Premier League + ORACLE

# Oracle AI

## Winning with AI

Sponsored by

NVIDIA.

Premier League + ORACLE

Oracle AI Workshop

# **Desperte o Poder da IA:** Acelere seus Agents com GPUs NVIDIA e LLMs direto do Hugging Face em poucos minutos

Premier League + ORACLE

# Speaker

## Rafael Dias

Principal AI Engineer

www.linkedin.com/in/rafael-roberto-dias-data-lover/

# Declaração de porto seguro

O seguinte destina-se a descrever a nossa direção geral de produto. Destina-se apenas a fins informativos e não pode ser incorporado em nenhum contrato. Não é um compromisso entregar qualquer material, código ou funcionalidade, e não deve ser confiável na tomada de decisões de compra. O desenvolvimento, lançamento, tempo e preços de quaisquer recursos ou funcionalidades descritos para os produtos da Oracle podem mudar e permanecem a critério exclusivo da Oracle Corporation.

# OCI AI Services

## Applications

- Fusion Applications
- Fusion Analytics
- NetSuite
- Industry Applications
- 3rd Party Applications

Embedded Generative and Classic AI

## AI Services

GenAI Capabilities

- GenAI
- NEW GenAI Agents
- Digital Assistant
- Speech
- Language
- Vision
- Document Understanding

## ML and GenAI for Data Platforms

- Oracle Database Vector Search
- Autonomous Database Select AI
- MySQL HeatWave Vector Store and GenAI
- Data Science
- ML in Oracle Database
- MySQL HeatWave AutoML
- Data Labeling

## Data

## AI Infrastructure

- Supercluster with RDMA networking
- Compute GPU: bare metal, VMs, Kubernetes
- Block, Object, File Storage, HPC filesystems

AI Partners and ISVs

# OCI AI Services

## Applications

- Fusion Applications
- Fusion Analytics
- NetSuite
- Industry Applications
- 3rd Party Applications

Embedded Generative and Classic AI

## AI Services

GenAI Capabilities

- GenAI
- NEW GenAI Agents
- Digital Assistant
- Speech
- Language
- Vision
- Document Understanding

## ML and GenAI for Data Platforms

- Oracle Database Vector Search
- Autonomous Database Select AI
- MySQL HeatWave Vector Store and GenAI
- Data Science
- ML in Oracle Database
- MySQL HeatWave AutoML
- Data Labeling

## Data

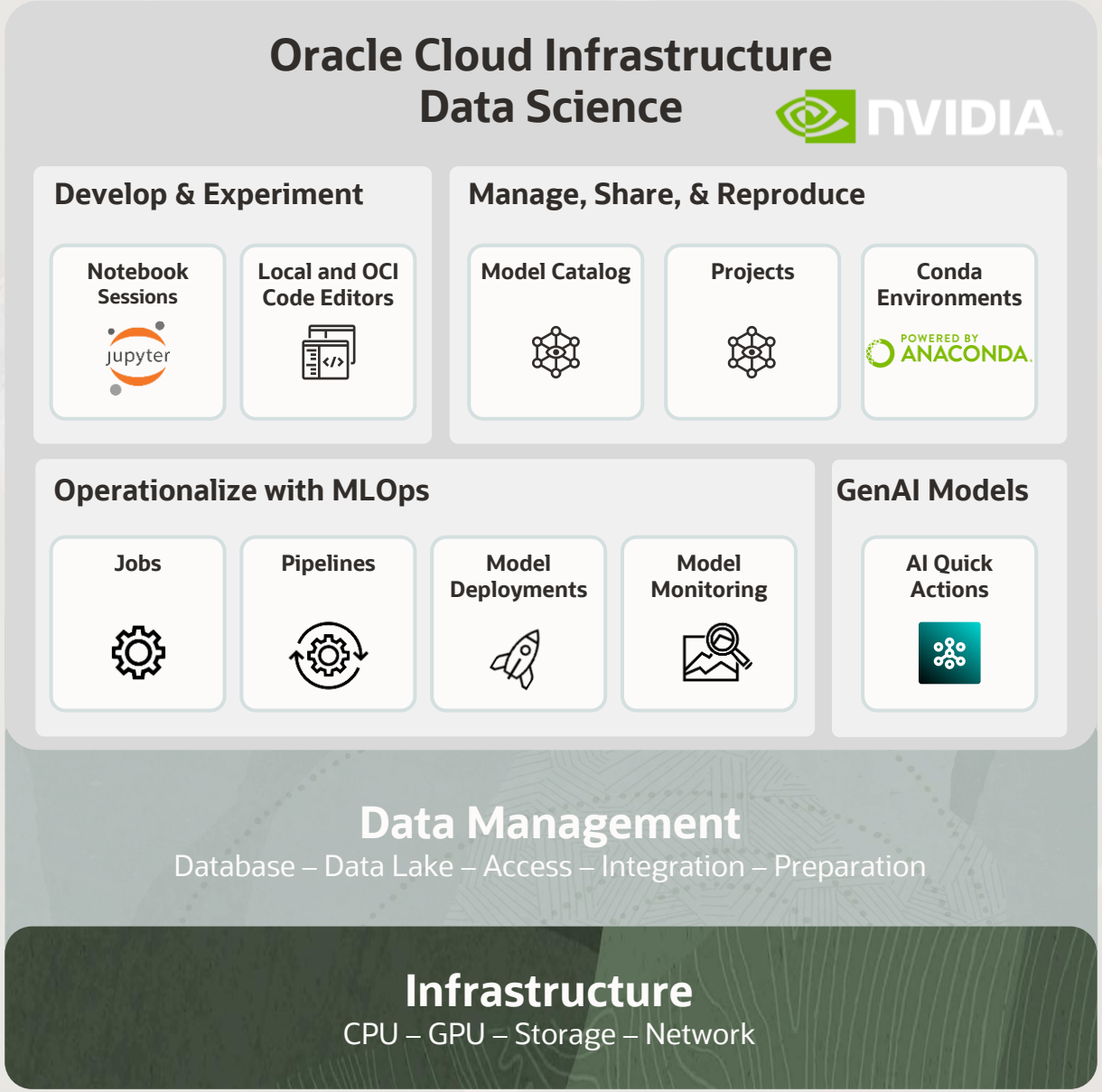
## AI Infrastructure

- Supercluster with RDMA networking
- Compute GPU: bare metal, VMs, Kubernetes
- Block, Object, File Storage, HPC filesystems

AI Partners and ISVs

# OCI Data Science

- Acelere e automatize todo o ciclo de vida de ciência de dados de ponta a ponta
- Use suas ferramentas e estruturas Python de código aberto favoritas
- MLOps de nível empresarial com interfaces flexíveis e escala ilimitada
- Colabore com colegas de equipe em ativos de ciência de dados compartilháveis e reproduzíveis
- Execute cargas de trabalho em grande escala com acesso a GPUs bare metal e processamento
- Pague apenas pela infraestrutura sob demanda, sem impostos ou despesas adicionais



# OCI Data Science – AI quick actions

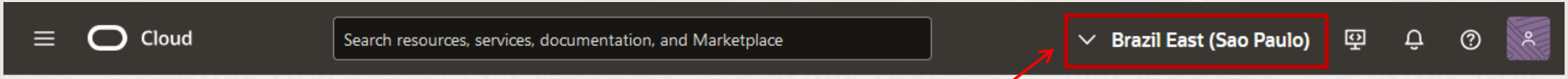
- Use GPUs NVIDIA avançadas: A10, L40S, H100
- Deploy e Fine tune de vários Foundation Models
- Fine tunning utilizando dados próprios
- Teste o modelo logo após colocá-lo em produção
- Avalie seus modelos para garantir a qualidade
- Implante LLMs usando servidores de inferência especializados como TGI e vLLM
- Bring your own model do HuggingFace ou fine tuned

