CLOUD COMPUTING ASSIGNMENT

Shreya Singh
21052875
CSE_28

Fine-Tune & Evaluate LLMs with Amazon SageMaker

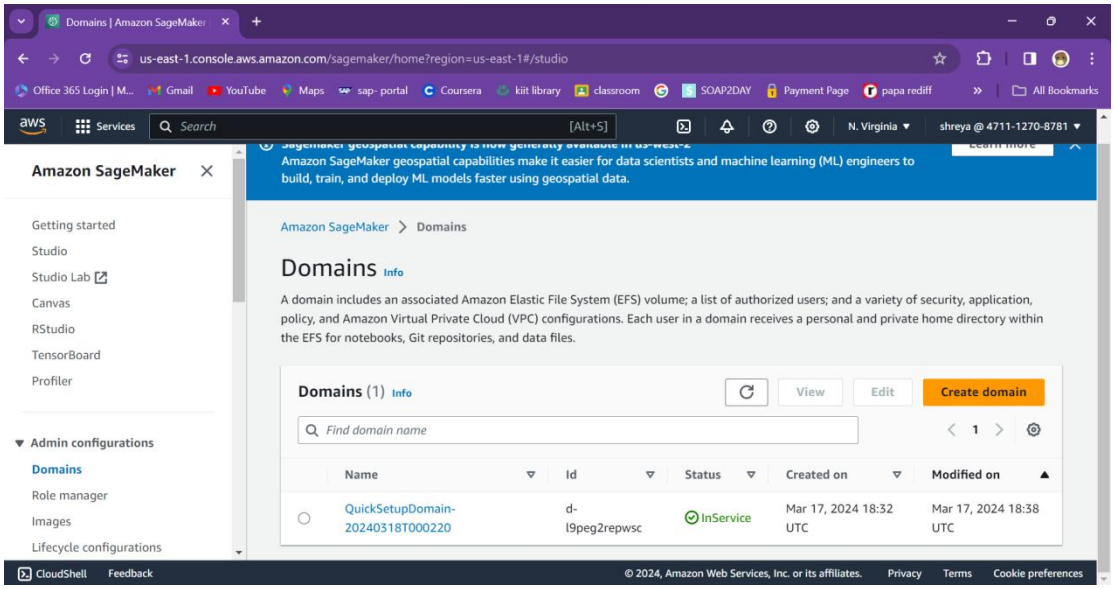
Introduction

Large Language Models (LLMs) have revolutionized the field of natural language processing (NLP) by demonstrating remarkable capabilities in tasks like text generation, translation, and question answering. However, their generic nature often requires fine-tuning for optimal performance on specific domains or tasks. This report explores the potential of Amazon SageMaker, a cloud-based machine learning platform, for fine-tuning and evaluating LLMs.

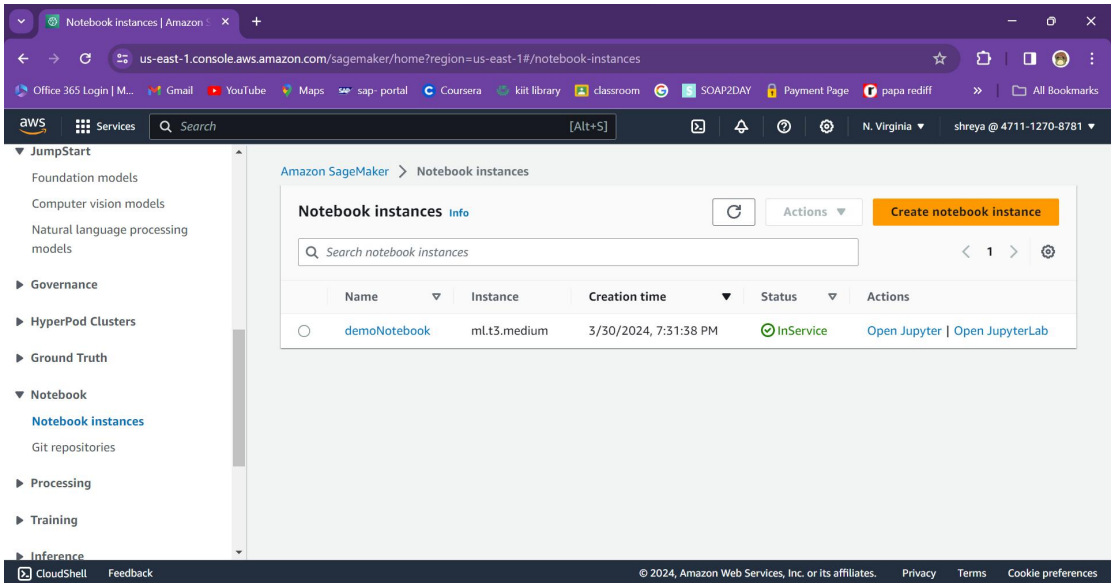
We will delve into the benefits of using SageMaker for LLM fine-tuning, including its integration with popular deep learning frameworks, ease of deployment, and scalability. We'll then showcase a practical example of fine-tuning an LLM on SageMaker for a specific task, demonstrating the process step-by-step. Finally, we'll discuss evaluation techniques for fine-tuned LLMs and how SageMaker can facilitate this process.

