Last updated: September 10, 2024

AWS Sagemaker Studio

Navigate to AWS access portal.

Find **portx-genai** and click on PTX-SuperAdmins.

Find **Amazon Sagemaker** and select it. You can find it by selecting “All services” in the side panel and then looking under “Machine Learning.” You can also simply search for it in the search bar.

On the side panel, find “Admin configurations” and click on “Domains.”

Click on the Domain called “QuickSetupDomain-20240605T144834”. Note: if you do not see this, make sure to change the location to “Oregon.” You can find this located in the top right by the gear symbol.

Navigate to “User profiles” in the horizontal menu.

Find the one named “default-20240605t144834” and click on “Launch” and then click on “Studio.” This will launch Sagemaker Studio.

Click on the JuptyerLab icon on the left side of the screen.

If you are just running an existing space, go to Run Space, else go to Setup Space.

Run Space

From here you can run one of the fine-tuning spaces. They are both the same besides the datasets that were used to fine-tune the models. Simply click “Run” and it will start the space.

Setup Space

Click on the “Create JupyterLab space” button on the right side.

Give it whatever name you would like and make it public.

Change the instance type to “ml.g4dn.xlarge” and storage to 100GB. Then, start the space.

Fine-tuning

Upload the app.py, local.py, and commands.txt files from this Github repository.

Run the first command found in the commands.txt file by opening a terminal window in the Sagemaker Studio.

Then, find your Hugging Face token by navigating to your Hugging Face profile settings and clicking on “Access Tokens.”

Create a new token by clicking on the button in the top right. Make sure you save this token in a secure location such as 1Password or another password keeper since you will not be able to view it again. You can always create a new token, though.

Now, go back to your Jupyter notebook and run the second command in the commands.txt file with your Hugging Face token. For example, export HF_TOKEN=XXXXXXXXXXXXXXXX.

Then, navigate to the app.py file and ensure everything is how you want it to be. Change all usernames to your username and the dataset to your desired dataset. These are marked with TODO comments.

Then, run the app.py script.

While this runs, open the local.py file and edit where you find TODO comments.

Once app.py is done running, run local.py file.

Then, run the command to create a Modelfile file.

Copy and paste the contents of the Modelfile from the Github repository into this file making sure to change the name of model in the first line.

Then, run the command to create the model with Ollama.

Then, you can run this model if you would like to.
Run the command to get the Ollama access key.

Copy this key and go to your profile settings in Ollama.

Navigate to “Ollama keys” and add the key here.

Finally, run the last command to push the fine-tuned model to Ollama.