehouse with
ehouse, dashboards, ETL, analytics,
View all Solutions
Products
Create a VM
Compute Engine
Train and host ML models
Vertex AI
Create a database
Cloud SQL
Analyze and manage data
BigQuery
Create a storage bucket
Cloud Storage
Run containerized apps
Cloud Run
Create a containerized app
Kubernetes Engine
Embed maps
Google Maps Platform

Google Cloud

Welcome AI Sandbox User 100!

Create and manage your Google Cloud instances, disks, networks, and other resources in one place.

AI Sandbox User 100
ai-sb-user-100@ai-trailblazer-sandbox.com

Country

Singapore

Terms of Service

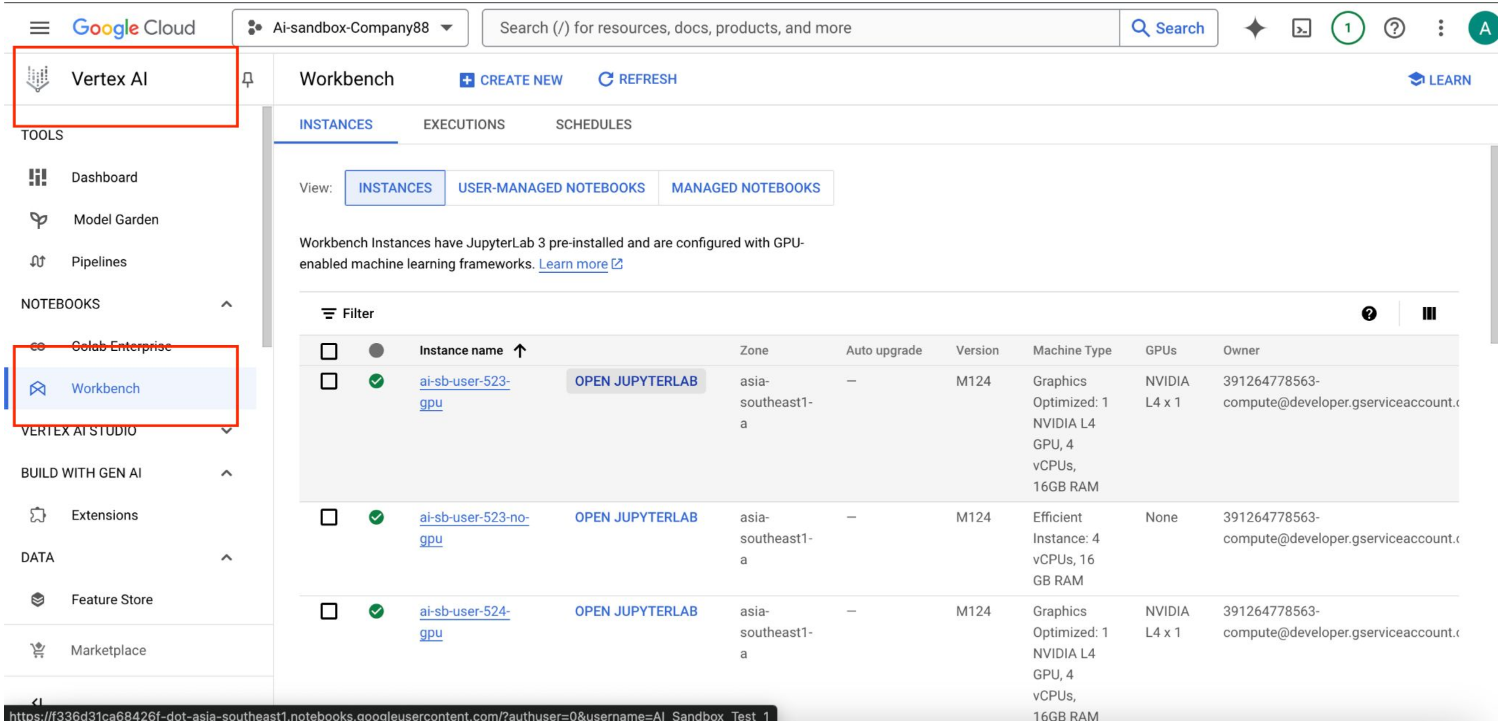
☑ I agree to the Google Cloud Platform Terms of Service, and the terms of service of any applicable services and APIs.

Email updates

□ I would like to receive periodic emails on news, product updates and special offers from Google Cloud and Google Cloud Partners.

Labs Access: Accessing Google Cloud Console

Navigate to Vertex AI and then Workbench



Workbench Instances have JupyterLab 3 pre-installed and are configured with GPU-enabled machine learning frameworks. Learn more

□	●	Instance name ↑		Zone	Auto upgrade	Version	Machine Type	GPUs	Owner
□	✓	ai-sb-user-523-gpu	OPEN JUPYTERLAB	asia-southeast1-a	–	M124	Graphics Optimized: 1 NVIDIA L4 GPU, 4 vCPUs, 16GB RAM	NVIDIA L4 x 1	391264778563-compute@developer.gserviceaccount.c
□	✓	ai-sb-user-523-no-gpu	OPEN JUPYTERLAB	asia-southeast1-a	–	M124	Efficient Instance: 4 vCPUs, 16 GB RAM	None	391264778563-compute@developer.gserviceaccount.c
□	✓	ai-sb-user-524-gpu	OPEN JUPYTERLAB	asia-southeast1-a	–	M124	Graphics Optimized: 1 NVIDIA L4 GPU, 4 vCPUs, 16GB RAM	NVIDIA L4 x 1	391264778563-compute@developer.gserviceaccount.c

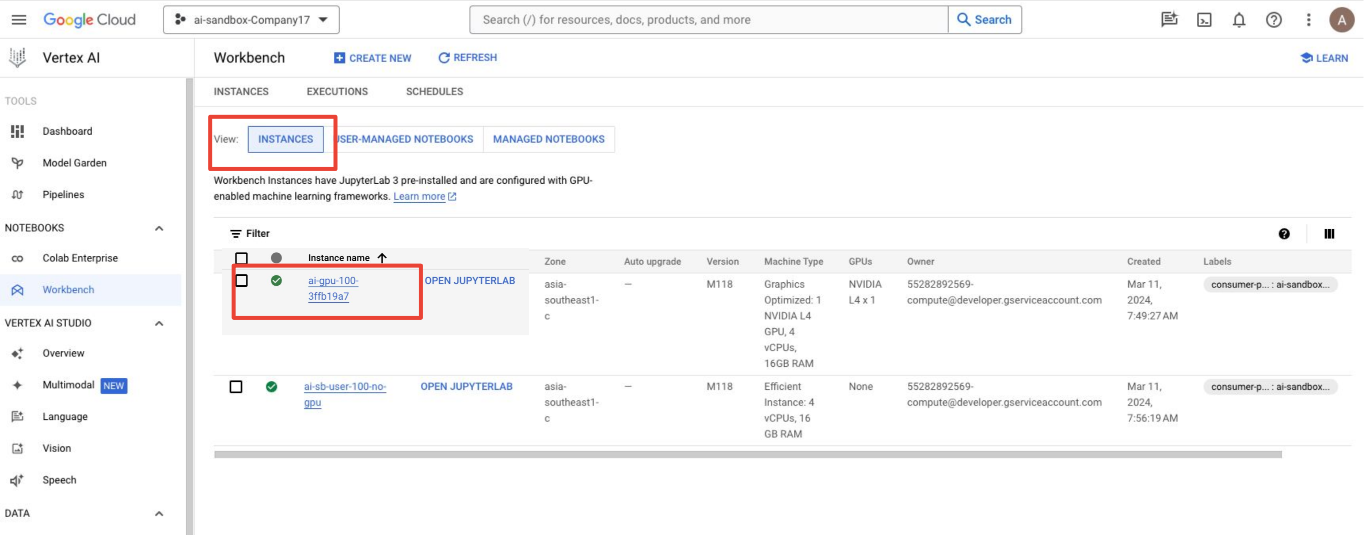
https://f336d31ca68426f-dot-asia-southeast1.notebooks.googleusercontent.com/?authuser=0&username=AI_Sandbox_Test_1

Labs Setup: Accessing Workbench Notebook

Step 4: Once at Workbench, ensure you are on the **INSTANCES** tab

Step 5: You should see that the notebooks created for your team.
Each user is assigned 2 notebooks. 1 with GPU and another without GPU.

GPU instance: ai-gpu-100-XXX. Non-GPU instance: ai-no-gpu-100--XXX



Google Cloud · ai-sandbox-Company17 · Search (/) for resources, docs, products, and more · Search

Vertex AI

Workbench · CREATE NEW · REFRESH · LEARN

TOOLS: Dashboard, Model Garden, Pipelines

NOTEBOOKS: Colab Enterprise, Workbench

VERTEX AI STUDIO: Overview, Multimodal NEW, Language, Vision, Speech

DATA

INSTANCES · EXECUTIONS · SCHEDULES

View: INSTANCES · USER-MANAGED NOTEBOOKS · MANAGED NOTEBOOKS

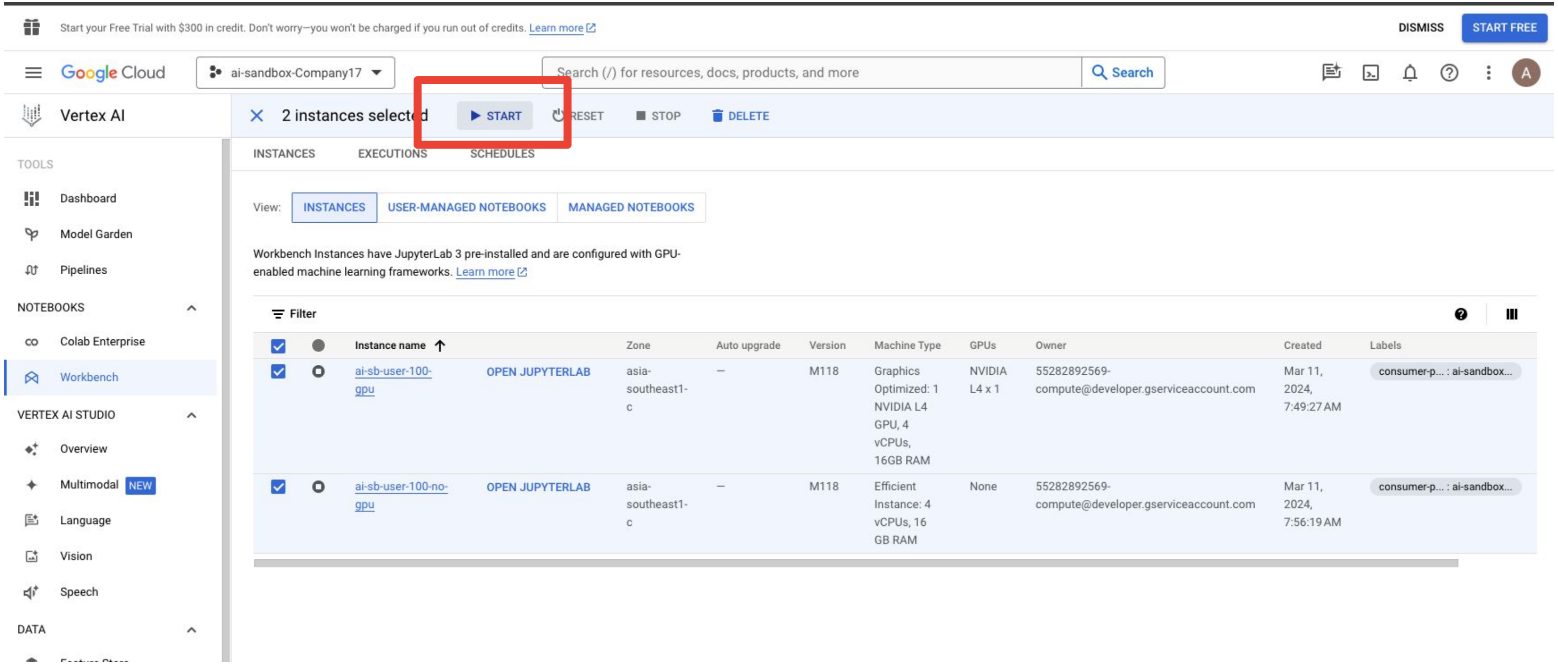
Workbench Instances have JupyterLab 3 pre-installed and are configured with GPU-enabled machine learning frameworks. Learn more

Filter

□	●	Instance name ↑		Zone	Auto upgrade	Version	Machine Type	GPUs	Owner	Created	Labels
□	✓	ai-gpu-100-3ffb19a7	OPEN JUPYTERLAB	asia-southeast1-c	—	M118	Graphics Optimized: 1 NVIDIA L4 GPU, 4 vCPUs, 16GB RAM	NVIDIA L4 x 1	55282892569-compute@developer.gserviceaccount.com	Mar 11, 2024, 7:49:27 AM	consumer-p... : ai-sandbox...
□	✓	ai-sb-user-100-no-gpu	OPEN JUPYTERLAB	asia-southeast1-c	—	M118	Efficient Instance: 4 vCPUs, 16 GB RAM	None	55282892569-compute@developer.gserviceaccount.com	Mar 11, 2024, 7:56:19 AM	consumer-p... : ai-sandbox...

