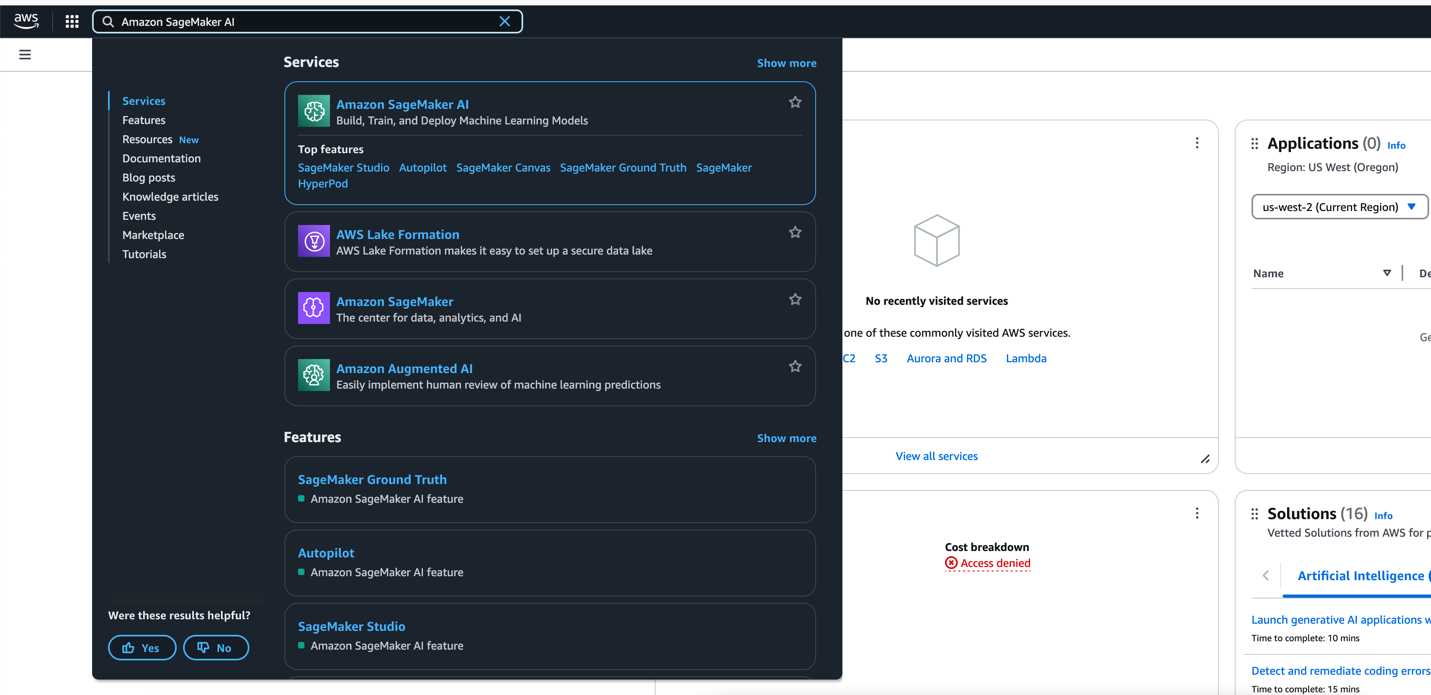
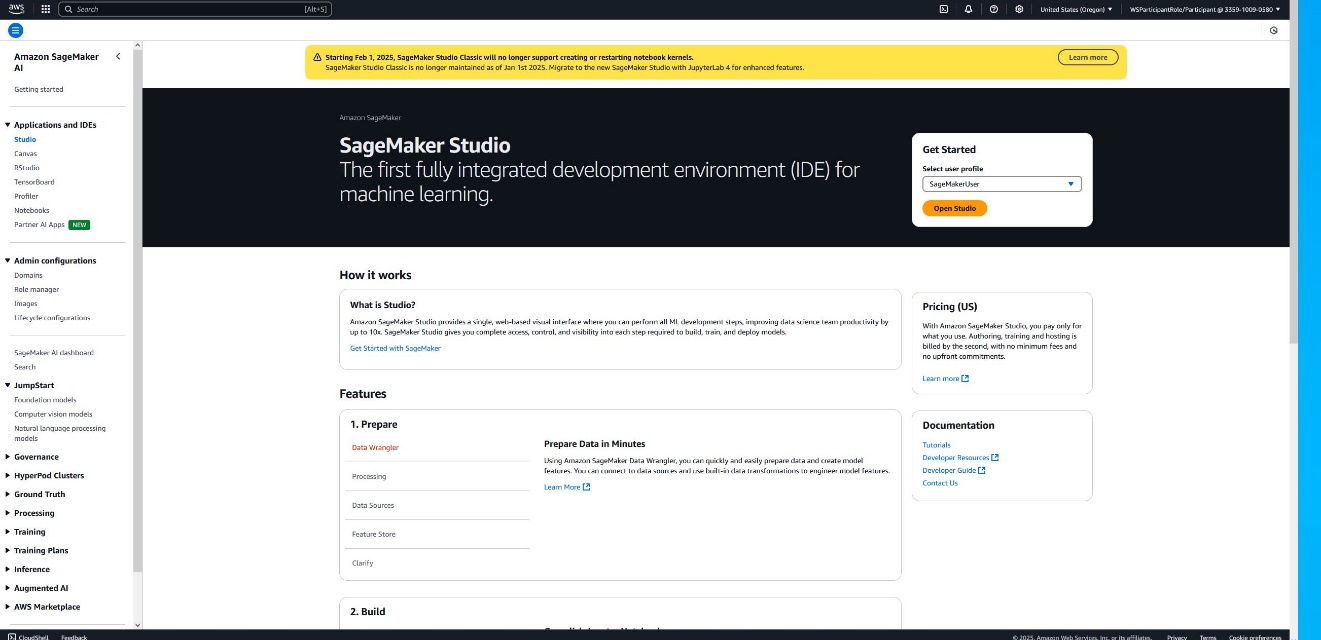
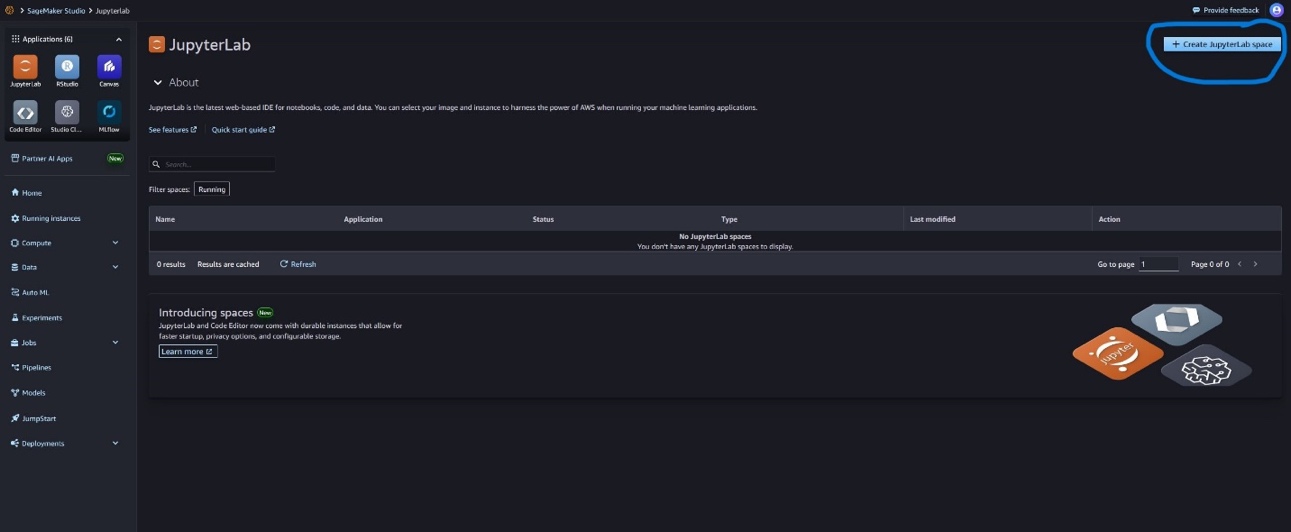
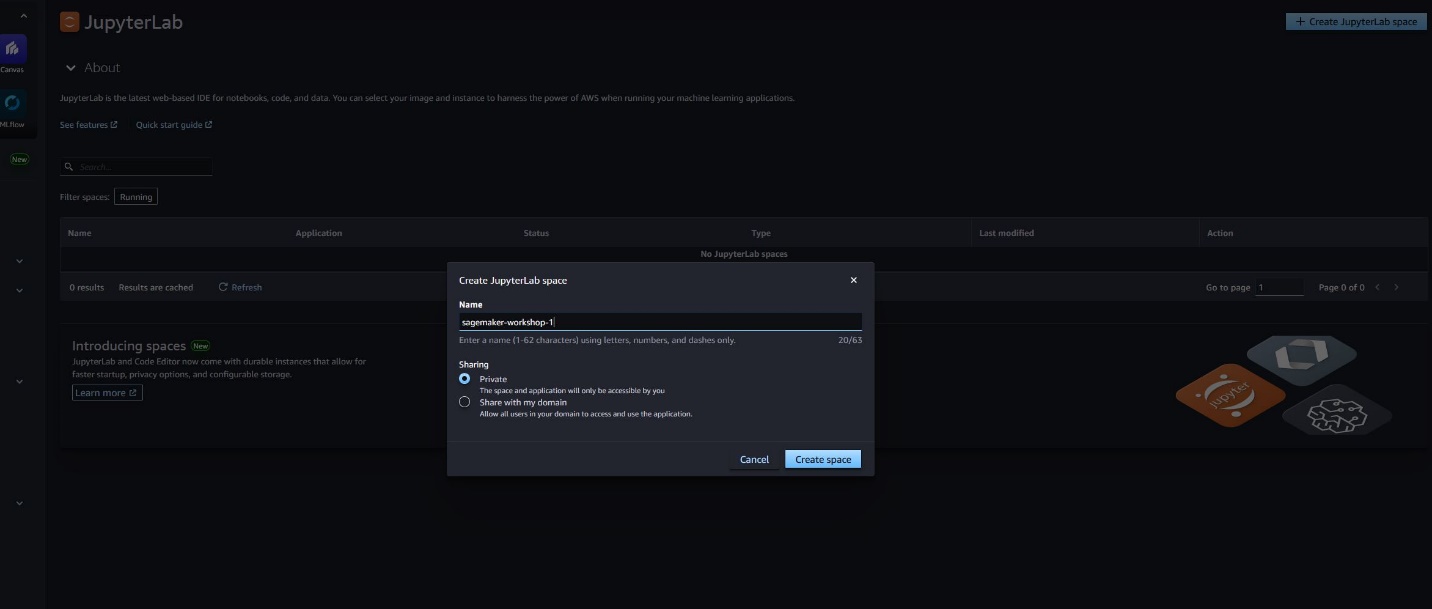
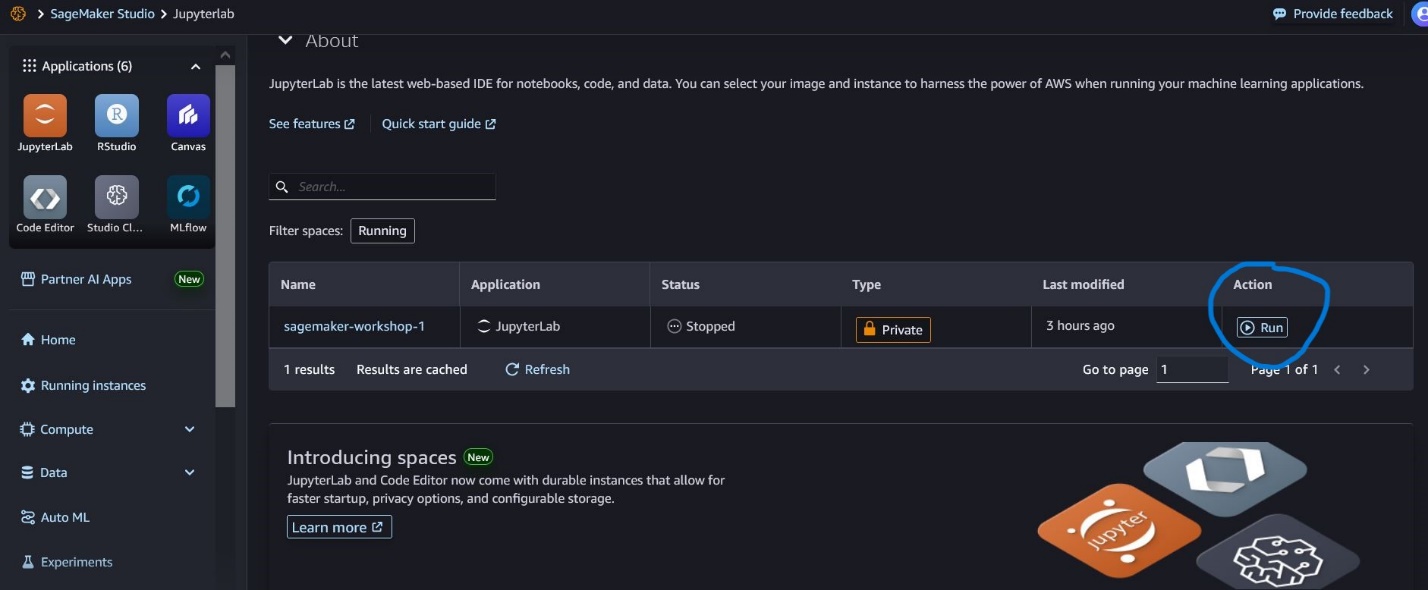
**Step 1: Deploying Embedding and BGE models via SageMaker's Jumpstart**

