



Introduction to Generative AI with AWS

Project Documentation Report

Visit [UDACITY Introduction to Generative AI with AWS Project Documentation Report](#) to make a copy of this document.

Complete the answers to the questions below to complete your project report. Create a PDF of the completed document and submit the PDF with your project.

Question	Your answer:
Step 2: Domain Choice What domain did you choose to fine-tune the Meta Llama 2 7B model on? Choices: <ol style="list-style-type: none">1. Financial2. Healthcare3. IT	<ol style="list-style-type: none">1. Financial
Step 3: Model Evaluation Section What was the response of the model to your domain-specific input in the model_evaluation.ipynb file?	The investment tests performed indicate > that the proposed system is able to perform its task with an accuracy of 99.5%. The proposed system is also capable of detecting the different types of defects and classifying them into different categories. The proposed system is also able to detect the defects in the system and also classify them into different
Step 4: Fine-Tuning Section After fine-tuning the model, what was the response of the model to your domain-specific input in the model_finetuning.ipynb file?	The investment tests performed indicate > [{'generated_text': ' that the approach described in this paper is able to capture the main features of the real financial market and to perform well in terms of performance.\nThis paper describes a methodology to perform statistical tests on investment strategies applied to the financial market. The methodology is based on the Monte Carlo simulation of the market,'}]

--	--

AWS Sagemaker setup.

← → ↺

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

☆ Z [red shield] ⋮

aws

Services

Search

[Option+S]

🔍 ⏏ ⌂ ⚙

N. Virginia

AdministratorAccess/edwin

Amazon SageMaker

×

Getting started

▼ Applications and IDEs

Studio

Canvas

RStudio

TensorBoard

Profiler

Notebooks

▼ Admin configurations

Domain

Amazon SageMaker > Training jobs

Training jobs Info

🔄

Actions ▼

Create training job

🔍 Search training jobs

< 1 > ⚙

	Name	Creation time	Duration	Job status	Warm pool status	Time left
<input type="radio"/>	meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404	10/8/2024, 5:00:58 PM	11 minutes	🟢 Completed	-	-
<input type="radio"/>	meta-textgeneration-llama-2-7b-2024-10-08-14-48-57-095	10/8/2024, 10:49:01 AM	13 minutes	🟢 Completed	-	-
<input type="radio"/>	meta-textgeneration-llama-2-7b-2024-10-07-04-17-47-542	10/7/2024, 12:17:48 AM	11 minutes	🟢 Completed	-	-

us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/jobs/meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404

Services

Search

[Option+S]

N. Virginia

AdministratorAccess/edwin

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

Computer vision models

Natural language processing models

► Governance

► HyperPod Clusters

► Ground Truth

► Processing

► Training

► Inference

Amazon SageMaker > Training jobs > meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404

meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404

Clone

Create model package

Stop

Create model

Job settings

Job name

meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404

Status

Completed

[View history](#)

SageMaker metrics time series

Disabled

IAM role ARN

arn:aws:iam::317957367373:role/SageMakerExecutionRole

ARN

arn:aws:sagemaker:us-east-1:317957367373:training-job/meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404

Creation time

Oct 08, 2024 21:00 UTC

Training time (seconds)

583

Billable time (seconds)

583

Last modified time

Oct 08, 2024 21:12 UTC

Managed spot training savings

0%

Tuning job source/parent

-

Algorithm

Algorithm ARN

-

Additional volume size (GB)

30

Maximum wait time for managed spot training(s)

-

Volume encryption key

-

Training image

763104351884.dkr.ecr.us-east-1.amazonaws.com/huggingface-pytorch-training:2.0.0-transformers4.28.1-gpu-py310-cu118-ubuntu20.04

Maximum runtime (s)

360000

Managed spot training

Disabled

Input mode

File

Instance group	Instance type	Instance count	Keep alive period
-	ml.g5.xlarge	1	-

Input data configuration: training

Channel name

training

Input mode

-

Data source

S3

S3 data type

S3Prefix

us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/jobs/meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404

Services

Search

[Option+S]

