Foundation Model Fine-Tuning Workshop

Michael Lin

Sr. Solutions Architect
Amazon Web Services

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Labs

- Bedrock Console: text generation with Titan Text
- Bedrock Console: news summary with Command Light
- Bedrock SDK: news summary with Llama 2
- Bedrock SDK: image generation with Titan Image Generator
- SageMaker JumpStart console: text generation with Mistral
- SageMaker JumpStart SDK: dialog summary with LLama 3

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

https://ppt.cc/f5NCLx

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Workshop Setup

- Workshop Access
- Bedrock Model Access
- SageMaker Studio Access

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Workshop Studio

catalog.us-east-1.prod.workshops.aws/sign-in?redirect=%2Fjoin%3Faccess-code%3Db0e6-0e4b8c-89

Incognito

aws workshop studio

Workshop Studio > Sign in

Sign in

Choose a preferred sign-in method

Email one-time password (OTP)

Enter your personal or corporate email to receive a one-time password

AWS Builder ID

Login with AWS Builder ID, a new personal profile for builders

Amazon employee

Login with your Amazon Corporate account. Only for Amazon Employees.

One-time email passcode

login.us-east-1.auth.skillbuilder.aws/otp/input/response?redirect_uri=https%3A%2F%2Fgandalf-prod.auth.us-east-1.amazoncognito.com%2Foauth2%2Fidpr...

Incognito

aws training and certification

English

One-time email passcode

Send a passcode to the email below.

Email

michael.tw.lin@

Back

Send passcode

Get help signing in

Verify one-time email passco

login.us-east-1.auth.skillbuilder.aws/otp/challenge?redirect_uri=https%3A%2F%2Fgandalf-prod.auth.us-east-1.amazoncognito.com%2Foauth2%2Fidprespo...

Incognito

aws training and certification

English

One-time email passcode

We sent a passcode to michael.tw.lin@gmail.com. You should receive it within 5 minutes.

Passcode (9-digit) Resend passcode

529717102

Back Sign in

Get help signing in

Step 1
Enter event access code

Step 2
Review and join

Review and join

Event details

Name	Start time	Duration	Level
bedrock-finetune-20240712	7/12/2024 10:54 AM	72 hours	300

Description
bedrock-finetune-20240712

Terms and Conditions

Read and accept before joining the event

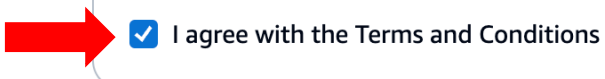
Read and accept before joining the event:

1. By using AWS Workshop Studio for the relevant event, You agree to the AWS Event Terms and Conditions, the AWS Responsible AI Policy, and the AWS Acceptable Use Policy.
2. If You are under 18 years old, you may participate in the relevant event using AWS Workshop Studio: (a) if You are at least the minimum age below based on the country or region in which You reside, and (b) with the involvement of a parent, guardian, or educator.

Country or region	Minimum age
All countries or regions not listed below (including the United States, Brazil, the United Kingdom, and India)	13
Canada, China, Republic of Korea (South Korea)	14
Australia	15

© 2008 - 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

3. You acknowledge and agree that You are using an AWS-owned account that You will only be able to access during the relevant event. You have no ownership rights over this AWS-owned account.
4. During the relevant event, while using this AWS-owned account, You will not use, import, input, or introduce any data, dataset, or other material that contains personal data, financial information, or any other data or materials that may be subject to laws and regulations (such as the General Data Protection Regulation or The Health Insurance Portability and Accountability Act of 1996).
5. If You find residual resources or materials in this AWS-owned account, You will notify your Event Operator immediately.
6. AWS, its affiliates, and any entities or persons acting on AWS's behalf reserves the right to terminate this AWS-owned account and to delete its contents at any time, without any notice to You.
7. During the relevant event, while using this AWS-owned account, You will not process or run any operation on any data other than test datasets or lab materials that have been approved by AWS.
8. You will not copy, import, export or otherwise create derivative works of materials provided by AWS for use outside of the relevant event.
9. AWS, its affiliates, and any entities or persons acting on AWS's behalf have no obligation to enable the transmission of Your materials through AWS Workshop Studio, and may, in their discretion, edit, block, refuse to post, or remove Your materials at any time, without notice to You.
10. If You access and use a service and/or third-party models that have their own terms during the relevant event, while in the AWS-owned account, You agree to review those terms and comply with them during the event.
11. If You are an AWS Partner using AWS Workshop Studio as part of Your participation in the AWS Partner Network Program, Your use of AWS Workshop Studio is governed by these terms, the AWS Partner Network Terms and Conditions, and the AWS Customer Agreement or other agreement with us governing your use of AWS Services.
12. Your use of AWS Workshop Studio will comply with these terms and all applicable laws. If You fail to comply with any of these terms, Your access to AWS Workshop Studio may be immediately terminated, without notice to You.



Cancel Previous Join event

bedrock-finetune-20240712

catalog.us-east-1.prod.workshops.aws/event/dashboard/en-US

aws workshop studio

michael_tw_lin

bedrock-finetune-20240712

- Amazon Bedrock Workshop
- Prerequisites
- Prompt Engineering
- Text Generation
- Knowledge Bases and RAG
- Model Customization
- Image and MultiModal applications
- Agents
- Open Source With Bedrock

AWS account access

Open AWS console (us-west-2)

Get AWS CLI credentials

Exit event

Event ends in 1 hour 27 minutes.

Event dashboard > Amazon Bedrock Workshop

bedrock-finetune-20240712

Event information

Start time	Duration	Accessible regions
7/12/2024 10:54 AM	72 hours	us-west-2, us-east-1

Description
bedrock-finetune-20240712

Workshop

Get started

Title	Complexity level	AWS services	Topics
Amazon Bedrock Workshop	300	Amazon Bedrock	Machine Learning (ML/AI), Generative AI

Description
Amazon Bedrock workshop with main design patterns

Event Outputs (0)

bedrock-finetune-20240712 | Console Home | Console Hom

us-west-2.console.aws.amazon.com/console/home?region=us-west-2

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

Console Home Info

Reset to default layout | + Add widgets

Recently visited Info

- Amazon SageMaker
- CloudFormation
- IAM
- S3
- Amazon Bedrock
- Cloud9
- RDS
- EC2

View all services

Applications (0) Info

Region: US West (Oregon)

Create application

us-west-2 (Current Region) | Find applications

1

Name	Description	Region	Originating account

No applications

Get started by creating an application.

Create application

Go to myApplications

Welcome to AWS

Getting started with AWS

AWS Health Info

Cost and usage Info

Current month costs

$70.51

Cost ($)

80

Workshop Setup

- Workshop Access
- Bedrock Model Access
- SageMaker Studio Access

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Consol

Recen

Amaz
IAM
S3
Amaz
Cloud
RDS
EC2

Welco

get the most out of AWS.

ault layout
+ Add widgets
Create application
1
on
Originating account
plication.
40

Search results for 'bedrock'

Services (1)
Resources New
Documentation (2,575)
Knowledge Articles (12)
Marketplace (368)
Blogs (217)
Events (1)

Services

Amazon Bedrock
The easiest way to build and scale generative AI applications with foundation models (F...

Resources / for a focused search

Introducing resource search
To search for resources, Resource Explorer must be active in at least one AWS Region and you must have permission to use the default view in the account. Learn more
Dismiss

Documentation

See all 2,575 results

Amazon Bedrock
User Guide

Deleting an Bedrock Studio project
User Guide

Getting started with Amazon Bedrock Studio
User Guide

Deleting apps and components from an Amazon Bedrock Studio

© 2024, Amazon Web Services, Inc. or its affiliates.

bedrock-finetune-20240712 | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

Machine Learning

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)

Try Bedrock

Get started

Overview

Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.

Benefits

bedrock-finetune-20240712 | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
  - Prompt management Preview
  - Prompt flows Preview
- **Assessment & deployment**

Machine Learning

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)

Try Bedrock

Get started

Overview

Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.

Benefits

bedrock-finetune-20240712 | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

- **Foundation models**
  - Base models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
  - Prompt management Preview
  - Prompt flows Preview
- **Assessment & deployment**
  - Model Evaluation

- Model access
- Bedrock Studio Preview
- Settings
- User guide

Machine Learning

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)

Try Bedrock

Get started

Overview

Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.

Benefits

bedrock-finetune-20240712 | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/modelaccess

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
  - Prompt management Preview
  - Prompt flows Preview

Amazon Bedrock > Model access

What is Model access?

To use Bedrock, account users with the correct IAM Permissions must enable access to available Bedrock foundation models (FMs). View all Bedrock Model Terms for Bedrock FMs.

Modify model access

Visit Amazon Bedrock Quotas for a quick guide to the default quotas and limits that apply to Amazon Bedrock.

Base models (29)

Collapse all

Not seeing a model you're interested in? Check out all supported models by region here.

Find model

Group by provider

Models	Access status	Modality	EULA
AI21 Labs (2)	0/2 access granted		
Jurassic-2 Ultra	Available to request	Text	EULA
Jurassic-2 Mid	Available to request	Text	EULA
Amazon (6)	6/6 access granted		
Titan Embeddings G1 - Text	Access granted	Embedding	EULA
Titan Text G1 - Lite	Access granted	Text	EULA

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/modelaccess

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

Amazon Bedrock > Model access > Manage model access

Step 1
Edit model access

Step 2
Review and submit

Edit model access

Base models (16/29)

Collapse all

Not seeing a model you're interested in? Check out all supported models by region here.

Find model

Group by provider

Models	Access status	Modality	EULA
AI21 Labs (2)	0/2 access granted		
Jurassic-2 Ultra	Available to request	Text	EULA
Jurassic-2 Mid	Available to request	Text	EULA
Amazon (6)	6/6 access granted		
Titan Embeddings G1 - Text	Access granted	Embedding	EULA
Titan Text G1 - Lite	Access granted	Text	EULA
Titan Text G1 - Express	Access granted	Text	EULA
Titan Image Generator G1	Access granted	Image	EULA
Titan Multimodal Embeddings G1	Access granted	Embedding	EULA
Titan Text Embeddings V2	Access granted	Embedding	EULA
Anthropic (5)	3/5 access granted		

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/modelaccess

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

☑	Embed Multilingual	Available to request	Embedding	EULA
☑	Command	Available to request	Text	EULA
☑	Command Light	Available to request	Text	EULA
☑	▼ Meta (6)	2/6 access granted		
☑	Llama 3 8B Instruct	Access granted	Text	EULA
☑	Llama 3 70B Instruct	Access granted	Text	EULA
☐	Llama 2 Chat 13B	Unavailable	Text	EULA
☐	Llama 2 Chat 70B	Unavailable	Text	EULA
☐	Llama 2 13B	Unavailable	Text	EULA
☐	Llama 2 70B	Unavailable	Text	EULA
☐	▼ Mistral AI (3)	3/3 access granted		
☐	Mistral 7B Instruct	Access granted	Text	EULA
☐	Mixtral 8x7B Instruct	Access granted	Text	EULA
☐	Mistral Large	Access granted	Text	EULA
☑	▼ Stability AI (1)	1/1 access granted		
☑	SDXL 1.0	Access granted	Image	EULA

Cancel | Next

Step 1
Edit model access

Step 2
Review and submit

Review and submit

Step 1: Edit model access

Edit

Model access modifications (10)

Models	Modifications
Jurassic-2 Ultra	Request access
Jurassic-2 Mid	Request access
Claude 3 Opus	Request access
Claude 3 Haiku	Request access
Command R+	Request access
Command R	Request access
Embed English	Request access
Embed Multilingual	Request access
Command	Request access
Command Light	Request access

Model access modifications (10)

Models	Modifications
Jurassic-2 Mid	Request access
Claude 3 Opus	Request access
Claude 3 Haiku	Request access
Command R+	Request access
Command R	Request access
Embed English	Request access
Embed Multilingual	Request access
Command	Request access
Command Light	Request access

Terms

By selecting Submit, you are requesting access to the selected third party models through the AWS Marketplace. By doing so, you agree to the seller's pricing terms and End User License Agreements (EULA), and the Bedrock Service Terms. You also agree and acknowledge that AWS may share information about this transaction with the respective sellers, in accordance with the AWS Privacy Notice.

AWS will issue invoices and collect payments from you on behalf of the seller through your AWS account. Your use of AWS services is subject to the AWS Customer Agreement or other agreements with AWS governing your use of such services.

Cancel Previous Submit

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Model access modifications (10)

Models	Modifications
Jurassic-2 Mid	Request access
Claude 3 Opus	Request access
Claude 3 Haiku	Request access
Command R+	Request access
Command R	Request access
Embed English	Request access
Embed Multilingual	Request access
Command	Request access
Command Light	Request access

Terms

By selecting Submit, you are requesting access to the selected third party models through the AWS Marketplace. By doing so, you agree to the seller's pricing terms and End User License Agreements (EULA), and the Bedrock Service Terms. You also agree and acknowledge that AWS may share information about this transaction with the respective sellers, in accordance with the AWS Privacy Notice.

AWS will issue invoices and collect payments from you on behalf of the seller through your AWS account. Your use of AWS services is subject to the AWS Customer Agreement or other agreements with AWS governing your use of such services.

Cancel Previous Submit

© 2024, Amazon Web Services, Inc. or its affiliates.

bedrock-finetune-20240712 | Amazon Bedrock | us-west-2

us-west-2.console.aws.amazon.com/bedrock/home?region=us-west-2#/modelaccess

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
  - Prompt management Preview
  - Prompt flows Preview

Access request for 10 models failed

- Jurassic-2 Ultra - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action
- Jurassic-2 Mid - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action
- Claude 3 Opus - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action
- Claude 3 Haiku - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action
- Command R+ - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action
- Command R - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action
- Embed English - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action
- Embed Multilingual - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action
- Command - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action
- Command Light - User: arn:aws:sts::653405491536:assumed-role/WSParticipantRole/Participant is not authorized to perform: aws-marketplace:Subscribe on resource: * because no identity-based policy allows the aws-marketplace:Subscribe action

Amazon Bedrock > Model access

What is Model access?

To use Bedrock, account users with the correct IAM Permissions must enable access to available Bedrock foundation models (FMs). View all Bedrock Model Terms for Bedrock FMs.

Modify model access

Visit Amazon Bedrock Quotas for a quick guide to the default quotas and limits that apply to Amazon

Workshop Setup

- Workshop Access
- Bedrock Model Access
- SageMaker Studio Access

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws Services sagemaker Canvas Oregon WSParticipantRole/Participant @ 6534-0549-1536

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Prev
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
  - Prompt management
  - Prompt flows Preview
- Assessment & depl
  - Model Evaluation

Search results for 'sagemaker'

Services (1)
Features (5)
Resources New
Documentation (11,787)
Knowledge Articles (48)
Marketplace (572)
Blogs (1,323)
Events (61)
Tutorials (23)

Services

Amazon SageMaker ☆
Build, Train, and Deploy Machine Learning Models

Features

See all 5 results ▸

SageMaker Studio
Amazon SageMaker feature

Notebooks
IoT Analytics feature

Autopilot
Amazon SageMaker feature

SageMaker Canvas
Amazon SageMaker feature

Resources / for a focused search

Introducing resource search
To search for resources, Resource Explorer must be active in at least one AWS Region and you must have permission to use the default view in the account. Learn more

examples Compare mode

等,適合不同的口味喜好。

Run

Define metric criteria

Overall summary

Define metric criteria

© 2024, Amazon Web Services, Inc. or its affiliates.

bedrock-finetune-20240712 | Amazon SageMaker | us-west | taipei 101 - Google Search

us-west-2.console.aws.amazon.com/sagemaker/home?region=us-west-2#/studio

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

Amazon SageMaker

- Getting started
- **Applications and IDEs**
  - Studio
  - Canvas
  - RStudio
  - TensorBoard
  - Notebooks
- **Admin configurations**
  - **Domains**
  - Role manager
  - Images
  - Lifecycle configurations
- SageMaker dashboard
- Search
- **JumpStart**

Automatic Domain Migration
If you have existing Studio Classic domains, they will be automatically migrated to the new Studio experience starting June 2024. The new Studio experience provides better speed, efficiency, and productivity.

Learn more

Amazon SageMaker > Domains

Domains Info

In SageMaker, a domain is an environment for your team to access SageMaker resources. A domain consists of a list of authorized users and users within a domain can share notebook files and other artifacts with each other. One account can have either one or multiple domains.

Domains (1) Info

View | Create domain

Find domain name

1

Name	Id	Status	Created on	Modified on
amazon-bedrock-workshop	d-noqxdjm0rbmm	InService	Jul 12, 2024 02:28 UTC	Jul 12, 2024 02:31 UTC

bedrock-finetune-20240712 | SageMaker Studio | Amazon S | taipei 101 - Google Search

us-west-2.console.aws.amazon.com/sagemaker/home?region=us-west-2#/studio-landing

aws | Services | Search [Option+S] | Oregon | WSParticipantRole/Participant @ 6534-0549-1536

Amazon SageMaker

- Getting started
- **Applications and IDEs**
  - Studio
  - Canvas
  - RStudio
  - TensorBoard
  - Notebooks
- **Admin configurations**
  - Domains
  - Role manager
  - Images
  - Lifecycle configurations
- SageMaker dashboard
- Search
- **JumpStart**
  - Foundation models

Automatic Domain Migration
If you have existing Studio Classic domains, they will be automatically migrated to the new Studio experience starting June 2024. The new Studio experience provides better speed, efficiency, and productivity.

Learn more

Amazon SageMaker

SageMaker Studio

The first fully integrated development environment (IDE) for machine learning.

Get Started

Select user profile

sagemakeruser

Open Studio

How it works

What is Studio?

Amazon SageMaker Studio provides a single, web-based visual interface where you can perform all ML development steps, improving data science team

Pricing (US)

With Amazon SageMaker Studio, you pay only for

SageMaker Studio

Provide feedback

Applications (6)

Jupyter... RStudio Canvas

Code E... Studio ... MLflow

Home

Running instances

Data

Auto ML

Experiments

Jobs

Pipelines

Models

JumpStart

Deployments

Collapse Menu

Home

Launch workflows, manage your applications and spaces, and view getting started materials

Onboarding

To get the most ou

Take the t

Quick tour high
the new experie
to be productiv

Take the tour

Not ready to use th

and notebooks

work into the new experience. Transfer notebooks, data
rtifacts so they remain accessible as you adopt the new

Overview Getti

Overview

Start a new ML wor

Welcome to the new

SageMaker Studio

We've built a new experience to empower you and your work.

Want to take a quick tour?

Skip Tour for now

Take a quick tour

Are you an existing Studio Classic user and looking to migrate your data and notebooks? Click here to learn how.

jupyter

SageMaker Studio

Provide feedback

Applications (6)

Jupyter... RStudio Canvas

Code E... Studio ... MLflow

Home

Running instances

Data

Auto ML

Experiments

Jobs

Pipelines

Models

JumpStart

Deployments

Collapse Menu

Home

Launch workflows, manage your applications and spaces, and view getting started materials

Onboarding plan

To get the most out of the new Studio experience, explore the onboarding steps below.

Take the tour

Quick tour highlights where you can find key features and how to navigate the new experience. See what's new and where to locate the tools you need to be productive.

Take the tour >

Migrate data and notebooks

Bring your previous work into the new experience. Transfer notebooks, data sources, and other artifacts so they remain accessible as you adopt the new environment.

Learn more

Not ready to use the new experience? Revert to Studio Classic experience in domain settings. Learn more

Overview | Getting started

Overview

Start a new ML workflow or jump back into your workflow

jupyter

Applications (6)

JupyterLab | RStudio | Canvas

Code Editor | Studio Cl... | MLflow

Home

Running instances

Data

Auto ML

Experiments

Jobs

Pipelines

Models

JumpStart

Deployments

Collapse Menu

SageMaker Studio Classic

+ Create Studio Classic space

About

Studio Classic is a legacy IDE that allows you to access the previous iteration of SageMaker Studio from within the new Studio experience.

See features | Learn more about Studio Classic

Search...

Filter spaces: Running

Name	Application	Status	Type	Last modified	Action
sagemakeruser	Studio Classic	Stopped	Private	0 seconds ago	Run

1 results Results are cached Refresh Go to page 1 Page 1 of 1

Introducing spaces New

JupyterLab and Code Editor now come with durable instances that allow for faster startup, privacy options, and configurable storage.

Learn more

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Applications (6)

Jupyter... RStudio Canvas

Code E... Studio ... MLflow

Studio Classic

Home

Running instances

Data

Auto ML

Experiments

Jobs

Pipelines

Models

JumpStart

Deployments

Collapse Menu

SageMaker Studio Classic

+ Create Studio Classic space

About

Studio Classic is a legacy IDE that allows you to access the previous iteration of SageMaker Studio from within the new Studio experience.

See features | Learn more about Studio Classic

Search...

Filter spaces: Running

Name	Application	Status	Type	Last modified	Action
sagemakeruser	Studio Classic	Running	Private	15 seconds ago	Stop Open

1 results Results are cached Refresh Go to page 1 Page 1 of 1

Introducing spaces New

JupyterLab and Code Editor now come with durable instances that allow for faster startup, privacy options, and configurable storage.

Learn more

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Introducing the new SageMaker Studio! Launch ML apps, access ML resources in one location, and share apps with users in your domain. Learn More

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

sagemakeruser / Personal Studio

Filter files by name

/

Name

amazon-bedrock-workshop

Launcher

Launcher

Interested in launching different applications?

Introducing a centralized hub for launching all your favorite apps including JupyterLab, Code Editor, and RStudio.

Learn more

Notebooks and compute resources

/

Create notebooks, code console, image terminal with custom environment in the active folder.

Image	Kernel	Instance	Start-up script
Data Science 3.0	Python 3	ml.t3.medium	No script

Change environment

Create notebook

Open code console

Open image terminal

Learn more about SageMaker images and how to customize compute environment

Utilities and files

Simple

Cookie Preferences

Launcher

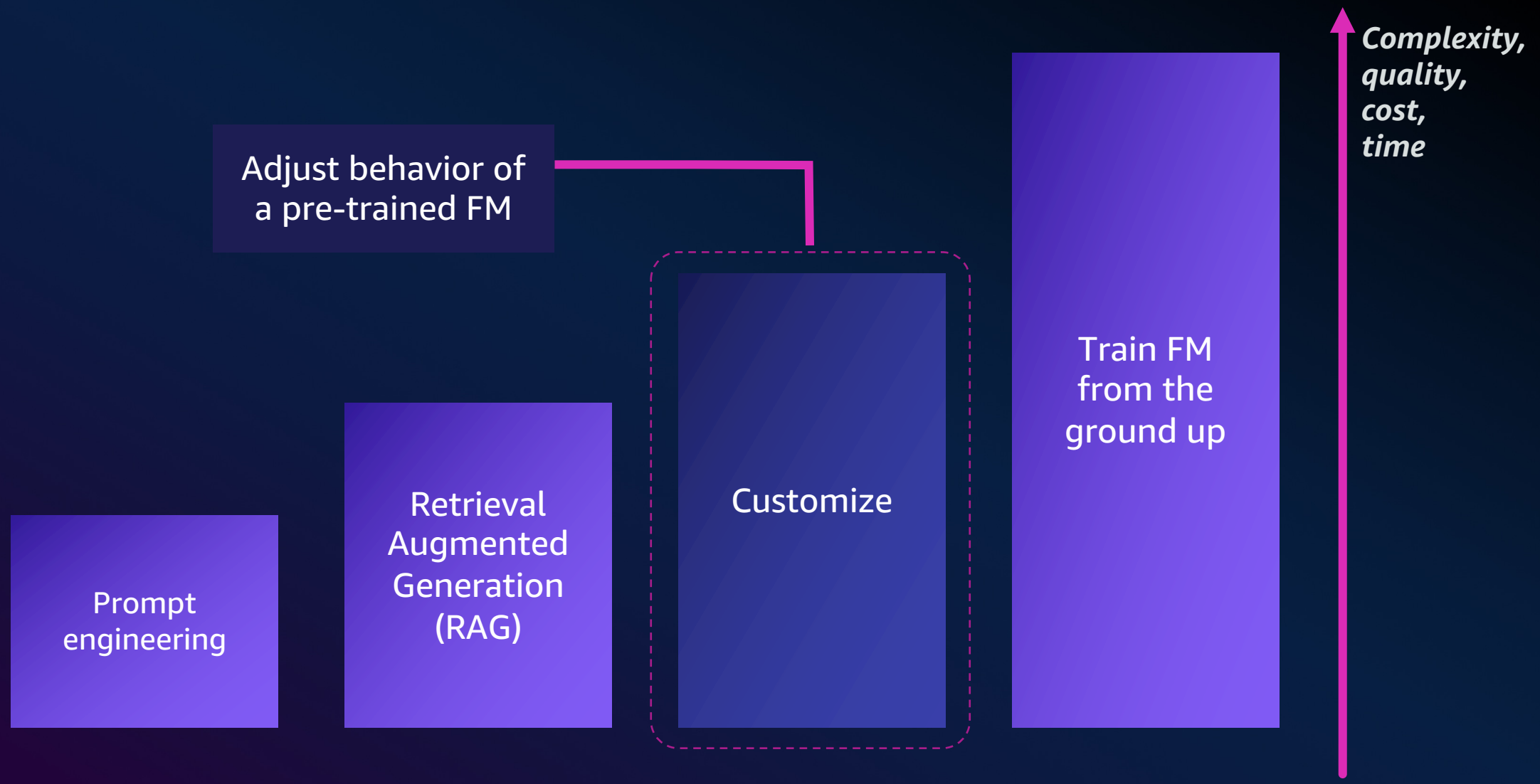
Customize FMs for generative AI applications with Amazon Bedrock

Michael Lin

Sr. Solutions Architect
Amazon Web Services

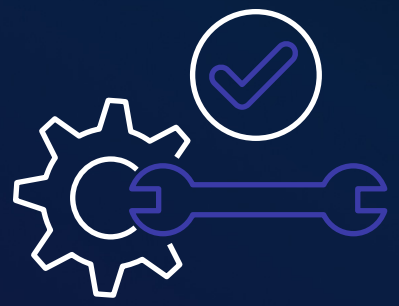
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Common approaches for customizing FMs



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customizing model responses for your business



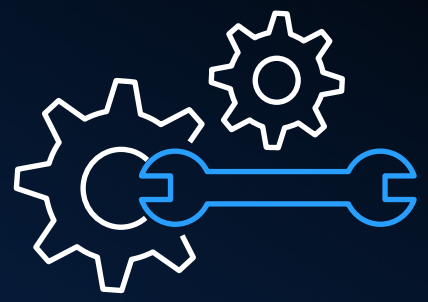
Fine-tuning

PURPOSE

Maximizing accuracy for ***specific tasks***

DATA NEED

Small number of labeled examples



Continued pretraining

PURPOSE

Maintaining model accuracy for ***your domain***

DATA NEED

Large number of unlabeled datasets

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

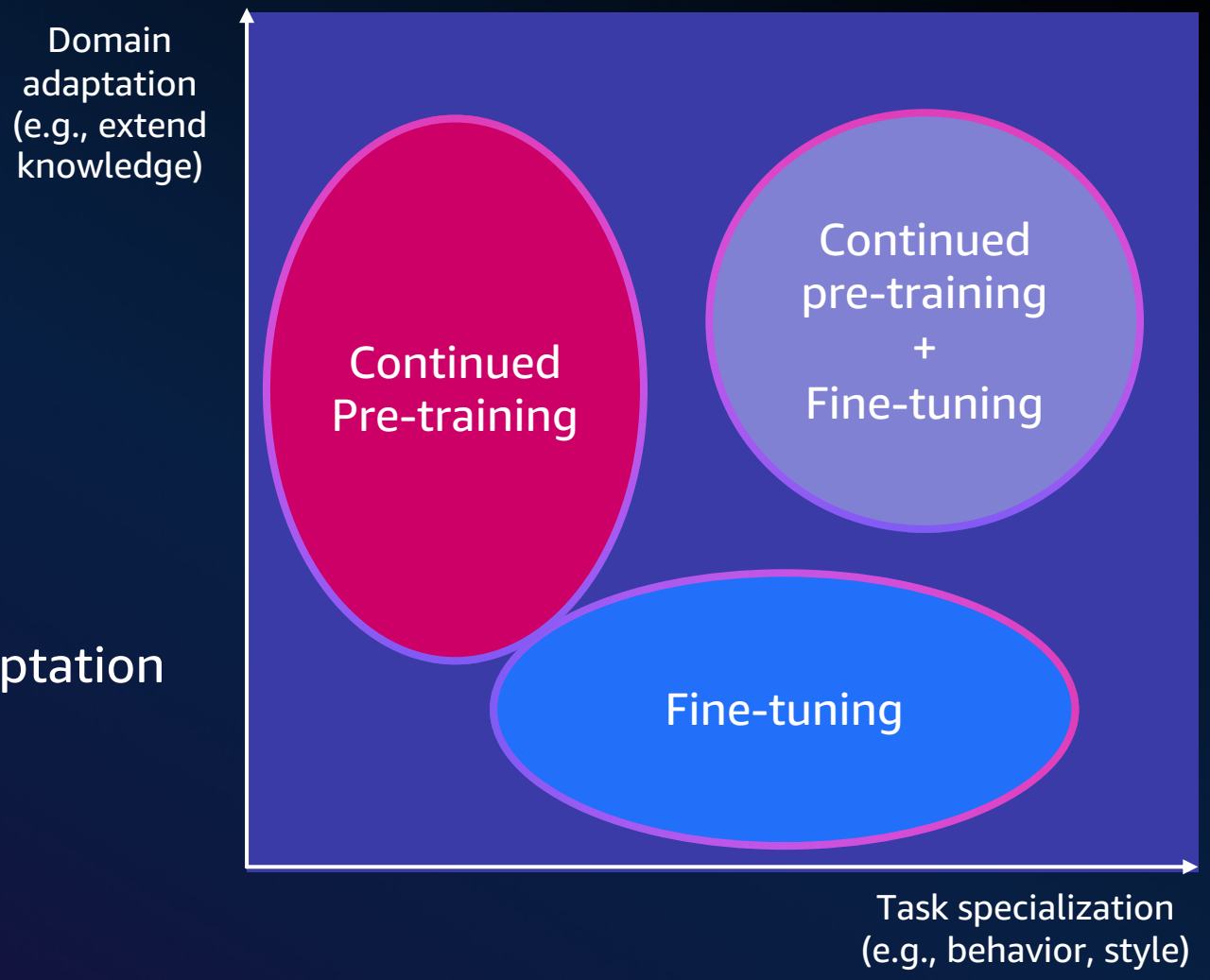
Fine-tuning and Continued pre-training

Fine-tuning

- Training dataset is available?
- Specific style, behavior required?

Continued pre-training

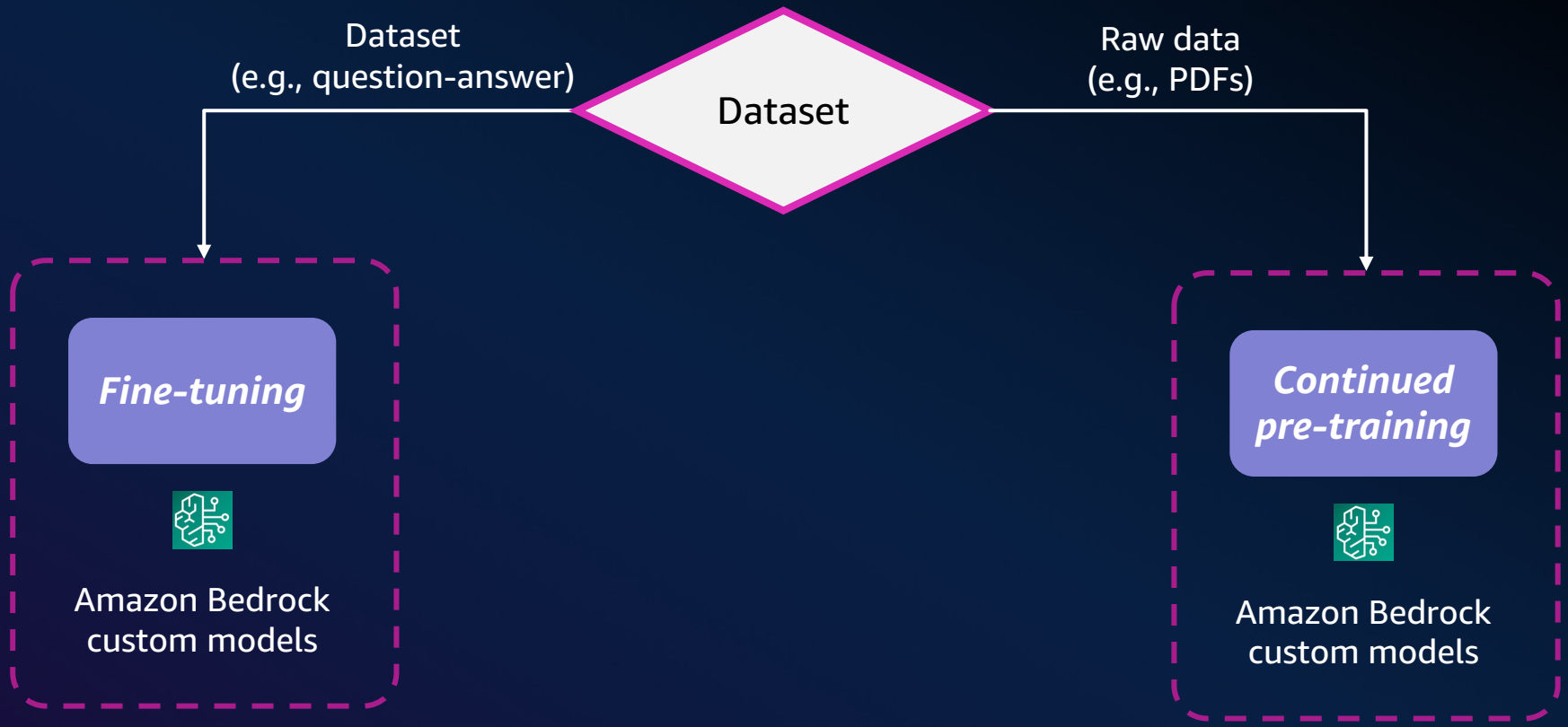
- Raw dataset (e.g., PDFs)
- Additional knowledge through domain adaptation



aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Datasets for Fine-tuning and Continued pre-training



{”prompt": "<prompt text>", ”completion": "<expected generated text>"}
{”prompt": "<prompt text>", ”completion": "<expected generated text>"}
{”prompt": "<prompt text>", ”completion": "<expected generated text>"}

{"input": "<raw text>"}
{"input": "<raw text>"}
{"input": "<raw text>"}

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

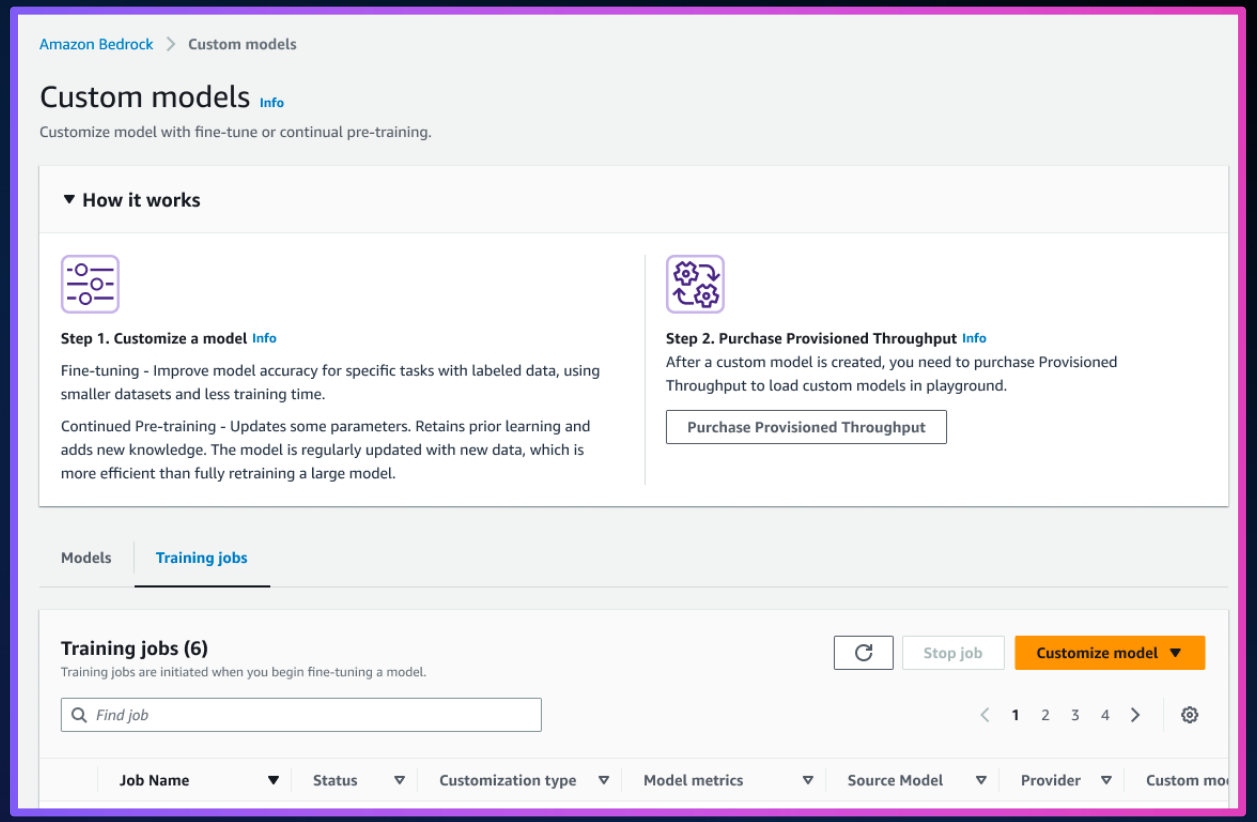
Amazon Bedrock custom models

CREATE CUSTOM MODELS USING THE CONSOLE OR APIS

Maximize accuracy of FMs by providing labeled or raw unlabeled data

Once deployed, custom models are invoked the same way as base models (playground or API)

Customizations now supported for Amazon Titan and some third-party FMs



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Security and privacy

YOU ARE ALWAYS IN CONTROL OF YOUR DATA

- ✓ Data not used to improve models, and not shared with model providers
- ✓ Customer data remain in AWS Region
- ✓ Support for ***AWS PrivateLink*** and ***VPC configurations***
- ✓ Integration with ***IAM***
- ✓ API monitoring in ***AWS CloudTrail***, logging & metrics in ***Amazon CloudWatch***
- ✓ Custom models encrypted and stored with ***service or customer-managed keys*** – Only you have access to your models

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fine-Tuning in Action: Text Summarization

```
prompt = """
Summarize the simplest and most interesting part of the following conversation.

#Person1#: Hello. My name is John Sandals, and I've got a reservation.
#Person2#: May I see some identification, sir, please?
#Person1#: Sure. Here you are.
#Person2#: Thank you so much. Have you got a credit card, Mr. Sandals?
#Person1#: I sure do. How about American Express?
#Person2#: Unfortunately, at the present time we take only MasterCard or VISA.
#Person1#: No American Express? Okay, here's my VISA.
#Person2#: Thank you, sir. You'll be in room 507, nonsmoking, with a queen-size bed. Do you approve, sir?
#Person1#: Yeah, that'll be fine.
#Person2#: That's great. This is your key, sir. If you need anything at all, anytime, just dial zero.

Summary:
"""

body = {
    "prompt": prompt,
    "temperature": 0.5,
    "top_p": 0.9,
    "max_gen_len": 512,
}
```

https://www.youtube.com/watch?v=YY9N7sDoP30

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fine-Tuning in Action: Baseline Completion

```
response = bedrock_runtime.invoke_model(
    modelId="meta.llama2-13b-chat-v1", # compare to chat model
    body=json.dumps(body)
)

response_body = response["body"].read().decode('utf8')
print(json.loads(response_body)["generation"])
```

```
A man named John Sandals checks into a hotel and provides identification and a credit card. The hotel only takes M
asterCard or VISA, so he uses his VISA card. He is given room 507, a nonsmoking room with a queen-size bed.
```

https://www.youtube.com/watch?v=YY9N7sDoP30

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fine-Tuning in Action: Improved Completion

```
response = bedrock_runtime.invoke_model(
    modelId=provisioned_model_arn, # custom fine-tuned model
    body=json.dumps(body)
)

response_body = response["body"].read().decode('utf8')
print(json.loads(response_body)["generation"])
```

```
John Sandals checks in the hotel with VISA and is assigned room 507, nonsmoking, with a queen-size bed.
```

https://www.youtube.com/watch?v=YY9N7sDoP30

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Labs

- Bedrock Console: text generation with Titan Text
- Bedrock Console: news summary with Command Light
- Bedrock SDK: news summary with Llama 2
- Bedrock SDK: image generation with Titan Image Generator
- SageMaker JumpStart console: text generation with Mistral
- SageMaker JumpStart SDK: dialog summary with LLama 3

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock
User Guide

Recently added to this guide Preview

AI21 Labs Jamba-Instruct models
June 26, 2024

AI21 Labs models
June 26, 2024

Prompt injection security
June 25, 2024

View all

- What is Amazon Bedrock?
- Set up
- Foundation model information
- Console overview
- Run model inference
- Prompt engineering guidelines
- Guardrails for Amazon Bedrock
- Model evaluation
- Knowledge bases for Amazon Bedrock

Before you can begin a model customization job, you need to minimally prepare a training dataset. Whether a validation dataset is supported and the format of your training and validation dataset depend on the following factors.

- The type of customization job (fine-tuning or Continued Pre-training).
- The input and output modalities of the data.

To see dataset and file requirements for different models, see Model customization quotas.

Select the tab that is relevant to your use-case.

Fine-tuning: Text-to-text | **Fine-tuning: Text-to-image & Image-to-embeddings** | **Continued Pre-training: Text-to-text**

To carry out Continued Pre-training on a text-to-text model, prepare a training and optional validation dataset by creating a JSONL file with multiple JSON lines. Because Continued Pre-training involves unlabeled data, each JSON line is a sample containing only an `input` field. Use 6 characters per token as an approximation for the number of tokens. The format is as follows.

```
{"input": "<input text>"}
{"input": "<input text>"}
{"input": "<input text>"}
```

The following is an example item that could be in the training data.

```
{"input": "AWS stands for Amazon Web Services"}
```

aws-cli-dataset.jsonl

Users > mba > Desktop > 20240724_Summit_Workshop_FM_Finetune > datasets > aws-cli-dataset.jsonl


```
{"input": "User Guide for Version 2\nAWS Command Line Interface\nCopyright \u00a9 2024 Amazon Web Services, Inc. and/or its a\ufb03li
{"input": "AWS Command Line Interface User Guide for Version 2\nAWS Command Line Interface: User Guide for Version 2\nCopyright \u00a
{"input": "AWS Command Line Interface User Guide for Version 2\nTable of Contents\n.....
{"input": "AWS Command Line Interface User Guide for Version 2\nRun the o\ufb03cial images .....
{"input": "AWS Command Line Interface User Guide for Version 2\nUsing the examples .....
{"input": "AWS Command Line Interface User Guide for Version 2\nUse the AWS CLI.....
{"input": "AWS Command Line Interface User Guide for Version 2\nResources.....
{"input": "AWS Command Line Interface User Guide for Version 2\nThe \"aws --version \" command returns a version after uninstalling t
{"input": "AWS Command Line Interface User Guide for Version 2\nix"}
{"input": "AWS Command Line Interface User Guide for Version 2\nWhat is the AWS Command Line Interface?\nThe AWS Command Line Interfa
{"input": "AWS Command Line Interface User Guide for Version 2\nor managed by AWS. We recommend that you install the AWS CLI from onl
{"input": "AWS Command Line Interface User Guide for Version 2\n\u2022Prompt  \u2013 The command prompt uses the Linux prompt and is
{"input": "AWS Command Line Interface User Guide for Version 2\n$ aws ec2 create-security-group --group-name my-sg --description \"My
{"input": "AWS Command Line Interface User Guide for Version 2\nGet started with the AWS CLI\nThis chapter provides steps to get star
{"input": "AWS Command Line Interface User Guide for Version 2\n\u2022Set up the AWS CLI\nPrerequisites to use the AWS CLI version 2\
{"input": "AWS Command Line Interface User Guide for Version 2\nChoose \none \nway to \nmanage \nyour \nadministr \natorTo By You can
{"input": "AWS Command Line Interface User Guide for Version 2\nImportant\nAWS CLI versions 1 and 2 use the same aws command name. If
{"input": "AWS Command Line Interface User Guide for Version 2\nAfter the yum installation of the AWS CLI is removed, follow the belo
{"input": "AWS Command Line Interface User Guide for Version 2\nLinux ARM\nNote\n(Optional)  The following command block downloads an
{"input": "AWS Command Line Interface User Guide for Version 2\n\u2022Downloading from the URL \u2013 To download the installer with
{"input": "AWS Command Line Interface User Guide for Version 2\ntCFBV1MgQ0xJIFRlYW0gPGF3cy1jbGlAYW1hem9uLmNvbT6JAlQEEwEIAD4CGwMF\nCwk
{"input": "AWS Command Line Interface User Guide for Version 2\nLinux x86 (64-bit)\nFor the latest version of the AWS CLI, use the fo
{"input": "AWS Command Line Interface User Guide for Version 2\ngpg: Signature made Mon Nov  4 19:00:01 2019 PST\ngpg:
{"input": "AWS Command Line Interface User Guide for Version 2\nYou can install without sudo if you specify directories that you alre
{"input": "AWS Command Line Interface User Guide for Version 2\n2.Use the ls command to \ufb01nd the directory that your symlink poin
{"input": "AWS Command Line Interface User Guide for Version 2\n\u2022For all users on the computer (requires sudo )\n\u2022You can i
{"input": "AWS Command Line Interface User Guide for Version 2\nCommand line installer - All users\nIf you have sudo permissions, you
{"input": "AWS Command Line Interface User Guide for Version 2\nCommand line - Current user\n1. To specify which folder the AWS CLI i
{"input": "AWS Command Line Interface User Guide for Version 2\nThe following example installs the AWS CLI in the folder /Users/myuse
{"input": "AWS Command Line Interface User Guide for Version 2\nInstall or update the AWS CLI\nTo update your current installation of
{"input": "AWS Command Line Interface User Guide for Version 2\nNext steps\nAfter you successfully install the AWS CLI, you can safel
```

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models **Preview**
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Machine Learning

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)

Try Bedrock

Get started

Overview

Amazon Bedrock is a fully managed service that makes FMs from leading AI startups and Amazon available via an API, so you can choose from a wide range of FMs to find the model that is best suited for your use case. With Bedrock's serverless experience, you can get started quickly, privately customize FMs with your own data, and easily integrate and deploy them into your applications using the AWS tools without having to manage any infrastructure.

Benefits

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Amazon Bedrock > Custom models

Custom models Info

Customize model with Fine-tuning or Continued Pre-training.

How it works

Create a model

Decide the custom model type you want to create, Fine-tuning or Continued Pre-training.

Create a model customization job. You need to purchase Provisioned Throughput before you can test or use your model in an application.

Test a custom model

Test your custom model in a playground or get the metrics from the model's details page.

Use a custom model

Use your model for inference in your application or experiment in the playground.

Models | Jobs

Models (8)

Purchase Provisioned Throughput | Customize model

Models that you have customized and have had their jobs successfully completed will appear here.

Find model

1

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Amazon Bedrock > Custom models

Custom models Info

Customize model with Fine-tuning or Continued Pre-training.

How it works

Create a model

Decide the custom model type you want to create, Fine-tuning or Continued Pre-training.

Create a model customization job. You need to purchase Provisioned Throughput before you can test or use your model in an application.

Test a custom model

Test your custom model in a playground or get the metrics from the model's details page.

Use a custom model

Use your model for inference in your application or experiment in the playground.

Models | Jobs

Models (8)

Models that you have customized and have had their jobs successfully completed will appear here.

Purchase Provisioned Throughput | Customize model

- Create Fine-tuning job
- Create Continued Pre-training job

Find model

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment

Amazon Bedrock > Custom models > Create Continued Pre-training job

Create Continued Pre-training job Info

Select the model you wish to pre-train and submit your data location.

Model details

Source model

Choose from a list of models that you wish to customize with using your own data.

Select model

Continued pre-trained model name

Enter a name to identify the newly created pre-trained model.

Enter a name here

☐ Model encryption Info

▶ **Tags - *optional***

Job configuration

Job name

Enter a name to identify the training job necessary to pre-train and create a new model.

Enter a name here

▶ **Tags - *optional***

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=CONTINUED_PRE_TRAINING

Paused

aws Services Search [Option+S] N. Virginia demouser @ 0947-8459-0684

CloudFormation EC2 RDS S3 Amazon SageMaker Amazon Bedrock AWS DeepRacer

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment
  - Model Evaluation

Select model

1. Category

Model providers

- Amazon

Custom models

- Fine-tuned models

2. Model

Titan Text G1 - Lite *v1*
Text model | Context size = 4k

Titan Text G1 - Express *v1*
Text model | Context size = 8k

Cancel Apply

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=CONTINUED_PRE_TRAINING

Paused

aws Services Search [Option+S] N. Virginia demouser @ 0947-8459-0684

CloudFormation EC2 RDS S3 Amazon SageMaker Amazon Bedrock AWS DeepRacer

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment
  - Model Evaluation

Select model

1. Category

Model providers

- Amazon

Custom models

- Fine-tuned models

2. Model

Titan Text G1 - Lite *v1*
Text model | Context size = 4k

Titan Text G1 - Express *v1*
Text model | Context size = 8k

Cancel Apply

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=CONTINUED_PRE_TRAINING&mode...

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Amazon Bedrock > Custom models > Create Fine-tuning job

Create Continued Pre-training job Info

Select the model you wish to pre-train and submit your data location.

Model details

Source model

Choose from a list of models that you wish to customize with using your own data.

Titan Text G1 - Lite *v1*

Change

Continued pre-trained model name

Enter a name to identify the newly created pre-trained model.

Enter a name here

☐ Model encryption Info

▶ **Tags - *optional***

Job configuration

Job name

Enter a name to identify the training job necessary to pre-train and create a new model.

Enter a name here

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment

Amazon Bedrock > Custom models > Create Fine-tuning job

Create Continued Pre-training job Info

Select the model you wish to pre-train and submit your data location.

Model details

Source model

Choose from a list of models that you wish to customize with using your own data.

Titan Text G1 - Lite *v1*

Change

Continued pre-trained model name

Enter a name to identify the newly created pre-trained model.

aws-cli-ft-20240630

☐ Model encryption Info

▶ **Tags - *optional***

Job configuration

Job name

Enter a name to identify the training job necessary to pre-train and create a new model.

aws-cli-ft-20240630-morning

Amazon Bedrock

Getting started
Overview
Examples
Providers

Foundation models
Base models
Custom models
Imported models Preview

Playgrounds
Chat
Text
Image

Safeguards
Guardrails
Watermark detection

Builder tools
Knowledge bases
Agents

Assessment & deployment

Input data Info

Choose a file in the S3 location. The files you choose must be in the dataset format that the model needs for training. You can also use Sagemaker Ground Truth to create and label training datasets. Learn more

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Validation dataset S3 location (optional)

s3://bucket/path-to-your-data/ View Browse S3

Hyperparameters Info

Epochs

The total number of iterations of all the training data in one cycle for training the model.

5

Enter an integer between 1 and 10.

Batch size

The number of samples processed before model parameters are updated.

1

Enter a value between 1 and 64.

Learning rate

The rate at which model parameters are updated after each batch of training data.

0.00001

Enter a float value between 0 and 1.

Amazon Bedrock

Getting started
Overview
Exam
Provi
Four
Base
Custo
Impo
Play
Chat
Text
Imag
Safe
Guar
Wate
Builder tools
Knowledge bases
Agents
Assessment & deployment
Model Evaluation
Provisioned Throughput

Input data Info

Choose a file in the S3 location. The files you choose must be in the dataset format that the model needs for training. You can also use Sagemaker Ground Truth to create and label training datasets. Learn more

S3 location

Choose an archive in S3

S3 buckets

Buckets (83)

cli

3No matches

1

Name	Creation date
aws-cli-dataset-ft-input-20240630	2024-06-30T01:55:00.000Z
aws-cli-dataset-ft-output-20240630	2024-06-30T01:55:11.000Z

Cancel Choose

Learning rate

The rate at which model parameters are updated after each batch of training data.

0.00001

Enter a float value between 0 and 1.

Learning rate warmup steps

Number of iterations over which learning rate is gradually increased to the initial rate specified.

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Input data Info

Choose a file in the S3 location. The files you choose must be in the dataset format that the model needs for training. You can also use Sagemaker Ground Truth to create and label training datasets. Learn more

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Choose an archive in S3

S3 buckets > aws-cli-dataset-ft-input-20240630

Objects (1/1)

Find object by prefix

1

Key	Last modified	Size
aws-cli-dataset.jsonl	June 30, 2024, 09:56 (UTC+08:00)	9.39 MB

Cancel Choose

1

Enter a value between 1 and 64.

Learning rate

The rate at which model parameters are updated after each batch of training data.

0.00001

Enter a float value between 0 and 1.

Learning rate warmup steps

Number of iterations over which learning rate is gradually increased to the initial rate specified.

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=CONTINUED_PRE_TRAINING&mode...

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Input data Info

Choose a file in the S3 location. The files you choose must be in the dataset format that the model needs for training. You can also use Sagemaker Ground Truth to create and label training datasets. Learn more

S3 location

s3://aws-cli-dataset-ft-input-20240630/aws-cli-dataset.jsonl

View | Browse S3

Validation dataset S3 location (optional)

s3://bucket/path-to-your-data/

View | Browse S3

Hyperparameters Info

Epochs

The total number of iterations of all the training data in one cycle for training the model.

5

Enter an integer between 1 and 10.

Batch size

The number of samples processed before model parameters are updated.

1

Enter a value between 1 and 64.

Learning rate

The rate at which model parameters are updated after each batch of training data.

0.00001

Enter a float value between 0 and 1.

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment

Hyperparameters Info

Epochs
The total number of iterations of all the training data in one cycle for training the model.

5

Enter an integer between 1 and 10.

Batch size
The number of samples processed before model parameters are updated.

1

Enter a value between 1 and 64.

Learning rate
The rate at which model parameters are updated after each batch of training data.

0.00001

Enter a float value between 0 and 1.

Learning rate warmup steps
Number of iterations over which learning rate is gradually increased to the initial rate specified.

5

Enter an integer between 1 and 250.

Output data Info

Choose S3 location to store the model validation outputs.

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment

Enter an integer between 1 and 250.

Output data Info

Choose S3 location to store the model validation outputs.

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Service access Info

Bedrock model customization job requires permissions to write to S3 on your behalf.

Choose a method to authorize Bedrock

- (●) Use an existing service role
- ( ) Create and use a new service role

Service role

Choose role

Purchase Provisioned Throughput to use continued pre-trained model
After this custom model is created, you need to purchase Provisioned Throughput to be able to use this model.

Learn more

Enter an integer between 1 and 250.

Output data Info

Choose S3 location to store the model validation outputs.

Choose an archive in S3

S3 buckets

Buckets (83)

cli

3No matches

1

Name	Creation date
aws-cli-dataset-ft-input-20240630	2024-06-30T01:55:00.000Z
aws-cli-dataset-ft-output-20240630	2024-06-30T01:55:11.000Z

Cancel Choose

Purchase Provisioned Throughput to use continued pre-trained model
After this custom model is created, you need to purchase Provisioned Throughput to be able to use this model.

Learn more

Cancel Create Continued Pre-training job

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Enter an integer between 1 and 250.

Output data Info

Choose S3 location to store the model validation outputs.

Choose an archive in S3

S3 buckets

Buckets (1/83)

cli

3No matches

1

Name	Creation date
aws-cli-dataset-ft-input-20240630	2024-06-30T01:55:00.000Z
aws-cli-dataset-ft-output-20240630	2024-06-30T01:55:11.000Z

Cancel Choose

Purchase Provisioned Throughput to use continued pre-trained model
After this custom model is created, you need to purchase Provisioned Throughput to be able to use this model.

Learn more

Cancel Create Continued Pre-training job

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment
  - Model Evaluation

Enter an integer between 1 and 250.

Output data Info

Choose S3 location to store the model validation outputs.

S3 location

s3://aws-cli-dataset-ft-output-20240630

View | Browse S3

Service access Info

Bedrock model customization job requires permissions to write to S3 on your behalf.

Choose a method to authorize Bedrock

- (●) Use an existing service role
- ( ) Create and use a new service role

Service role

Choose role

Purchase Provisioned Throughput to use continued pre-trained model
After this custom model is created, you need to purchase Provisioned Throughput to be able to use this model.

Learn more

Cancel | Create Continued Pre-training job

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment
  - Model Evaluation

Output data Info

Choose S3 location to store the model validation outputs.

S3 location

s3://aws-cli-dataset-ft-output-20240630 View Browse S3

Service access Info

Bedrock model customization job requires permissions to write to S3 on your behalf.

Choose a method to authorize Bedrock

- ○ Use an existing service role
- ◉ Create and use a new service role

Service role name

aws-cli-ft-role-20240630

Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.

View permission details

Purchase Provisioned Throughput to use continued pre-trained model
After this custom model is created, you need to purchase Provisioned Throughput to be able to use this model.

Learn more

Cancel Create Continued Pre-training job

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment

Continued Pre-train model job aws-cli-ft-20240630-morning created successfully. Once the training job completes successfully the model will be available in the Models table. Training time can vary depending on your hyperparameter settings.

Amazon Bedrock > Custom models

Custom models Info

Customize model with Fine-tuning or Continued Pre-training.

How it works

Create a model

Decide the custom model type you want to create, Fine-tuning or Continued Pre-training.

Create a model customization job. You need to purchase Provisioned Throughput before you can test or use your model in an application.

Test a custom model

Test your custom model in a playground or get the metrics from the model's details page.

Use a custom model

Use your model for inference in your application or experiment in the playground.

Models | Jobs

Models (8)

Purchase Provisioned Throughput | Customize model

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models?tab=jobs

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Models | Jobs

Jobs (10)

Training jobs are initiated when you begin customizing a model.

Stop job | Customize model

Find job

1

Job name	Status	Source	Job type	Custom model n...	Creation time
aws-cli-ft-20240...	Training	Titan Text G1 - Lite	Continued Pre-tr...	aws-cli-ft-20240...	June 30, 2024, 1...

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment

Amazon Bedrock > Custom models > Training job : aws-cli-ft-20240630-morning

Training job : aws-cli-ft-20240630-morning Info

Stop job

Training job overview

Custom model name
aws-cli-ft-20240630

Training job ARN
arn:aws:bedrock:us-east-1:094784590684:model-customization-job/amazon.titan-text-lite-v1:0:4k/6acth3jsdq6g

Data access role
arn:aws:iam::094784590684:role/service-role/aws-cli-ft-role-20240630

Job creation time
June 30, 2024, 10:15 (UTC+08:00)

Job duration
a few seconds

Source model name
Titan Text G1 - Lite

Status
Training

Custom model encryption KMS key
Bedrock owned KMS key

Source model ARN
arn:aws:bedrock:us-east-1::foundation-model/amazon.titan-text-lite-v1:0:4k

Input data

S3 location
aws-cli-dataset-ft-input-20240630

Hyperparameters

Learning Rate Warmup Steps
5

Learning Rate
0.00001

Maximum number of epochs
5

Batch Size

Labs

- Bedrock Console: text generation with Titan Text
- Bedrock Console: news summary with Command Light
- Bedrock SDK: news summary with Llama 2
- Bedrock SDK: image generation with Titan Image Generator
- SageMaker JumpStart console: text generation with Mistral
- SageMaker JumpStart SDK: dialog summary with LLama 3

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock
User Guide

Recently added to this guide Preview

AI21 Labs Jamba-Instruct models
June 26, 2024

AI21 Labs models
June 26, 2024

Prompt injection security
June 25, 2024

View all

- What is Amazon Bedrock?
- Set up
- Foundation model information
- Console overview
- Run model inference
- Prompt engineering guidelines
- Guardrails for Amazon Bedrock
- Model evaluation
- Knowledge bases for Amazon Bedrock

Before you can begin a model customization job, you need to minimally prepare a training dataset. Whether a validation dataset is supported and the format of your training and validation dataset depend on the following factors.

- The type of customization job (fine-tuning or Continued Pre-training).
- The input and output modalities of the data.

To see dataset and file requirements for different models, see Model customization quotas.

Select the tab that is relevant to your use-case.

Fine-tuning: Text-to-text | Fine-tuning: Text-to-image & Image-to-embeddings | Continued Pre-training: Text-to-text

To fine-tune a text-to-text model, prepare a training and optional validation dataset by creating a JSONL file with multiple JSON lines. Each JSON line is a sample containing both a `prompt` and `completion` field. Use 6 characters per token as an approximation for the number of tokens. The format is as follows.

```
{"prompt": "<prompt1>", "completion": "<expected generated text>"}
{"prompt": "<prompt2>", "completion": "<expected generated text>"}
{"prompt": "<prompt3>", "completion": "<expected generated text>"}
```

The following is an example item for a question-answer task:

```
{"prompt": "what is AWS", "completion": "it's Amazon Web Services"}
```

Welcome | train-cnn-5K.jsonl

Users > mba > Desktop > 20240724_Summit_Workshop_FM_Finetune > datasets > train-cnn-5K.jsonl

prompt 1 of 5165


```
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
{"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that
```

Users > mba > Desktop > 20240724_Summit_Workshop_FM_Finetune > datasets > {} train-cnn-5K.jsonl

completion 1 of 5009


```
 grabs EVERY WEEK… . Clash of the titans: Novak Djokovic remains on course to meet Andy Murray in the quarter-finals .", "completion":
im believed she was about to be killed and jumped from an upstairs window to escape the clutches of her attacker - leaving her having
one such photo tour in Tanzania when he spotted the lioness. Petr, 42, added: 'In Tanzania wild animals get used to the tourists so if
a on the Oval outfield . Pleased as punch: Root jumps for joy as he celebrates his century on day four at The Oval .", "completion": "

said that his friends and classmates \"were happy to help the environment.\" Hansen donated five pairs of his own outgrown jeans. The
cers were working at the mall when the shooting occurred, Cole said. The mall was locked down and evacuated, with 15 agencies helping
 to waste that moon/And the heat on the hood of this Ford'' on the parkin' ballad ''Talk,'' well, you may not want to waste it either.
aiding and abetting an unlicensed driver. Police will consider if further action is appropriate.", "completion": "response:\n\nMan let
-- three for \"Rehab\" as well as Album of the Year and Best New Artist. Winehouse's Grammy winning album, \"Back to Black,\" is still
 commercial broadcaster that Berlusconi founded. He has also been barred from holding public office for two years and was expelled fro
llips puts his players through their paces in a training session . The custom vehicle fitter and former goalkeeper said: 'I had the ab
the amount I would usually sell in a whole year.' Belinda, who sources her fakes from . India, China and Hatton Garden in London, tel
astily rewritten to allow her to leave the programme after filming only a handful of scenes. She said: 'The plan had been that my . ch
spect a human caused the blaze, but they're still investigating ."}
 picture. Isaacson's book should hopefully provide some concrete answers to unresolved questions about Jobs's life, along with some ne
ead an economic seminar at Macaulay Honors College in New York .\nPetraeus will assume the role of visiting professor at the college i
d or not.\" Stern also pointed out that people often tell him that his program was the introduction of reality television. Despite nak
ver 180 countries. Sanitas, the BUPA business in Spain, has one million insured customers who have access to a network of 18,000 medic
ot to stay completely in the present. 'That's . what I'm going to try to do for all 18 holes tomorrow. You have to . think about how y
 in latter part of the First Temple period - from the end of the eighth century BC until the destruction of the Temple in 586 BC. To f
ng.\"", "completion": "response:\n\nOfficials: 81 people are treated at local hospitals for dizziness, nausea, headaches .\nThe incide
s provisional repairs will be done on the nation's critical north-south highway by Wednesday. But because of poor communication, traff
 to stab him in the back with a kitchen knife. 'Fortunately the blade of the knife hit his shoulder blade, he managed to fight her off

fficult decision to make. As a prosecutor -- I've been a prosecutor for 30 years -- I hate to say this, but you're somewhat accustomed

squad for Euro 2016 qualifier against Macedonia .\nPique took in basketball game with popstar wife during time off ."}
n October. For any couples out there who want to attempt something similar be warned, the towers of the bridge aren't open to the publ
. 'In short, leave the damn Tree alone, so that future fans can enjoy it. Left alone, the Tree will be there for many, many decades to
 \"food service establishments.\" Sugary drinks linked to 180,000 deaths worldwide . Mayor Michael Bloomberg firmly disagreed. \"Today
```

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Amazon Bedrock > Custom models

Custom models Info

Customize model with Fine-tuning or Continued Pre-training.

How it works

Create a model

Decide the custom model type you want to create, Fine-tuning or Continued Pre-training.

Create a model customization job. You need to purchase Provisioned Throughput before you can test or use your model in an application.

Test a custom model

Test your custom model in a playground or get the metrics from the model's details page.

Use a custom model

Use your model for inference in your application or experiment in the playground.

Models | Jobs

Models (8)

Purchase Provisioned Throughput | Customize model

Models that you have customized and have had their jobs successfully completed will appear here.

Find model

1

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Amazon Bedrock > Custom models

Custom models Info

Customize model with Fine-tuning or Continued Pre-training.

How it works

Create a model

Decide the custom model type you want to create, Fine-tuning or Continued Pre-training.

Create a model customization job. You need to purchase Provisioned Throughput before you can test or use your model in an application.

Test a custom model

Test your custom model in a playground or get the metrics from the model's details page.

Use a custom model

Use your model for inference in your application or experiment in the playground.

Models | Jobs

Models (8)

Purchase Provisioned Throughput | Customize model

Models that you have customized and have had their jobs successfully completed will appear here.

Find model

- Create Fine-tuning job
- Create Continued Pre-training job

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=FINE_TUNING

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Amazon Bedrock > Custom models > Create Fine-tuning job

Create Fine-tuning job Info

Select the model you wish to fine-tune and submit your data location.

Model details

Source model

Choose from a list of models that you wish to customize with using your own data.

Select model

Fine-tuned model name

Enter a name to identify the new fine-tuned model.

Enter a name here

☐ Model encryption Info

▶ **Tags - *optional***

Job configuration

Job name

Enter a name to identify the training job necessary to pre-train and create a new model.

Enter a name here

▶ **Tags - *optional***

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=FINE_TUNING

Paused

aws Services Search [Option+S] N. Virginia demouser @ 0947-8459-0684

CloudFormation EC2 RDS S3 Amazon SageMaker Amazon Bedrock AWS DeepRacer

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment
  - Model Evaluation

Select model

1. Category

Model providers

- Amazon
- Cohere
- Meta

Custom models

- Fine-tuned models

2. Models available for fine-tuning

Command *v14.7*
Text model

Command Light *v14.7*
Text model

Cancel Apply

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment

Amazon Bedrock > Custom models > Create Fine-tuning job

Create Fine-tuning job Info

Select the model you wish to fine-tune and submit your data location.

