Deploying LLM from AI Quick Actions Catalog

Referenced Documentation

https://docs.oracle.com/en-us/iaas/data-science/using/ai-quick-actions-model-deploy.htm

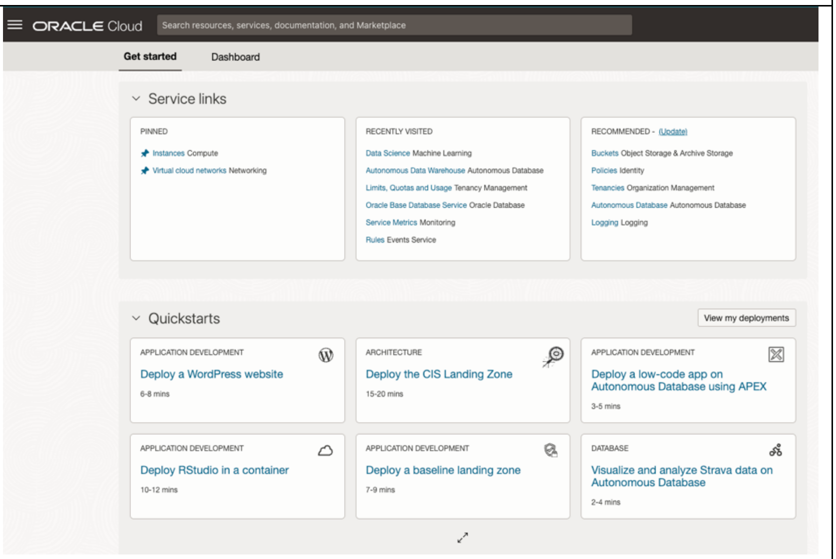
Description

You can create a Model Deployment from the foundation models with the tag Ready to Deploy in the Model Explorer, or with fine-tuned models. When you create a Model Deployment in AI Quick Actions, you're creating an OCI Data Science Model Deployment, which is a managed resource in the OCI Data Science Service. You can deploy the model as HTTP endpoints in OCI.

Pre-Requisites

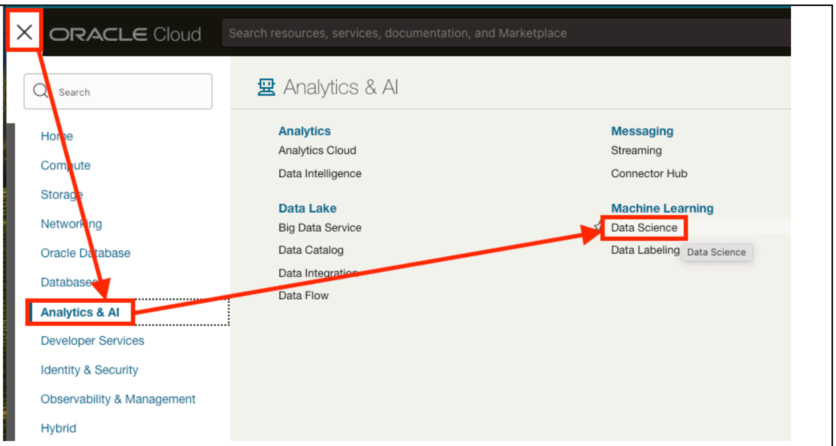
- Implement the required policies - https://docs.oracle.com/en-us/iaas/data-science/using/ai-quick-actions-set-up.htm
- Ensure you have your OCI Data Science GPU service limits raised for the GPU Shapes you plan to use. This can be done from OCI Console.
- Provisioned OCI Data Science Project and Notebook Session (Must be deactivated and reactivated if created before the policies where implemented).
- OCI Log Group & Log Created (Optional)

