# CSE665: Large Language Models

## Assignment 3
## Fine Tuning Large Language Models

**Maximum Marks: 25**

- Deadline: Strictly enforced. No late submissions will be accepted.
- GitHub Repository: It is mandatory to maintain a GitHub repository for all assignments, as subsequent tasks will require updates to the same files and functions.
- Evaluation: Marks will only be awarded if you can successfully explain your approach to the task during the evaluation by the TA.
- Submission Format: Submit a ZIP file named as ROLL_NUMBER.zip (e.g., PhDXXXXX.zip). The ZIP file should contain:
  - A PDF with a clear explanation of your results and approach.
  - Code files in .py or .ipynb formats only. Colab links will not be accepted—download your Colab files and include them in the ZIP.

---

**Reading:** Familiarize yourself with LoRA and its variants: [10 Marks- Viva in Evaluation]

- LoRA Developer Guide
- LoRA Conceptual Guide

---

**Task:** Fine-tune a Large Language Model (LLM) using **QLoRA** for the task of **Natural Language Inference (NLI)**. Finetune for 5 epochs and make sure that you save the model after every epoch. Save the final trained model, you will be asked to get some inferences while evaluations [12 Marks]

**Dataset:**

- **Training:** Select 1000 samples from the SNLI dataset by choosing every 550th sample from a total of 550k.
- **Testing:** Select 100 samples by choosing every 100th sample from a total of 10k.
- **Validation:** Select 100 samples by choosing every 100th sample from a total of 10k.

Dataset: SNLI on Hugging Face

**Model:**

- Use **Phi2** from Hugging Face (Phi2 Model).

**Note: Make sure that you change Colab run time to GPU only when you are sure of your code, till then use CPU**

---

**Report:** Your report should include the following: [3 Marks]

1. Accuracy comparison between the **pretrained** and **fine-tuned** models on the test set.
2. **Time taken** to fine-tune the model using QLoRA.
3. Total **parameters** in the model and the number of parameters fine-tuned.
4. **Resources used** (e.g., hardware, memory) during fine-tuning.
5. Failure cases of the pretrained model that were **corrected** by the fine-tuned model, as well as those that were **not corrected**. Provide possible explanations for both.

---

**Additional Resources:**

- Finetuning on colab
- Fine-tune LLM with QLoRA
- Example Notebook for Fine-Tuning
- LoRA and QLoRA Details
- How to Fine-Tune LLaVA

---