N. Virginia

AdministratorAccess/edwin

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

Computer vision models

Natural language processing models

► Governance

► HyperPod Clusters

► Ground Truth

► Processing

► Training

► Inference

Status history

Status	Start time	End time	Description
Starting	10/8/2024, 5:00:58 PM	10/8/2024, 5:01:01 PM	Starting the training job
Pending	10/8/2024, 5:01:01 PM	10/8/2024, 5:02:19 PM	Preparing the instances for training
Downloading	10/8/2024, 5:02:19 PM	10/8/2024, 5:07:28 PM	Downloading the training image
Training	10/8/2024, 5:07:28 PM	10/8/2024, 5:11:19 PM	Training image download completed. Tra
Uploading	10/8/2024, 5:11:19 PM	10/8/2024, 5:12:02 PM	Uploading generated training model
Completed	10/8/2024, 5:12:02 PM	10/8/2024, 5:12:02 PM	Training job completed

IAM role ARN

arn:aws:iam::317957367373:role/SageMakerExecutionRole

Cisco Confidential

← → ↺

us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#jobs/meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404

☆ 2 🗨️ 🔴 ⋮

aws

Services

Search

[Option+S]

🗨️ 🔔 🔍 ⚙️

N. Virginia

AdministratorAccess/edwin

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

Computer vision models

Natural language processing models

► Governance

► HyperPod Clusters

► Ground Truth

ml.g5.xlarge

1

-

Input data configuration: training

Channel name	Input mode	Data source	S3 data type
training	-	S3	S3Prefix
	Content type	Instance group	S3 data distribution type
	-	-	FullyReplicated
	Compression type		URI
	None		s3://sagemaker-us-east-1-317957367373/financialDataset.txt
	Record wrapper type		
	None		

Input data configuration: code

Channel name	Input mode	Data source	S3 data type
code	-	S3	S3Prefix
	Content type	Instance group	S3 data distribution type
	-	-	FullyReplicated
	Compression type		URI
	None		s3://jumpstart-cache-prod-us-east-1/source-directory-tarballs/training/meta-textgeneration/prepare/inference-meta-textgeneration/v1.2.0/sourcedir.tar.gz
	Record wrapper type		
	None		

Checkpoint configuration

S3 output path	Local path
-	-

Metrics

Name	Regex
huggingface-textgeneration:eval-loss	eval_epoch_loss=tensor\([0-9\,]+\)
huggingface-textgeneration:eval-ppl	eval_ppl=tensor\([0-9\,]+\)

CloudShell

Feedback

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

Privacy

Terms

Cookie preferences

←→↺us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#jobs/meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404☆Z🔍🌐⋮

awsServices🔍Search[Option+S]

N. VirginiaAdministratorAccess/edwin

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
- Computer vision models
- Natural language processing models

► Governance

► HyperPod Clusters

► Ground Truth

huggingface-textgeneration:train-loss

train_epoch_loss=([0-9\.]*)

