Building with Foundation Models on Amazon SageMaker Studio

Building with Foundation Models on Amazon SageMaker Studio
Introduction to Workshop Studio Setup
SageMaker Spaces: JupyterLab and Code Editor
Lab 0 - Deploy Llama2 and Embedding Models
Lab 1 - Setup an LLM Playground on Studio
Lab 2 - Prompt Engineering with LLMs
Lab 3 - Retrieval Augmented Generation (RAG) using PySpark on EMR
Lab 4 - Fine-Tune Gen AI Models on Studio
Studio Notebook HuggingFace Fine-Tuning
Studio Notebook HuggingFace Fine-Tuning
Trainium JumpStart Fine-Tuning
Lab 5 - Foundation Model Evaluation

AWS account access
Open AWS console (us-east-1)
Get AWS CLI credentials
Get EC2 SSH key

Exit event

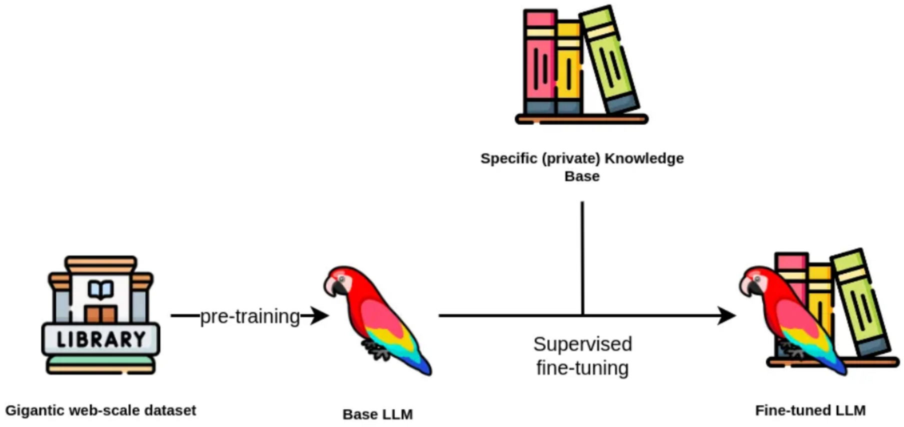
Event dashboard > Lab 4 - Fine-Tune Gen AI Models on Studio

Lab 4 - Fine-Tune Gen AI Models on Studio

Contents

Overview

LLM fine-tuning involves training a pre-existing large language model, like `Llama2`, `Mistral`, `Falcon`, on a specific dataset to improve its performance in a particular domain or task. This process adjusts the model's parameters to better understand and generate text relevant to specialized fields, such as legal, medical, or technical content. Fine-tuning makes the model more accurate and effective in handling the nuances of the targeted area. It is a crucial step in customizing general-purpose language models for specific applications or industries.



(Image credits: Neo4j)

Fine-tuning is one step below a Pre-Training a LLM and a step above Retrieval Augmented Generation with LLMs. In this lab you're going to learn how to,

- Fine-tune a LLM like `Llama2`/or `Llama2 variant` using custom dataset
- Fine-tune LLMs on AWS Silicon or NVIDIA GPUs- Trainium instances (`trn1` \ `trn1n`) or SageMaker `g4dn` Instances
- Deploy a Fine-tuned model to SageMaker Endpoints for Large Language Model Serving (inference)

Fine-Tuning Labs

There are 2 labs that demonstrate LLM fine-tuning,

1. **Studio Notebook HuggingFace Fine-Tuning**: This lab demonstrates how to fine-tune a Llama2 variant on Studio's Code Editor backed by a `ml.g4dn.xlarge` instance. Here you learn how to download a `Llama2` variant from HuggingFace hub, quantize the model into 4bit using `bitsandbytes`, fine-tune a model using custom dataset and deploy the model as a SageMaker Endpoint for model inference.
2. **Inferentia2 JumpStart Fine-Tuning**: This lab demonstrates how to fine-tune Llama2 JumpStart model using your custom dataset on `trn1`/`trn1n` instances as a SageMaker training job. This lab also demonstrates how to deploy a trained JumpStart model as a SageMaker Endpoint on `inf2` instances.

> ⚠ **Important**
> Due to Workshop limitations, please run **Studio Notebook HuggingFace Fine-Tuning** only!

- Studio Notebook HuggingFace Fine-Tuning
- Trainium JumpStart Fine-Tuning

Previous Next