



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	3
Step 3: Model Evaluation Section What was the response of the model to your domain-specific input in the model_evaluation.ipynb file?	Traditional approaches to data management such as > SQL and NoSQL are not designed to handle the scale and complexity of modern applications. Data Virtualization is a modern approach to data management that enables organizations to access and analyze all of their data, regardless of location or format. Data Virtualization provides a unified view of data, allowing organizations to make better decisions
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?	Traditional approaches to data management such as > [{'generated_text': " relational databases are no longer adequate to meet the needs of today's enterprise.\n\nIn this webinar, you'll learn about the benefits of the NoSQL database.\n\nThe NoSQL database is

an alternative to the traditional relational database, which is inefficient for handling unstructured data"]}]

- Model in SageMaker:

Amazon SageMaker

Getting started

▼ Applications and IDEs

Studio

Canvas

RStudio

TensorBoard

Notebooks

▼ Admin configurations

Amazon SageMaker

Models

Search models

Create endpoint

Create endpoint configuration

Actions

Create model

	Name	ARN	Creation time
	meta-textgeneration-llama-2-7b-2024-07-04-18-27-37-349	arn:aws:sagemaker:us-east-1:677389480787:model/meta-textgeneration-llama-2-7b-2024-07-04-18-27-37-349	7/4/2024, 10:27:38 PM
	meta-textgeneration-llama-2-7b-2024-07-02-22-24-12-115	arn:aws:sagemaker:us-east-1:677389480787:model/meta-textgeneration-llama-2-7b-2024-07-02-22-24-12-115	7/3/2024, 2:24:12 AM
	meta-textgeneration-llama-2-7b-2024-07-02-22-17-58-696	arn:aws:sagemaker:us-east-1:677389480787:model/meta-textgeneration-llama-2-7b-2024-07-02-22-17-58-696	7/3/2024, 2:17:59 AM
	meta-textgeneration-llama-2-7b-2024-07-02-22-16-39-895	arn:aws:sagemaker:us-east-1:677389480787:model/meta-textgeneration-llama-2-7b-2024-07-02-22-16-39-895	7/3/2024, 2:16:40 AM

meta-textgeneration-llama-2-7b-2024-07-04-18-27-37-349

Actions

Create batch transform job

Create endpoint

Model settings

Name

meta-textgeneration-llama-2-7b-2024-07-04-18-27-37-349

ARN

arn:aws:sagemaker:us-east-1:677389480787:model/meta-textgeneration-llama-2-7b-2024-07-04-18-27-37-349

Creation time

7/4/2024, 10:27:38 PM

IAM role ARN

arn:aws:iam::677389480787:role/service-role/SageMaker-ProjectManager

Container 1

Container Name

Container 1

Image

763104551884.dkr.ecr.us-east-1.amazonaws.com/huggingface-pytorch-fgi-inference:2.1.1-tgi2.0.0-gpu-py310-cu121-ubuntu22.04

Training job

-

S3 URI

s3://sagemaker-us-east-1-677389480787/meta-textgeneration-llama-2-7b-2024-07-04-18-12-09-077/output/model/

Environment variables

Key	Value
ENDPOINT_SERVER_TIMEOUT	3600
HF_MODEL_ID	/opt/ml/model
MAX_INPUT_LENGTH	4095
MAX_TOTAL_TOKENS	4096
MODEL_CACHE_ROOT	/opt/ml/model

Model data location

-

Mode

Single model

Compression Type

None

S3 data type

S3Prefix