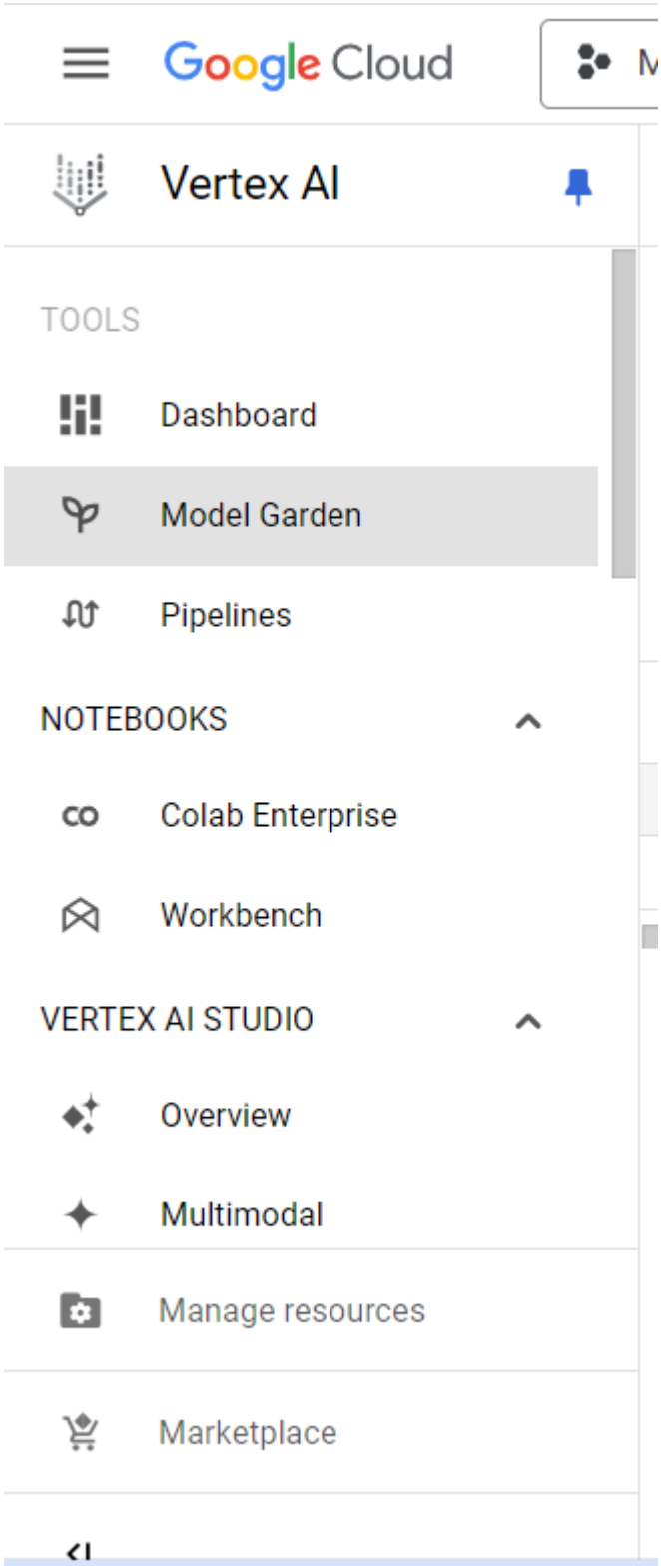
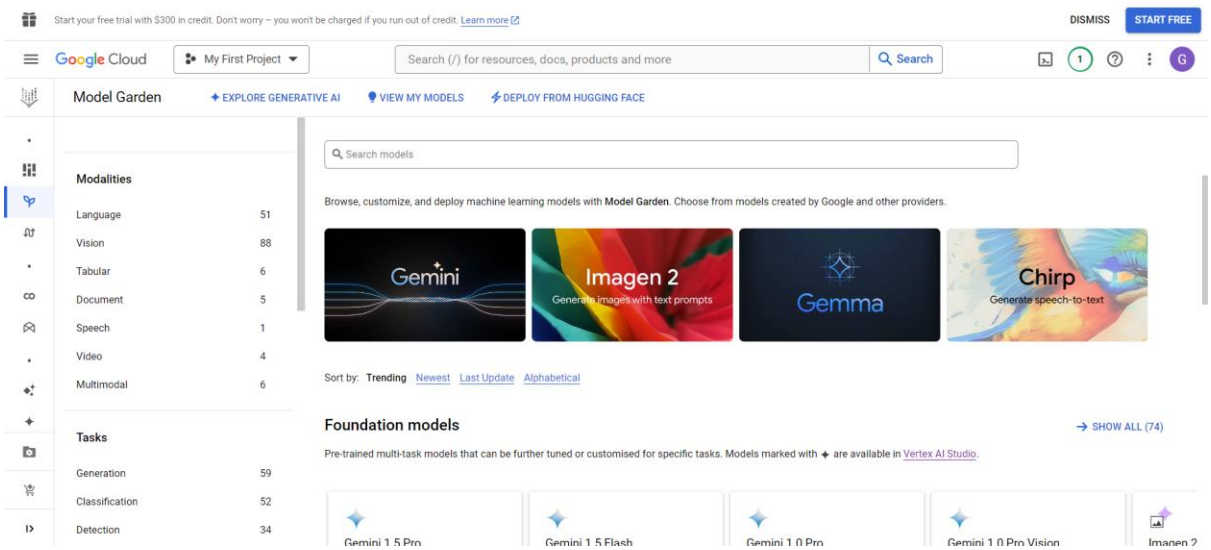
