



# Introducing Generative AI with AWS

## Project: Building a Domain Expert Model

### Environment and Project Setup

Configured and completed the below steps and used Aws US West Oregon (us-west-2) Region.

- An AWS SageMaker IAM Role
- An AWS SageMaker Notebook Instance
- A GPU instance for fine-tuning training
- Downloaded the [project starter files](#).

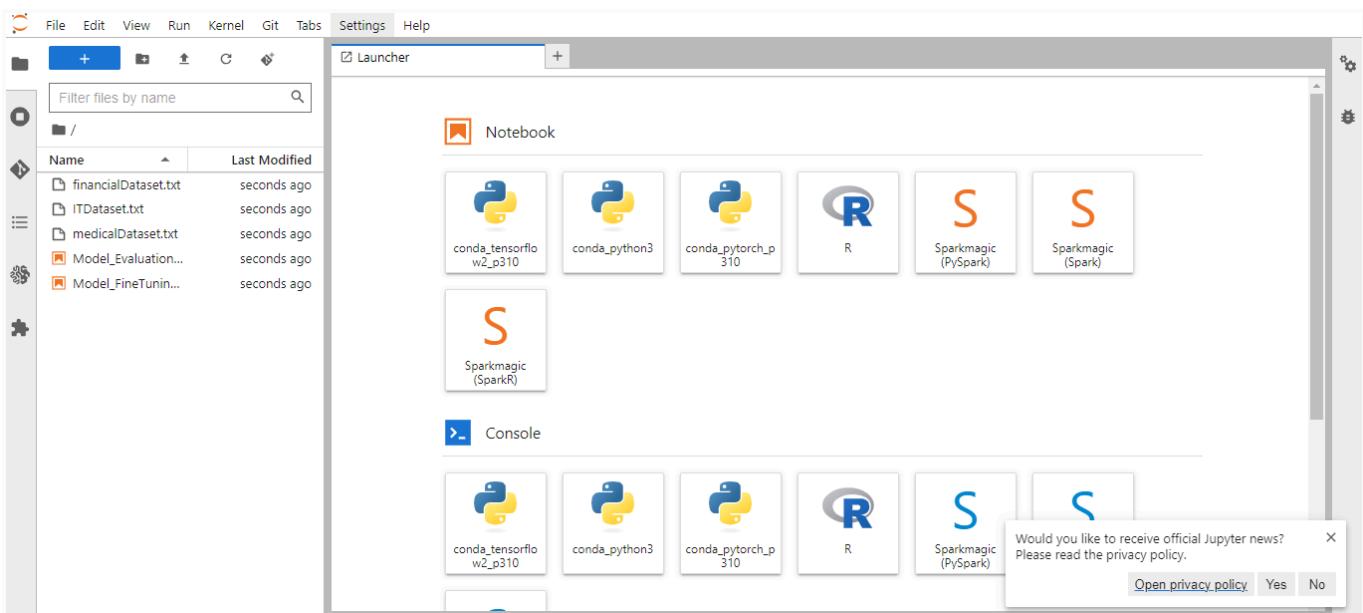
### Step 1: Upload Project Starter Files

- Creating and Running the Fine-tuningProject Instance

A screenshot of the AWS SageMaker console. The top navigation bar shows 'Services' and the search bar. The main content area is titled 'Amazon SageMaker > Notebook instances'. It displays a table of notebook instances with columns: Name, Instance, Creation time, Status, and Actions. One instance, 'Fine-tuningProject', is listed with details: ml.t3.medium, 5/17/2024, 11:20:10 AM, InService, and links to 'Open Jupyter' and 'Open JupyterLab'. On the left sidebar, there's a 'Getting started' section with links to 'Studio', 'Studio Lab', 'Canvas', 'RStudio', and 'TensorBoard'. Below that is an 'Admin configurations' section with links to 'Domains', 'Role manager', 'Images', and 'Lifecycle configurations'.

Name	Instance	Creation time	Status	Actions
Fine-tuningProject	ml.t3.medium	5/17/2024, 11:20:10 AM	InService	<a href="#">Open Jupyter</a>   <a href="#">Open JupyterLab</a>

- Uploading the Python notebook files (.ipynb)