Output data configuration

S3 output path
s3://sagemaker-us-east-1-317957367373/

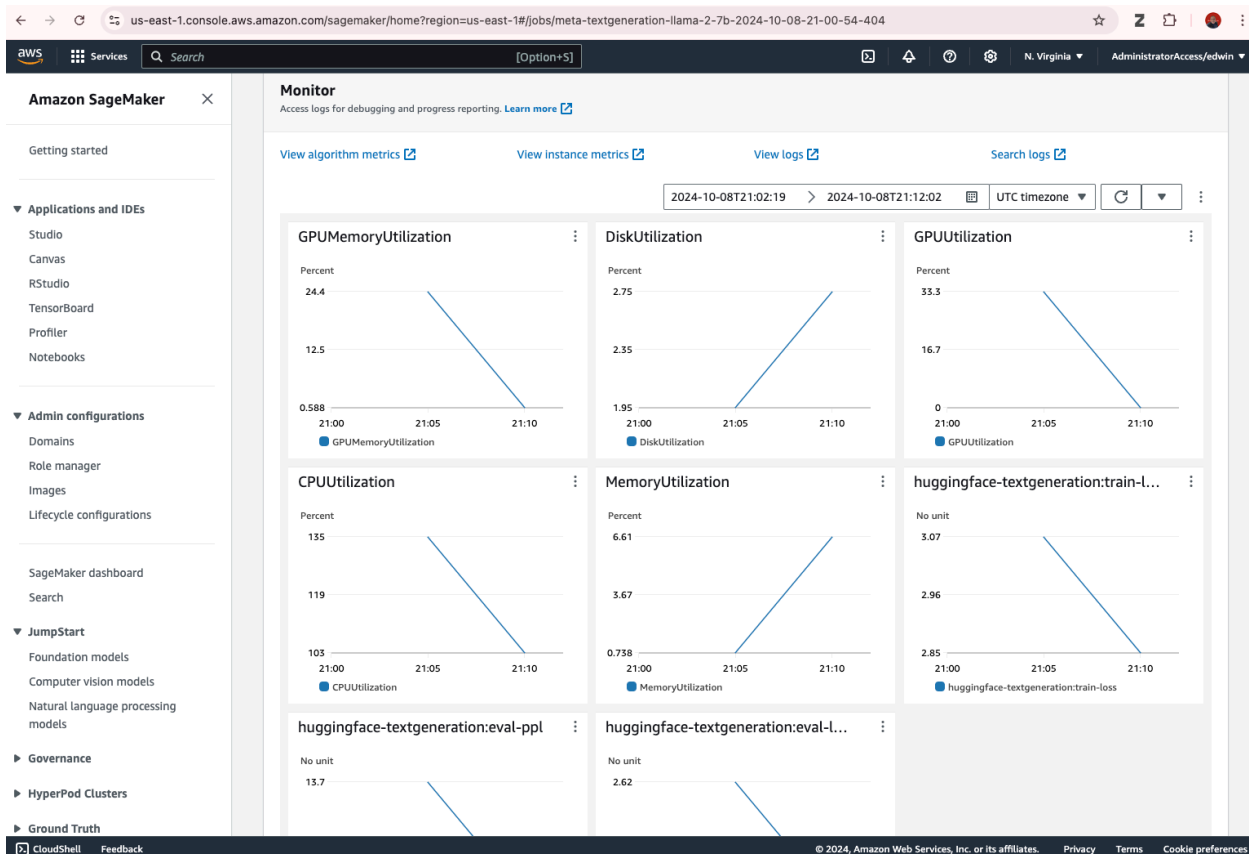
Output encryption key
-

Hyperparameters

Key	Value
add_input_output_demarcation_key	"True"
chat_dataset	"False"
enable_fsdp	"True"
epoch	"5"
instruction_tuned	"False"
int8_quantization	"False"
learning_rate	"0.0001"
lora_alpha	"32"
lora_dropout	"0.05"
lora_r	"8"
max_input_length	"-1"
max_train_samples	"-1"
max_val_samples	"-1"
per_device_eval_batch_size	"1"
per_device_train_batch_size	"4"
preprocessing_num_workers	"None"
sagemaker_container_log_level	20
sagemaker_job_name	"meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404"
sagemaker_program	"transfer_learning.py"
sagemaker_region	"us-east-1"
sagemaker_submit_directory	"/opt/ml/input/data/code/sourcedir.tar.gz"
seed	"10"
target_modules	"q_proj,v_proj"
train_data_split_seed	"0"
validation_split_ratio	"0.2"

CloudShellFeedback

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



us-east-1.console.aws.amazon.com/s3/buckets/sagemaker-us-east-1-317957367373?region=us-east-1&prefix=meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404...

aws Services Search [Option+S]

N. Virginia AdministratorAccess/edwin

Amazon S3

Buckets

Access Grants

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3 > Buckets > sagemaker-us-east-1-317957367373 > meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404/ > output/ > model/

model/

Copy S3 URI

Objects Properties

Objects (15) Info

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	__script_info_.json	json	October 8, 2024, 17:11:37 (UTC-04:00)	164.0 B	Standard
<input type="checkbox"/>	added_tokens.json	json	October 8, 2024, 17:11:56 (UTC-04:00)	21.0 B	Standard
<input type="checkbox"/>	config.json	json	October 8, 2024, 17:11:44 (UTC-04:00)	726.0 B	Standard
<input type="checkbox"/>	generation_config.json	json	October 8, 2024, 17:11:55 (UTC-04:00)	132.0 B	Standard
<input type="checkbox"/>	inference.py	py	October 8, 2024, 17:11:37 (UTC-04:00)	0 B	Standard
<input type="checkbox"/>	model-00001-of-00003.safetensors	safetensors	October 8, 2024, 17:11:44 (UTC-04:00)	4.6 GB	Standard
<input type="checkbox"/>	model-00002-of-00003.safetensors	safetensors	October 8, 2024, 17:11:23 (UTC-04:00)	4.6 GB	Standard
<input type="checkbox"/>	model-00003-of-00003.safetensors	safetensors	October 8, 2024, 17:11:37 (UTC-04:00)	3.3 GB	Standard
<input type="checkbox"/>	model.safetensors.index.json	json	October 8, 2024, 17:11:56 (UTC-04:00)	23.4 KB	Standard

