



Oracle AI

—
Winning with AI

Sponsored by





Oracle AI Workshop

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





Speaker



Rafael Dias

Principal AI Engineer

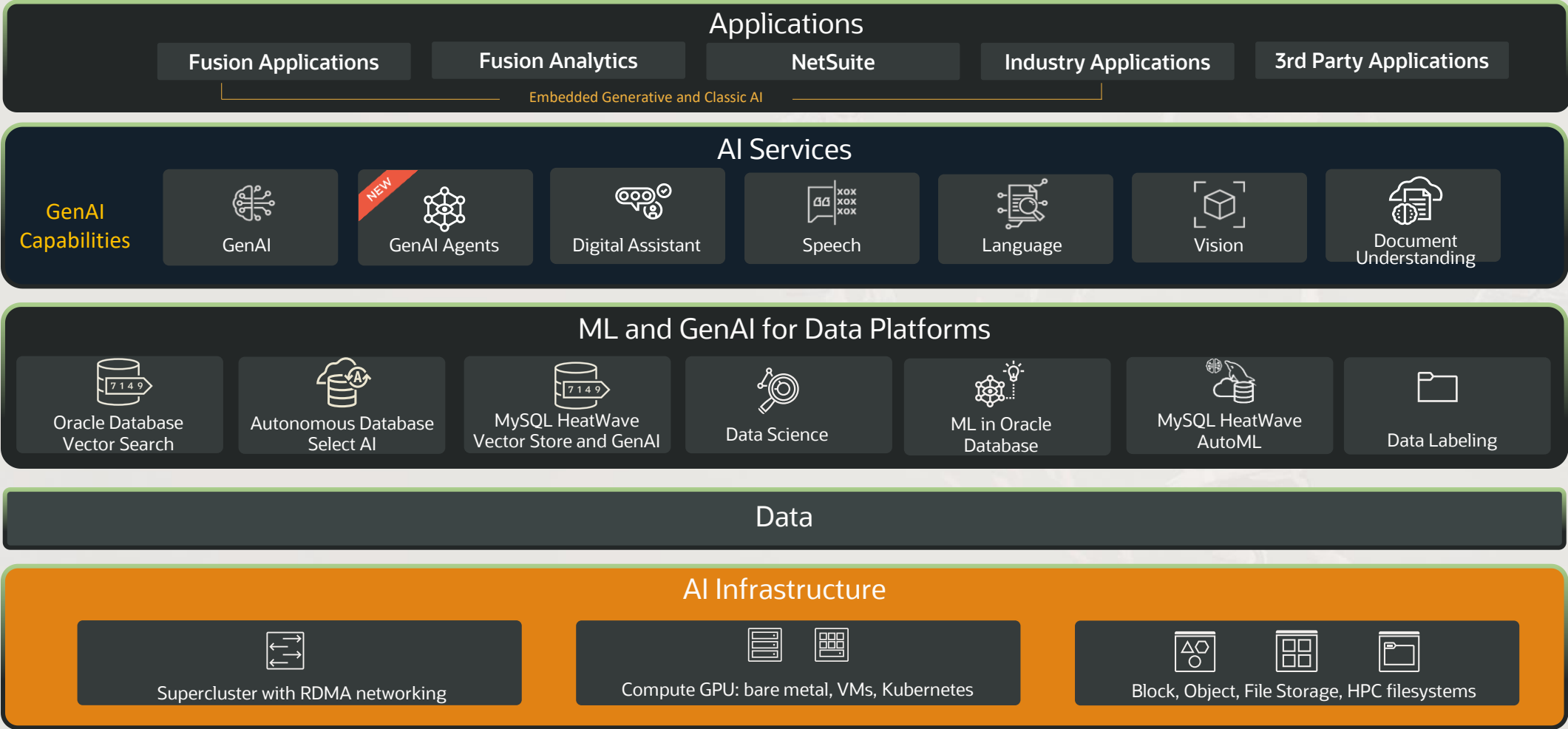
www.linkedin.com/in/rafael-roberto-dias-data-lover/



Declaração de porto seguro

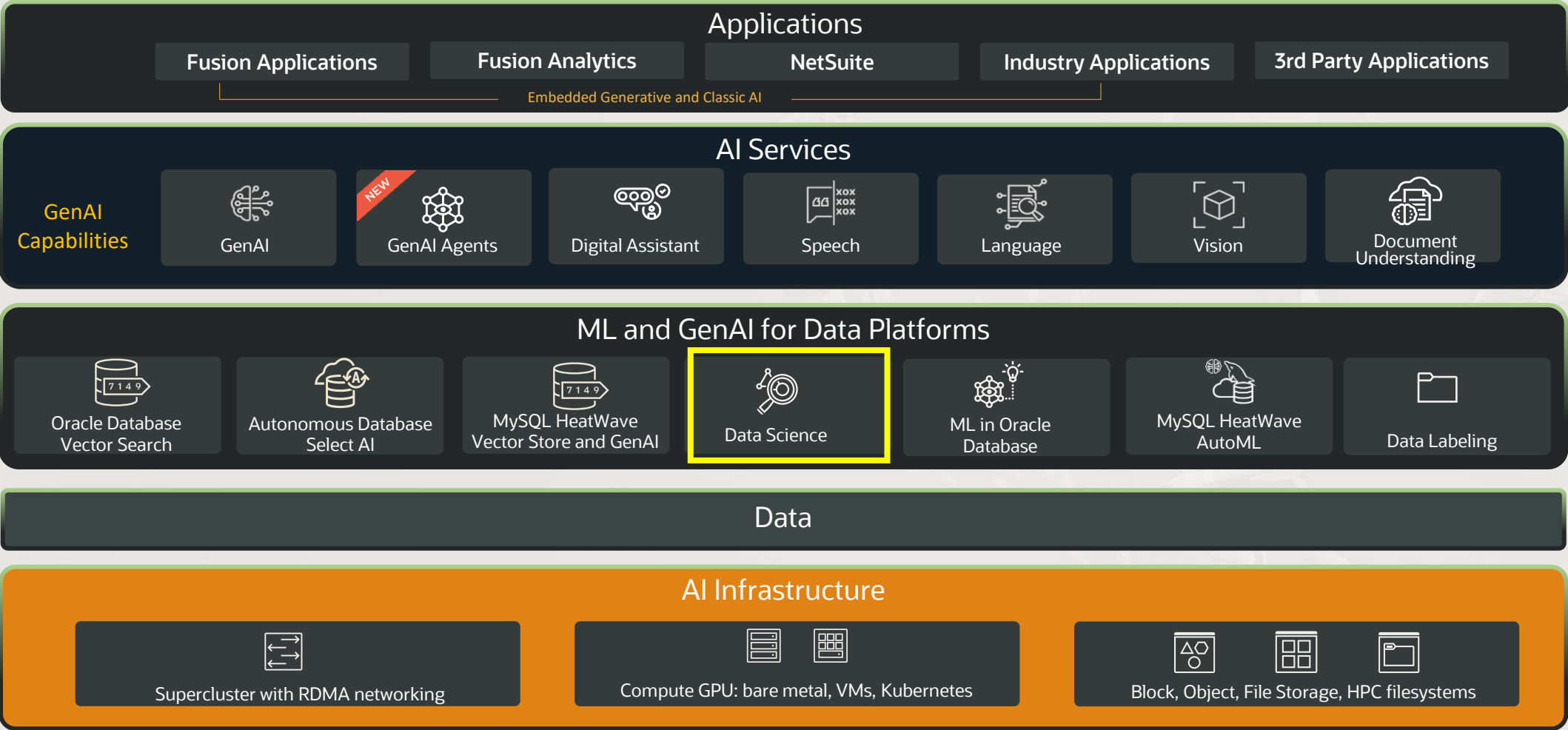
O seguinte destina-se a descrever a nossa direção geral de produto. Destina-se apenas a fins informativos e não pode ser incorporado em nenhum contrato. Não é um compromisso entregar qualquer material, código ou funcionalidade, e não deve ser confiável na tomada de decisões de compra. O desenvolvimento, 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



