Assignment 3

100 points

Implement the "Self Alignment with Instruction Backtranslation" paper. When fine tuning the model, use LoRA. You will not be able to do full finetuning because there is not enough memory.

Link to paper: https://arxiv.org/pdf/2308.06259.pdf
Colab's GPU usage is limited. Try to first prototype and get things working on the CPU first before training on the GPU with the full dataset. If you are not able to connect to a GPU on colab, you can try to create a PyTorch Lightning Studio or a Kaggle notebook.

In particular:

1. Finetune the base language model (llama2 7B) with (output, instruction) pairs {(yi, xi)} from the seed data to obtain a backward model Myx := p(x|y). In other words, finetune a model that uses the output to predict the instruction. Use the openassistant-guanaco training set dataset. (25 points)
    a. Push the backwards model to HF and paste url here
2. Self-Augmentation -- Randomly sample a subset of size 150 and generate **instructions** from the LIMA dataset's completions and filtering out any mutli-turn examples. Print out 5 examples of generated instructions. (25 points)
    a. (generated instructions from backwards model, response is from LIMA) pairs
    b. Single turn:
        i. Single turn: (What is the capital of France?, Paris)
        ii. Multi turn: (What is the meaning of life, 42, Why is it 42?, That's universe, ...)
3. Self curation (selecting high quality examples) using few shot prompting in addition to the prompt in Table 1 of the paper. Print out 5 examples of high quality examples and 5 examples of low quality examples. (25 points)
    a. Push the dataset to HF hub and paste the url here
    b. Goal is to filter out bad samples
    c. Method: using an LLM to rate the example
        i. LLM (meta/llama-7b-chat-hf): LLM("Evaluate the quality of the instruction/response pair" + example." Rate it from 1-5)
4. Finetune base model on dataset generated by step 3. Print out 5 example responses. (25 points)
    a. Push the instruction fine tuned model to HF hub and paste the url here

Please include a link to your colab notebook here: