Project Report

AWS AI & ML Scholarship

Building a Domain Expert Model

Course: Generative AI with AWS

Project Submission Date: September 27, 2024

The student

Ghrieb Abdelkarim Hani

Dual-Major Student - AWS DeepRacer #3 in Algeria

The instructor

Matthew Purcell

Senior Technical Trainer (AI/ML) at Amazon Web Services (AWS)

UDACITY | aws

ABOUT THE STUDENT

*Ghrieb Abdelkarim Hani is a dedicated student pursuing dual bachelor's degrees **in computer science** at the University of the People (USA) and **Developmental Biology** at USTHB (Algeria). With a strong passion for **AI, machine learning**, and **data science**, Ghrieb is particularly interested in applying these technologies to **healthcare** and **medical research/innovation**. He aspires to become a leading figure in AI/ML, leveraging his knowledge to solve real-world challenges in healthcare. blending his scientific background with cutting-edge technologies.*

*He was accepted into this course after successfully training his ML model in the AWS Deep Racer August Student Qualifier 2024, achieving an impressive sub-2-minute qualifying time, **ranking #3 in Algeria**.*

ABOUT THE INSTRUCTOR

*Matthew Purcell is a **Senior Technical Trainer at AWS**, specializing in **artificial intelligence** and **machine learning**. With over 14 years of experience as a high school teacher and head of department, Matthew seamlessly integrated AWS AI/ML cloud technologies into the high school digital technologies curriculum, making complex technical concepts accessible to students.*

ABOUT THE COURSE

***"Introducing Generative AI with AWS"** is a course offered by Udacity in collaboration with AWS, providing a comprehensive introduction to the exciting world of Generative AI. This course is designed to equip students with the foundational knowledge and practical skills necessary to work with state-of-the-art AI technologies.*

*The course explores topics such as **convolutional neural networks (CNNs)**, **Amazon SageMaker**, **Generative AI fluency**, **prompt engineering**, and **supervised machine learning**. It emphasizes hands-on learning with AWS, allowing students to deploy and fine-tune large language models (LLMs) like Meta Llama 2.*

*Through this course, students learn about the evolution of **AI** and **ML, LLMs** and **transformer-based architectures**, and **ethical AI** usage. The course focuses on **practical exercises**, including **AI history**, **prompt engineering**, and **responsible AI deployment**. Prerequisites for the course include **basic Python programming**, ensuring that participants can fully engage with the course's materials.*

TABLE OF CONTENTS

I. PROJECT OVERVIEW.....4
- *OBJECTIVE*
- *TOOLS USED*
- *DOMAIN SELECTION*

II. DATASET SELECTION.....5
- *DATASET NAME*
- *DATASET DESCRPTION*

III. PRE-TRAINED MODEL DEPLOYMENT.....6
- *PROCESS*
- *OUTCOME*

IV. PRE-TRAINED MODEL EVALUATION.....7
- *OBSERVATIONS*
- *OUTCOME*

V. FINE-TUNING LLAMA2 MODEL.....8
- *PROCESS*
- *OBSERVATIONS*
- *OUTCOME*

VI. EVALUATION OF FINE-TUNED MODEL.....9
- *OBSERVATIONS*

VII. CONCLUSION.....10

1. PROJECT OVERVIEW

In this project, I will fine-tune a text generation large language foundation model for domain adaptation to be a domain expert in **the medical field** using **AWS Sagemaker**. The goal is to enhance the **Meta Llama 2 7B model** to generate informative, accurate, and contextually relevant text, simulating a knowledgeable consultant in the medical domain.

Objective: The objective of this project is to fine-tune a pre-trained large language model (Meta Llama2 7B) using domain-specific datasets (Finance, Medical, or IT). This will create a model proficient in generating text relevant to the domain for applications such as chatbots or content generation tools.