Model details

Source model

Choose from a list of models that you wish to customize with using your own data.

Command Light *v14.7*

Change

Fine-tuned model name

Enter a name to identify the new fine-tuned model.

Enter a name here

☐ Model encryption Info

▶ **Tags - *optional***

Job configuration

Job name

Enter a name to identify the training job necessary to pre-train and create a new model.

Enter a name here

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=FINE_TUNING&modelId=cohere.co...

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Amazon Bedrock > Custom models > Create Fine-tuning job

Create Fine-tuning job Info

Select the model you wish to fine-tune and submit your data location.

Model details

Source model

Choose from a list of models that you wish to customize with using your own data.

Command Light *v14.7*

Change

Fine-tuned model name

Enter a name to identify the new fine-tuned model.

cnn-daily-ft-20240630

☐ Model encryption Info

▶ **Tags - *optional***

Job configuration

Job name

Enter a name to identify the training job necessary to pre-train and create a new model.

cnn-daily-ft-20240630-morning

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment

Input data Info

Choose a file in the S3 location. The files you choose must be in the dataset format that the model needs for training. You can also use Sagemaker Ground Truth to create and label training datasets. Learn more

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Validation dataset S3 location (optional)

s3://bucket/path-to-your-data/ View Browse S3

Evaluate the model by providing the S3 location of a validation dataset or by using a percentage of the training dataset. If you don't provide a validation dataset, specify the evaluation percentage in the Hyperparameters section.

Hyperparameters Info

Epochs

The total number of iterations of all the training data in one cycle for training the model.

1

Enter an integer between 1 and 100.

Batch size

The number of samples processed before model parameters are updated.

32

Amazon Bedrock

Getting started
Overview
Examples
Provi
Four
Base
Custo
Impo
Play
Chat
Text
Imag
Safe
Guar
Watermark detection
Builder tools
Knowledge bases
Agents
Assessment & deployment
Model Evaluation
Provisioned Throughput

Input data Info

Choose a file in the S3 location. The files you choose must be in the dataset format that the model needs for training. You can also use Sagemaker Ground Truth to create and label training datasets. Learn more

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Choose an archive in S3

S3 buckets

Buckets (83)

cnn 2No matches 1

Name	Creation date
cnn-daily-ft-input-20240630	2024-06-30T02:05:26.000Z
cnn-daily-ft-output-20240630	2024-06-30T02:05:36.000Z

Cancel Choose

Enter an integer between 1 and 100.

Batch size

The number of samples processed before model parameters are updated.

32

Learning rate

The rate at which model parameters are updated after each batch of training data.

0.0001

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Amazon Bedrock

Getting started
Overview
Examples
Provid
Four
Base
Custo
Impo
Play
Chat
Text
Imag
Safe
Guar
Watermark detection
Builder tools
Knowledge bases
Agents
Assessment & deployment
Model Evaluation
Provisioned Throughput

Input data Info

Choose a file in the S3 location. The files you choose must be in the dataset format that the model needs for training. You can also use Sagemaker Ground Truth to create and label training datasets. Learn more

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Choose an archive in S3

S3 buckets > cnn-daily-ft-input-20240630

Objects (1/1)

Find object by prefix

1

Key	Last modified	Size
train-cnn-5K.jsonl	June 30, 2024, 10:08 (UTC+08:00)	11.36 MB

Cancel Choose

1

Enter an integer between 1 and 100.

Batch size

The number of samples processed before model parameters are updated.

32

Learning rate

The rate at which model parameters are updated after each batch of training data.

0.0001

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models/jobs/create?jobType=FINE_TUNING&modelId=cohere.co...

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Input data Info

Choose a file in the S3 location. The files you choose must be in the dataset format that the model needs for training. You can also use Sagemaker Ground Truth to create and label training datasets. Learn more

S3 location

s3://cnn-daily-ft-input-20240630/train-cnn-5K.jsonl

View | Browse S3

Validation dataset S3 location (optional)

s3://bucket/path-to-your-data/

View | Browse S3

Evaluate the model by providing the S3 location of a validation dataset or by using a percentage of the training dataset. If you don't provide a validation dataset, specify the evaluation percentage in the Hyperparameters section.

Hyperparameters Info

Epochs

The total number of iterations of all the training data in one cycle for training the model.

1

Enter an integer between 1 and 100.

Batch size

The number of samples processed before model parameters are updated.

32

Amazon Bedrock

Getting started
Overview
Examples
Providers

Foundation models
Base models
Custom models
Imported models Preview

Playgrounds
Chat
Text
Image

Safeguards
Guardrails
Watermark detection

Builder tools
Knowledge bases
Agents

Assessment & deployment

Hyperparameters Info

Epochs
The total number of iterations of all the training data in one cycle for training the model.

10

Enter an integer between 1 and 100.

Batch size
The number of samples processed before model parameters are updated.

32

Learning rate
The rate at which model parameters are updated after each batch of training data.

0.0001

Enter a float value between 0.00005 and 0.1.

Early stopping threshold
The minimum amount of validation loss improvement required to continue training the model.

0.01

Enter a value between 0 and 0.1.

Early stopping patience
The number of epochs to wait for an improvement with validation loss before stopping training. If the validation loss improvement has not reached the early stopping threshold within the patience period, the training will be stopped.

6

Enter an integer between 1 and 10.

Evaluation percentage
The percentage of the dataset allocated for model evaluation if a separate validation dataset is not provided.

Enter a value between 5% and 50%

Output data Info

Choose S3 location to store the model validation outputs.

S3 location

s3://bucket/path-to-your-data/ View Browse S3

Service access Info

Bedrock model customization job requires permissions to write to S3 on your behalf.

Choose a method to authorize Bedrock

- (●) Use an existing service role
- ( ) Create and use a new service role

Service role

Choose role

Purchase Provisioned Throughput to use fine-tuned model
After this custom model is created, you need to purchase Provisioned Throughput to be able to use this model.

Learn more

Cancel Create Fine-tuning job

Enter a value between 5% and 50%

Output data Info

Choose S3 location to store the model validation outputs.

S3 location

Choose an archive in S3

S3 buckets

Buckets (1/83)

cnn

2No matches

1

Name	Creation date
cnn-daily-ft-input-20240630	2024-06-30T02:05:26.000Z
cnn-daily-ft-output-20240630	2024-06-30T02:05:36.000Z

Cancel Choose

Purchase Provisioned Throughput to use fine-tuned model
After this custom model is created, you need to purchase Provisioned Throughput to be able to use this model.

Learn more

Cancel Create Fine-tuning job

© 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

Enter a value between 5% and 50%

Output data Info

Choose S3 location to store the model validation outputs.

S3 location

s3://cnn-daily-ft-output-20240630 View Browse S3

Service access Info

Bedrock model customization job requires permissions to write to S3 on your behalf.

Choose a method to authorize Bedrock

- (●) Use an existing service role
- ( ) Create and use a new service role

Service role

Choose role

Purchase Provisioned Throughput to use fine-tuned model
After this custom model is created, you need to purchase Provisioned Throughput to be able to use this model.

Learn more

Cancel Create Fine-tuning job

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment
  - Model Evaluation

Output data Info

Choose S3 location to store the model validation outputs.

S3 location

s3://cnn-daily-ft-output-20240630 View Browse S3

Service access Info

Bedrock model customization job requires permissions to write to S3 on your behalf.

Choose a method to authorize Bedrock

- ○ Use an existing service role
- ◉ Create and use a new service role

Service role name

cnn-daily-ft-service-role-20240630

Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.

View permission details

Purchase Provisioned Throughput to use fine-tuned model
After this custom model is created, you need to purchase Provisioned Throughput to be able to use this model.

Learn more

Cancel Create Fine-tuning job

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Create Fine-tuning job job cnn-daily-ft-20240630-morning created successfully. Once the training job completes successfully the model will be available in the Models table. Training time can vary depending on your hyperparameter settings.

Amazon Bedrock > Custom models

Custom models Info

Customize model with Fine-tuning or Continued Pre-training.

How it works

Create a model

Decide the custom model type you want to create, Fine-tuning or Continued Pre-training.

Create a model customization job. You need to purchase Provisioned Throughput before you can test or use your model in an application.

Test a custom model

Test your custom model in a playground or get the metrics from the model's details page.

Use a custom model

Use your model for inference in your application or experiment in the playground.

Models | Jobs

Models (8)

Purchase Provisioned Throughput | Customize model

Amazon Bedrock | us-east-1

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models?tab=jobs

Paused

aws | Services | Search [Option+S] | N. Virginia | demouser @ 0947-8459-0684

CloudFormation | EC2 | RDS | S3 | Amazon SageMaker | Amazon Bedrock | AWS DeepRacer

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
- **Assessment & deployment**

Models | Jobs

Jobs (10)

Training jobs are initiated when you begin customizing a model.

Stop job | Customize model

Find job

1

Job name	Status	Source	Job type	Custom model n...	Creation time
cnn-daily-ft-202...	Training	Command Light	Fine-tuning	cnn-daily-ft-202...	June 30, 2024, 1...

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
- Assessment & deployment

Training job : cnn-daily-ft-20240630-morning Info

Stop job

Training job overview

Custom model name
cnn-daily-ft-20240630

Training job ARN
arn:aws:bedrock:us-east-1:094784590684:model-customization-job/cohere.command-light-text-v14:7:4k/kkgw68zdf9vr

Data access role
arn:aws:iam::094784590684:role/service-role/cnn-daily-ft-service-role-20240630

Job creation time
June 30, 2024, 10:24 (UTC+08:00)

Job duration
a few seconds

Source model name
Command Light

Status
Training

Custom model encryption KMS key
Bedrock owned KMS key

Source model ARN
arn:aws:bedrock:us-east-1::foundation-model/cohere.command-light-text-v14:7:4k

Input data

S3 location
cnn-daily-ft-input-20240630

Hyperparameters

Learning Rate
0.0001

Evaluation percentage
20.0

Maximum number of epochs
10

Labs

- Bedrock Console: text generation with Titan Text
- Bedrock Console: news summary with Command Light
- Bedrock SDK: news summary with Llama 2
- Bedrock SDK: image generation with Titan Image Generator
- SageMaker JumpStart console: text generation with Mistral
- SageMaker JumpStart SDK: dialog summary with LLama 3

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

https://github.com/aws-samples/amazon-bedrock-workshop/tree/main/03_Model_customization

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon WorkSpaces

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 00_setup | Domain- | Domain- | The Ama

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Filter files by name

/ amazon-bedrock-workshop-20240712 / 03_Model_customization /

Name	Last Modified
data	18 hours ago
fine-tuning-datasets	a day ago
=7,	a day ago
00_setup.ipynb	2 hours ago
01_fine-tuning-titan-lite.ipynb	2 hours ago
02_fine-tuning_llama2.ipynb	2 hours ago
03_continued_pretraining_titan...	2 hours ago
04_cleanup.ipynb	a day ago
README.md	a day ago

00_setup.ipynb | 01_fine-tuning-titan- | 02_fine-tuning_llama | 03_continued_pretrai | 1-TIGFT-customizatio | 2-TIGFT-provisioned-

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Setup for running customization notebooks both for fine-tuning and continued pre-training using Amazon Bedrock

In this notebook, we will create set of roles and an s3 bucket which will be used for other notebooks in this module.

> This notebook should work well with the `Data Science 3.0`, `Python 3`, and `ml.c5.2xlarge` kernel in SageMaker Studio

Prerequisites

Custom job role

The notebook allows you to either create a Bedrock role for running customization jobs in the **Create IAM customisation job role** section or you can skip this section and create Bedrock Service role for customization jobs following instructions on managing permissions for customization jobs. If you want to using an existing custom job role please edit the variable **customization_role** and also ensure it has access to the S3 bucket which is created containing the dataset.

Create IAM Pre-requisites

This notebook requires permissions to:

- create and delete Amazon IAM roles
- create, update and delete Amazon S3 buckets
- access Amazon Bedrock

If you are running this notebook without an Admin role, make sure that your role include the following managed policies:

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Setup

Install and import all the needed libraries and dependencies to complete this notebook.

Warning: Please ignore error messages related to pip's dependency resolver.

[2]:
```
!pip install --upgrade pip
%pip install --no-build-isolation --force-reinstall \
    "boto3>=1.28.57" \
    "awscli>=1.29.57" \
    "botocore>=1.31.57"
!pip install -qU --force-reinstall langchain typing_extensions pypdf urllib3==2.1.0
!pip install -qU ipywidgets>=7,<8
!pip install jsonlines
!pip install datasets==2.15.0
!pip install pandas==2.1.3
!pip install matplotlib==3.8.2
```

```
Requirement already satisfied: pip in /opt/conda/lib/python3.10/site-packages (24.1.2)
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your system unusabl
e.It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what you are doing and want to s
uppress this warning.
Collecting boto3>=1.28.57
  Using cached boto3-1.34.143-py3-none-any.whl.metadata (6.6 kB)
Collecting awscli>=1.29.57
  Using cached awscli-1.33.25-py3-none-any.whl.metadata (11 kB)
Collecting botocore>=1.31.57
  Using cached botocore-1.34.143-py3-none-any.whl.metadata (5.7 kB)
Collecting jmespath<2.0.0,>=0.7.1 (from boto3>=1.28.57)
  Using cached jmespath-1.0.1-py3-none-any.whl.metadata (7.6 kB)
Collecting s3transfer<0.11.0,>=0.10.0 (from boto3>=1.28.57)
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[3]:
```
# restart kernel for packages to take effect
from IPython.core.display import HTML
HTML("<script>Jupyter.notebook.kernel.restart()</script>")
```

[3]:

[4]:
```
import warnings
warnings.filterwarnings('ignore')
import json
import os
import sys
import boto3
import time
import pprint
from datasets import load_dataset
import random
import jsonlines
```

[5]:
```
session = boto3.session.Session()
region = session.region_name
sts_client = boto3.client('sts')
account_id = sts_client.get_caller_identity()["Account"]
s3_suffix = f"{region}-{account_id}"
bucket_name = f"bedrock-customization-{s3_suffix}"
s3_client = boto3.client('s3')
bedrock = boto3.client(service_name="bedrock")
bedrock_runtime = boto3.client(service_name="bedrock-runtime")
iam = boto3.client('iam', region_name=region)
```

[6]:
```
import uuid
suffix = str(uuid.uuid4())
role_name = "BedrockRole-" + suffix
s3_bedrock_finetuning_access_policy="BedrockPolicy-" + suffix
customization_role = f"arn:aws:iam::{account_id}:role/{role_name}"
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Testing boto3 connection

We will list the foundation models to test the bot3 connection and make sure bedrock client has been successfully created.

[7]:
```
for model in bedrock.list_foundation_models(
    byCustomizationType="FINE_TUNING")["modelSummaries"]:
    for key, value in model.items():
        print(key, ":", value)
    print("-----\n")
```

```
modelArn : arn:aws:bedrock:us-east-1::foundation-model/amazon.titan-image-generator-v1:0
modelId : amazon.titan-image-generator-v1:0
modelName : Titan Image Generator G1
providerName : Amazon
inputModalities : ['TEXT', 'IMAGE']
outputModalities : ['IMAGE']
customizationsSupported : ['FINE_TUNING']
inferenceTypesSupported : ['PROVISIONED']
modelLifecycle : {'status': 'ACTIVE'}
-----

modelArn : arn:aws:bedrock:us-east-1::foundation-model/amazon.titan-text-lite-v1:0:4k
modelId : amazon.titan-text-lite-v1:0:4k
modelName : Titan Text G1 - Lite
providerName : Amazon
inputModalities : ['TEXT']
outputModalities : ['TEXT']
responseStreamingSupported : True
customizationsSupported : ['FINE_TUNING', 'CONTINUED_PRE_TRAINING']
inferenceTypesSupported : ['PROVISIONED']
modelLifecycle : {'status': 'ACTIVE'}
-----
```

Amazon WorkSpaces

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 00_setup | Domain- | Domain- | The Ama

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
modelArn : arn:aws:bedrock:us-east-1::foundation-model/cohere.command-light-text-v14:7:4k
modelId : cohere.command-light-text-v14:7:4k
modelName : Command Light
providerName : Cohere
inputModalities : ['TEXT']
outputModalities : ['TEXT']
responseStreamingSupported : True
customizationsSupported : ['FINE_TUNING']
inferenceTypesSupported : ['PROVISIONED']
modelLifecycle : {'status': 'ACTIVE'}
-----

modelArn : arn:aws:bedrock:us-east-1::foundation-model/meta.llama2-13b-v1:0:4k
modelId : meta.llama2-13b-v1:0:4k
modelName : Llama 2 13B
providerName : Meta
inputModalities : ['TEXT']
outputModalities : ['TEXT']
responseStreamingSupported : True
customizationsSupported : ['FINE_TUNING']
inferenceTypesSupported : []
modelLifecycle : {'status': 'LEGACY'}
-----

modelArn : arn:aws:bedrock:us-east-1::foundation-model/meta.llama2-70b-v1:0:4k
modelId : meta.llama2-70b-v1:0:4k
modelName : Llama 2 70B
providerName : Meta
inputModalities : ['TEXT']
outputModalities : ['TEXT']
responseStreamingSupported : True
customizationsSupported : ['FINE_TUNING']
inferenceTypesSupported : []
modelLifecycle : {'status': 'LEGACY'}
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Create s3 bucket

In this step we will create a s3 bucket, which will be used to store data for fine-tuning and continued pre-training notebooks.

[8]:
```
# Create S3 bucket for knowledge base data source
s3bucket = s3_client.create_bucket(
    Bucket=bucket_name,
    ## Uncomment the following if you run into errors
    # CreateBucketConfiguration={
    #     'LocationConstraint':region,
    # },
)
```

Creating role and policies required to run customization jobs with Amazon Bedrock

This JSON object defines the trust relationship that allows the bedrock service to assume a role that will give it the ability to talk to other required AWS services. The conditions set restrict the assumption of the role to a specfic account ID and a specific component of the bedrock service (model_customization_jobs)

[9]:
```
ROLE_DOC = f"""{{
    "Version": "2012-10-17",
    "Statement": [
        {{
            "Effect": "Allow",
            "Principal": {{
                "Service": "bedrock.amazonaws.com"
            }},
            "Action": "sts:AssumeRole",
            "Condition": {{
                "StringEquals": {{
                    "aws:SourceAccount": "{account_id}"
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Creating role and policies required to run customization jobs with Amazon Bedrock

This JSON object defines the trust relationship that allows the bedrock service to assume a role that will give it the ability to talk to other required AWS services. The conditions set restrict the assumption of the role to a specfic account ID and a specific component of the bedrock service (model_customization_jobs)

[9]:
```
ROLE_DOC = f"""{{
    "Version": "2012-10-17",
    "Statement": [
        {{
            "Effect": "Allow",
            "Principal": {{
                "Service": "bedrock.amazonaws.com"
            }},
            "Action": "sts:AssumeRole",
            "Condition": {{
                "StringEquals": {{
                    "aws:SourceAccount": "{account_id}"
                }},
                "ArnEquals": {{
                    "aws:SourceArn": "arn:aws:bedrock:{region}:{account_id}:model-customization-job/*"
                }}
            }}
        }}
    ]
}}
"""
```

This JSON object defines the permissions of the role we want bedrock to assume to allow access to the S3 bucket that we created that will hold our fine-tuning datasets and allow certain bucket and object manipulations.

[10]:
```
ACCESS_POLICY_DOC = f"""{{
```

Amazon WorkSpaces

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 00_setup | Domain- | Domain- | The Ama

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynb | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[10]:
```
ACCESS_POLICY_DOC = f"""{{
    "Version": "2012-10-17",
    "Statement": [
        {{
            "Effect": "Allow",
            "Action": [
                "s3:AbortMultipartUpload",
                "s3:DeleteObject",
                "s3:PutObject",
                "s3:GetObject",
                "s3:GetBucketAcl",
                "s3:GetBucketNotification",
                "s3:ListBucket",
                "s3:PutBucketNotification"
            ],
            "Resource": [
                "arn:aws:s3:::{bucket_name}",
                "arn:aws:s3:::{bucket_name}/*"
            ]
        }}
    ]
}}"""
```

[11]:
```
response = iam.create_role(
    RoleName=role_name,
    AssumeRolePolicyDocument=ROLE_DOC,
    Description="Role for Bedrock to access S3 for finetuning",
)
pprint.pp(response)
```

```
{'Role': {'Path': '/',
          'RoleName': 'BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac',
          'RoleId': 'AROARMEM2CNOOMCZTHO5I',
          'Arn': 'arn:aws:iam::094784590684:role/BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac',
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[11]:
```
response = iam.create_role(
    RoleName=role_name,
    AssumeRolePolicyDocument=ROLE_DOC,
    Description="Role for Bedrock to access S3 for finetuning",
)
pprint.pp(response)
```

```
{'Role': {'Path': '/',
          'RoleName': 'BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac',
          'RoleId': 'AROARMEM2CNOOMCZTHO5I',
          'Arn': 'arn:aws:iam::094784590684:role/BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac',
          'CreateDate': datetime.datetime(2024, 7, 12, 5, 35, 7, tzinfo=tzlocal()),
          'AssumeRolePolicyDocument': {'Version': '2012-10-17',
                                       'Statement': [{'Effect': 'Allow',
                                                      'Principal': {'Service': 'bedrock.amazonaws.com'},
                                                      'Action': 'sts:AssumeRole',
                                                      'Condition': {'StringEquals': {'aws:SourceAccount': '094784590684'},
                                                                    'ArnEquals': {'aws:SourceArn': 'arn:aws:bedrock:us-east-1:094784590684:model-customization-job/*'}}}]}},
 'ResponseMetadata': {'RequestId': '8b973123-4d38-47f2-9340-a5bd2907c10b',
                      'HTTPStatusCode': 200,
                      'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 05:35:07 GMT',
                                      'x-amzn-requestid': '8b973123-4d38-47f2-9340-a5bd2907c10b',
                                      'content-type': 'text/xml',
                                      'content-length': '1812'},
                      'RetryAttempts': 0}}
```

[12]:
```
role_arn = response["Role"]["Arn"]
pprint.pp(role_arn)
```

```
'arn:aws:iam::094784590684:role/BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac'
```

[13]:
```
response = iam.create_policy(
    PolicyName=s3_bedrock_finetuning_access_policy,
    PolicyDocument=ACCESS_POLICY_DOC,
)
pprint.pp(response)
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
)
pprint.pp(response)
```

```
{'Policy': {'PolicyName': 'BedrockPolicy-257549b9-3aff-4c30-86e9-530c1416e6ac',
            'PolicyId': 'ANPARMEM2CNOPFWJXG3E2',
            'Arn': 'arn:aws:iam::094784590684:policy/BedrockPolicy-257549b9-3aff-4c30-86e9-530c1416e6ac',
            'Path': '/',
            'DefaultVersionId': 'v1',
            'AttachmentCount': 0,
            'PermissionsBoundaryUsageCount': 0,
            'IsAttachable': True,
            'CreateDate': datetime.datetime(2024, 7, 12, 5, 35, 7, tzinfo=tzlocal()),
            'UpdateDate': datetime.datetime(2024, 7, 12, 5, 35, 7, tzinfo=tzlocal())},
 'ResponseMetadata': {'RequestId': '8bffbb0b-4903-4692-82e5-de0f1a92d5ce',
                      'HTTPStatusCode': 200,
                      'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 05:35:07 GMT',
                                      'x-amzn-requestid': '8bffbb0b-4903-4692-82e5-de0f1a92d5ce',
                                      'content-type': 'text/xml',
                                      'content-length': '835'},
                      'RetryAttempts': 0}}
```

[14]:
```
policy_arn = response["Policy"]["Arn"]
pprint.pp(policy_arn)
```

```
'arn:aws:iam::094784590684:policy/BedrockPolicy-257549b9-3aff-4c30-86e9-530c1416e6ac'
```

[15]:
```
iam.attach_role_policy(
    RoleName=role_name,
    PolicyArn=policy_arn,
)
```

[15]:
```
{'ResponseMetadata': {'RequestId': '081e20e5-d600-4e22-9ae1-758abafc53c8',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 05:35:07 GMT',
   'x-amzn-requestid': '081e20e5-d600-4e22-9ae1-758abafc53c8',
   'content-type': 'text/xml',
   'content-length': '212'},
```

Amazon WorkSpaces

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 00_setup | Domain- | Domain- | The Ama

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Prepare CNN news article dataset for fine-tuning job and evaluation

The dataset that will be used is a collection of new articles from CNN and the associated highlights from that article. More information can be found at huggingface: https://huggingface.co/datasets/cnn_dailymail

[16]:
```
#Load cnn dataset from huggingface
dataset = load_dataset("cnn_dailymail",'3.0.0')
```

View the structure of the dataset

[17]:
```
print(dataset)
```

```
DatasetDict({
    train: Dataset({
        features: ['article', 'highlights', 'id'],
        num_rows: 287113
    })
    validation: Dataset({
        features: ['article', 'highlights', 'id'],
        num_rows: 13368
    })
    test: Dataset({
        features: ['article', 'highlights', 'id'],
        num_rows: 11490
    })
})
```

Prepare the Fine-tuning Dataset In this example, we are using a .jsonl dataset following example format:

{"prompt": "", "completion": ""}

The following is an example item for a question-answer task:{"prompt": "prompt is AWS", "completion": "it's Amazon Web Services"}

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[18]:
```
instruction='''Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

instruction:

Summarize the news article provided below.

input:

'''
```

For the 'completion' component we will attach the word "response" and new lines together with the summary/highlights of the article. The transformation of each datapoint is performed with the code below

[19]:
```
datapoints_train=[]
for dp in dataset['train']:
    temp_dict={}
    temp_dict['prompt']=instruction+dp['article']
    temp_dict['completion']='response:\n\n'+dp['highlights']
    datapoints_train.append(temp_dict)
```

An example of a processed datapoint can be printed below

[20]:
```
print(datapoints_train[4]['prompt'])
```

```
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

instruction:

Summarize the news article provided below.

input:
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

An example of a processed datapoint can be printed below

[20]:
```
print(datapoints_train[4]['prompt'])
```

```
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

instruction:

Summarize the news article provided below.

input:

(CNN)  -- The National Football League has indefinitely suspended Atlanta Falcons quarterback Michael Vick without pay, officials with the league said Friday. NFL star Mic
hael Vick is set to appear in court Monday. A judge will have the final say on a plea deal. Earlier, Vick admitted to participating in a dogfighting ring as part of a plea
agreement with federal prosecutors in Virginia. "Your admitted conduct was not only illegal, but also cruel and reprehensible. Your team, the NFL, and NFL fans have all be
en hurt by your actions," NFL Commissioner Roger Goodell said in a letter to Vick. Goodell said he would review the status of the suspension after the legal proceedings ar
e over. In papers filed Friday with a federal court in Virginia, Vick also admitted that he and two co-conspirators killed dogs that did not fight well. Falcons owner Arth
ur Blank said Vick's admissions describe actions that are "incomprehensible and unacceptable." The suspension makes "a strong statement that conduct which tarnishes the go
od reputation of the NFL will not be tolerated," he said in a statement.  Watch what led to Vick's suspension » . Goodell said the Falcons could "assert any claims or reme
dies" to recover $22 million of Vick's signing bonus from the 10-year, $130 million contract he signed in 2004, according to The Associated Press. Vick said he would plead
guilty to one count of "Conspiracy to Travel in Interstate Commerce in Aid of Unlawful Activities and to Sponsor a Dog in an Animal Fighting Venture" in a plea agreement f
iled at U.S. District Court in Richmond, Virginia. The charge is punishable by up to five years in prison, a $250,000 fine, "full restitution, a special assessment and 3 y
ears of supervised release," the plea deal said. Federal prosecutors agreed to ask for the low end of the sentencing guidelines. "The defendant will plead guilty because t
```

The same processing is done for the validation and test sets as well.

[21]:
```
datapoints_valid=[]
for dp in dataset['validation']:
    temp_dict={}
    temp_dict['prompt']=instruction+dp['article']
    temp_dict['completion']='response:\n\n'+dp['highlights']
    datapoints_valid.append(temp_dict)
```

[22]:
```
datapoints_test=[]
```

Amazon WorkSpaces

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | train-cnn | Domain- | Domain- | The Ama

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/fine-tuning-datasets/train-cnn-...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Filter files by name

/ ··· / 03_Model_customization / fine-tuning-datasets /

Name	Last Modified
model-customization-job-hj6q5...	a day ago
test-cnn-10.jsonl	a day ago
train-cnn-5K.jsonl	a day ago
validation-cnn-1K.jsonl	a day ago

00_setup.ipyr | 01_fine-tunin | 02_fine-tunin | test-cnn-10.js | train-cnn-5K.j | 03_continued | 1-TIGFT-custc | 2-TIGFT-provi

1 {"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.\n\ninstruction:\n\nSummarize the news article provided below.\n\ninput:\n\n(CNN) -- Slim Whitman, the high-pitched yodeling county music singer whose productive career spanned decades, died Wednesday, his son-in-law told CNN. He was 90. The singer-songwriter, born Ottis Dewey Whitman Jr., died of heart failure at Orange Park Medical Center in Florida, Roy Beagle said. Whitman gained fame in Europe as well as the United States. \"Love Song of the Waterfall,\" which a Country Music Television biography calls his \"breakthrough\" hit, was released in the early '50s. His next single \"Indian Love Call\" brought him stardom, according to the bio. \"Whitman joined the Grand Ole Opry, and then went to Britain in 1956 as the first country singer to play the London Palladium. Throughout the late '50s and early '60s, he had a string of British hits, including 'Tumbling Tumbleweeds,' 'Unchain My Heart,' and 'I'll Take You Home Again Kathleen,\" the bio said. He gained cult status after he filmed a TV commercial that touted a released of his top greatest hits, a compilation that was a great success. \"Between 1980 and 1984, Whitman had a small run of minor hits, highlighted by 1980's number 15 hit 'When.' In the late '80s, he returned to television-marketed albums, releasing Slim Whitman: Best Loved Favorites in 1989 and 20 Precious Memories in 1991. During the '90s, Whitman recorded infrequently but continued to tour successfully, particularly in Europe and Australia,\" CMT said. His music was featured in the 1996 film \"Mars Attacks!\" In the film, the sound of Whitman's \"Indian Love Call\" made the heads of invading Martians explode. People we've lost in 2013: The lives they lived .", "completion": "response:\n\nCountry music singer Slim Whitman died at 90 Wednesday .\nThe cause was heart failure .\nThe yodeling singer gained fame in Europe as well as in the U.S."}

2 {"prompt": "Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.\n\ninstruction:\n\nSummarize the news article provided below.\n\ninput:\n\nA young schoolgirl was horrified to discover six metal nails in her chicken and chips from the family's favourite local takeaway. Mollie Howe, 11, had been just about to tuck in to her tea when she spotted the inch-long nails in the £1.30 meal. Her furious mother, Sharon Howe, 39, has now blasted the Red Rooster fast food outlet, in Dagenham, Essex, which served her daughter the dangerous meal. Mollie Howe (left) and her mother Sharon Howe (right) were horrified to discover six nails in her takeaway from Red Rooster in Dagenham, Essex . Ms Howe said: 'I'm just so glad Mollie didn't eat it - who knows what might have happened? She came home, opened it up and there were these six nails - just like the ones you hammer into the wall. 'It's even more worrying because my dad could have eaten it as he often goes there too. He's 81 and lives on his own.' Ms Howe, of Dagenham, Essex, added that her family had been regular customers at the Red Rooster takeaway for years - but none of them will be going back. 'It's almost an everyday thing because it's cheap and cheerful. But I definitely won't be going there again - it just isn't worth the risk,' said the mother-of-two. The lethal inch-long metal nails were discovered nestled amongst the 11-year-old's chicken and chips . Maumand Khan, who works at Red Rooster (pictured) apologised and said he was baffled as to how the nails had got into the schoolgirl's food . 'None of my family and friends will be using it either. 'I just want people to be aware of what happened because it could have been a lot worse if poor Mollie had started eating and swallowed those nails.' Ms Howe said she was so outraged that she immediately went to confront staff at Red Rooster with her son Reece Hare, 21. 'I was so angry that my son wouldn't even let me go in there to see them,' she admitted. 'He went in instead and told them about it and they just wanted to offer him more food - it's not good enough.' Maumand Khan, who works at the takeaway, said he was baffled as to how the nails had got into Mollie's food. Mr Khan said: 'We are terribly sorry, but we can't understand how this has happened. 'We don't have that kind of nail anywhere in the shop.' He said they had not had any other complaints of this nature since the Red Rooster opened back in 2000. Furious Ms Howe

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

The same processing is done for the validation and test sets as well.

[21]:
```
datapoints_valid=[]
for dp in dataset['validation']:
    temp_dict={}
    temp_dict['prompt']=instruction+dp['article']
    temp_dict['completion']='response:\n\n'+dp['highlights']
    datapoints_valid.append(temp_dict)
```

[22]:
```
datapoints_test=[]
for dp in dataset['test']:
    temp_dict={}
    temp_dict['prompt']=instruction+dp['article']
    temp_dict['completion']='response:\n\n'+dp['highlights']
    datapoints_test.append(temp_dict)
```

Here we define some helper functions to process our datapoints further by modifying the number of datapoints we want to include in each set and the max string length of the datapoints we want to include. The final function will convert our datasets into JSONL files.

[23]:
```
def dp_transform(data_points,num_dps,max_dp_length):
    lines=[]
    for dp in data_points:
        if len(dp['prompt']+dp['completion'])<=max_dp_length:
                lines.append(dp)
    random.shuffle(lines)
    lines=lines[:num_dps]
    return lines
```

[24]:
```
def jsonl_converter(dataset,file_name):
    print(file_name)
    with jsonlines.open(file_name, 'w') as writer:
        for line in dataset:
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Process data partitions. Every LLM may have different input token limits and what string of characters represents a token is defined by a particular model's vocabulary. For simplicity, we have restricted each datapoint to be <=3,000 characters.

