



## 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:
<b>Step 2: Domain Choice</b> What domain did you choose to fine-tune the Meta Llama 2 7B model on? Choices: <ol style="list-style-type: none"><li>1. Financial</li><li>2. Healthcare</li><li>3. IT</li></ol>	I choose the IT domain to fine-tune the Meta Llama 2 7B model
<b>Step 3: Model Evaluation Section</b> What was the response of the model to your domain-specific input in the <b>model_evaluation.ipynb</b> file?	<p>For the prompts:"</p> <p>Traditional approaches to data management such as" - response was " data warehousing and data marts are ineffective in today's world of big data. In this session, we will discuss the challenges of data management in the big data world, the role of data virtualization in addressing those challenges, and how to get started with data virtual</p> <ul style="list-style-type: none"><li>- Prompt "because ubiquitous computing is intended to"</li><li>- Response "be used by people with a wide range of capabilities. This is not to say that ubiquitous computing cannot be used by people with disabilities. Rather, it is to say that it must be used by people with disabilities in order to be useful to people with disabilities. As a result"</li><li>- Prompt "A second important aspect of ubiquitous computing</li></ul>

	<p>environments is”</p> <ul style="list-style-type: none"> <li>- Response “that they must be designed to be as user-friendly as possible. The user-friendly aspect of ubiquitous computing is a result of the fact that the user should be able to use the system without being aware of the underlying technology. In this way, the system should appear to be an ordinary application“</li> <li>- Prompt “outline the key aspects of ubiquitous computing from a data management perspective.”</li> <li>- We highlight the unique requirements for the management of ubiquitous data and the challenges that are posed by the requirements. We argue that ubiquitous computing and data management are complementary and that ubiquitous computing is a key enabler of data management. We provide a set of guideline</li> <li>-</li> </ul>
<p><b>Step 4: Fine-Tuning Section</b> After fine-tuning the model, what was the response of the model to your domain-specific input in the <b>model_finetuning.ipynb</b> file?</p>	<ul style="list-style-type: none"> <li>- Prompt for fine-tuned model “Traditional approaches to data management such as”</li> <li>- Response “ETL (Extract, Transform and Load) and ELT (Extract, Load and Transform) are becoming less and less efficient as the volume and variety of data grows. Traditional data warehouse solutions are also becoming obsolete. In the past, data was stored in one or more</li> <li>- Prompt “A second important aspect of ubiquitous computing environments is”</li> <li>- Response “that they are not limited to a single user. In fact, many applications can be used by multiple users simultaneously. For example, a user might want to share a personal digital assistant with other members of a family or</li> </ul>

	<p>office. As another example, a number of users might want to share a set of virtual reality goggles</p> <ul style="list-style-type: none"><li>- Prompt "because ubiquitous computing is intended to"</li><li>- Response "be everywhere. The key to the success of ubiquitous computing is to make computing devices and services so seamless that users are unaware of their existence.  <a href="https://github.com/yangshun/yangshun.github.io/blob">https://github.com/yangshun/yangshun.github.io/blob</a></li><li>- Prompt "outline the key aspects of ubiquitous computing from a data management perspective."</li><li>- Response "In this paper, we focus on the challenges and issues of data management in ubiquitous computing. We focus on the key issues of data management in ubiquitous computing. We first discuss the challenges of ubiquitous computing in the context of data management. We then focus on the key"</li></ul>
--	---

## Base model deployed

**The next cell will take some time to run.** It is deploying a large language model, and that takes time. You'll see dashes (--) while it is being deployed. Please be patient! You'll see an exclamation point at the end of the dashes (---!) when the model is deployed and then you can continue running the next cells.

You might see a warning "For forward compatibility, pin to model\_version..." You can ignore this warning, just wait for the model to deploy.

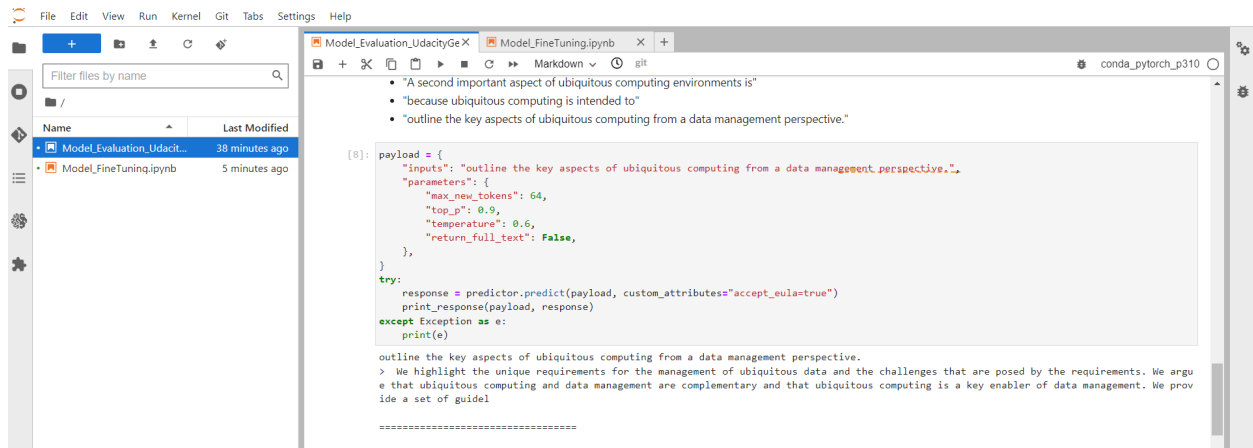
```
[3]: from sagemaker.jumpstart.model import JumpStartModel
```

```
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 upgrades 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.  
Note that models may have different input/output signatures after a major version upgrade.  
-----!

-----!

## Deployed base model response



## Deployed fine tuned model

### ▼ 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.

```
4]: 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-2024-03-10-07-44-55-888  
INFO:sagemaker:Creating endpoint-config with name meta-textgeneration-llama-2-7b-2024-03-10-07-44-55-885  
INFO:sagemaker:Creating endpoint with name meta-textgeneration-llama-2-7b-2024-03-10-07-44-55-885  
-----!

## Deployed fine tuned model response

**for llm domain:**

"inputs": "Replace with sentence below from text"

- "Traditional approaches to data management such as"
- "A second important aspect of ubiquitous computing environments is"
- "because ubiquitous computing is intended to"
- "outline the key aspects of ubiquitous computing from a data management perspective."

```
[12]: payload = {
    "inputs": "outline the key aspects of ubiquitous computing from a data management perspective.",
    "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)
```

outline the key aspects of ubiquitous computing from a data management perspective.

>

In this paper, we focus on the challenges and issues of data management in ubiquitous computing. We focus on the key issues of data management in ubiquitous computing. We first discuss the challenges of ubiquitous computing in the context of data management. We then focus on the key

## Model output saved in s3 bucket

The screenshot shows the Amazon S3 console interface. The breadcrumb navigation indicates the path: Amazon S3 > Buckets > sagemaker-us-west-2-059030573350 > meta-textgeneration-llama-2-7b-2024-03-10-07-30-39-027/ > output/ > model/. The 'model/' folder is selected, and the 'Objects' tab is active. A table lists the objects in the folder:

	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	__script_info__.json	json	March 10, 2024, 03:43:35 (UTC-04:00)	170.0 B	Standard
<input type="checkbox"/>	added_tokens.json	json	March 10, 2024, 03:44:06 (UTC-04:00)	21.0 B	Standard