UDACITY

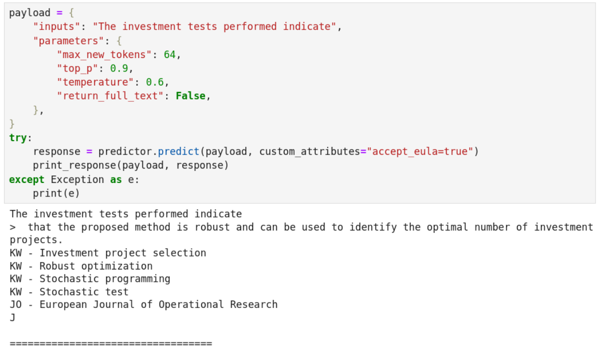
Introduction to Generative AI with AWS

Project Documentation Report

By: Ahmed Shaban

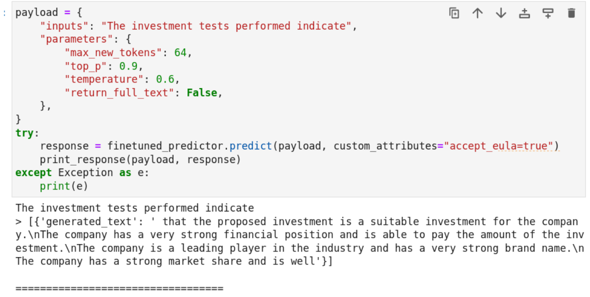
Complete the answers to the questions below to complete your project report. Create a PDF of the completed document and submit the PDF with your project.

Question	Your answer:
Step 2: Domain Choice What domain did you choose to fine-tune the Meta Llama 2 7B model on? Choices: 1. Financial 2. Healthcare 3. IT	Financial Domain

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 method is robust and can be used to identify the optimal number of investment projects.
	KW - Investment project selection
	KW - Robust optimization
	KW - Stochastic programming
	KW - Stochastic test
	JO - European Journal of Operational Research
	J
	
	`payload = {`
	`    "inputs": "The investment tests performed indicate",`
	`    "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)`
	`The investment tests performed indicate`
	`>  that the proposed method is robust and can be used to identify the optimal number of investment projects.`
	`KW - Investment project selection`
	`KW - Robust optimization`
	`KW - Stochastic programming`
	`KW - Stochastic test`
	`JO - European Journal of Operational Research`
	`J`
	`==================================`

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 proposed investment is a suitable investment for the company.\nThe company has a very strong financial position and is able to pay the amount

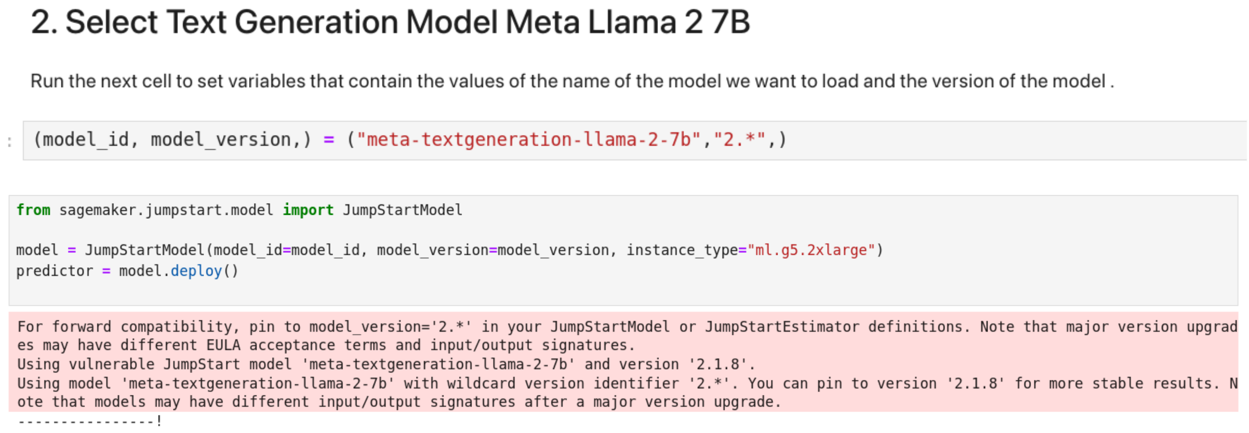
of the investment.\nThe company is a leading player in the industry and has a very strong brand name.\nThe company has a strong market share and is well'}]



```
payload = {
    "inputs": "The investment tests performed indicate",
    "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)
```

```
The investment tests performed indicate
> [{'generated_text': ' that the proposed investment is a suitable investment for the compan
y.\nThe company has a very strong financial position and is able to pay the amount of the inv
estment.\nThe company is a leading player in the industry and has a very strong brand name.\n
The company has a strong market share and is well'}]

==================================
```

