DUE DATE: NOV/24/2025 23:59

Decoding, Prompting and Instruction Tuning (5 Points)

1 Coding: implementation and observation

Instruction: Please submit an .ipynb file named HW3_generation.ipynb via LMS.

Introduction

Large language models (LLMs) demonstrate remarkable capabilities across various tasks, but their performance heavily depends on the decoding strategy and prmopt. For example, **greedy decoding** deterministically picks the most probable token, while **beam search** explores multiple hypotheses to find higher-likelihood sequences.

In this assignment, you will implement and compare these decoding methods to observe their effects on fluency, diversity, factuality, and length, and understand why decoding, prompting, and instruction tuning are crucial for effectively steering LLM behavior.

[**Tip: Include screenshots of your generations in the notebook—you will refer to them in your short write-up.**]

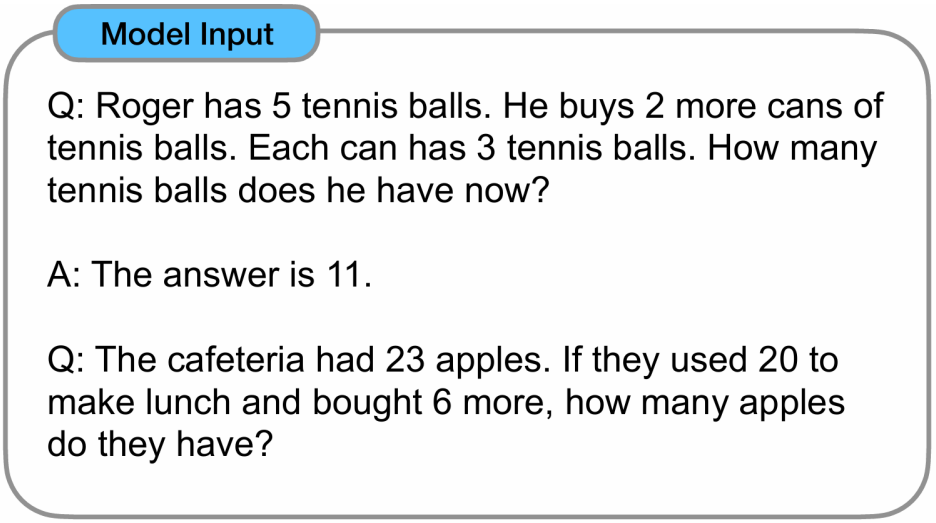

Figure 1: An example of a question that involves complex reasoning

1.1 Part1 : Implementing Decoding Strategies

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])
- 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.)
- Try and come up with 3 questions that requires complex step-by-step reasoning. (An example of such question is illustrated in Figure1.)
- Observe how the model reacts to your questions. (As for decoding, use one of the five strategies from above except greedy decoding.)
- Using the three candidate sentences you came up with earlier as input, generate some text using the five decoding strategies.)
- Choose a decoding strategy that you think will work best for your 3 candidate sentences. Utilize **2 different 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].)

1.2 Part2 : The Story Cloze Test

In cloze tests, a segment of text is removed and the person taking the test is asked to fill in the blank. In the Story Cloze Test, the ending to the 5-sentence story is missing and the model has to generate the last sentence. Examples of the task can be found here. https://cs.rochester.edu/nlp/rocstories/ We will be using the first 4-sentences of this dataset as input and generate the last (5th) sentence using a LLM model of your choice.

Detailed Requirements:

- Download the Cloze Test dataset. Refer to the StoryClozeTest.ipynb file.
- Select one story(data sample) from the Cloze Test dataset to use as the story you're generating on. (We will be using the implementation that you did in Part 1.)

[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

- Using the **five different decoding strategies** you implemented earlier, generate the 5th/final sentence for the story you have chosen. (You should have 5 different generated outputs for each decoding strategy)
- Evaluate the 5 generated outputs against the **gold 5th sentence** using **at least 2** different evaluation metrics. (BLEU, ROUGE, ...)

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.**

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 should be related to the task you are doing.)
- **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 the prompting methods you used to help the model handle difficult questions and evaluate the generated results.
- **Evaluating The Story Cloze Test:** Explain how each decoding strategy performed on the evaluation metrics. Mention in detail why you think each decoding strategy got the score that it did, and which decoding strategy metrics you think is the best for this task.

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