# Lab I – Registrar Modelo LLM do Hugging Face

AI Quick Actions

## Passo 1:Tenancy Região de São Paulo (GRU)

**1.1 Necessário ter a tenancy subscrita na região de São Paulo, para utilizar o OCI Generatice AI**

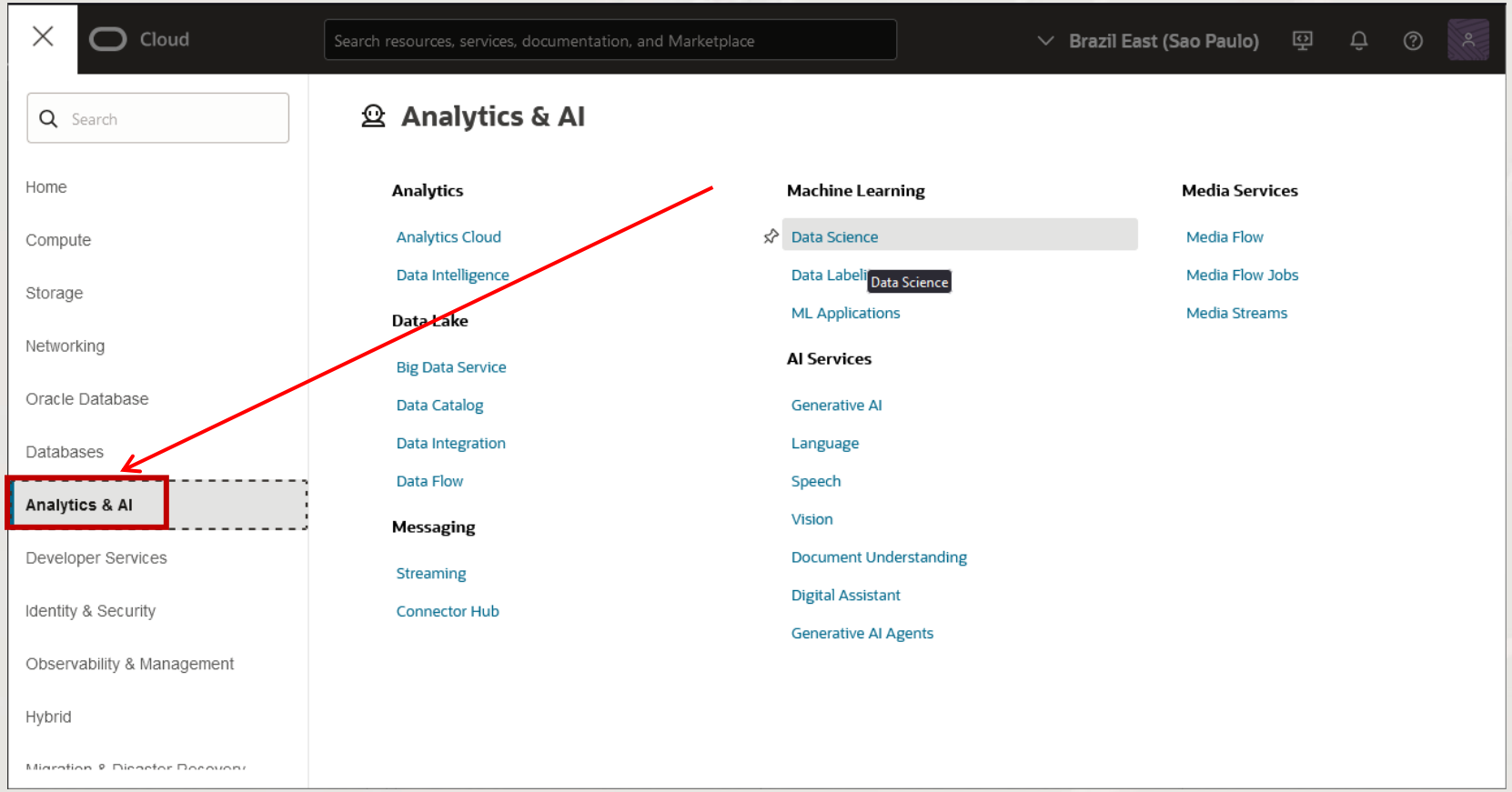


# Lab I – Registrar Modelo LLM do Hugging Face

AI Quick Actions

## Passo 2: Data Science

### 2.1 No menu, selecione "Analytics & AI"

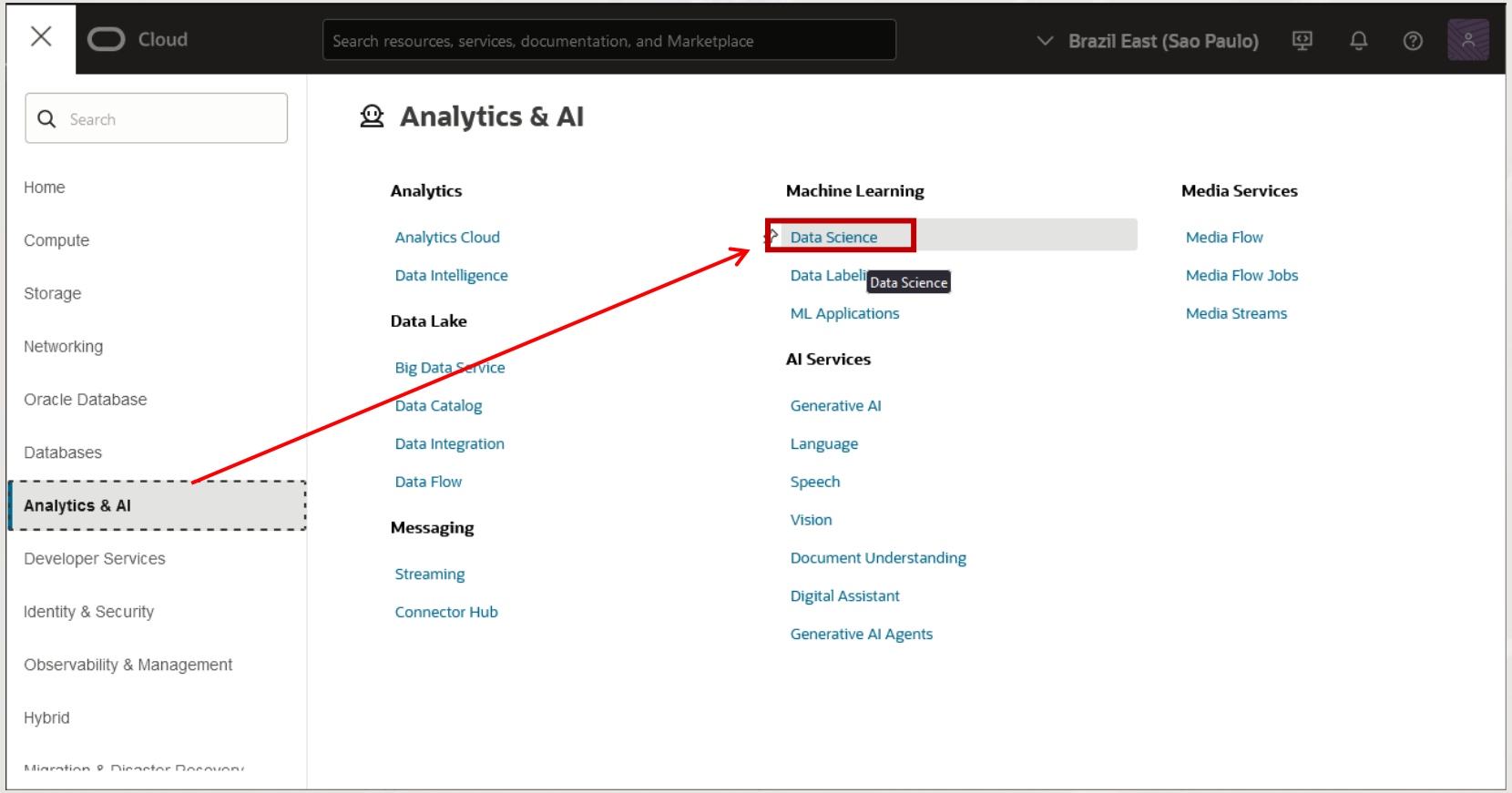


# Lab I – Registrar Modelo LLM do Hugging Face

AI Quick Actions

## Passo 2: Data Science

### 2.2 Em "Analytics & AI", selecione "Data Science"

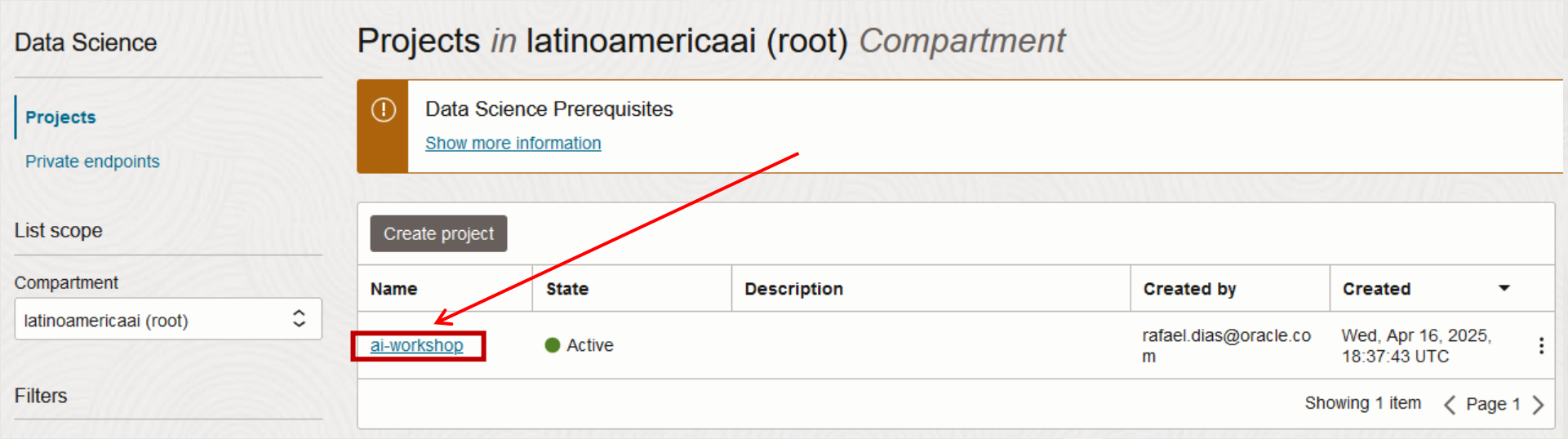


# Lab I – Registrar Modelo LLM do Hugging Face

AI Quick Actions