Guide

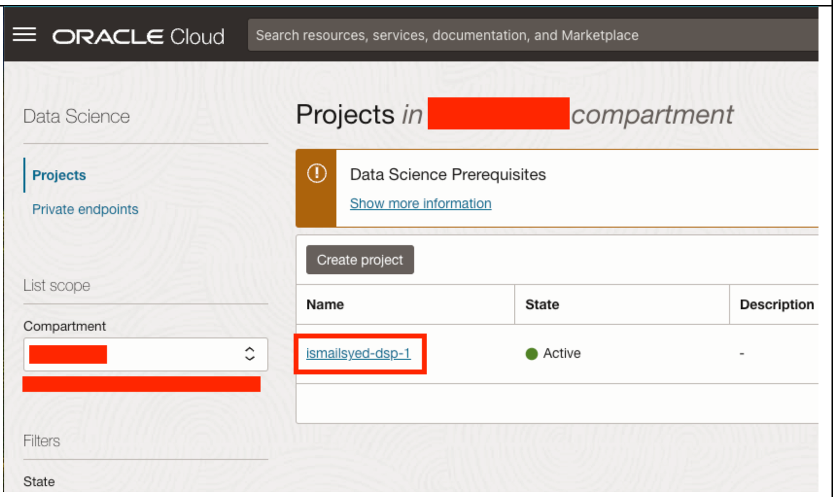
Step	Screenshot
Login to the Cloud Console. cloud.oracle.com	 ORACLE Cloud Search resources, services, documentation, and Marketplace Get started Dashboard Service links PINNED Instances Compute Virtual cloud networks Networking RECENTLY VISITED Data Science Machine Learning Autonomous Data Warehouse Autonomous Database Limits, Quotas and Usage Tenancy Management Oracle Base Database Service Oracle Database Service Metrics Monitoring Rules Events Service RECOMMENDED - (Update) Buckets Object Storage & Archive Storage Policies Identity Tenancies Organization Management Autonomous Database Autonomous Database Logging Logging Quickstarts View my deployments APPLICATION DEVELOPMENT Deploy a WordPress website 6-8 mins ARCHITECTURE Deploy the CIS Landing Zone 15-20 mins APPLICATION DEVELOPMENT Deploy a low-code app on Autonomous Database using APEX 3-5 mins APPLICATION DEVELOPMENT Deploy RStudio in a container 10-12 mins APPLICATION DEVELOPMENT Deploy a baseline landing zone 7-9 mins DATABASE Visualize and analyze Strava data on Autonomous Database 2-4 mins

Navigate to your Data Science Projects.

OCI Menu > Analytics & AI > Data Science.



Open up your existing Data Science Project.



Projects *in* [redacted] *compartment*

Data Science Prerequisites
Show more information

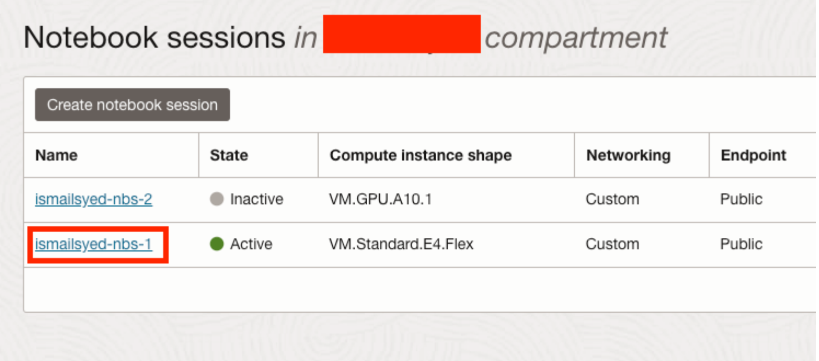
Compartment: [redacted]

[redacted]

Name	State	Description
ismailsyed-dsp-1	● Active	-

Click on your existing Data Science Notebook Session.

Note – This does not have to be a GPU Shape.



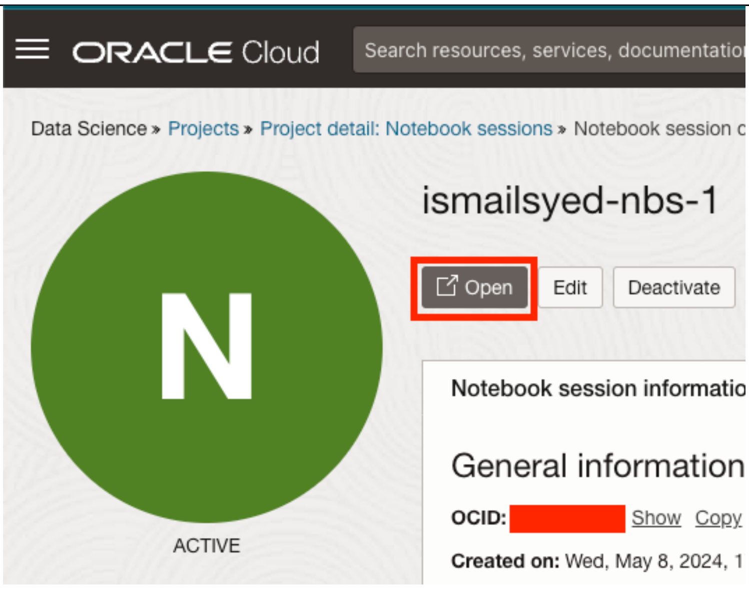
Notebook sessions *in* [redacted] *compartment*