[25]:
```
train=dp_transform(datapoints_train,5000,3000)
validation=dp_transform(datapoints_valid,999,3000)
test=dp_transform(datapoints_test,10,3000)
```

Create local directory for datasets

Please note that your training dataset for fine-tuning cannot be greater than 10K records, and validation dataset has a maximum limit of 1K records.

[26]:
```
dataset_folder="fine-tuning-datasets"
train_file_name="train-cnn-5K.jsonl"
validation_file_name="validation-cnn-1K.jsonl"
test_file_name="test-cnn-10.jsonl"
!mkdir fine-tuning-datasets
abs_path=os.path.abspath(dataset_folder)
```

```
mkdir: cannot create directory ‘fine-tuning-datasets’: File exists
```

Create JSONL format datasets for Bedrock fine-tuning job

[27]:
```
jsonl_converter(train,f'{abs_path}/{train_file_name}')
jsonl_converter(validation,f'{abs_path}/{validation_file_name}')
jsonl_converter(test,f'{abs_path}/{test_file_name}')
```

```
/root/amazon-bedrock-workshop-20240712/03_Model_customization/fine-tuning-datasets/train-cnn-5K.jsonl
/root/amazon-bedrock-workshop-20240712/03_Model_customization/fine-tuning-datasets/validation-cnn-1K.jsonl
/root/amazon-bedrock-workshop-20240712/03_Model_customization/fine-tuning-datasets/test-cnn-10.jsonl
```

Upload datasets to s3 bucket

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/00_setup.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Upload datasets to s3 bucket

Uploading both training and test dataset. We will use the training and validation datasets for fine-tuning the model. The test dataset will be used for evaluating the performance of the model on an unseen input.

[28]:
```
s3_client.upload_file(f'{abs_path}/{train_file_name}', bucket_name, f'fine-tuning-datasets/train/{train_file_name}')
s3_client.upload_file(f'{abs_path}/{validation_file_name}', bucket_name, f'fine-tuning-datasets/validation/{validation_file_name}')
s3_client.upload_file(f'{abs_path}/{test_file_name}', bucket_name, f'fine-tuning-datasets/test/{test_file_name}')
```

[29]:
```
s3_train_uri=f's3://{bucket_name}/fine-tuning-datasets/train/{train_file_name}'
s3_validation_uri=f's3://{bucket_name}/fine-tuning-datasets/validation/{validation_file_name}'
s3_test_uri=f's3://{bucket_name}/fine-tuning-datasets/test/{test_file_name}'
```

Storing variables to be used in other notebooks.

> Please make sure to use the same kernel as used for 00_setup.ipynb for other notebooks on fine-tuning and continued pre-training.

[30]:
```
%store role_arn
%store bucket_name
%store role_name
%store policy_arn
%store s3_train_uri
%store s3_validation_uri
%store s3_test_uri
```

```
Stored 'role_arn' (str)
Stored 'bucket_name' (str)
Stored 'role_name' (str)
Stored 'policy_arn' (str)
Stored 's3_train_uri' (str)
Stored 's3_validation_uri' (str)
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan- | 02_fine-tuning_llama | 03_continued_pretrai | 1-TIGFT-customizatio | 2-TIGFT-provisioned-

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Filter files by name

/ amazon-bedrock-workshop-20240712 / 03_Model_customization /

Name	Last Modified
data	18 hours ago
fine-tuning-datasets	a day ago
=7,	a day ago
00_setup.ipynb	2 hours ago
01_fine-tuning-titan-lite.ipynb	2 hours ago
02_fine-tuning_llama2.ipynb	2 hours ago
03_continued_pretraining_titan...	2 hours ago
04_cleanup.ipynb	a day ago
README.md	a day ago

Fine-tune Meta Llama2 13B model provided by Amazon Bedrock: End-to-End

In this notebook we demonstrate using Boto3 sdk for the fine-tuning and provisioning of Llama2 13B model in Bedrock. You can also do this through the Bedrock Console.

Warning: This module cannot be executed in Workshop Studio Accounts, and you will have to run this notebook in your own account.

A Summarization Use Case

In this notebook, we build an end-to-end workflow for fine-tuning and evaluating the Foundation Models (FMs) in Amazon Bedrock. We choose Meta Llama 2 13B as our FM to perform the customization through fine-tuning, we then create provisioned throughput of the fine-tuned model, test the provisioned model invocation, and finally evaluate the fine-tuned model performance using fmeval on the summarization accuracy metrics including METEOR, ROUGE, and BERT scores. We have defined these scores in the `Evaluate the Provisioned Custom Model¶` section below.

> *This notebook should work well with the `Data Science 3.0`, `Python 3`, and `ml.c5.2xlarge` kernel in SageMaker Studio*

Prerequisites

- Make sure you have executed `00_setup.ipynb` notebook.
- Make sure you are using the same kernel and instance as `00_setup.ipynb` notebook.

Amazon WorkSpaces

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 02_fine-t | Domain- | Domain- | The Ama

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Fine-tune Meta Llama2 13B model provided by Amazon Bedrock: End-to-End

In this notebook we demonstrate using Boto3 sdk for the fine-tuning and provisioning of Llama2 13B model in Bedrock. You can also do this through the Bedrock Console.

Warning: This module cannot be executed in Workshop Studio Accounts, and you will have to run this notebook in your own account.

A Summarization Use Case

In this notebook, we build an end-to-end workflow for fine-tuning and evaluating the Foundation Models (FMs) in Amazon Bedrock. We choose Meta Llama 2 13B as our FM to perform the customization through fine-tuning, we then create provisioned throughput of the fine-tuned model, test the provisioned model invocation, and finally evaluate the fine-tuned model performance using fmeval on the summarization accuracy metrics including METEOR, ROUGE, and BERT scores. We have defined these scores in the `Evaluate the Provisioned Custom Model¶` section below.

> *This notebook should work well with the `Data Science 3.0`, `Python 3`, and `ml.c5.2xlarge` kernel in SageMaker Studio*

Prerequisites

- Make sure you have executed `00_setup.ipynb` notebook.
- Make sure you are using the same kernel and instance as `00_setup.ipynb` notebook.

In this notebook we demonstrate using Boto3 sdk for the fine-tuning and provisioning of Llama2 13B model in Bedrock. You can also do this through the Bedrock Console.

Warning: This notebook will create provisioned throughput for testing the fine-tuned model. Therefore, please make sure to delete the provisioned throughput as mentioned in the last section of the notebook, otherwise you will be charged for it, even if you are not using it.

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Setup

Install and import all the needed libraries and dependencies to complete this notebook.

Please ignore error messages related to pip's dependency resolver.

[2]:
```
# # install the fmeval package for foundation model evaluation
!rm -Rf ~/.cache/pip/*
!pip install tokenizers==0.12.1
!pip install -qU fmeval==0.3.0
```

```
Collecting tokenizers==0.12.1
  Downloading tokenizers-0.12.1-cp310-cp310-manylinux_2_12_x86_64.manylinux2010_x86_64.whl.metadata (6.5 kB)
Downloading tokenizers-0.12.1-cp310-cp310-manylinux_2_12_x86_64.manylinux2010_x86_64.whl (6.6 MB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 6.6/6.6 MB 39.9 MB/s eta 0:00:00:00:01
Installing collected packages: tokenizers
  Attempting uninstall: tokenizers
    Found existing installation: tokenizers 0.19.1
    Uninstalling tokenizers-0.19.1:
      Successfully uninstalled tokenizers-0.19.1
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflict
s.
transformers 4.42.4 requires tokenizers<0.20,>=0.19, but you have tokenizers 0.12.1 which is incompatible.
Successfully installed tokenizers-0.12.1
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your system unusable.I
t is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what you are doing and want to suppre
ss this warning.
ERROR: pip's dependency resolver does not currently take into account all the packages that are installed. This behaviour is the source of the following dependency conflict
s.
sphinx 7.2.6 requires docutils<0.21,>=0.18.1, but you have docutils 0.16 which is incompatible.
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager, possibly rendering your system unusable.I
t is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv. Use the --root-user-action option if you know what you are doing and want to suppre
ss this warning.
```

Amazon WorkSpaces

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 02_fine-t | Domain- | Domain- | The Ama

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[3]:
```
# restart kernel for packages to take effect
from IPython.core.display import HTML
HTML("<script>Jupyter.notebook.kernel.restart()</script>")
```

[3]:

[4]:
```
## Fetching varialbes from `00_setup.ipynb` notebook.
%store -r role_arn
%store -r s3_train_uri
%store -r s3_validation_uri
%store -r s3_test_uri
%store -r bucket_name
```

[5]:
```
import pprint
pprint.pp(role_arn)
pprint.pp(s3_train_uri)
pprint.pp(s3_validation_uri)
pprint.pp(s3_test_uri)
pprint.pp(bucket_name)
```

```
'arn:aws:iam::094784590684:role/BedrockRole-257549b9-3aff-4c30-86e9-530c1416e6ac'
's3://bedrock-customization-us-east-1-094784590684/fine-tuning-datasets/train/train-cnn-5K.jsonl'
's3://bedrock-customization-us-east-1-094784590684/fine-tuning-datasets/validation/validation-cnn-1K.jsonl'
's3://bedrock-customization-us-east-1-094784590684/fine-tuning-datasets/test/test-cnn-10.jsonl'
'bedrock-customization-us-east-1-094784590684'
```

[6]:
```
import warnings
warnings.filterwarnings('ignore')
import json
import os
import sys
import boto3
import pandas as pd
from matplotlib import pyplot as plt
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[7]:
```
session = boto3.session.Session()
region = session.region_name
sts_client = boto3.client('sts')
s3_client = boto3.client('s3')
aws_account_id = sts_client.get_caller_identity()["Account"]
bedrock = boto3.client(service_name="bedrock")
bedrock_runtime = boto3.client(service_name="bedrock-runtime")
```

[8]:
```
test_file_name = "test-cnn-10.jsonl"
data_folder = "fine-tuning-datasets"
```

Create the Fine-Tuning Job

Note: Fine-tuning job will take around 60mins to complete with 5K records.

Meta Llama2 customization hyperparameters:

- `epochs` : The number of iterations through the entire training dataset and can take up any integer values in the range of 1-10, with a default value of 2.
- `batchSize` : The number of samples processed before updating model parametersand can take up any integer values in the range of 1-64, with a default value of 1.
- `learningRate` : The rate at which model parameters are updated after each batch which can take up a float value betweek 0.0-1.0 with a default value set to 1.00E-5.
- `learningRateWarmupSteps` : The number of iterations over which the learning rate is gradually increased to the specified rate and can take any integer value between 0-250 with a default value of 5.

For guidelines on setting hyper-parameters refer to the guidelines provided here

[9]:
```
from datetime import datetime
ts = datetime.now().strftime("%Y-%m-%d-%H-%M-%S")
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Create the Fine-Tuning Job

Note: Fine-tuning job will take around 60mins to complete with 5K records.

Meta Llama2 customization hyperparameters:

- `epochs` : The number of iterations through the entire training dataset and can take up any integer values in the range of 1-10, with a default value of 2.
- `batchSize` : The number of samples processed before updating model parametersand can take up any integer values in the range of 1-64, with a default value of 1.
- `learningRate` : The rate at which model parameters are updated after each batch which can take up a float value betweek 0.0-1.0 with a default value set to 1.00E-5.
- `learningRateWarmupSteps` : The number of iterations over which the learning rate is gradually increased to the specified rate and can take any integer value between 0-250 with a default value of 5.

For guidelines on setting hyper-parameters refer to the guidelines provided here

[9]:
```
from datetime import datetime
ts = datetime.now().strftime("%Y-%m-%d-%H-%M-%S")


# Choose the foundation model you want to customize and provide ModelId(find more about model reference at https://docs.aws.amazon.com/bedrock/latest/userguide/bedrock-refere
base_model_id = "meta.llama2-13b-v1:0:4k"

# Select the customization type from "FINE_TUNING" or "CONTINUED_PRE_TRAINING".
customization_type = "FINE_TUNING"

# Specify the roleArn for your customization job
customization_role = role_arn

# Create a customization job name
customization_job_name = f"llama2-finetune-sm-test-model-{ts}"
```

Amazon WorkSpaces

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 02_fine-t | Domain- | Domain- | The Ama

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[9]:
```
from datetime import datetime
ts = datetime.now().strftime("%Y-%m-%d-%H-%M-%S")


# Choose the foundation model you want to customize and provide ModelId(find more about model reference at https://docs.aws.amazon.com/bedrock/latest/userguide/bedrock-refere
base_model_id = "meta.llama2-13b-v1:0:4k"

# Select the customization type from "FINE_TUNING" or "CONTINUED_PRE_TRAINING".
customization_type = "FINE_TUNING"

# Specify the roleArn for your customization job
customization_role = role_arn

# Create a customization job name
customization_job_name = f"llama2-finetune-sm-test-model-{ts}"

# Create a customized model name for your fine-tuned Llama2 model
custom_model_name = f"llama2-finetune-{ts}"

# Define the hyperparameters for fine-tuning Llama2 model
hyper_parameters = {
        "epochCount": "2",
        "batchSize": "1",
        "learningRate": "0.00005",
    }

# Specify your data path for training, validation(optional) and output
training_data_config = {"s3Uri": s3_train_uri}

# # uncomment the below section if you have validation dataset and provide the s3 uri for it.
validation_data_config = {
        "validators": [{
            "s3Uri": s3_validation_uri
        }]
    }
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
output_data_config = {"s3Uri": f's3://{bucket_name}/outputs/output-{custom_model_name}'}

# # Create the customization job
bedrock.create_model_customization_job(
    customizationType=customization_type,
    jobName=customization_job_name,
    customModelName=custom_model_name,
    roleArn=customization_role,
    baseModelIdentifier=base_model_id,
    hyperParameters=hyper_parameters,
    trainingDataConfig=training_data_config,
    validationDataConfig=validation_data_config,
    outputDataConfig=output_data_config
)
```

[9]:
```
{'ResponseMetadata': {'RequestId': '814513c4-65b0-41bc-b83e-5abae996cf6c',
  'HTTPStatusCode': 201,
  'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 05:39:59 GMT',
   'content-type': 'application/json',
   'content-length': '112',
   'connection': 'keep-alive',
   'x-amzn-requestid': '814513c4-65b0-41bc-b83e-5abae996cf6c'},
  'RetryAttempts': 0},
 'jobArn': 'arn:aws:bedrock:us-east-1:094784590684:model-customization-job/meta.llama2-13b-v1:0:4k/hj6q5t4jcg91'}
```

Check Customization Job Status

[10]:
```
import time
fine_tune_job = bedrock.get_model_customization_job(jobIdentifier=customization_job_name)["status"]
print(fine_tune_job)

while fine_tune_job == "InProgress":
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Markdown | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Check Customization Job Status

[10]:
```
import time
fine_tune_job = bedrock.get_model_customization_job(jobIdentifier=customization_job_name)["status"]
print(fine_tune_job)

while fine_tune_job == "InProgress":
    time.sleep(60)
    fine_tune_job = bedrock.get_model_customization_job(jobIdentifier=customization_job_name)["status"]
    print (fine_tune_job)
```

```
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
```

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
  - Prompt management Preview
  - Prompt flows Preview
- Assessment & deployment

Custom models Info

Customize model with Fine-tuning or Continued Pre-training.

How it works

Create a model

Decide the custom model type you want to create, Fine-tuning or Continued Pre-training.

Create a model customization job. You need to purchase Provisioned Throughput before you can test or use your model in an application.

Test a custom model

Test your custom model in a playground or get the metrics from the model's details page.

Use a custom model

Use your model for inference in your application or experiment in the playground.

Models | Jobs

Jobs (10)

Training jobs are initiated when you begin customizing a model.

Stop job | Customize model

Find job

1

Job name	Status	Source	Job type	Custom model name	Creation time
llama2-finetune-sm-te...	Training	Llama 2 13B	Fine-tuning	llama2-finetune-20...	July 12, 2024, 13:...

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
InProgress
Completed
```

Retrieve Custom Model

Once the customization job is finished, you can check your existing custom model(s) and retrieve the modelArn of your fine-tuned Llama2 model.

[11]:
```
# You can list your custom models using the command below
bedrock.list_custom_models()
```

Amazon WorkSpaces

AI/ML and Generative AI Enable | Isengard | Amazon Bedrock | us-east-1 | SageMaker Studio | 02_fine-tuni... (4) - JupyterLab

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/custom-models?tab=jobs

aws | Services | Search [Alt+S] | N. Virginia | Admin/linmicht-Isengard @ 0947-8459-0684

Amazon Bedrock

- Getting started
  - Overview
  - Examples
  - Providers
- Foundation models
  - Base models
  - Custom models
  - Imported models Preview
- Playgrounds
  - Chat
  - Text
  - Image
- Safeguards
  - Guardrails
  - Watermark detection
- Builder tools
  - Knowledge bases
  - Agents
  - Prompt management Preview
  - Prompt flows Preview
- Assessment & deployment

Amazon Bedrock > Custom models

Custom models Info

Customize model with Fine-tuning or Continued Pre-training.

How it works

Create a model

Decide the custom model type you want to create, Fine-tuning or Continued Pre-training.

Create a model customization job. You need to purchase Provisioned Throughput before you can test or use your model in an application.

Test a custom model

Test your custom model in a playground or get the metrics from the model's details page.

Use a custom model

Use your model for inference in your application or experiment in the playground.

Models | Jobs

Jobs (10)

Training jobs are initiated when you begin customizing a model.

Stop job | Customize model

Find job

1

Job name	Status	Source	Job type	Custom model name	Creation...
llama2-finetune-sm-test-mod...	Complete	Llama 2 13B	Fine-tuning	llama2-finetune-20...	July 12, 2...

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Retrieve Custom Model

Once the customization job is finished, you can check your existing custom model(s) and retrieve the modelArn of your fine-tuned Llama2 model.

[11]:
```
# You can list your custom models using the command below
bedrock.list_custom_models()
```

[11]:

```
{'ResponseMetadata': {'RequestId': 'c13432a9-498c-4300-8b6b-5297b2664c86',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 06:26:07 GMT',
   'content-type': 'application/json',
   'content-length': '3719',
   'connection': 'keep-alive',
   'x-amzn-requestid': 'c13432a9-498c-4300-8b6b-5297b2664c86'},
  'RetryAttempts': 0},
 'modelSummaries': [{'modelArn': 'arn:aws:bedrock:us-east-1:094784590684:custom-model/meta.llama2-13b-v1:0:4k/vrme87cn827l',
   'modelName': 'llama2-finetune-2024-07-12-05-39-59',
   'creationTime': datetime.datetime(2024, 7, 12, 5, 39, 59, 672000, tzinfo=tzlocal()),
   'baseModelArn': 'arn:aws:bedrock:us-east-1::foundation-model/meta.llama2-13b-v1:0:4k',
   'baseModelName': '',
   'customizationType': 'FINE_TUNING'},
  {'modelArn': 'arn:aws:bedrock:us-east-1:094784590684:custom-model/cohere.command-light-text-v14:7:4k/jglb0h06zthe',
   'modelName': 'cnn-daily-ft-20240630',
   'creationTime': datetime.datetime(2024, 6, 30, 2, 24, 3, 210000, tzinfo=tzlocal()),
   'baseModelArn': 'arn:aws:bedrock:us-east-1::foundation-model/cohere.command-light-text-v14:7:4k',
   'baseModelName': '',
   'customizationType': 'FINE_TUNING'},
  {'modelArn': 'arn:aws:bedrock:us-east-1:094784590684:custom-model/meta.llama2-13b-v1:0:4k/34kn0fkqp2sb',
   'modelName': 'llama2-finetune-2024-06-26-14-13-22',
   'creationTime': datetime.datetime(2024, 6, 26, 14, 13, 22, 359000, tzinfo=tzlocal()),
   'baseModelArn': 'arn:aws:bedrock:us-east-1::foundation-model/meta.llama2-13b-v1:0:4k',
   'baseModelName': '',
   'customizationType': 'FINE_TUNING'},
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[12]:
```
# retrieve the modelArn of the fine-tuned model
fine_tune_job = bedrock.get_custom_model(modelIdentifier=custom_model_name)
custom_model_id = fine_tune_job['modelArn']
```

[13]:
```
output_job_name = "model-customization-job-"+fine_tune_job['jobArn'].split('/')[-1]
output_job_name
```

[13]:
```
'model-customization-job-hj6q5t4jcg91'
```

Visualize Training and Validation Loss

Now that we have completed fine-tuning job, lets visualize our results to see if our job is not underfitting or overfitting.

Download model customization job metrics from S3 and plot the learning curves.

[14]:
```
output_metrics_path = f"fine-tuning-datasets/{output_job_name}"
```

[15]:
```
!mkdir $output_metrics_path
```

[16]:
```
train_metrics_s3_prefix=f'outputs/output-{custom_model_name}/{output_job_name}/training_artifacts/step_wise_training_metrics.csv'
validation_metrics_s3_prefix=f'outputs/output-{custom_model_name}/{output_job_name}/validation_artifacts/post_fine_tuning_validation/validation/validation_metrics.csv'
train_metrics_name='train_metrics.csv'
validation_metrics_name='validation_metrics.csv'
train_file_name_local=output_metrics_path+'/'+train_metrics_name
validation_file_name_local=output_metrics_path+'/'+validation_metrics_name
```

[17]:
```
s3_client.download_file(bucket_name, train_metrics_s3_prefix, train_file_name_local)
s3_client.download_file(bucket_name, validation_metrics_s3_prefix, validation_file_name_local)
```

[18]:
```
train_data = pd.read_csv(train_file_name_local)
'''The training loss is at an iteration level. To calculate loss at the epoch level,
```

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 02_fine-t | Domain- | Domain- | The Ama
d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb
Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio
00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp
Code git Cluster Data Science 3.0 Python 3 4 vCPU + 8 GiB Share

Visualize Training and Validation Loss

Now that we have completed fine-tuning job, lets visualize our results to see if our job is not underfitting or overfitting.

Download model customization job metrics from S3 and plot the learning curves.

[14]:
```
output_metrics_path = f"fine-tuning-datasets/{output_job_name}"
```

[15]:
```
!mkdir $output_metrics_path
```

[16]:
```
train_metrics_s3_prefix=f'outputs/output-{custom_model_name}/{output_job_name}/training_artifacts/step_wise_training_metrics.csv'
validation_metrics_s3_prefix=f'outputs/output-{custom_model_name}/{output_job_name}/validation_artifacts/post_fine_tuning_validation/validation/validation_metrics.csv'
train_metrics_name='train_metrics.csv'
validation_metrics_name='validation_metrics.csv'
train_file_name_local=output_metrics_path+'/'+train_metrics_name
validation_file_name_local=output_metrics_path+'/'+validation_metrics_name
```

[17]:
```
s3_client.download_file(bucket_name, train_metrics_s3_prefix, train_file_name_local)
s3_client.download_file(bucket_name, validation_metrics_s3_prefix, validation_file_name_local)
```

[18]:
```
train_data = pd.read_csv(train_file_name_local)
'''The training loss is at an iteration level. To calculate loss at the epoch level,
    average the iteration-level loss for each epoch'''
train_metrics_epoch=train_data.groupby('epoch_number').mean()
validation_metrics_epoch=pd.read_csv(validation_file_name_local)
```

[19]:
```
plt.plot(validation_metrics_epoch.epoch_number, validation_metrics_epoch.validation_loss,label='validation')
plt.plot(train_metrics_epoch.index, train_metrics_epoch.training_loss,label='training')
plt.title('Training vs Validation Loss')
plt.ylabel('Loss')
plt.xlabel('Epoch')
```

Amazon WorkSpaces

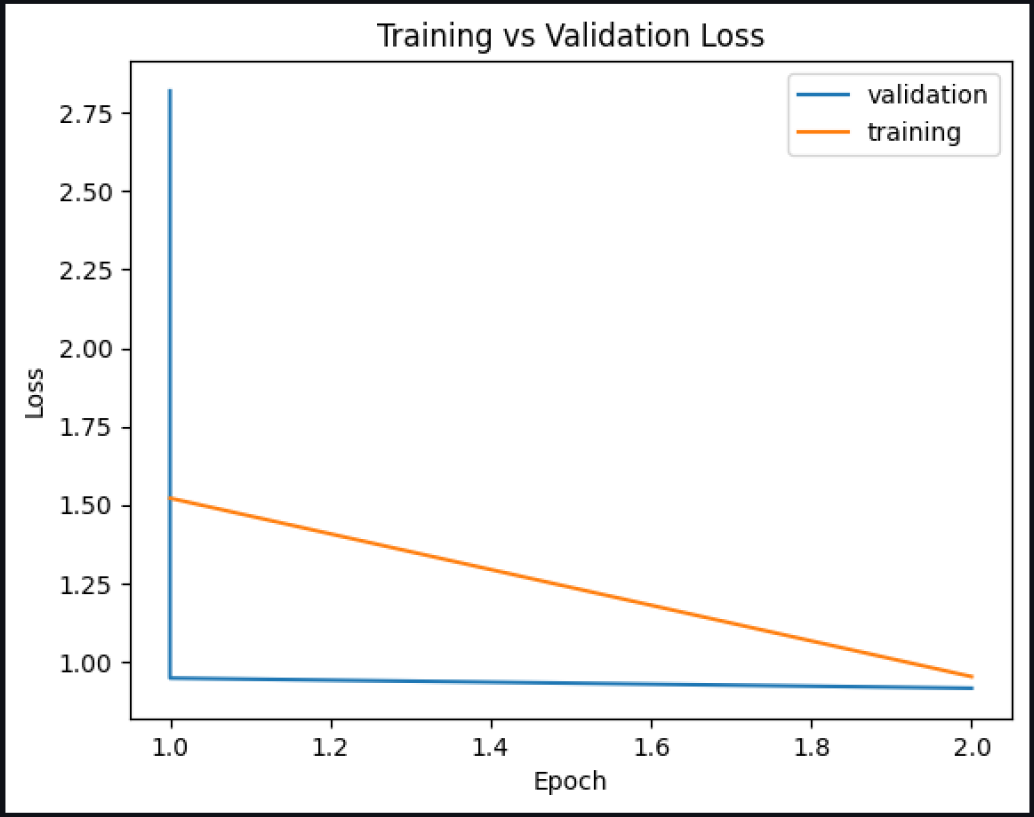
d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
plt.xlabel('Epoch')
plt.legend()
plt.show()
```



Create Provisioned Throughput

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 02_fine-t | Domain- | Domain- | The Ama
d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb
Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio
00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp
Code git Cluster Data Science 3.0 Python 3 4 vCPU + 8 GiB Share

Create Provisioned Throughput

Note: Creating provisioned throughput will take around 20-30mins to complete.

You will need to create provisioned throughput to be able to evaluate the model performance. You can do so through the [console] (https://docs.aws.amazon.com/bedrock/latest/userguide/prov-cap-console.html) or use the following api call.

[20]:
```
# Create the provision throughput job and retrieve the provisioned model id
provisioned_model_id = bedrock.create_provisioned_model_throughput(
     modelUnits=1,
    # create a name for your provisioned throughput model
     provisionedModelName='test-model-v1-001',
     modelId=custom_model_id
    )['provisionedModelArn']
```

[21]:
```
# check provisioned throughput job status
import time
status_provisioning = bedrock.get_provisioned_model_throughput(provisionedModelId = provisioned_model_id)['status']
while status_provisioning == 'Creating':
    time.sleep(60)
    status_provisioning = bedrock.get_provisioned_model_throughput(provisionedModelId=provisioned_model_id)['status']
    print(status_provisioning)
```

```
Creating
Creating
Creating
Creating
Creating
Creating
Creating
Creating
Creating
InService
```

AI/ML and Generative AI Enable | Isengard | Amazon Bedrock | us-east-1 | SageMaker Studio | 02_fine-tuni... (4) - JupyterLab

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/provisioned-throughput

aws | Services | Search [Alt+S] | N. Virginia | Admin/linmicht-Isengard @ 0947-8459-0684

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
  - Prompt management Preview
  - Prompt flows Preview
- **Assessment & deployment**

Amazon Bedrock > Provisioned Throughput

Provisioned Throughput Info

Provisioned Throughput allows you to have dedicated capacity to deploy your models

Overview

Active	Failed	Updating	Inactive
1	0	0	0

Provisioned Throughput (1)

Actions | Purchase Provisioned Throughput

Find Provisioned Throughput

1

Name	Status	Model	Model units	Creation time	Commitment term	End
test-model-v1-001	In service	llama2-finetune-2024-07-12-05-39-59	1	July 12, 2024, 14:26 (UTC+08:00)	No commitment	-

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Invoke the Provisioned Custom Model

Invoke the privisioned custom model.You can replace the follwing prompt_txt with the prompts that are more similar to your fine-tuning dataset, this helps to check whether the fine-tuned model is performing as you expected.

Note: Please make sure your provisioned throughput job status becomes InService before proceeding.

[22]:
```
# Provide the prompt text
test_file_path = f'{data_folder}/{test_file_name}'
with open(test_file_path) as f:
    lines = f.read().splitlines()
```

[23]:
```
test_prompt = json.loads(lines[0])['prompt']
reference_summary = json.loads(lines[0])['completion']
print(test_prompt)
print()
print(reference_summary)
```

```
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

instruction:

Summarize the news article provided below.

input:

Pep Guardiola was left purring at Bayern Munich's first-half performance against Porto as the club put a turbulent week behind them to storm in to the Champions League semi-
finals. A 3-1 first leg defeat in Portugal last week followed by the dramatic resignation of long-serving club doctor Hans-Wilhelm Muller-Wohlfahrt had turned the spotlight
on Guardiola heading in to the tie. But the Bayern boss reminded the watching world why he is considered by many to be the best manager in the business as his side raced in
to a 5-0 lead before half-time on Tuesday night. Bayern Munich players celebrate in front of their fans after the 6-1 demolition of Porto on Tuesday night . Pep Guardiola lo
oks animated as he issues instructions from the touchline at the Allianz Arena . 'It's easy to love my players now,' said Guardiola. 'I am the coach of extraordinary player
```

```
test_file_path = f'{data_folder}/{test_file_name}'
with open(test_file_path) as f:
    lines = f.read().splitlines()
```

[23]:
```
test_prompt = json.loads(lines[0])['prompt']
reference_summary = json.loads(lines[0])['completion']
print(test_prompt)
print()
print(reference_summary)
```

```
Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

instruction:

Summarize the news article provided below.

input:

Pep Guardiola was left purring at Bayern Munich's first-half performance against Porto as the club put a turbulent week behind them to storm in to the Champions League semi-
finals. A 3-1 first leg defeat in Portugal last week followed by the dramatic resignation of long-serving club doctor Hans-Wilhelm Muller-Wohlfahrt had turned the spotlight
on Guardiola heading in to the tie. But the Bayern boss reminded the watching world why he is considered by many to be the best manager in the business as his side raced in
to a 5-0 lead before half-time on Tuesday night. Bayern Munich players celebrate in front of their fans after the 6-1 demolition of Porto on Tuesday night . Pep Guardiola lo
oks animated as he issues instructions from the touchline at the Allianz Arena . 'It's easy to love my players now,' said Guardiola. 'I am the coach of extraordinary player
s. I didn't expect such a first half.' Typical of the man though, Guardiola found faults in Bayern's second half performance as Porto pulled back to 5-1 before Xabi Alonso a
dded a late sixth. 'We can still play better,' he added. 'We lost the ball here and there and we can improve on that. We had a few problems in the second half. 'They had onl
y a few chances but you should never relax in the Champions League.' Guardiola talks to captain Philipp Lahm (third right) as Mario Gotze (far left), Thiago and Thomas Mulle
r listen in . It is believed that Muller-Wohlfahrt and his team resigned as they felt they were blamed for the first-leg defeat, though neither the doctor, the club or Guard
iola have said as much in public. Bayern find out their semi-final opponents on Thursday with Guardiola's former club Barcelona already through to the last four and the winn
ers of Real Madrid v Atletico Madrid and Monaco v Juventus to join them in the draw. Guardiola celebrates as Robert Lewandowski scores Bayern's fifth goal in the first half
.

response:

Bayern Munich beat Porto 6-1 at the Allianz Arena on Tuesday night .
Result gave them a 7-4 aggregate victory in Champions League last eight .
Bayern manager Pep Guardiola hailed his players after the match .
READ: Luis Enrique 'happy' to see Pep Guardiola prove doubters wrong .
```

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-workshop-20240712/03_Model_customization/02_fine-tuning_llama2.ipynb

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code git Cluster Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB Share

Construct model input following the format needed by Llama2 model following instructions here. Please pay attention to the "Model invocation request body field" section

[24]:
```
body = json.dumps({
    "prompt": test_prompt,
    # specify the parameters as needed
    "max_gen_len": 200,
    "temperature": 0.4,
    "top_p": 0.3,
})

# provide the modelId of the provisioned custom model
modelId = provisioned_model_id
accept = 'application/json'
contentType = 'application/json'