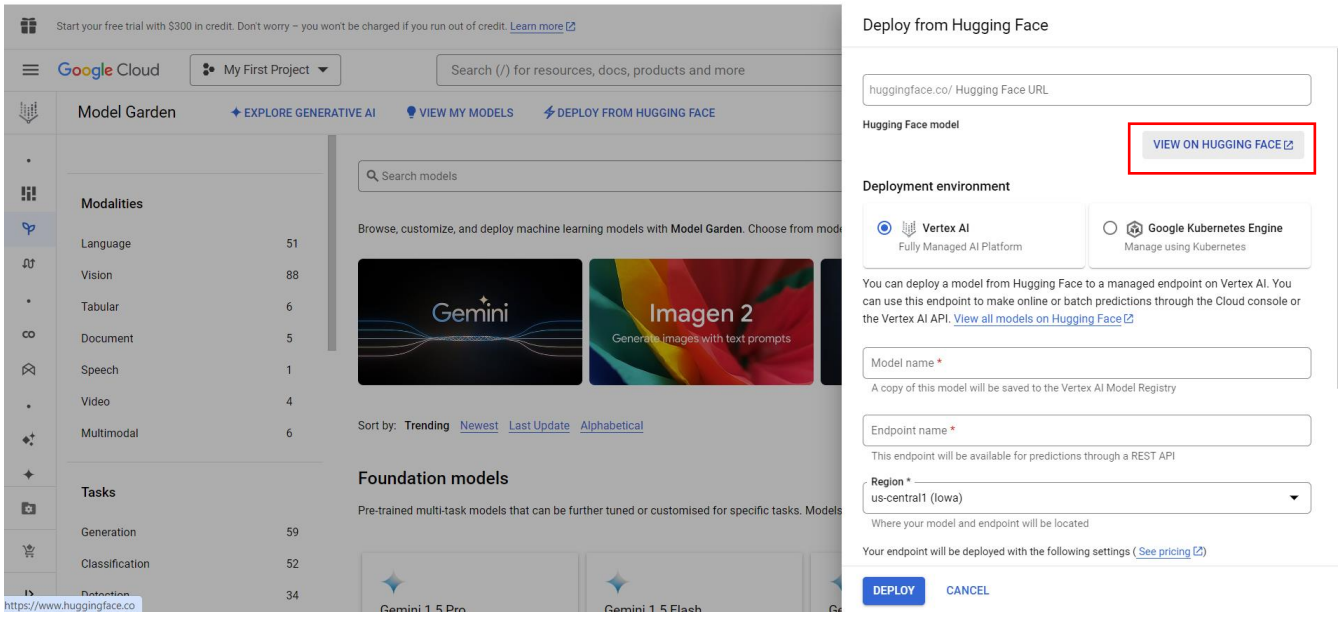
Go to Vertex AI and click Model Garden



