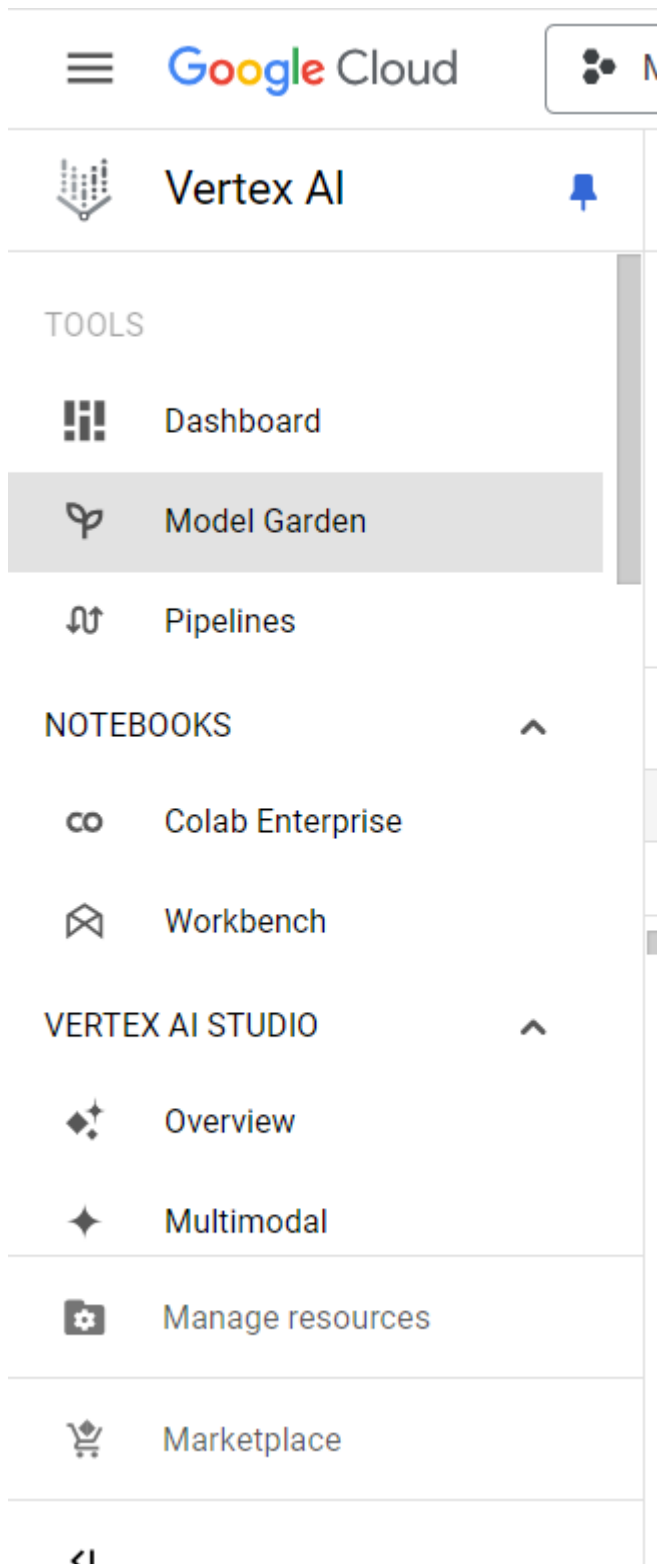
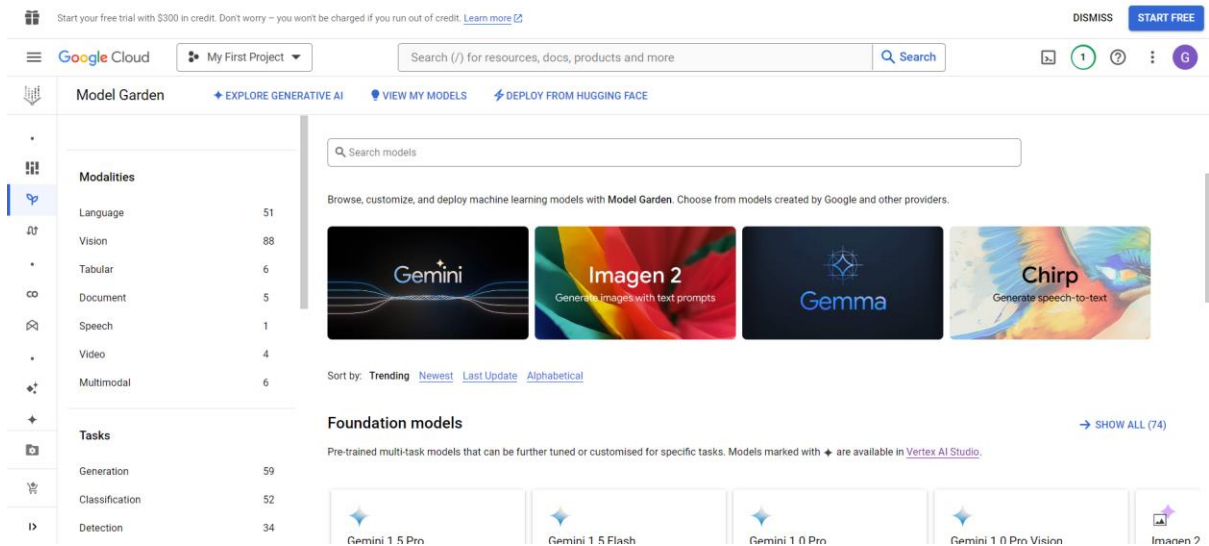