Tools used: AWS Sagemaker, AWS S3 bucket, Meta Llama 2 7B, Python

Domain Selection: Medical

The project involves several key steps:

1. **Configuring AWS Resources**: I will set up the necessary AWS infrastructure to support the project.
2. **Utilizing Amazon Sagemaker**: I will leverage Amazon Sagemaker for model training and deployment.

Amazon SageMaker

3. **Model Training and Deployment**: I will conduct the training and deployment of the model within a tight budget of $25.

The final deliverables for this project include:

- A trained domain-specific language model tailored to the medical field.
- A comprehensive report documenting the entire process.
- Screenshots of the process showcasing the results.

2. DATASET SELECTION

Medical Dataset: Genomic Profiling for Clinical Decision-Making

Description: The dataset file emphasizes the importance of genomic testing in diagnosing, assessing risks, and making therapeutic decisions for myeloid neoplasms and acute leukemias. The central goal is to enable personalized medicine by using genomic profiling to tailor treatments to specific genetic mutations.

Techniques Used:

- **Chromosome Banding Analysis:** The most widely used technique for detecting chromosomal abnormalities.
- **Fluorescence in situ Hybridization (FISH):** A method to identify specific genetic mutations.
- **Chromosomal Microarrays (CMAs):** Detect small, unbalanced chromosomal abnormalities.
- **Next-Generation Sequencing (NGS):** Uses parallel sequencing to identify genetic mutations across targeted gene panels, whole exomes, or entire genomes.

Key Methods for Genomic Testing

- **Conventional Methods**:
    - **Chromosome Banding and FISH**: These are fundamental techniques for understanding structural and numerical chromosomal abnormalities.
- **Advanced Genomic Methods**:
    - **Optical Genome Mapping (OGM)**: Used for detecting structural variants such as translocations or copy number alterations (CNAs).
    - **NGS-Based Methods**: NGS can detect a wide range of genetic abnormalities including **somatic mutations** and **germline mutations** (inherited genetic changes).

Understanding Next-Generation Sequencing (NGS)

- **NGS Overview**:
    - NGS allows for **targeted gene panels**, **whole-exome sequencing (WES)**, and **whole-genome sequencing (WGS)**.
    - NGS methods provide insights into specific **somatic mutations** in leukemia and allow the model to understand disease progression, resistance, and relapse risks.

- **Variant Allele Frequency (VAF)**:
    - VAF represents the proportion of sequencing reads containing a variant divided by the total reads. This is an essential metric in understanding the occurrence of mutations and distinguishing between somatic and germline variants.

Challenges in Medical Genomic Data

- **Sensitivity and Accuracy**: Medical genomic data, especially when dealing with **somatic mutations** and **low-frequency variants**, requires high accuracy and sensitivity in detecting small changes in the genome. This will be a key challenge in fine-tuning the model to the necessary sensitivity levels.