Name	State	Compute instance shape	Networking	Endpoint
ismailsyed-nbs-2	● Inactive	VM.GPU.A10.1	Custom	Public
ismailsyed-nbs-1	● Active	VM.Standard.E4.Flex	Custom	Public

Click on Open.

This will open up your Data Science Notebook Session.

You will have to reauthenticate.



Data Science » Projects » Project detail: Notebook sessions » Notebook session c

ismailsyed-nbs-1

Notebook session informatio

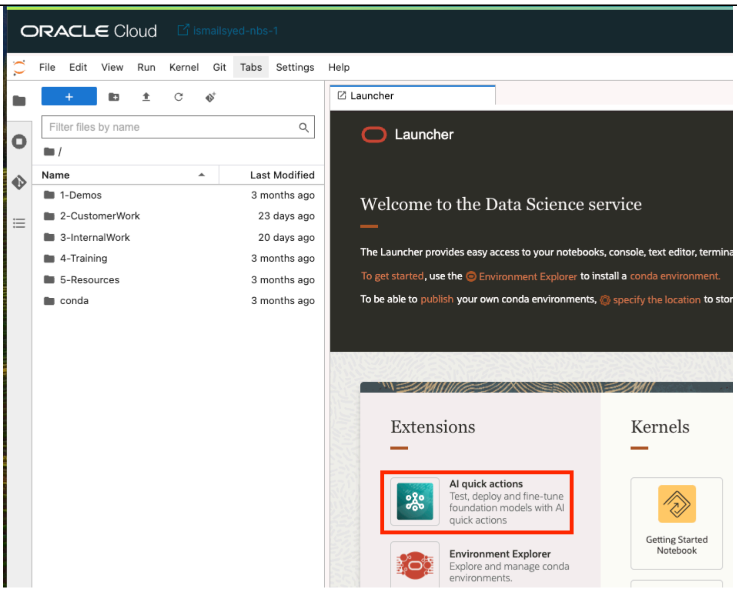
General information

OCID: [redacted] Show Copy

Created on: Wed, May 8, 2024, 1

If the policies within the pre-requisites have been implemented correctly you should be able to open up the AI Quick Actions Extension within the Launcher.

Click AI quick actions.



Name	Last Modified
1-Demos	3 months ago
2-CustomerWork	23 days ago
3-InternalWork	20 days ago
4-Training	3 months ago
5-Resources	3 months ago
conda	3 months ago

Launcher

Welcome to the Data Science service

The Launcher provides easy access to your notebooks, console, text editor, termina

To get started, use the Environment Explorer to install a conda environment.

To be able to publish your own conda environments, specify the location to stor

Extensions

AI quick actions Test, deploy and fine-tune foundation models with AI quick actions

Environment Explorer Explore and manage conda environments.

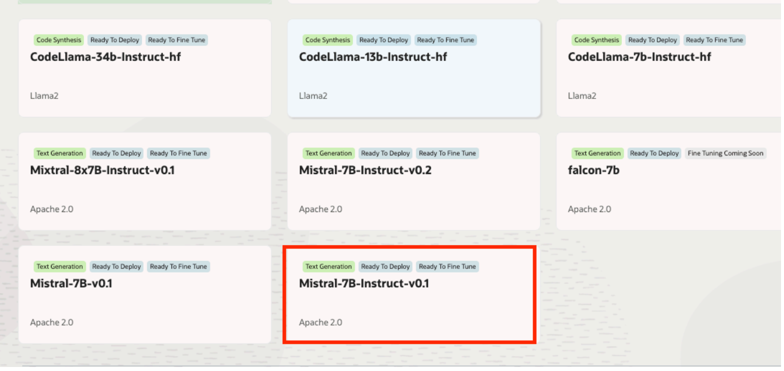
Kernels

Getting Started Notebook

The AI Quick Actions Catalog will be displayed.