AI Partners and ISVs

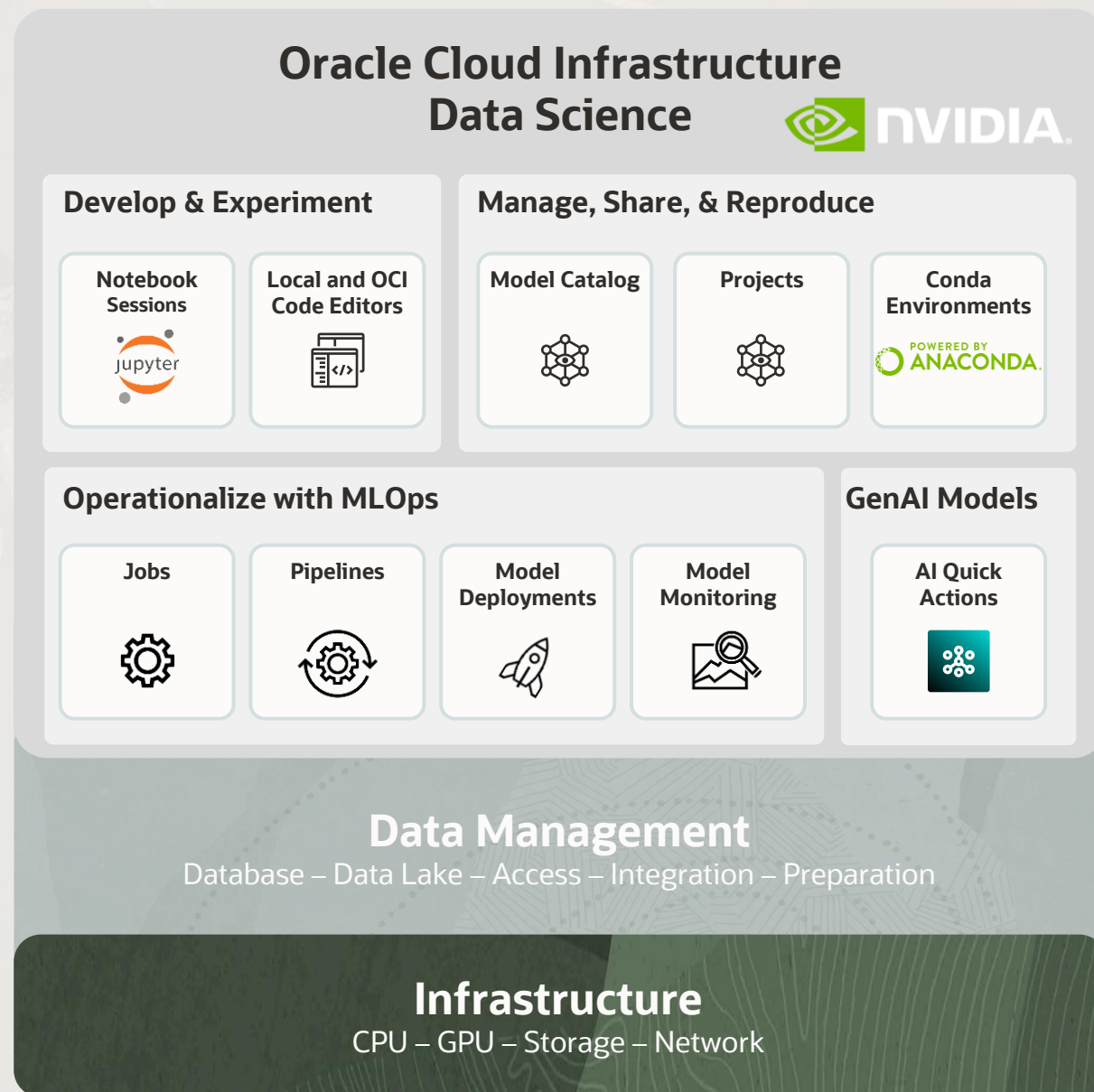
OCI AI Services



AI Partners and ISVs

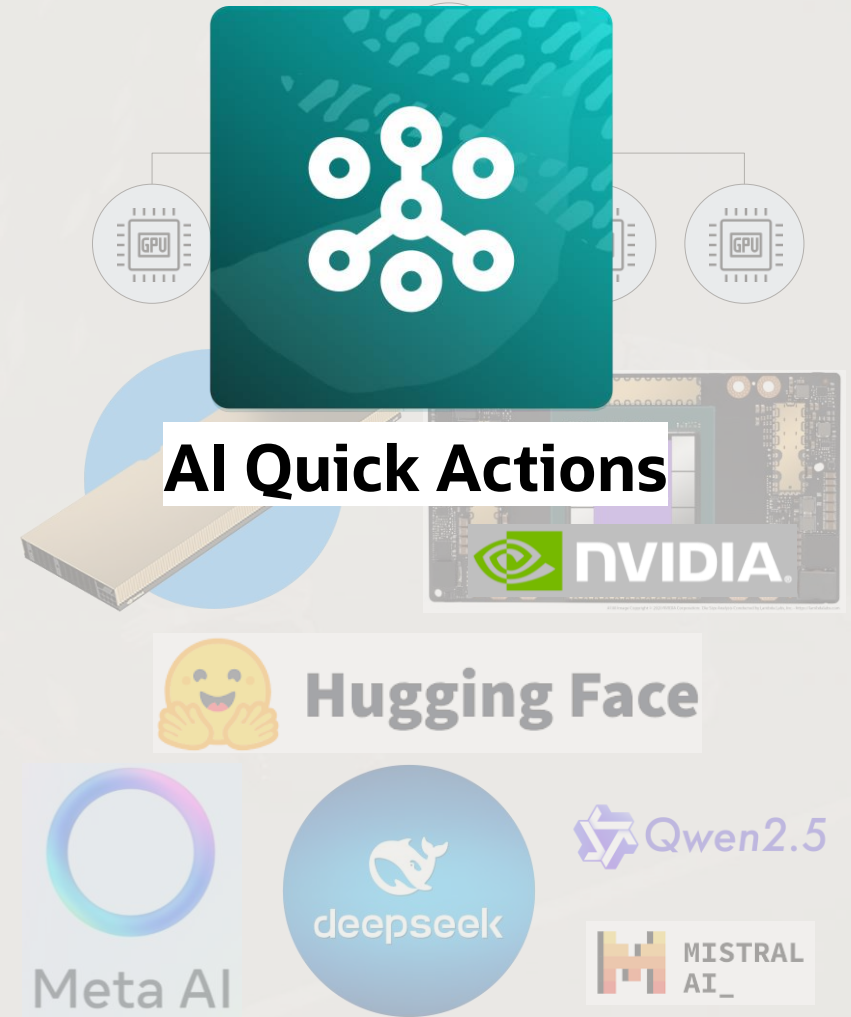
OCI Data Science

- 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 tuning 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