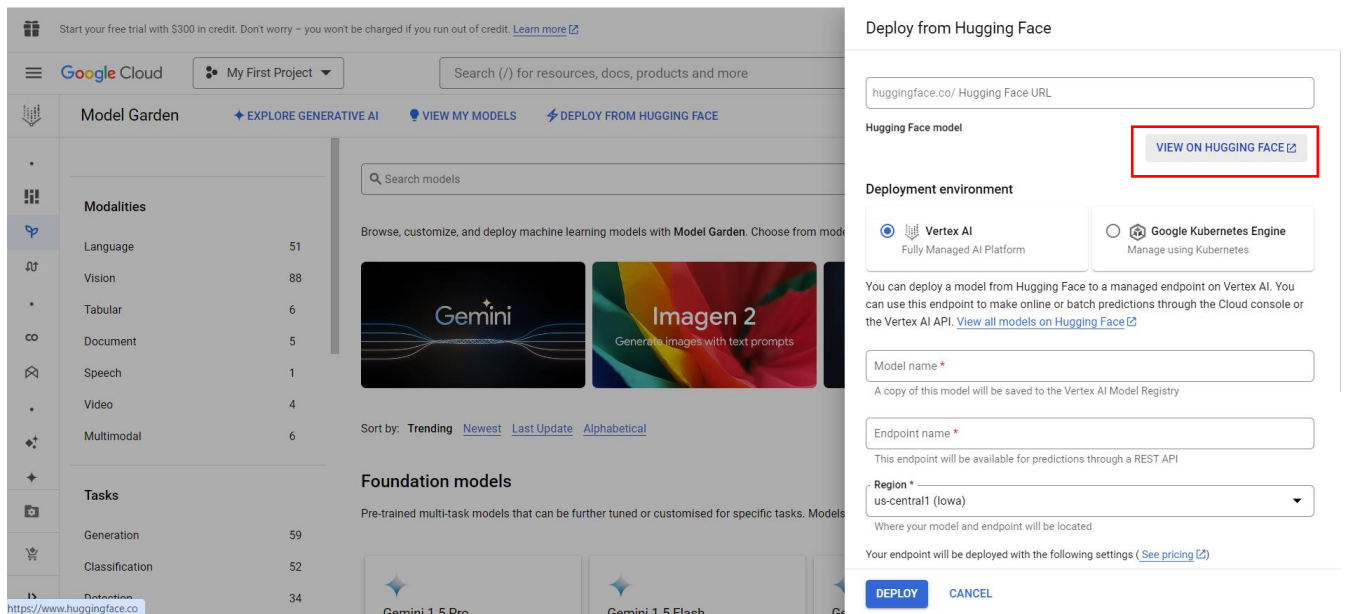
Go to Vertex AI and click Model Garden



Here you can choose a model to deploy



Personally I deploy model through Hugging Face.