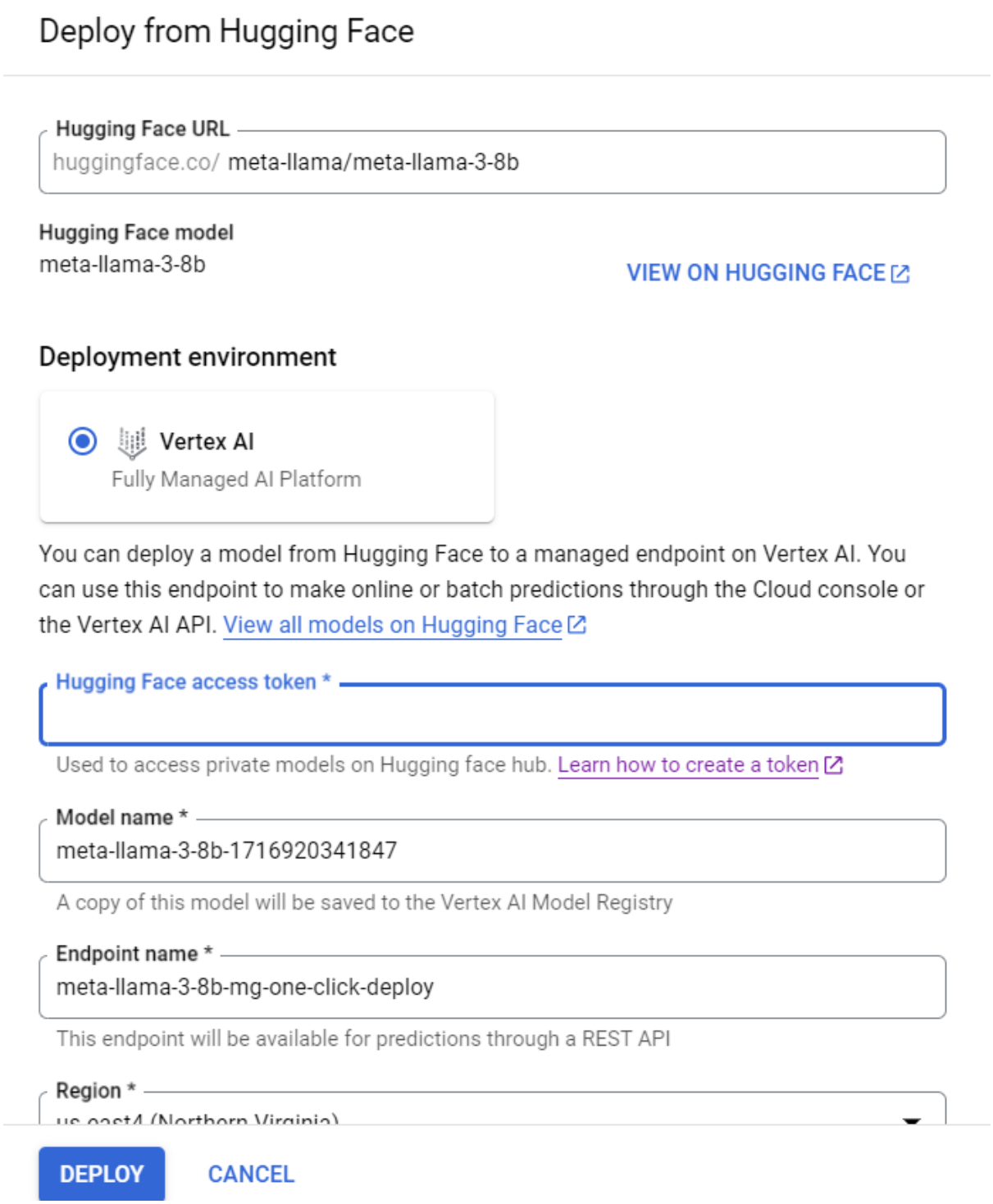
Here you can choose a model to deploy



Personally I deploy model through Hugging Face.