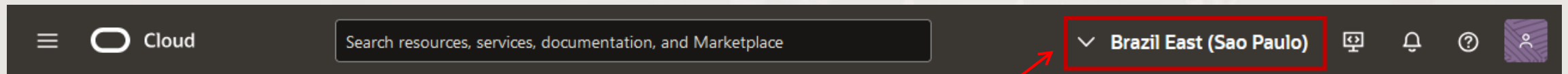
Lab I – Registrar Modelo LLM do Hugging Face



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 Generative AI



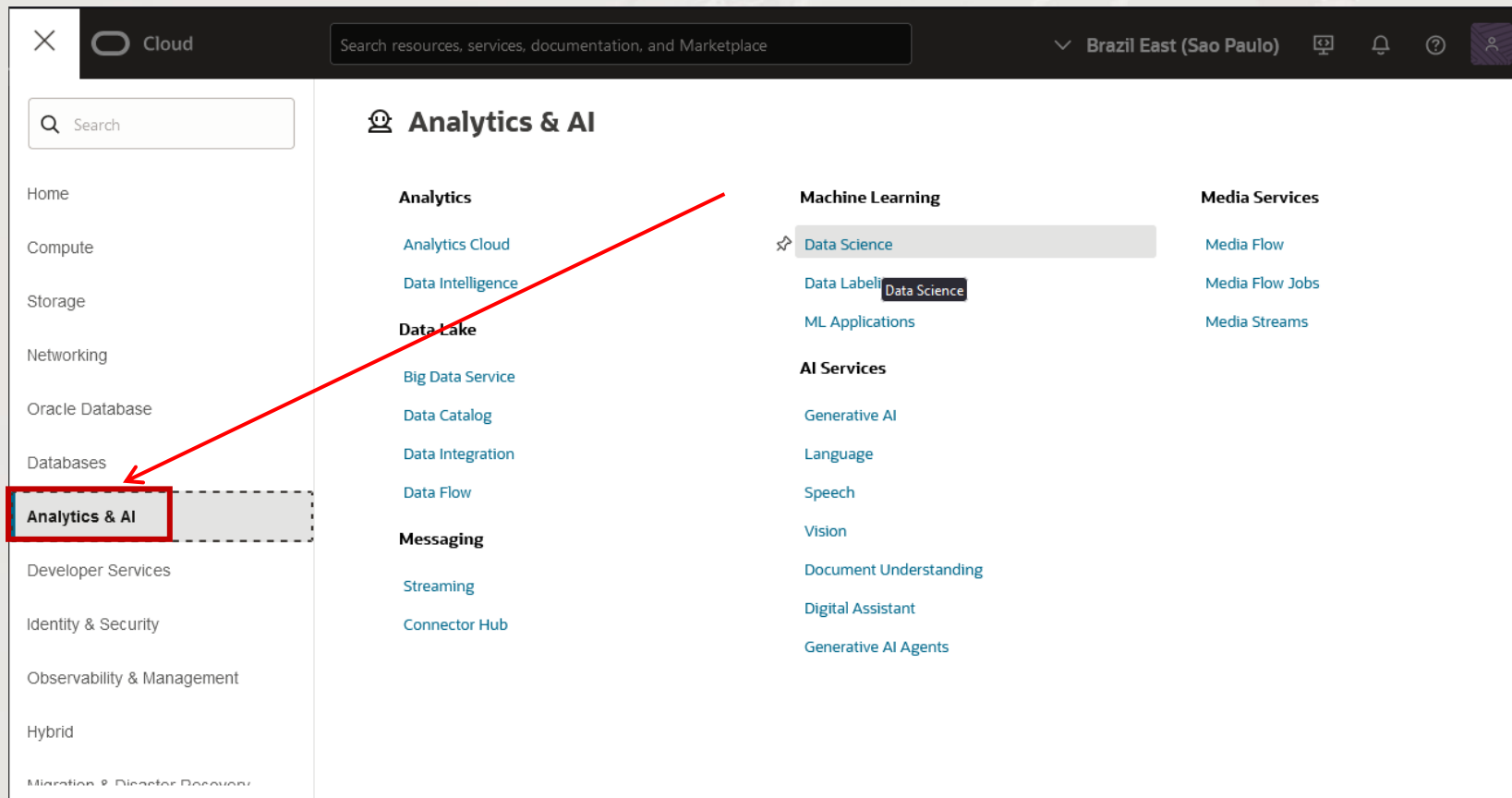
Lab I – Registrar Modelo LLM do Hugging Face