In this tutorial we will deploy the ***Mistral-7B-Instruct-v0.1 model***.

Select this model.



Code Synthesis Ready To Deploy Ready To Fine Tune
CodeLlama-34b-Instruct-hf
Llama2

Code Synthesis Ready To Deploy Ready To Fine Tune
CodeLlama-13b-Instruct-hf
Llama2

Code Synthesis Ready To Deploy Ready To Fine Tune
CodeLlama-7b-Instruct-hf
Llama2

Text Generation Ready To Deploy Ready To Fine Tune
Mixtral-8x7B-Instruct-v0.1
Apache 2.0

Text Generation Ready To Deploy Ready To Fine Tune
Mistral-7B-Instruct-v0.2
Apache 2.0

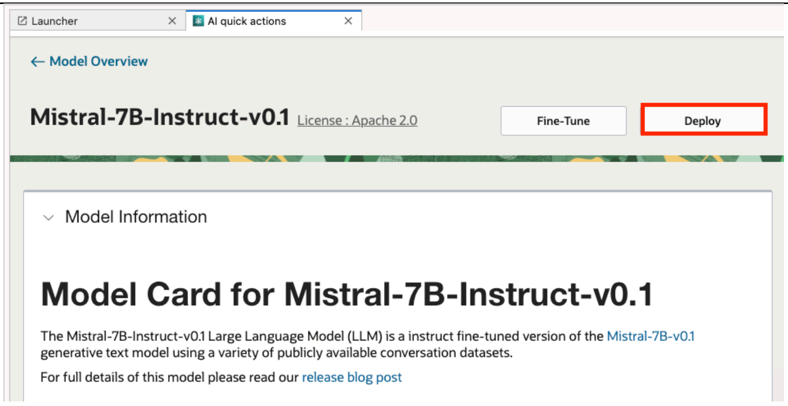
Text Generation Ready To Deploy Fine Tuning Coming Soon
falcon-7b
Apache 2.0

Text Generation Ready To Deploy Ready To Fine Tune
Mistral-7B-v0.1
Apache 2.0

Text Generation Ready To Deploy Ready To Fine Tune
Mistral-7B-Instruct-v0.1
Apache 2.0

We will then be taken to the Model card which gives you a bit of information about the Model.

Click on Deploy.



Mistral-7B-Instruct-v0.1 License : Apache 2.0

Model Card for Mistral-7B-Instruct-v0.1

The Mistral-7B-Instruct-v0.1 Large Language Model (LLM) is a instruct fine-tuned version of the Mistral-7B-v0.1 generative text model using a variety of publicly available conversation datasets.

For full details of this model please read our release blog post

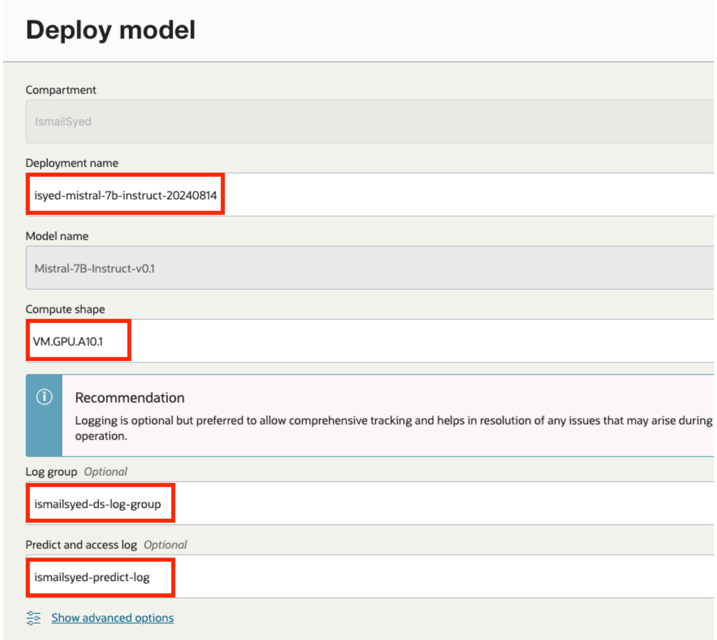
Enter:

Deployment Name

Compute Shape – leave the recommend shape selected (VM.GPU.A10.1)

As an optional step, select your Log Group and Log to store your logging information.

