Yavar Internship Selection – May 2025

Assessment Problem Statement-2

Title: Image Captioning from Contextual Metadata Using Vision-Language Models (VLMs)

Objective:

Develop a fine-tuned open-source Vision-Language Model (VLM) that generates both **concise** and **detailed** captions for a given image, using its **visual content** and **surrounding textual metadata** (section headers, caption, above/below text around the image, footnotes, etc.).

Caption should match not just what's in the image, but what part of the context refers to that image. Generate both captions **grounded** in:

- Visual content of the image
- Surrounding context: section_header, above_text, below_text, footnote, and optional caption

Constraints & Requirements:

- Images can be tables, graphs, charts, pictures, layout diagrams, photos, logos etc
- Use only open-source models and libraries
- Fine-tuning is mandatory
- Preprocessing images essential
- Use BLEU/ROUGE/custom semantic similarity measurements

Inputs:

- img_folder/: Folder containing image files (figures, graphs, tables, circuit diagrams, etc.) **[To test, images would be loaded here]**
- metadata_folder/: Corresponding .txt files (one per image) with structure:
    - section_header: null or string
    - above_text: null or string

- caption: null or string
- picture_id: #/pictures/0
- footnote: null or string
- below_text: null or string

[To test the solution, metadata files would be added here]

Expected Outputs:

In the output_folder/:

For each image in img_folder/:

- **Overlaid image with:**
  - Concise caption (short, summary-style) in one color (e.g., blue)
  - Detailed caption (explains structure/trend/function) in another color (e.g., Red)
- Save to output_folder/:
  - Annotated image file [Image + captions]
  - captions.json with both captions + confidence scores

Verifiability Requirements:

- Include confidence scores (0–1) for each caption
- Log inconsistencies or low-confidence outputs
- Check that generated captions are consistent with given metadata (e.g., do not contradict section headers) and image
- Highlight or underline the caption text if confidence is low

Evaluation Criteria

- Relevance: How well the caption aligns with both the image and the provided context
- Fine tuning methodologies adopted
- Clean documentation of approach adopted

Timeline for completing assessment:

Time of expected completion: EoD: 30-May-2025

(a) A writeup explaining crisply the approach adopted, models used, fine tuning, preprocessing done etc

(b) Solution to be uploaded in Github and the link to be provided to the Yavar team with appropriate permissions.