AI Quick Actions

Passo 2: Data Science

2.1 No menu, selecione "Analytics & AI"



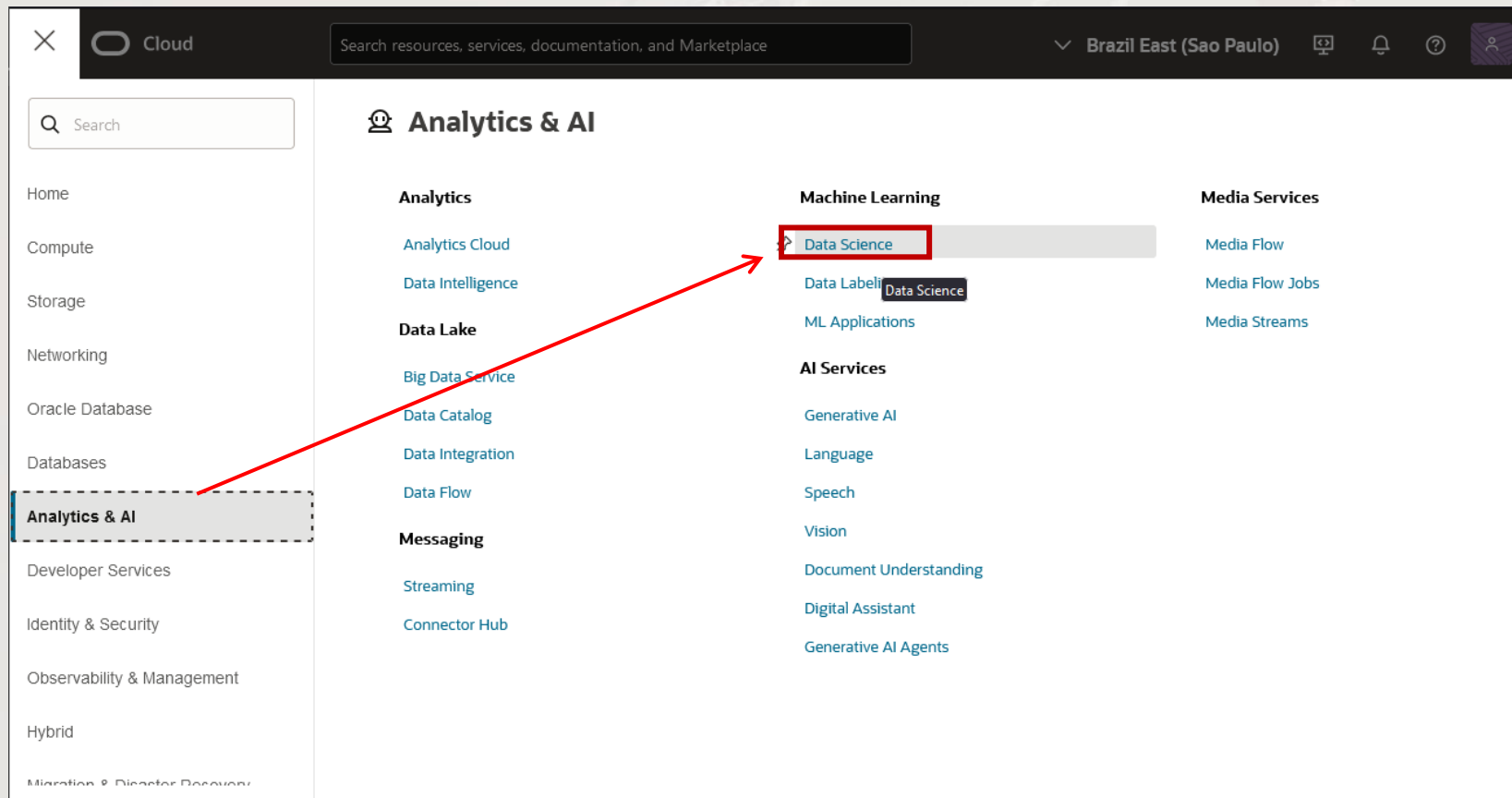
Lab I – Registrar Modelo LLM do Hugging Face



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'

Data Science

Projects

[Private endpoints](#)

List scope

Compartment

latinoamericaai (root)

Filters

Projects *in* latinoamericaai (root) *Compartment*

! Data Science Prerequisites
[Show more information](#)

Create project

Name	State	Description	Created by	Created
ai-workshop	● Active		rafael.dias@oracle.com	Wed, Apr 16, 2025, 18:37:43 UTC

Showing 1 item < Page 1 >

Lab I – Registrar Modelo LLM do Hugging Face




AI Quick Actions

Passo 2: Data Science

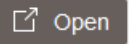

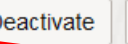
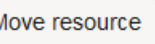

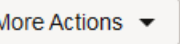
2.5 Em " nb-lab-ai-quick-actions ", clique em 'Open'

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



ACTIVE

nb-lab-ai-quick-actions

 Open  Edit  Deactivate  Move resource  Add tags  More Actions ▼

Notebook session information

Storage mounts Runtime configuration Tags

General Information

OCID: ...mokw62gq [Show](#) [Copy](#)