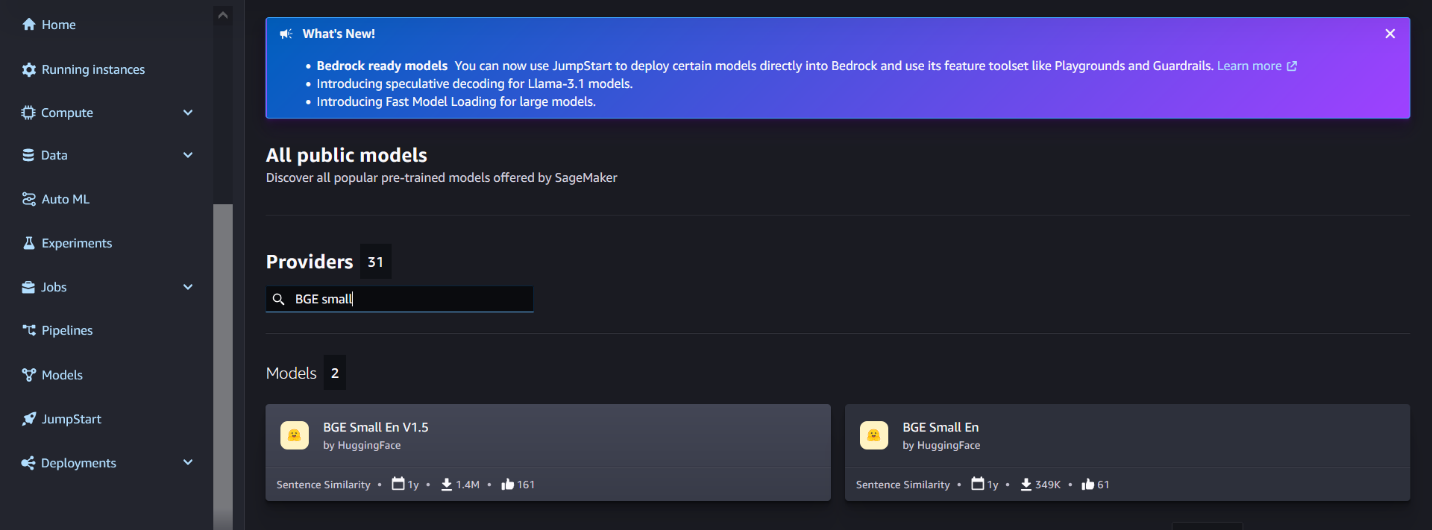
**Lab 1.1: Deploying Embedding BGE Embedding Model via Jumpstar**

