Assignment 1: Large Language Models for Text Generation (15 points)

CS 410/510 Large Language Models, Fall 2024

In this assignment, you will explore large language models (LLMs) from two model families – Llama and Phi. We choose to use open-source models in this assignment because there tends to be more information available about how these models were trained than there is for 'closed-source' proprietary models. The goal is to generate text using these models, analyze the generated outputs, and build an understanding of some of the strengths and limitations of LLMs.

The deliverable for this assignment is a Google Colab notebook converted to a PDF report discussing your observations for questions described below. When appropriate, you should use figures and tables to support your answers. Please submit the PDF on Canvas[1].

Instructions

You will explore 3 different models in this assignment. These include:

1. Llama-3.2-1B: https://huggingface.co/meta-llama/Llama-3.2-1B
2. Llama-3.2-3B: https://huggingface.co/meta-llama/Llama-3.2-3B
3. Phi-3.5-mini-instruct: https://huggingface.co/microsoft/Phi-3.5-mini-instruct

These three models are gated on Hugging Face[2], meaning you need to request access on each model's page. Click the links above to make the requests. It may take around an hour for access to be granted. Once approved, you will receive an email notification. After that, go to your Hugging Face account to generate an access token (Link), which you will need to use these models.

---

[1] We have done our best to make sure all homework problems are solvable and that the LLMs will behave in a relatively predictable way. However, you are free to experiment and get creative to solve the assignment. If you are struggling to get the models to behave in a way that allows you to answer any of the questions, please let us know by posting on Slack.

[2] https://huggingface.co/

Analysis

Q1. [3 points] Describe three differences between Llama 3.2 models and Phi-3.5 model.

Q2. [6 points] Generate a story of 200 words that starts with the words "*Once upon a time*" using each of these models. You should have 3 outputs in total.

a) Show the process you used to generate these. Discuss any variations or challenges encountered in the generation process of each model.
b) Compute the following two metrics -- type-token ratio (TTR) and perplexity -- for your generated text outputs which help to **quantitatively** evaluate the generated texts and discuss the results.
c) Conduct a **qualitative** analysis of the text generations, highlighting any patterns or notable variations between these models. Share your observations on the creativity, coherence, and overall quality of the generated stories.

Q3. [1 point] Investigate the relationship between your quality metrics (type-token ratio and perplexity) and the size of these models. What trends do you notice? Discuss the implications of these findings in terms of model scalability and performance.

Q4. [1 points] The Phi-3.5 model includes "instruct" in its name. Did you notice any differences in the generated text compared to the other models? Did you have to modify your approach for this model?

Q5. [3 points] For your best performing model, how would you improve its performance further? Experiment with a few parameters that control the generation strategy (e.g., sampling or beam search) and/or parameters that manipulate the model output logits (temperature, top-k decoding, top-p decoding, repetition penalty, etc.). Present your results obtained using different settings. Describe what you tried and whether it improved your results. Explain why your results improved or did not improve.

[1 point] for a clearly documented, well-formatted report.

All the best!