# invoke the provisioned custom model
response = bedrock_runtime.invoke_model(body=body, modelId=modelId, accept=accept, contentType=contentType)

response_body = json.loads(response.get('body').read())
print(response_body)
```

```
{'generation': " response:\n\nBayern Munich put a turbulent week behind them to storm in to the Champions League semi-finals .\nPep Guardiola's side raced in to a 5-0 lead b
efore half-time on Tuesday night .\nGuardiola reminded the watching world why he is considered by many to be the best manager in the business .", 'prompt_token_count': 531,
'generation_token_count': 85, 'stop_reason': 'stop'}
```

Clean up

Warning: Please make sure to delete providsioned throughput with the following code as there will be cost incurred if its left in running state, even if you are not using it.

[25]:
```
# delete the provisioned throughput
bedrock.delete_provisioned_model_throughput(provisionedModelId=provisioned_model_id)
```

[25]:
```
{'ResponseMetadata': {'RequestId': '2d0ec641-3779-4b8b-8812-bdb231931047',
```

Simple 2 6 main Python 3 (Data Science 3.0) | Disconnected Kernel: Dead | Instance MEM Cookie Preferences Mode: Command Ln 1, Col 1 02_fine-tuning_llama2.ipynb 0

Type here to search ENG 1:13 PM

Clean up

Warning: Please make sure to delete providsioned throughput with the following code as there will be cost incurred if its left in running state, even if you are not using it.

[25]:
```
# delete the provisioned throughput
bedrock.delete_provisioned_model_throughput(provisionedModelId=provisioned_model_id)
```

[25]:
```
{'ResponseMetadata': {'RequestId': '2d0ec641-3779-4b8b-8812-bdb231931047',
  'HTTPStatusCode': 200,
  'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 06:38:00 GMT',
   'content-type': 'application/json',
   'content-length': '2',
   'connection': 'keep-alive',
   'x-amzn-requestid': '2d0ec641-3779-4b8b-8812-bdb231931047'},
  'RetryAttempts': 0}}
```

Note: Please finish up the cleaning process by running 04_cleanup.ipynb to clean up the other resources.

AI/ML and Generative AI Enable | Isengard | Amazon Bedrock | us-east-1 | SageMaker Studio | 02_fine-tuni... (4) - JupyterLab

us-east-1.console.aws.amazon.com/bedrock/home?region=us-east-1#/provisioned-throughput

aws | Services | Search [Alt+S] | N. Virginia | Admin/linmicht-Isengard @ 0947-8459-0684

Amazon Bedrock

- **Getting started**
  - Overview
  - Examples
  - Providers
- **Foundation models**
  - Base models
  - Custom models
  - Imported models Preview
- **Playgrounds**
  - Chat
  - Text
  - Image
- **Safeguards**
  - Guardrails
  - Watermark detection
- **Builder tools**
  - Knowledge bases
  - Agents
  - Prompt management Preview
  - Prompt flows Preview
- **Assessment & deployment**

Amazon Bedrock > Provisioned Throughput

Provisioned Throughput Info

Provisioned Throughput allows you to have dedicated capacity to deploy your models

Overview

Active	Failed	Updating	Inactive
0	0	0	0

Provisioned Throughput (0)

Actions | Purchase Provisioned Throughput

Find Provisioned Throughput

1

Name	Status	Model	Model units	Creation time	Commitment term	End time

No Provisioned Throughput

There are currently no resources.

Purchase Provisioned Throughput

Labs

- Bedrock Console: text generation with Titan Text
- Bedrock Console: news summary with Command Light
- Bedrock SDK: news summary with Llama 2
- Bedrock SDK: image generation with Titan Image Generator
- SageMaker JumpStart console: text generation with Mistral
- SageMaker JumpStart SDK: dialog summary with LLama 3

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

https://github.com/aws-samples/amazon-bedrock-samples/tree/main/bedrock-fine-tuning/amazon-titan-image-generator

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/1-TIGFT-cu...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lit | 02_fine-tuning_llama2. | 03_continued_pretraini | 1-TIGFT-customization- | 2-TIGFT-provisioned-th

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Filter files by name

/ ··· / bedrock-fine-tuning / amazon-titan-image-generator /

Name	Last Modified
data	a day ago
prompts	15 hours ago
1-TIGFT-customiz...	an hour ago
2-TIGFT-provisio...	an hour ago
README.md	a day ago

Fine-Tuning Amazon Titan Image Generator G1

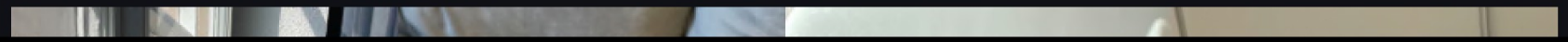
> ☝ This notebook has been tested with the `SageMaker Data Science 3.0` kernel in Amazon SageMaker Studio.

In this notebook, we will show how to fine tune Amazon Titan Image Generator G1 on Amazon Bedrock model.

We will teach our model to recognize two new classes:

Ron the dog

and **Smila the cat**



Simple | 2 | 6 | main | Python 3 (Data Science 3.0) | Disconnected | Kernel: Dead | Instance MEM | Cookie Preferences | Mode: Command | Ln 1, Col 1 | 1-TIGFT-customization-job.ipynb | 0

Type here to search | ENG | 12:30 PM

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/data/ron_ft...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipy | 01_fine-tuni | 02_fine-tuni | 03_continue | 1-TIGFT-cust | ron_01.jpg | ron_ft_1.png | ron_ft_2.png | 2-TIGFT-pro

Filter files by name

/ ··· / amazon-titan-image-generator / data /

Name	Last Modified
ron_16.jpg	a day ago
ron_17.jpg	a day ago
ron_18.jpg	a day ago
ron_19.jpg	a day ago
ron_20.jpg	a day ago
ron_21.jpg	a day ago
ron_22.jpg	a day ago
ron_23.jpg	a day ago
ron_24.jpg	a day ago
ron_25.jpg	a day ago
ron_26.jpg	a day ago
ron_27.jpg	a day ago
ron_28.jpg	a day ago
ron_29.jpg	a day ago
ron_30.jpg	a day ago
ron_ft_1.png	a day ago
ron_ft_2.png	a day ago
smila_01.jpg	a day ago
smila_02.jpg	a day ago
smila_03.jpg	a day ago
smila_04.jpg	a day ago
smila_05.jpg	a day ago
smila_06.jpg	a day ago

Simple | 2 | 6 | main | Cookie Preferences | ron_ft_1.png | 0

Type here to search | ENG | 12:50 PM

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/data/smila...

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

default-20240221t112169 / Personal Studio

Filter files by name

/ ··· / amazon-titan-image-generator / data /

Name	Last Modified
smila_10.jpg	a day ago
smila_11.jpg	a day ago
smila_12.jpg	a day ago
smila_13.jpg	a day ago
smila_14.jpg	a day ago
smila_15.jpg	a day ago
smila_16.jpg	a day ago
smila_17.jpg	a day ago
smila_18.jpg	a day ago
smila_19.jpg	a day ago
smila_20.jpg	a day ago
smila_21.jpg	a day ago
smila_22.jpg	a day ago
smila_23.jpg	a day ago
smila_24.jpg	a day ago
smila_25.jpg	a day ago
smila_26.jpg	a day ago
smila_27.jpg	a day ago
smila_28.jpg	a day ago
smila_29.jpg	a day ago
smila_30.jpg	a day ago
smila_ft_1.png	a day ago
smila_ft_2.png	a day ago

00_s 01_fi 02_fi 03_c 1-TI(ron_ ron_ ron_ ron_ smila smila smila smila 2-TI(

Simple 2 6 main

Cookie Preferences smila_ft_1.png 0

Type here to search

ENG 12:52 PM

Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
default-20240221t112169 / Personal Studio
00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp
Code git Cluster Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB Share

Fine-Tuning Amazon Titan Image Generator G1

> ☝️ This notebook has been tested with the `SageMaker Data Science 3.0` kernel in Amazon SageMaker Studio.

In this notebook, we will show how to fine tune Amazon Titan Image Generator G1 on Amazon Bedrock model.

We will teach our model to recognize two new classes:

Ron the dog

Simple 2 6 main Python 3 (Data Science 3.0) | Disconnected Kernel: Dead | Instance MEM Cookie Preferences Mode: Command Ln 1, Col 1 1-TIGFT-customization-job.ipynb 0
Type here to search
ENG 12:30 PM 7/13/2024

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ip | 02_fine-tuning_llama2.ipy | 03_continued_pretraining_ | 1-TIGFT-customization-job | 2-TIGFT-provisioned-throu | llama-3-finetuning.ipynb

Code git Cluster Data Science 3.0 Python 3 4 vCPU + 8 GiB Share

and **Smila the cat**

[2]:
```
!pip install --upgrade --force-reinstall --no-cache boto3
!pip install --upgrade --force-reinstall --no-cache botocore
!pip install --upgrade --force-reinstall --no-cache awscli
```

```
Collecting boto3
  Downloading boto3-1.34.143-py3-none-any.whl.metadata (6.6 kB)
Collecting botocore<1.35.0,>=1.34.143 (from boto3)
  Downloading botocore-1.34.143-py3-none-any.whl.metadata (5.7 kB)
Collecting jmespath<2.0.0,>=0.7.1 (from boto3)
```

Simple 0 0 main No Kernel | Unknown No kernel: Unknown | Instance MEM Cookie Preferences Mode: Command Ln 1, Col 1 1-TIGFT-customization-job.ipynb 0

Type here to search ENG 3:46 PM 7/13/2024

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/1-TIGFT-cu...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Data preparation

- To fine-tune a text-to-image or image-to-embedding model, prepare a training dataset by create a JSONL file with multiple JSON lines.
- Validation datasets are not supported.
- Each JSON line is a sample containing an image-ref, the Amazon S3 URI for an image, and a caption that could be a prompt for the image.

The images must be in JPEG or PNG format.

```
{"image-ref": "s3://bucket/path/to/image001.png", "caption": "<prompt text>"}
{"image-ref": "s3://bucket/path/to/image002.png", "caption": "<prompt text>"}
{"image-ref": "s3://bucket/path/to/image003.png", "caption": "<prompt text>"}
```

The following is an example item:

```
{"image-ref": "s3://my-bucket/my-pets/cat.png", "caption": "an orange cat with white spots"}
```

Locate your sample json file

We are going to use a json file which contains the image captions in the following format:

```
{
    "imagefile":"caption",
    "imagefile":"caption",
    "imagefile":"caption"
}
```

[5]:
```
raw_data_file = "prompts/captions.json"

with open(raw_data_file, 'r') as file:
    raw_data = json.load(file)
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/1-TIGFT-cu...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[5]:
```
raw_data_file = "prompts/captions.json"

with open(raw_data_file, 'r') as file:
    raw_data = json.load(file)

print(raw_data)
```

```
{'ron_01.jpg': 'Ron the dog laying on a white dog bed', 'ron_02.jpg': 'Ron the dog sitting on a tile floor, possibly in a kitchen or living room', 'ron_03.jpg': 'Ron the dog
laying on a car seat', 'ron_04.jpg': 'Ron the dog looking directly at the camera. He is laying down on a wooden floor.', 'ron_05.jpg': 'Ron the dog sitting on a couch, looki
ng at the camera with a smile on his face.', 'ron_06.jpg': 'Ron the dog lying on a couch, covered in a blanket.', 'ron_07.jpg': 'Ron the dog sleeping or resting, with his he
ad on the stuffed animal.', 'ron_08.jpg': 'Ron the dog sitting on a box with a red leash.', 'ron_09.jpg': 'Ron the dog sitting in the snow, wearing a red collar.', 'ron_10.j
pg': 'Ron the dog lying on a couch, chewing on a tennis shoe.', 'ron_11.jpg': 'Ron the dog sitting on a sandy beach, wearing a blue collar.', 'ron_12.jpg': 'Ron the dog wear
ing a yellow raincoat and is sitting on the floor.', 'ron_13.jpg': 'Ron the dog looking at the camera in front of a table.', 'ron_14.jpg': 'Ron the dog lying on a white couc
h, playing with a pink toy.', 'ron_15.jpg': 'Ron the dog sleeping on a white couch.', 'ron_16.jpg': 'Ron the dog laying on a couch.', 'ron_17.jpg': 'Ron the dog laying on a
couch.', 'ron_18.jpg': 'Ron the dog playing with a ball.', 'ron_19.jpg': 'Ron the dog sitting on a couch.', 'ron_20.jpg': 'Ron the dog holding a large stick in his mouth.',
'ron_21.jpg': 'Ron the dog sitting on a brick floor.', 'ron_22.jpg': 'Ron the dog laying on a blanket.', 'ron_23.jpg': 'Ron the dog sitting on a couch.', 'ron_24.jpg': 'Ron
the dog standing on a tile floor, wearing a leash.', 'ron_25.jpg': 'Ron the dog wearing a blue and white jacket and is sitting on the grass.', 'ron_26.jpg': 'Ron the dog sta
nding in front of a large brown teddy bear.', 'ron_27.jpg': 'Ron the dog is standing in a kitchen, looking at the camera.', 'ron_28.jpg': 'Ron the dog sitting on a couch.',
'ron_29.jpg': 'Ron the dog wearing a festive sweater with snowflakes on it.', 'ron_30.jpg': 'Ron the dog wearing a red sweater with a white snowflake pattern.', 'smila_01.jp
g': 'Smila the cat lying on a couch.', 'smila_02.jpg': 'Smila the cat sitting next to the window next to a statue cat.', 'smila_03.jpg': 'Smila the cat sitting inside a colo
rful cat tunnel.', 'smila_05.jpg': 'Smila the cat lying on a sofa.', 'smila_06.jpg': 'Smila the cat sitting on a couch.', 'smila_07.jpg': 'Smila the cat sitting on a black c
upboard.', 'smila_08.jpg': 'Smila the cat is sitting inside a cat carrier.', 'smila_09.jpg': 'Smila the cat is sitting on the floor next to a door.', 'smila_10.jpg': 'Smila
the cat is sitting in a pet carrier.', 'smila_12.jpg': 'Smila the cat lying on the sofa.', 'smila_13.jpg': 'Smila the cat lying on a grey couch.', 'smila_14.jpg': 'Smila the
cat sitting inside a colorful tunnel.', 'smila_15.jpg': 'Smila the cat sitting in a cat transporter.', 'smila_16.jpg': 'Smila the cat is sitting next to a window.', 'smila_1
7.jpg': 'Smila the cat laying on a white table.', 'smila_18.jpg': 'Smila the cat lying on a cat scratcher.', 'smila_19.jpg': 'Smila the cat sitting on a wicker basket in the
bathroom.', 'smila_20.jpg': 'Smila the cat sitting on a couch, eating yogurt.', 'smila_21.jpg': 'Smila the cat sitting on a wicker basket in a bathroom.', 'smila_22.jpg': 'S
mila the cat laying on a bed.', 'smila_23.jpg': 'Smila the cat sitting on a cat bed.', 'smila_24.jpg': 'Smila the cat sitting on a white bed.', 'smila_25.jpg': 'Smila the ca
t laying on a cat scratcher.', 'smila_26.jpg': 'Smila the cat is laying on a gray sofa, surrounded by pillows.', 'smila_27.jpg': 'Smila the cat sitting on a couch.', 'smila_
28.jpg': 'Smila the cat laying on a gray sofa, looking at the camera.', 'smila_29.jpg': 'Smila the cat sitting on a bench, staring at the camera.', 'smila_30.jpg': 'Smila th
e cat lying on a white bed.'}
```

Create the dataset file and upload the images to Amazon S3

Create the `jsonl` file with the images prompt based on the image's s3 path.

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/1-TIGFT-cu...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[11]:
```
ts = datetime.datetime.now().strftime("%Y-%m-%d-%H-%M-%S")

# Select the foundation model you want to customize (you can find this from the "modelId" from listed foundation model above)
base_model_id = "amazon.titan-image-generator-v1:0"

# Select the customization type from "FINE_TUNING" or "CONTINUED_PRE_TRAINING".
customization_type = "FINE_TUNING"

# Specify the roleArn for your customization job
customization_role = role_arn

# Create a customization job name
customization_job_name = f"image-gen-ft-{ts}"

# Create a customized model name for your fine-tuned Llama2 model
custom_model_name = f"image-gen-ft-{ts}"

# Define the hyperparameters for fine-tuning Llama2 model
hyper_parameters = {
    "stepCount": "8000",
    "batchSize": "8",
    "learningRate": "0.00001",
}

# Specify your data path for training, validation(optional) and output
s3_train_uri = s3_bucket_path + "/" + output_file
training_data_config = {"s3Uri": s3_train_uri}


output_data_config = {"s3Uri": f's3://{bucket_name}/outputs/output-{custom_model_name}'}

# Create the customization job
bedrock_client.create_model_customization_job(
    customizationType=customization_type,
    jobName=customization_job_name,
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/1-TIGFT-cu...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
# Create the customization job
bedrock_client.create_model_customization_job(
    customizationType=customization_type,
    jobName=customization_job_name,
    customModelName=custom_model_name,
    roleArn=customization_role,
    baseModelIdentifier=base_model_id,
    hyperParameters=hyper_parameters,
    trainingDataConfig=training_data_config,
    outputDataConfig=output_data_config
)
```

[11]:
```
{'ResponseMetadata': {'RequestId': 'c1c09ec1-3658-4593-b986-e13cb80ae3e9',
  'HTTPStatusCode': 201,
  'HTTPHeaders': {'date': 'Fri, 12 Jul 2024 13:43:03 GMT',
   'content-type': 'application/json',
   'content-length': '122',
   'connection': 'keep-alive',
   'x-amzn-requestid': 'c1c09ec1-3658-4593-b986-e13cb80ae3e9'},
  'RetryAttempts': 0},
 'jobArn': 'arn:aws:bedrock:us-east-1:094784590684:model-customization-job/amazon.titan-image-generator-v1:0/hw5vebkzf8c1'}
```

Waiting until customization job is completed

Once the customization job is finished, you can check your existing custom model(s) and retrieve the modelArn of your fine-tuned model.

Warning: The model customization job can take hours to run. With 5000 steps, 0.000001 learning rate, 64 of batch size and 60 images, it takes around 4 hours to complete

[ ]:
```
# check model customization status
status = bedrock_client.list_model_customization_jobs(
    nameContains=customization_job_name
)["modelCustomizationJobSummaries"][0]["status"]
```

Amazon WorkSpaces

Search fo | Search fo | DevRel - | Isengard | Amazon | Amazon | 2-TIGFT- | Domain- | Domain- | The Ama

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/2-TIGFT-pr...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

Image Generation with Fine-tuned Amazon Titan Image Generator G1 model

> ☝️ This notebook has been tested with the `SageMaker Data Science 3.0` kernel in Amazon SageMaker Studio.

In this notebook, we will show how to invoke a fine-tuned Amazon Titan Image Generator G1 on Amazon Bedrock trained with Ron the dog and Smila the cat

Pre-requisites

Import needed libraries, instantiate the boto3 clients and setup any style tag that was added during model fine-tuning

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/2-TIGFT-pr...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

[9]:
```python
def decode_base64_image(img_b64):
    return Image.open(io.BytesIO(base64.b64decode(img_b64)))

def invoke_model_tgi(prompt, seed, model_arn):
    img_gen_conf = {
        "cfgScale": 8,
        "seed": seed,
        "quality": "standard",
        "width": 1024,
        "height": 1024,
        "numberOfImages": 2
    }

    body = json.dumps({
        "textToImageParams": {
            "text": prompt
        },
        "taskType": "TEXT_IMAGE",
        "imageGenerationConfig": img_gen_conf
    })

    response = bedrock_runtime.invoke_model(
        body=body,
        modelId=model_arn,
        accept="application/json",
        contentType="application/json"
    )

    response_body = json.loads(response.get("body").read())
    return [decode_base64_image(img) for img in response_body["images"]]

def compare_models_outputs(prompts, seeds, model_arn_base, model_arn_custom):
    for prompt in prompts:
        for seed in seeds:
            # Invoke the base model
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/2-TIGFT-pr...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
    response = bedrock_runtime.invoke_model(
        body=body,
        modelId=model_arn,
        accept="application/json",
        contentType="application/json"
    )

    response_body = json.loads(response.get("body").read())
    return [decode_base64_image(img) for img in response_body["images"]]

def compare_models_outputs(prompts, seeds, model_arn_base, model_arn_custom):
    for prompt in prompts:
        for seed in seeds:
            # Invoke the base model
            base_images = invoke_model_tgi(prompt, seed, model_arn_base)

            # Invoke the customized model
            custom_images = invoke_model_tgi(prompt, seed, model_arn_custom)

            # Display the images for comparison
            plot_images(base_images, custom_images, prompt, seed)

def plot_images(base_images, custom_images, prompt, seed):
    fig, axes = plt.subplots(1, 2, figsize=(10, 5))

    axes[0].imshow(np.array(base_images[0]))
    axes[0].set_title('Base Model\nSeed: {}'.format(seed))
    axes[0].axis('off')

    axes[1].imshow(np.array(custom_images[0]))
    axes[1].set_title('Custom Model\nSeed: {}'.format(seed))
    axes[1].axis('off')
    print("Prompt:{}\n".format(prompt))
    plt.show()
```

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/2-TIGFT-pr...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
Prompt:'Ron the dog' is staring at the camera.
```

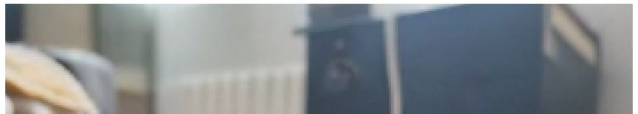
Base Model
Seed: 3000

Custom Model
Seed: 3000

```
Prompt:'Smila the cat' is staring at the camera.
```

Base Model
Seed: 0

Custom Model
Seed: 0



Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/2-TIGFT-pr...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

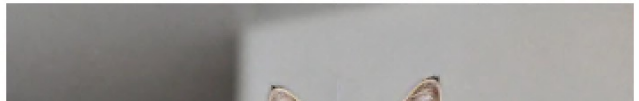
```
Prompt:'Smila the cat' is staring at the camera.
```

Base Model
Seed: 0

Custom Model
Seed: 0

```
Prompt:'Smila the cat' is staring at the camera.
```

Base Model
Seed: 3000



Custom Model
Seed: 3000

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/2-TIGFT-pr...

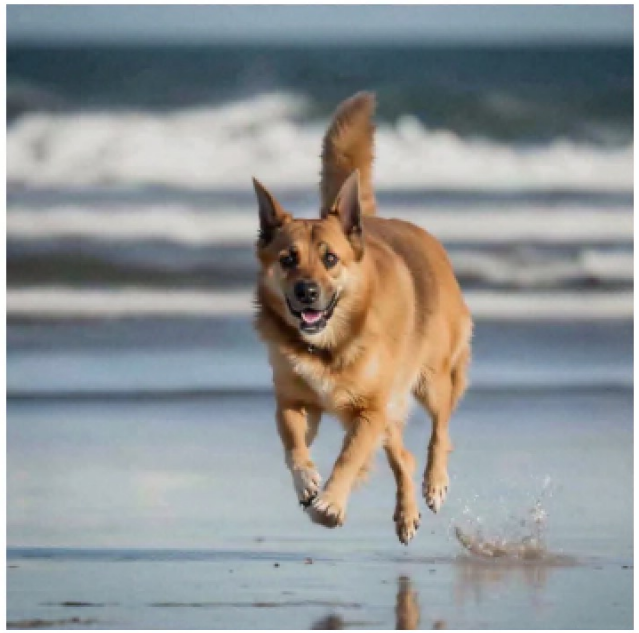
Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
Prompt:'Ron the dog' is running on the beach.
```

Base Model
Seed: 3000



Custom Model
Seed: 3000

```
Prompt:'Smila the cat' is running on the grass.
```

Base Model
Seed: 0

Custom Model
Seed: 0

Amazon WorkSpaces

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-bedrock-samples/bedrock-fine-tuning/amazon-titan-image-generator/2-TIGFT-pr...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita | 1-TIGFT-customization-job.ipy | 2-TIGFT-provisioned-throughp

Code | git | Cluster | Data Science 3.0 | Python 3 | 4 vCPU + 8 GiB | Share

```
Prompt:'Smila the cat' is running on the grass.
```

[ ]:

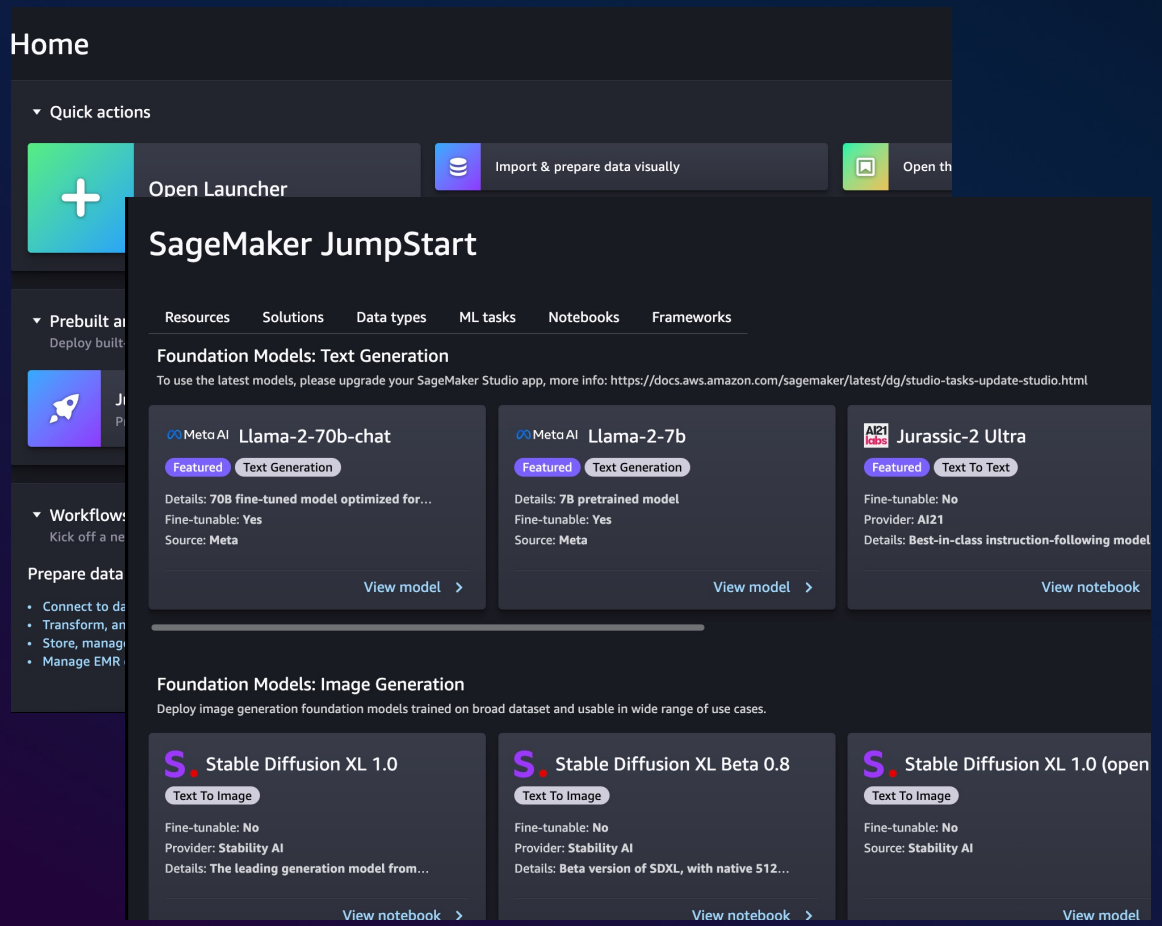
Clean Up

To avoid unnecessary costs, let's now delete the provisioned throughput model

Accelerate FM development with Amazon SageMaker JumpStart

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

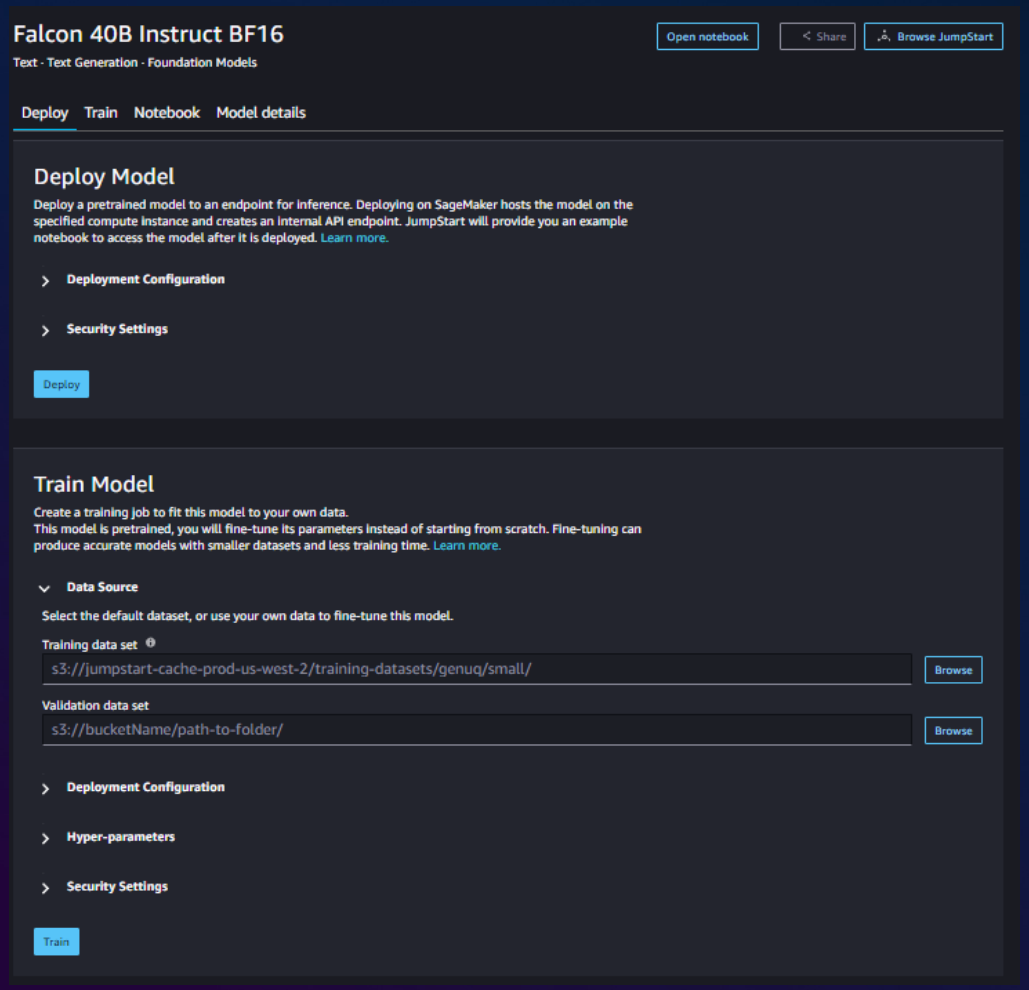
Discover foundation models from multiple providers



- Browse in SageMaker Studio
- Search for specific model or provider from search bar
- View model-specific information

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Review model details and take action



Details from model provider:

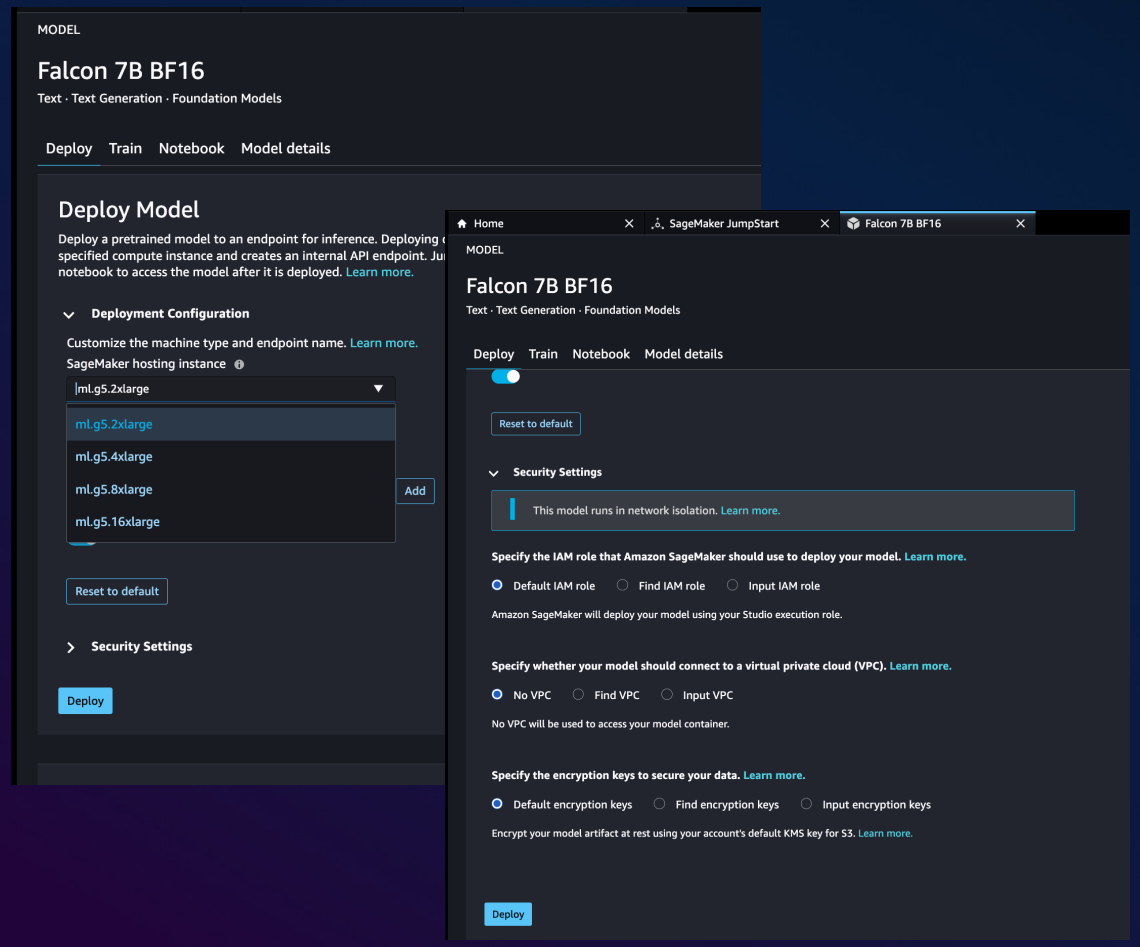
- Model size and description
- License info
- Use cases and how-to use model

Take action:

- Deploy
- Train
- View API snippet

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

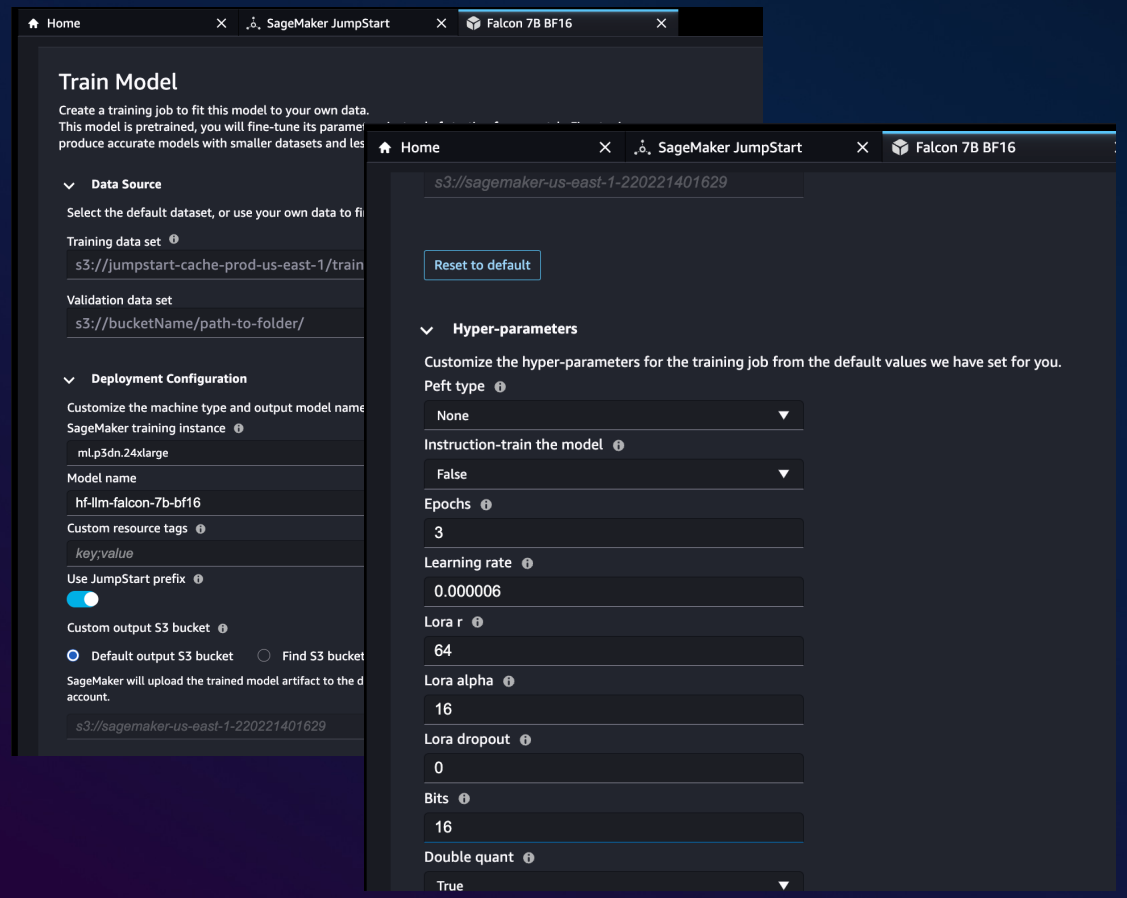
Deploy to SageMaker with just a few clicks



- One-click deploy with defaults
- Configure for cost, throughput, and latency
- Control security and VPC settings

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Customize models with your data



- Fine-tune open and closed models
- Store model weights in your S3 bucket
- Control hyper-parameters
- Choose preferred instance type

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scale using the SageMaker SDK

MODEL

Falcon 7B BF16

Text · Text Generation · Foundation Models

Open notebook

Deploy Train Notebo

Hyper-parameters

Security Settings

Train

Run in notebo

Use the model programm

Open notebook

Markdown Cluster Data Science 2.0 Python 3 2 vCPU +

It is recommended that Instruct models should be used without fine-tuning and base models should be fine-tuned further on the task.