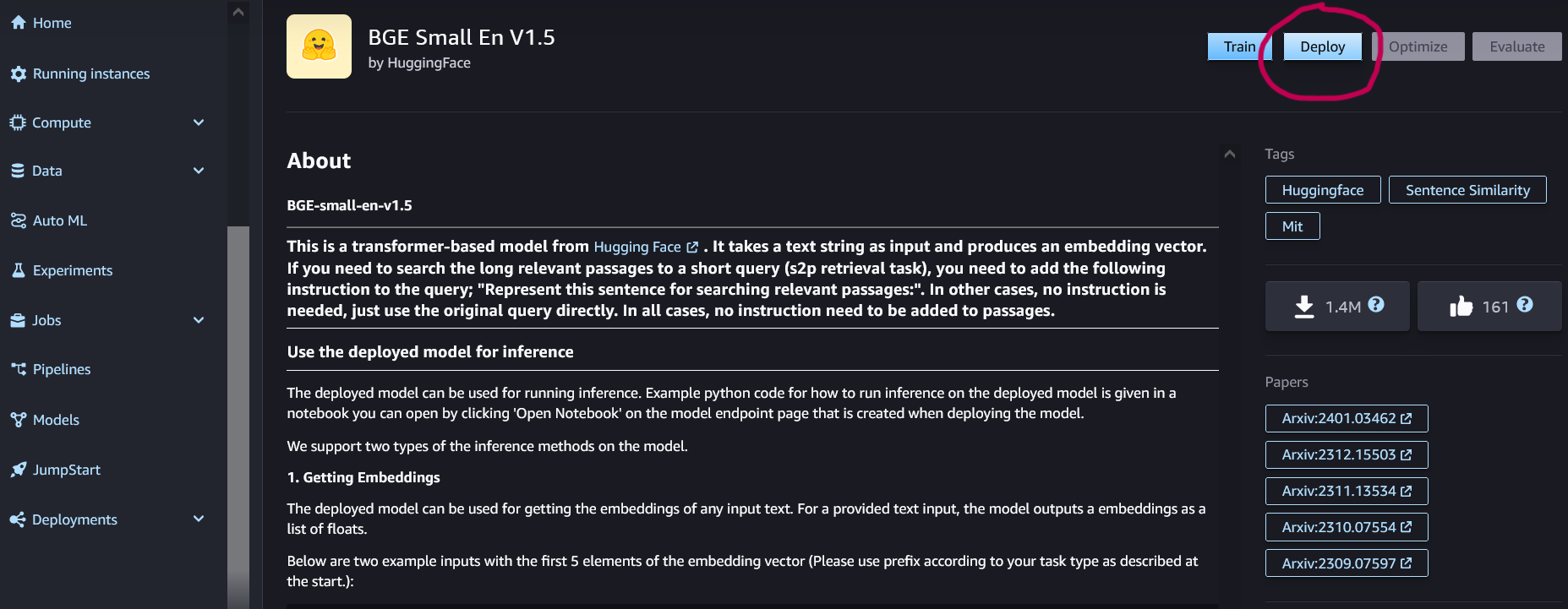
If it is not already open, go to [Amazon SageMaker AI < Studio](https://us-east-1.console.aws.amazon.com/sagemaker/home#/studio-landing)



