Llama 2 Lab on GCP

Sign up for Trial of Google Cloud

****

**A screenshot of a computer

Description automatically generated**

Enter payment info for trial (instructor will load via preloaded cards)

**A screenshot of a computer

Description automatically generated**

Deploy > New Project

A screenshot of a computer

Description automatically generated

Create

A screenshot of a computer

Description automatically generated

Vertex AI endpoint

A screenshot of a computer

Description automatically generated

Select 2 NVIDIA > Deploy

Wait

A screenshot of a computer

Description automatically generated

Search for Endpoint Attachments

A screenshot of a cloud

Description automatically generated

Cloud overview > Model Registry

A screenshot of a computer

Description automatically generated

Check for deployed status

A screenshot of a computer

Description automatically generated

Click Cloud overview > Service Accounts

A screenshot of a computer

Description automatically generated

Click Create service account

A screenshot of a computer

Description automatically generated

Select Vertex AI

A screenshot of a computer

Description automatically generated

Assign the listed roles

A screenshot of a computer

Description automatically generated

Verify account created

A screenshot of a computer

Description automatically generated

Click Keys > Add key

A screenshot of a computer

Description automatically generated

Return to Model Registry and select Llama2

Select sample request

A screenshot of a computer

Description automatically generated

Copy the first gcloud snippet

A screenshot of a computer

Description automatically generated



Launch the CLI and past the auth code snippet

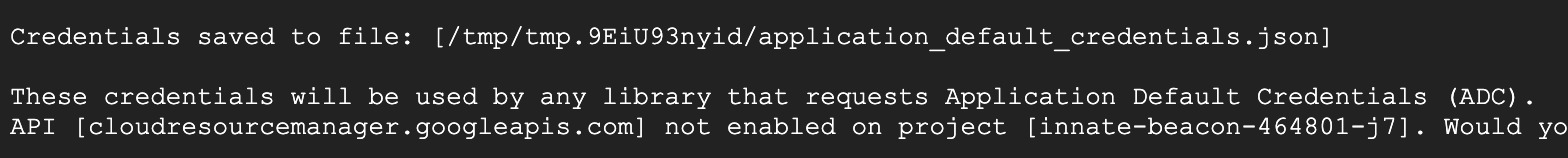
Click the link returned in the CLI and follow the popout window prompts

Copy the authorization code (last page) and put back into CLI

A white background with black text

Description automatically generated

You should see a confirmation:



Copy the second snipped from sample request and run the prediction.

A screenshot of a computer

Description automatically generated

Submit the a response from the prediction.