Click Deploy



Deploy model

Compartment: IsmailSyed

Deployment name: isyed-mistral-7b-instruct-20240814

Model name: Mistral-7B-Instruct-v0.1

Compute shape: VM.GPU.A10.1

Recommendation

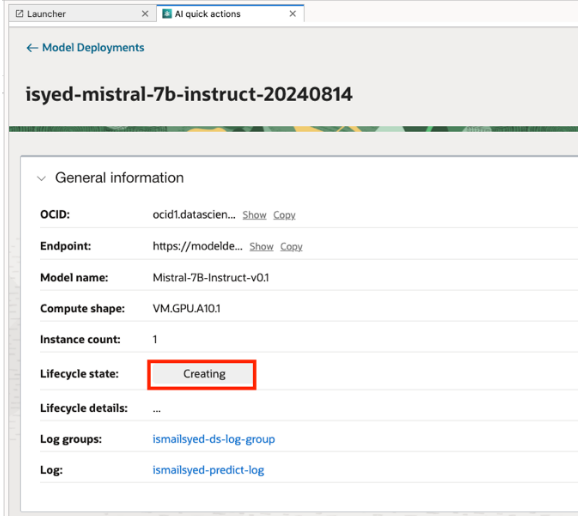
Logging is optional but preferred to allow comprehensive tracking and helps in resolution of any issues that may arise during operation.

Log group *Optional*: ismailsyed-ds-log-group

Predict and access log *Optional*: ismailsyed-predict-log

The Model will then enter the ***Creating*** Lifecycle state.

You can continue to do other tasks within OCI Data Science while the deployment is happening.



isyed-mistral-7b-instruct-20240814

OCID: ocid1.datascien... Show Copy

Endpoint: https://modelde... Show Copy

Model name: Mistral-7B-Instruct-v0.1

Compute shape: VM.GPU.A10.1

Instance count: 1

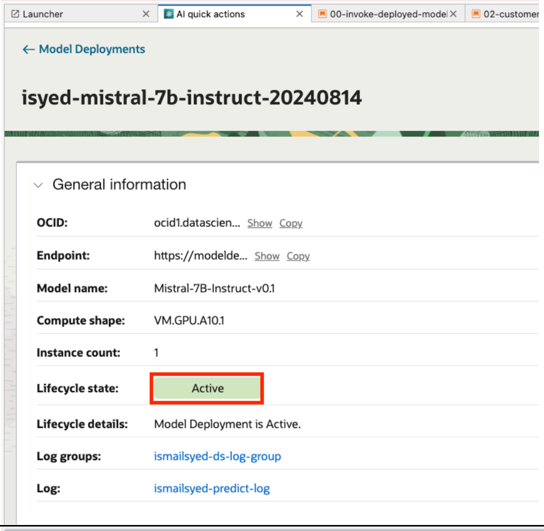
Lifecycle state: Creating

Lifecycle details: ...

Log groups: ismailsyed-ds-log-group

Log: ismailsyed-predict-log

Once the deployment is complete the lifecycle state will update to ***Active***.



← Model Deployments

isyed-mistral-7b-instruct-20240814

General information

OCID: ocid1.datascien... Show Copy

Endpoint: https://modelde... Show Copy

Model name: Mistral-7B-Instruct-v0.1

Compute shape: VM.GPU.A10.1

Instance count: 1

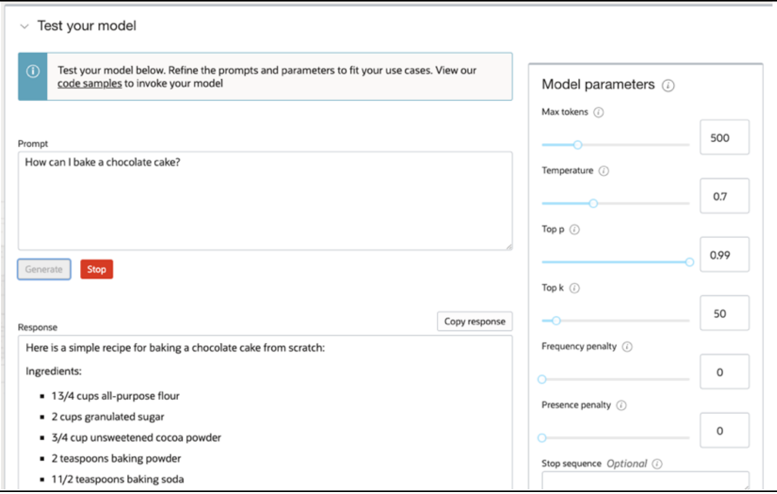
Lifecycle state: Active

Lifecycle details: Model Deployment is Active.