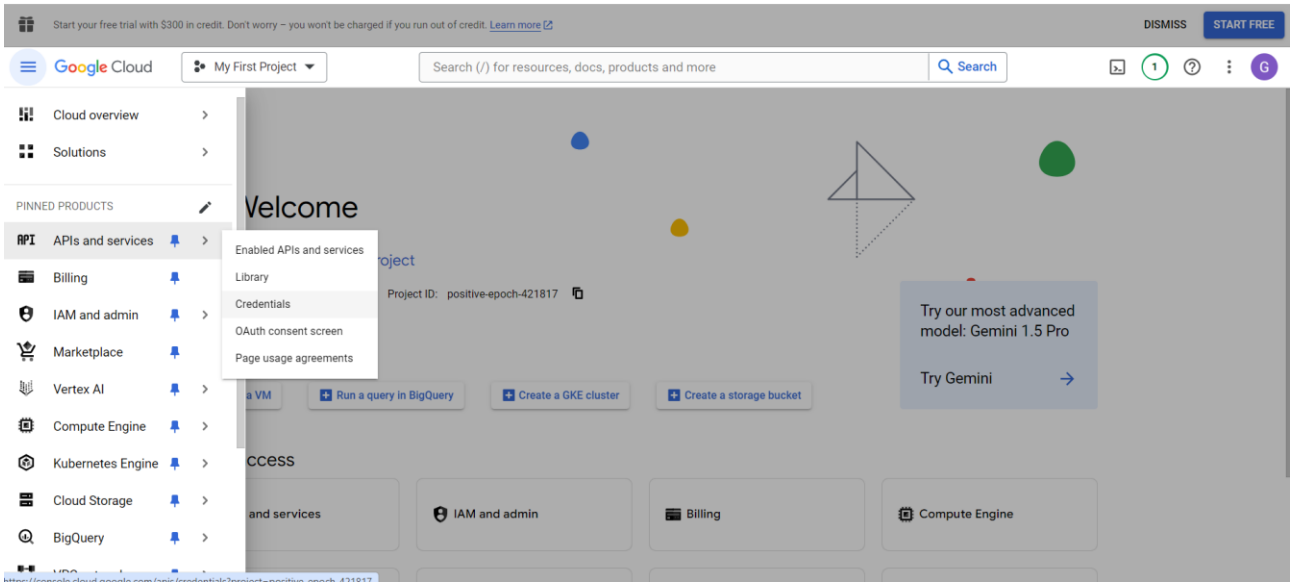


Also don‘t forget to provide a Token

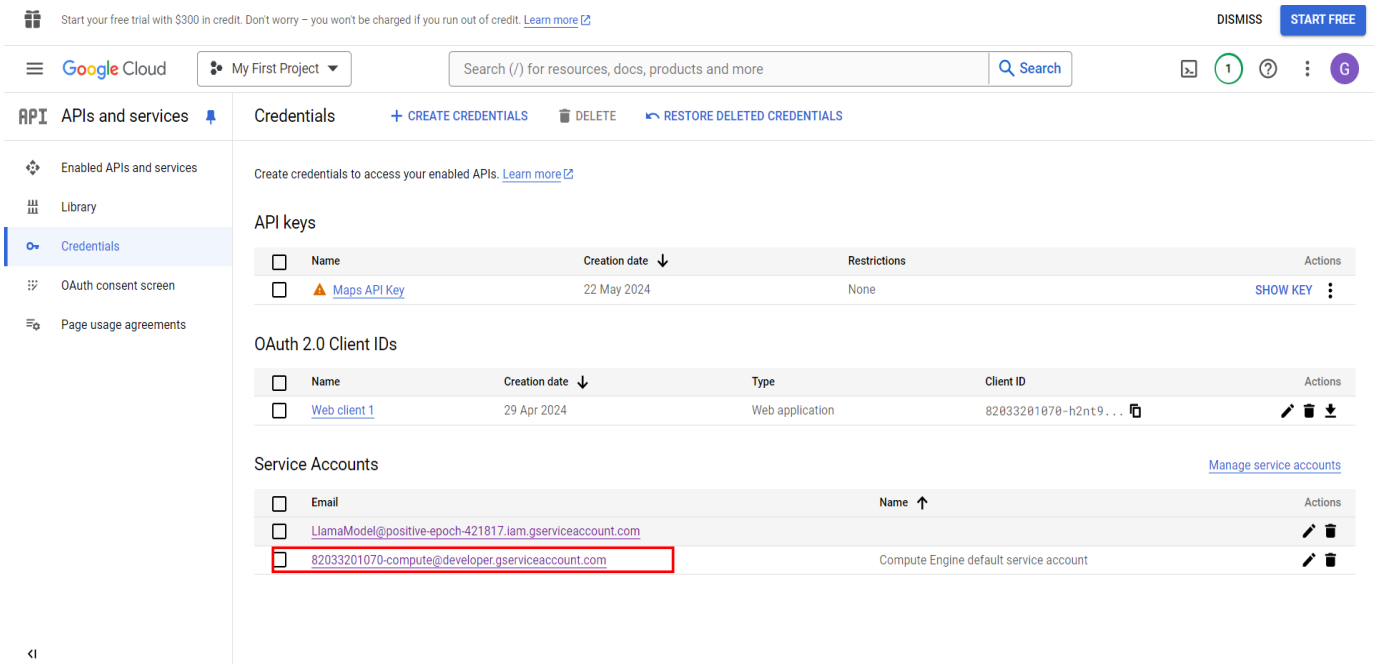


Note, model takes some time to be deployed (approximately 30 minutes)