Labs Setup: Accessing Workbench Notebook

Step 6: Select the 2 instances assigned and click on **START** in the top menu. If the instances are already started, skip to step 8.



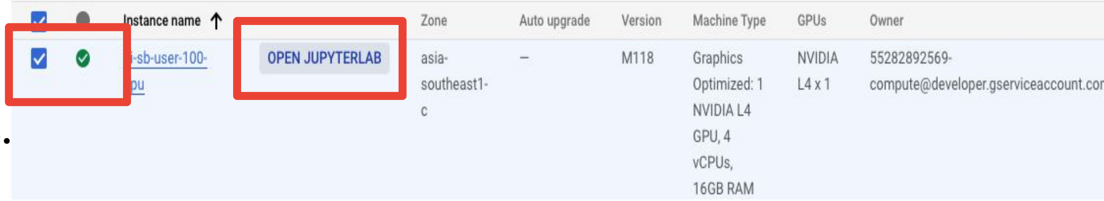
Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more

Workbench Instances have JupyterLab 3 pre-installed and are configured with GPU-enabled machine learning frameworks. Learn more

☑	●	Instance name ↑		Zone	Auto upgrade	Version	Machine Type	GPUs	Owner	Created	Labels
☑	○	ai-sb-user-100-gpu	OPEN JUPYTERLAB	asia-southeast1-c	—	M118	Graphics Optimized: 1 NVIDIA L4 GPU, 4 vCPUs, 16GB RAM	NVIDIA L4 x 1	55282892569-compute@developer.gserviceaccount.com	Mar 11, 2024, 7:49:27 AM	consumer-p... : ai-sandbox...
☑	○	ai-sb-user-100-no-gpu	OPEN JUPYTERLAB	asia-southeast1-c	—	M118	Efficient Instance: 4 vCPUs, 16 GB RAM	None	55282892569-compute@developer.gserviceaccount.com	Mar 11, 2024, 7:56:19 AM	consumer-p... : ai-sandbox...

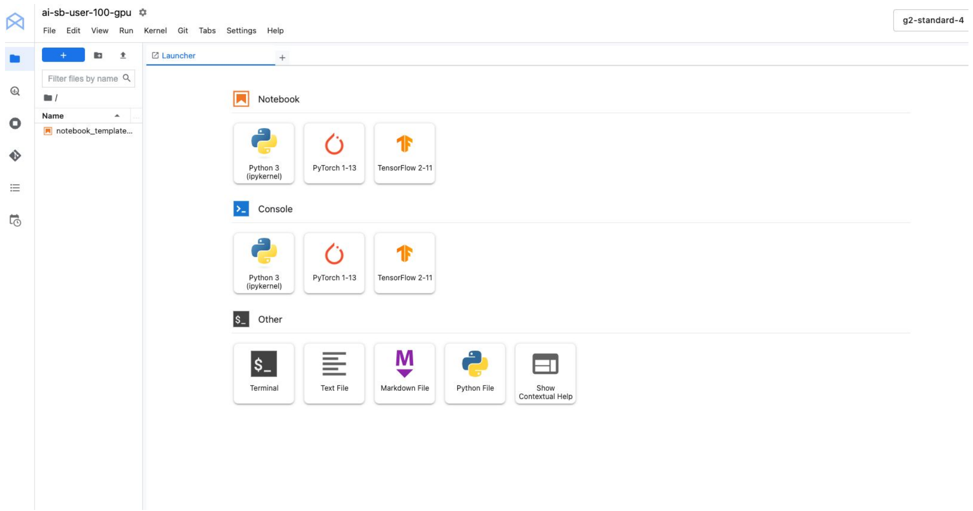
Labs Setup: Accessing Workbench Notebook

Step 7: Wait 1-3 minutes for the instances to start. Verify that there is a green tick beside the instances.



☑	●	Instance name ↑		Zone	Auto upgrade	Version	Machine Type	GPUs	Owner
☑	✓	i-sb-user-100-pu	OPEN JUPYTERLAB	asia-southeast1-c	–	M118	Graphics Optimized: 1 NVIDIA L4 GPU, 4 vCPUs, 16GB RAM	NVIDIA L4 x 1	55282892569-compute@developer.gserviceaccount.co

Step 8: Click on **OPEN JUPYTERLAB** for the **gpu instance** <ai-gpu-100-XXX>.
A new tab will open with access to the Jupyter Notebook.

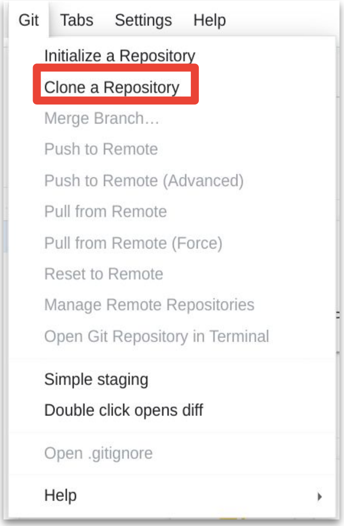


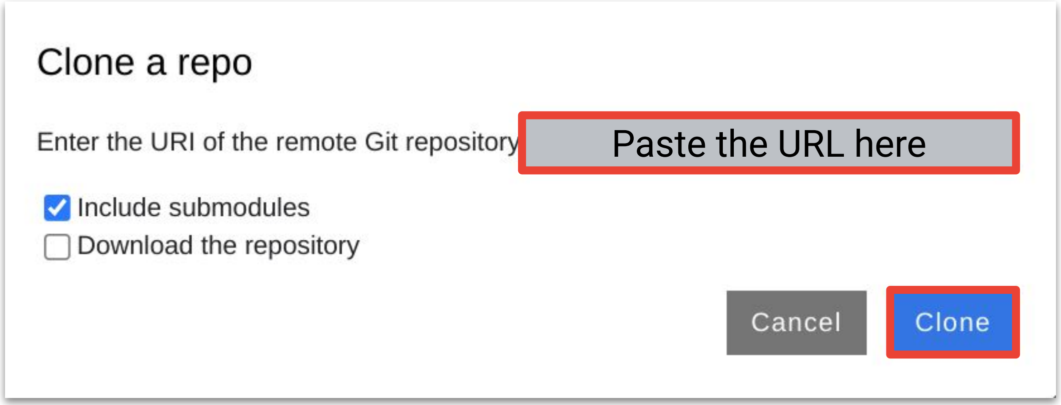
Labs Setup: Accessing Workbench Notebook

Step 9: Expand the **Git** menu and click on **Clone a Repository**

Step 10: Paste the URL provided for you into the text box and click **Clone**

URL = https://github.com/analyticsrepo01/pru-ai-labs.git





Enter the URI of the remote Git repository

Paste the URL here

☑ Include submodules

□ Download the repository

Labs Setup: Setting Up Notebooks

Step 12: Expand the **Run** menu and select **Run All Cells**

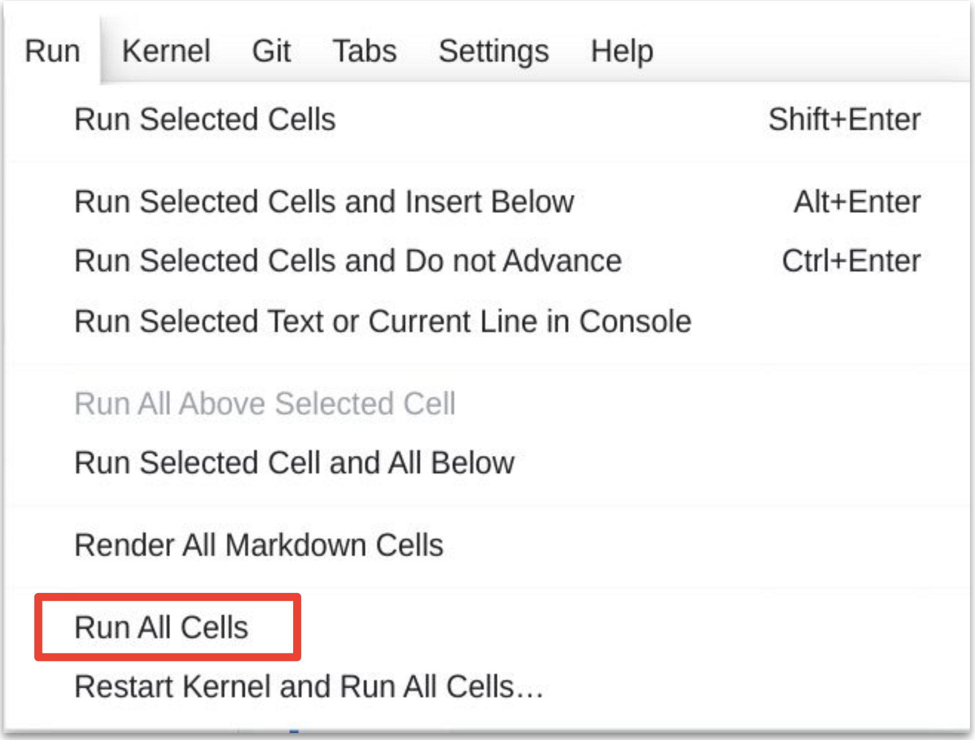
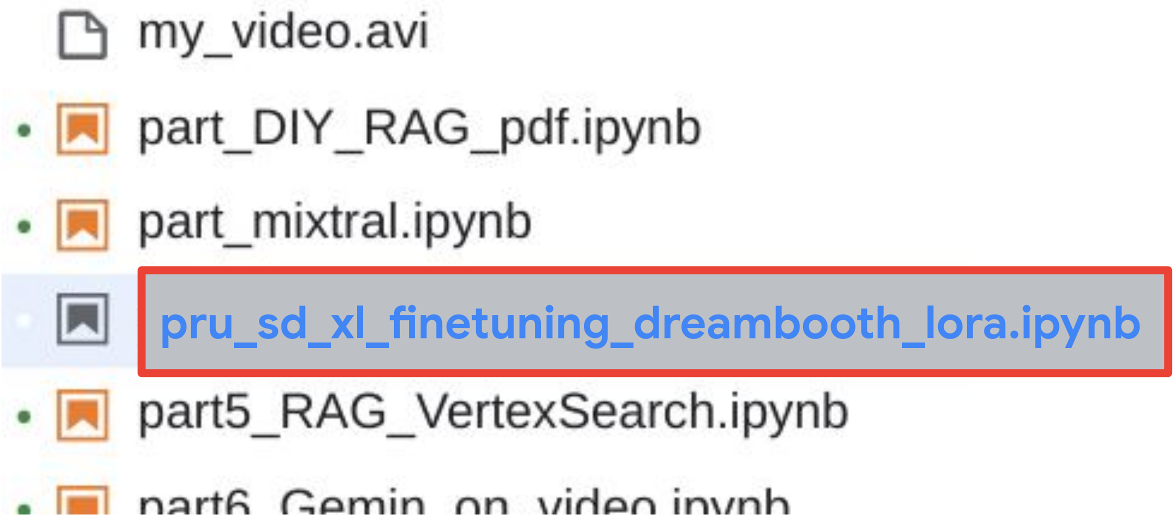