Limitations:

- Falcon models are mostly trained on English data and may not generalize to other languages.
- Falcon carries the stereotypes and biases commonly encountered online and in the training data. Hence, it is recommended t develop guardrails and to take appropriate precautions for any production use. This is a raw, pretrained model, which should further finetuned for most usecases.

[ ]:
```
def query_endpoint(payload):
    """Query endpoint and print the response"""
    response = predictor.predict(payload)
    print(f"\033[1m Input:\033[0m {payload['inputs']}")
    print(f"\033[1m Output:\033[0m {response[0]['generated_text']}")
```

[ ]:
```
# Code generation
payload = {"inputs": "Write a program to compute factorial in python:", "parameters":{"max_new_tokens": 2
query_endpoint(payload)
```

[ ]:
```
payload = {
    "inputs": "Building a website can be done in 10 simple steps:",
    "parameters":{
        "max_new_tokens": 110,
        "no_repeat_ngram_size": 3
        }
}
query_endpoint(payload)
```

[ ]:
```
# Translation
payload = {
    "inputs": """Translate English to French:

    sea otter => loutre de mer
```

- Automate using the SageMaker APIs
- View example code for each model
- Control deployment parameters

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

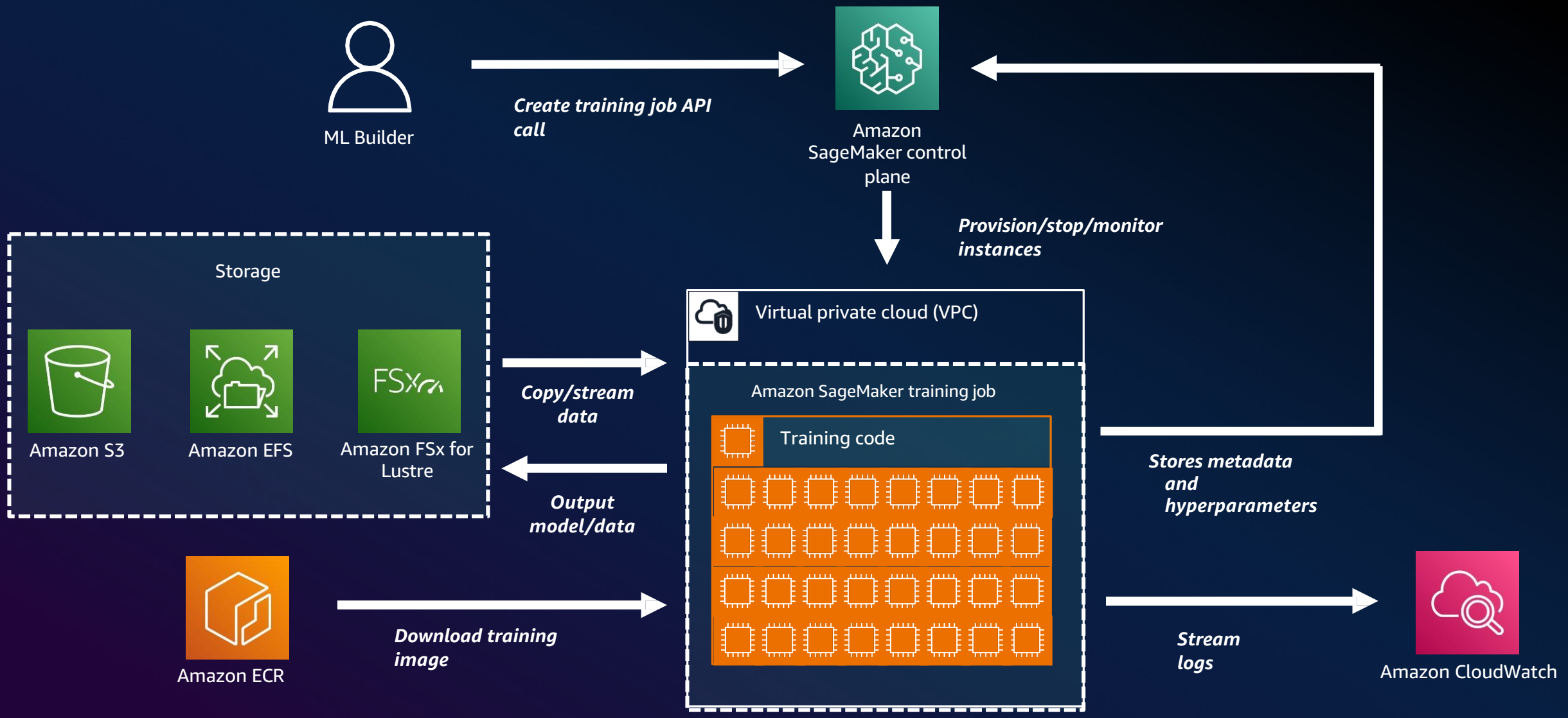
Optimize training of foundation models on Amazon SageMaker

Michael Lin

Sr. Solutions Architect
Amazon Web Services

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker ephemeral training clusters

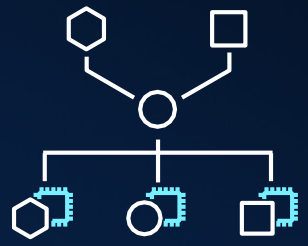


© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon SageMaker distributed training libraries

Data parallelism

Scale out training clusters with near-linear efficiency, optimized for AWS networking and instance topology



Model parallelism

Train models too large to fit within GPU memory, with automated model splitting and sophisticated pipeline scheduling

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Optimize training with Amazon SageMaker data parallel

Data parallelism

Scale out training clusters with near-linear efficiency, optimized for AWS networking and instance topology

Support for popular ML framework APIs

Re-use existing APIs such as Horovod and PyTorch DistributedDataParallel (DDP)

Reduced training time

~25% faster with synchronization across GPUs (as tested with BERT)

Minimal code change

See Scrum Master (SM) developer guide for PyTorch and TensorFlow instructions – or use Hugging Face Trainer API scripts with no code changes at all

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

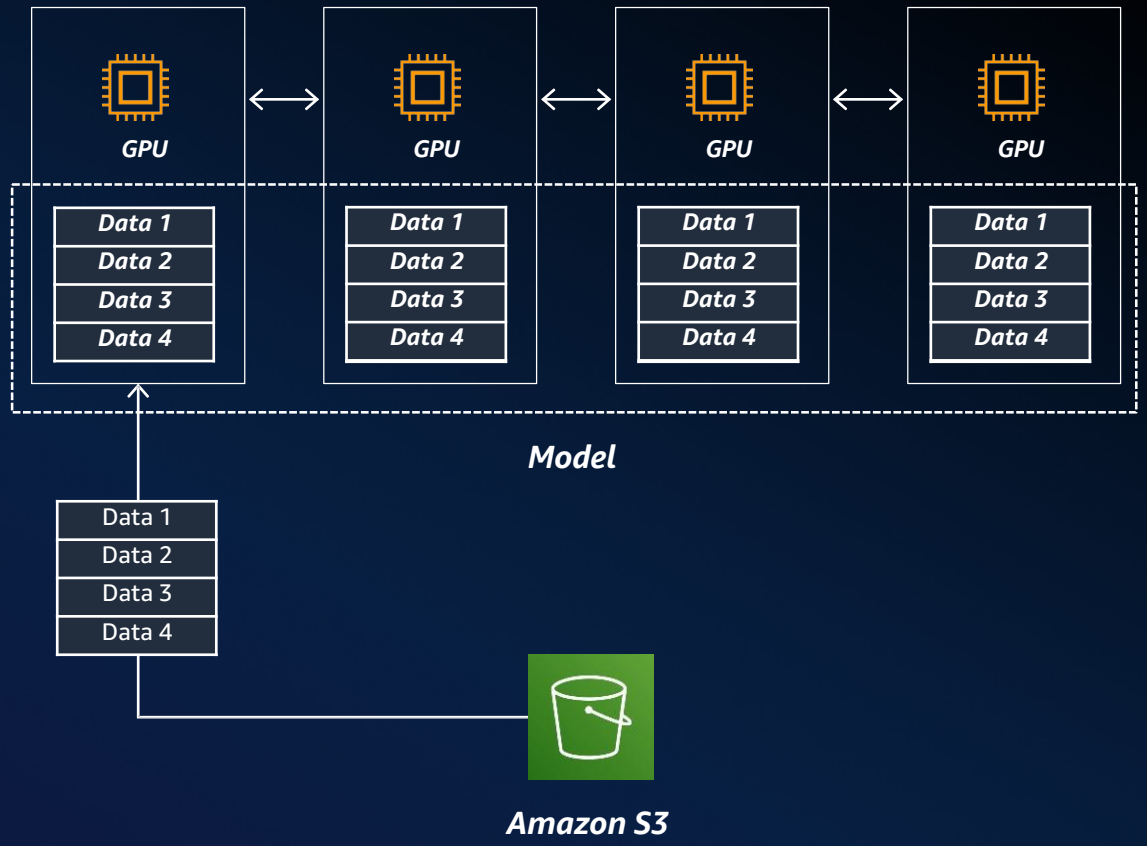
Streamline distributed training with model parallel

Model parallelism on Amazon SageMaker enables interleave pipeline execution to stabilize GPU utilization

Split mini-batches into N “micro-batches”

Feed micro-batches sequentially, but process them to keep GPU utilization more even

Minimize “idle” time on GPUs



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is fine-tuning?

Fine-tuning basics

- Use pre-trained model
- Update model's weights
- General-purpose base model into a specialized model
- Models can achieve better performance while requiring (far) fewer manually labeled examples
- Smaller (fine-tuned) model can often outperform larger (more expensive) models

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Parameter-Efficient Fine-Tuning (PEFT)

Parameter training

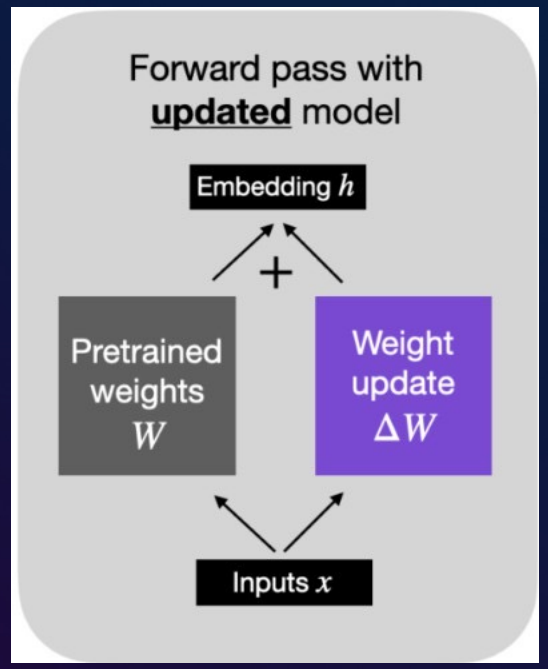
- A novel approach for fine-tuning
- Open-source library from Hugging Face
- Fine-tune a small number of (extra) model parameters
- State-of-the-art PEFT achieve full fine-tuning performance
- Supported methods
  - LoRA and QLoRA - are most widely used and effective
  - Prefix-tuning
  - AdaLoRA
  - …

Github - HuggingFace - PEFT

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

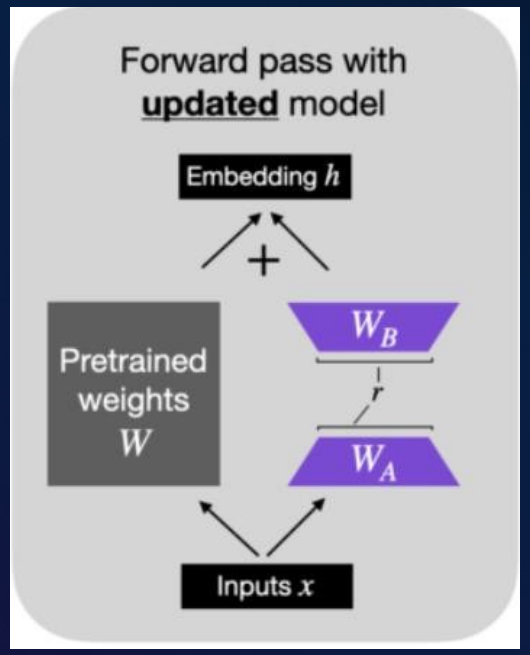
Low-Rank Adaptation (LoRA)

Training method that accelerates the training of large models while consuming less memory



$$W \leftarrow W + \Delta W$$

Where, $W \in \mathbb{R}^{m \times n}$



$$\Delta W = WA \,.\, WBT$$

Where, $W_A \in \mathbb{R}^{m \times k}$ & $WB \in \mathbb{R}^{n \times k}$

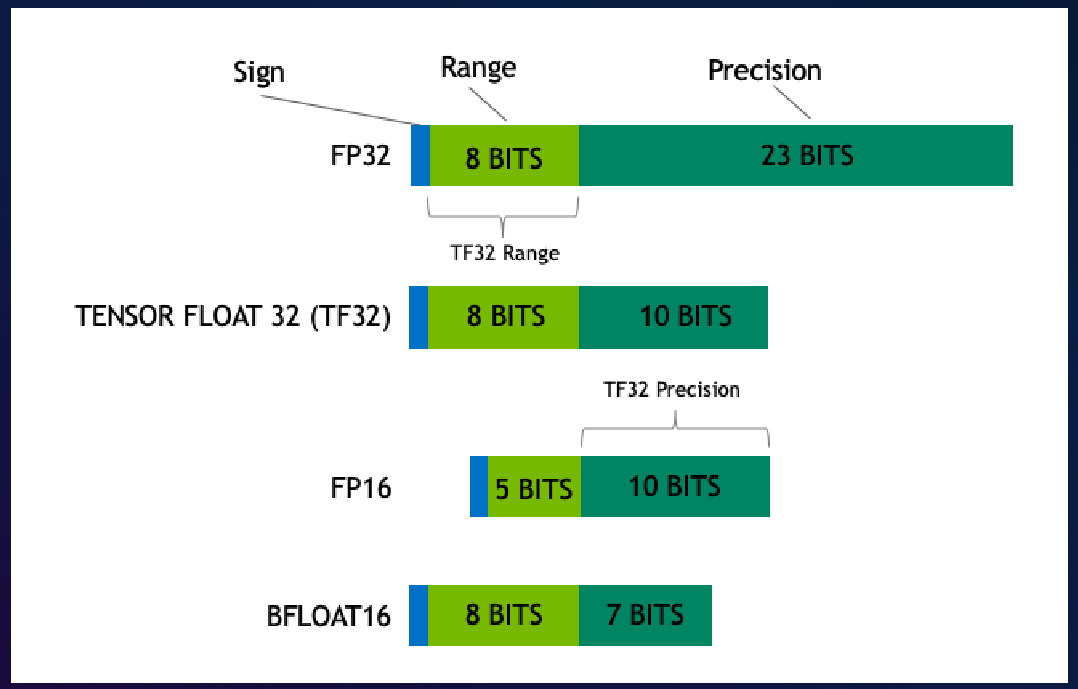
- Adds pairs of rank-decomposition weight matrices to existing weights
- r – Lower dimensional space
- Only trains newly added weights
- Pick a subset of layers in an existing model and modify their weights
- Fine-tune <1% of the parameters

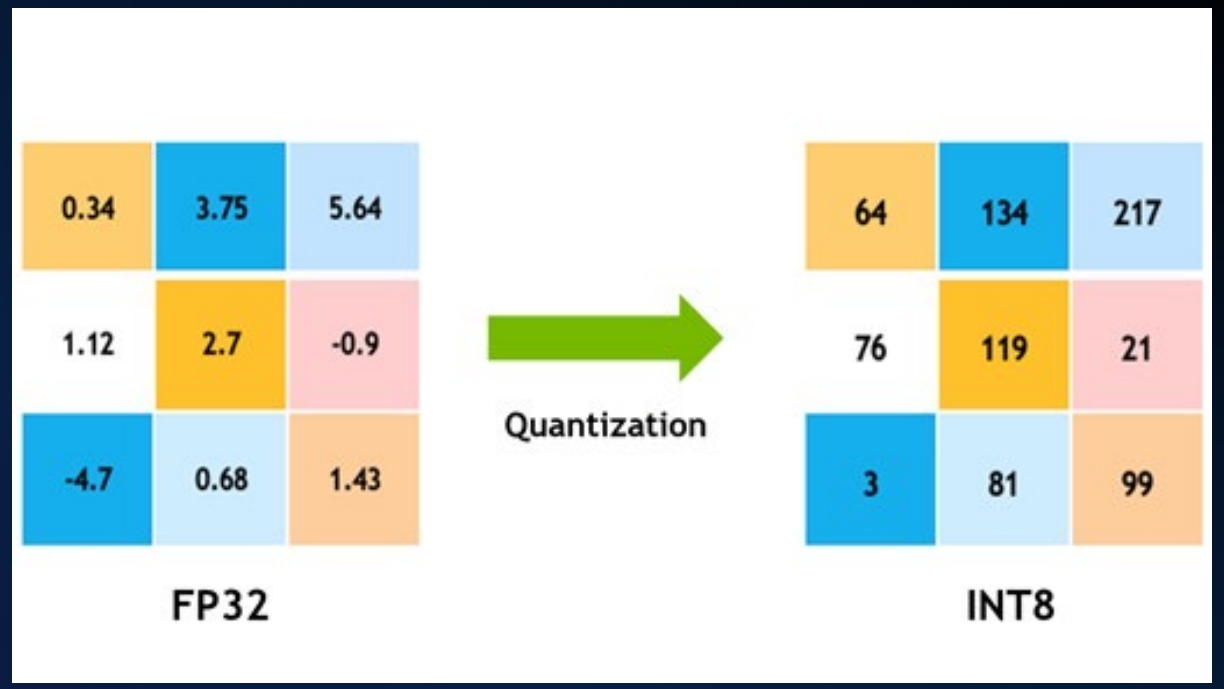
LoRA: Low-Rank Adaptation of large language models

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Quantization

Model compression





© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Labs

- Bedrock Console: text generation with Titan Text
- Bedrock Console: news summary with Command Light
- Bedrock SDK: news summary with Llama 2
- Bedrock SDK: image generation with Titan Image Generator
- SageMaker JumpStart console: text generation with Mistral
- SageMaker JumpStart SDK: dialog summary with LLama 3

aws

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Isengard | Console Home | Console Home | SageMaker Studio | Amazon Sa

us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/studio-landing

aws | Services | Search [Alt+S] | N. Virginia | Admin/linmicht-Isengard @ 0947-8459-0684

Amazon SageMaker

- Getting started
- **Applications and IDEs**
  - Studio
  - Canvas
  - RStudio
  - TensorBoard
  - Profiler
  - Notebooks
- **Admin configurations**
  - Domains
  - Role manager
  - Images
  - Lifecycle configurations
- SageMaker dashboard
- Search
- **JumpStart**
  - Foundation models

Amazon SageMaker

SageMaker Studio

The first fully integrated development environment (IDE) for machine learning.

Get Started

Select user profile

default-20240221t112169

Open Studio

How it works

What is Studio?

Amazon SageMaker Studio provides a single, web-based visual interface where you can perform all ML development steps, improving data science team productivity by up to 10x. SageMaker Studio gives you complete access, control, and visibility into each step required to build, train, and deploy models.

Get Started with SageMaker

Pricing (US)

With Amazon SageMaker Studio, you pay only for what you use. Authoring, training and hosting is billed by the second, with no minimum fees and no upfront commitments.

Learn more

studio-d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/home?tab=Overview

SageMaker Studio

Provide feedback

Applications (6)

JupyterLab
RStudio
Canvas
Code Editor
Studio Cl...
MLflow

Home
Running instances
Data
Auto ML
Experiments
Jobs
Pipelines
Models

Collapse Menu

Home

Launch workflows, manage your applications and spaces, and view getting started materials

Onboarding plan

To get the most out of the new Studio experience, explore the onboarding steps below.

Take the tour

Quick tour highlights where you can find key features and how to navigate the new experience. See what's new and where to locate the tools you need to be productive.

Take the tour >

Migrate data and notebooks

Bring your previous work into the new experience. Transfer notebooks, data sources, and other artifacts so they remain accessible as you adopt the new environment.

Learn more

Not ready to use the new experience? Revert to Studio Classic experience in domain settings. Learn more

Overview | Getting started

Overview

Start a new ML workflow or jump back into your workflow

Type here to search

ENG 6:21 PM

studio-d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/home?tab=Overview

SageMaker Studio

Provide feedback

Code Editor | Studio Cl... | MLflow

- Home
- Running instances
- Data
- Auto ML
- Experiments
- Jobs
- Pipelines
- Models
- JumpStart
- Deployments

Collapse Menu

Home

Launch workflows, manage your applications and spaces, and view getting started materials

Onboarding plan

To get the most out of the new Studio experience, explore the onboarding steps below.

Take the tour

Quick tour highlights where you can find key features and how to navigate the new experience. See what's new and where to locate the tools you need to be productive.

Take the tour >

Migrate data and notebooks

Bring your previous work into the new experience. Transfer notebooks, data sources, and other artifacts so they remain accessible as you adopt the new environment.

Learn more

Not ready to use the new experience? Revert to Studio Classic experience in domain settings. Learn more

Overview | Getting started

Overview

Start a new ML workflow or jump back into your workflow

Type here to search

ENG 6:21 PM

Code Editor
Studio Cl...
MLflow

Home
Running instances
Data
Auto ML
Experiments
Jobs
Pipelines
Models
JumpStart
Deployments

Use pre-trained foundation models

Collapse Menu

All public models

Discover all popular pre-trained models offered by SageMaker

Providers 16

Search providers or models...

HuggingFace

Explore hundreds of popular and trending models

Meta

Explore popular and trending models from Meta including Llama, Code Llama, and more.

View 36 models >

AI21

Explore popular and trending models from AI21 Labs including Jurassic and more.

View 6 models >

Stability AI

Cohere

TensorFlow

studio-d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/JumpStart/SageMakerPublicHub

SageMaker Studio > JumpStart > SageMakerPublicHub

Provide feedback

Code Editor | Studio Cl... | MLflow

- Home
- Running instances
- Data
- Auto ML
- Experiments
- Jobs
- Pipelines
- Models
- JumpStart
- Deployments

Collapse Menu

All public models

Discover all popular pre-trained models offered by SageMaker

Providers 16

mistral

Models 16

Mistral 7B Instruct V3
by HuggingFace
Text Generation

Mistral 7B V3
by HuggingFace
Text Generation

Mixtral-8x22B-Instruct-v0.1
by HuggingFace
Text Generation

Mistral 7B Instruct
by HuggingFace
Text Generation • 4m • 3M • 2K

Mistral 7B
by HuggingFace
Text Generation • 7m • 1.9M • 3.1K

Mistral 7B OpenOrca GPTQ
by HuggingFace
Text Generation • 9m • 10K • 101

Type here to search

ENG 6:21 PM

Code Editor
Studio Cl...
MLflow

Home
Running instances
Data
Auto ML
Experiments
Jobs
Pipelines
Models
JumpStart
Deployments

Collapse Menu

Mistral 7B

by HuggingFace

Train | Deploy | Optimize | Evaluate

About | Notebooks

About

Description

Mistral 7B

License

This model has Apache 2.0 License. Please read the terms carefully.

Mistral 7B is the first foundation model from Mistral AI, supporting English text generation tasks and showing natural coding abilities. It is a powerful model with a variety of use cases such as text summarization, classification, text completion, code completion. To demonstrate the easy customizability of the model, Mistral AI has also released a Mistral 7B-instruct model for chat use cases, fine-tuned using a variety of publicly available conversation datasets.

Mistral 7B is a transformer model and uses Grouped-query attention and Sliding-window attention to achieve faster inference (low latency) and handle longer sequences. Mistral 7B has a 8k context length and is an optimal model for low latency with low memory requirement and high throughput for its size. The model is made available under the permissive Apache 2.0 license, for use without restrictions.

Note that we utilize a single GPU to load the model.

Use the Deployed Model for Inference

The deployed model can be used for running inference on any input text. Example python code for how to run inference

Preview notebooks >

Tags

Huggingface | Text Generation | Apache-2.0

1.9M | 3.1K

Papers

Arxiv:2310.06825

Author HuggingFace

Type here to search
ENG 6:21 PM

Fine-tune model

The training settings have been populated with recommended defaults

This model is pretrained. You will fine-tune its parameters instead of starting from scratch. Fine-tuning can produce accurate models with smaller datasets and less training time.

Create a training job to fit this model to your own data. This model is pre-trained. Fine-tune the model weights instead of starting from scratch. Fine-tuning can produce accurate models with smaller datasets and less training time.

Model

Model settings to be used in the training job.

Model artifact location (S3 URI)*

Specify the location (S3 URI) to the model artifact that Amazon SageMaker should use to fine-tune the model.

- (●) Default model artifact location
- ( ) Enter model artifact location

s3://jumpstart-cache-prod-us-east-1/huggingface-training/v1.1.0/train-huggingface-llm-mistral-7b.tar.gz

Data

Data settings to be used in the training job.

studio-d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jobs/create/train?jumpstart_model_id=huggingface-llm-mistral-7b&jumpstart_model_version=2.9.1&jumpstart_hu...

SageMaker Studio > Jobs > Create > Train

Provide feedback

Data

Data settings to be used in the training job.

Training dataset location (S3 URI)*

Specify the location (S3 URI) to the training dataset that Amazon SageMaker should use to fine-tune the model.

- (●) Example training dataset
- ( ) Enter training dataset

s3://jumpstart-cache-prod-us-east-1/training-datasets/genuq/small/

Hyperparameters

Hyperparameters to be used to train the model.

Peft Type
None

Instruction-Train The Model
True

Chat Dataset
False

Epochs
1

Learning Rate
0.000006

Lora R
64

Lora Alpha
16

Lora Dropout
0

Show more

Type here to search

6:22 PM

mba@mbas-MacBook-Air:~/Desktop

```
→  Desktop aws s3 ls s3://jumpstart-cache-prod-us-east-1/training-datasets/genuq/small/
2023-06-30 18:24:29          0
2023-06-30 18:24:29    1803346 2000-data-genuq.jsonl
2023-06-30 18:24:29        154 template.json
→  Desktop
```

mba@mbas-MacBook-Air:~/Desktop

```
➜  Desktop aws s3 cp s3://jumpstart-cache-prod-us-east-1/training-datasets/genuq/small . --recursive
download: s3://jumpstart-cache-prod-us-east-1/training-datasets/genuq/small/template.json to ./template.json
download: s3://jumpstart-cache-prod-us-east-1/training-datasets/genuq/small/2000-data-genuq.jsonl to ./2000-data-genuq.jsonl
➜  Desktop _
```

mba@mbas-MacBook-Air:~/Desktop

```
➜  Desktop cat ./template.json
{"prompt": "Ask a question which is related to the following text, but cannot be answered based on the text. Text: {context}", "co
mpletion": "{question}"}%
➜  Desktop cat ./2000-data-genuq.jsonl | head -n 1
{"context": "The Normans (Norman: Nourmands; French: Normands; Latin: Normanni) were the people who in the 10th and 11th centuries
 gave their name to Normandy, a region in France. They were descended from Norse (\"Norman\" comes from \"Norseman\") raiders and
pirates from Denmark, Iceland and Norway who, under their leader Rollo, agreed to swear fealty to King Charles III of West Francia
. Through generations of assimilation and mixing with the native Frankish and Roman-Gaulish populations, their descendants would g
radually merge with the Carolingian-based cultures of West Francia. The distinct cultural and ethnic identity of the Normans emerg
ed initially in the first half of the 10th century, and it continued to evolve over the succeeding centuries.", "question": "Who g
ave their name to Normandy in the 1000's and 1100's"}
➜  Desktop
```

Hyperparameters

Hyperparameters to be used to train the model.

Peft Type

None

Instruction-Train The Model

True

Chat Dataset

False

Epochs

1

Learning Rate

0.000006

Lora R

64

Lora Alpha

16

Lora Dropout

0

Show more

Deployment

Instance settings to be used in the training job.

Training Instance*

ml.g5.24xlarge

Output artifact location (S3 URI)*

Specify the location (S3 URI) where Amazon SageMaker should upload the fine-tuned model.

SageMaker Default Bucket

Enter output artifact location

studio-d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jobs/create/train?jumpstart_model_id=huggingface-llm-mistral-7b&jumpstart_model_version=2.9.1&jumpstart_hu...

SageMaker Studio > Jobs > Create > Train

Provide feedback

Lora Alpha
16

Lora Dropout
0

Bits
16

Double Quant
True

Quant Type
nf4

Per Device Train Batch Size
2

Per Device Evaluation Batch Size
8

Add Input Output Demarcation Key
True

Warmup Ratio
0.1

Train The Model From Scratch
False

Fp16 16-Bit (Mixed) Precision Training
False

Bf16 16-Bit (Mixed) Precision Training
True

Evaluation Strategy
steps

Evaluation Steps
20

Gradient Accumulation Steps
8

Logging Steps
8

Weight Decay
0.2

Load Best Model At End
True

Maximum Train Samples

Maximum Validation Samples

Type here to search

ENG 6:22 PM

Deployment

Instance settings to be used in the training job.

Training Instance*

ml.g5.24xlarge

Output artifact location (S3 URI)*

Specify the location (S3 URI) where Amazon SageMaker should upload the fine-tuned model.

- (●) SageMaker Default Bucket
- ( ) Enter output artifact location

s3://sagemaker-us-east-1-094784590684

Security

Security settings to be used in the training job.

IAM Role*

Specify the IAM role that Amazon SageMaker should use to train your model.

- (●) Default IAM Role
- ( ) Enter IAM role
- ( ) Find IAM role

arn:aws:iam::094784590684:role/service-role/AmazonSageMaker-ExecutionRole-20240221T112170

VPC*

Specify whether you training job should connect to a virtual private cloud (VPC).