Log groups: ismailsyed-ds-log-group

Log: ismailsyed-predict-log

If you scroll down, you will see a sandpit environment to start testing your model along with editing the Model Parameters on the right-hand side.



Test your model below. Refine the prompts and parameters to fit your use cases. View our code samples to invoke your model

Prompt: How can I bake a chocolate cake?

Response:

Here is a simple recipe for baking a chocolate cake from scratch:

Ingredients:

- 13/4 cups all-purpose flour
- 2 cups granulated sugar
- 3/4 cup unsweetened cocoa powder
- 2 teaspoons baking powder
- 11/2 teaspoons baking soda

Model parameters

Max tokens: 500

Temperature: 0.7

Top p: 0.99

Top k: 50

Frequency penalty: 0

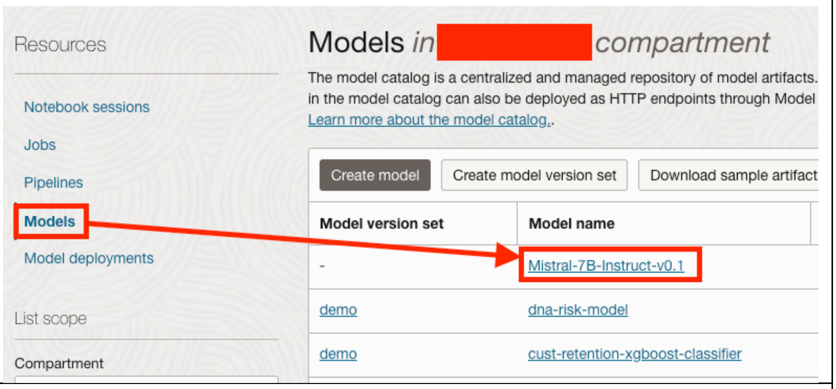
Presence penalty: 0

Stop sequence *Optional*

Head back to the OCI Console within our Data Project.

Navigate to Models.

Here you will see our new LLM in the Model Catalog.



Models *in* [redacted] *compartment*

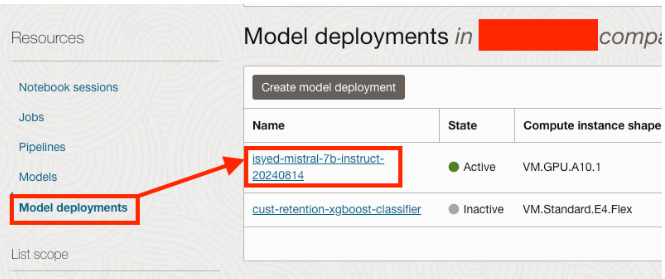
The model catalog is a centralized and managed repository of model artifacts. in the model catalog can also be deployed as HTTP endpoints through Model Learn more about the model catalog.

Model version set	Model name
-	Mistral-7B-Instruct-v0.1
demo	dna-risk-model
demo	cust-retention-xgboost-classifier

We can then ***navigate to Model Deployments.***

Here we can see our deployed Model.

Select our deployed model.

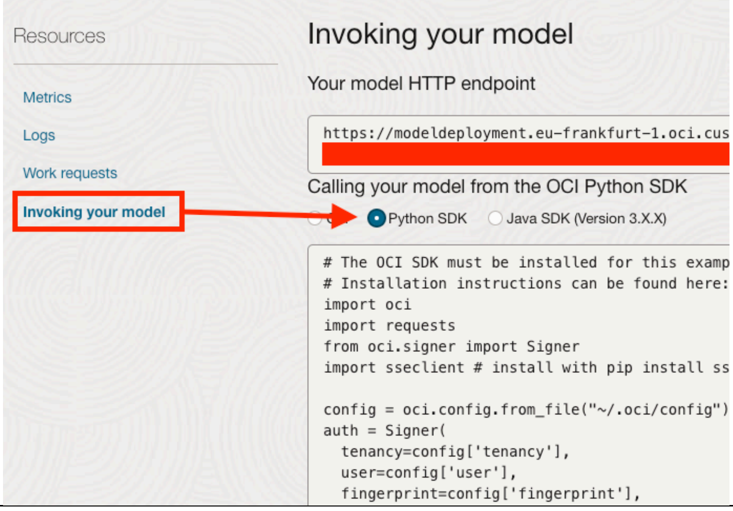


Model deployments *in* [redacted] *compa*

Name	State	Compute instance shape
isyed-mistral-7b-instruct-20240814	● Active	VM.GPU.A10.1
cust-retention-xgboost-classifier	● Inactive	VM.Standard.E4.Flex

Navigate to Invoking your Model under Resources.

