CSE665: Large Language Models

Assignment 2
Trade off between Model size, Prompt type, Time Taken and Quality

Maximum Marks: 25

- ❖ The deadline is strict and late submissions will not be accepted since the LLM class schedule is already discussed in class.
- ❖ It is mandatory to maintain a github repository for assignments since subsequent assignments will require the same files and functions for update.
- ❖ The marks of each task of assignment will be provided only if the student is also able to answer questions asked by TA related to the task in evaluation.
- ❖ You need to submit a zip with name ROLL_NUMBER.zip (eg:PhDXXXXX.zip) which should have:
    - A pdf which should have all your results and conclusions mentioned and link to the github repository.
    - Code files in .py/.ipynb format only, colab links will not be accepted. (Download your collab file and put in zip)

- **Task**:

    - Three publicly available LLMs of different sizes:
        - https://huggingface.co/google/gemma-2b-it
        - https://huggingface.co/microsoft/Phi-3.5-mini-instruct
        - https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

        (NOTE: In case you observe that inference of 8B model is taking time you can use free api credits from https://together.ai/ , you can use these and many more similar api key exists)

    - Use this https://huggingface.co/datasets/cais/mmlu/viewer/college_mathematics dataset of Mathematics question answers to do this task.

    - Implement functions to perform inference using below type of prompts for all three LLMs.
        - **Zero Shot :** “Choose the answer to the given question from below options. [Question][Option 1][Option 2][Option 3][Option 4]” [3 MARKS]

- **Chain of Thought (Zero shot) : "**Choose answer of given question from below options.[Question][Option 1][Option 2][Option 3][Option 4]. Think step by step" [6 MARKS]
- **ReAct Prompting** (https://github.com/dair-ai/Prompt-Engineering-Guide/blob/main/notebooks/react.ipynb) [6 MARKS]

  For more clarity you can refer : https://www.promptingguide.ai/

Evaluate and compare the inference time for each LLM using above prompts. Additionally, assess the accuracy of the generated outputs and discuss the trade-offs between model size, inference speed, prompt used and output quality. [3 MARKS]

Read technical reports and papers related to given 3 LLM's and give reasons why Model X performed better than Model Y or comparable in your case by citing relevant reasons from verified resources. [7 MARKS]