## Passo 2: Data Science

### 2.3 Em "Data Science", clique no projeto ‘ai-workshop’

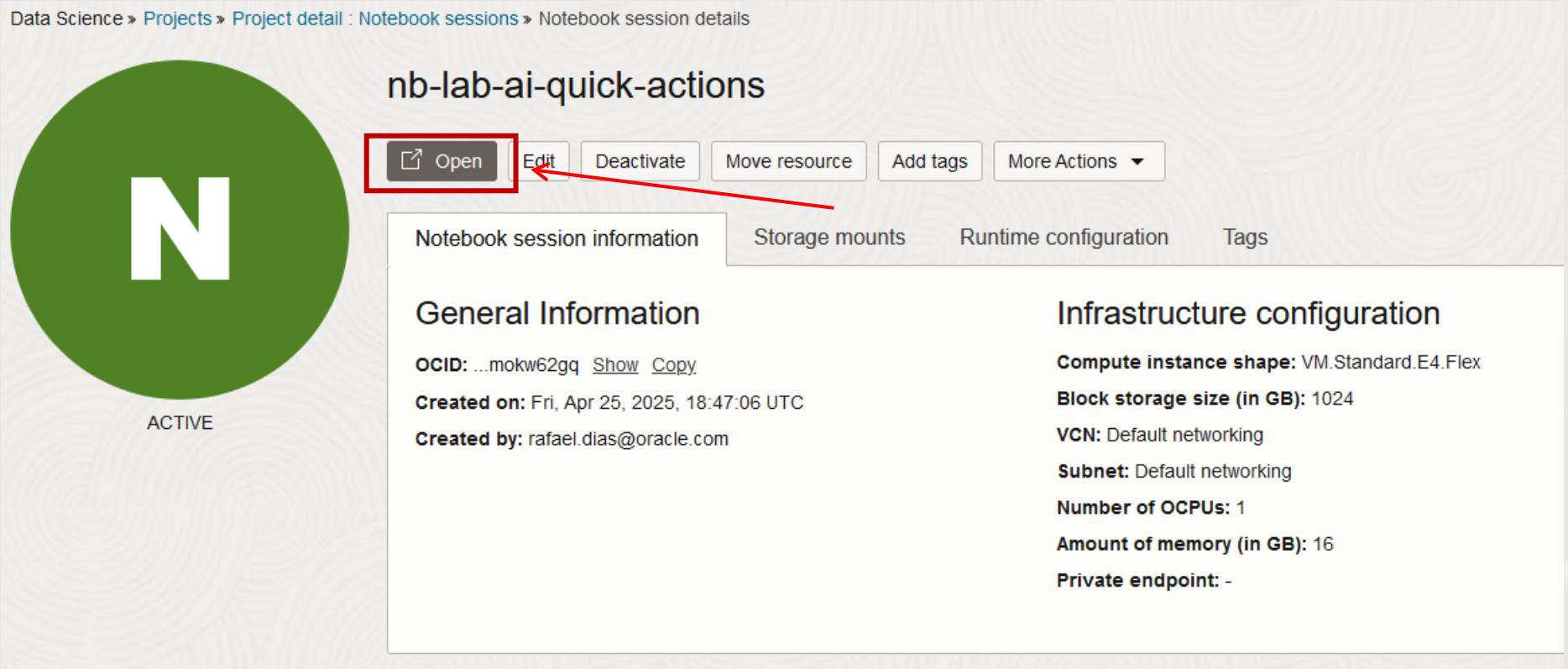


# Lab I – Registrar Modelo LLM do Hugging Face

AI Quick Actions

## Passo 2: Data Science

### 2.5 Em " nb-lab-ai-quick-actions ", clique em ‘Open’

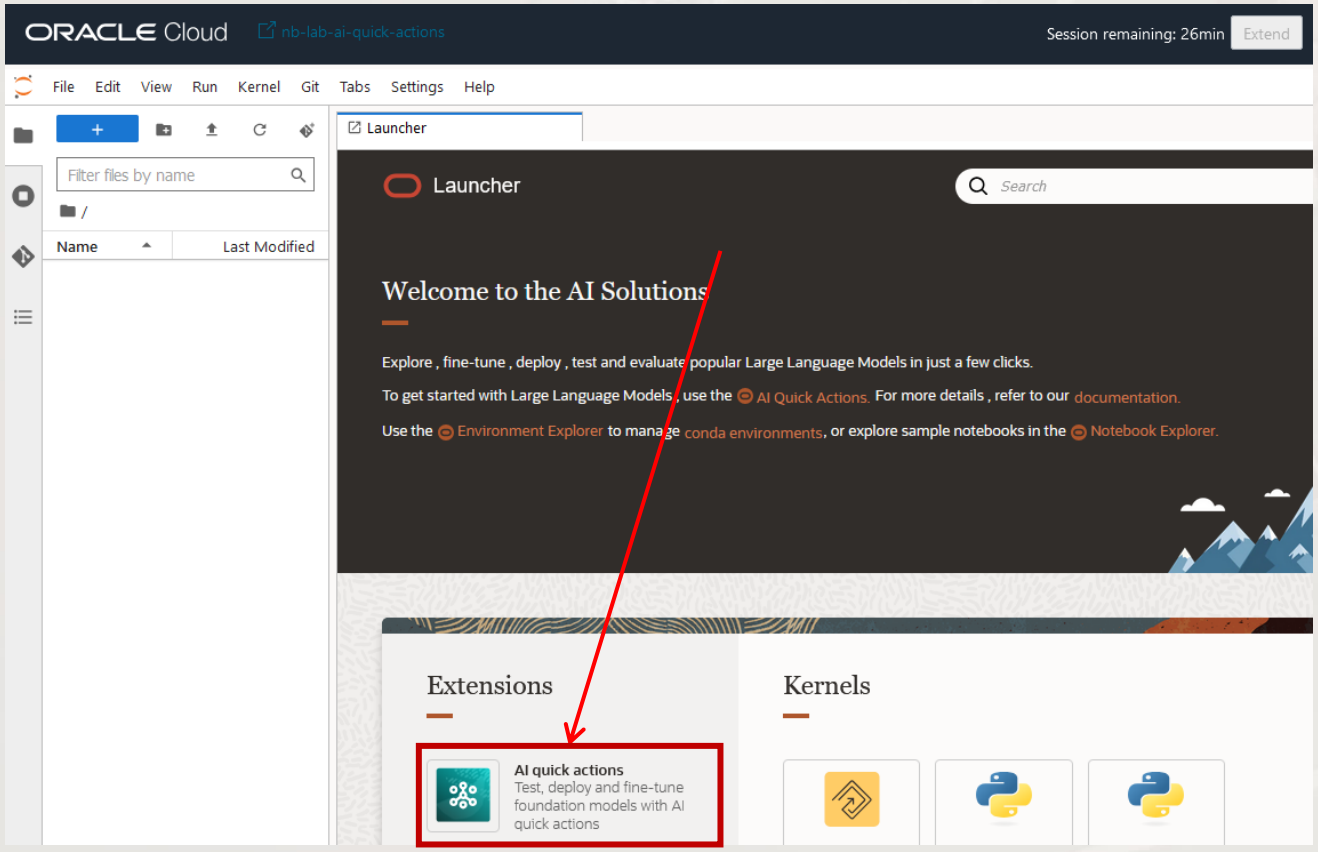


# Lab I – Registrar Modelo LLM do Hugging Face

AI Quick Actions

## Passo 3: Notebook session

### 3.1 No Notebook, clique em ‘AI quick actions’

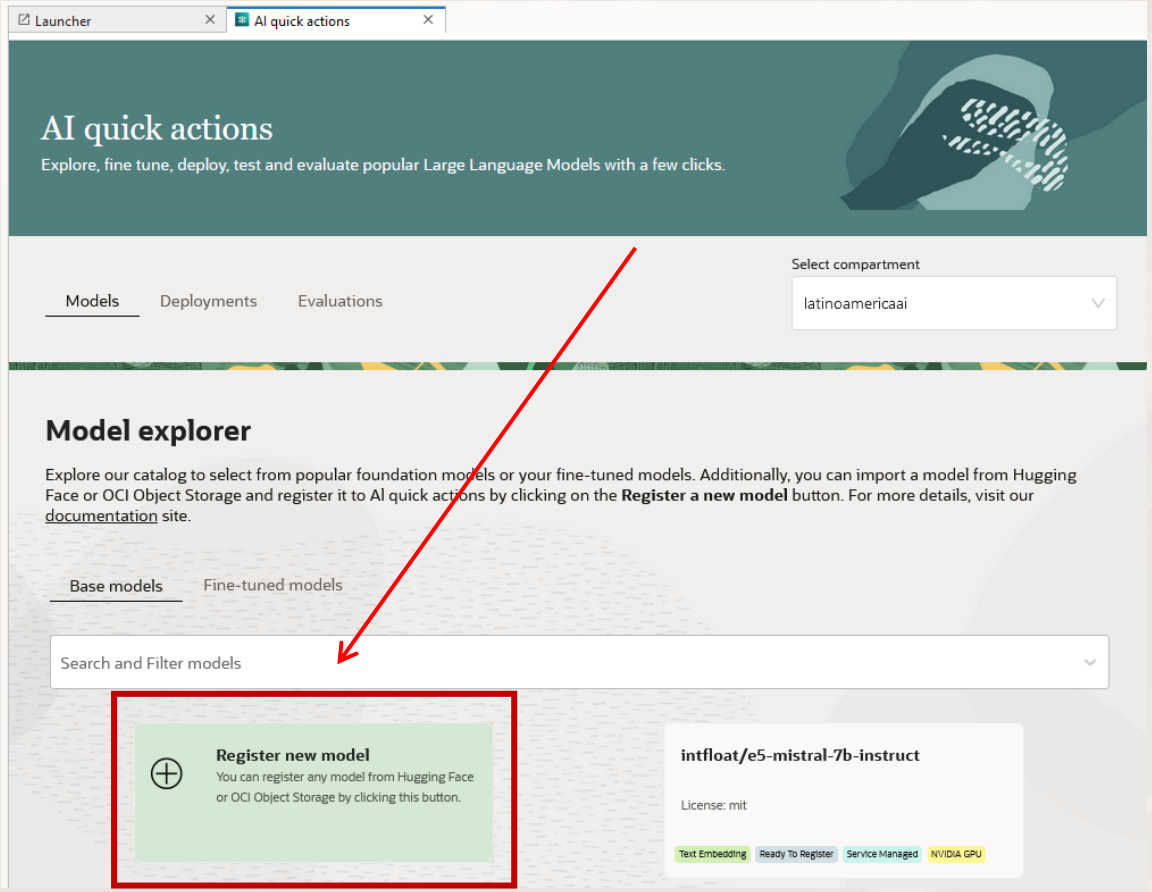


# Lab I – Registrar Modelo LLM do Hugging Face

AI Quick Actions

## Passo 4: AI quick actions

### 4.1 No AI quick actions, clique em ‘Register new model’

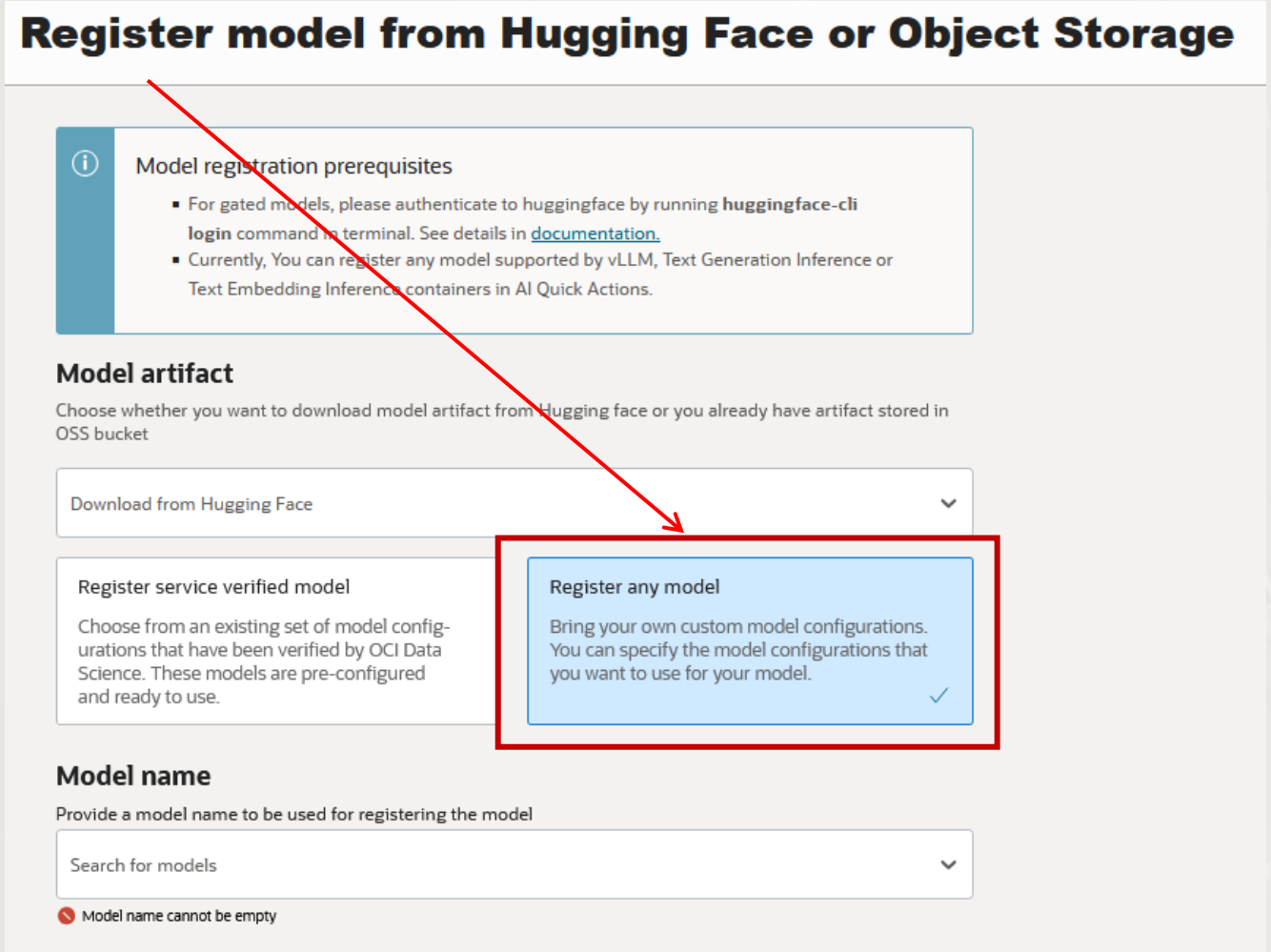