## Step 2: Choose your Dataset

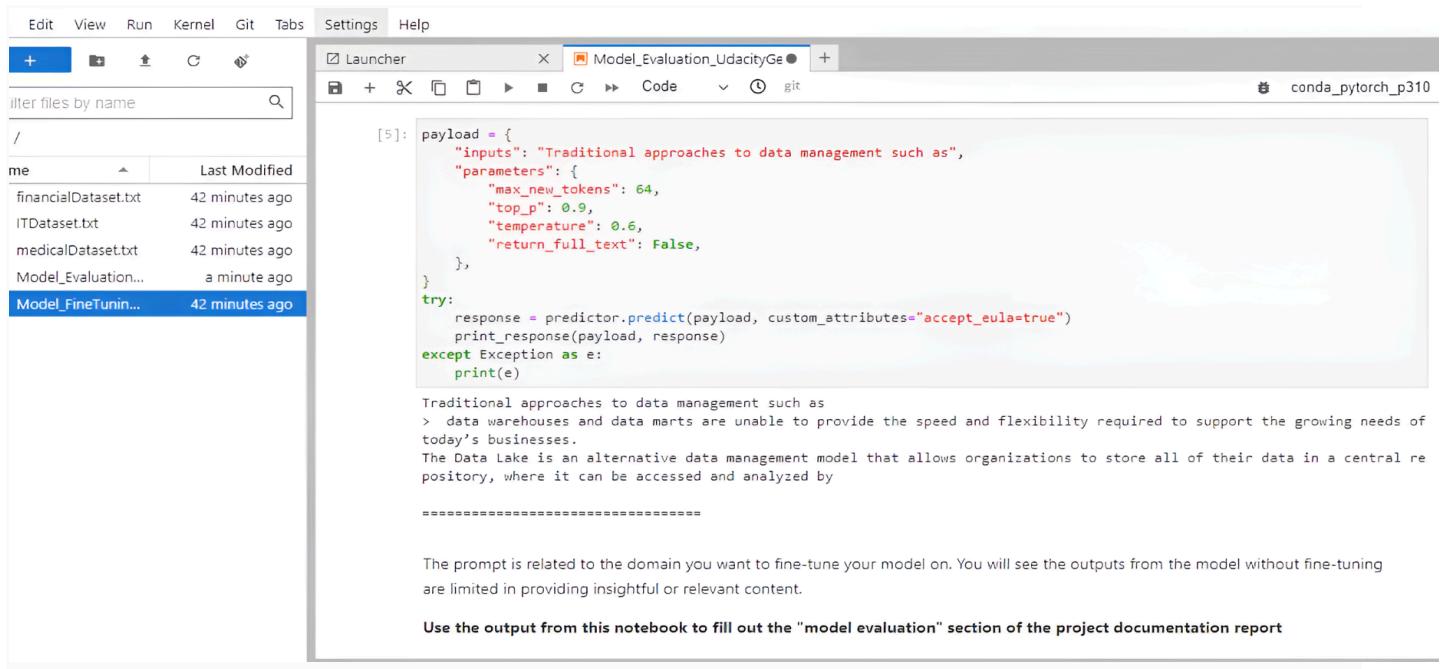
I have opted for a dataset within the domain of **information technology (IT)**.

## Step 3: Deploy and Evaluate the model (Model\_Evaluation.ipynb)

- Completed and ran the cells in the Model\_Evaluation.ipynb file
- Screenshot of the Model Deployment of Model\_Evaluation.ipynb

The screenshot shows the AWS SageMaker console under the 'Endpoints' section. The left sidebar includes links for 'Getting started', 'Studio', 'Studio Lab', 'Canvas', 'RStudio', 'TensorBoard', 'Admin configurations' (with 'Domains', 'Role manager', 'Images', 'Lifecycle configurations'), 'SageMaker dashboard', and 'Search'. The main 'Endpoints' table has columns for Name, ARN, Creation time, Status, and Last updated. One endpoint is listed: 'meta-textgeneration-llama-2-7b-2024-05-17-06-41-07-811' with ARN 'arn:aws:sagemaker:us-west-2:979246567187:endpoint/meta-textgeneration-llama-2-7b-2024-05-17-06-41-07-811', created on 5/17/2024 at 12:11:09 PM, and status 'InService'.