Run
Kernel
Git
Tabs
Settings
Help
Run Selected Cells Shift+Enter
Run Selected Cells and Insert Below Alt+Enter
Run Selected Cells and Do not Advance Ctrl+Enter
Run Selected Text or Current Line in Console
Run All Above Selected Cell
Run Selected Cell and All Below
Render All Markdown Cells
Run All Cells
Restart Kernel and Run All Cells…

Image model with custom training

Labs Setup: Setting Up Notebooks

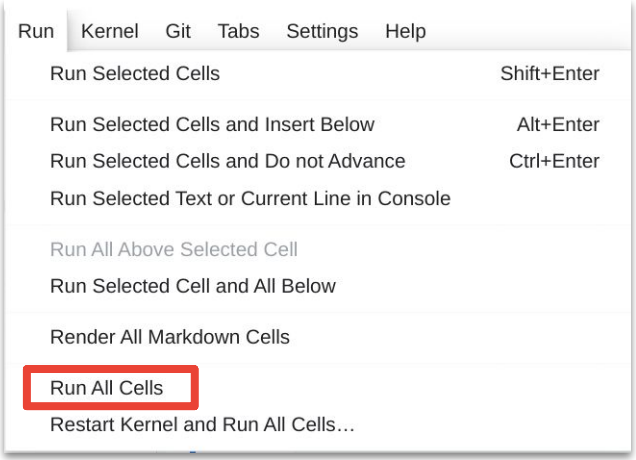
Step 11: Double click pru_sd_xl_finetuning_dreambooth_lora.ipynb to open the next notebook



my_video.avi
part_DIY_RAG_pdf.ipynb
part_mixtral.ipynb
pru_sd_xl_finetuning_dreambooth_lora.ipynb
part5_RAG_VertexSearch.ipynb
part6_Gemin_on_video.ipynb

Labs Setup: Setting Up Notebooks

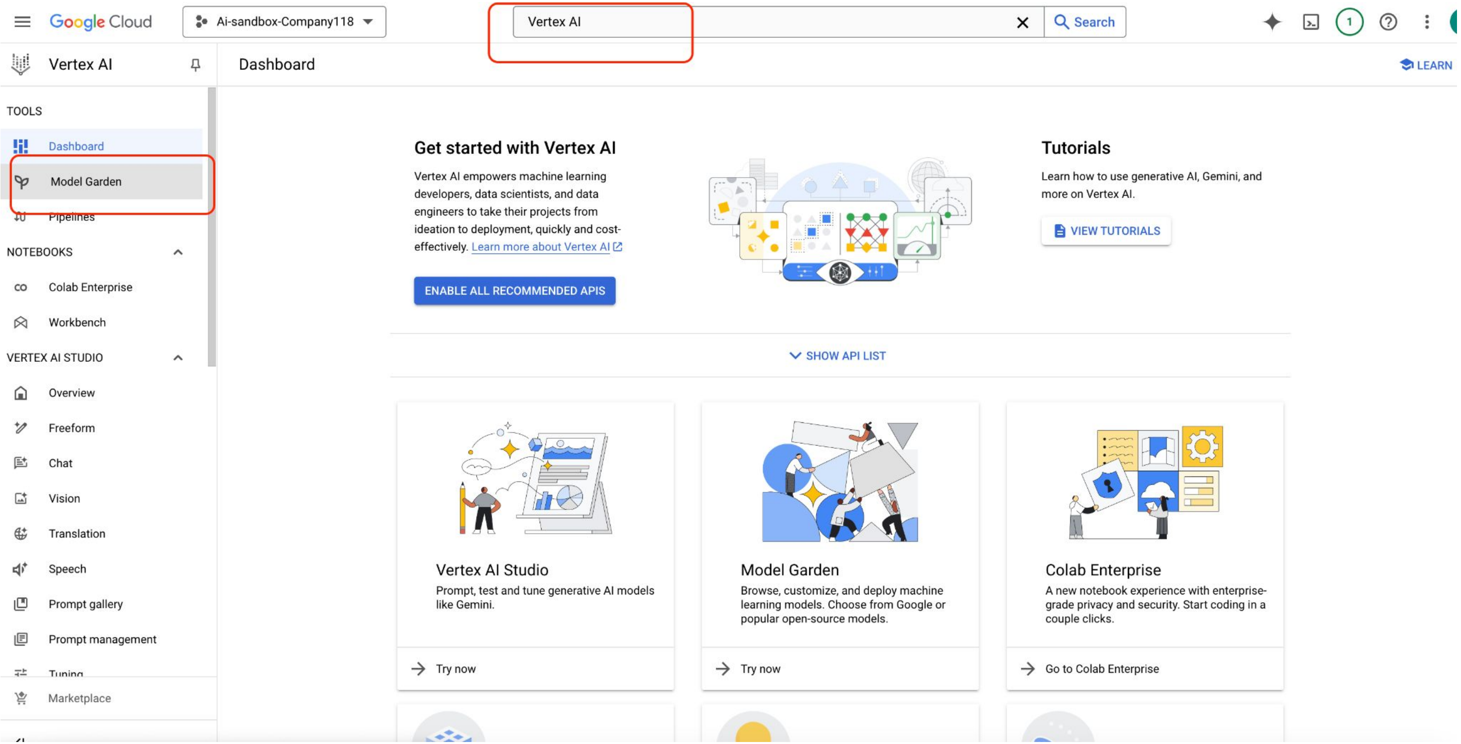
Step 12: Expand the **Run** menu and select **Run All Cells**.


Run
Kernel
Git
Tabs
Settings
Help
Run Selected Cells Shift+Enter
Run Selected Cells and Insert Below Alt+Enter
Run Selected Cells and Do not Advance Ctrl+Enter
Run Selected Text or Current Line in Console
Run All Above Selected Cell
Run Selected Cell and All Below
Render All Markdown Cells
Run All Cells
Restart Kernel and Run All Cells…

Open LLM with custom training

Labs Setup: Tuning a Open source LLM

Step 13: On Google Cloud Console, Navigate to Vertex AI → Model Garden



Get started with Vertex AI

Vertex AI empowers machine learning developers, data scientists, and data engineers to take their projects from ideation to deployment, quickly and cost-effectively. Learn more about Vertex AI

Tutorials

Learn how to use generative AI, Gemini, and more on Vertex AI.

Vertex AI Studio

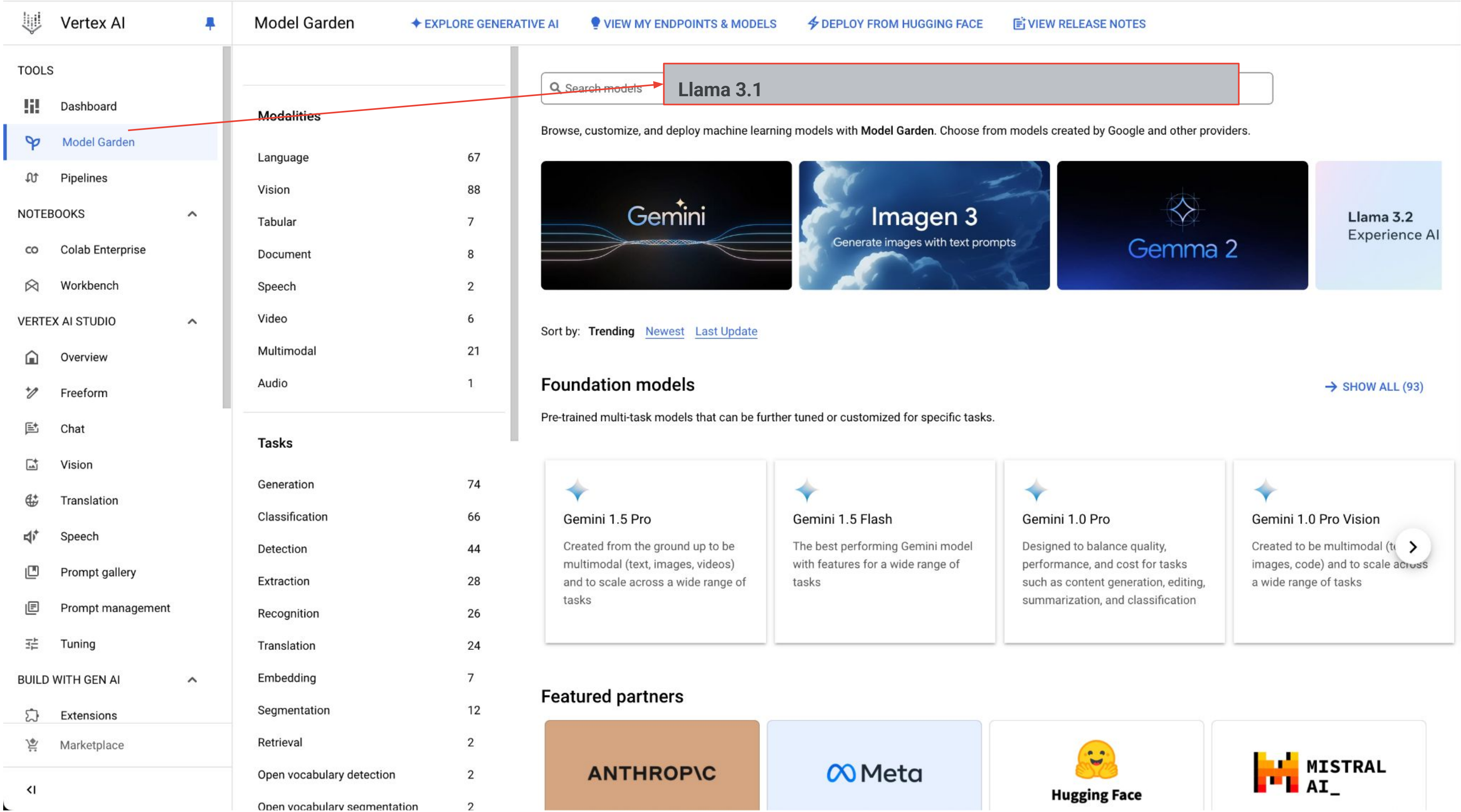
Prompt, test and tune generative AI models like Gemini.

Model Garden

Browse, customize, and deploy machine learning models. Choose from Google or popular open-source models.

Colab Enterprise

A new notebook experience with enterprise-grade privacy and security. Start coding in a couple clicks.



Vertex AI

Model Garden

✦ EXPLORE GENERATIVE AI

VIEW MY ENDPOINTS & MODELS

DEPLOY FROM HUGGING FACE

VIEW RELEASE NOTES

TOOLS

- Dashboard
- Model Garden
- Pipelines

NOTEBOOKS

- Colab Enterprise
- Workbench

VERTEX AI STUDIO

- Overview
- Freeform
- Chat
- Vision
- Translation
- Speech
- Prompt gallery
- Prompt management
- Tuning

BUILD WITH GEN AI

- Extensions
- Marketplace

Modalities

Language	67
Vision	88
Tabular	7
Document	8
Speech	2
Video	6
Multimodal	21
Audio	1

Tasks

Generation	74
Classification	66
Detection	44
Extraction	28
Recognition	26
Translation	24
Embedding	7
Segmentation	12
Retrieval	2
Open vocabulary detection	2
Open vocabulary segmentation	2

Search models

Llama 3.1

Browse, customize, and deploy machine learning models with **Model Garden**. Choose from models created by Google and other providers.