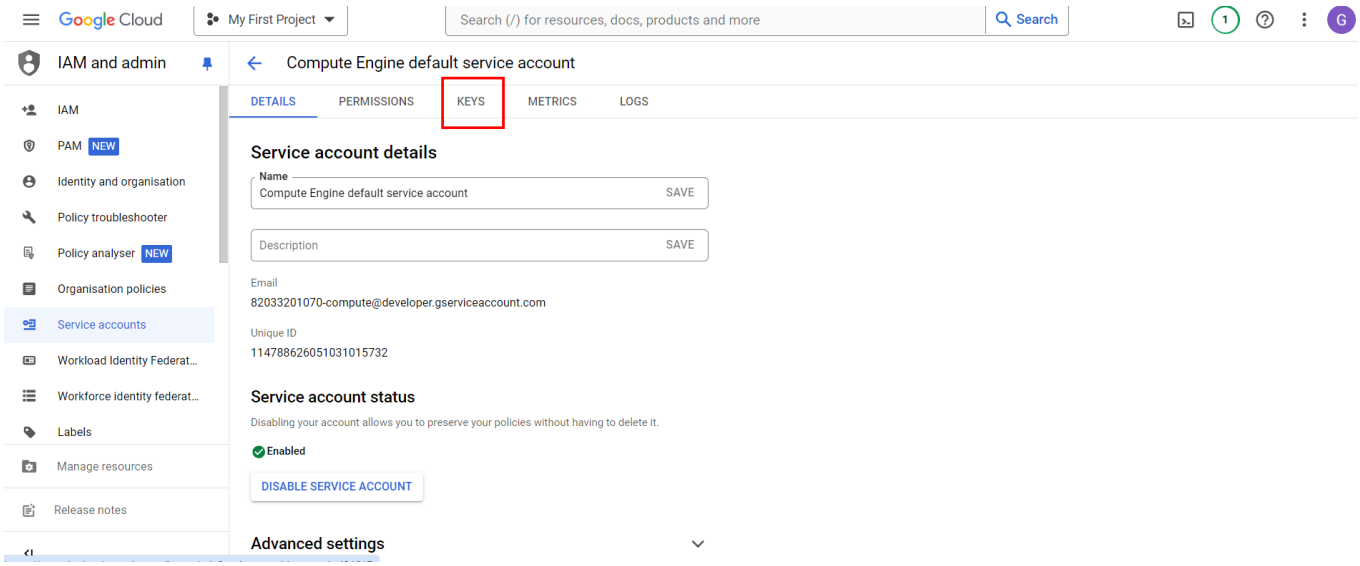
Go to Credentials



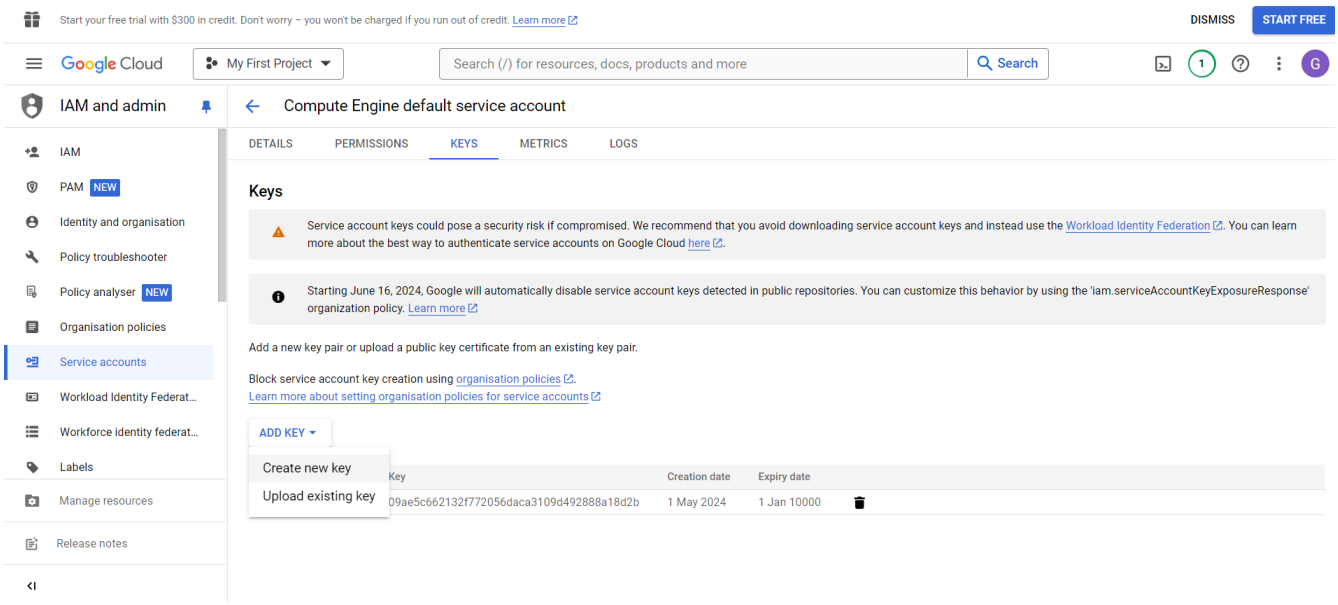
Click on the Service Account that starts with numbers