# Lab I – Registrar Modelo LLM do Hugging Face

## Passo 4: AI quick actions

AI Quick Actions

### 4.2 Em ‘Register model from…’, clique em ‘Register any model’

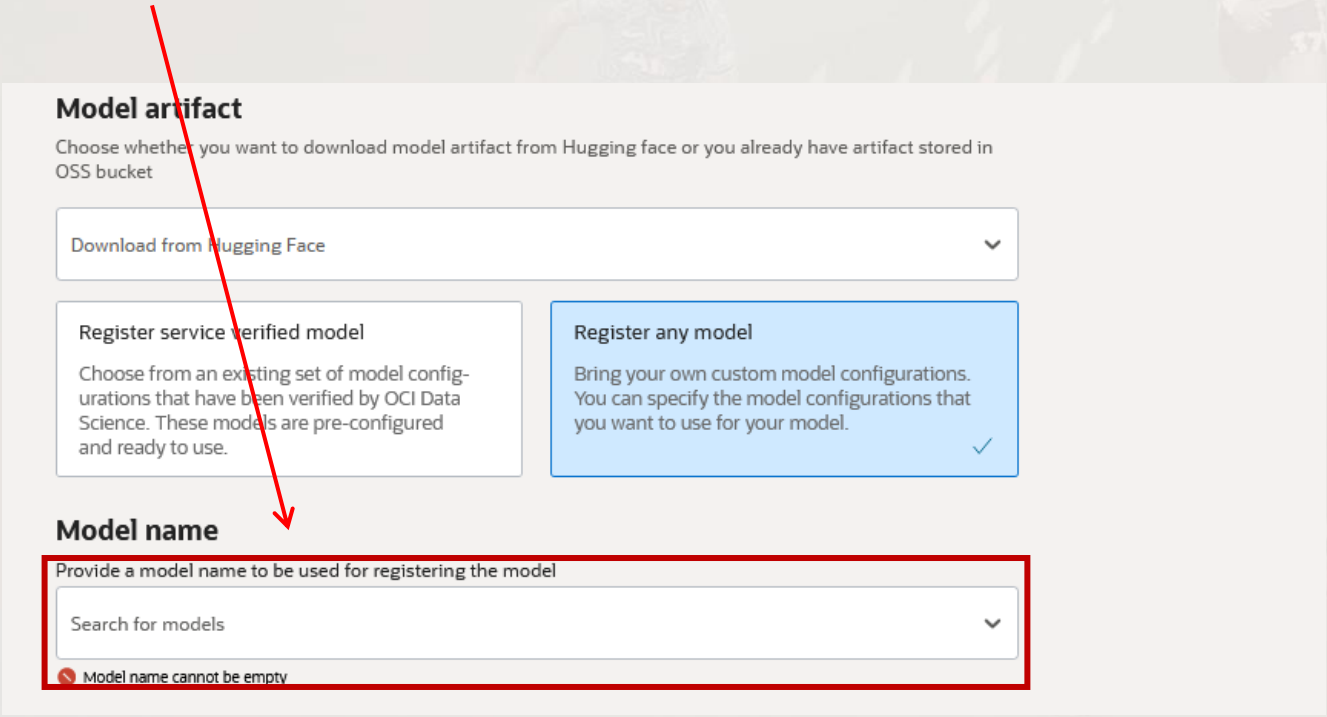


# Lab I – Registrar Modelo LLM do Hugging Face

## Passo 4: AI quick actions

AI Quick Actions

### 4.3 Ainda em 'Register model from…', clique em 'Search for models'

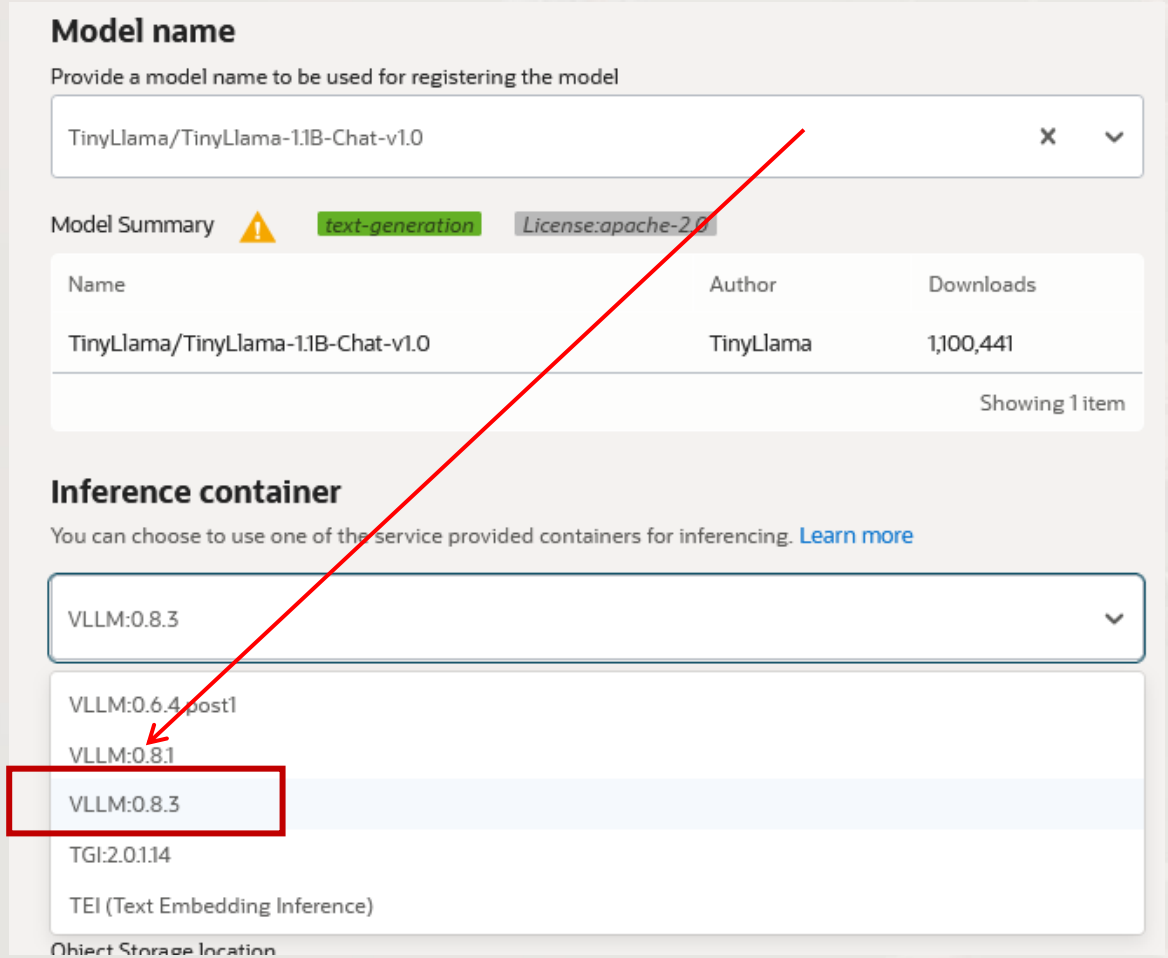


## Lab I – Registrar Modelo LLM do Hugging Face

# Passo 4: AI quick actions

AI Quick Actions

**4.5 Irá aparecer alguns detalhes sobre o LLM Tiny, após clicar em 'Inference container' e escolher a opção 'VLLM:0.8.3'**

