Assignment 4

In this assignment, we will be generating a preference dataset with PairRM and fine tuning a model with DPO. This is a powerful training recipe that is behind some of the top models according to Alpaca Eval.
You may use llama-3.2 1B or llama-3.2 3B.

Preference Dataset Collection and DPO Model Training

Part 1: Dataset Generation and Judge Implementation (40 points)
Create two separate preference datasets using different collection methods:

a) LLM Judge-Based Collection (20 points)
- Implement an LLM-based judge system
- Document your reasoning for the judge's prompt design
- Explain how you ensure consistent and reliable preference judgments
- Include examples of the judge's evaluation process
- You can choose between using local inference on Colab/Lightning studio or a 3rd party provider like fireworks ai/openai/together ai

b) PairRM-Based Collection (20 points)
- Extract 50 instructions from the Lima dataset
- Generate 5 responses per instruction using the llama-3.2 chat template
- Apply PairRM to create preference pairs
- Upload dataset to HuggingFace
- Submit repository link

Part 2: Model Training and Evaluation (60 points)

a) DPO Fine-tuning (40 points)
- Fine-tune llama-3.2 using PairRM preference dataset
- Fine-tune llama-3.2 using LLM Judge preference dataset
- Document training parameters and process
- Upload PEFT adapters to HuggingFace
- Submit repository links

b) Comparative Analysis (20 points)
- Select 10 novel instructions (not in training data)
- Generate completions using:
  * Original llama-3.2
  * DPO fine-tuned model (LLM judge dataset)
  * DPO fine-tuned model (PairRM dataset)
- Present results in a pandas DataFrame
- Analyze and compare the quality of completions

- Include quantitative and qualitative observations

Address the following points:
1. Qualitative differences in model outputs
2. Training stability across iterations
3. Computational efficiency considerations
4. Potential limitations and failure modes
5. Suggestions for improvement

The comparative analysis must be original work. No LLM assistance is permitted. Responses will be screened through AI detection tools.

Grading Criteria for Free Response:
- Depth of technical understanding
- Critical analysis of results
- Clear articulation of observations
- Original insights and suggestions
- Proper technical writing style

Extra Credit: Iterative DPO Implementation and Analysis (30 points)

a) Implementation (20 points)
- Implement the iterative DPO algorithm as described in "Self Rewarding Language Models"
- Train multiple iterations of the model (minimum 2 iterations)
- Document:
  * Implementation details
  * Training parameters

b) Comparative Analysis (10 points)
Free Response Question (~250 words)
Compare and analyze the performance and behavioral differences against the base llama-3.2 model, the DPO-PairRM model, and DPO-LLM-judge model