Deploy the Llama2 Model on AWS Sagemaker



2. Select Text Generation Model Meta Llama 2 7B

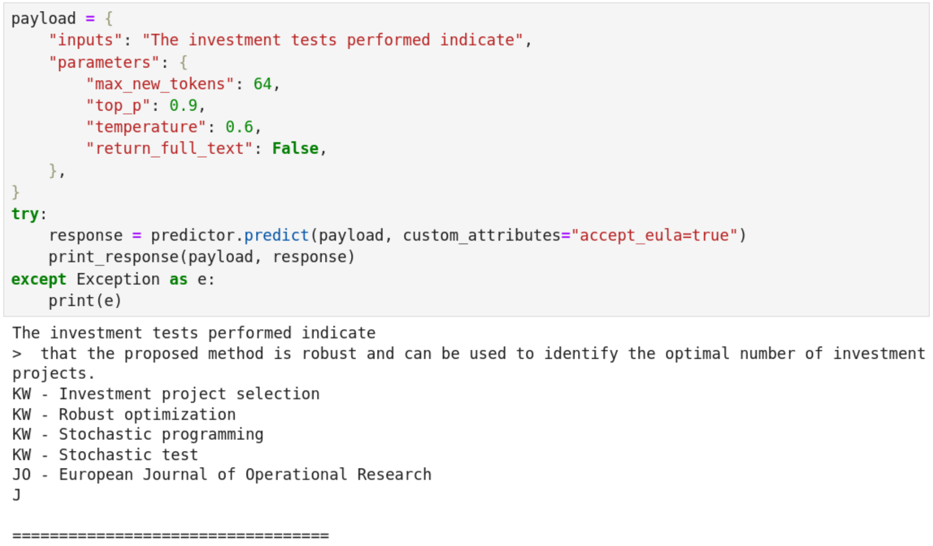
Run the next cell to set variables that contain the values of the name of the model we want to load and the version of the model .

```
(model_id, model_version,) = ("meta-textgeneration-llama-2-7b","2.*",)
```

```
from sagemaker.jumpstart.model import JumpStartModel

model = JumpStartModel(model_id=model_id, model_version=model_version, instance_type="ml.g5.2xlarge")
predictor = model.deploy()
```

```
For forward compatibility, pin to model_version='2.*' in your JumpStartModel or JumpStartEstimator definitions. Note that major version upgrad
es may have different EULA acceptance terms and input/output signatures.
Using vulnerable JumpStart model 'meta-textgeneration-llama-2-7b' and version '2.1.8'.
Using model 'meta-textgeneration-llama-2-7b' with wildcard version identifier '2.*'. You can pin to version '2.1.8' for more stable results. N
ote that models may have different input/output signatures after a major version upgrade.
----------------!
```

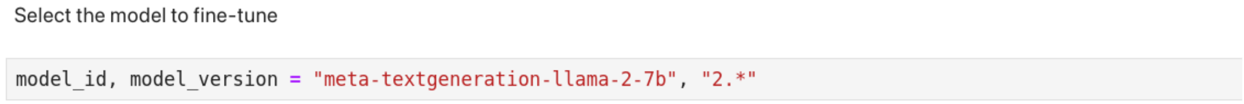
Screenshot (below) of Step 3: **Model Evaluation Section:** What was the response of the model to your domain-specific input in the **model_evaluation.ipynb file**?



```
payload = {
    "inputs": "The investment tests performed indicate",
    "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)
```