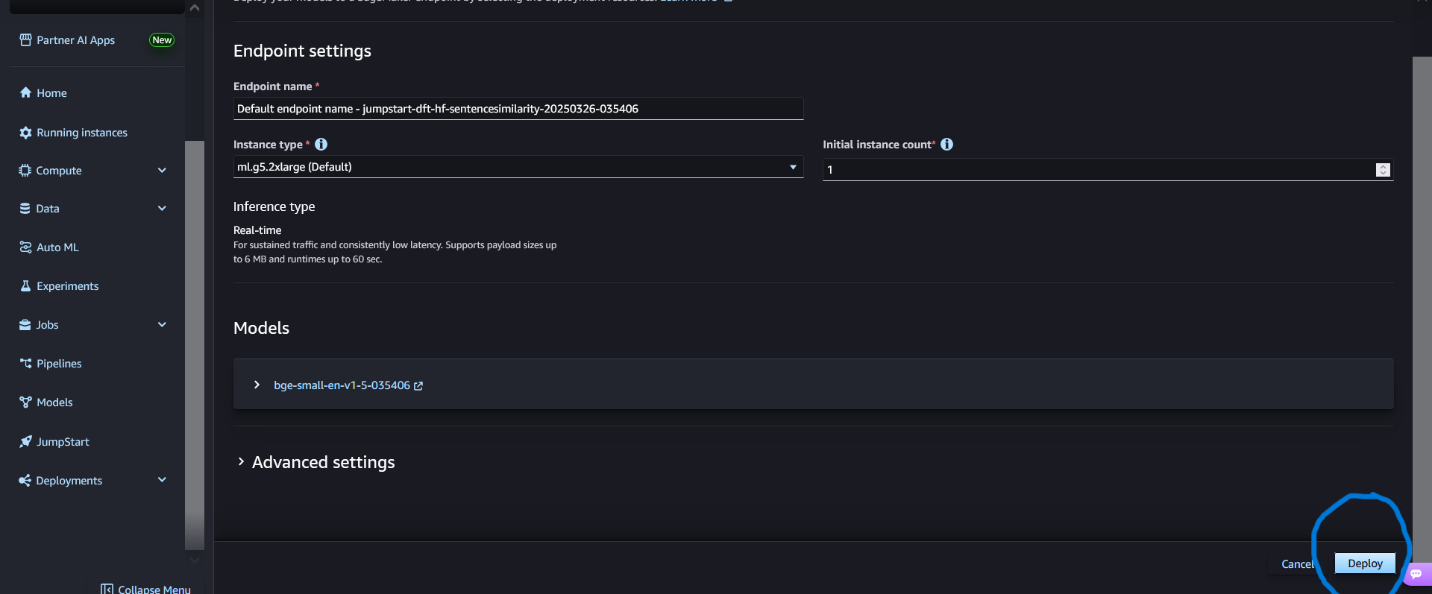
* And click the button **Open Studio**.
* Once, click on the **JupyterLab** Icon on the top left.
* Click on Create **JupyterLab Space** on Top Left Corner.
* Add Name “sagemaker-workshop-1”, Click on “Create Space”
* Next, click on the **Run** Icon to activate classic studio. This will take a few minutes.

