Assignment 2

Please submit the code and report (in pdf format) on Blackboard system before 23:59 Nov. 15. Report can be written either in English or Chinese. Name your report as `studentID_Name.pdf`.

Note:

- If you use ChatGPT etc., list all the prompts you use. Note that the LLM may have hallucination problem and you are responsible for the correctness of the answers.
- You can use your laptop or Google Colab computing platform, etc.

Q1

Explain and answer the following questions in your own words. You can read the blog and other papers mentioned in the survey for more details.

(1) [2 points] Can LoRA match the performance of full fine-tuning, and if so, under which conditions?

(2) [1 point] How does learning rate and batch size effect the performance of LoRA-based PEFT methods?

(3) [2 points] Implement the TODOs in the attached python file `lora_training.py`. Then run the training script with `Qwen/Qwen2.5-0.5B` and show the printed results. Please refer to the notes in the code file for more instructions.

Q2

Explain and answer the following questions in your own words.

(1) [2 points] The principles behind different sampling methods: **Greedy**, **Top-K**, **Top-P**, and **Min-P** decoding.

(2) [1 point] How changing parameters such as `temperature`, `k`, `p`, and `p_min` influences the generated results for each decoding algorithm (if applied).

(3) [2 points] Complete the missing parts of the provided Python file **`decoding.py`**. Then run the training script with `Qwen/Qwen2.5-0.5B-instruct` and show the printed results.