Gemini

Imagen 3
Generate images with text prompts

Gemma 2

Llama 3.2
Experience AI

Sort by: **Trending** Newest Last Update

Foundation models

→ SHOW ALL (93)

Pre-trained multi-task models that can be further tuned or customized for specific tasks.

Gemini 1.5 Pro
Created from the ground up to be multimodal (text, images, videos) and to scale across a wide range of tasks

Gemini 1.5 Flash
The best performing Gemini model with features for a wide range of tasks

Gemini 1.0 Pro
Designed to balance quality, performance, and cost for tasks such as content generation, editing, summarization, and classification

Gemini 1.0 Pro Vision
Created to be multimodal (t… images, code) and to scale across a wide range of tasks

Featured partners

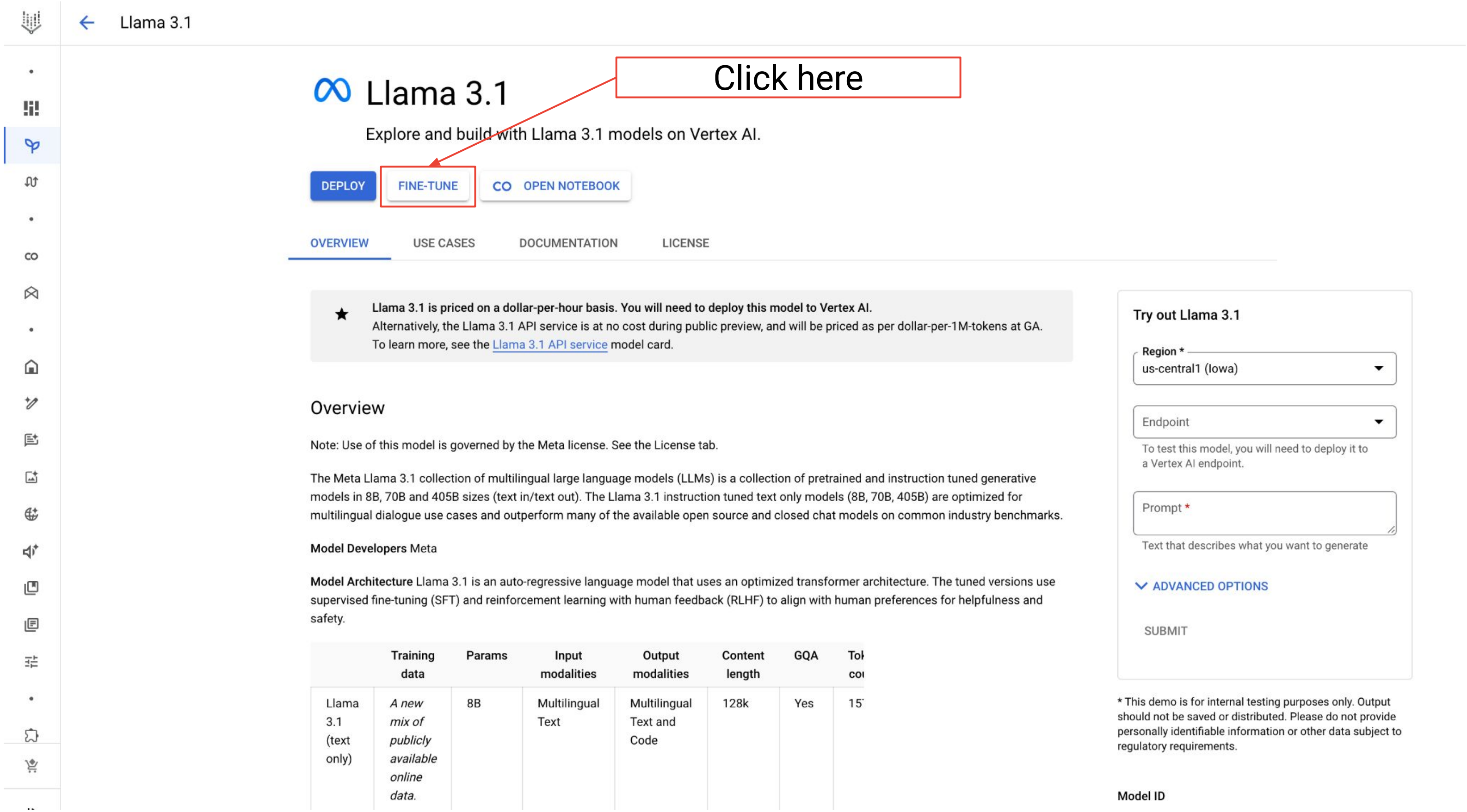
ANTHROPIC

Meta

Hugging Face

MISTRAL AI_

Step 14: Choose Llama 3.1 model, Click on “FINE-TUNE”



Click here

Llama 3.1

Explore and build with Llama 3.1 models on Vertex AI.

★ **Llama 3.1 is priced on a dollar-per-hour basis. You will need to deploy this model to Vertex AI.**
Alternatively, the Llama 3.1 API service is at no cost during public preview, and will be priced as per dollar-per-1M-tokens at GA.
To learn more, see the Llama 3.1 API service model card.

Overview

Note: Use of this model is governed by the Meta license. See the License tab.

The Meta Llama 3.1 collection of multilingual large language models (LLMs) is a collection of pretrained and instruction tuned generative models in 8B, 70B and 405B sizes (text in/text out). The Llama 3.1 instruction tuned text only models (8B, 70B, 405B) are optimized for multilingual dialogue use cases and outperform many of the available open source and closed chat models on common industry benchmarks.

Model Developers Meta

Model Architecture Llama 3.1 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety.

	Training data	Params	Input modalities	Output modalities	Content length	GQA	Tok co
Llama 3.1 (text only)	*A new mix of publicly available online data.*	8B	Multilingual Text	Multilingual Text and Code	128k	Yes	15

Try out Llama 3.1

Region *
us-central1 (Iowa)

Endpoint

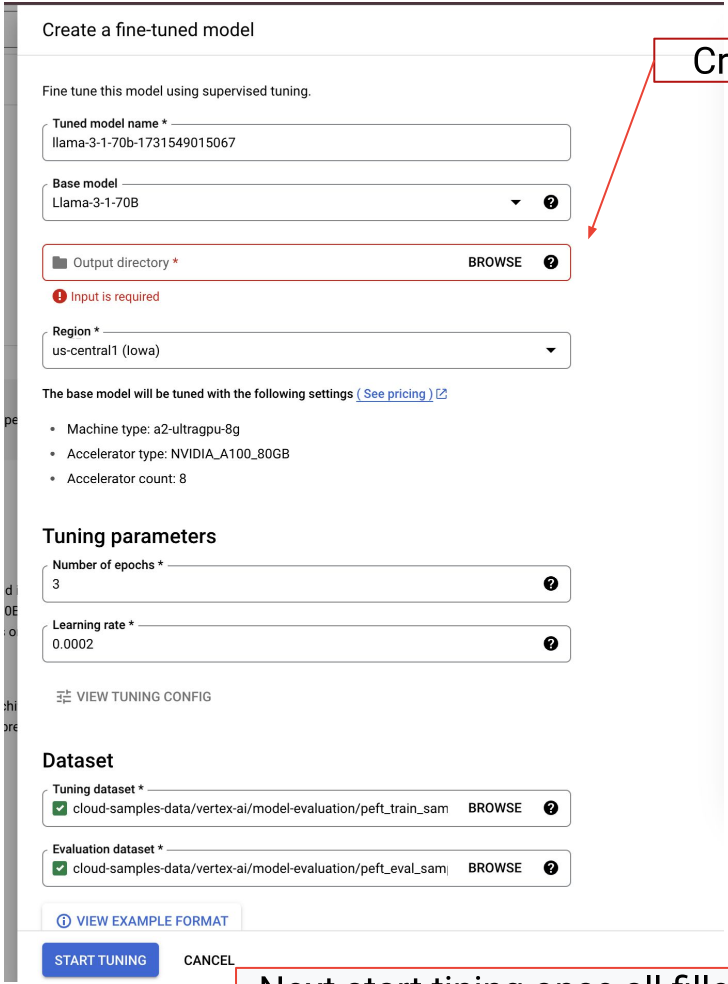
To test this model, you will need to deploy it to a Vertex AI endpoint.

Prompt *

Text that describes what you want to generate

* This demo is for internal testing purposes only. Output should not be saved or distributed. Please do not provide personally identifiable information or other data subject to regulatory requirements.

Model ID



Create a fine-tuned model

Fine tune this model using supervised tuning.

Tuned model name *: llama-3-1-70b-1731549015067

Base model: Llama-3-1-70B

Output directory *

Input is required

Region *: us-central1 (Iowa)

The base model will be tuned with the following settings (See pricing)

- Machine type: a2-ultragpu-8g
- Accelerator type: NVIDIA_A100_80GB
- Accelerator count: 8

Tuning parameters

Number of epochs *: 3

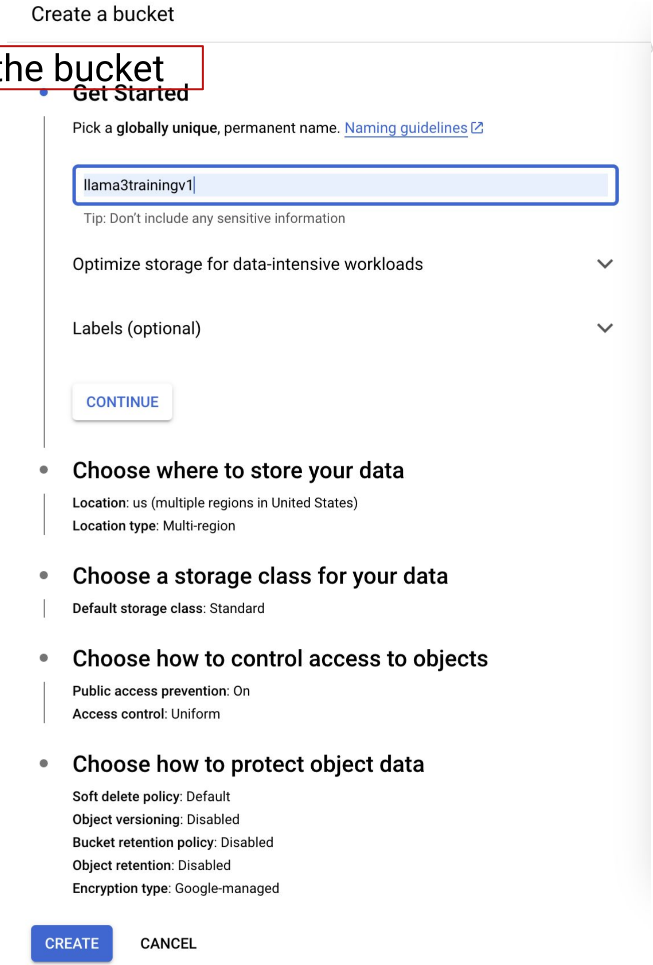
Learning rate *: 0.0002

Dataset

Tuning dataset *: ☑ cloud-samples-data/vertex-ai/model-evaluation/peft_train_sam

Evaluation dataset *: ☑ cloud-samples-data/vertex-ai/model-evaluation/peft_eval_sam

Create or select the bucket



Create a bucket

Get Started

Pick a **globally unique**, permanent name. Naming guidelines

llama3trainingv1

Tip: Don't include any sensitive information

Optimize storage for data-intensive workloads

Labels (optional)