Working

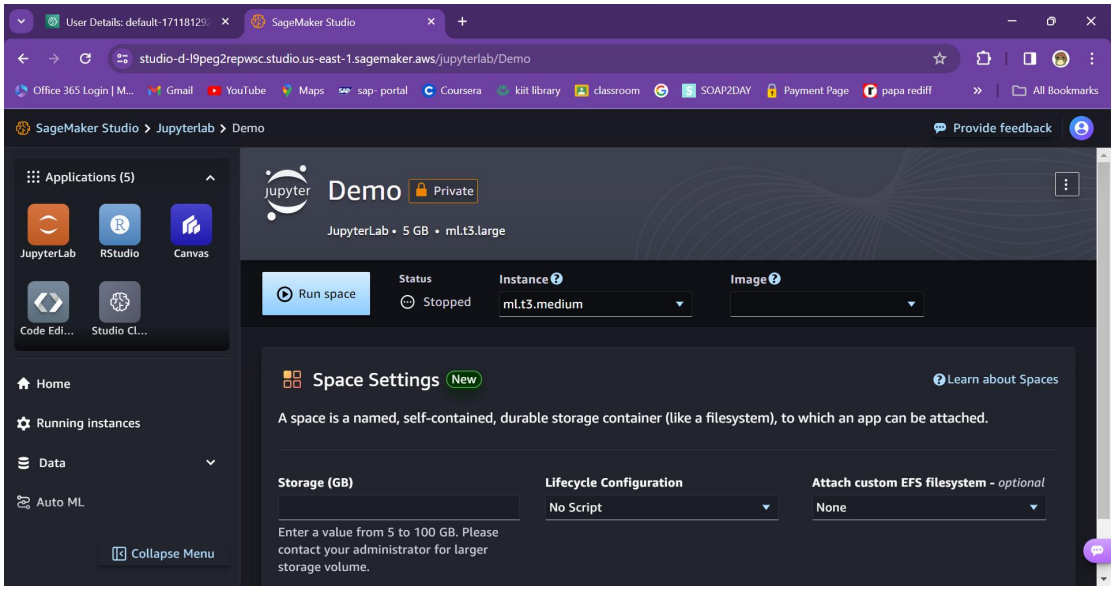
1. Creating SageMaker Domain:



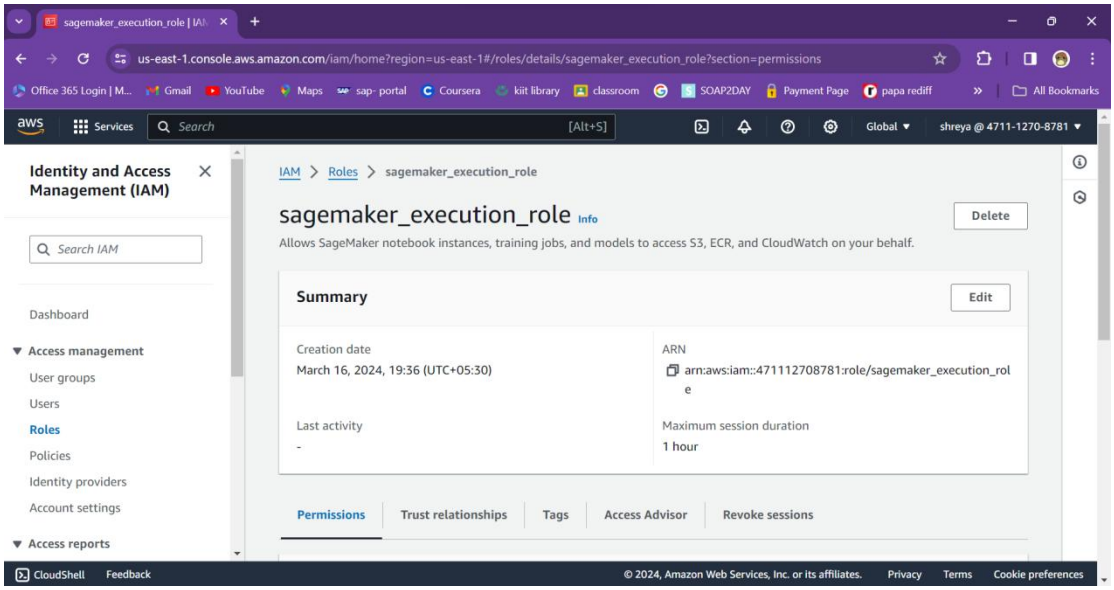
2. Creating Notebook Instance:



3. We deploy our model through the given code via SageMaker Studio's Jupyter Lab



4. We create a another User with SageMaker access to test/predict from the deployed model.



Conclusion

This report has highlighted the capabilities of Amazon SageMaker as a powerful platform for fine-tuning and evaluating LLMs. By leveraging SageMaker's infrastructure and tools, developers can streamline the LLM fine-tuning process, ensuring efficient development cycles and optimal model performance. The provided example demonstrated the practical application of SageMaker for fine-tuning, and the discussion on evaluation techniques emphasized the importance of rigorous model assessment. Moving forward, SageMaker can be expected to continue evolving with new functionalities and integration, further simplifying and accelerating LLM development for various NLP applications.