Homework 3

DUE DATE: NOV/20/2024 23:59

Decoding, Prompting and Instruction Tuning (5 Points)

1 Coding: implementation and observation

Instruction: Please submit an ipynb file named generation.ipynb via LMS.

Large language models (LLMs) show amazing capabilities throughout various tasks using their superior knowledge of the world and language acquired through the process of training. However, it is difficult to fully utilize their potential without the help of suitable decoding, prompting strategies and instruction tuning. These techniques can steer the model towards the right direction and help us achieve what we want from our model.

The primary objective of this assignment is to familiarize you with such methods. By implementing and comparing these methods yourself, you will be able to better understand the way LLMs work and why such concepts of decoding, prompting and instruction tuning are necessary.

[Tip: Make sure to leave screenshots of your generation results. You are going to need them for your written report!]

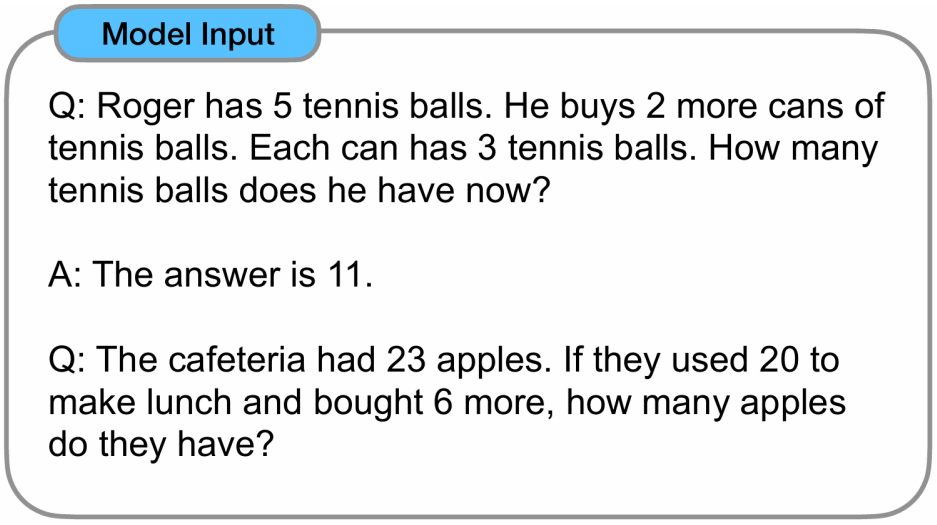

Figure 1: An example of a question that involves complex reasoning

Detailed Requirements:

- Choose a suitable model from the Hugging Face hub. Make sure to choose a **"Text Generation"** model that supports both the **base** version and the **instruction fine-tuned** version. (One good example is the Llama model[1])
- Prepare at least **three unfinished sentences** that you would like to use as input for the model. (e.g. "Hanyang University is ")
- Implement at least **five different decoding strategies**. Greedy decoding must be included within the five. (You can refer to the official Hugging Face document[2] or this post[3] on decoding strategies.)
- Using the three candidate sentences you came up with earlier as input, generate some text using the five decoding strategies. **(Use the base model for this part)**
- Try and come up with one or more questions that requires complex step-by-step reasoning. (An example of such question is illustrated in Figure1.)
- Observe how the **base** model reacts to your questions. (As for decoding, use one of the five strategies from above except greedy decoding.)
- Utilize various **prompting** techniques such as few-shot prompting and chain-of-thought prompting to help your model better solve the given questions. (You can refer to the official Hugging Face LLM prompting guide[4].)
- With all other conditions fixed, generate answers to the same questions using the **instruct** model, applying all the **prompting** techniques used above. (If your GPU is not big enough to accomodate two models simultaneously you may need to restart your Jupyter notebook session for this part.)

2 Written: writing a report on your work

Instruction: Please submit a pdf file named report.pdf via LMS.

Write a report that provides detailed explanations to justify your decisions. You may also include some stories about the difficulties you faced during the above coding part of the assignment. **Please include screenshots of your generation results for each corresponding part.**

[1] https://huggingface.co/meta-llama
[2] https://huggingface.co/docs/transformers/generation_strategies
[3] https://huggingface.co/blog/how-to-generate
[4] https://huggingface.co/docs/transformers/tasks/prompting

Detailed Requirements: Your submission must include the following information.

- **Model information:** Provide some information about the model you used for the above coding part of your assignment. You should also mention **why** you chose that model. (The reason does not have to be too formal, you can just say that you chose a certain model because it is small enough to fit on your GPU. Any "reasonable" explanation should do.)
- **Decoding strategies:** Describe the decoding strategies of your choice, focusing on **how each one works** along with its **pros and cons**. Based on your explanation, compare the generation outcomes across different decoding tactics.
- **Prompting:** Explain about the prompting methods you used to help the model handle difficult questions and evaluate the generated results.
- **Instruction tuning:** Regarding your generated results, explain the key differences between base models and instruction fine-tuned models. Additionally, discuss which type of model you believe is more suitable for commercial applications.

How to Submit: Please upload your code and report as two separate files through LMS.[5] Do **not** compress them into a single zip file.

End of document.

[5] https://lms.hanyang.ac.kr