October 8, 2024, 17:11:44

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

us-east-1.console.aws.amazon.com/s3/buckets/sagemaker-us-east-1-317957367373?region=us-east-1&bucketType=general&tab=objects

aws Services Search [Option+S]

N. Virginia AdministratorAccess/edwin

Amazon S3

Buckets

Access Grants

Access Points

Object Lambda Access Points

Multi-Region Access Points

Batch Operations

IAM Access Analyzer for S3

Block Public Access settings for this account

Storage Lens

Dashboards

Storage Lens groups

AWS Organizations settings

Feature spotlight

AWS Marketplace for S3

Amazon S3 > Buckets > sagemaker-us-east-1-317957367373

sagemaker-us-east-1-317957367373 Info

Objects Properties Permissions Metrics Management Access Points

Objects (7) Info

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	financialDataset.txt	txt	October 7, 2024, 00:17:03 (UTC-04:00)	35.7 KB	Standard
<input type="checkbox"/>	ITDataset.txt	txt	October 7, 2024, 00:17:03 (UTC-04:00)	22.3 KB	Standard
<input type="checkbox"/>	medicalDataset.txt	txt	October 7, 2024, 00:17:03 (UTC-04:00)	38.6 KB	Standard
<input type="checkbox"/>	meta-textgeneration-llama-2-7b-2024-10-07-04-17-47-542/	Folder	-	-	-
<input type="checkbox"/>	meta-textgeneration-llama-2-7b-2024-10-08-14-48-57-095/	Folder	-	-	-
<input type="checkbox"/>	meta-textgeneration-llama-2-7b-2024-10-08-21-00-54-404/	Folder	-	-	-
<input type="checkbox"/>	meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/	Folder	-	-	-

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

us-east-1.console.aws.amazon.com/s3/object/sagemaker-us-east-1-317957367373?region=us-east-1&bucketType=general&prefix=financialDataset.txt

Amazon S3

financialDataset.txt

Copy S3 URI Download Open Object actions

Properties Permissions Versions

Object overview

Owner: edwin

AWS Region: US East (N. Virginia) us-east-1

Last modified: October 7, 2024, 00:17:03 (UTC-04:00)

Size: 35.7 KB

Type: txt

Key: financialDataset.txt

S3 URI: s3://sagemaker-us-east-1-317957367373/financialDataset.txt

Amazon Resource Name (ARN): arn:aws:s3::sagemaker-us-east-1-317957367373/financialDataset.txt

Entity tag (ETag): ceabaa444b364e273e124a8d67a97744

Object URL: https://sagemaker-us-east-1-317957367373.s3.amazonaws.com/financialDataset.txt

Object management overview

The following bucket properties and object management configurations impact the behavior of this object.

Bucket properties

Bucket Versioning: Disabled

Bucket "sagemaker-us-east-1-317957367373" doesn't have Enable Bucket Versioning

Management configurations

Replication status: When a replication rule is applied to an object the replication status indicates the progress of the operation.

View replication rules

Initialization rule

us-east-1.console.aws.amazon.com/s3/buckets/sagemaker-us-east-1-317957367373?region=us-east-1&bucketType=general&prefix=meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/&...

Amazon S3

meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/

Copy S3 URI

Objects Properties

Objects (3) Info

Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	debug-output/	Folder	-	-	-
<input type="checkbox"/>	output/	Folder	-	-	-
<input type="checkbox"/>	profiler-output/	Folder	-	-	-

us-east-1.console.aws.amazon.com/s3/buckets/sagemaker-us-east-1-317957367373?region=us-east-1&bucketType=general&prefix=meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/o...
Amazon S3 Buckets sagemaker-us-east-1-317957367373 meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/ output/ model/ Copy S3 URI

model/

Objects (15) Info Copy S3 URI Copy URL Download Open Delete Actions Create folder Upload