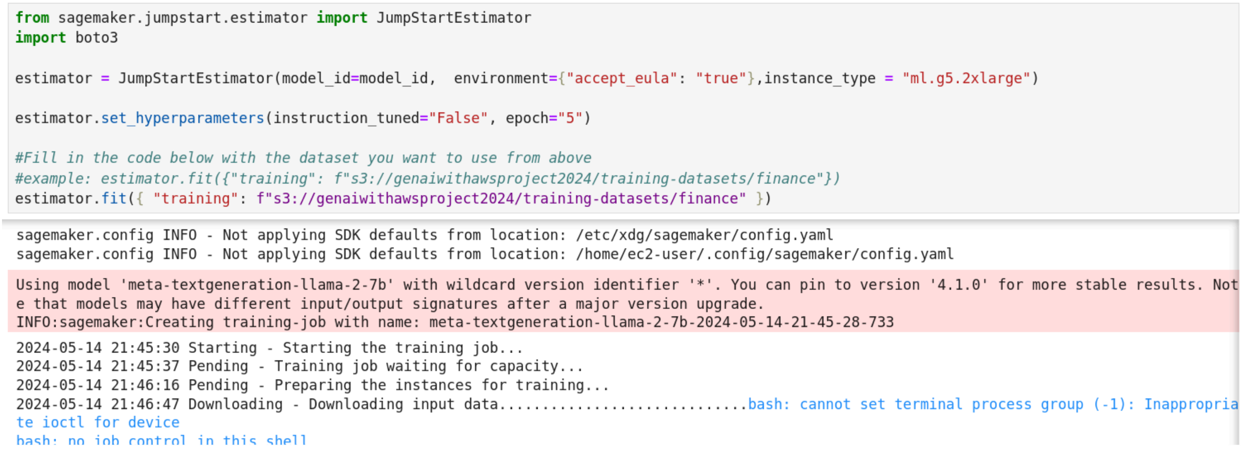
```
The investment tests performed indicate
>  that the proposed method is robust and can be used to identify the optimal number of investment
projects.
KW - Investment project selection
KW - Robust optimization
KW - Stochastic programming
KW - Stochastic test
JO - European Journal of Operational Research
J

==================================
```

Fine-tune a Large Language Model with a Domain-Specific Dataset (finance)



Select the model to fine-tune

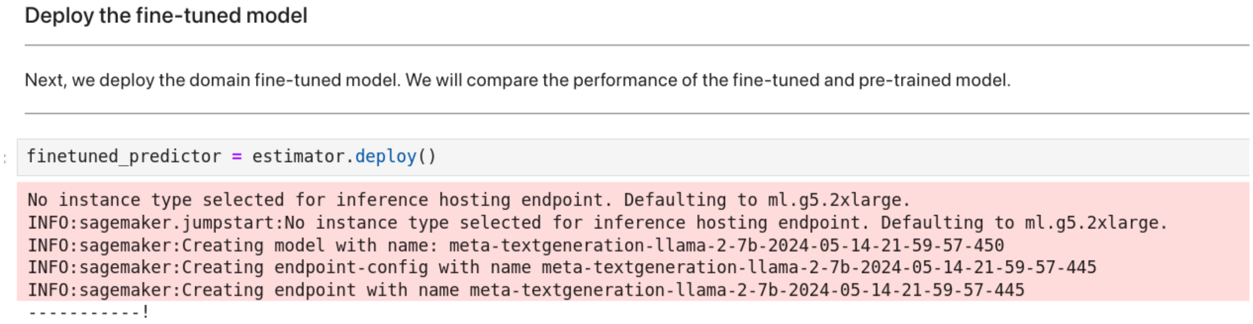
```
model_id, model_version = "meta-textgeneration-llama-2-7b", "2.*"
```



```
from sagemaker.jumpstart.estimator import JumpStartEstimator
import boto3

estimator = JumpStartEstimator(model_id=model_id,  environment={"accept_eula": "true"},instance_type = "ml.g5.2xlarge")

estimator.set_hyperparameters(instruction_tuned="False", epoch="5")

#Fill in the code below with the dataset you want to use from above
#example: estimator.fit({"training": f"s3://genaiwithawsproject2024/training-datasets/finance"})
estimator.fit({ "training": f"s3://genaiwithawsproject2024/training-datasets/finance" })
```

```
sagemaker.config INFO - Not applying SDK defaults from location: /etc/xdg/sagemaker/config.yaml
sagemaker.config INFO - Not applying SDK defaults from location: /home/ec2-user/.config/sagemaker/config.yaml
Using model 'meta-textgeneration-llama-2-7b' with wildcard version identifier '*'. You can pin to version '4.1.0' for more stable results. Not
e that models may have different input/output signatures after a major version upgrade.
INFO:sagemaker:Creating training-job with name: meta-textgeneration-llama-2-7b-2024-05-14-21-45-28-733
2024-05-14 21:45:30 Starting - Starting the training job...
2024-05-14 21:45:37 Pending - Training job waiting for capacity...
2024-05-14 21:46:16 Pending - Preparing the instances for training...
2024-05-14 21:46:47 Downloading - Downloading input data............................bash: cannot set terminal process group (-1): Inappropria
te ioctl for device
bash: no job control in this shell
```