- Saved and download Model\_Evaluation.ipynb with the cell output, uploaded in the zip file and file name is ***Model\_Evaluation\_UdacityGenAIAWS.ipynb***
- Screenshot of the Model\_Evaluation.ipynb file with the cell output as proof



```

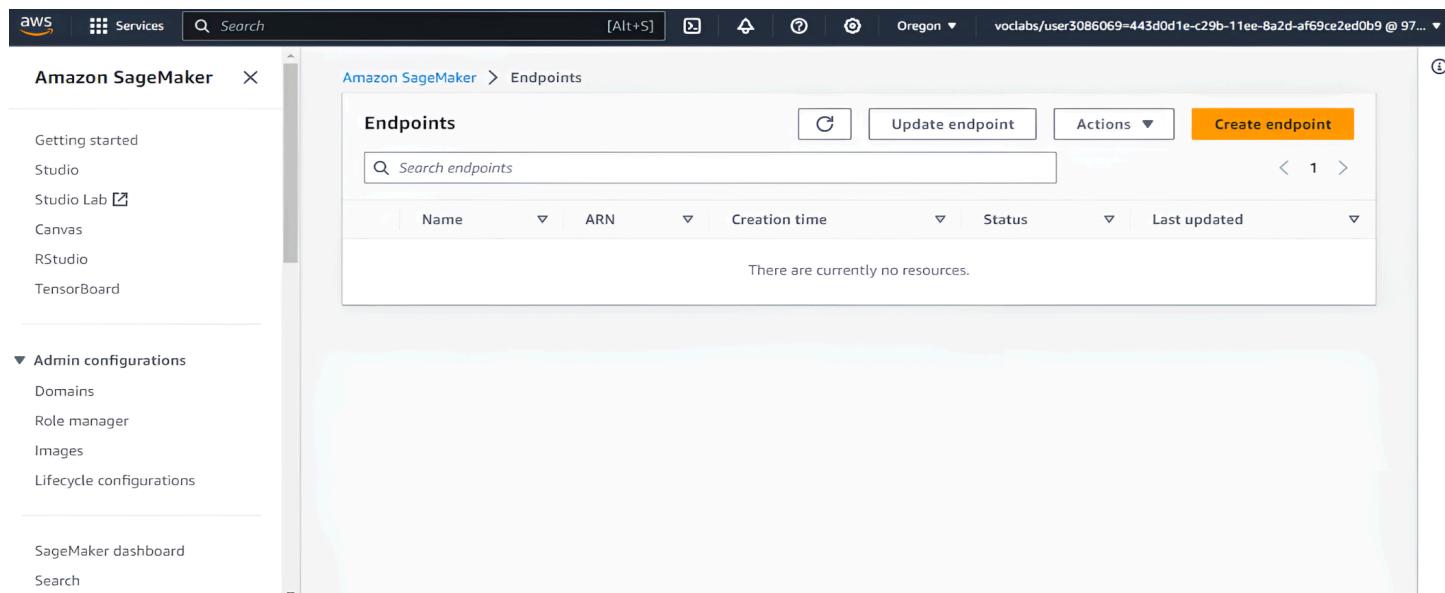
Edit View Run Kernel Git Tabs Settings Help
+ Launcher Model_Evaluation_UdacityGe + conda_pytorch_p310
filter files by name
/ me Last Modified
financialDataset.txt 42 minutes ago
ITDataset.txt 42 minutes ago
medicalDataset.txt 42 minutes ago
Model_Evaluation... a minute ago
Model_FineTunin... 42 minutes ago
[5]: payload = {
    "inputs": "Traditional approaches to data management such as",
    "parameters": {
        "max_new_tokens": 64,
        "top_p": 0.9,
        "temperature": 0.6,
        "return_full_text": False,
    },
}
try:
    response = predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)

Traditional approaches to data management such as
> data warehouses and data marts are unable to provide the speed and flexibility required to support the growing needs of
today's businesses.
The Data Lake is an alternative data management model that allows organizations to store all of their data in a central re
pository, where it can be accessed and analyzed by
=====
The prompt is related to the domain you want to fine-tune your model on. You will see the outputs from the model without fine-tuning
are limited in providing insightful or relevant content.

Use the output from this notebook to fill out the "model evaluation" section of the project documentation report

```

- Screenshot of Deleted the Model Deployment and endpoints



- Updated the Project Documentation section about the evaluation of the model's text generation capabilities and knowledge.

## Step 4: Fine-tune the Model (Model\_FineTuning.ipynb)

- Completed and ran the cells in the Model\_FineTuning.ipynb
- Saved and download Model\_Evaluation.ipynb with the cell output, uploaded in the zip file and file name is **Model\_FineTuning.ipynb**
- screenshot of the Model\_FineTuning.ipynb file with the cell output as proof

```
[5]: payload = {
    "inputs": "Traditional approaches to data management such as",
    "parameters": {
        "max_new_tokens": 64,
        "top_p": 0.9,
        "temperature": 0.6,
        "return_full_text": False,
    },
}
try:
    response = finetuned_predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)

Traditional approaches to data management such as
> [{"generated_text": "relational databases and data warehouses have become obsolete in today's dynamic business environment.\nThe NoSQL movement has grown in response to the need for more agile data management systems that can adapt to changing business needs.\nNoSQL is a term used to describe a class of non-relational databases."}]

=====
Do the outputs from the fine-tuned model provide domain-specific insightful and relevant content? You can continue experimenting with the inputs of the model to test its domain knowledge.
```

- Updated the Project Documentation Report section about fine-tuning the model

## Step 5: Deploy and Evaluate the Fine-tuned Model

- Screenshot of Visiting the AWS S3 bucket where my fine-tuned model weights are stored after training for your submission.

