Assignment: Text Extraction using Vision Language Models

1. Dataset

- The dataset to be used for the assignment can be found at this link.
- Create appropriate training and test splits of the data for baseline inference and subsequent fine-tuning.
- For fine-tuning and evaluation, create necessary training and test data. You may use a combination of human annotation and extraction of text using methods such as OCR.

2. Baseline Inference

- Use the dataset to run inference using Llama 3.2 11B Vision model. That is, extract the text from images using the pretrained model. Feel free to use a quantized version of the model.
- Evaluate the performance of the extraction using following metrics:
    - Word Error Rate
    - Character Error Rate

4. Text Organization

- Process and structure the extracted text for better readability and usability.
- If needed, apply post-processing techniques to refine the text output.

5. Fine-Tuning

- Fine-tune the Llama 3.2 11B Vision model on the training split.
- Use parameter-efficient tuning methods (e.g., LoRA) to optimize performance to avoid excessive computational cost.

6. Comparison

- Compare the baseline (pre-trained) model's inference performance with the fine-tuned model.
- Report differences in extraction accuracy using the evaluation metrics.

Share your end-to-end reproducible code as a Python notebook, preferably using Google Colab.

(Optional) Extra Credit

Similar to the above exercise, extract text from documents in Indic language using VLMs. The dataset can be found here. You may use models such as Qwen2-VL for this. Feel free to use others that perform extraction better. Perform finetuning and report comparisons.