Choose where to store your data

Location: us (multiple regions in United States)
Location type: Multi-region

Choose a storage class for your data

Default storage class: Standard

Choose how to control access to objects

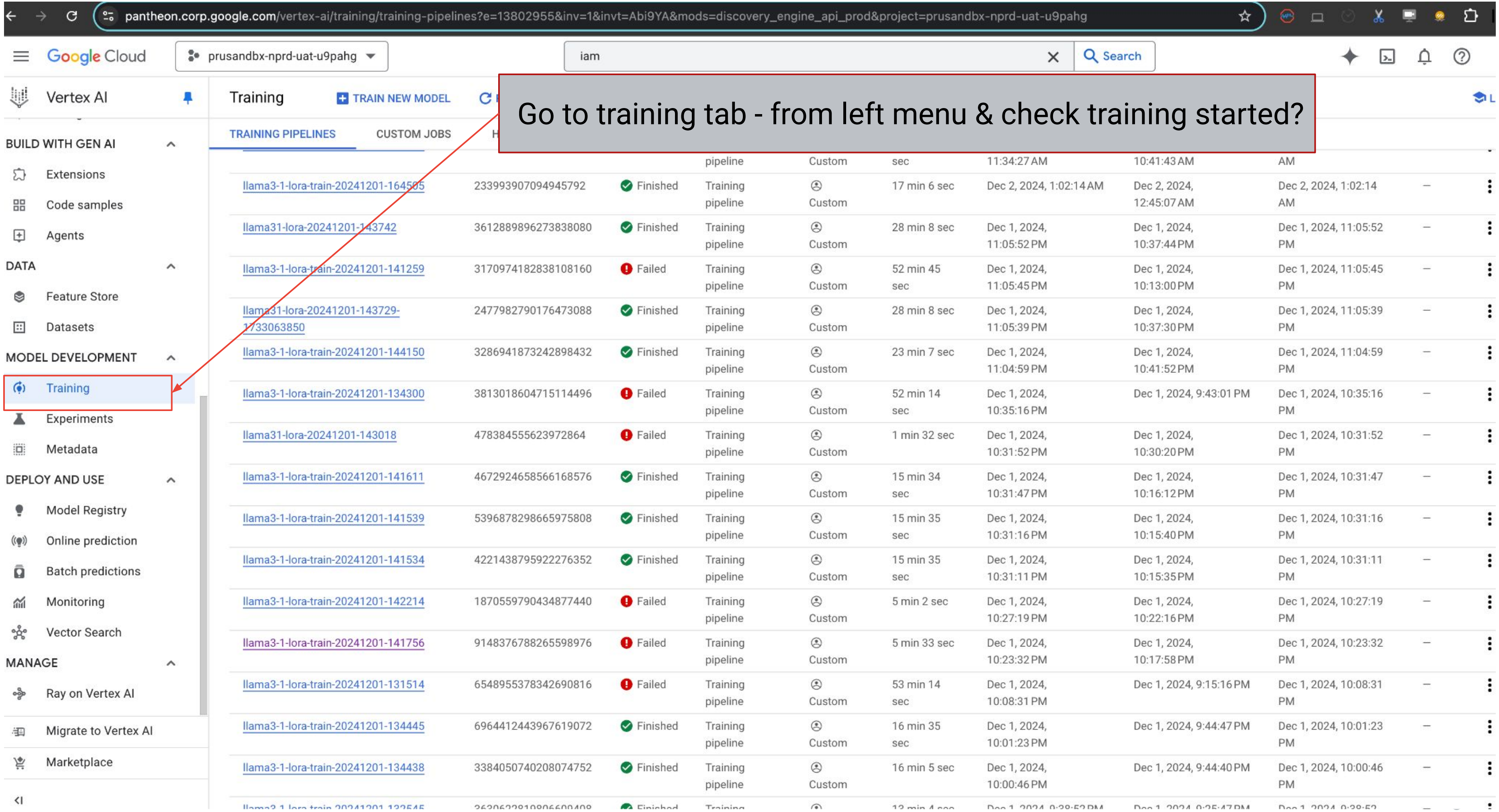
Public access prevention: On
Access control: Uniform

Choose how to protect object data

Soft delete policy: Default
Object versioning: Disabled
Bucket retention policy: Disabled
Object retention: Disabled
Encryption type: Google-managed

Next start tining once all filled up

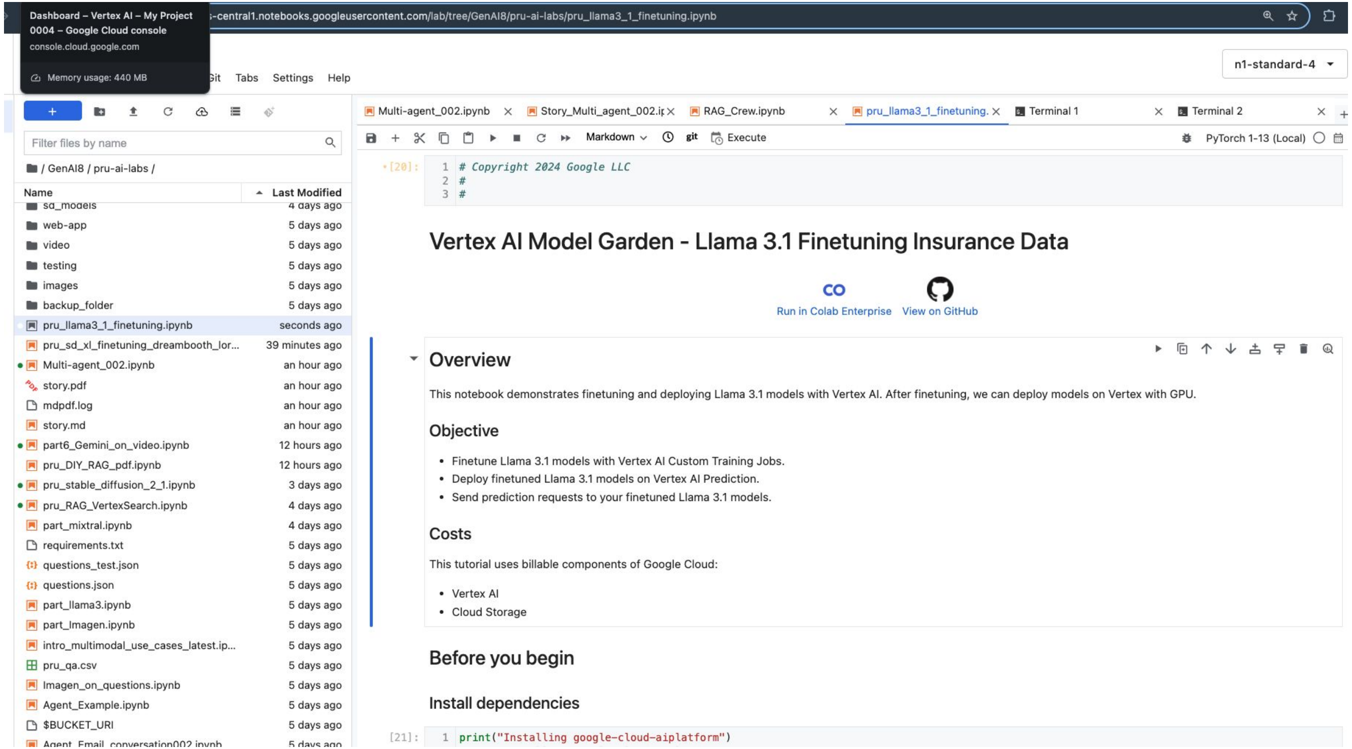
Go to training tab - from left menu & check training started?



			pipeline	Custom	sec	11:34:27 AM	10:41:43 AM	AM	
llama3-1-lora-train-20241201-164505	233993907094945792	✓ Finished	Training pipeline	Custom	17 min 6 sec	Dec 2, 2024, 1:02:14 AM	Dec 2, 2024, 12:45:07 AM	Dec 2, 2024, 1:02:14 AM	–
llama31-lora-20241201-143742	3612889896273838080	✓ Finished	Training pipeline	Custom	28 min 8 sec	Dec 1, 2024, 11:05:52 PM	Dec 1, 2024, 10:37:44 PM	Dec 1, 2024, 11:05:52 PM	–
llama3-1-lora-train-20241201-141259	3170974182838108160	! Failed	Training pipeline	Custom	52 min 45 sec	Dec 1, 2024, 11:05:45 PM	Dec 1, 2024, 10:13:00 PM	Dec 1, 2024, 11:05:45 PM	–
llama31-lora-20241201-143729-1733063850	2477982790176473088	✓ Finished	Training pipeline	Custom	28 min 8 sec	Dec 1, 2024, 11:05:39 PM	Dec 1, 2024, 10:37:30 PM	Dec 1, 2024, 11:05:39 PM	–
llama3-1-lora-train-20241201-144150	3286941873242898432	✓ Finished	Training pipeline	Custom	23 min 7 sec	Dec 1, 2024, 11:04:59 PM	Dec 1, 2024, 10:41:52 PM	Dec 1, 2024, 11:04:59 PM	–
llama3-1-lora-train-20241201-134300	3813018604715114496	! Failed	Training pipeline	Custom	52 min 14 sec	Dec 1, 2024, 10:35:16 PM	Dec 1, 2024, 9:43:01 PM	Dec 1, 2024, 10:35:16 PM	–
llama31-lora-20241201-143018	478384555623972864	! Failed	Training pipeline	Custom	1 min 32 sec	Dec 1, 2024, 10:31:52 PM	Dec 1, 2024, 10:30:20 PM	Dec 1, 2024, 10:31:52 PM	–
llama3-1-lora-train-20241201-141611	4672924658566168576	✓ Finished	Training pipeline	Custom	15 min 34 sec	Dec 1, 2024, 10:31:47 PM	Dec 1, 2024, 10:16:12 PM	Dec 1, 2024, 10:31:47 PM	–
llama3-1-lora-train-20241201-141539	5396878298665975808	✓ Finished	Training pipeline	Custom	15 min 35 sec	Dec 1, 2024, 10:31:16 PM	Dec 1, 2024, 10:15:40 PM	Dec 1, 2024, 10:31:16 PM	–
llama3-1-lora-train-20241201-141534	4221438795922276352	✓ Finished	Training pipeline	Custom	15 min 35 sec	Dec 1, 2024, 10:31:11 PM	Dec 1, 2024, 10:15:35 PM	Dec 1, 2024, 10:31:11 PM	–
llama3-1-lora-train-20241201-142214	1870559790434877440	! Failed	Training pipeline	Custom	5 min 2 sec	Dec 1, 2024, 10:27:19 PM	Dec 1, 2024, 10:22:16 PM	Dec 1, 2024, 10:27:19 PM	–
llama3-1-lora-train-20241201-141756	9148376788265598976	! Failed	Training pipeline	Custom	5 min 33 sec	Dec 1, 2024, 10:23:32 PM	Dec 1, 2024, 10:17:58 PM	Dec 1, 2024, 10:23:32 PM	–
llama3-1-lora-train-20241201-131514	6548955378342690816	! Failed	Training pipeline	Custom	53 min 14 sec	Dec 1, 2024, 10:08:31 PM	Dec 1, 2024, 9:15:16 PM	Dec 1, 2024, 10:08:31 PM	–
llama3-1-lora-train-20241201-134445	6964412443967619072	✓ Finished	Training pipeline	Custom	16 min 35 sec	Dec 1, 2024, 10:01:23 PM	Dec 1, 2024, 9:44:47 PM	Dec 1, 2024, 10:01:23 PM	–
llama3-1-lora-train-20241201-134438	3384050740208074752	✓ Finished	Training pipeline	Custom	16 min 5 sec	Dec 1, 2024, 10:00:46 PM	Dec 1, 2024, 9:44:40 PM	Dec 1, 2024, 10:00:46 PM	–
llama3-1-lora-train-20241201-132545	3630622810806600408	✓ Finished	Training		13 min 4 sec	Dec 1, 2024, 9:38:52 PM	Dec 1, 2024, 9:25:47 PM	Dec 1, 2024, 9:38:52	–