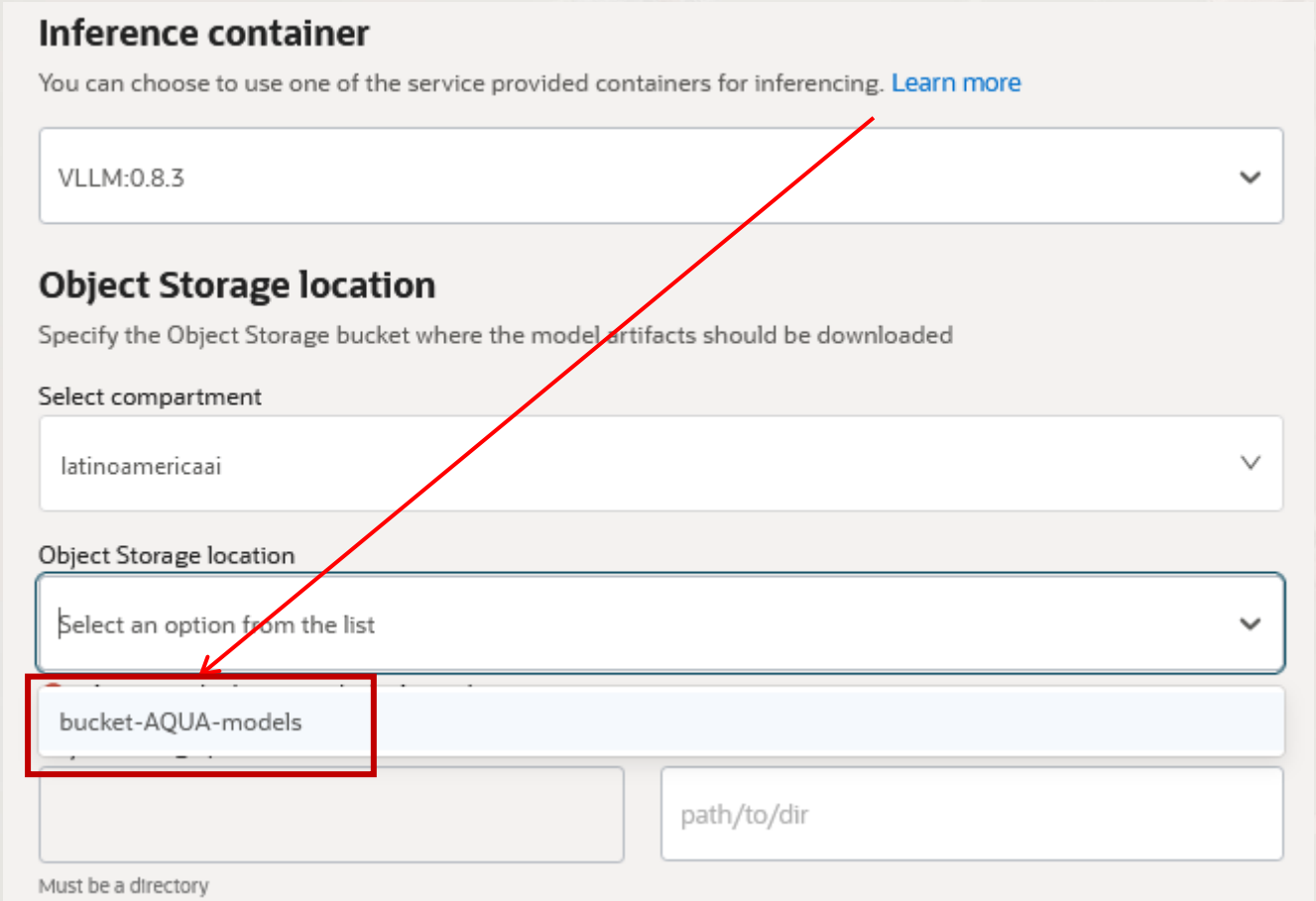


## Lab I – Registrar Modelo LLM do Hugging Face

# Passo 4: AI quick actions

AI Quick Actions

### 4.6 Logo abaixo selecionar em ‘Object Storage location’ o ‘bucket-AQUA-models’

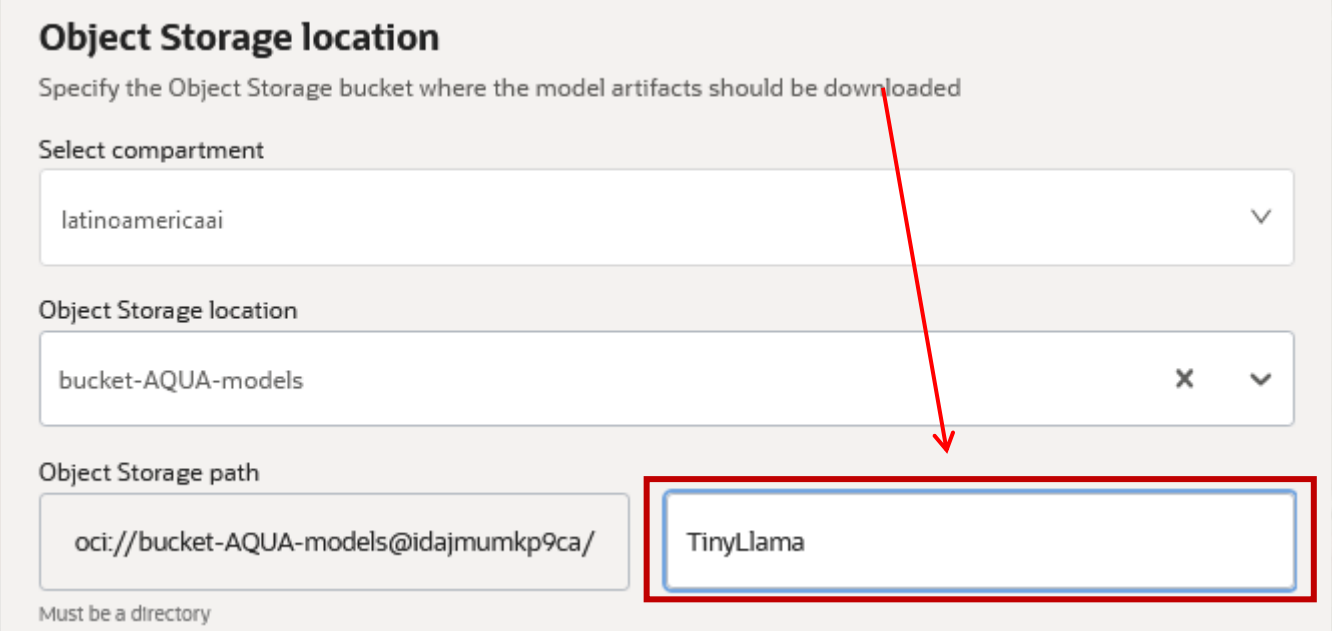


## Lab I – Registrar Modelo LLM do Hugging Face

# Passo 4: AI quick actions

AI Quick Actions

### 4.7 E em ‘path/to/dir’ colocar o nome ‘TinyLlama’

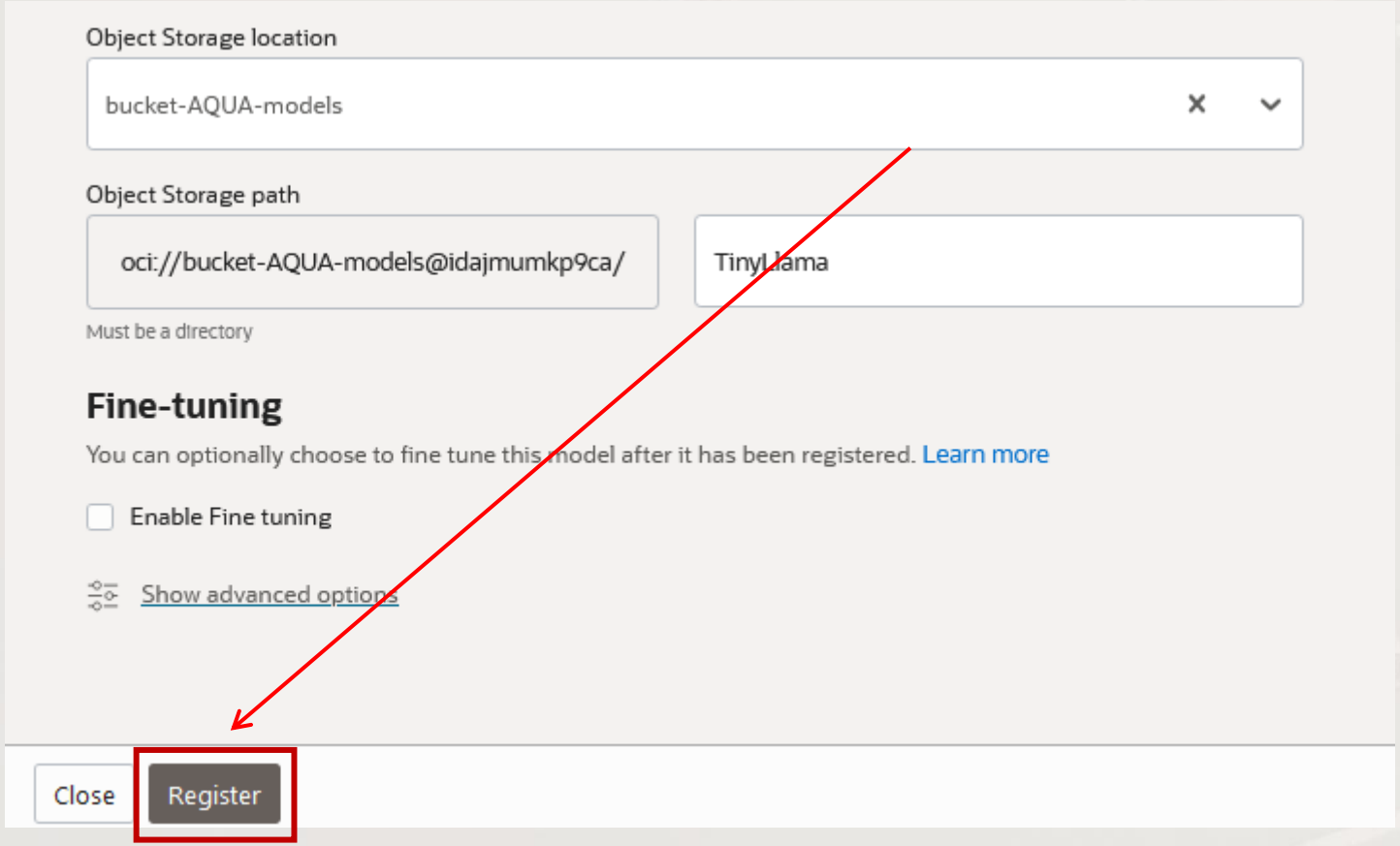


# Lab I – Registrar Modelo LLM do Hugging Face

## Passo 4: AI quick actions

AI Quick Actions

### 4.8 E finalmente clicar em ‘Register’



# Lab I – Registrar Modelo LLM do Hugging Face

## Passo 4: AI quick actions

AI Quick Actions

### 4.9 Em poucos minutos finalizará o registro e aparecerá conforme abaixo

Launcher | AI quick actions | datascience@:~

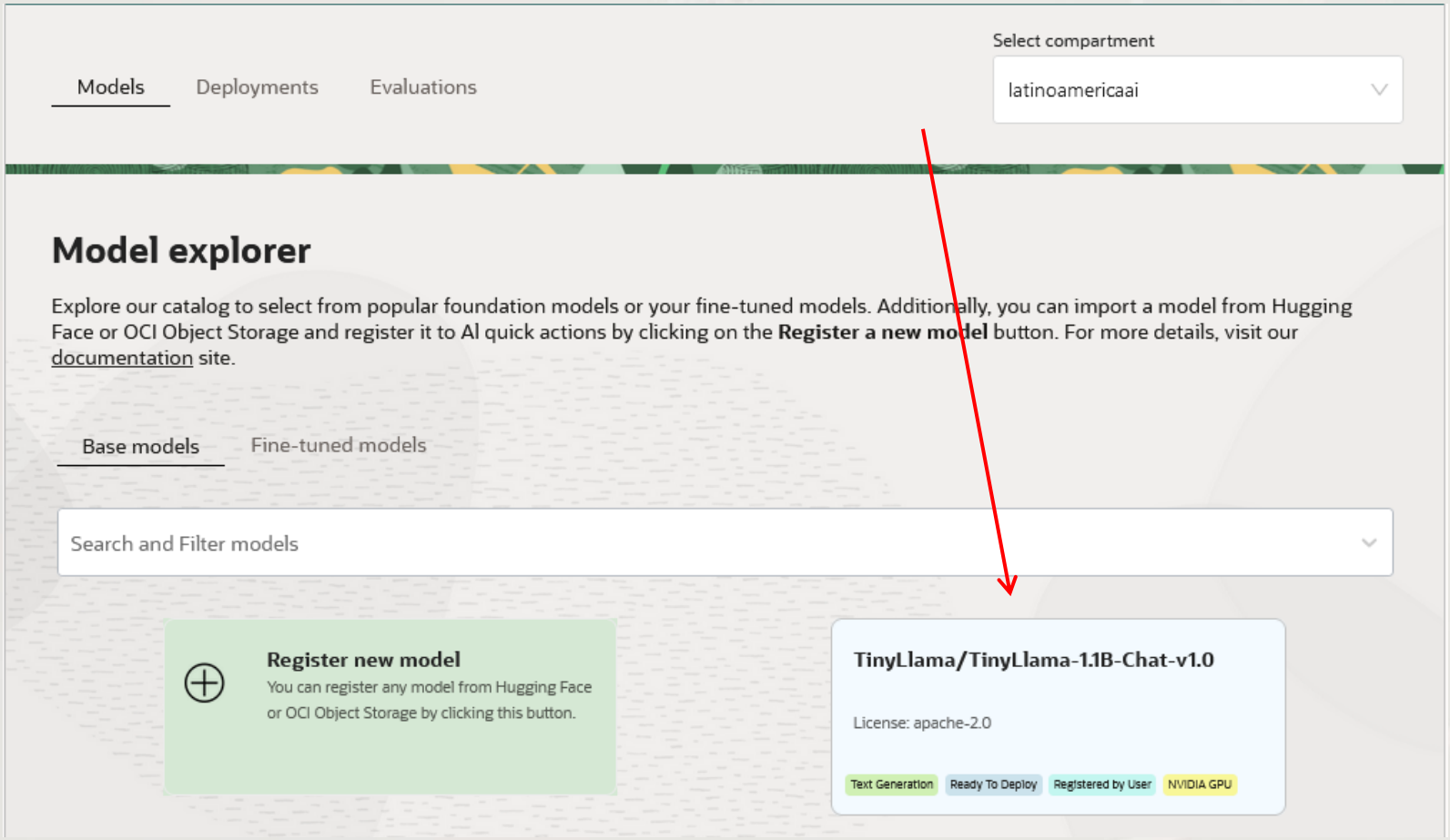

```
    "TinyLlama/TinyLlama/TinyLlama-1.1B-Chat-v1.0/README.md",
    "TinyLlama/TinyLlama/TinyLlama-1.1B-Chat-v1.0/eval_results.json"
  ],
  "upload-failures": {},
  "uploaded-objects": {}
}
{
    "compartment_id": "ocid1.tenancy.oc1..aaaaaaaakzuzfystmhyj7je5x7ymwd3ofd7wb6rtsfrx7nvbo272vta3rwna",
    "icon": "",
    "id": "ocid1.datasciencemodel.oc1.sa-saopaulo-1.amaaaaaad6nji3aaadrwq4u6gxpulnjkpool6efykjxc6penrw5tnm552nla",
    "is_fine_tuned_model": false,
    "license": "apache-2.0",
    "name": "TinyLlama/TinyLlama-1.1B-Chat-v1.0",
    "organization": "TinyLlama",
    "project_id": "ocid1.datascienceproject.oc1.sa-saopaulo-1.amaaaaaad6nji3aarpr54ridavh63zdopcca7ftl2xeaz75n2dvdkoa7alnq",