* Inside the **Sagemaker Studio** tab, navigate to **Home** tab and click on **Jumpstart**
* Search for **BGE Small En** and click on **BGE Small En V1.5**

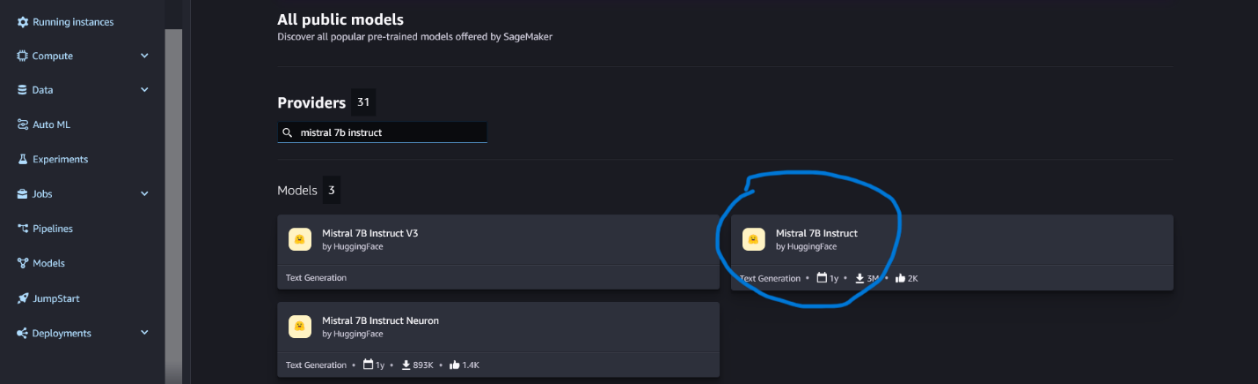
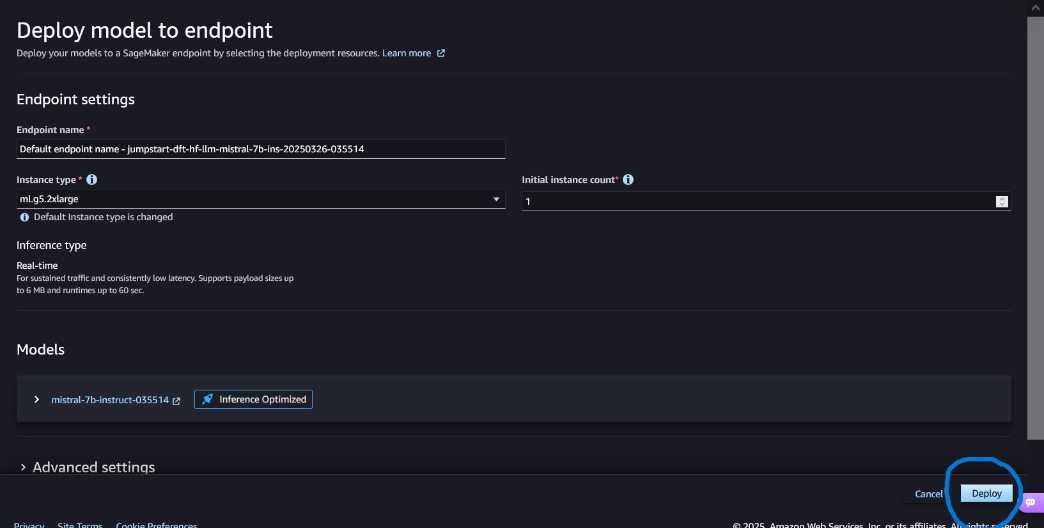


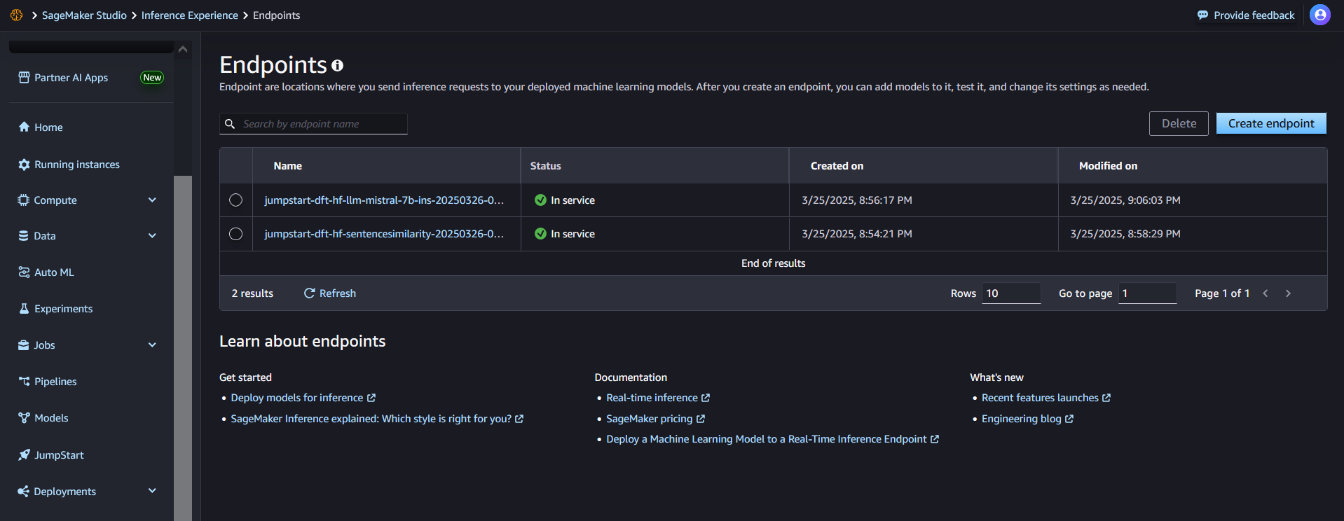


* **Under Deployment Configuration, select ml.g5.2xlarge for the SageMaker hosting instance**
* **Leave the rest as default and click Deploy**
* A deployment window will open. Deployment takes about 5-10 minutes. Once deployed, it will reflect as **"In Service"**. You can continue to the next step while waiting for JumpStart deployment to complete.