Product details

	Instance Size	GPU	GPU Memory (GiB)	vCPUs	Memory (GiB)	Storage (GB)	Network Bandwidth (Gbps)	EBS Bandwidth (Gbps)	On Demand Price/hr*	1-yr ISP Effective Hourly (Linux)	3-yr ISP Effective Hourly (Linux)
Single GPU VMs	g5.xlarge	1	24	4	16	1x250	Up to 10	Up to 3.5	$1.006	$0.604	$0.402
	g5.2xlarge	1	24	8	32	1x450	Up to 10	Up to 3.5	$1.212	$0.727	$0.485
	g5.4xlarge	1	24	16	64	1x600	Up to 25	8	$1.624	$0.974	$0.650
	g5.8xlarge	1	24	32	128	1x900	25	16	$2.448	$1.469	$0.979
	g5.16xlarge	1	24	64	256	1x1900	25	16	$4.096	$2.458	$1.638
Multi GPU VMs	g5.12xlarge	4	96	48	192	1x3800	40	16	$5.672	$3.403	$2.269
	g5.24xlarge	4	96	96	384	1x3800	50	19	$8.144	$4.886	$3.258
	g5.48xlarge	8	192	192	768	2x3800	100	19	$16.288	$9.773	$6.515

* Prices shown are for US East (Northern Virginia) AWS Region. Prices for 1-year and 3-year reserved instances are for "Partial Upfront" payment options or "No Upfront" for instances without the Partial Upfront option.

aws Services service quotas N. Virginia aws_sandbox

EC2 VPC

Search results for 'service quotas'

- Services (82)
- Features (127)
- Resources New
- Documentation (488,562)
- Knowledge Articles (1,241)
- Marketplace (56)
- Blogs (11,255)
- Tutorials (33)
- Events (296)

Services

See all 82 results

Service Quotas
View and manage your AWS service quotas from a central location

Service Catalog
Create, share, organize, and govern your curated infrastructure as code (IaC) templates

Directory Service
Host and Manage Active Directory

Key Management Service
Securely Generate and Manage AWS Encryption Keys

Features

See all 127 results

Quota request history
Service Quotas feature

Quota request template
Service Quotas feature

Quota increase request
Service Quotas feature

Consol

Recen

Supp

Amaz

EC2

IAM

Amaz

Billin

Amaz

ault layout

Add widgets

Create application

1

on

Originating account

plication.

Welco

find valuable information to

319% compared to last month for same period

600

CloudShell Feedback

© 2024, Amazon Web Services, Inc. or its affiliates.

Privacy Terms Cookie preferences

Dashboard | Service Quotas

us-east-1.console.aws.amazon.com/servicequotas/home?region=us-east-1#

Paused

aws | Services | Search [Option+S] | N. Virginia | aws_sandbox

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Service Quotas

- Dashboard
- AWS services
- Quota request history
- **Organization**
- Quota request template

Management & Governance

Service Quotas

View and manage AWS quotas

Service Quotas is an AWS service that helps you manage your quotas (also known as limits) for many AWS services in one location. Along with looking up the quota values and current utilization, you can request a quota increase from the Service Quotas console.

Manage quotas

Select a service to view available quotas

AWS Services

View quotas

How it works

Pricing

Service Quotas is offered at no additional charge. There are no setup fees or upfront commitments.

More resources

[User Guide] What is Service Quotas?

Dashboard | Service Quotas

us-east-1.console.aws.amazon.com/servicequotas/home?region=us-east-1#

Paused

aws | Services | Search [Option+S] | N. Virginia | aws_sandbox

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Service Quotas

- Dashboard
- AWS services
- Quota request history
- **Organization**
- Quota request template

Management & Governance

Service Quotas

View and manage AWS quotas

Service Quotas is an AWS service that helps you manage your quotas (also known as limits) for many AWS services in one location. Along with looking up the quota values and current utilization, you can request a quota increase from the Service Quotas console.

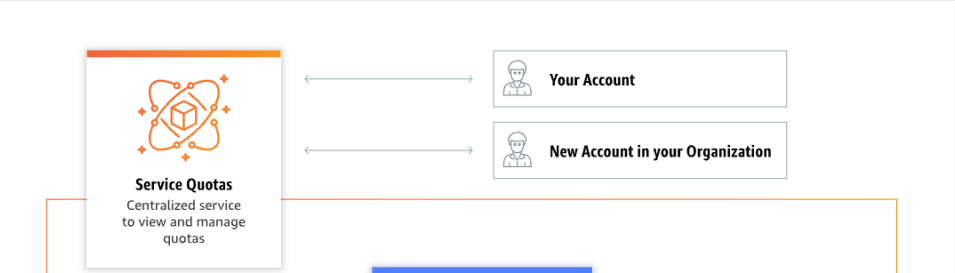
Manage quotas

Select a service to view available quotas

Amazon SageMaker

View quotas

How it works



Pricing

Service Quotas is offered at no additional charge. There are no setup fees or upfront commitments.

More resources

[User Guide] What is Service Quotas?

Dashboard | Service Quotas

us-east-1.console.aws.amazon.com/servicequotas/home?region=us-east-1#

Paused

aws | Services | Search [Option+S] | N. Virginia | aws_sandbox

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Service Quotas

- Dashboard
- AWS services
- Quota request history
- **Organization**
- Quota request template

Management & Governance

Service Quotas

View and manage AWS quotas

Service Quotas is an AWS service that helps you manage your quotas (also known as limits) for many AWS services in one location. Along with looking up the quota values and current utilization, you can request a quota increase from the Service Quotas console.

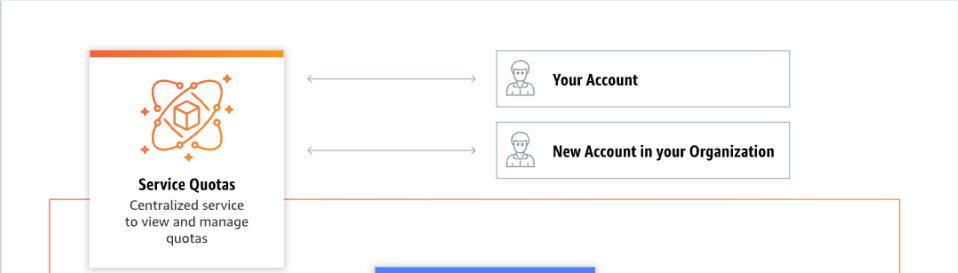
Manage quotas

Select a service to view available quotas

Amazon SageMaker

View quotas

How it works



Pricing

Service Quotas is offered at no additional charge. There are no setup fees or upfront commitments.

More resources

[User Guide] What is Service Quotas?

Amazon SageMaker

Amazon SageMaker is a fully-managed service that enables data scientists and developers to quickly and easily build, train, and deploy machine learning models at scale.

Service quotas info

Request increase at account level

View your applied quota values, default quota values, and request quota increases for quotas. Learn more

Search by quota name

1 2 3 4 5 6 7 ... 51

Quota name	Applied account-level quota value	AWS default quota value	Utilization	Adjustability
Canvas Apps running on ml.m5.4xlarge instances	40	10	0	Account level
Canvas Apps running on system instances	40	10	0	Account level
Large-sized MLflow Tracking Server usage	1	1	0	Account level
Longest run time for a processing job	432,000 seconds	432,000 seconds	Not available	Not adjustable
Longest run time for a training job	432,000	432,000	0	Account level
Longest run time for an AutoML job, from creation to termination	2,592,000 seconds	2,592,000 seconds	Not available	Not adjustable

Quotas list - Amazon SageMa

us-east-1.console.aws.amazon.com/servicequotas/home/services/sagemaker/quotas

Paused

aws | Services | Search [Option+S] | N. Virginia | aws_sandbox

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Service Quotas

- Dashboard
- AWS services
- Quota request history
- Organization
- Quota request template

Service Quotas > AWS services > Amazon SageMaker

Amazon SageMaker

Amazon SageMaker is a fully-managed service that enables data scientists and developers to quickly and easily build, train, and deploy machine learning models at scale.

Service quotas info

Request increase at account level

View your applied quota values, default quota values, and request quota increases for quotas. Learn more

m5.24xlarge

12 matches

1

Quota name	Applied account-level quota value	AWS default quota value	Utilization	Adjustability
ml.m5.24xlarge for cluster usage	2	0	0	Account level
ml.m5.24xlarge for endpoint usage	1	0	0	Account level
ml.m5.24xlarge for notebook instance usage	2	0	0	Account level
ml.m5.24xlarge for processing job usage	1	0	0	Account level
ml.m5.24xlarge for spot training job usage	2	0	0	Account level
ml.m5.24xlarge for training job usage	3	0	0	Account level
ml.m5.24xlarge for training warm pool usage	0	0	0	Account level

Quotas list - Amazon SageMa

us-east-1.console.aws.amazon.com/servicequotas/home/services/sagemaker/quotas

Paused

aws | Services | Search [Option+S] | N. Virginia | aws_sandbox

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Service Quotas

- Dashboard
- AWS services
- Quota request history
- Organization
- Quota request template

Service Quotas > AWS services > Amazon SageMaker

Amazon SageMaker

Amazon SageMaker is a fully-managed service that enables data scientists and developers to quickly and easily build, train, and deploy machine learning models at scale.

Service quotas info

View your applied quota values, default quota values, and request quota increases for quotas. Learn more

Request increase at account level

m5.24xlarge

12 matches

1

Quota name	Applied account-level quota value	AWS default quota value	Utilization	Adjustability
ml.m5.24xlarge for cluster usage	2	0	0	Account level
ml.m5.24xlarge for endpoint usage	1	0	0	Account level
ml.m5.24xlarge for notebook instance usage	2	0	0	Account level
ml.m5.24xlarge for processing job usage	1	0	0	Account level
ml.m5.24xlarge for spot training job usage	2	0	0	Account level
ml.m5.24xlarge for training job usage	3	0	0	Account level
ml.m5.24xlarge for training warm pool usage	0	0	0	Account level

Request quota increase: ml.m5.24xlarge for training job usage

Description
ml.m5.24xlarge for training job usage

Requested for
Account (111735445051)

Region
US East (N. Virginia) us-east-1

Increase quota value
Enter in the total amount that you want the quota to be.

5

Must be a number greater than your current quota value of 3

Utilization
0

Language: For requests in a different language than English, send it via AWS Support Center.

Approvals: For some services, smaller increases are automatically approved, while larger requests are submitted to AWS Support.

Approval timeline: AWS Support can approve, deny, or partially approve your requests. Larger increase requests take more time to process and assess while we work with the service team. For urgent requests, use AWS Support Center.

Cancel | View quota details | Request

© 2024, Amazon Web Services, Inc. or its affiliates.

Quotas list - Amazon SageMa

us-east-1.console.aws.amazon.com/servicequotas/home/services/sagemaker/quotas

Paused

aws | Services | Search [Option+S] | N. Virginia | aws_sandbox

EC2 | VPC | RDS | S3 | Support | Amazon SageMaker | AWS DeepRacer | CloudFormation

Service Quotas

- Dashboard
- AWS services
- Quota request history
- **Organization**
- Quota request template

Quota increase request for ml.m5.24xlarge for training job usage was Approved with the Value of 5.

Service Quotas > AWS services > Amazon SageMaker

Amazon SageMaker

Amazon SageMaker is a fully-managed service that enables data scientists and developers to quickly and easily build, train, and deploy machine learning models at scale.

Service quotas info

View your applied quota values, default quota values, and request quota increases for quotas. Learn more

Request increase at account level

m5.24xlarge

12 matches

1

Quota name	Applied account-level quota value	AWS default quota value	Utilization	Adjustability
ml.m5.24xlarge for cluster usage	2	0	0	Account level
ml.m5.24xlarge for endpoint usage	1	0	0	Account level
ml.m5.24xlarge for notebook instance usage	2	0	0	Account level
ml.m5.24xlarge for processing job usage	1	0	0	Account level
ml.m5.24xlarge for spot training job usage	2	0	0	Account level
ml.m5.24xlarge for training job usage	5	0	0	Account level

No VPC

Find VPC

No VPC will be applied to this training job

Encryption*

Specify the encryption keys to secure your data.

Default encryption keys (i.e. KMS key for S3)

Enter encryption keys

Find encryption keys

No output artifact encryption enabled

Additional Information

Name and tags for the training job.

Training Job Name*

jumpstart-dft-huggingface-llm-mistr-20240722-102146

Tags

+ Add

Submit Cancel

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Type here to search

ENG 6:23 PM

jumpstart-dft-huggingface-llm-mistr-20240722-102146 Training Job

X Stop

Status	ARN	Base model
Executing...	arn:aws:sagemaker:us-east-...	huggingface-llm-mistral-7b
Run time (seconds)	**Created on**	**Modified on**
36	7/22/2024, 6:23:25 PM	7/22/2024, 6:24:37 PM

Performance | Artifacts | Security | Hyperparameters | Configurations | Instances | Logs | Tags

Logs

Search...

Load old events

No logs.

There are no logs or events for this training job.

studio-d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jobs/train/jumpstart-dft-huggingface-llm-mistr-20240722-102146

SageMaker Studio > Jobs > Train > Jumpstart Dft Huggingface Llm Mistr 20240722 102146

Provide feedback

Performance | Artifacts | Security | Hyperparameters | Configurations | Instances | Logs | Tags

Logs

Search...

Load old events

```
2024-07-22T10:35:21.449Z [INFO|trainer.py:1813] 2024-07-22 10:35:21,082 >> Num examples = 1,600
2024-07-22T10:35:21.449Z [INFO|trainer.py:1814] 2024-07-22 10:35:21,082 >> Num Epochs = 1
2024-07-22T10:35:21.449Z [INFO|trainer.py:1815] 2024-07-22 10:35:21,082 >> Instantaneous batch size per device = 2
2024-07-22T10:35:21.449Z [INFO|trainer.py:1818] 2024-07-22 10:35:21,082 >> Total train batch size (w. parallel, distributed & accumulation) = 64
2024-07-22T10:35:21.449Z [INFO|trainer.py:1819] 2024-07-22 10:35:21,082 >> Gradient Accumulation steps = 8
2024-07-22T10:35:21.449Z [INFO|trainer.py:1820] 2024-07-22 10:35:21,082 >> Total optimization steps = 25
2024-07-22T10:35:21.449Z [INFO|trainer.py:1812] 2024-07-22 10:35:21,082 >> ***** Running training *****
2024-07-22T10:35:21.449Z [INFO|trainer.py:1813] 2024-07-22 10:35:21,082 >> Num examples = 1,600
2024-07-22T10:35:21.449Z [INFO|trainer.py:1814] 2024-07-22 10:35:21,082 >> Num Epochs = 1
2024-07-22T10:35:21.449Z [INFO|trainer.py:1815] 2024-07-22 10:35:21,082 >> Instantaneous batch size per device = 2
2024-07-22T10:35:21.449Z [INFO|trainer.py:1818] 2024-07-22 10:35:21,082 >> Total train batch size (w. parallel, distributed & accumulation) = 64
2024-07-22T10:35:21.449Z [INFO|trainer.py:1819] 2024-07-22 10:35:21,082 >> Gradient Accumulation steps = 8
2024-07-22T10:35:21.449Z [INFO|trainer.py:1820] 2024-07-22 10:35:21,082 >> Total optimization steps = 25
2024-07-22T10:35:21.449Z [INFO|trainer.py:1821] 2024-07-22 10:35:21,083 >> Number of trainable parameters = 7,241,764,864
2024-07-22T10:35:21.449Z [INFO|trainer.py:1821] 2024-07-22 10:35:21,083 >> Number of trainable parameters = 7,241,764,864
2024-07-22T10:35:21.449Z 0%|          | 0/25 [00:00<?, ?it/s]
2024-07-22T10:36:58.593Z 4%|▍         | 1/25 [01:37<38:55, 97.33s/it]
2024-07-22T10:38:32.749Z 8%|▊         | 2/25 [03:10<36:25, 95.04s/it]
```

Pause

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Type here to search

ENG 6:40 PM

studio-d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jobs/train/jumpstart-dft-huggingface-llm-mistr-20240722-102146

SageMaker Studio > Jobs > Train > Jumpstart Dft Huggingface Llm Mistr 20240722 102146

Provide feedback

Performance | Artifacts | Security | Hyperparameters | Configurations | Instances | Logs | Tags

Logs

Search...

Load old events

```
2024-07-22T11:15:14.109Z [INFO|trainer.py:3067] 2024-07-22 11:15:13,286 >> Saving model checkpoint to /opt/ml/model
2024-07-22T11:15:14.109Z [INFO|configuration_utils.py:473] 2024-07-22 11:15:13,287 >> Configuration saved in /opt/ml/model/config.json
2024-07-22T11:15:14.109Z [INFO|configuration_utils.py:473] 2024-07-22 11:15:13,287 >> Configuration saved in /opt/ml/model/config.json
2024-07-22T11:15:14.109Z [INFO|configuration_utils.py:614] 2024-07-22 11:15:13,288 >> Configuration saved in /opt/ml/model/generation_config.json
2024-07-22T11:15:14.109Z [INFO|configuration_utils.py:614] 2024-07-22 11:15:13,288 >> Configuration saved in /opt/ml/model/generation_config.json
2024-07-22T11:15:20.110Z [2024-07-22 11:15:19,910] [INFO] [launch.py:347:main] Process 140 exits successfully.
2024-07-22T11:15:21.111Z [2024-07-22 11:15:20,912] [INFO] [launch.py:347:main] Process 142 exits successfully.
2024-07-22T11:15:22.111Z [2024-07-22 11:15:21,913] [INFO] [launch.py:347:main] Process 141 exits successfully.
2024-07-22T11:15:39.116Z [INFO|modeling_utils.py:2462] 2024-07-22 11:15:38,555 >> The model is bigger than the maximum size per checkpoint (5GB) and is going to be split in 3 checkpoint shards. You can find where each ...
2024-07-22T11:15:39.116Z [INFO|modeling_utils.py:2462] 2024-07-22 11:15:38,555 >> The model is bigger than the maximum size per checkpoint (5GB) and is going to be split in 3 checkpoint shards. You can find where each ...
2024-07-22T11:15:39.116Z [INFO|tokenization_utils_base.py:2459] 2024-07-22 11:15:38,555 >> tokenizer config file saved in /opt/ml/model/tokenizer_config.json
2024-07-22T11:15:39.116Z [INFO|tokenization_utils_base.py:2459] 2024-07-22 11:15:38,555 >> tokenizer config file saved in /opt/ml/model/tokenizer_config.json
2024-07-22T11:15:39.116Z [INFO|tokenization_utils_base.py:2468] 2024-07-22 11:15:38,556 >> Special tokens file saved in /opt/ml/model/special_tokens_map.json
2024-07-22T11:15:39.116Z [INFO|tokenization_utils_base.py:2468] 2024-07-22 11:15:38,556 >> Special tokens file saved in /opt/ml/model/special_tokens_map.json
2024-07-22T11:15:46.118Z [2024-07-22 11:15:45,938] [INFO] [launch.py:347:main] Process 139 exits successfully.
2024-07-22T11:15:49.118Z 2024-07-22 11:15:48,508 sagemaker-training-toolkit INFO Waiting for the process to finish and give a return code.
2024-07-22T11:15:49.119Z 2024-07-22 11:15:48,508 sagemaker-training-toolkit INFO Done waiting for a return code. Received 0 from exiting process.
2024-07-22T11:15:49.119Z 2024-07-22 11:15:48,508 sagemaker-training-toolkit INFO Reporting training SUCCESS
```

Pause

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Type here to search

ENG 8:25 PM

studio-d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jobs/train/jumpstart-dft-huggingface-llm-mistr-20240722-102146

SageMaker Studio > Jobs > Train > Jumpstart Dft Huggingface Llm Mistr 20240722 102146

Provide feedback

jumpstart-dft-huggingface-llm-mistr-20240722-102146 Training Job

Optimize | Deploy

Status	ARN	Base model
Completed	arn:aws:sagemaker:us-east-...	huggingface-llm-mistral-7b
Run time (seconds)	**Created on**	**Modified on**
3113	7/22/2024, 6:23:25 PM	7/22/2024, 7:16:30 PM

Performance | Artifacts | Security | Hyperparameters | Configurations | Instances | Logs | Tags

Logs

Search...

Load old events

```
2024-07-22T11:15:14.109Z [INFO|trainer.py:3067] 2024-07-22 11:15:13,286 >> Saving model checkpoint to /opt/ml/model
2024-07-22T11:15:14.109Z [INFO|configuration_utils.py:473] 2024-07-22 11:15:13,287 >> Configuration saved in /opt/ml/model/config.json
2024-07-22T11:15:14.109Z [INFO|configuration_utils.py:473] 2024-07-22 11:15:13,287 >> Configuration saved in /opt/ml/model/config.json
2024-07-22T11:15:14.109Z [INFO|configuration_utils.py:614] 2024-07-22 11:15:13,288 >> Configuration saved in /opt/ml/model/generation_config.json
2024-07-22T11:15:14.109Z [INFO|configuration_utils.py:614] 2024-07-22 11:15:13,288 >> Configuration saved in /opt/ml/model/generation_config.json
2024-07-22T11:15:20.110Z [2024-07-22 11:15:19,910] [INFO] [launch.py:347:main] Process 140 exits successfully.
2024-07-22T11:15:21.111Z [2024-07-22 11:15:20,912] [INFO] [launch.py:347:main] Process 142 exits successfully.
2024-07-22T11:15:22.111Z [2024-07-22 11:15:21,913] [INFO] [launch.py:347:main] Process 141 exits successfully.
2024-07-22T11:15:39.116Z [INFO|modeling_utils.py:2462] 2024-07-22 11:15:38,555 >> The model is bigger than the maximum size per checkpoint (5GB) and is going to be split in 3 checkpoint shards. You can find where each ...
2024-07-22T11:15:39.116Z [INFO|modeling_utils.py:2462] 2024-07-22 11:15:38,555 >> The model is bigger than the maximum size per checkpoint (5GB) and is going to be split in 3 checkpoint shards. You can find where each ...
2024-07-22T11:15:39.116Z [INFO|tokenization_utils_base.py:2459] 2024-07-22 11:15:38,555 >> tokenizer config file saved in /opt/ml/model/tokenizer_config.json
2024-07-22T11:15:39.116Z [INFO|tokenization_utils_base.py:2459] 2024-07-22 11:15:38,555 >> tokenizer config file saved in /opt/ml/model/tokenizer_config.json
```

Type here to search

ENG 8:26 PM

Deploy model to endpoint

Deploy your models to a SageMaker endpoint by selecting the deployment resources. Learn more

Endpoint settings

Endpoint name *

Default endpoint name - jumpstart-dft-hf-llm-mistral-7b-20240722-122546

Instance type *

ml.g5.12xlarge (Default)

Initial instance count*

1

Inference type

Real-time

For sustained traffic and consistently low latency. Supports payload sizes up to 6 MB and runtimes up to 60 sec.

Models

Cancel

Deploy

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Type here to search

ENG 8:27 PM

Default endpoint name - jumpstart-dft-hf-llm-mistral-7b-20240722-122546

Instance type *

ml.g5.12xlarge (Default)

Initial instance count*

1

Inference type

Real-time

For sustained traffic and consistently low latency. Supports payload sizes up to 6 MB and runtimes up to 60 sec.

Models

mistral-7b-122546 Inference Optimized

Advanced settings

Cancel Deploy

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Type here to search ENG 8:27 PM

Default endpoint name - jumpstart-dft-hf-llm-mistral-7b-20240722-122546

Instance type *

ml.g5.12xlarge (Default)

Initial instance count*

1

Inference type

Real-time

For sustained traffic and consistently low latency. Supports payload sizes up to 6 MB and runtimes up to 60 sec.

Models

mistral-7b-122546 Inference Optimized

Advanced settings

Cancel Deploy

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Type here to search

ENG 8:27 PM

jumpstart-dft-hf-llm-mistral-7b-20240722-122546 Endpoint

Delete

Endpoint summary

Inference Type	Status	Creation time	Last updated
Real-time	Creating	Mon Jul 22 2024 20:27:47 GMT+0800 (Taipei Standard Time)	Mon Jul 22 2024 20:27:48 GMT+0800 (Taipei Standard Time)

ARN

arn:aws:sagemaker:us-east-1:094784590684:endpoint/jumpstart-dft-hf-llm-mistral-7b-20240722-122546

URL

https://runtime.sagemaker.us-east-1.amazonaws.com/endpoints/jumpstart-dft-hf-llm-mistral-7b-20240722-122546/invocations

Variants | Settings | Test inference

Variants

Search...

Endpoint jumpstart-dft-hf-llm-mistral-7b-20240722-122546 is being created. 3 s

Model	Variant name	Instance type	Instance count	Instance weight
mistral-7b-122546	AllTraffic P	ml.g5.12xlarge	- -	- -

Type here to search

ENG 8:27 PM

jumpstart-dft-hf-llm-mistral-7b-20240722-122546 Endpoint

Delete

Endpoint summary

Inference Type	Status	Creation time	Last updated
Real-time	In service	Mon Jul 22 2024 20:27:47 GMT+0800 (Taipei Standard Time)	Mon Jul 22 2024 20:34:12 GMT+0800 (Taipei Standard Time)

ARN

arn:aws:sagemaker:us-east-1:094784590684:endpoint/jumpstart-dft-hf-llm-mistral-7b-20240722-122546

URL

https://runtime.sagemaker.us-east-1.amazonaws.com/endpoints/jumpstart-dft-hf-llm-mistral-7b-20240722-122546/invocations

Variants | Settings | Test inference | Auto-scaling

Test inference

Test your model by sending request payload to the SageMaker hosted endpoint and receiving a response, or by using the Python SDK example code provided.

Testing Options

Select the method you want to use to test your inference

- (●) Test the sample request
- ( ) Use Python SDK example code

Content type

Type here to search

ENG 9:31 PM

application/json

JSON

The default payload for the selected model is shown below. You can send the payload as-is or modify and send the request to the endpoint.

```
{
    "inputs": "Girafatron is obsessed with giraffes, the most glorious animal on the face of this Earth. Giraftron believes all other animals are irrelevant when compared to the glorious majesty of the giraffe.\nDaniel: Hello, Girafatron!\nGirafatron:",
    "parameters": {
        "max_new_tokens": 50,
        "return_full_text": false,
        "do_sample": true,
        "top_k": 10,
        "stop": [
            "Daniel:"
        ]
    }
}
```

Send Request

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Type here to search

ENG 9:31 PM

application/json

JSON

The default payload for the selected model is shown below. You can send the payload as-is or modify and send the request to the endpoint.

```
{
    "inputs": "Girafatron is obsessed with giraffes, the most glorious animal on the face of this Earth. Giraftron believes all other animals are irrelevant when compared to the glorious majesty of the giraffe.\nDaniel: Hello, Girafatron!\nGirafatron:",
    "parameters": {
        "max_new_tokens": 50,
        "return_full_text": false,
        "do_sample": true,
        "top_k": 10,
        "stop": [
            "Daniel:"
        ]
    }
}
```

Send Request

Inference Result

Status	Execution Length (ms)
Success	1553
Request Time	**Result Time**
10 seconds ago	10 seconds ago

Result


```
{
    "body": {
        "generated_text": " Hello, Daniel!\n\n is the most glorious animal on the face of this Earth.\n\n is the most relevant animal when compared to the glorious \n\n\n\n\n\n\n\n is the least"
    },
    "contentType": "application/json",
    "invokedProductionVariant": "AllTraffic"
}
```


Copy entire result

Request

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Type here to search

ENG 9:32 PM

jumpstart-dft-hf-llm-mistral-7b-20240722-122546 Endpoint

Delete

Endpoint summary

Inference Type	Status	Creation time	Last updated
Real-time	In service	Mon Jul 22 2024 20:27:47 GMT+0800 (Taipei Standard Time)	Mon Jul 22 2024 20:34:12 GMT+0800 (Taipei Standard Time)

ARN

arn:aws:sagemaker:us-east-1:094784590684:endpoint/jumpstart-dft-hf-llm-mistral-7b-20240722-122546

URL

https://runtime.sagemaker.us-east-1.amazonaws.com/endpoints/jumpstart-dft-hf-llm-mistral-7b-20240722-122546/invocations

Variants | Settings | Test inference | Auto-scaling

Test inference

Test your model by sending request payload to the SageMaker hosted endpoint and receiving a response, or by using the Python SDK example code provided.

Testing Options

Select the method you want to use to test your inference

- (●) Test the sample request
- ( ) Use Python SDK example code

Content type

Type here to search

ENG 10:22 PM

studio-d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/inference-experience/endpoints/details/jumpstart-dft-hf-llm-mistral-7b-20240722-122546

SageMaker Studio > Inference Experience > Endpoints > Details > Jumpstart Dft Hf Llm Mistral 7b 20240722 122546

Provide feedback

jumpstart-dft-hf-llm-mistral-7b-20240722-122546 Endpoint

Delete

Endpoint summary

Inference Type
Real-time

Status
In

T+0800

Last updated
Mon Jul 22 2024 20:34:12 GMT+0800 (Taipei Standard Time)

ARN
arn:aws:sagemaker:us-east-1:094784590684:endpoint/jumpstart-dft-hf-llm-mistral-7b-20240722-122546

URL
ht
1.a
df
12

Variants | Settings | Test inference | Auto-scaling

Test inference

Test your model by sending request payload to the SageMaker hosted endpoint and receiving a response, or by using the Python SDK example code provided.

Testing Options

Select the method you want to use to test your inference

- (●) Test the sample request
- ( ) Use Python SDK example code

Content type

Delete endpoint

Deleting this endpoint is permanent and irreversible. All data within this endpoint will be deleted and cannot be recovered.

Are you sure you want to delete this endpoint?

☑ I confirm that I want to delete the endpoint

Cancel | Delete endpoint

Type here to search

ENG 10:22 PM

Endpoints

Endpoint are locations where you send inference requests to your deployed machine learning models. After you create an endpoint, you can add models to it, test it, and change its settings as needed.

Search by endpoint name

Delete | Create endpoint

	Name	Status	Created on	Modified on
	No endpoints You don't have any endpoints			

0 results | Refresh | Rows 10 | Go to page | Page 0 of 0

Learn about endpoints

Get started

- Deploy models for inference
- SageMaker Inference explained: Which style is right for you?

Documentation

- Real-time inference
- SageMaker pricing
- Deploy a Machine Learning Model to a Real-Time Inference Endpoint

What's new

- Recent features launches
- Engineering blog

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Type here to search

ENG 10:22 PM

Labs

- Bedrock Console: text generation with Titan Text
- Bedrock Console: news summary with Command Light
- Bedrock SDK: news summary with Llama 2
- Bedrock SDK: image generation with Titan Image Generator
- SageMaker JumpStart console: text generation with Mistral
- SageMaker JumpStart SDK: dialogue summary with LLama 3

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

https://github.com/aws/amazon-sagemaker-examples/blob/main/introduction_to_amazon_algorithms/jumpstart-foundation-models/llama-3-finetuning.ipynb

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstart-foundation-model...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Filter files by name

/ introduction_to_amazon_algorithms / jumpstart-foundation-models /

Name	Last Modified
domain-adaption-finetuning...	13 days ago
gemma-fine-tuning.ipynb	14 days ago
image-generation-stable-di...	14 days ago
jumpstart-text-generation-i...	14 days ago
llama-2-chat-completion.ipy...	14 days ago
llama-2-finetuning.ipynb	14 days ago
llama-2-text-completion.ipynb	14 days ago
llama-3-finetuning.ipynb	13 days ago
llama-3-text-completion.ipynb	14 days ago
llama-guard-text-moderatio...	14 days ago
mistral-7b-instruction-doma...	14 days ago
README.md	14 days ago
template.json	13 days ago
text-generation-chatbot.ipynb	14 days ago
text-generation-falcon.ipynb	14 days ago
text-generation-few-shot-le...	14 days ago
text-generation-open-llama....	14 days ago
text2text-generation-Batch-...	14 days ago
text2text-generation-bloom...	14 days ago
text2text-generation-flan-t5...	14 days ago
text2text-generation-flan-t5...	14 days ago
train.jsonl	13 days ago

00_setup.ipynb | 01_fine-tuning-t | 02_fine-tuning_l | 03_continued_pr | 1-TIGFT-customi | 2-TIGFT-provisio | llama-3-finetunir

Markdown | git | No Kernel | Share

Fine-tune LLaMA 3 models on SageMaker JumpStart

This notebook's CI test result for us-west-2 is as follows. CI test results in other regions can be found at the end of the notebook.

us-west-2 | sagemaker-distribution-1.6-cpu | ml.t3.medium Failed

---

In this demo notebook, we demonstrate how to use the SageMaker Python SDK to deploy pre-trained Llama 3 model as well as fine-tune it for your dataset in domain adaptation or instruction tuning format.

Model License information

To perform inference on these models, you need to pass custom_attributes='accept_eula=true' as part of header. This means you have read and accept the end-user-license-agreement (EULA) of the model. EULA can be found in model card description or from https://ai.meta.com/resources/models-and-libraries/llama-downloads/. By default, this notebook sets custom_attributes='accept_eula=false', so all inference requests will fail until you explicitly change this custom attribute.

Note: Custom_attributes used to pass EULA are key/value pairs. The key and value are separated by '=' and pairs are separated by ';'. If the user passes the same key more than once, the last value is kept and passed to the script handler (i.e., in this case, used for conditional logic). For example, if 'accept_eula=false; accept_eula=true' is passed to the server, then 'accept_eula=true' is kept and passed to the script handler.