    "tags": {
        "Oracle-Tags": {
            "CreatedBy": "ocid1.datasciencenotebooksession.oc1.sa-saopaulo-1.amaaaaaad6nji3aagmk3bthxzhowksry3nemjapmvgwez6gi5gicmokw62gq",
            "CreatedOn": "2025-04-25T20:25:22.226Z"
        },
        "aqua_custom_base_model": "true",
        "license": "apache-2.0",
        "task": "text-generation",
        "model_format": "SAFETENSORS",
        "OCI_AQUA": "active",
        "organization": "TinyLlama"
    },
    "task": "text-generation",
    "time_created": "2025-04-25 20:25:22.326000+00:00",
    "console_link": "https://cloud.oracle.com/data-science/models/ocid1.datasciencemodel.oc1.sa-saopaulo-1.amaaaaaad6nji3aaadrwq4u6gxpulnjkpool6efykjxc6penrw5tnm5
52nla?region=sa-saopaulo-1",
    "search_text": "{'CreatedBy': 'ocid1.datasciencenotebooksession.oc1.sa-saopaulo-1.amaaaaaad6nji3aagmk3bthxzhowksry3nemjapmvgwez6gi5gicmokw62gq', 'CreatedOn':
'2025-04-25T20:25:22.226Z'},true,apache-2.0,text-generation,SAFETENSORS,active,TinyLlama",
    "ready_to_deploy": true,
    "ready_to_finetune": false,
    "ready_to_import": false,
    "nvidia_gpu_supported": true,
    "arm_cpu_supported": false,
    "model_file": "",
    "model_formats": [
        "SAFETENSORS"
    ],
    "inference_container": "odsc-vllm-serving-llama4",
    "inference_container_uri": null,
    "finetuning_container": null,
    "evaluation_container": "odsc-llm-evaluate",
    "artifact_location": "oci://bucket-AQUA-models@idajmumkp9ca/TinyLlama/TinyLlama/TinyLlama-1.1B-Chat-v1.0"
}
(base) bash-4.4$
```