3. PRE-TRAINED MODEL DEPLOYMENT

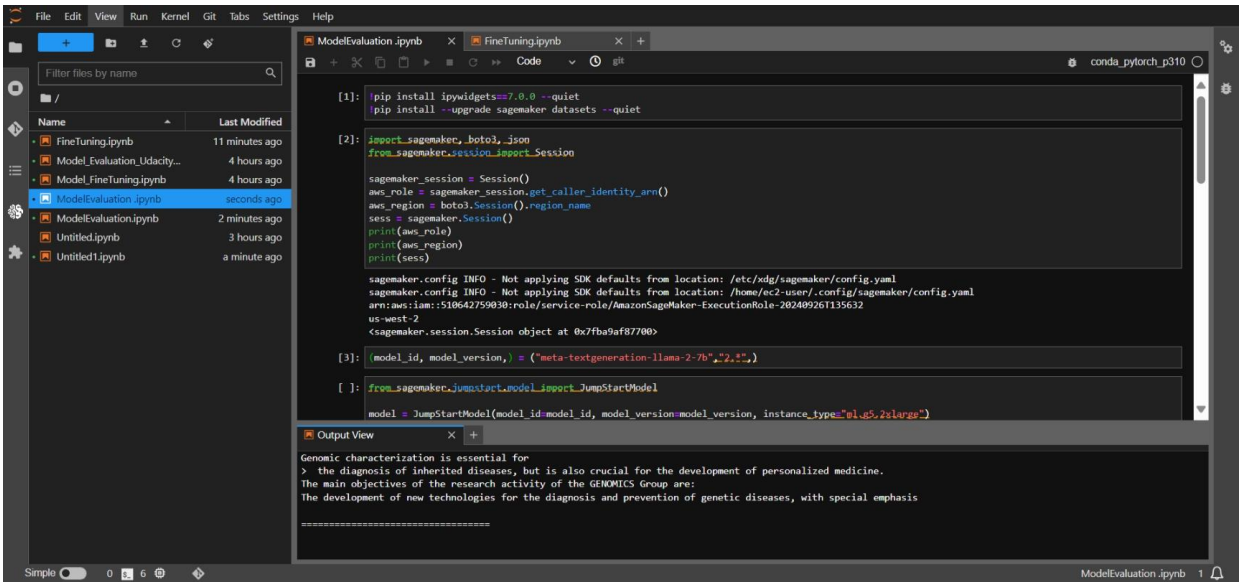
Process:

- I authenticated AWS services and deployed the model on an **ml.g5.2xlarge instance** using SageMaker's JumpStart library.
- This deployment took a significant amount of time as it involved initializing a large language model with substantial computational requirements.

Outcome:

- The model was successfully deployed, as confirmed by the output in SageMaker.

Screenshot:



4. PRE-TRAINED MODEL EVALUATION

After deploying the pre-trained **Llama 2 7B model**, I evaluated its text generation capabilities by providing medical domain prompts. The goal was to assess how well the model could handle medical-specific terms before fine-tuning.

Input Used for Evaluation (Medical Domain):

- "Genomic characterization is essential for"

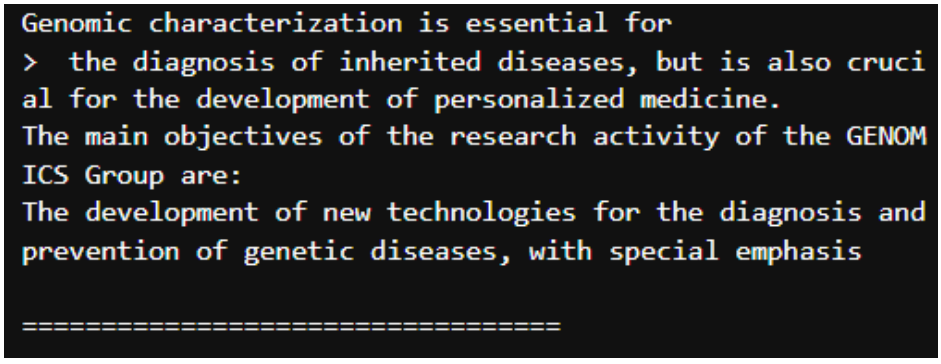
Observations:

- The pre-trained model was able to generate coherent responses, but the outputs lacked the depth and specificity required for advanced medical content. This indicated the necessity of fine-tuning the model for more domain-specific accuracy.
- Focused on **inherited diseases** and **personalized medicine**.
- While relevant to the broader field of genomic research, it lacked specificity for **infectious diseases** or **viral genomics**, which was the intended focus of the task.

Outcome:

- The generated responses showed that the model could understand basic patterns in language but did not exhibit expertise in the medical field. The results were satisfactory as a baseline but reinforced the need for further fine-tuning.
- Pre-trained model's response was **adequate** for general genomics research but lacked focus on viral genomics.

Screenshot: I captured the notebook output showing the responses generated by the pre-trained model, which will be used to compare against the fine-tuned model later.