Also don't forget to provide a Token

Deploy from Hugging Face

Hugging Face URL

huggingface.co/ meta-llama/meta-llama-3-8b

Hugging Face model

meta-llama-3-8b

[VIEW ON HUGGING FACE](#)

Deployment environment



Vertex AI

Fully Managed AI Platform

You can deploy a model from Hugging Face to a managed endpoint on Vertex AI. You can use this endpoint to make online or batch predictions through the Cloud console or the Vertex AI API. [View all models on Hugging Face](#)

Hugging Face access token *

Used to access private models on Hugging face hub. [Learn how to create a token](#)

Model name *

meta-llama-3-8b-1716920341847

A copy of this model will be saved to the Vertex AI Model Registry

Endpoint name *

meta-llama-3-8b-mg-one-click-deploy

This endpoint will be available for predictions through a REST API

Region *

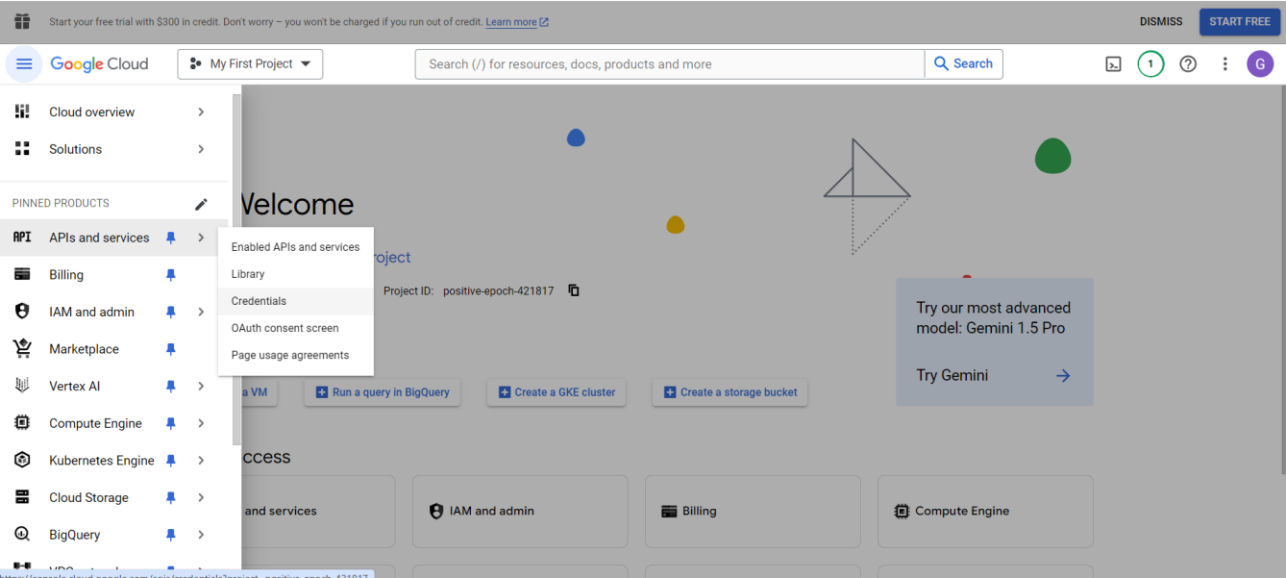
us-east4 (Northern Virginia)