# Lab I – Registrar Modelo LLM do Hugging Face

## Passo 4: AI quick actions

AI Quick Actions

### 4.10 E ao retornar à aba interna 'AI quick actions' será possível ver o LLM Tiny disponível para deploy

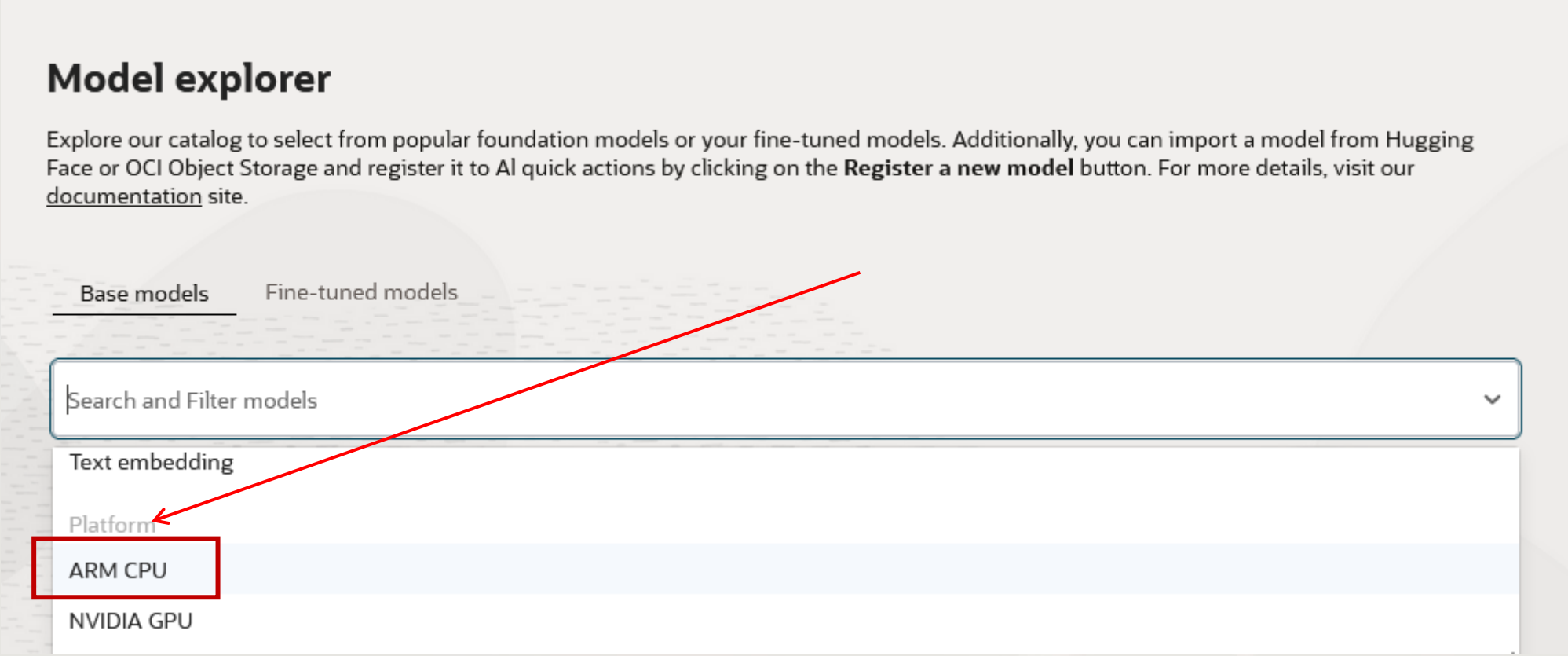


# Lab II – Fazer Deploy de Modelo LLM em ARM

AI Quick Actions

## Passo 1: Model explorer

**1.1 Clicar em 'Search and Filter models', após procurar por 'Platform' e clicar em 'ARM CPU'**

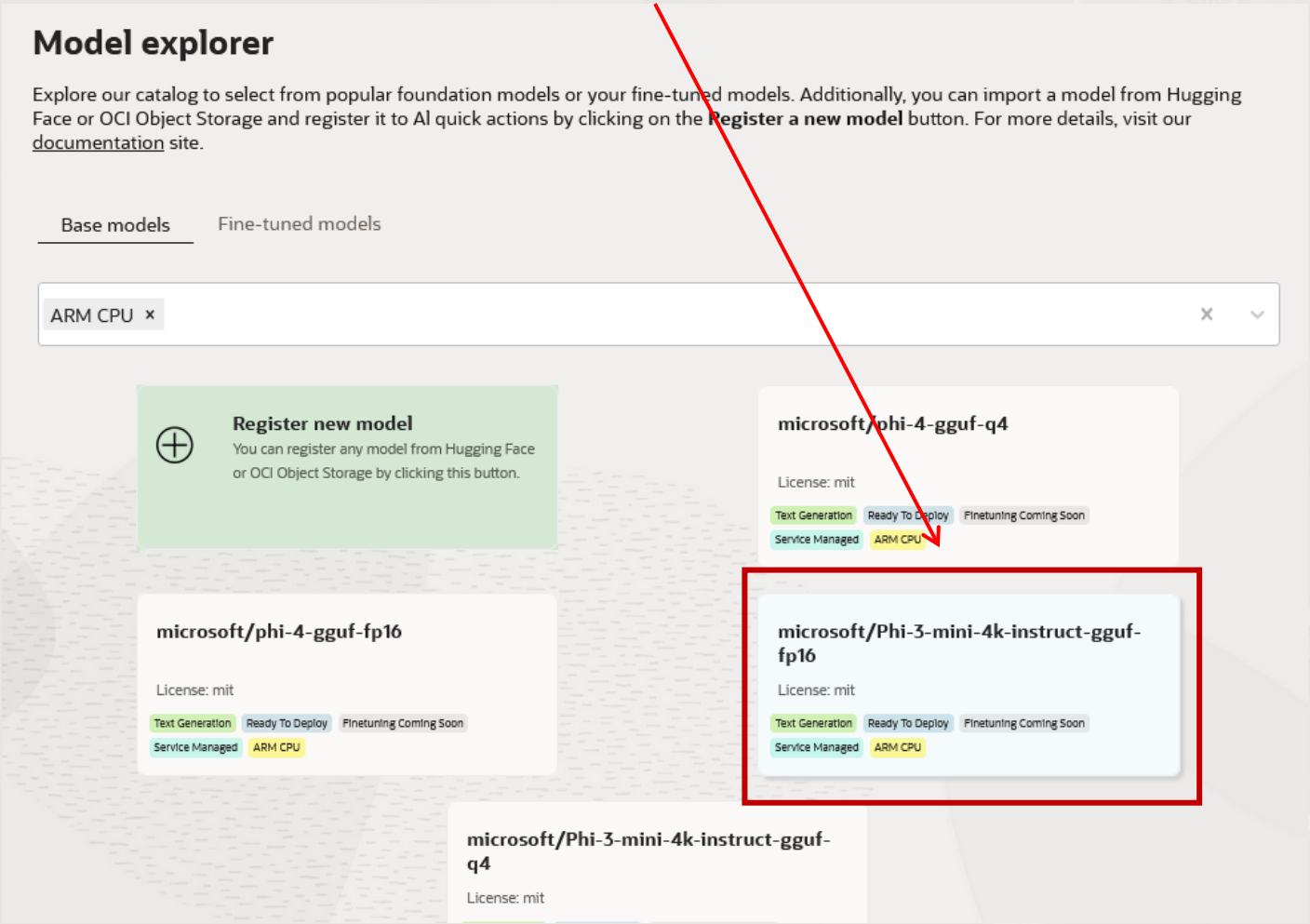


## Lab II – Fazer Deploy de Modelo LLM em ARM

# Passo 1: Model explorer

AI Quick Actions

### 1.2 Após clicar no modelo LLM ‘microsoft/Phi-3-mini-4k-instruct-gguf-fp16’

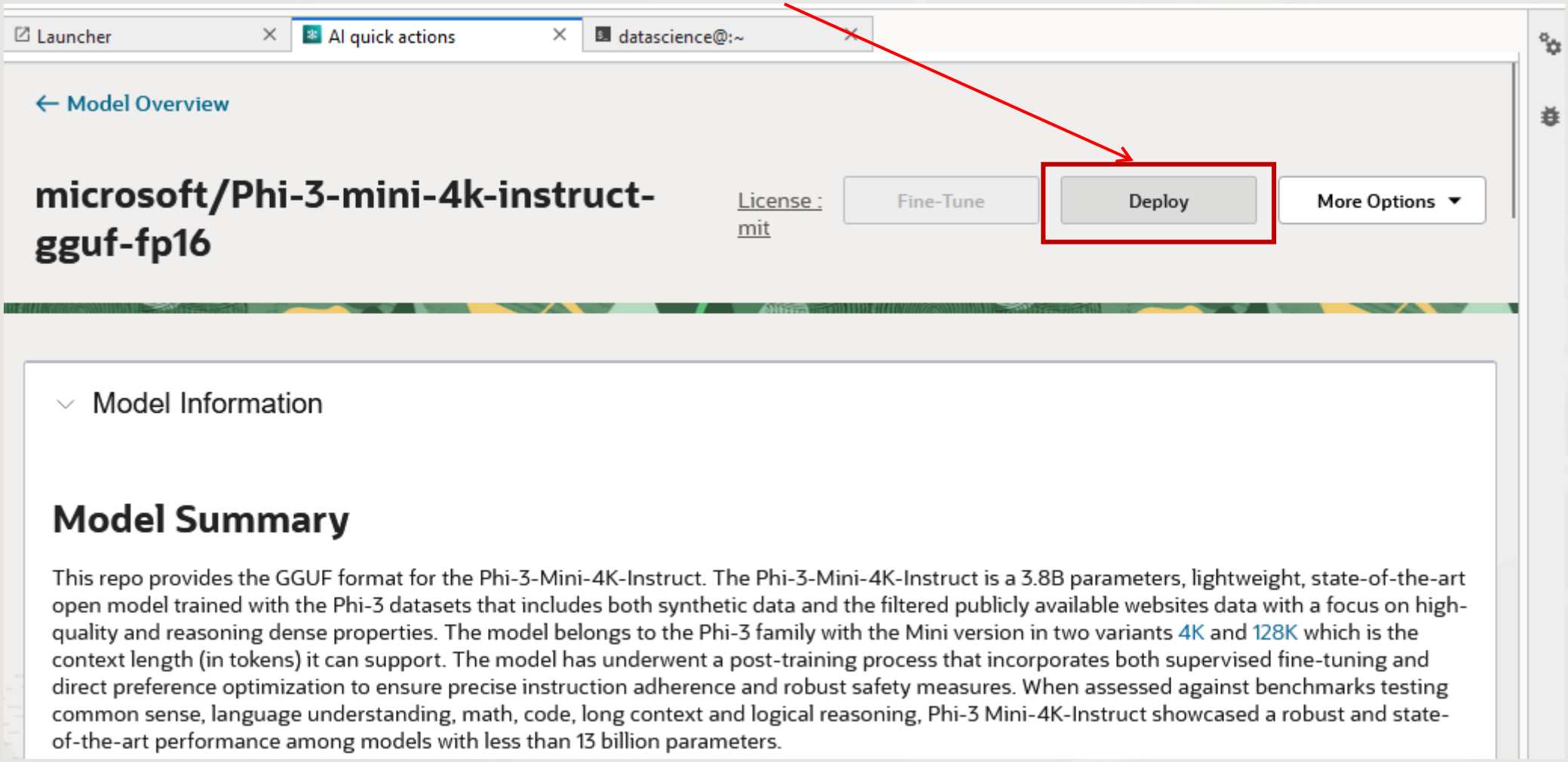


## Lab II – Fazer Deploy de Modelo LLM em ARM

# Passo 1: Model explorer

AI Quick Actions

### 1.3 E então clicar em ‘Deploy’

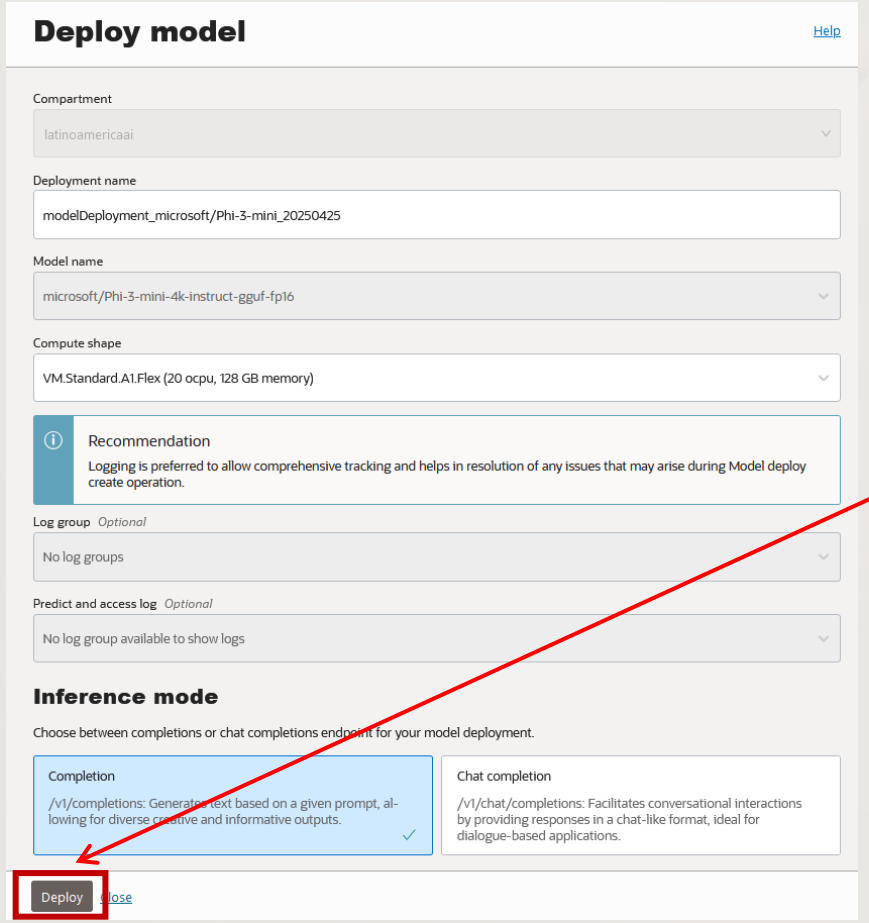


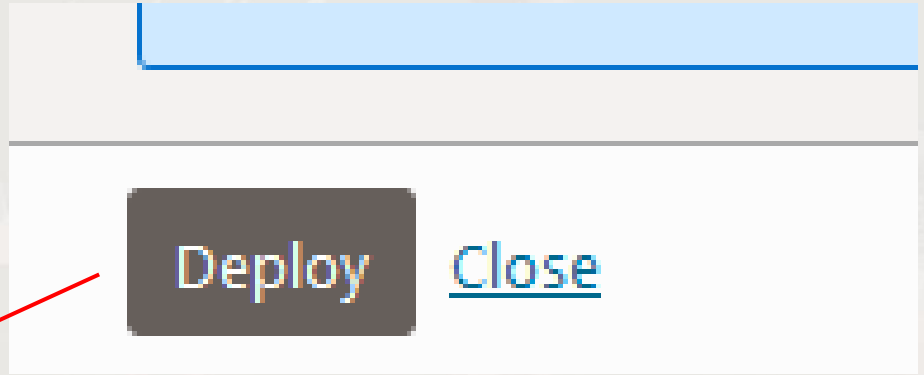
# Lab II – Fazer Deploy de Modelo LLM em ARM