## Lab 1.2: Deploying Mistral 7B Instruct Large Language Model(LLM) via Jumpstart

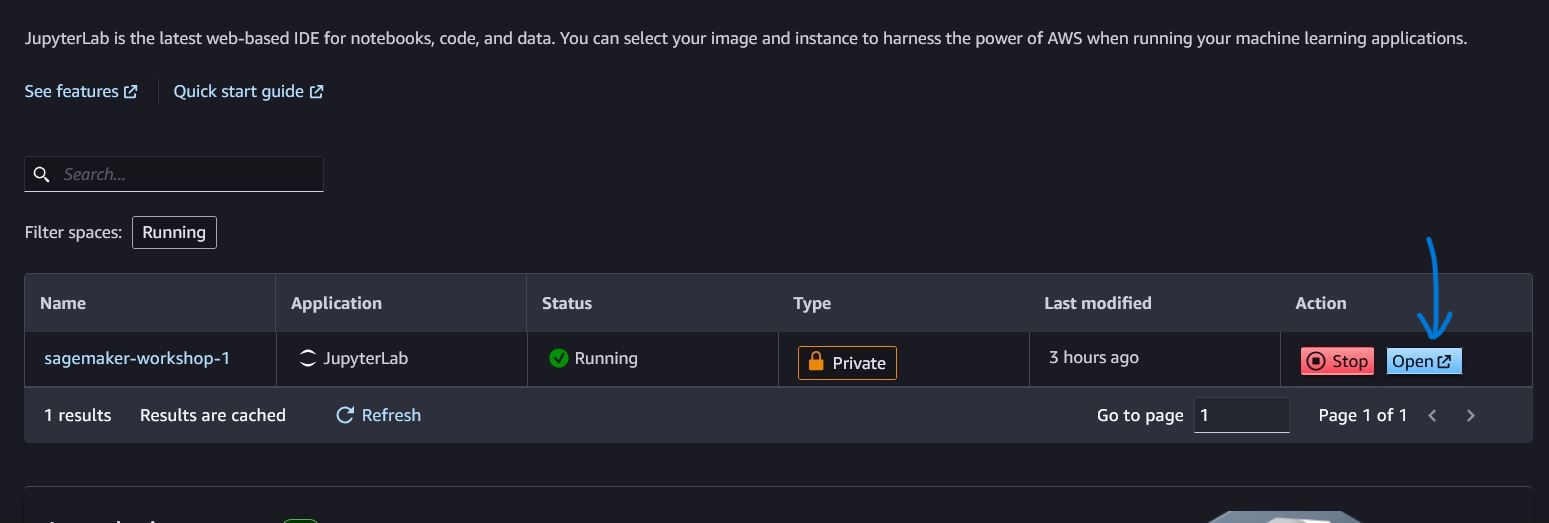
* Click on **SageMaker JumpStart** tab
* Search for **instruct** and click on **Mistral 7B Instruct**
* Under Deployment Configuration, select **ml.g5.2xlarge** for the SageMMaker hosting instance
* Leave the rest as default, and click **Deploy**
* Deployment window will open, deployment takes about 5-10 minutes.Once deployed, it'll reflect as **"In Service"**



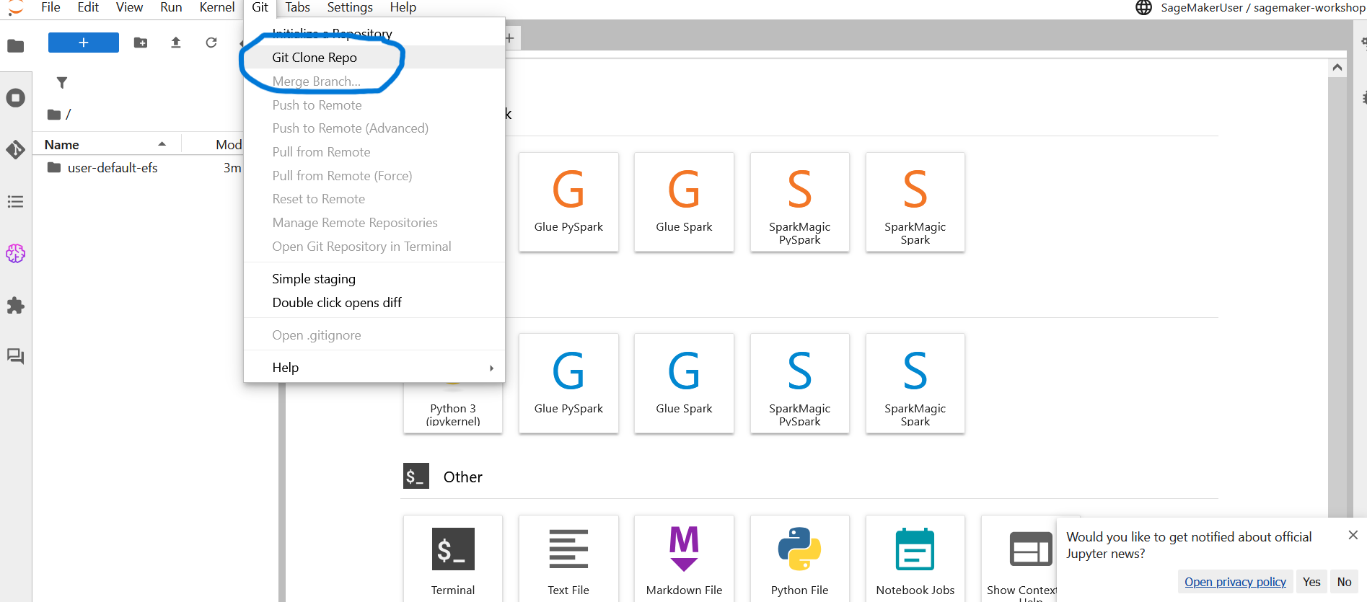
# Step 2: Using Sagemaker Studio to run the RAG enabled question answering