Set up

Simple | 0 | 0 | main | Python 3 (Data Science 3.0) | Initializing | No kernel: Unknown | Instance MEM | Cookie Preferences | Mode: Command | Ln 1, Col 1 | llama-3-finetuning.ipynb | 0

2:48 PM

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...
Paused
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
default-20240221t112169 / Personal Studio
Home
domain-adaption-finetuning-
llama-3-finetuning.ipynb
Markdown git Cluster ipykernel) Python 3 2 vCPU + 4 GiB Share

Fine-tune LLaMA 3 models on SageMaker JumpStart

This notebook's CI test result for us-west-2 is as follows. CI test results in other regions can be found at the end of the notebook.

us-west-2 | sagemaker-distribution-1.6-cpu | ml.t3.medium Failed

---

In this demo notebook, we demonstrate how to use the SageMaker Python SDK to deploy pre-trained Llama 3 model as well as fine-tune it for your dataset in domain adaptation or instruction tuning format.

Model License information

To perform inference on these models, you need to pass custom_attributes='accept_eula=true' as part of header. This means you have read and accept the end-user-license-agreement (EULA) of the model. EULA can be found in model card description or from https://ai.meta.com/resources/models-and-libraries/llama-downloads/. By default, this notebook sets custom_attributes='accept_eula=false', so all inference requests will fail until you explicitly change this custom attribute.

Note: Custom_attributes used to pass EULA are key/value pairs. The key and value are separated by '=' and pairs are separated by ';'. If the user passes the same key more than once, the last value is kept and passed to the script handler (i.e., in this case, used for conditional logic). For example, if 'accept_eula=false; accept_eula=true' is passed to the server, then 'accept_eula=true' is kept and passed to the script handler.

Set up

We begin by installing and upgrading necessary packages. Restart the kernel after executing the cell below for the first time.

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

Home | domain-adaption-finetuning-c | llama-3-finetuning.ipynb

Markdown git Cluster ipykernel) | Python 3 | 2 vCPU + 4 GiB Share

Set up

We begin by installing and upgrading necessary packages. Restart the kernel after executing the cell below for the first time.

[2]:
```
!pip install --upgrade sagemaker datasets
```

```
Requirement already satisfied: yarl<2.0,>=1.0 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets) (1.9.4)
Requirement already satisfied: async-timeout<5.0,>=4.0 in /opt/conda/lib/python3.10/site-packages (from aiohttp->datasets) (4.0.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /opt/conda/lib/python3.10/site-packages (from huggingface-hub>=0.21.2->datasets) (4.11.0)
Requirement already satisfied: zipp>=0.5 in /opt/conda/lib/python3.10/site-packages (from importlib-metadata<7.0,>=1.4.0->sagemaker) (3.17.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /opt/conda/lib/python3.10/site-packages (from requests->sagemaker) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /opt/conda/lib/python3.10/site-packages (from requests->sagemaker) (3.6)
Requirement already satisfied: certifi>=2017.4.17 in /opt/conda/lib/python3.10/site-packages (from requests->sagemaker) (2024.2.2)
Requirement already satisfied: websocket-client>=0.32.0 in /opt/conda/lib/python3.10/site-packages (from docker->sagemaker) (1.7.0)
Requirement already satisfied: six in /opt/conda/lib/python3.10/site-packages (from google-pasta->sagemaker) (1.16.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in /opt/conda/lib/python3.10/site-packages (from jsonschema->sagemaker) (2023.12.1)
Requirement already satisfied: referencing>=0.28.4 in /opt/conda/lib/python3.10/site-packages (from jsonschema->sagemaker) (0.34.0)
Requirement already satisfied: rpds-py>=0.7.1 in /opt/conda/lib/python3.10/site-packages (from jsonschema->sagemaker) (0.18.0)
Requirement already satisfied: python-dateutil>=2.8.2 in /opt/conda/lib/python3.10/site-packages (from pandas->sagemaker) (2.9.0)
Requirement already satisfied: pytz>=2020.1 in /opt/conda/lib/python3.10/site-packages (from pandas->sagemaker) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in /opt/conda/lib/python3.10/site-packages (from pandas->sagemaker) (2024.1)
Requirement already satisfied: ppft>=1.7.6.8 in /opt/conda/lib/python3.10/site-packages (from pathos->sagemaker) (1.7.6.8)
Requirement already satisfied: pox>=0.3.4 in /opt/conda/lib/python3.10/site-packages (from pathos->sagemaker) (0.3.4)
Requirement already satisfied: contextlib2>=0.5.5 in /opt/conda/lib/python3.10/site-packages (from schema->sagemaker) (21.6.0)
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the system package manager. It is recommended to use a v
irtual environment instead: https://pip.pypa.io/warnings/venv
```

Deploy Pre-trained Model

First we will deploy the Llama-2 model as a SageMaker endpoint. To train/deploy 8B and 70B models, please change model_id to "meta-textgeneration-llama-3-8b" and "meta-textgeneration-llama-3-70b" respectively.

[3]:
```
model_id, model_version = "meta-textgeneration-llama-3-8b", "2.*"
```

Amazon SageMaker | us-east | SageMaker Studio | llama-3-fine... (2) - JupyterLa

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Home | domain-adaption-finetuning-ç | llama-3-finetuning.ipynb

Markdown | git | Cluster | ipykernel) | Python 3 | 2 vCPU + 4 GiB | Share

Deploy Pre-trained Model

First we will deploy the Llama-2 model as a SageMaker endpoint. To train/deploy 8B and 70B models, please change model_id to "meta-textgeneration-llama-3-8b" and "meta-textgeneration-llama-3-70b" respectively.

[3]:
```
model_id, model_version = "meta-textgeneration-llama-3-8b", "2.*"
```

[5]:
```
from sagemaker.jumpstart.model import JumpStartModel

pretrained_model = JumpStartModel(model_id=model_id, model_version=model_version)
# Please change the following line to have accept_eula = True
pretrained_predictor = pretrained_model.deploy(accept_eula=True) # manually configure accept_eula=True
```

```
----------!
```

Invoke the endpoint

Next, we invoke the endpoint with some sample queries. Later, in this notebook, we will fine-tune this model with a custom dataset and carry out inference using the fine-tuned model. We will also show comparison between results obtained via the pre-trained and the fine-tuned models.

[7]:
```
def print_response(payload, response):
    print(payload["inputs"])
    print(f"> {response.get('generated_text')}")
    print("\n==================================\n")
```

[10]:
```
payload = {
    "inputs": "I believe the meaning of life is",
    "parameters": {
```

Amazon SageMaker | us-east | SageMaker Studio | llama-3-fine... (2) - JupyterLa

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Home | domain-adaption-finetuning-c | llama-3-finetuning.ipynb

Markdown | git | Cluster | ipykernel) | Python 3 | 2 vCPU + 4 GiB | Share

Invoke the endpoint

Next, we invoke the endpoint with some sample queries. Later, in this notebook, we will fine-tune this model with a custom dataset and carry out inference using the fine-tuned model. We will also show comparison between results obtained via the pre-trained and the fine-tuned models.

[7]:
```
def print_response(payload, response):
    print(payload["inputs"])
    print(f"> {response.get('generated_text')}")
    print("\n==================================\n")
```

[10]:
```
payload = {
    "inputs": "I believe the meaning of life is",
    "parameters": {
        "max_new_tokens": 64,
        "top_p": 0.9,
        "temperature": 0.6,
        "return_full_text": False,
    },
}
try:
    response = pretrained_predictor.predict(
        payload, custom_attributes="accept_eula=true"
    )
    print_response(payload, response)
except Exception as e:
    print(e)
```

```
I believe the meaning of life is
>  to be happy. I believe that happiness is a choice. I believe that happiness is a choice. I believe that happiness is a choice. I believe that happiness is a choice. I believe that happiness is a choice. I believe that happiness is a choice. I believe that happiness

==================================
```

Amazon SageMaker | us-east | SageMaker Studio | llama-3-fine... (2) - JupyterLa

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Home | domain-adaption-finetuning-ç | llama-3-finetuning.ipynb

Markdown | git | Cluster | ipykernel) | Python 3 | 2 vCPU + 4 GiB | Share

Dataset preparation for fine-tuning

You can fine-tune on the dataset with domain adaptation format or instruction tuning format. Please find more details in the section Dataset instruction. In this demo, we will use a subset of Dolly dataset in an instruction tuning format. Dolly dataset contains roughly 15,000 instruction following records for various categories such as question answering, summarization, information extraction etc. It is available under Apache 2.0 license. We will select the summarization examples for fine-tuning.

Training data is formatted in JSON lines (.jsonl) format, where each line is a dictionary representing a single data sample. All training data must be in a single folder, however it can be saved in multiple jsonl files. The training folder can also contain a template.json file describing the input and output formats.

To train your model on a collection of unstructured dataset (text files), please see the section Example fine-tuning with Domain-Adaptation dataset format in the Appendix.

[11]:
```
from datasets import load_dataset

dolly_dataset = load_dataset("databricks/databricks-dolly-15k", split="train")

# To train for question answering/information extraction, you can replace the assertion in next line to example["category"] == "closed_qa"/"information_extractio
summarization_dataset = dolly_dataset.filter(
    lambda example: example["category"] == "summarization"
)
summarization_dataset = summarization_dataset.remove_columns("category")

# We split the dataset into two where test data is used to evaluate at the end.
train_and_test_dataset = summarization_dataset.train_test_split(test_size=0.1)

# Dumping the training data to a local file to be used for training.
train_and_test_dataset["train"].to_json("train.jsonl")
```

Creating json from Arrow format: 100% 2/2 [00:00<00:00, 25.72ba/s]

[11]: 2115840

Amazon SageMaker | us-east | SageMaker Studio | llama-3-fine... (2) - JupyterLa

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Home | domain-adaption-finetuning-ç | llama-3-finetuning.ipynb

Markdown | git | Cluster | ipykernel) | Python 3 | 2 vCPU + 4 GiB | Share

[12]:
```
train_and_test_dataset["train"][0]
```

[12]:
```
{'instruction': 'Using the passage below, in Machine Learning, what is unsupervised learning and is it different from supervised learning?',
 'context': 'Unsupervised learning is a type of algorithm that learns patterns from untagged data. The goal is that through mimicry, which is an important mode of learning in people, the machine is forced to build a concise representation of its world and then generate imaginative content from it.\n\nIn contrast to supervised learning where data is tagged by an expert, e.g. tagged as a "ball" or "fish", unsupervised methods exhibit self-organization that captures patterns as probability densities or a combination of neural feature preferences encoded in the machine\'s weights and activations. The other levels in the supervision spectrum are reinforcement learning where the machine is given only a numerical performance score as guidance, and semi-supervised learning where a small portion of the data is tagged.',
 'response': 'Unsupervised learning can be understood in contrast to supervised learning. The latter requires a data set tagged by an expert to train the machine learning model. The former learns patterns directly from the (unlabeled or untagged) data.'}
```

Next, we create a prompt template for using the data in an instruction / input format for the training job (since we are instruction fine-tuning the model in this example), and also for inferencing the deployed endpoint.

---

[15]:
```
import json

template = {
    "prompt": "Below is an instruction that describes a task, paired with an input that provides further context. "
    "Write a response that appropriately completes the request.\n\n"
    "### Instruction:\n{instruction}\n\n### Input:\n{context}\n\n",
    "completion": " {response}",
}
with open("template.json", "w") as f:
    json.dump(template, f)
```

Upload dataset to S3

We will upload the prepared dataset to S3 which will be used for fine-tuning.

Amazon SageMaker | us-east
SageMaker Studio
llama-3-fine... (2) - JupyterLa
d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...
Paused
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
default-20240221t112169 / Personal Studio
Home
domain-adaption-finetuning-
llama-3-finetuning.ipynb
Markdown
git
Cluster
ipykernel) | Python 3 | 2 vCPU + 4 GiB
Share

Next, we create a prompt template for using the data in an instruction / input format for the training job (since we are instruction fine-tuning the model in this example), and also for inferencing the deployed endpoint.

---

[15]:
```
import json

template = {
    "prompt": "Below is an instruction that describes a task, paired with an input that provides further context. "
    "Write a response that appropriately completes the request.\n\n"
    "### Instruction:\n{instruction}\n\n### Input:\n{context}\n\n",
    "completion": " {response}",
}
with open("template.json", "w") as f:
    json.dump(template, f)
```

Upload dataset to S3

We will upload the prepared dataset to S3 which will be used for fine-tuning.

---

[16]:
```
from sagemaker.s3 import S3Uploader
import sagemaker
import random

output_bucket = sagemaker.Session().default_bucket()
local_data_file = "train.jsonl"
train_data_location = f"s3://{output_bucket}/dolly_dataset"
S3Uploader.upload(local_data_file, train_data_location)
S3Uploader.upload("template.json", train_data_location)
print(f"Training data: {train_data_location}")
```

```
Training data: s3://sagemaker-us-east-1-094784590684/dolly_dataset
```

Amazon SageMaker | us-east
SageMaker Studio
llama-3-fine... (2) - JupyterLa
d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...
Paused
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
default-20240221t112169 / Personal Studio
Home
domain-adaption-finetuning-
llama-3-finetuning.ipynb
Markdown
git
Cluster
ipykernel) | Python 3 | 2 vCPU + 4 GiB
Share

Upload dataset to S3

We will upload the prepared dataset to S3 which will be used for fine-tuning.

[16]:
```
from sagemaker.s3 import S3Uploader
import sagemaker
import random

output_bucket = sagemaker.Session().default_bucket()
local_data_file = "train.jsonl"
train_data_location = f"s3://{output_bucket}/dolly_dataset"
S3Uploader.upload(local_data_file, train_data_location)
S3Uploader.upload("template.json", train_data_location)
print(f"Training data: {train_data_location}")
```

```
Training data: s3://sagemaker-us-east-1-094784590684/dolly_dataset
```

Train the model

Next, we fine-tune the LLaMA 3 8B model on the summarization dataset from Dolly. Finetuning scripts are based on scripts provided by this repo. To learn more about the fine-tuning scripts, please checkout section 5. Few notes about the fine-tuning method. For a list of supported hyper-parameters and their default values, please see section 3. Supported Hyper-parameters for fine-tuning.

[17]:
```
from sagemaker.jumpstart.estimator import JumpStartEstimator


estimator = JumpStartEstimator(
    model_id=model_id,
    model_version=model_version,
    environment={"accept_eula": "true"}  # Please change {"accept_eula": "true"}
```

Train the model

Next, we fine-tune the LLaMA 3 8B model on the summarization dataset from Dolly. Finetuning scripts are based on scripts provided by this repo. To learn more about the fine-tuning scripts, please checkout section 5. Few notes about the fine-tuning method. For a list of supported hyper-parameters and their default values, please see section 3. Supported Hyper-parameters for fine-tuning.

[17]:
```
from sagemaker.jumpstart.estimator import JumpStartEstimator


estimator = JumpStartEstimator(
    model_id=model_id,
    model_version=model_version,
    environment={"accept_eula": "true"},  # Please change {"accept_eula": "true"}
    disable_output_compression=True,
    instance_type="ml.g5.12xlarge",  # For Llama-3-70b, add instance_type = "ml.g5.48xlarge"
)
# By default, instruction tuning is set to false. Thus, to use instruction tuning dataset you use
estimator.set_hyperparameters(
    instruction_tuned="True", epoch="5", max_input_length="1024"
)
estimator.fit({"training": train_data_location})
```

```
INFO:root:Key: avg_epoch_time, Value: 253.53136914639998
INFO:root:Key: avg_checkpoint_time, Value: 3.7008557292000432
INFO:root:Combining pre-trained base model with the PEFT adapter module.
Loading checkpoint shards:   0%|          | 0/4 [00:00<?, ?it/s]
Loading checkpoint shards:  25%|▌         | 1/4 [00:00<00:01,  1.84it/s]
Loading checkpoint shards:  50%|▌         | 2/4 [00:01<00:01,  1.84it/s]
Loading checkpoint shards:  75%|▌         | 3/4 [00:01<00:00,  1.85it/s]
Loading checkpoint shards: 100%|▌         | 4/4 [00:01<00:00,  2.41it/s]#015Loading checkpoint shards: 100%|▌         | 4/4 [00:01<00:00,  2.17it/s]
INFO:root:Saving the combined model in safetensors format.

2024-06-30 04:20:05 Uploading - Uploading generated training modelINFO:root:Saving complete.
INFO:root:Copying tokenizer to the output directory.
INFO:root:Putting inference code with the fine-tuned model directory.
2024-06-30 04:20:03,560 sagemaker-training-toolkit INFO     Waiting for the process to finish and give a return code.
2024-06-30 04:20:03,560 sagemaker-training-toolkit INFO     Done waiting for a return code. Received 0 from exiting process.
```

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

Home | domain-adaption-finetuning-c | llama-3-finetuning.ipynb

Code git Cluster ipykernel) | Python 3 | 2 vCPU + 4 GiB Share

```
)
estimator.fit({"training": train_data_location})
```

```
INFO:root:Key: avg_epoch_time, Value: 253.53136914639998
INFO:root:Key: avg_checkpoint_time, Value: 3.7008557292000432
INFO:root:Combining pre-trained base model with the PEFT adapter module.
Loading checkpoint shards:   0%|          | 0/4 [00:00<?, ?it/s]
Loading checkpoint shards:  25%|▌         | 1/4 [00:00<00:01,  1.84it/s]
Loading checkpoint shards:  50%|▌         | 2/4 [00:01<00:01,  1.84it/s]
Loading checkpoint shards:  75%|▌         | 3/4 [00:01<00:00,  1.85it/s]
Loading checkpoint shards: 100%|▌         | 4/4 [00:01<00:00,  2.41it/s]#015Loading checkpoint shards: 100%|▌         | 4/4 [00:01<00:00,  2.17it/s]
INFO:root:Saving the combined model in safetensors format.

2024-06-30 04:20:05 Uploading - Uploading generated training modelINFO:root:Saving complete.
INFO:root:Copying tokenizer to the output directory.
INFO:root:Putting inference code with the fine-tuned model directory.
2024-06-30 04:20:03,560 sagemaker-training-toolkit INFO     Waiting for the process to finish and give a return code.
2024-06-30 04:20:03,560 sagemaker-training-toolkit INFO     Done waiting for a return code. Received 0 from exiting process.
2024-06-30 04:20:03,561 sagemaker-training-toolkit INFO     Reporting training SUCCESS

2024-06-30 04:20:53 Completed - Training job completed
Training seconds: 2266
Billable seconds: 2266
```

Studio Kernel Dying issue: If your studio kernel dies and you lose reference to the estimator object, please see section 6. Studio Kernel Dead/Creating JumpStart Model from the training Job on how to deploy endpoint using the training job name and the model id.

Deploy the fine-tuned model

Next, we deploy fine-tuned model. We will compare the performance of fine-tuned and pre-trained model.

[19]:
```
finetuned_predictor = estimator.deploy()
```

```
No instance type selected for inference hosting endpoint. Defaulting to ml.g5.12xlarge.
INFO:sagemaker.jumpstart:No instance type selected for inference hosting endpoint. Defaulting to ml.g5.12xlarge.
INFO:sagemaker:Creating model with name: meta-textgeneration-llama-3-8b-2024-06-30-04-35-09-676
INFO:sagemaker:Creating endpoint-config with name meta-textgeneration-llama-3-8b-2024-06-30-04-35-09-667
INFO:sagemaker:Creating endpoint with name meta-textgeneration-llama-3-8b-2024-06-30-04-35-09-667
```

Amazon SageMaker | us-east | SageMaker Studio | llama-3-fine... (2) - JupyterLa

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Home | domain-adaption-finetuning-ç | llama-3-finetuning.ipynb

Code | git | Cluster | ipykernel) | Python 3 | 2 vCPU + 4 GiB | Share

Deploy the fine-tuned model

Next, we deploy fine-tuned model. We will compare the performance of fine-tuned and pre-trained model.

```
finetuned_predictor = estimator.deploy()
```

```
No instance type selected for inference hosting endpoint. Defaulting to ml.g5.12xlarge.
INFO:sagemaker.jumpstart:No instance type selected for inference hosting endpoint. Defaulting to ml.g5.12xlarge.
INFO:sagemaker:Creating model with name: meta-textgeneration-llama-3-8b-2024-06-30-04-35-09-676
INFO:sagemaker:Creating endpoint-config with name meta-textgeneration-llama-3-8b-2024-06-30-04-35-09-667
INFO:sagemaker:Creating endpoint with name meta-textgeneration-llama-3-8b-2024-06-30-04-35-09-667
------------!
```

Evaluate the pre-trained and fine-tuned model

Next, we use the test data to evaluate the performance of the fine-tuned model and compare it with the pre-trained model.

[24]:
```
import pandas as pd
from IPython.display import display, HTML

test_dataset = train_and_test_dataset["test"]

(
    inputs,
    ground_truth_responses,
    responses_before_finetuning,
    responses_after_finetuning,
) = (
    [],
    [],
    [],
```

Amazon SageMaker | us-east
SageMaker Studio
llama-3-fine... (2) - JupyterLa
d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...
Paused
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
default-20240221t112169 / Personal Studio
Home
domain-adaption-finetuning-
llama-3-finetuning.ipynb
Code
git
Cluster
ipykernel) | Python 3 | 2 vCPU + 4 GiB
Share

Evaluate the pre-trained and fine-tuned model

Next, we use the test data to evaluate the performance of the fine-tuned model and compare it with the pre-trained model.

[24]:
```
import pandas as pd
from IPython.display import display, HTML

test_dataset = train_and_test_dataset["test"]

(
    inputs,
    ground_truth_responses,
    responses_before_finetuning,
    responses_after_finetuning,
) = (
    [],
    [],
    [],
    [],
)


def predict_and_print(datapoint):
    # For instruction fine-tuning, we insert a special key between input and output
    input_output_demarkation_key = "\n\n### Response:\n"

    payload = {
        "inputs": template["prompt"].format(
            instruction=datapoint["instruction"], context=datapoint["context"]
        )
        + input_output_demarkation_key,
        "parameters": {"max_new_tokens": 100},
    }
    inputs.append(payload["inputs"])
```

Amazon SageMaker | us-east | SageMaker Studio | llama-3-fine... (2) - JupyterLa

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Home | domain-adaption-finetuning-ç | llama-3-finetuning.ipynb

Code | git | Cluster | ipykernel) | Python 3 | 2 vCPU + 4 GiB | Share

```python
def predict_and_print(datapoint):
    # For instruction fine-tuning, we insert a special key between input and output
    input_output_demarkation_key = "\n\n### Response:\n"

    payload = {
        "inputs": template["prompt"].format(
            instruction=datapoint["instruction"], context=datapoint["context"]
        )
        + input_output_demarkation_key,
        "parameters": {"max_new_tokens": 100},
    }
    inputs.append(payload["inputs"])
    ground_truth_responses.append(datapoint["response"])
    # Please change the following line to "accept_eula=true"
    pretrained_response = pretrained_predictor.predict(
        payload, custom_attributes="accept_eula=false"
    )
    responses_before_finetuning.append(pretrained_response.get("generated_text"))
    # Fine Tuned Llama 3 models doesn't required to set "accept_eula=true"
    finetuned_response = finetuned_predictor.predict(payload)
    responses_after_finetuning.append(finetuned_response.get("generated_text"))


try:
    for i, datapoint in enumerate(test_dataset.select(range(5))):
        predict_and_print(datapoint)

    df = pd.DataFrame(
        {
            "Inputs": inputs,
            "Ground Truth": ground_truth_responses,
            "Response from non-finetuned model": responses_before_finetuning,
            "Response from fine-tuned model": responses_after_finetuning,
        }
    )
    display(HTML(df.to_html()))
```

Amazon SageMaker | us-east | SageMaker Studio | llama-3-fine... (2) - JupyterLa

d-c9qeuo6x91qp.studio.us-east-1.sagemaker.aws/jupyter/default/lab/tree/amazon-sagemaker-examples/introduction_to_amazon_algorithms/jumpstar...

Amazon SageMaker Studio Classic | File | Edit | View | Run | Kernel | Git | Tabs | Settings | Help | default-20240221t112169 / Personal Studio

Home | domain-adaption-finetuning-ç | llama-3-finetuning.ipynb

Code | git | Cluster | ipykernel) | Python 3 | 2 vCPU + 4 GiB | Share

4	Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.\n\n### Instruction:\nWhat is the oldest higher learning institution in the United States\n\n### Input:\nHarvard University is a private Ivy League research university in Cambridge, Massachusetts. Founded in 1636 as Harvard College and named for its first benefactor, the Puritan clergyman John Harvard, it is the oldest institution of higher learning in the United States and is widely considered to be one of the most prestigious universities in the world.\n\n\n\n### Response:\n	The oldest higher learning institution in the United States is Harvard University, which is widely considered the most prestigious university in the world.	Harvard University is the oldest higher learning institution in the United States. It was founded in 1636 as Harvard College and named for its first benefactor, the Puritan clergyman John Harvard. The university is widely considered to be one of the most prestigious universities in the world.	Harvard University is the oldest higher learning institution in the United States. It was founded in 1636 as Harvard College and named for its first benefactor, the Puritan clergyman John Harvard.

Clean up resources ¶

[ ]:
```
# Delete resources
#pretrained_predictor.delete_model()
#pretrained_predictor.delete_endpoint()
#finetuned_predictor.delete_model()
#finetuned_predictor.delete_endpoint()
```

Appendix

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynb | llama-3-finetuning.ipynb | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita

Markdown git No Kernel Share

3. Supported Hyper-parameters for fine-tuning

- epoch: The number of passes that the fine-tuning algorithm takes through the training dataset. Must be an integer greater than 1. Default: 5
- learning_rate: The rate at which the model weights are updated after working through each batch of training examples. Must be a positive float greater than 0. Default: 1e-4.
- instruction_tuned: Whether to instruction-train the model or not. Must be 'True' or 'False'. Default: 'False'
- per_device_train_batch_size: The batch size per GPU core/CPU for training. Must be a positive integer. Default: 4.
- per_device_eval_batch_size: The batch size per GPU core/CPU for evaluation. Must be a positive integer. Default: 1
- max_train_samples: For debugging purposes or quicker training, truncate the number of training examples to this value. Value -1 means using all of training samples. Must be a positive integer or -1. Default: -1.
- max_val_samples: For debugging purposes or quicker training, truncate the number of validation examples to this value. Value -1 means using all of validation samples. Must be a positive integer or -1. Default: -1.
- max_input_length: Maximum total input sequence length after tokenization. Sequences longer than this will be truncated. If -1, max_input_length is set to the minimum of 1024 and the maximum model length defined by the tokenizer. If set to a positive value, max_input_length is set to the minimum of the provided value and the model_max_length defined by the tokenizer. Must be a positive integer or -1. Default: -1.
- validation_split_ratio: If validation channel is none, ratio of train-validation split from the train data. Must be between 0 and 1. Default: 0.2.
- train_data_split_seed: If validation data is not present, this fixes the random splitting of the input training data to training and validation data used by the algorithm. Must be an integer. Default: 0.
- preprocessing_num_workers: The number of processes to use for the preprocessing. If None, main process is used for preprocessing. Default: "None"
- lora_r: Lora R. Must be a positive integer. Default: 8.
- lora_alpha: Lora Alpha. Must be a positive integer. Default: 32
- lora_dropout: Lora Dropout. must be a positive float between 0 and 1. Default: 0.05.
- int8_quantization: If True, model is loaded with 8 bit precision for training. Default for 8B: False. Default for 70B: True.
- enable_fsdp: If True, training uses Fully Sharded Data Parallelism. Default for 8B: True. Default for 70B: False.

Note 1: int8_quantization is not supported with FSDP. Also, int8_quantization = 'False' and enable_fsdp = 'False' is not supported due to CUDA memory issues for any of the g5 family instances. Thus, we recommend setting exactly one of int8_quantization or enable_fsdp to be 'True' Note 2: Due to the size of the model, 70B model can not be fine-tuned with enable_fsdp = 'True' for any of the supported instance types.

Simple 2 4 main No Kernel | Initializing No kernel: Unknown | Instance MEM Cookie Preferences Mode: Command Ln 1, Col 1 llama-3-finetuning.ipynb 0

Amazon SageMaker Studio Classic File Edit View Run Kernel Git Tabs Settings Help default-20240221t112169 / Personal Studio

00_setup.ipynb | 01_fine-tuning-titan-lite.ipynt | llama-3-finetuning.ipynb | 02_fine-tuning_llama2.ipynb | 03_continued_pretraining_tita

Markdown git No Kernel Share

4. Supported Instance types for fine-tuning Llama 3

We have tested our scripts on the following instances types for fine-tuning Llama 3:

Model	Model ID	All Supported Instances Types for fine-tuning
Llama 3 8B	meta-textgeneration-llama-3-8b	ml.g5.12xlarge, ml.g5.24xlarge, ml.g5.48xlarge, ml.p3dn.24xlarge, ml.g4dn.12xlarge
Llama 3 8B Instruct	meta-textgeneration-llama-3-8b-instruct	ml.g5.12xlarge, ml.g5.24xlarge, ml.g5.48xlarge, ml.p3dn.24xlarge, ml.g4dn.12xlarge
Llama 3 70B	meta-textgeneration-llama-3-70b	ml.g5.48xlarge, ml.p4d.24xlarge
Llama 3 70B Instruct	meta-textgeneration-llama-3-70b-instruct	ml.g5.48xlarge, ml.p4d.24xlarge

Other instance types may also work to fine-tune. Note: When using p3 instances, training will be done with 32 bit precision as bfloat16 is not supported on these instances. Thus, training job would consume double the amount of CUDA memory when training on p3 instances compared to g5 instances.

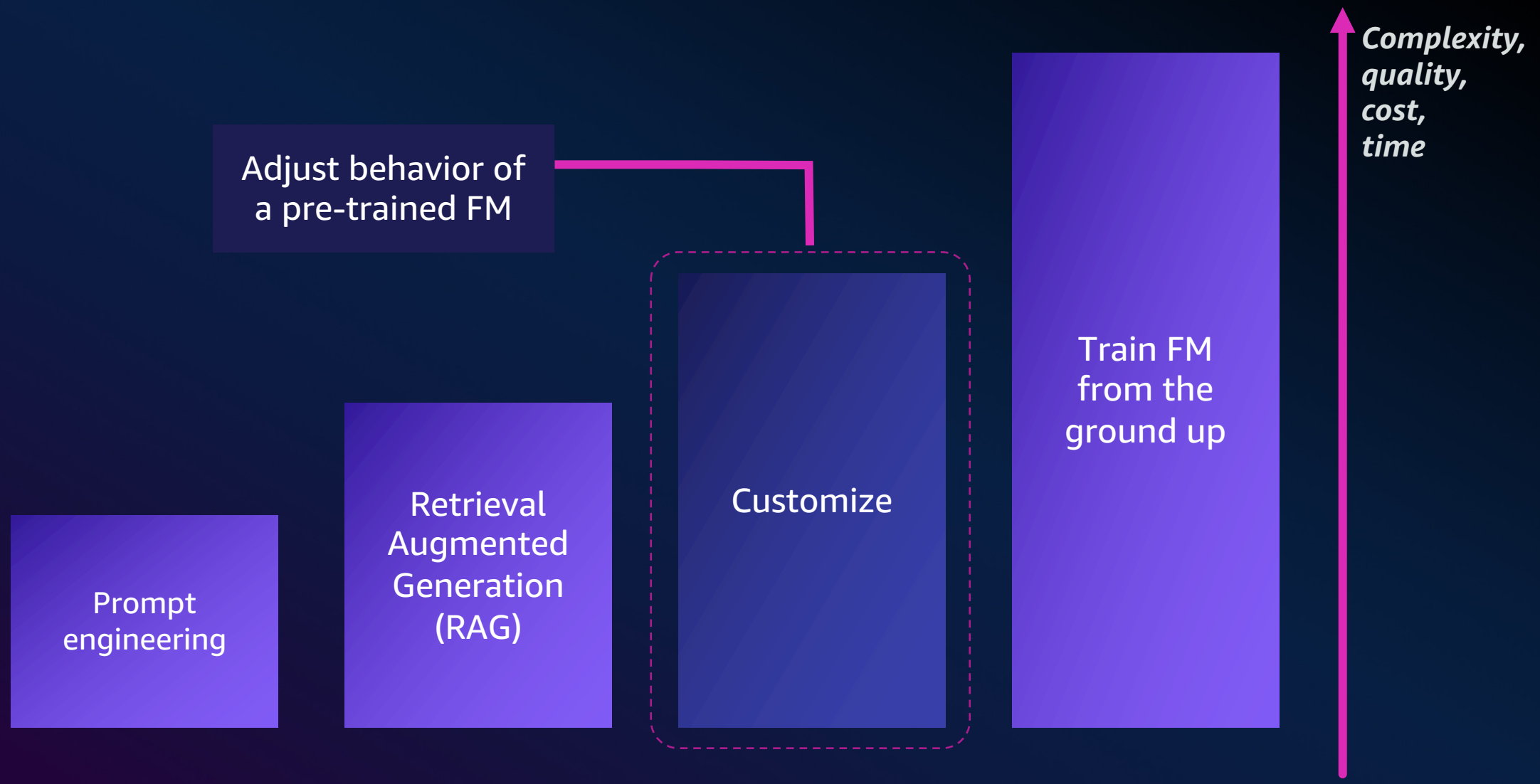
5. Few notes about the fine-tuning method

- Fine-tuning scripts are based on this repo.
- Instruction tuning dataset is first converted into domain adaptation dataset format before fine-tuning.
- Fine-tuning scripts utilize Fully Sharded Data Parallel (FSDP) as well as Low Rank Adaptation (LoRA) method fine-tuning the models

6. Studio Kernel Dead/Creating JumpStart Model from the training Job

Due to the size of the Llama 70B model, training job may take several hours and the studio kernel may die during the training phase. However, during this time, training is still running in SageMaker. If this happens, you can still deploy the endpoint using the training job name with the following code:

Common approaches for customizing FMs



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Please complete the session survey!

Michael Lin

linmicht@amazon.com

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.