Created on: Fri, Apr 25, 2025, 18:47:06 UTC

Created by: raphael.dias@oracle.com

Infrastructure configuration

Compute instance shape: VM.Standard.E4.Flex

Block storage size (in GB): 1024

VCN: Default networking

Subnet: Default networking

Number of OCPUs: 1

Amount of memory (in GB): 16

Private endpoint: -

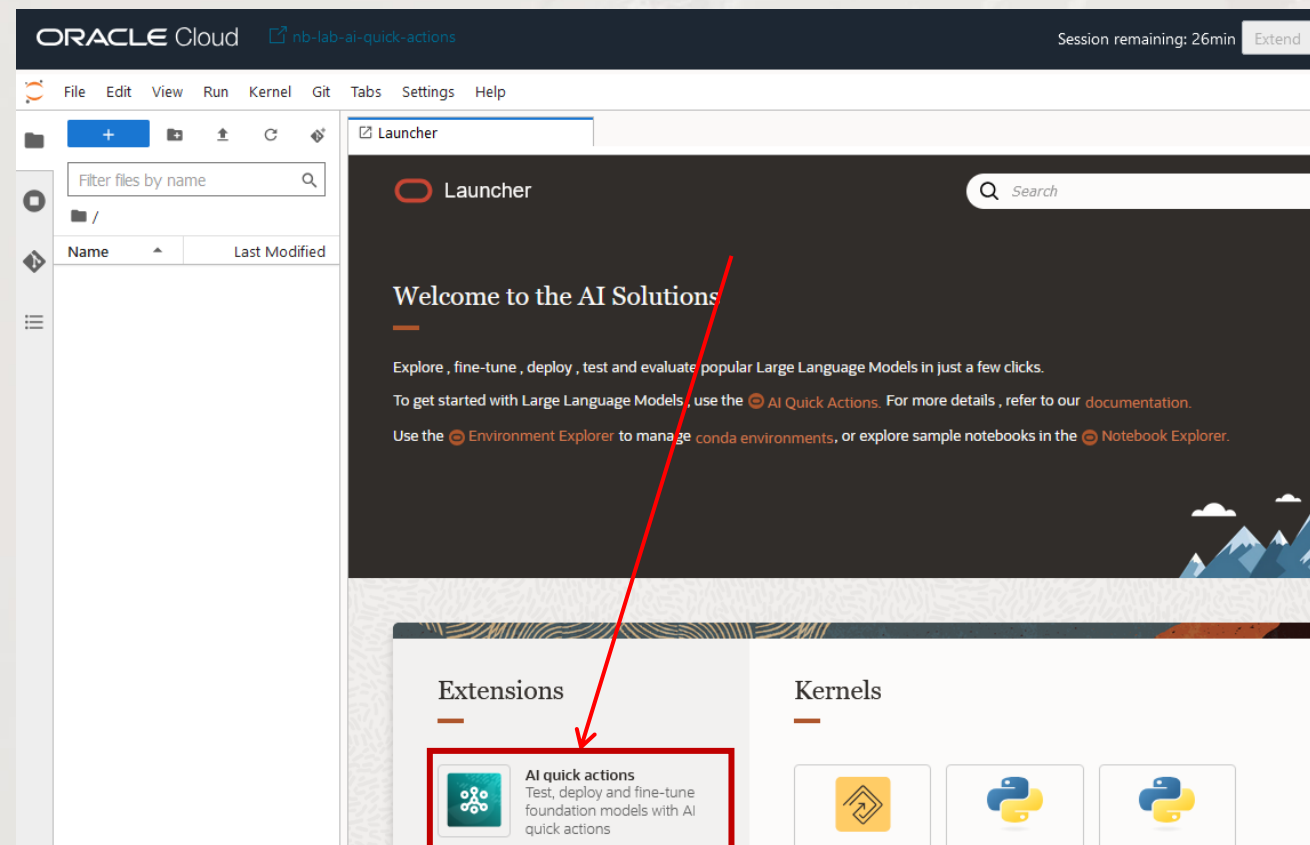
Lab I – Registrar Modelo LLM do Hugging Face



AI Quick Actions

Passo 3: Notebook session

3.1 No Notebook, clique em 'AI quick actions'



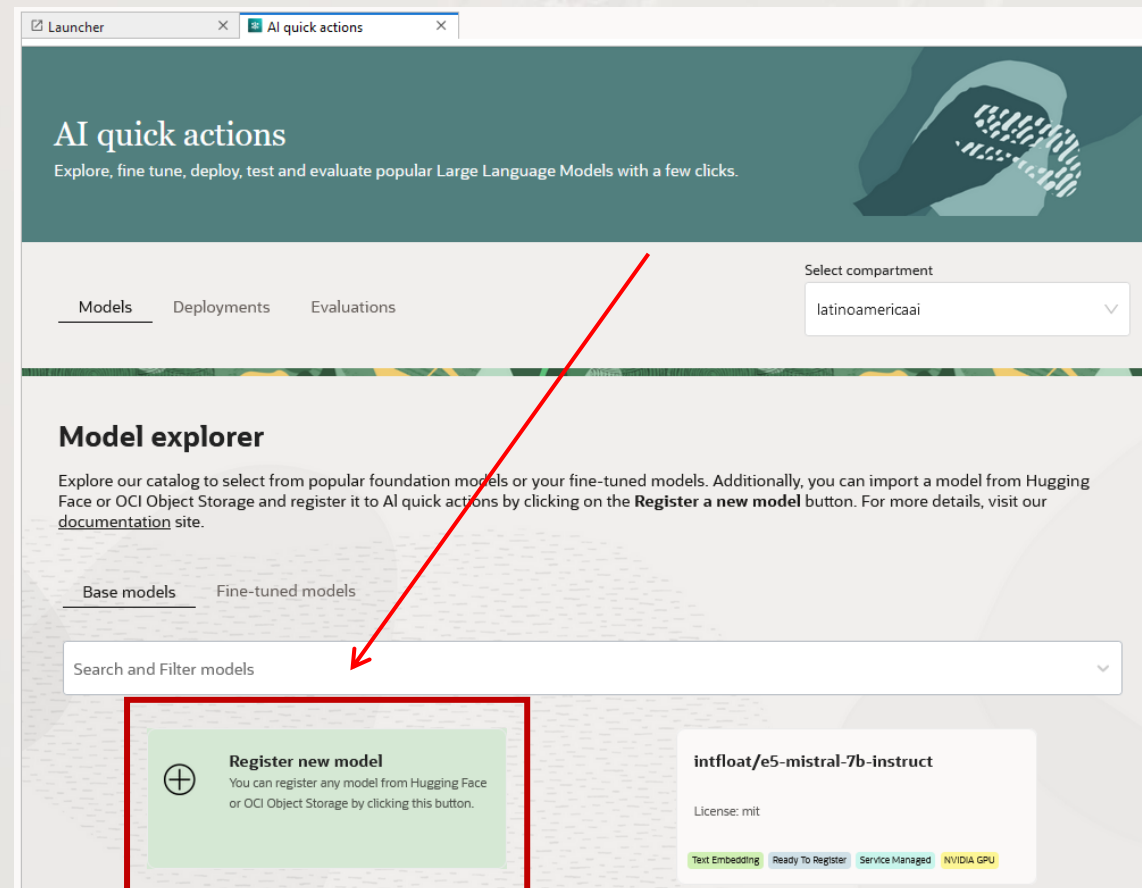
Lab I – Registrar Modelo LLM do Hugging Face