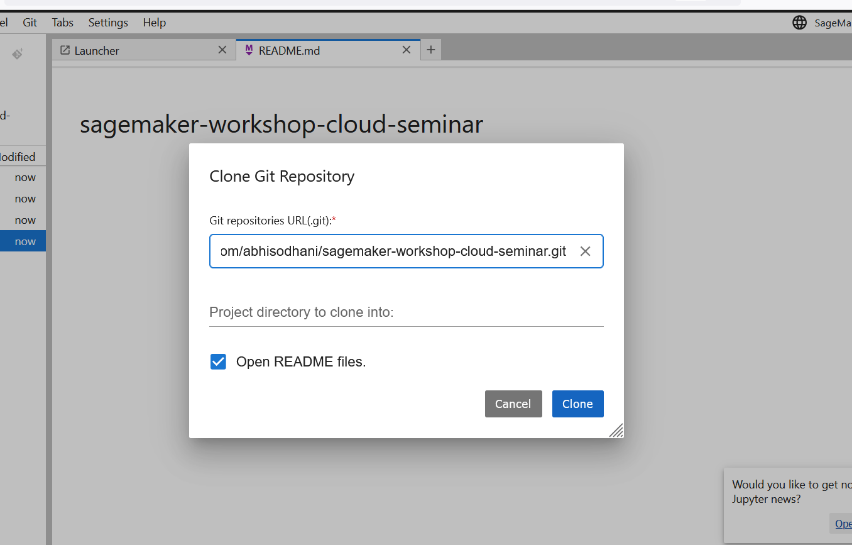
## [Lab 1.3: Setup of Sagemaker JupyterLab's notebook](https://catalog.us-east-1.prod.workshops.aws/workshops/43bac1c0-de0f-4b76-a1ac-a195de4c5647/en-US/content/lab-1/step-2" \l "lab-1.3:-setup-of-sagemaker-studio's-notebook)

* Inside SageMaker Studio, at the top left navigation bar, click on **JupyterLab**
* Next, click on the **Open** Icon to open classic studio. A new tab will open.



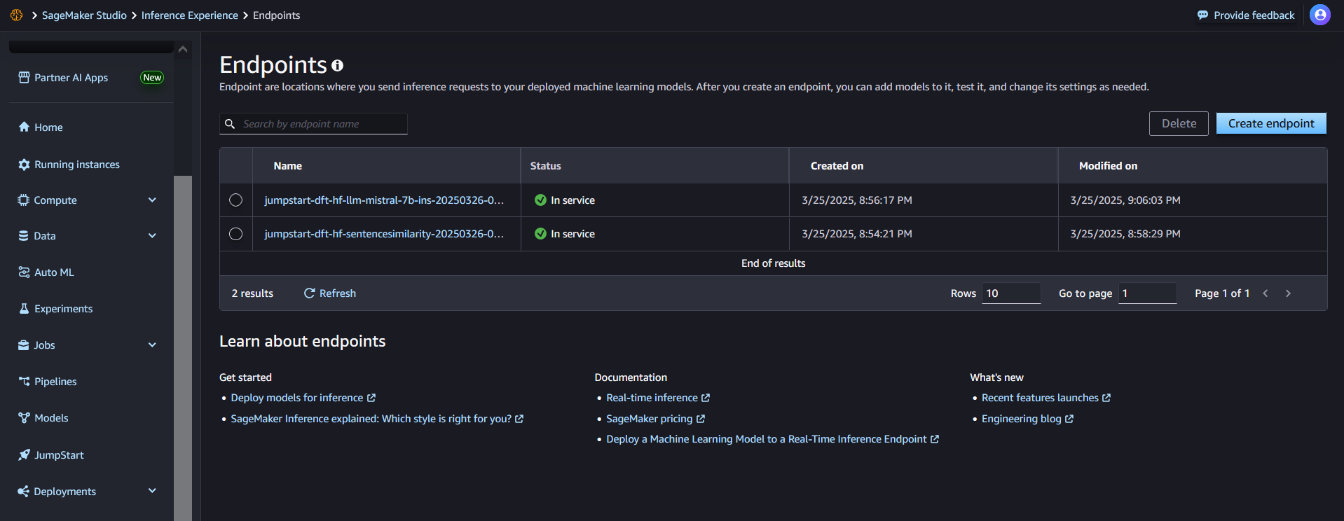
* Next, click on the **Git** Icon to Git Clone Repo - https://github.com/abhisodhani/sagemaker-workshop-cloud-seminar.git