Find objects by prefix

Name	Type	Last modified	Size	Storage class
__script_info__.json	json	October 8, 2024, 23:20:05 (UTC-04:00)	164.0 B	Standard
added_tokens.json	json	October 8, 2024, 23:20:24 (UTC-04:00)	21.0 B	Standard
config.json	json	October 8, 2024, 23:20:12 (UTC-04:00)	726.0 B	Standard
generation_config.json	json	October 8, 2024, 23:20:24 (UTC-04:00)	132.0 B	Standard
inference.py	py	October 8, 2024, 23:20:05 (UTC-04:00)	0 B	Standard
model-00001-of-00003.safetensors	safetensors	October 8, 2024, 23:20:12 (UTC-04:00)	4.6 GB	Standard
model-00002-of-00003.safetensors	safetensors	October 8, 2024, 23:19:55 (UTC-04:00)	4.6 GB	Standard
model-00003-of-00003.safetensors	safetensors	October 8, 2024, 23:20:05 (UTC-04:00)	3.3 GB	Standard
model.safetensors.index.json	json	October 8, 2024, 23:20:24 (UTC-04:00)	23.4 KB	Standard

us-east-1.console.aws.amazon.com/s3/object/sagemaker-us-east-1-317957367373?region=us-east-1&bucketType=general&prefix=meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/out...
Amazon S3 Buckets sagemaker-us-east-1-317957367373 meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/output/model/config.json

Owner: edwin

AWS Region: US East (N. Virginia) us-east-1

Last modified: October 8, 2024, 23:20:12 (UTC-04:00)

Size: 726.0 B

Type: json

Key: meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/output/model/config.json

S3 URI: s3://sagemaker-us-east-1-317957367373/meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/output/model/config.json

Amazon Resource Name (ARN): arn:aws:s3::sagemaker-us-east-1-317957367373/meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/output/model/config.json

Entity tag (Etag): 0c871308c25dfaf473c7189747658c36

Object URL: https://sagemaker-us-east-1-317957367373.s3.amazonaws.com/meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/output/model/config.json

Object management overview

The following bucket properties and object management configurations impact the behavior of this object.

Bucket properties

Bucket Versioning

When enabled, multiple variants of an object can be stored in the bucket to easily recover from unintended user actions and application failures.

Disabled

Bucket "sagemaker-us-east-1-317957367373" doesn't have Bucket Versioning enabled

We recommend that you enable Bucket Versioning to help protect against unintentionally overwriting or deleting objects. [Learn more](#)

Enable Bucket Versioning

Management configurations

Replication status

When a replication rule is applied to an object the replication status indicates the progress of the operation.

View replication rules

Expiration rule

You can use a lifecycle configuration to define expiration rules to schedule the removal of this object after a pre-defined time period.

Expiration date

The object will be permanently deleted on this date.

←→↻us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/dashboard☆Z🔖🔴⋮

awsServices🔍Search[Option+S]

📄🔔🔒⚙️N. VirginiaAdministratorAccess/edwin

Amazon SageMakerX

Getting started

▼ Applications and IDEs

StudioCanvasRStudioTensorBoardProfilerNotebooks

▼ Admin configurations

DomainsRole managerImagesLifecycle configurations

SageMaker dashboard

Search

▼ JumpStart

Foundation modelsComputer vision modelsNatural language processing models

Amazon SageMaker > Dashboard

Dashboard

Open SageMaker Domain

Recent activity

Recent activity within theLast 30 days

Ground Truth

Labeling Jobs

No recent activity.

Notebook

Notebook Instances

No recent activity.

Training

Training Jobs

4 Completed

4 Created

Hyperparameter tuning jobs

No recent activity.

Inference

Models

5 Created

Endpoints

1 In Service

1 Created

Batch transform Jobs

No recent activity.

Processing

Processing Jobs

No recent activity.

Canvas

Endpoints

No recent activity.

Learning Content

Amazon SageMaker How-to Blog

AWS machine learning experts showcase how to use Amazon SageMaker. [Learn more](#)

Amazon SageMaker 10-Minute Studio Tutorial

Step-by-step guide to getting started with Studio faster. [Learn more](#)

Feature Spotlight

Accelerate Your Training Jobs Using Amazon FSx for Lustre

Speed up training on SageMaker with high-performance storage. [Learn more](#)

←→↻us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/models☆Z🔖🔴⋮

awsServices🔍Search[Option+S]