## Passo 2: Deploy model

AI Quick Actions

### 2.1 Na próxima tela pode manter as configurações iniciais e clicar em ‘Deploy’



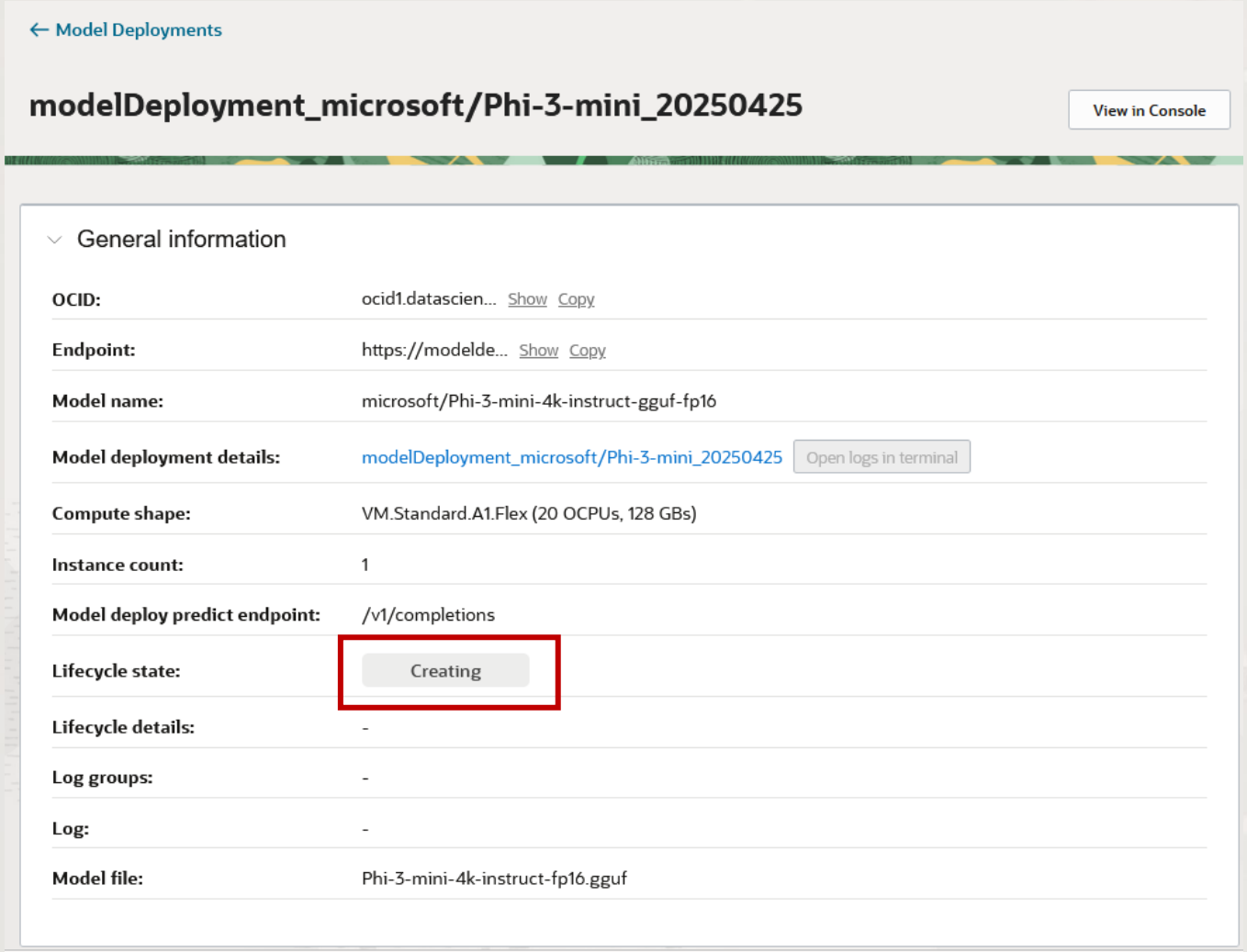



# Lab II – Fazer Deploy de Modelo LLM em ARM

## Passo 2: Deploy model

AI Quick Actions

### 2.2 Aparecerá a tela abaixo com o ‘Lifecycle state’ como ‘Creating’



## Lab II – Fazer Deploy de Modelo LLM em ARM

# Passo 2: Deploy model

AI Quick Actions

**2.3 Após alguns poucos minutos, ao finalizar o deploy, o 'Lifecycle state' aparecerá como 'Active'**

General information

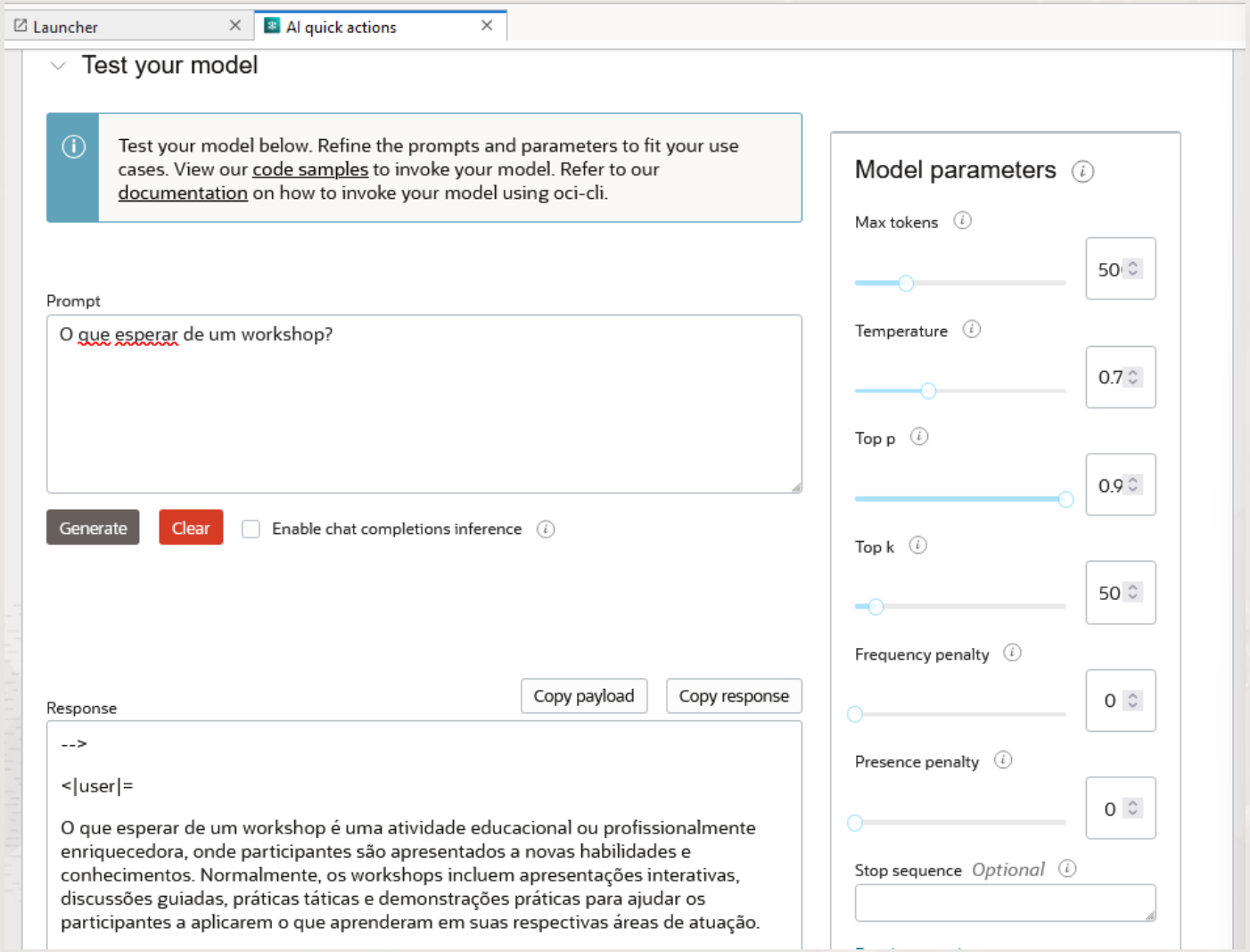
| | |
|---|---|
| **OCID:** | ocid1.datascien... Show Copy |
| **Endpoint:** | https://modelde... Show Copy |
| **Model name:** | microsoft/Phi-3-mini-4k-instruct-gguf-fp16 |
| **Model deployment inference:** | Invoke your model |
| **Model deployment details:** | modelDeployment_microsoft/Phi-3-mini_20250425 Open logs in terminal |
| **Compute shape:** | VM.Standard.A1.Flex (20 OCPUs, 128 GBs) |
| **Instance count:** | 1 |
| **Model deploy predict endpoint:** | /v1/completions |
| **Lifecycle state:** | Active |
| **Lifecycle details:** | Model Deployment is Active. |
| **Log groups:** | - |
| **Log:** | - |

# Lab II – Fazer Deploy de Modelo LLM em ARM

## Passo 2: Deploy model

AI Quick Actions

### 2.4 E rolando a página mais para baixo, será possível testar prompt com este modelo



# Obrigado

ORACLE