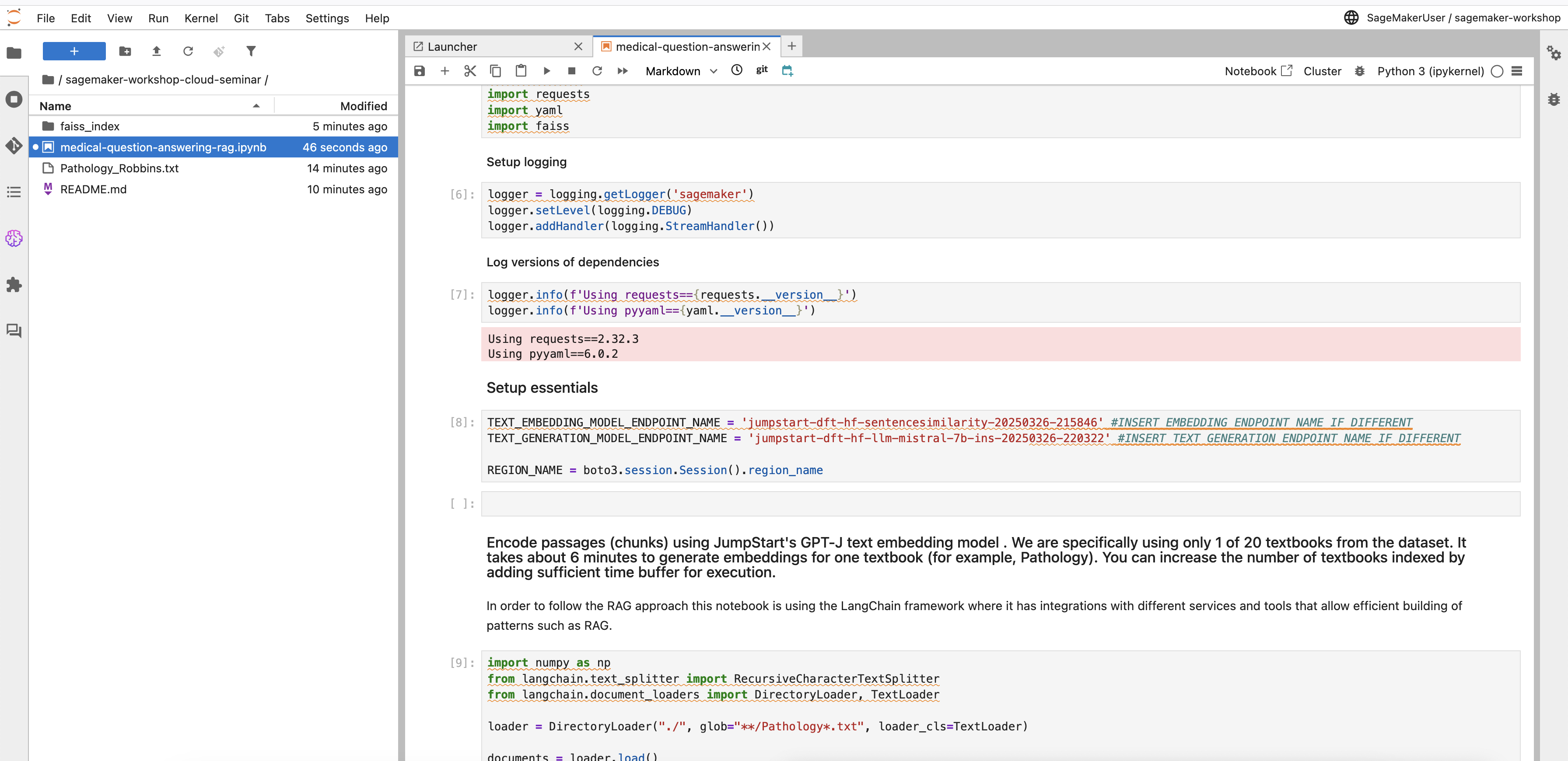
* Inside the classic SageMaker Studio tab, use the left navigation menu and go into **"lab-1"** folder
* Navigate to **mistral-rag.ipynb** below and open it
* Copy Endpoints of **BGE Model and MISTRAL Model**

\*NOTE: It might take 5-10 minutes to be in **In Services**



* Replace **TEXT\_EMBEDDING\_MODEL\_ENDPOINT\_NAME** with **BGE Model** deployed Endpoint
* And Replace **TEXT\_GENERATION\_MODEL\_ENDPOINT\_NAME** with **MISTRAL Model** deployed Endpoint

(NOTE: You need to replace model endpoints 1 times in the document)



## Follow the instructor as he/she goes through the notebook in the following 2 steps.

### [Step 1 : Index documents in the vector store](https://catalog.us-east-1.prod.workshops.aws/workshops/43bac1c0-de0f-4b76-a1ac-a195de4c5647/en-US/content/lab-1/step-2" \l "step-1-:-index-documents-in-the-vector-store)

Before being able to answer the questions, the documents must be processed and a stored in a document store index. Load the documents. Process and split them into smaller chunks to create a numerical vector representation of each chunk using BGE model. Create an index using the chunks and the corresponding embeddings.

### [Step 2 : Conversation with users with a large language model and an existing knowledge base](https://catalog.us-east-1.prod.workshops.aws/workshops/43bac1c0-de0f-4b76-a1ac-a195de4c5647/en-US/content/lab-1/step-2" \l "step-2-:-conversation-with-users-with-a-large-language-model-and-an-existing-knowledge-base)

When the documents index has been prepared, you are ready to ask the questions and relevant documents will be fetched based on the question being asked. The following steps will be executed:

1. Create an embedding of the input question.
2. Compare the question embedding with the FAISS vector store embeddings for the (top N) relevant document chunks.
3. Add those chunks as part of the context in the prompt.
4. Send the prompt to the model at the Amazon SageMaker endpoint using the contextual answer based on the documents retrieved

You can now run through the notebook on your own. With the RAG approach, try asking different questions.