Assignment 2 AMS 560
Fall 2024

Name: _____ SBU ID: _____

The Assignment is due on October 11, 2024 11:59 PM. Please read the instructions carefully first and then answer the following questions.

https://github.com/yl1127/llama-models

Submit your assignment to Brightspace.
Please include Answers to all the questions asked.

In this assignment, you will explore and implement key tasks related to running large-scale machine learning models on CloudLab infrastructure using a GPU. The focus is on setting up the environment, downloading the Llama3.1-8B-Instruct model, and running an example script that leverages the model's capabilities. Additionally, you will explore the model's internal structure by analyzing tokenization outputs, attention blocks and transformer layers. This assignment helps develop practical experience with large language models (LLMs) while ensuring proper configuration of cloud resources and GPU acceleration. An additional reading is provided for context but is not required.

Section 1: Download and run (75 points)

1. (15 points) Do you successfully set up a Cloudlab with GPU? (Yes or No and a screenshot)
2. (15 points) Do you successfully print the model list? (Yes or No and a screenshot)
3. (15 points) Do you successfully download the **Llama3.1-8B-Instruct** model? Where you download it? (Yes or No and a screenshot)
4. (15 points) Do you successfully install GPU driver? Show your GPU usage. (Yes or No and a screenshot)
5. (15 points) Do you successfully run the **example_chat_completion.py**? (Yes or No and three screenshots)

Section 2: Explore (25 points)

6. (5 points) How many files in **/Llama3.1-8B-Instruct**? What are they?
7. (5 points) How many heads of each multi-head attention block?
8. (5 points) What's the tokenizer encode output of "hello world!"?
9. (5 points) How many transformer layers of this model?

10. (5 points) What's the shape of Attention weight W_o of layer 21?

Section 3: Additional reading (Not required)

The Llama 3 Herd of Models

https://ai.meta.com/research/publications/the-llama-3-herd-of-models/