Example on custom data: Llama3.1 on Insurance data

Open and run all the notebook pru_llama3_1_finetuning.ipynb



Dashboard – Vertex AI – My Project 0004 – Google Cloud console
console.cloud.google.com
Memory usage: 440 MB

-central1.notebooks.googleusercontent.com/lab/tree/GenAI8/pru-ai-labs/pru_llama3_1_finetuning.ipynb

/ GenAI8 / pru-ai-labs /

Name	Last Modified
sd_models	4 days ago
web-app	5 days ago
video	5 days ago
testing	5 days ago
images	5 days ago
backup_folder	5 days ago
pru_llama3_1_finetuning.ipynb	seconds ago
pru_sd_xl_finetuning_dreambooth_lor...	39 minutes ago
Multi-agent_002.ipynb	an hour ago
story.pdf	an hour ago
mdpdf.log	an hour ago
story.md	an hour ago
part6_Gemini_on_video.ipynb	12 hours ago
pru_DIY_RAG_pdf.ipynb	12 hours ago
pru_stable_diffusion_2_1.ipynb	3 days ago
pru_RAG_VertexSearch.ipynb	4 days ago
part_mixtral.ipynb	4 days ago
requirements.txt	5 days ago
questions_test.json	5 days ago
questions.json	5 days ago
part_llama3.ipynb	5 days ago
part_Imagen.ipynb	5 days ago
intro_multimodal_use_cases_latest.ip...	5 days ago
pru_qa.csv	5 days ago
Imagen_on_questions.ipynb	5 days ago
Agent_Example.ipynb	5 days ago
$BUCKET_URI	5 days ago
Agent_Email_conversation002.ipynb	5 days ago

[20]:
```
# Copyright 2024 Google LLC
#
#
```

Vertex AI Model Garden - Llama 3.1 Finetuning Insurance Data

Overview

This notebook demonstrates finetuning and deploying Llama 3.1 models with Vertex AI. After finetuning, we can deploy models on Vertex with GPU.

Objective

- Finetune Llama 3.1 models with Vertex AI Custom Training Jobs.
- Deploy finetuned Llama 3.1 models on Vertex AI Prediction.
- Send prediction requests to your finetuned Llama 3.1 models.

Costs

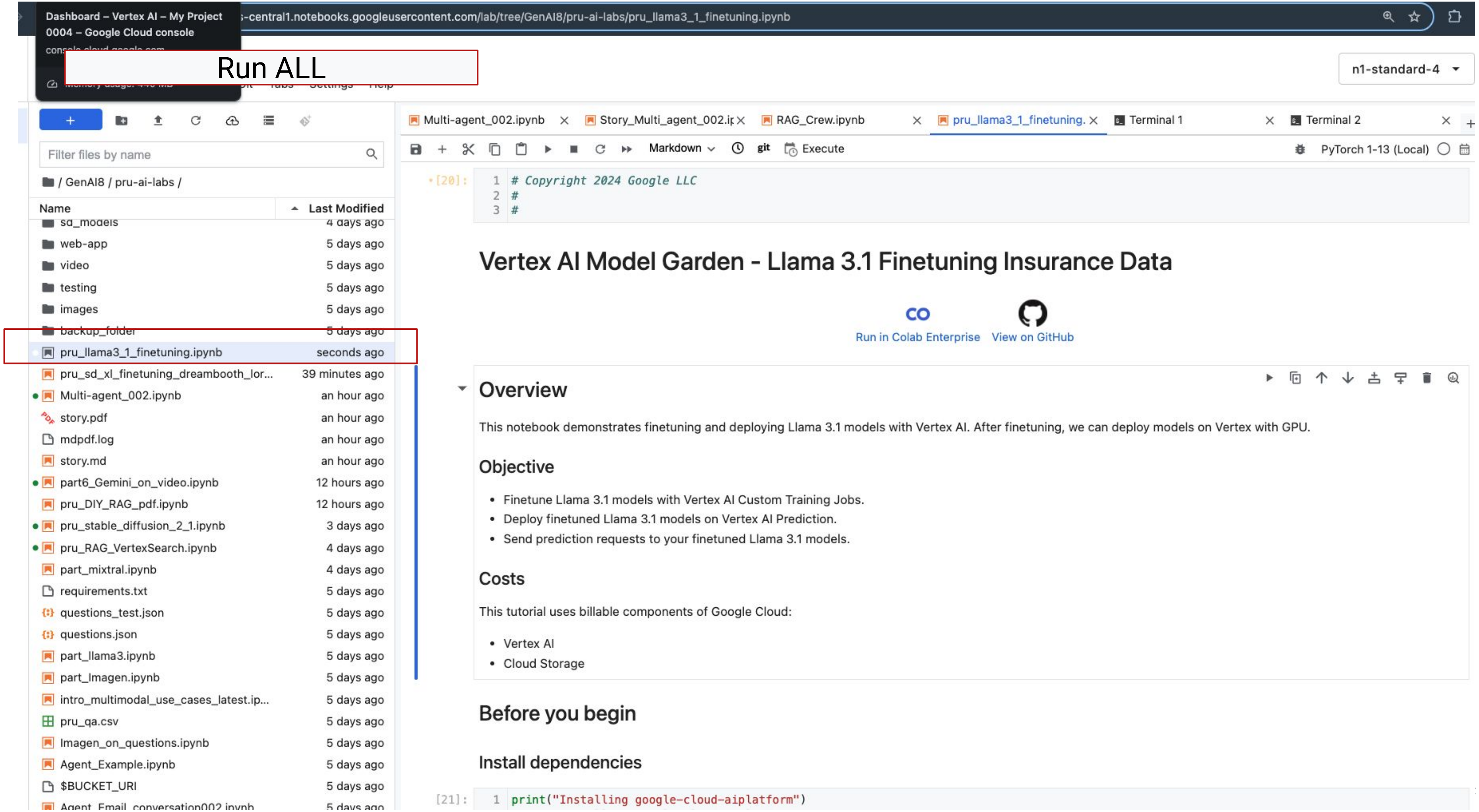
This tutorial uses billable components of Google Cloud:

- Vertex AI
- Cloud Storage

Before you begin

Install dependencies

[21]:
```
print("Installing google-cloud-aiplatform")
```



Run ALL

Name	Last Modified
sd_models	4 days ago
web-app	5 days ago
video	5 days ago
testing	5 days ago
images	5 days ago
backup_folder	5 days ago
pru_llama3_1_finetuning.ipynb	seconds ago
pru_sd_xl_finetuning_dreambooth_lor...	39 minutes ago
Multi-agent_002.ipynb	an hour ago
story.pdf	an hour ago
mdpdf.log	an hour ago
story.md	an hour ago
part6_Gemini_on_video.ipynb	12 hours ago
pru_DIY_RAG_pdf.ipynb	12 hours ago
pru_stable_diffusion_2_1.ipynb	3 days ago
pru_RAG_VertexSearch.ipynb	4 days ago
part_mixtral.ipynb	4 days ago
requirements.txt	5 days ago
questions_test.json	5 days ago
questions.json	5 days ago
part_llama3.ipynb	5 days ago
part_Imagen.ipynb	5 days ago
intro_multimodal_use_cases_latest.ip...	5 days ago
pru_qa.csv	5 days ago
Imagen_on_questions.ipynb	5 days ago
Agent_Example.ipynb	5 days ago
$BUCKET_URI	5 days ago
Agent_Email_conversation002.ipynb	5 days ago

[20]:
```
# Copyright 2024 Google LLC
#
#
```

Vertex AI Model Garden - Llama 3.1 Finetuning Insurance Data

Run in Colab Enterprise   View on GitHub

Overview

This notebook demonstrates finetuning and deploying Llama 3.1 models with Vertex AI. After finetuning, we can deploy models on Vertex with GPU.

Objective

- Finetune Llama 3.1 models with Vertex AI Custom Training Jobs.
- Deploy finetuned Llama 3.1 models on Vertex AI Prediction.
- Send prediction requests to your finetuned Llama 3.1 models.

Costs

This tutorial uses billable components of Google Cloud:

- Vertex AI
- Cloud Storage

Before you begin

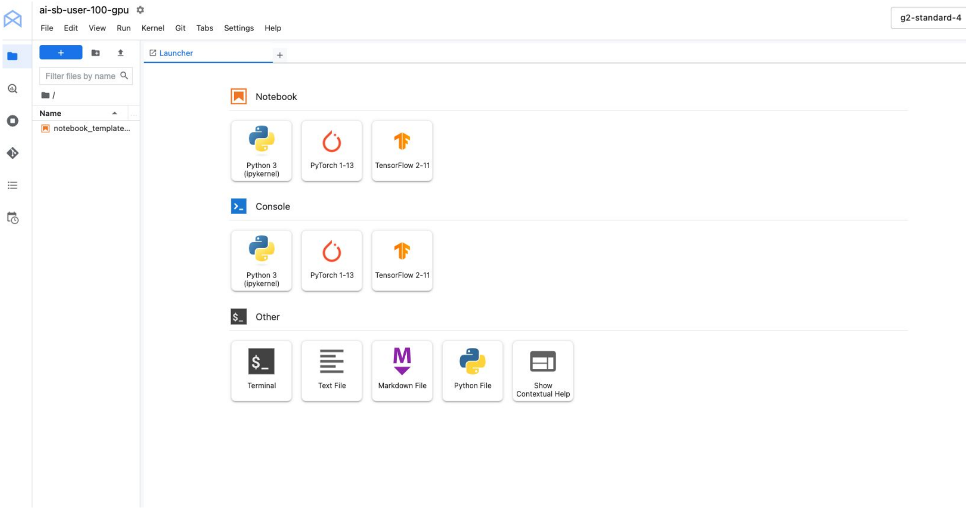
Install dependencies

[21]:
```
print("Installing google-cloud-aiplatform")
```

RAG for Prudential Use case

Labs Setup: Accessing Workbench Notebook

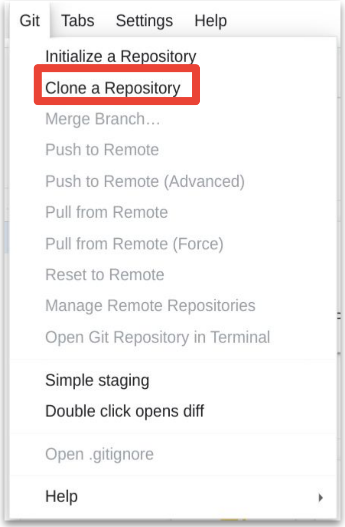
Step 15: Click on **OPEN JUPYTERLAB** for the **no-gpu** instance <ai-no-gpu-100-XXX>. A new tab will open with access to the Jupyter Notebook.

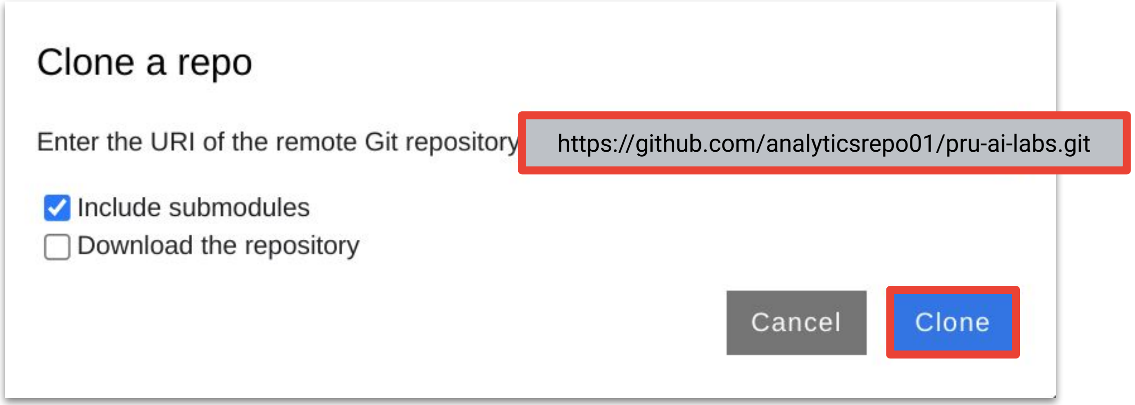


Labs Setup: Accessing Workbench Notebook

Step 16: Expand the **Git** menu and click on **Clone a Repository**

Step 17: Paste the URL provided for you into the text box and click **Clone**
URL = https://github.com/analyticsrepo01/pru-ai-labs.git





Clone a repo

Enter the URI of the remote Git repository: https://github.com/analyticsrepo01/pru-ai-labs.git

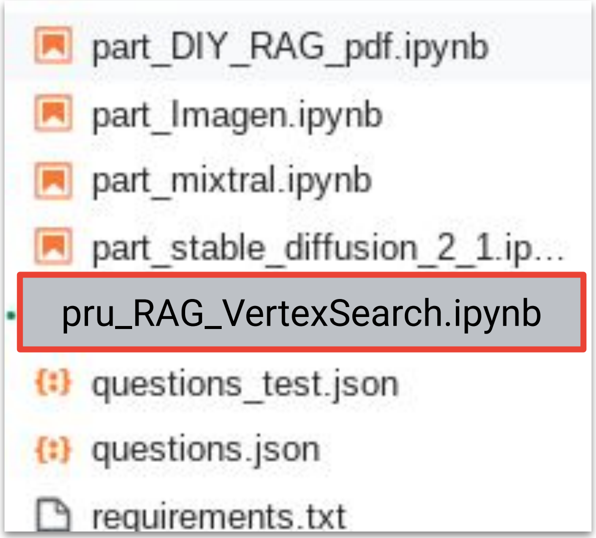
☑ Include submodules

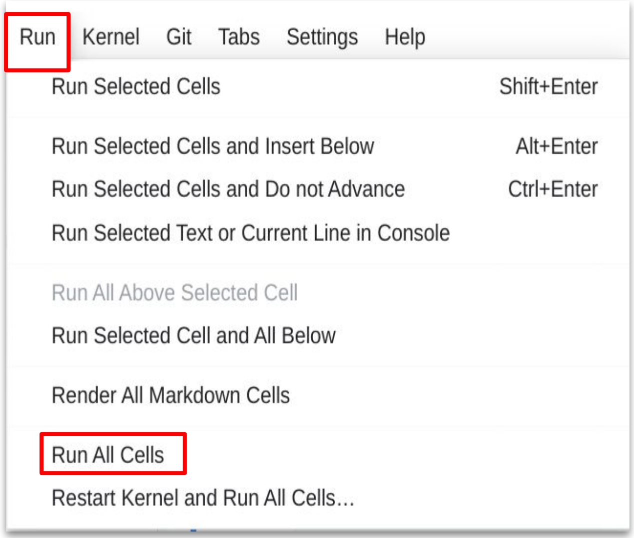
☐ Download the repository

Labs Setup: Vertex AI Search

Step 18: Double click **pru_RAG_VertexSearch.ipynb** to open the Jupyter Notebook

Step 19: In the top menu bar, expand the **Run** menu and select **Run All Cells**



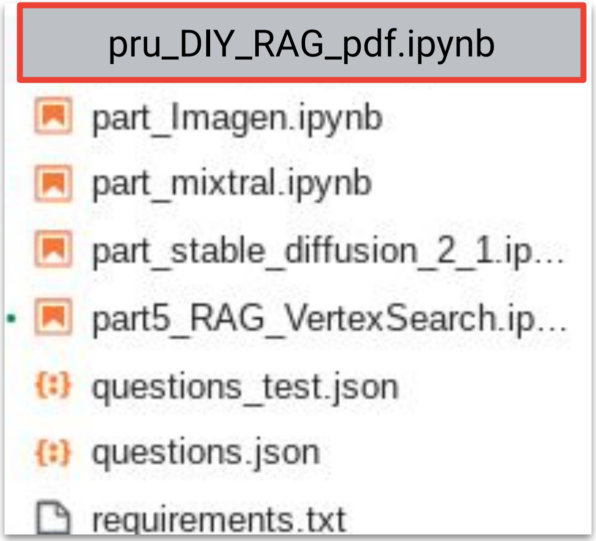

Run
Kernel
Git
Tabs
Settings
Help
Run Selected Cells Shift+Enter
Run Selected Cells and Insert Below Alt+Enter
Run Selected Cells and Do not Advance Ctrl+Enter
Run Selected Text or Current Line in Console
Run All Above Selected Cell
Run Selected Cell and All Below
Render All Markdown Cells
Run All Cells
Restart Kernel and Run All Cells...

DIY RAG

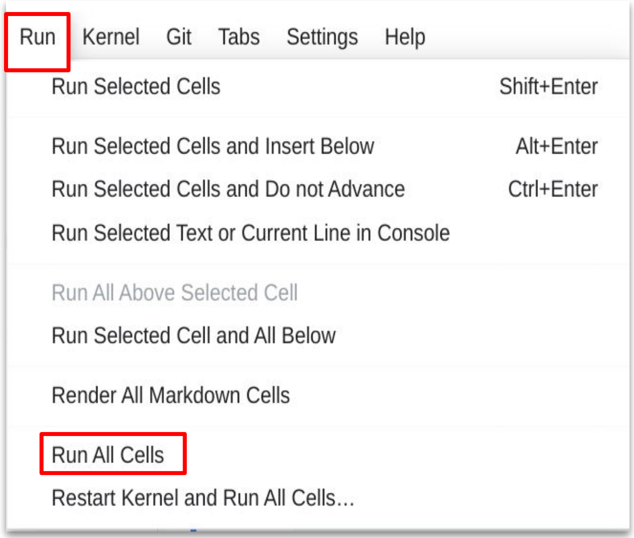
Labs Setup: Vector AI Search

Step 20: Double click pru_DIY_RAG_pdf.ipynb to open the Jupyter Notebook

Step 21: In the top menu bar, expand the **Run** menu and select **Run All Cells**



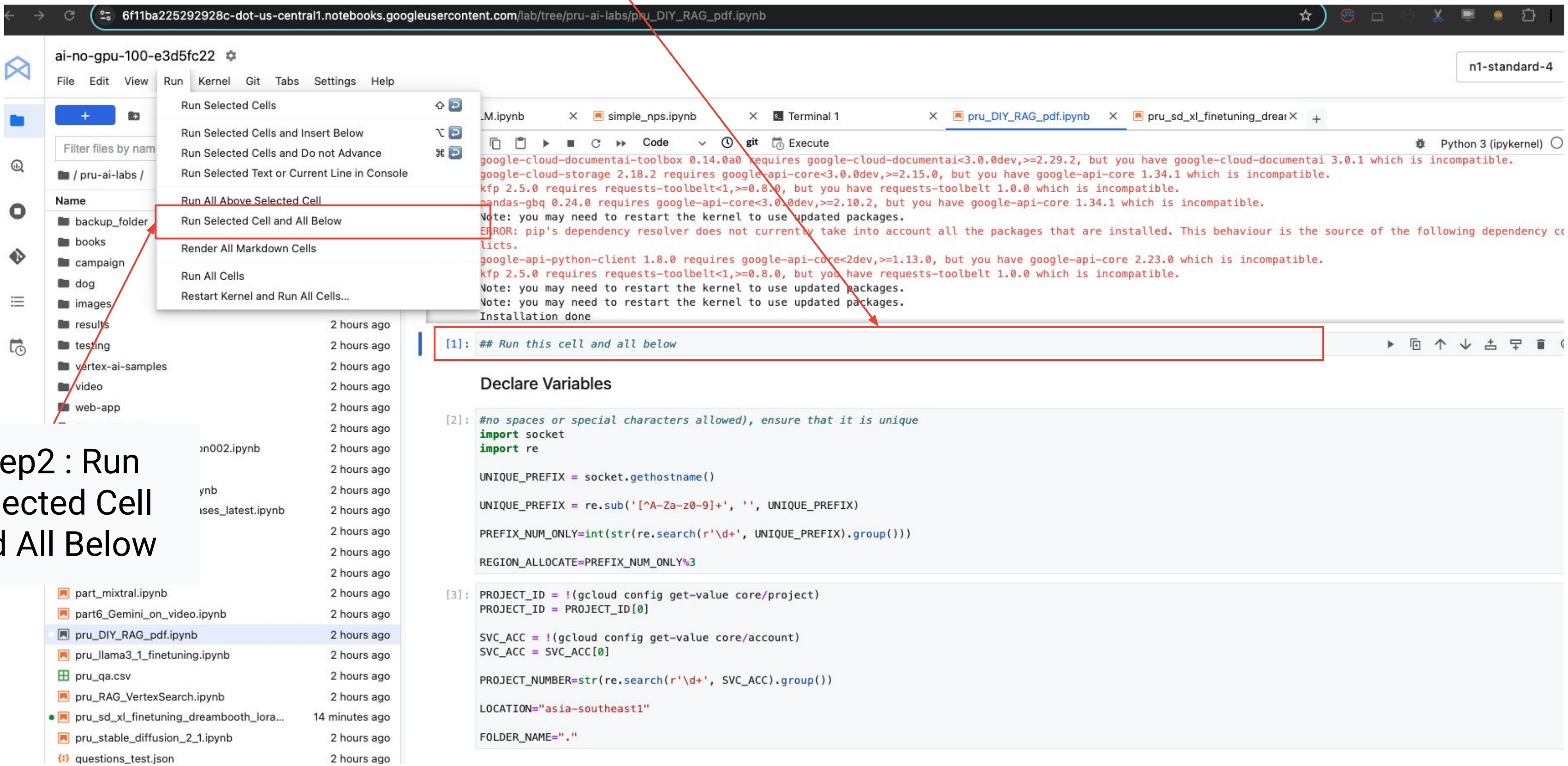
pru_DIY_RAG_pdf.ipynb
part_Imagen.ipynb
part_mixtral.ipynb
part_stable_diffusion_2_1.ip…
part5_RAG_VertexSearch.ip…
questions_test.json
questions.json
requirements.txt



Run
Kernel
Git
Tabs
Settings
Help
Run Selected Cells Shift+Enter
Run Selected Cells and Insert Below Alt+Enter
Run Selected Cells and Do not Advance Ctrl+Enter
Run Selected Text or Current Line in Console
Run All Above Selected Cell
Run Selected Cell and All Below
Render All Markdown Cells
Run All Cells
Restart Kernel and Run All Cells…

Kernel Restarts after installation

Step1 : Click on this cell after restart


6f11ba225292928c-dot-us-central1.notebooks.googleusercontent.com/lab/tree/pru-ai-labs/pru_DIY_RAG_pdf.ipynb
ai-no-gpu-100-e3d5fc22
n1-standard-4
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Run Selected Cells
Run Selected Cells and Insert Below
Run Selected Cells and Do not Advance
Run Selected Text or Current Line in Console
Run All Above Selected Cell
Run Selected Cell and All Below
Render All Markdown Cells
Run All Cells
Restart Kernel and Run All Cells...
+
Filter files by nam
/ pru-ai-labs /
.M.ipynb
simple_nps.ipynb
Terminal 1
pru_DIY_RAG_pdf.ipynb
pru_sd_xl_finetuning_drear
Code
git
Execute
Python 3 (ipykernel)