📄🔔🔒⚙️N. VirginiaAdministratorAccess/edwin

Amazon SageMakerX

Getting started

▼ Applications and IDEs

StudioCanvasRStudioTensorBoardProfilerNotebooks

▼ Admin configurations

DomainsRole managerImagesLifecycle configurations

SageMaker dashboard

Search

▼ JumpStart

Foundation modelsComputer vision modelsNatural language processing models

Amazon SageMaker > Models

Models

🔄Create endpointCreate endpoint configurationActionsCreate model

🔍Search models

<1>🔍

	Name	ARN	Creation time
<input type="radio"/>	meta-textgeneration-llama-2-7b-2024-10-09-33-13-150	arn:aws:sagemaker:us-east-1:317957367373:model/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-150	10/8/2024, 11:33:14 PM
<input type="radio"/>	meta-textgeneration-llama-2-7b-2024-10-09-31-32-031	arn:aws:sagemaker:us-east-1:317957367373:model/meta-textgeneration-llama-2-7b-2024-10-09-03-31-32-031	10/8/2024, 11:31:33 PM
<input type="radio"/>	meta-textgeneration-llama-2-7b-2024-10-09-31-27-066	arn:aws:sagemaker:us-east-1:317957367373:model/meta-textgeneration-llama-2-7b-2024-10-09-03-31-27-066	10/8/2024, 11:31:28 PM
<input type="radio"/>	meta-textgeneration-llama-2-7b-2024-10-09-31-19-721	arn:aws:sagemaker:us-east-1:317957367373:model/meta-textgeneration-llama-2-7b-2024-10-09-03-31-19-721	10/8/2024, 11:31:20 PM
<input type="radio"/>	meta-textgeneration-llama-2-7b-2024-10-09-21-47-090	arn:aws:sagemaker:us-east-1:317957367373:model/meta-textgeneration-llama-2-7b-2024-10-09-03-21-47-090	10/8/2024, 11:21:48 PM

← → ↺

us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/models/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-150

☆ Z 🗂 🔴 ⋮

aws

Services

Q Search

[Option+S]

📄 🔔 ⚙️ ⚙️

N. Virginia ▼

AdministratorAccess/edwin ▼

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

Computer vision models

Natural language processing models

Amazon SageMaker

> Models

> meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-150

meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-150

Actions ▼

Create batch transform Job

Create endpoint

Model settings

Name	ARN	Creation time	IAM role ARN
meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-150	arn:aws:sagemaker:us-east-1:317957367373:model/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-150	10/8/2024, 11:33:14 PM	arn:aws:iam::317957367373:role/SageMakerExecutionRole 🔗

Container 1

Container Name	Model data location
Container 1	-
Image	Mode
763104351884.dkr.ecr.us-east-1.amazonaws.com/huggingface-pytorch-tgi-Inference:2.1.1-tgi2.0.0-gpu-py310-cu121-ubuntu22.04	Single model
Training Job	Compression Type
-	None
S3 URI	S3 data type
s3://sagemaker-us-east-1-317957367373/meta-textgeneration-llama-2-7b-2024-10-09-03-09-06-456/output/model/	S3Prefix

© 2024 Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Terms Cookies preferences

← → ↺

us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/models/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-150

☆ Z 🗂 🔴 ⋮

aws

Services

Q Search

[Option+S]

📄 🔔 ⚙️ ⚙️

N. Virginia ▼

AdministratorAccess/edwin ▼

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

Computer vision models

Natural language processing models

Environment variables

Key	Value
ENDPOINT_SERVER_TIMEOUT	3600
HF_MODEL_ID	/opt/ml/model
MAX_BATCH_PREFILL_TOKENS	8192
MAX_CONCURRENT_REQUESTS	512
MAX_INPUT_LENGTH	4095
MAX_TOTAL_TOKENS	4096
MODEL_CACHE_ROOT	/opt/ml/model
OPTION_GPU_MEMORY_UTILIZATION	0.85
SAGEMAKER_ENV	1
SAGEMAKER_MODEL_SERVER_WORKERS	1
SAGEMAKER_PROGRAM	inference.py
SM_NUM_GPUS	1

Network

Enable network isolation

True

© 2024 Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Terms Cookies preferences

←→↻

us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/models/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-150

☆Z🔍🔒🔑

aws

Services

Search

[Option+S]

📧🔔🔒⚙️

N. Virginia

AdministratorAccess/edwin

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

Computer vision models

Natural language processing models

Network

Enable network isolation

True

No custom VPC settings applied.