DEPLOY

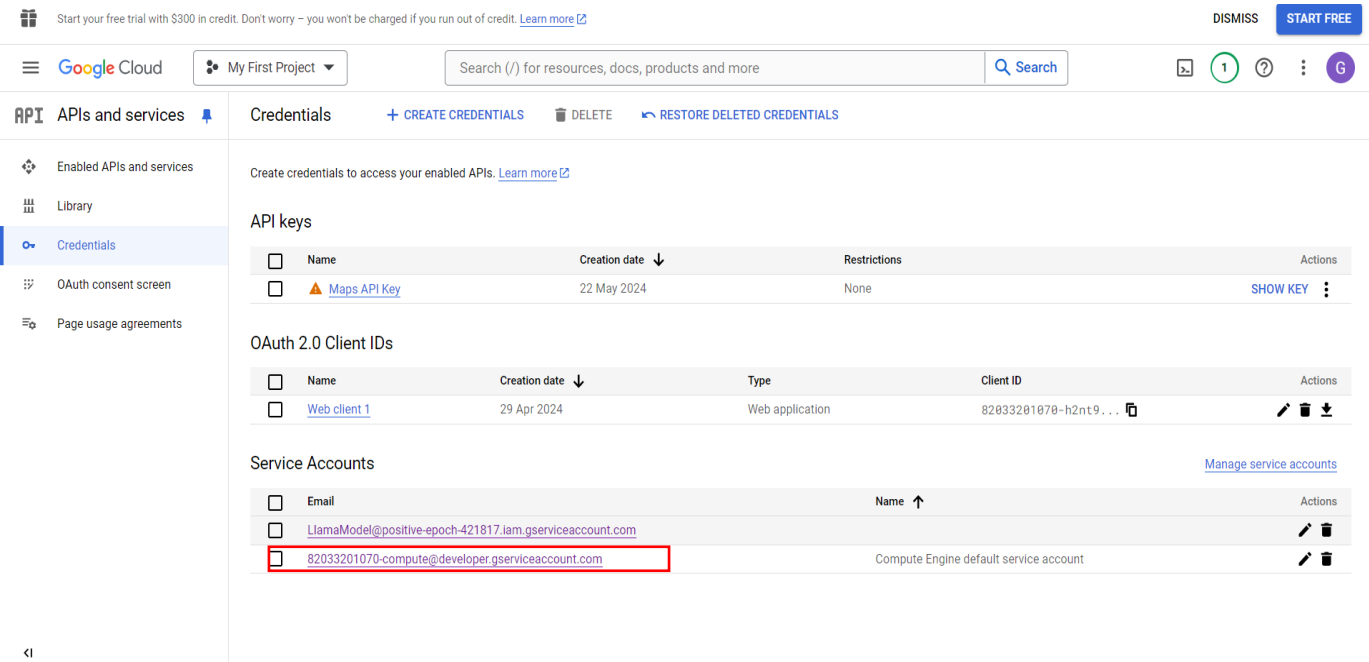
CANCEL

Note, model takes some time to be deployed (approximately 30 minutes)

Go to Credentials



Click on the Service Account that starts with numbers



Click on Keys

Google Cloud | My First Project | Search (/) for resources, docs, products and more

IAM and admin | Compute Engine default service account

DETAILS | PERMISSIONS | KEYS | METRICS | LOGS

Service account details

Name: Compute Engine default service account [SAVE]

Description: [SAVE]

Email: 82033201070-compute@developer.gserviceaccount.com

Unique ID: 114788626051031015732

Service account status

Disabling your account allows you to preserve your policies without having to delete it.

Enabled

[DISABLE SERVICE ACCOUNT]

Advanced settings

Click on ADD KEY -> Create new key

Start your free trial with \$300 in credit. Don't worry - you won't be charged if you run out of credit. [Learn more](#)

DISMISS | START FREE

Google Cloud | My First Project | Search (/) for resources, docs, products and more

IAM and admin | Compute Engine default service account

DETAILS | PERMISSIONS | KEYS | METRICS | LOGS

Keys

Service account keys could pose a security risk if compromised. We recommend that you avoid downloading service account keys and instead use the [Workload Identity Federation](#). You can learn more about the best way to authenticate service accounts on Google Cloud [here](#).

Starting June 16, 2024, Google will automatically disable service account keys detected in public repositories. You can customize this behavior by using the 'iam.serviceAccountKeyExposureResponse' organization policy. [Learn more](#)

Add a new key pair or upload a public key certificate from an existing key pair.

Block service account key creation using [organisation policies](#). [Learn more about setting organisation policies for service accounts](#)

[ADD KEY]

Create new key

Upload existing key

| Key | Creation date | Expiry date | |
|--|---------------|-------------|----------|
| 09ae5c662132f772056daca3109d492888a18d2b | 1 May 2024 | 1 Jan 10000 | [Delete] |

Create private key for 'Compute Engine default service account'

Downloads a file that contains the private key. Store the file securely because this key cannot be recovered if lost.

Key type

☒ JSON

Recommended

☐ P12

For backward compatibility with code using the P12 format

CANCEL

CREATE

Click create

A key should be downloaded and saved in a directory where the API script will be saved.