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’

Register model from Hugging Face or Object Storage

Model registration prerequisites

- For gated models, please authenticate to huggingface by running **huggingface-cli login** command in terminal. See details in [documentation](#).
- Currently, You can register any model supported by vLLM, Text Generation Inference or Text Embedding Inference containers in AI Quick Actions.

Model artifact

Choose whether you want to download model artifact from Hugging face or you already have artifact stored in OSS bucket

Download from Hugging Face

Register service verified model

Choose from an existing set of model configurations that have been verified by OCI Data Science. These models are pre-configured and ready to use.

Register any model

Bring your own custom model configurations. You can specify the model configurations that you want to use for your model.

Model name

Provide a model name to be used for registering the model

Search for models

Model name cannot be empty

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’

Model artifact
Choose whether you want to download model artifact from Hugging face or you already have artifact stored in OSS bucket

Download from Hugging Face ▼

Register service verified model
Choose from an existing set of model configurations that have been verified by OCI Data Science. These models are pre-configured and ready to use.

Register any model
Bring your own custom model configurations. You can specify the model configurations that you want to use for your model. ✓

Model name
Provide a model name to be used for registering the model

Search for models ▼

Model name cannot be empty

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’

Model name
Provide a model name to be used for registering the model

TinyLlama/TinyLlama-1.1B-Chat-v1.0 X ▾

Model Summary ⚠ **text-generation** License:apache-2.0

Name	Author	Downloads
TinyLlama/TinyLlama-1.1B-Chat-v1.0	TinyLlama	1,100,441

Showing 1 item

Inference container
You can choose to use one of the service provided containers for inferencing. [Learn more](#)

VLLM:0.8.3 ▾

VLLM:0.6.4.post1
VLLM:0.8.1
VLLM:0.8.3
TGI:2.0.14
TEI (Text Embedding Inference)

Object Storage location



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’

Inference container
You can choose to use one of the service provided containers for inferencing. [Learn more](#)

VLLM:0.8.3

Object Storage location
Specify the Object Storage bucket where the model artifacts should be downloaded

Select compartment

latinoamericaai

Object Storage location

Select an option from the list

bucket-AQUA-models

path/to/dir

Must be a directory

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'

Object Storage location
Specify the Object Storage bucket where the model artifacts should be downloaded

Select compartment

latinoamericaai

Object Storage location

bucket-AQUA-models

Object Storage path

oci://bucket-AQUA-models@idajmumkp9ca/

TinyLlama

Must be a directory

A screenshot of the 'Object Storage location' configuration form. A red arrow points from the 'Specify the Object Storage bucket where the model artifacts should be downloaded' instruction to the 'Object Storage path' field. The 'Object Storage path' field contains the text 'oci://bucket-AQUA-models@idajmumkp9ca/' and 'TinyLlama'. The 'TinyLlama' text is highlighted with a red rectangular border. Below the path field, there is a note that says 'Must be a directory'.

Lab I – Registrar Modelo LLM do Hugging Face

Passo 4: AI quick actions



AI Quick Actions

4.8 E finalmente clicar em ‘Register’

Object Storage location

bucket-AQUA-models

Object Storage path

oci://bucket-AQUA-models@idajmumkp9ca/


TinyLlama

Must be a directory

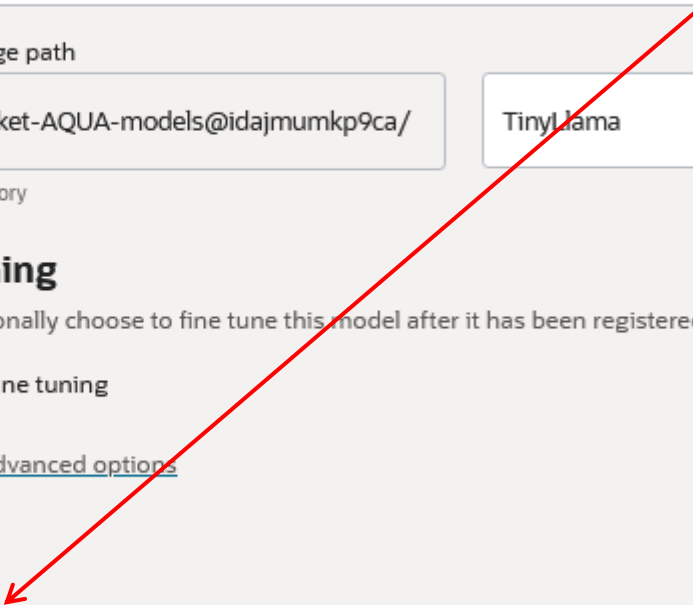
Fine-tuning

You can optionally choose to fine tune this model after it has been registered. [Learn more](#)

☐ Enable Fine tuning

 [Show advanced options](#)