Name	
backup_folder	
books	
campaign	
dog	
images	
results	2 hours ago
testing	2 hours ago
vertex-ai-samples	2 hours ago
video	2 hours ago
web-app	2 hours ago
	2 hours ago
on002.ipynb	2 hours ago
	2 hours ago
ynb	2 hours ago
ases_latest.ipynb	2 hours ago
	2 hours ago
	2 hours ago
	2 hours ago
part_mixtral.ipynb	2 hours ago
part6_Gemini_on_video.ipynb	2 hours ago
pru_DIY_RAG_pdf.ipynb	2 hours ago
pru_llama3_1_finetuning.ipynb	2 hours ago
pru_qa.csv	2 hours ago
pru_RAG_VertexSearch.ipynb	2 hours ago
pru_sd_xl_finetuning_dreambooth_lora...	14 minutes ago
pru_stable_diffusion_2_1.ipynb	2 hours ago
questions_test.json	2 hours ago

```
google-cloud-documentai-toolbox 0.14.0a0 requires google-cloud-documentai<3.0.0dev,>=2.29.2, but you have google-cloud-documentai 3.0.1 which is incompatible.
google-cloud-storage 2.18.2 requires google-api-core<3.0.0dev,>=2.15.0, but you have google-api-core 1.34.1 which is incompatible.
kfp 2.5.0 requires requests-toolbelt<1,>=0.8.0, but you have requests-toolbelt 1.0.0 which is incompatible.
pandas-gbq 0.24.0 requires google-api-core<3.0.0dev,>=2.10.2, but you have google-api-core 1.34.1 which is incompatible.
Note: you may need to restart the kernel to use updated packages.
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency co
licts.
google-api-python-client 1.8.0 requires google-api-core<2dev,>=1.13.0, but you have google-api-core 2.23.0 which is incompatible.
kfp 2.5.0 requires requests-toolbelt<1,>=0.8.0, but you have requests-toolbelt 1.0.0 which is incompatible.
Note: you may need to restart the kernel to use updated packages.
Note: you may need to restart the kernel to use updated packages.
Installation done
```

```
[1]: ## Run this cell and all below
```

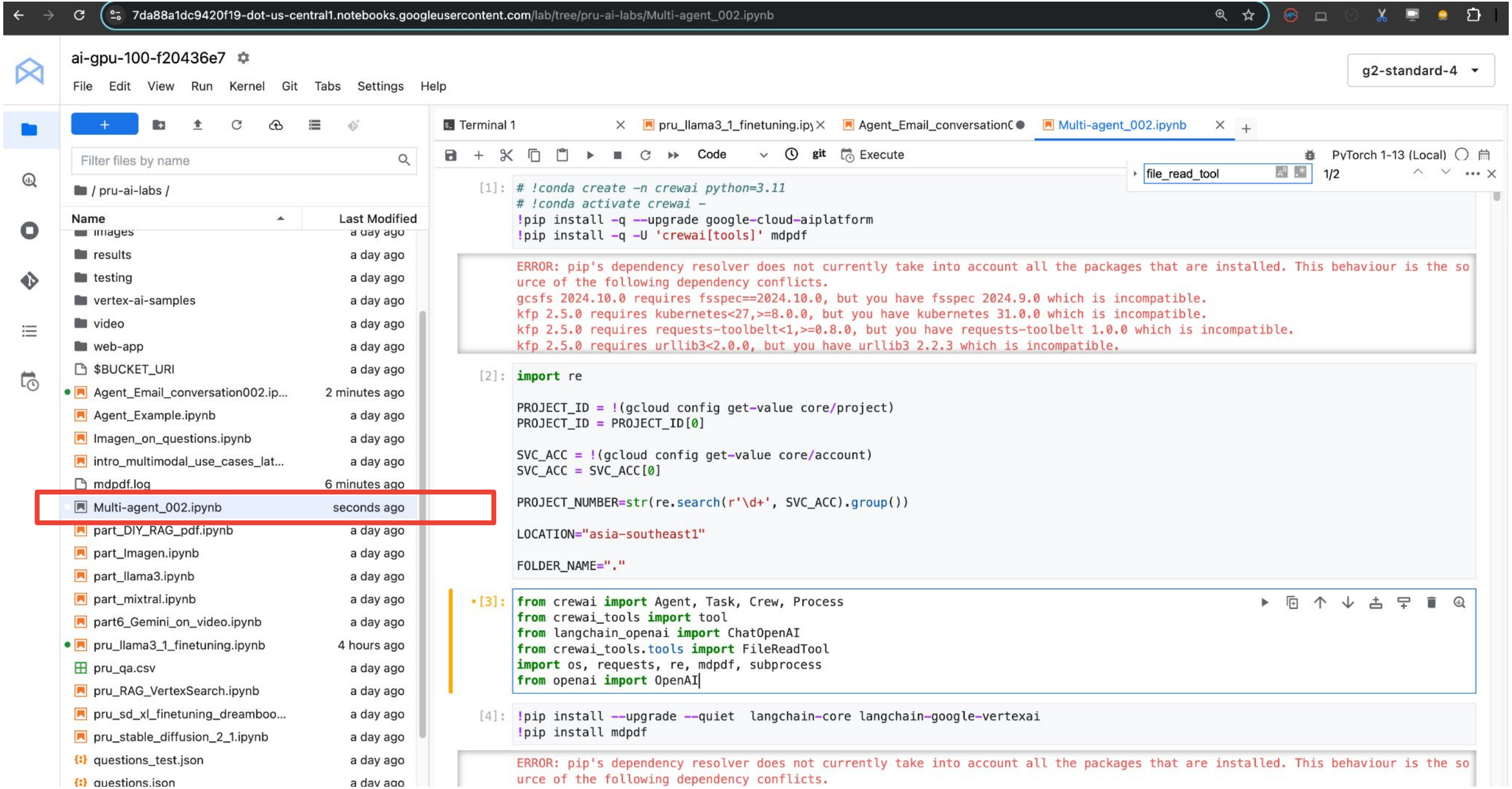
Declare Variables

```
[2]: #no spaces or special characters allowed), ensure that it is unique
     import socket
     import re

     UNIQUE_PREFIX = socket.gethostname()

     UNIQUE_PREFIX = re.sub('[^A-Za-z0-9]+', '', UNIQUE_PREFIX)

     PREFIX_NUM_ONLY=int(str(re.search(r'\d+', UNIQUE_PREFIX).group()))

     REGION_ALLOCATE=PREFIX_NUM_ONLY%3
```

```
[3]: PROJECT_ID = !(gcloud config get-value core/project)
     PROJECT_ID = PROJECT_ID[0]

     SVC_ACC = !(gcloud config get-value core/account)
     SVC_ACC = SVC_ACC[0]

     PROJECT_NUMBER=str(re.search(r'\d+', SVC_ACC).group())

     LOCATION="asia-southeast1"

     FOLDER_NAME="."
```

Step2 : Run Selected Cell and All Below

Multi Agents



Name	Last Modified
images	a day ago
results	a day ago
testing	a day ago
vertex-ai-samples	a day ago
video	a day ago
web-app	a day ago
$BUCKET_URI	a day ago
Agent_Email_conversation002.ip...	2 minutes ago
Agent_Example.ipynb	a day ago
Imagen_on_questions.ipynb	a day ago
intro_multimodal_use_cases_lat...	a day ago
mdpdf.log	6 minutes ago
Multi-agent_002.ipynb	seconds ago
part_DIY_RAG_pdf.ipynb	a day ago
part_Imagen.ipynb	a day ago
part_llama3.ipynb	a day ago
part_mixtral.ipynb	a day ago
part6_Gemini_on_video.ipynb	a day ago
pru_llama3_1_finetuning.ipynb	4 hours ago
pru_qa.csv	a day ago
pru_RAG_VertexSearch.ipynb	a day ago
pru_sd_xl_finetuning_dreamboo...	a day ago
pru_stable_diffusion_2_1.ipynb	a day ago
questions_test.json	a day ago
questions.json	a day ago

[1]:

```
# !conda create -n crewai python=3.11
# !conda activate crewai -
!pip install -q --upgrade google-cloud-aiplatform
!pip install -q -U 'crewai[tools]' mdpdf
```

```
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the so
urce of the following dependency conflicts.
gcsfs 2024.10.0 requires fsspec==2024.10.0, but you have fsspec 2024.9.0 which is incompatible.
kfp 2.5.0 requires kubernetes<27,>=8.0.0, but you have kubernetes 31.0.0 which is incompatible.
kfp 2.5.0 requires requests-toolbelt<1,>=0.8.0, but you have requests-toolbelt 1.0.0 which is incompatible.
kfp 2.5.0 requires urllib3<2.0.0, but you have urllib3 2.2.3 which is incompatible.
```

[2]:

```
import re

PROJECT_ID = !(gcloud config get-value core/project)
PROJECT_ID = PROJECT_ID[0]

SVC_ACC = !(gcloud config get-value core/account)
SVC_ACC = SVC_ACC[0]

PROJECT_NUMBER=str(re.search(r'\d+', SVC_ACC).group())

LOCATION="asia-southeast1"

FOLDER_NAME="."
```

[3]:

```
from crewai import Agent, Task, Crew, Process
from crewai_tools import tool
from langchain_openai import ChatOpenAI
from crewai_tools.tools import FileReadTool
import os, requests, re, mdpdf, subprocess
from openai import OpenAI
```

[4]:

```
!pip install --upgrade --quiet  langchain-core langchain-google-vertexai
!pip install mdpdf
```

```
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the so
urce of the following dependency conflicts.
```

Proprietary + Confidential

Optional Lab :

Go to this github URL and clone

https://github.com/analyticsrepo01/ai_agents_v2

```
git clone <URL>
```

Labs Setup: Setting Up Notebooks

Note:

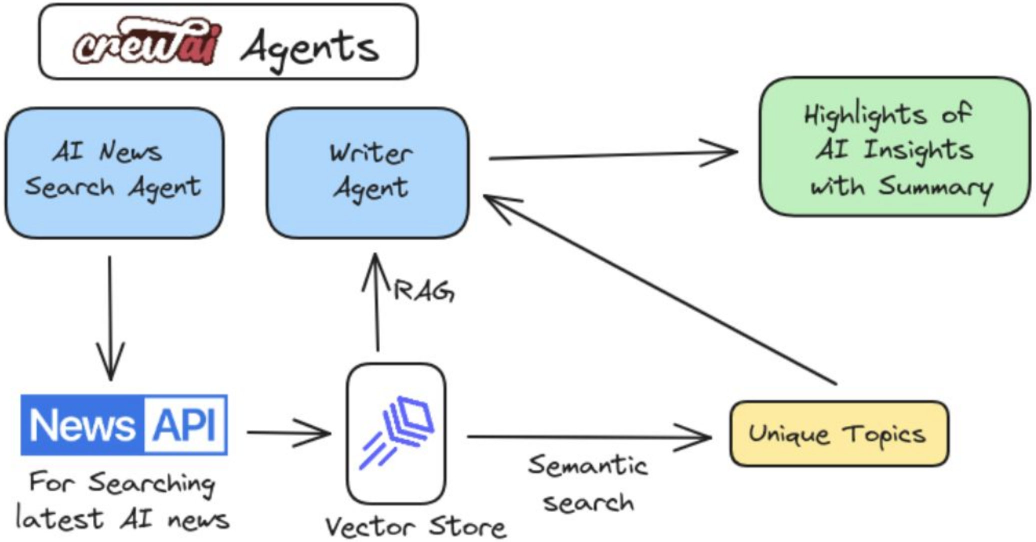
Avoid letting your laptops/computers enter sleep mode to prevent problems arising when running the notebooks

Thank you

Multi-Agent Lab: Generating AI News Insights

Objectives:

- Retrieve news about AI using RAG
- Generate summarized reports on AI trends


crewai Agents
AI News Search Agent
Writer Agent
Highlights of AI Insights with Summary
RAG
News API
For Searching latest AI news
Vector Store
Semantic search
Unique Topics