Here we can view the deployed Model Endpoint along with some sample code to make inference against the model.



Invoking your model

Your model HTTP endpoint

```
https://modeldeployment.eu-frankfurt-1.oci.cus
[redacted]
```

Calling your model from the OCI Python SDK

● Python SDK ○ Java SDK (Version 3.X.X)

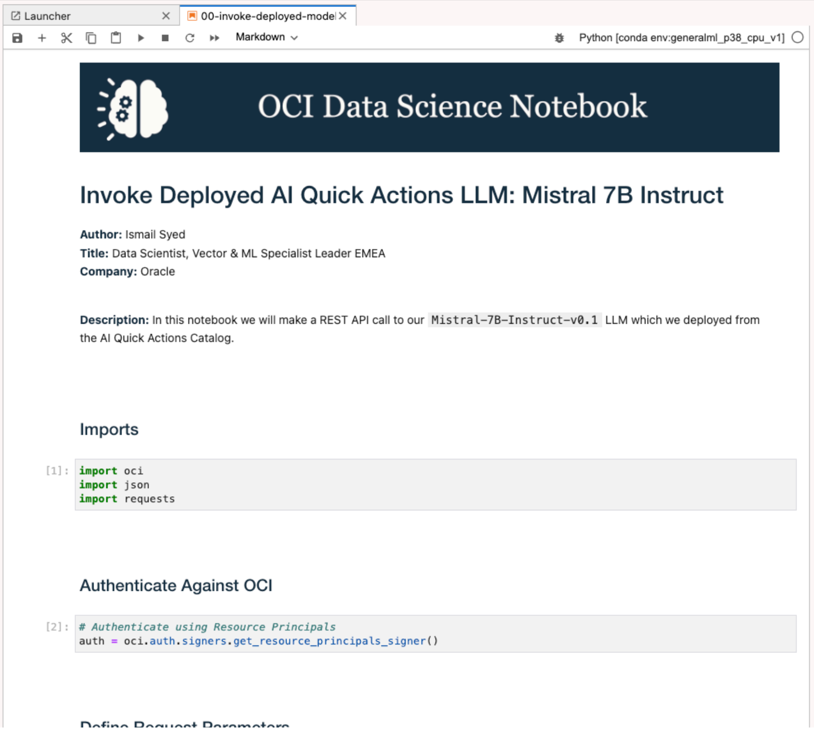
```
# The OCI SDK must be installed for this examp
# Installation instructions can be found here:
import oci
import requests
from oci.signer import Signer
import ssecleint # install with pip install ss

config = oci.config.from_file("~/.oci/config")
auth = Signer(
  tenancy=config['tenancy'],
  user=config['user'],
  fingerprint=config['fingerprint'],
```

Assuming you have the necessary policies and dynamic groups in place to use a resource principal.

Please refer to the following Notebook which provides some sample code.

00-invoke-deployed-model.ipynb

Remember to replace the Model Endpoint with your model along with editing the JSON Body with the parameters you would like sent to the LLM.



Launcher
00-invoke-deployed-model
Markdown
Python [conda env:generalml_p38_cpu_v1]
OCI Data Science Notebook

Invoke Deployed AI Quick Actions LLM: Mistral 7B Instruct

Author: Ismail Syed
Title: Data Scientist, Vector & ML Specialist Leader EMEA
Company: Oracle

Description: In this notebook we will make a REST API call to our `Mistral-7B-Instruct-v0.1` LLM which we deployed from the AI Quick Actions Catalog.

Imports

```
[1]: import oci
     import json
     import requests
```

Authenticate Against OCI

```
[2]: # Authenticate using Resource Principals
     auth = oci.auth.signers.get_resource_principals_signer()
```

Define Request Parameters

If you are unable to use Resource Principals, then use the code provided under the ***'Invoking your Model'*** section on your Model Deployment Page.