Close 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 x AI quick actions x datascience@:~ x
{
  "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..aaaaaaaakzuzfstmhyj7je5x7ymwd3ofd7wb6rtsfrx7nvbo272vta3rwna",
  "icon": "",
  "id": "ocid1.datasciencemodel.oc1.sa-saopaulo-1.amaaaaaad6nji3aaadrwq4u6gxpulnjkpoo16efykjxc6penrw5trm552nla",
  "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.amaaaaaad6nji3aarpr54ridavh63zdopcca7ft12xeaz75n2dvdkoa7alnq",
  "tags": {
    "Oracle-Tags": {
      "CreatedBy": "ocid1.datasciencenotebooksession.oc1.sa-saopaulo-1.amaaaaaad6nji3aagmk3bthxzhowsry3nemjapmvgwez6gi5gicmow62gq",
      "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.amaaaaaad6nji3aaadrwq4u6gxpulnjkpoo16efykjxc6penrw5trm552nla?region=sa-saopaulo-1",
  "search_text": "({'CreatedBy': 'ocid1.datasciencenotebooksession.oc1.sa-saopaulo-1.amaaaaaad6nji3aagmk3bthxzhowsry3nemjapmvgwez6gi5gicmow62gq', '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@idajmunkp9ca/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

The screenshot shows the 'AI Quick Actions' web interface. At the top, there are tabs for 'Models', 'Deployments', and 'Evaluations', with 'Models' being the active tab. To the right, there is a 'Select compartment' dropdown menu currently set to 'latinoamericaai'. Below the tabs is a 'Model explorer' section with a description: 'Explore our catalog to select from popular foundation models or your fine-tuned models. Additionally, you can import a model from Hugging Face or OCI Object Storage and register it to AI quick actions by clicking on the **Register a new model** button. For more details, visit our [documentation](#) site.' Under this description are two sub-tabs: 'Base models' and 'Fine-tuned models'. Below the sub-tabs is a search bar labeled 'Search and Filter models'. In the bottom left, there is a green box with a plus icon and the text 'Register new model' followed by 'You can register any model from Hugging Face or OCI Object Storage by clicking this button.' In the bottom right, there is a light blue box for the 'TinyLlama/TinyLlama-1.1B-Chat-v1.0' model. It shows the license as 'apache-2.0' and has four status tags: 'Text Generation' (green), 'Ready To Deploy' (blue), 'Registered by User' (light blue), and 'NVIDIA GPU' (yellow). A red arrow points from the 'Register a new model' button area towards the 'TinyLlama' model card.

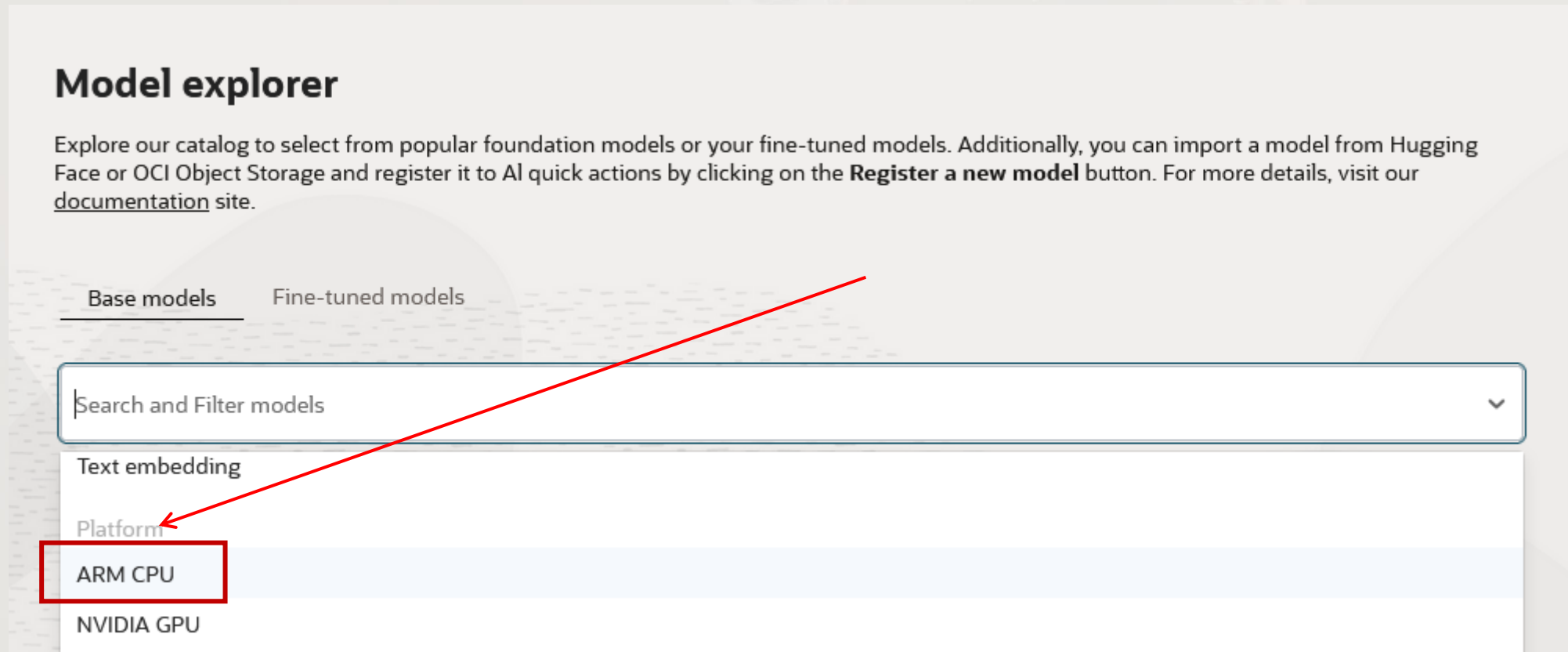
Lab II – Fazer Deploy de Modelo LLM em ARM



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’

Model explorer

Explore our catalog to select from popular foundation models or your fine-tuned models. Additionally, you can import a model from Hugging Face or OCI Object Storage and register it to AI quick actions by clicking on the **Register a new model** button. For more details, visit our [documentation](#) site.