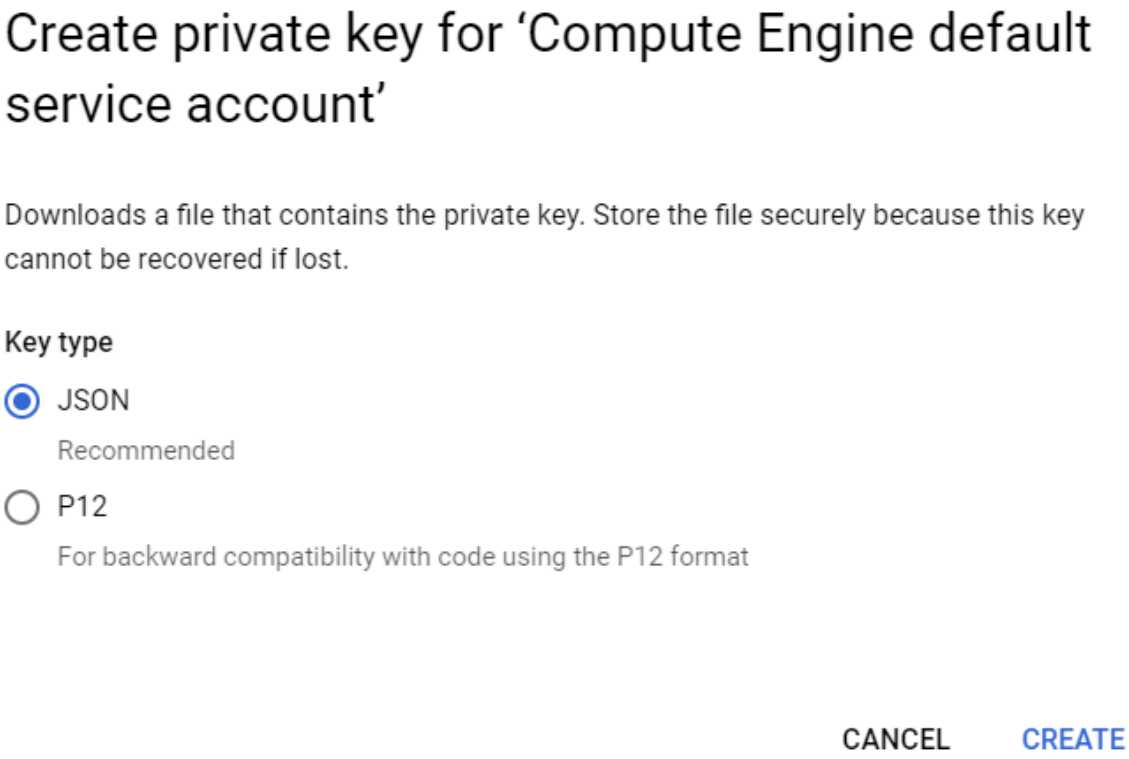


Click on Keys



Click on ADD KEY -> Create new key





Click create

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

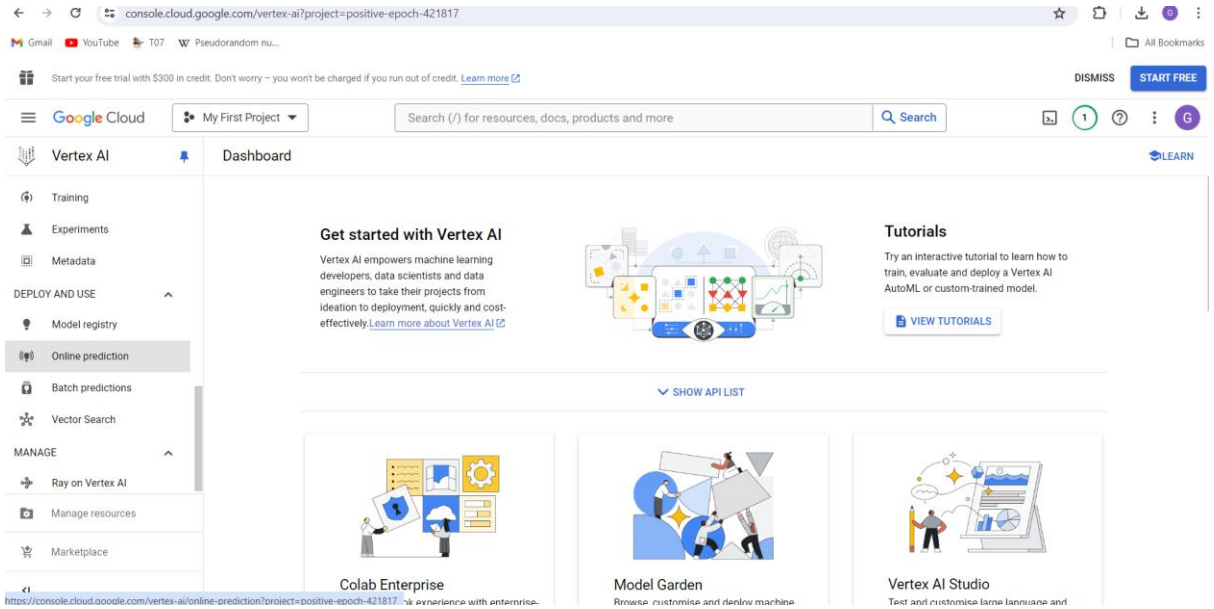
API SCRIPT

```
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()
```

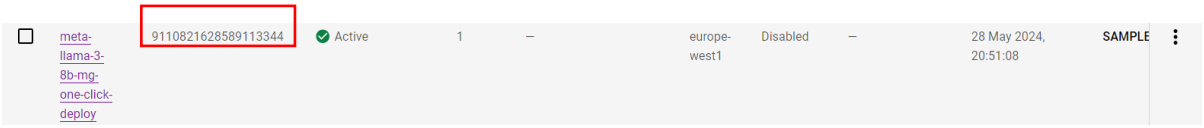
```
cred.refresh(auth_req)

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,
            "tempreture": 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



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