The screenshot shows the AWS S3 console under the 'General purpose buckets' tab. There is one bucket listed:

Name	AWS Region	IAM Access Analyzer	Creation date
sagemaker-us-west-2-979246567187	US West (Oregon) us-west-2	<a href="#">View analyzer for us-west-2</a>	May 17, 2024, 12:35:21 (UTC+05:30)

**ARN** - arn:aws:s3:::sagemaker-us-west-2-979246567187

- Screenshot of Model\_FineTuning.ipynb file about deploying and evaluating the fine-tuned model.

The screenshot shows the Amazon SageMaker console under the 'Endpoints' tab. One endpoint is listed:

Name	ARN	Creation time	Status	Last updated
meta-textgeneration-llama-2-7b-2024-05-17-07-20-14-714	arn:aws:sagemaker:us-west-2:979246567187:endpoint/meta-textgeneration-llama-2-7b-2024-05-17-07-20-14-714	5/17/2024, 12:50:16 PM	<span>InService</span>	5/17/2024, 12:55:41 PM

- Screenshot of the Model\_FineTuning.ipynb file with the cell output as proof

The screenshot shows a Jupyter Notebook interface with several tabs at the top: File, Edit, View, Run, Kernel, Git, Tabs, Settings, Help. The current tab is 'Model\_FineTuning.ipynb'. On the left, there's a file browser showing local files like financialDataset.txt, ITDataset.txt, medicalDataset.txt, Model\_Evaluation..., and Model\_FineTuning.ipynb. The main area displays a code cell and its output. The code cell contains Python code for fine-tuning a model. The output shows the generated text response from the model.

```
[5]: payload = {
    "inputs": "Traditional approaches to data management such as",
    "parameters": {
        "max_new_tokens": 64,
        "top_p": 0.9,
        "temperature": 0.6,
        "return_full_text": False,
    }
}
try:
    response = finetuned_predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)

Traditional approaches to data management such as
> [{"generated_text": " relational databases and data warehouses have become obsolete in today's dynamic business environment.\nThe NoSQL movement has grown in response to the need for more agile data management systems that can adapt to changing business needs.\nNoSQL is a term used to describe a class of non-rel"}]
```

Do the outputs from the fine-tuned model provide domain-specific insightful and relevant content? You can continue experimenting with the inputs of the model to test it's domain knowledge.

- Screenshot of Deleted the Model Deployment and endpoints(Cell Output)

The screenshot shows a Jupyter Notebook interface with several tabs at the top: File, Edit, View, Run, Kernel, Git, Tabs, Settings, Help. The current tab is 'Model\_FineTuning.ipynb'. On the left, there's a file browser showing local files like ITDataset.txt, medicalDataset.txt, Model\_Evaluation..., and Model\_FineTuning.ipynb. The main area displays a code cell and its output. The code cell contains Python code for deleting the model and endpoint. The output shows the log messages from AWS Sagemaker confirming the deletion.

```
[6]: finetuned_predictor.delete_model()
finetuned_predictor.delete_endpoint()

INFO:sagemaker:Deleting model with name: meta-textgeneration-llama-2-7b-2024-05-17-07-20-14-718
INFO:sagemaker:Deleting endpoint configuration with name: meta-textgeneration-llama-2-7b-2024-05-17-07-20-14-714
INFO:sagemaker:Deleting endpoint with name: meta-textgeneration-llama-2-7b-2024-05-17-07-20-14-714
```

- Screenshot of Deleted the Model Deployment and endpoints(Cell Output)

The screenshot shows the AWS SageMaker console. The left sidebar includes 'Natural language processing' under 'models', and sections for Governance, HyperPod Clusters, Ground Truth, Notebook, Processing, Training, Inference, Augmented AI, and AWS Marketplace. Under 'Tutorials', there's a link to 'Documentation'. The main content area is titled 'Amazon SageMaker > Endpoints'. It shows a table with columns: Name, ARN, Creation time, Status, and Last updated. A message at the bottom says 'There are currently no resources.'

- Updated the Project Documentation Report section about fine-tuning the model.

## Step 6: Collect Project Documentation and Submit

**Zip File Named Project\_Building\_a\_Domain\_Expert\_Model.zip is uploaded and consists of**

- Model\_evaluation.ipynb with cell output (File\_Name:-Model\_evaluation.ipynb).
- Model\_FineTuning.ipynb with cell output (File\_Name:-Model\_FineTuning.ipynb).
- Screenshots of both notebooks with cell output.
- The completed Project Documentation Report (File\_Name:- UDACITY Introduction to Generative AI with AWS Project Documentation Report).
- Snapshots folder consists of both the notebooks with outputs for better visibility.