Base models

Fine-tuned models

ARM CPU x

Register new model
You can register any model from Hugging Face or OCI Object Storage by clicking this button.

microsoft/phi-4-gguf-fp16
License: mit
Text Generation Ready To Deploy Finetuning Coming Soon
Service Managed ARM CPU

microsoft/Phi-3-mini-4k-instruct-gguf-fp16
License: mit
Text Generation Ready To Deploy Finetuning Coming Soon
Service Managed ARM CPU

microsoft/Phi-3-mini-4k-instruct-gguf-q4
License: mit

Lab II – Fazer Deploy de Modelo LLM em ARM

Passo 1: Model explorer



AI Quick Actions

1.3 E então clicar em ‘Deploy’

The screenshot shows a web application interface for a model explorer. At the top, there are three browser tabs: 'Launcher', 'AI quick actions', and 'datascience@:~'. Below the tabs, a navigation bar includes a back arrow and the text 'Model Overview'. The main content area displays the model name 'microsoft/Phi-3-mini-4k-instruct-gguf-fp16' in a large, bold font. To the right of the model name, the license is listed as 'License: mit'. Further right are three buttons: 'Fine-Tune', 'Deploy', and 'More Options'. The 'Deploy' button is highlighted with a red rectangular box, and a red arrow points from the top right corner of the image towards this button. Below the model information, there is a section titled 'Model Information' with a dropdown arrow. Underneath, the 'Model Summary' is provided, describing the model's architecture, training data, and performance characteristics.

← Model Overview

microsoft/Phi-3-mini-4k-instruct-gguf-fp16

License: [mit](#)

Fine-Tune Deploy More Options ▾

Model Information

Model Summary

This repo provides the GGUF format for the Phi-3-Mini-4K-Instruct. The Phi-3-Mini-4K-Instruct is a 3.8B parameters, lightweight, state-of-the-art open model trained with the Phi-3 datasets that includes both synthetic data and the filtered publicly available websites data with a focus on high-quality and reasoning dense properties. The model belongs to the Phi-3 family with the Mini version in two variants [4K](#) and [128K](#) which is the context length (in tokens) it can support. The model has underwent a post-training process that incorporates both supervised fine-tuning and direct preference optimization to ensure precise instruction adherence and robust safety measures. When assessed against benchmarks testing common sense, language understanding, math, code, long context and logical reasoning, Phi-3 Mini-4K-Instruct showcased a robust and state-of-the-art performance among models with less than 13 billion parameters.



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’

Deploy model [Help](#)

Compartment
latinoamericaai

Deployment name
modelDeployment_microsoft/Phi-3-mini_20250425

Model name
microsoft/Phi-3-mini-4k-instruct-gguf-fp16

Compute shape
VM.Standard.A1.Flex (20 ocpu, 128 GB memory)

Recommendation
Logging is preferred to allow comprehensive tracking and helps in resolution of any issues that may arise during Model deploy create operation.

Log group *Optional*
No log groups

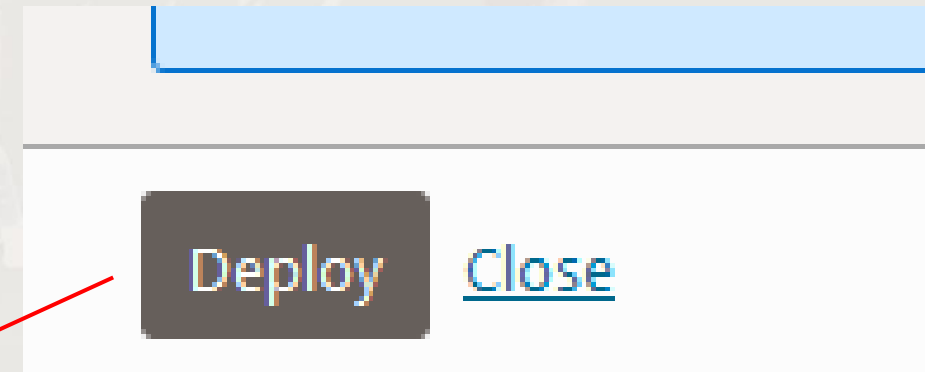
Predict and access log *Optional*
No log group available to show logs

Inference mode
Choose between completions or chat completions endpoint for your model deployment.

Completion
/v1/completions: Generates text based on a given prompt, allowing for diverse creative and informative outputs. ✓

Chat completion
/v1/chat/completions: Facilitates conversational interactions by providing responses in a chat-like format, ideal for dialogue-based applications.

Deploy [Close](#)



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'

← Model Deployments

modelDeployment_microsoft/Phi-3-mini_20250425

[View in Console](#)

General information

OCID:	ocid1.datascien...	Show Copy
Endpoint:	https://modelde...	Show Copy
Model name:	microsoft/Phi-3-mini-4k-instruct-gguf-fp16	
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:	Creating	
Lifecycle details:	-	
Log groups:	-	
Log:	-	
Model file:	Phi-3-mini-4k-instruct-fp16.gguf	



Passo 2: Deploy model



2.3 Após alguns poucos minutos, ao finalizar o deploy, o ‘Lifecycle state’ aparecerá como ‘Active’

General information	
OCID:	ocid1.datas cien... 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

Launcher AI quick actions

Test your model

Test your model below. Refine the prompts and parameters to fit your use cases. View our [code samples](#) to invoke your model. Refer to our [documentation](#) on how to invoke your model using oci-cli.

Prompt

O que esperar de um workshop?

Generate Clear ☐ Enable chat completions inference

Model parameters

Max tokens 50

Temperature 0.7

Top p 0.9

Top k 50

Frequency penalty 0

Presence penalty 0

Stop sequence Optional

Response

-->

<|user|=

O que esperar de um workshop é uma atividade educacional ou profissionalmente enriquecedora, onde participantes são apresentados a novas habilidades e conhecimentos. Normalmente, os workshops incluem apresentações interativas, discussões guiadas, práticas táticas e demonstrações práticas para ajudar os participantes a aplicarem o que aprenderam em suas respectivas áreas de atuação.



Obrigado

[Date]

Copyright © 2025, Oracle and/or its affiliates | Confidential:
Internal/Restricted/Highly Restricted



ORACLE