Deploy the Fine-tuned Llama2 Model on AWS Sagemaker



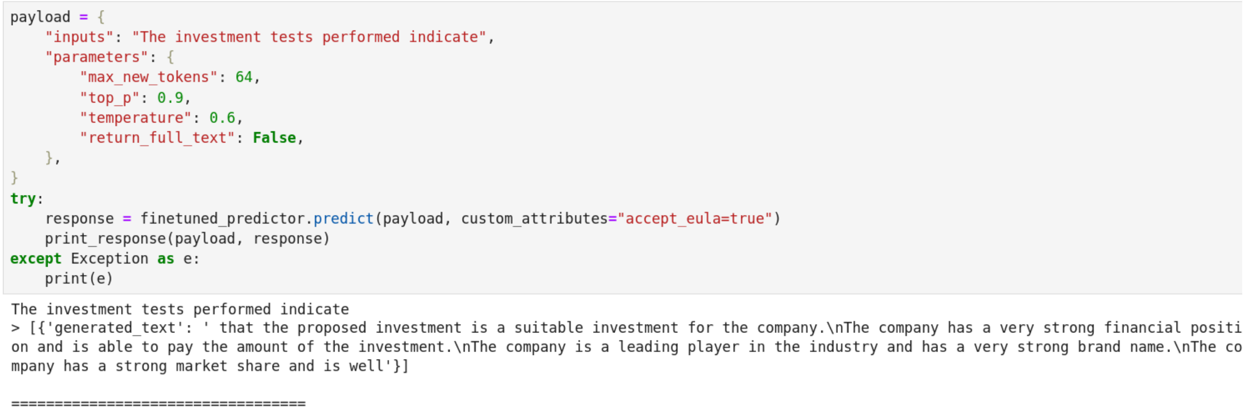
Deploy the fine-tuned model

Next, we deploy the domain fine-tuned model. We will compare the performance of the fine-tuned and pre-trained model.

```
finetuned_predictor = estimator.deploy()
```

```
No instance type selected for inference hosting endpoint. Defaulting to ml.g5.2xlarge.
INFO:sagemaker.jumpstart:No instance type selected for inference hosting endpoint. Defaulting to ml.g5.2xlarge.
INFO:sagemaker:Creating model with name: meta-textgeneration-llama-2-7b-2024-05-14-21-59-57-450
INFO:sagemaker:Creating endpoint-config with name meta-textgeneration-llama-2-7b-2024-05-14-21-59-57-445
INFO:sagemaker:Creating endpoint with name meta-textgeneration-llama-2-7b-2024-05-14-21-59-57-445
-----------!
```

Screenshot (below) of 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**?



```
payload = {
    "inputs": "The investment tests performed indicate",
    "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)
```

```
The investment tests performed indicate
> [{'generated_text': ' that the proposed investment is a suitable investment for the company.\nThe company has a very strong financial positi
on and is able to pay the amount of the investment.\nThe company is a leading player in the industry and has a very strong brand name.\nThe co
mpany has a strong market share and is well'}]

==================================
```