Prospective instances to deploy model

Run Inference recommender job

Loading resources

Tags

Edit

Key	Value
sagemaker-sdk:jumpstart-model-id	meta-textgeneration-llama-2-7b
sagemaker-sdk:jumpstart-inference-config-name	tgi
sagemaker-sdk:jumpstart-model-version	4.9.0

←→↻

us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/endpoints?fo=kvm%253ACreationTimeAfter%253A%202024%2003%253A50%20UTC

☆Z🔍🔒🔑

aws

Services

Search

[Option+S]

📧🔔🔒⚙️

N. Virginia

AdministratorAccess/edwin

RStudio

TensorBoard

Profiler

Notebooks

▼ Admin configurations

Domains

Role manager

Images

Lifecycle configurations

Amazon SageMaker > Endpoints

Endpoints

🔄

Update endpoint

Actions

Create endpoint

Search endpoints

< 1 >

	Name	ARN	Creation time	Status	Last updated
○	meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-148	arn:aws:sagemaker:us-east-1:317957367373:endpoint/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-148	10/8/2024, 11:33:15 PM	InService	10/8/2024, 11:39:02 PM

us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/endpoints/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-148

AWS Services Search [Option+S]

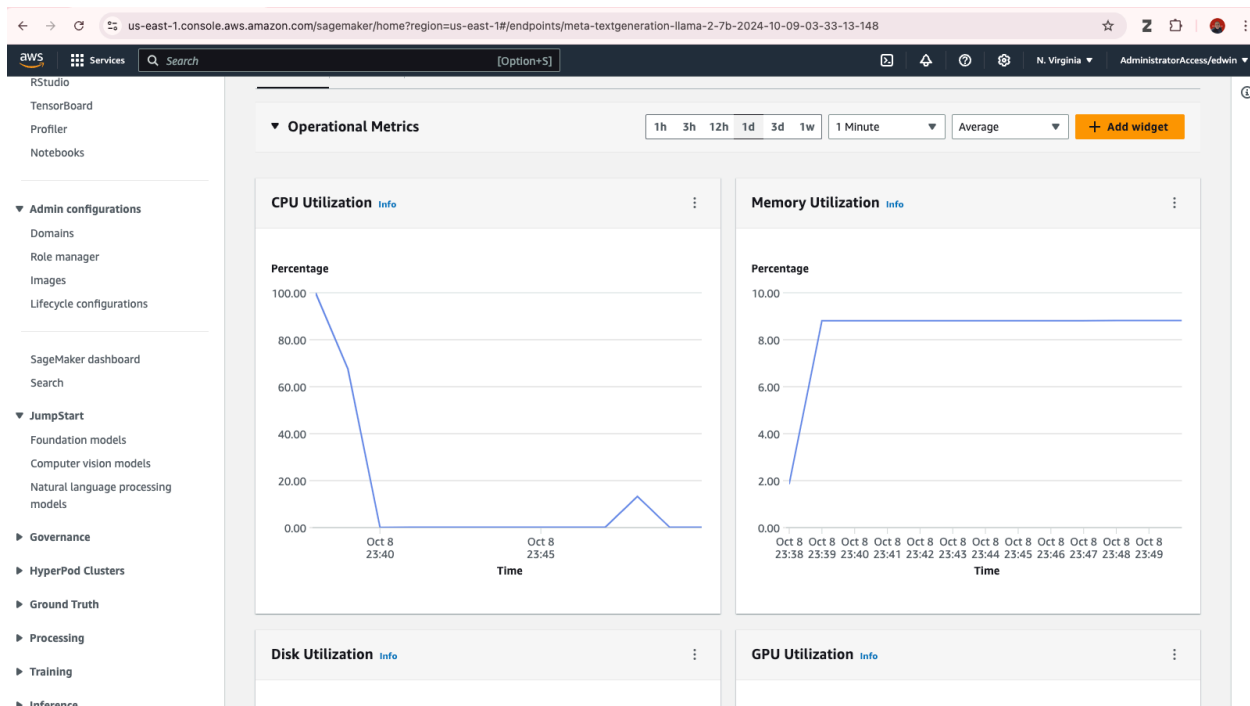
AdministratorAccess/edwin

meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-148 Delete

Endpoint summary

Name meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-148	Status InService	Type Real-time
ARN arn:aws:sagemaker:us-east-1:317957367373:endpoint/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-148	Creation time Tue Oct 08 2024 23:33:15 GMT-0400 (Eastern Daylight Time)	Last updated Tue Oct 08 2024 23:39:02 GMT-0400 (Eastern Daylight Time)
URL https://runtime.sagemaker.us-east-1.amazonaws.com/endpoints/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-148/invocations Learn more about the API	Model container logs /aws/sagemaker/endpoints/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-148	Alarms 0 alarms

