Weekly Progress Report

Group Members:

- Geron Simon Javier
- Mhar Andrei Macapallag
- Seanrei Ethan Valdeabella

Work Completed:

This week, we created a synthetic dataset for general computer science questions of students with the help of an LLM called DataLLM. As of writing, we generated 40,000 rows of data that has a feature of input and a label called output. We also created a demo chatbot that is fine-tuned by our generated dataset.

Challenges Encountered:

We encountered a problem when we are trying to generate a demo chatbot. It turns out that Google Colab has a GPU usage limit. Unfortunately, our dataset is too large which makes the training of our chatbot resource-heavy, leading to restrictions on further model training in this environment.

Solutions Implemented:

The solution we came up with is using a technique called Mixed Precision Training. Mixed precision training offers significant computational speedup by performing operations in half-precision format, while storing minimal information in single-precision to retain as much information as possible in critical parts of the network. As a result, our training process is now significantly faster and more resource-efficient, allowing us to continue refining the chatbot within our resource constraints.

Tasks for Next Week:

Next week, we will enter the pretraining phase, focusing on optimizing hyperparameter testing. We will set up controlled experiments to identify the best hyperparameters that enhance the performance and accuracy of the chatbot. We will test various learning rates, batch sizes, and layer configurations to determine which factors most significantly affect the model's efficiency and output quality. By the end of the week, we aim to fine-tune a few sets of refined hyperparameters and continue training, laying a solid foundation for future development and deployment.

Instructor's Feedback:

____

____

____

Instructor's Signature: ____________________
Date: November 1, 2024