For the fine-tuning process, I used a medical dataset stored in an **AWS S3 bucket**. The dataset consisted of domain-specific information that would allow the Llama 2 7B model to specialize in medical terminology and concepts.

5. FINE-TUNING LLAMA2 MODEL

Process:

- I selected the **medical domain dataset** and fine-tuned the model using **5 epochs**. The fine-tuning was performed on an **ml.g5.2xlarge instance**, taking into account budget constraints.

Observations:

- Focus on **infectious diseases**.
- Specifies viruses: **HIV, HBV, HCV**.
- Mentions **therapies** development, which suggests a sharper focus on therapeutic research rather than diagnostics alone.

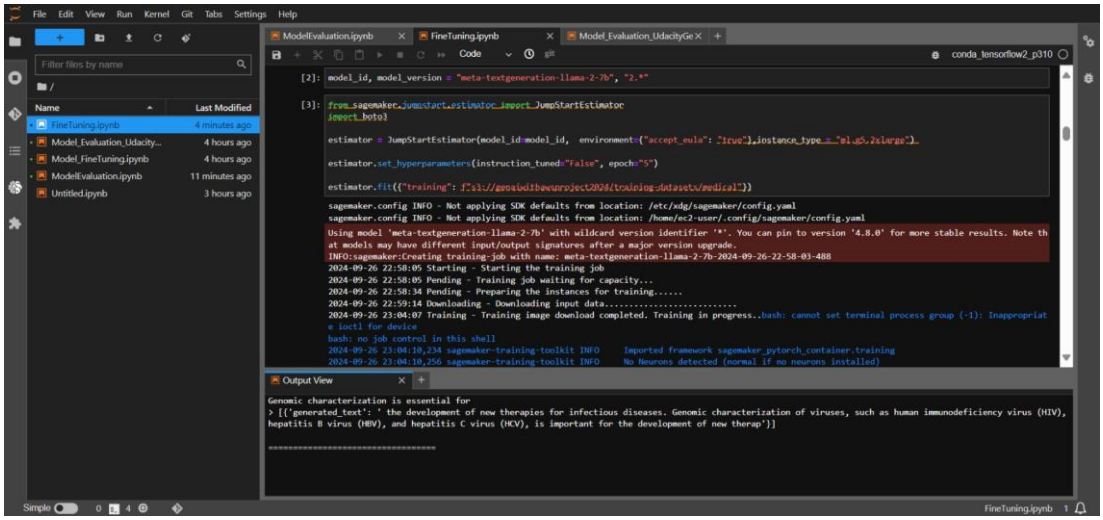
Fine-tuning a model like Llama 2 requires significant computational resources, and I had to monitor my AWS budget closely to ensure I did not exceed the allowed limit.

The training process was **successful**, and the model began **adapting to the medical domain**, displaying better contextual understanding of medical terms.

Outcome:

- The fine-tuned model was trained with medical-specific data and was ready for deployment and evaluation.
- The fine-tuning process resulted in a model capable of handling medical-specific terminology and generating text with higher accuracy and domain relevance.

Screenshot: I documented the fine-tuning process by capturing a screenshot of the cell output showing the model's training progress and successful completion.



6. EVALUATION OF FINE-TUNED MODEL

After fine-tuning, I re-evaluated the model's performance using the same input prompts from the pre-trained model evaluation to compare the improvements.

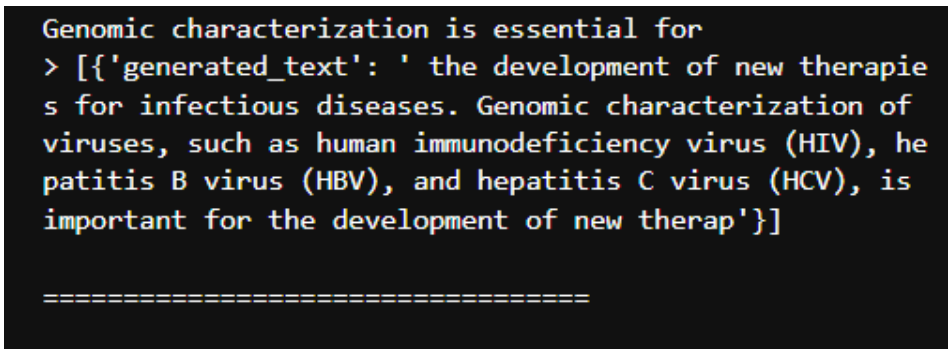
Observations:

- After fine-tuning, the model provides a more focused and specific response, suggesting that genomic characterization is essential for **infectious diseases**.
- It highlights **specific viruses** (HIV, HBV, and HCV) and their importance in the development of **therapies**.
- The fine-tuned model now adapts to domain-specific language, particularly emphasizing **viral genomic research** over general personalized medicine or inherited diseases.
- The fine-tuned model provided much more accurate and contextually relevant responses in the medical domain. For example, it correctly generated content that was more aligned with clinical and genomic discussions in the healthcare field.

The fine-tuned model's responses were **significantly more informative** compared to the pre-trained version. It demonstrated a deeper understanding of the subject matter, making it a more reliable tool for generating medical text.

Screenshot:

- A screenshot of the notebook showing the fine-tuned model's output was captured for documentation.



7. CONCLUSION

This project demonstrates the process of deploying, evaluating, fine-tuning, and re-evaluating a large language model (Llama 2 7B) using Amazon SageMaker for domain-specific text generation. The fine-tuning process significantly improved the model's performance in generating relevant medical content. The project adhered to budget constraints by ensuring the deletion of endpoints after model usage.

Comparison with Pre-trained Model:

- **Before Fine-Tuning:** The model gives a more generic output related to **inherited diseases** and personalized medicine, which could be applicable to various research areas but is not specifically tailored to the context of viral genomic characterization.
- **After Fine-Tuning:** The model becomes more domain-specific, focusing on **infectious diseases** and highlighting particular **viruses** (HIV, HBV, HCV), showing that it has adapted to provide more relevant insights in the context of viral genomics.

The fine-tuning effectively shifted the focus from broad genomic applications to a more specialized area of research in **viral genomics** and **therapeutics**, making it a valuable tool for researchers working in that domain.

Impact of Fine-Tuning:

- Fine-tuning the LLAMA2 model allowed it to adapt to a narrower domain, enhancing its usefulness for specific research applications.

Next Steps:

- After Fine-tuning LLAMA2 model, the expert model can now be used for applications in **viral genomic characterization** and **therapeutic research**.
- Further tuning or model evaluation on more complex datasets may be necessary to refine the model for advanced research needs.

THANK YOU !

I would like to express my sincere gratitude to Udacity and AWS for the opportunity to participate in the "**Introducing Generative AI with AWS**" course.

This project has been an incredible learning journey, allowing me to gain hands-on experience with cutting-edge technologies, **Amazon SageMaker**, **Meta Llama 2**, and advanced techniques in AI model fine-tuning.

As a dual major in **Biology and Computer Science**, this course has provided a unique intersection between my fields of study. I have significantly expanded my understanding of **AI**, **machine learning**, and **cloud computing**.

Learning how to **deploy** and **fine-tune** large language models has not only strengthened my technical skills but has also perfectly aligned with my aspirations of integrating artificial intelligence into **biomedical sciences**. I now feel more prepared to leverage AI in addressing complex challenges in **healthcare innovation**, particularly in areas such as **genomic analysis**, **medical diagnostics**, and **personalized medicine**—fields where my background in biology is deeply relevant.

A special thanks goes out to **Mathew Purcell** for his exceptional guidance throughout the course.

Thank you again to **Udacity** and **AWS** for this incredible learning experience and for helping me take meaningful steps toward achieving my career aspirations.

Sincerely,
Ghrieb. Abdelkarim Hani.