CSE665: Large Language Models

Assignment 1
Exploring and Probing Large Language Models

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 approach clearly mentioned and discussed.
    - Code files in .py/.ipynb format only, colab links will not be accepted. (Download your colab file and put in zip)

PART 1- Exploration (10 Marks)

Objective: You have to explore and analyze the phenomenon of hallucinations in two Large Language Models (LLMs). You'll identify specific examples of hallucinations of different types, and apply RAG to reduce these issues.

LLMs to Use:

- LLAMA 3.1: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
- OpenHathi: https://huggingface.co/sarvamai/OpenHathi-7B-Hi-v0.1-Base

Task 1 (4 Marks) :

- Identify 3 examples of the Self-consistency and Fact Checking each, perform this for both LLMs, resulting in 12 examples in total. (3 Marks)
- Write a short report analyzing the types of hallucinations encountered in these models. (1 Mark)

Task 2 (6 Marks) :

- Use Retrieval-Augmented Generation (RAG) techniques to minimize or solve all the hallucinations identified in the previous step. (6 Marks)

Make logical assumptions wherever needed.

References:LLM hansons slides

PART 2 - Probing (15 Marks)

Objective: Your task is to explore how well Large Language Models (LLMs) encode information about various topics or entities at different layers. You will use probing techniques to analyze the model's ability to retain and predict specific information based on a dataset of your choice. This exercise aims to deepen your understanding of how LLMs represent knowledge and how it can be extracted.

Steps:

1. Select a Dataset:
   - Choose a dataset that contains structured information about a specific topic or entities (e.g., historical figures, geographical locations, scientific concepts, people).
   - Ensure your dataset has several fields that can be predicted (e.g.,, population, notable achievements, dates, any number, any class).
2. Design a Prompt:
   - Create prompts that query the LLM about the entities or topics in your dataset (e.g., If it's a database of geographical locations "Tell me about Paris?").
3. Extract Embeddings: (Up to this step - 1.5 Marks)
   - Use your designed prompts to perform a forward pass through an LLM (Use LLAMA 3) and extract the embedding of the final token.
4. Set Up a Linear Regression and Classification model: (This step 3 Marks)
   - Use the extracted token embeddings as inputs for a linear regression model and a classification model (Regressor and classifier are just like a head that can extract info from embedding). You will aim to predict several fields from your chosen dataset, such as:
     - A numeric field using regression: (e.g., birth year, population, discovery date, number of mentions, page views, citations, etc).
     - A class variable:(e.g: Gender, profession, etc).
5. Evaluate the Probing Results: (1.5 Marks)
   - Analyze the results by evaluating how well your regression models predict both the fields.
   - Compare the performance across model first layer embeddings, mid layer embeddings and final layer embeddings

6. Discussion: (1.5 Mark)
    - Reflect on your findings. What do the results indicate about the LLM's ability to encode the information in your dataset?
    - Discuss any patterns or anomalies you observe, such as differences in performance between various models or layers.

NOTE:

For Datasets you can explore hugging face for datasets like IMDB, DBPEDIA, Wiki_BIO, or any other .

You can use different dataset for regression and classification based probing.

In above there are 2 parts regressor based probing and classification based probing and both have 7.5 marks each.