API SCRIPT

```
import requests
```

```
import google.auth
```

```
import google.auth.transport.requests
```

```
from google.oauth2 import service_account
```

```
SCOPES = ['https://www.googleapis.com/auth/cloud-platform']
```

```
SERVICE_ACCOUNT_FILE = 'positive-epoch-421817-09ae5c662132.json'
```

```
cred = service_account.Credentials.from_service_account_file(SERVICE_ACCOUNT_FILE,  
scopes=SCOPES)
```

```
auth_req = google.auth.transport.requests.Request()
```

```
cred.refresh(auth_req)
```

```
bearer_token = cred.token
```

```
base_url = "https://europe-west1-  
aiplatform.googleapis.com/v1beta1/projects/{project_id}/locations/europe-  
west1/endpoints/{endpoint_id}:predict"
```

```
project_id = "positive-epoch-421817"
```

```
endpoint_id = "5966922060591529984"
```

```
prompt = input("Enter prompt: ")
```

```
request_body = {  
    "instances": [  
        {  
            "inputs": prompt,  
            "max_length": 10,  
            "temperature": 0.4  
        }  
    ]  
}
```

```
full_url = base_url.format(project_id=project_id, endpoint_id=endpoint_id)
```

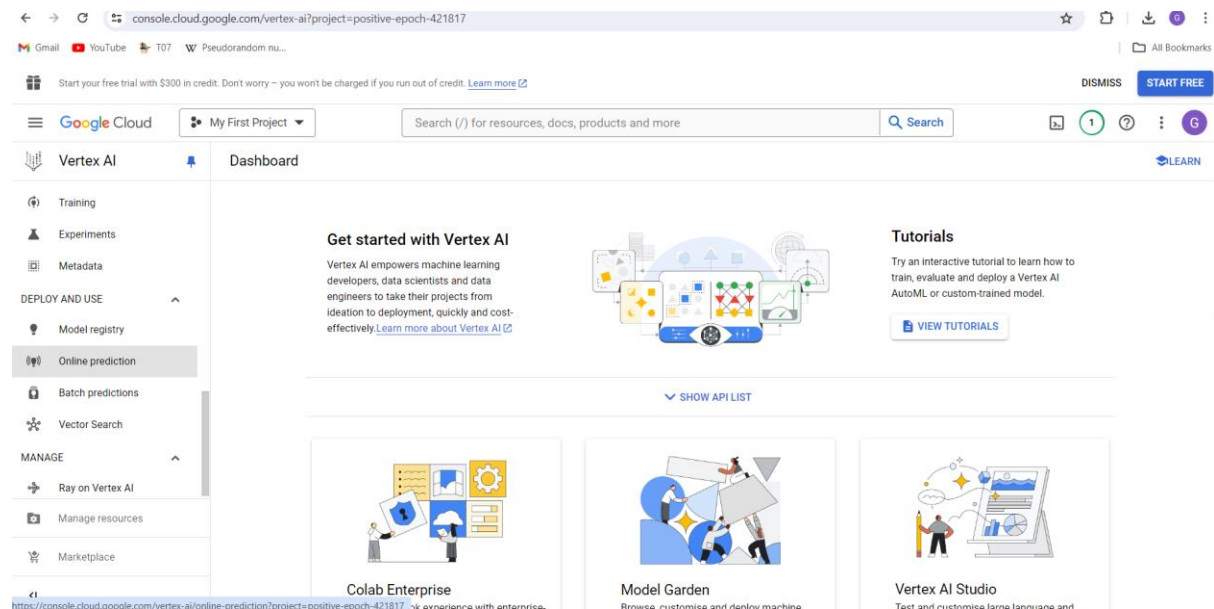
```
headers = {  
    "Authorization": "Bearer {bearer_token}".format(bearer_token=bearer_token),  
    "Content-Type": "application/json"  
}
```

```
resp = requests.post(full_url, json=request_body, headers=headers)
```

```
print(resp.json())
```

The only thing that you need to change is endpoint id.

Go to Vertex Ai and click „Online prediction“



Copy ID of your created endpoint

| | | | | | | | | | | | |
|--------------------------|---|---------------------|--------|---|---|--------------|----------|---|-----------------------|--------|---|
| <input type="checkbox"/> | meta-llama-3-8b-mg-one-click-deploy | 9110821628589113344 | Active | 1 | — | europa-west1 | Disabled | — | 28 May 2024, 20:51:08 | SAMPLE | ⋮ |
|--------------------------|---|---------------------|--------|---|---|--------------|----------|---|-----------------------|--------|---|

And change the ID in the script

Now you can run the script

If you change a deployment location than you also need to change base_url in script according to location