Operational Metrics 1h 3h 12h 1d 3d 1w 1 Minute Average + Add widget



us-east-1 console.aws.amazon.com/cloudwatch/home?region=us-east-1#logs-log-group-log-group/\$252Faws-fargate-m252Fmeta-text-generation-llama-2-7b-2024-10-09-03-33-13-148

aws

Services

Search

[Option+5]

🔍

🔔

🔄

⚙️

N. Virginia

AdministratorAccess/edwin

CloudWatch

CloudWatch

Log groups

/aws/sagemaker/Endpoints/meta-textgeneration-llama-2-7b-2024-10-09-03-33-13-148

AllTraffic/I-03633b7a2a6c0ca04

Log events

🔄

Actions

Start tailing

Create metric filter

You can use the filter bar below to search for and match terms, phrases, or values in your log events. [Learn more about filter patterns](#)

Clear

1m

30m

1h

12h

Custom

UTC timezone

Display

⚙️

▶

Timestamp

Message

No older events at this moment. [Retry](#)

▶

2024-10-09T03:38:03.757Z

#033[Zm2024-10-09T03:38:03.049482Z#033[0m #033[32m INFO#033[0m #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m Args : mo...

▶

2024-10-09T03:38:03.757Z

#033[Zm2024-10-09T03:38:03.049543Z#033[0m #033[32m INFO#033[0m #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m Using def...

▶

2024-10-09T03:38:07.974Z

#033[Zm2024-10-09T03:38:03.049603Z#033[0m #033[32m INFO#033[0m #033[Zmload#033[0m: #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m

▶

2024-10-09T03:38:08.476Z

#033[Zm2024-10-09T03:38:07.799307Z#033[0m #033[32m INFO#033[0m #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m Files are...

▶

2024-10-09T03:38:08.476Z

#033[Zm2024-10-09T03:38:08.354800Z#033[0m #033[32m INFO#033[0m #033[Zmload#033[0m: #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m

▶

2024-10-09T03:38:12.757Z

#033[Zm2024-10-09T03:38:08.354965Z#033[0m #033[32m INFO#033[0m #033[Zmshard-manager#033[0m: #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m

▶

2024-10-09T03:38:18.748Z

#033[Zm2024-10-09T03:38:18.365267Z#033[0m #033[32m INFO#033[0m #033[Zmshard-manager#033[0m: #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m

▶

2024-10-09T03:38:18.748Z

#033[Zm2024-10-09T03:38:18.506723Z#033[0m #033[32m INFO#033[0m #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m Server st...

▶

2024-10-09T03:38:18.748Z

#033[Zm2024-10-09T03:38:18.565467Z#033[0m #033[32m INFO#033[0m #033[Zmshard-manager#033[0m: #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m

▶

2024-10-09T03:38:18.748Z

#033[Zm2024-10-09T03:38:18.664383Z#033[0m #033[32m INFO#033[0m #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m Starting ...

▶

2024-10-09T03:38:18.748Z

#033[Zm2024-10-09T03:38:18.719782Z#033[0m #033[32m WARN#033[0m #033[Zmtokenizers::tokenizer::serialization#033[0m#033[Zm: #033[0m

▶

2024-10-09T03:38:18.748Z

#033[Zm2024-10-09T03:38:18.720084Z#033[0m #033[32m INFO#033[0m #033[Zmtext_generation_router#033[0m#033[Zm: #033[0m #033[Zmrou...

▶

2024-10-09T03:38:18.748Z

#033[Zm2024-10-09T03:38:18.720099Z#033[0m #033[32m INFO#033[0m #033[Zmtext_generation_router#033[0m#033[Zm: #033[0m #033[Zmrou...

▶

2024-10-09T03:38:18.748Z

#033[Zm2024-10-09T03:38:18.720112Z#033[0m #033[32m WARN#033[0m #033[Zmtext_generation_router#033[0m#033[Zm: #033[0m #033[Zmrou...

▶

2024-10-09T03:38:22.504Z

#033[Zm2024-10-09T03:38:18.722958Z#033[0m #033[32m INFO#033[0m #033[Zmtext_generation_router#033[0m#033[Zm: #033[0m #033[Zmrou...

▶

2024-10-09T03:38:23.505Z

#033[Zm2024-10-09T03:38:22.368372Z#033[0m #033[32m INFO#033[0m #033[Zmtext_generation_launcher#033[0m#033[Zm: #033[0m Cuda Gran...