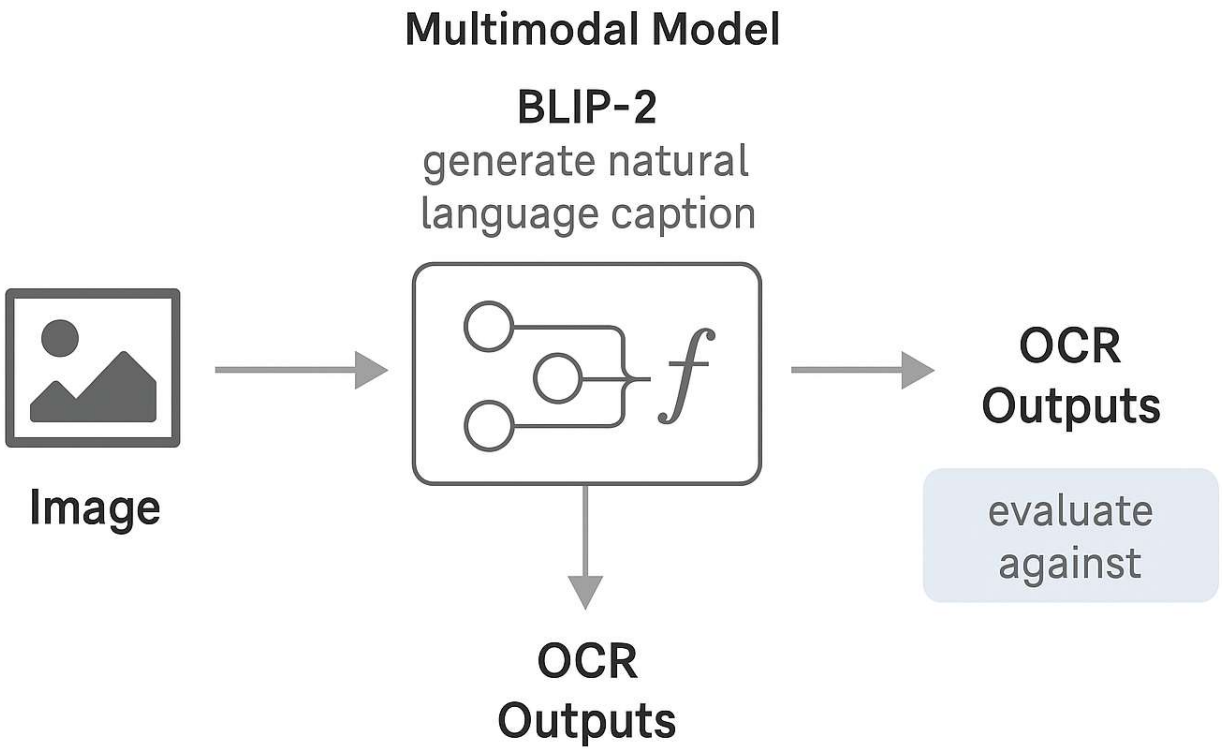
Multimodal Language Usage:

In this study, a multimodal approach was employed by integrating EasyOCR and the BLIP-2(Bootstrapping Language-Image Pretraining) model to process and interpret image data, specifically handwritten medical prescriptions. EasyOCR, a reliable open-source optical character recognition engine, was used to extract textual information from each image in the dataset. This extracted text, along with the corresponding image, was then passed to BLIP-2—a state-of-the-art vision-language model—for conditional caption generation.

The objective was to evaluate how effectively BLIP-2 could utilize OCR-derived textual content in tandem with visual features to produce meaningful and contextually accurate image captions. This approach mirrors human visual-linguistic reasoning, where textual elements within an image contribute substantially to semantic interpretation.

To assess performance, the outputs of EasyOCR and BLIP-2 were compared using a comprehensive set of evaluation metrics, including Fuzzy Match Score, Cosine Similarity, BLEU Score, Levenshtein Ratio, and Jaccard Similarity. The analysis revealed an average cosine similarity of 0.91 and a Levenshtein ratio of 0.77, indicating strong alignment between OCR extractions and BLIP-2 captions. Outlier analysis and metric-based heatmaps further highlighted discrepancies and model behaviour across diverse image types.



My Evaluation Strategy for the Multimodal Model

For this project, I used EasyOCR for optical character recognition (OCR) and BLIP-2, a vision-language model, to perform text extraction and image captioning tasks. The goal was to assess how well these models can work together to understand both the visual and textual content in images. My evaluation strategy involved several key steps:

1. Data Preparation

I used a diverse set of images that included different types of text-heavy content, such as scanned documents, signs, and printed text. This variety helped me see how the models handle different formats of text in various contexts. First, I ran the images through EasyOCR to extract the text. Then, I fed the extracted text and the images into BLIP-2 to generate captions based on both the text and the visual elements.

2. Evaluation Metrics

To measure the performance of the OCR and image captioning, I relied on a few different metrics that help assess both the quality and accuracy of the outputs:

- **Fuzzy Match Score**: This metric checks how closely the OCR output from EasyOCR matches the caption generated by BLIP-2. A higher fuzzy-match score means the extracted text and the generated caption are more aligned.
- **Cosine Similarity**: This one looks at the similarity between the vector representations of the OCR text and the generated caption. Higher cosine similarity suggests the two pieces of text are closer in meaning or context.
- **BLEU Score**: I used BLEU (Bilingual Evaluation Understudy) score to evaluate how similar the BLIP-2 caption is to reference text. A score closer to 1 indicates the caption closely matches the reference text.
- **Levenshtein Ratio**: This metric measures how many character edits (insertions, deletions, or substitutions) are needed to transform one string into the other. A higher Levenshtein ratio means fewer edits are needed.
- **Jaccard Similarity**: This one compares the overlap of words between the OCR output and the BLIP-2 generated caption. The higher the Jaccard similarity, the more words the two texts share.

3. Outlier Detection and Analysis

Next, I performed outlier detection to flag any images where the model was performing poorly on one or more of the metrics. If a specific image had unusually low scores across multiple metrics, I flagged it as an outlier for further investigation. This helped me identify certain types of images or text formats where the models struggled, and it gave me insight into areas for improvement.

4. Statistical Analysis

I calculated the average score for each of the metrics to get a sense of the overall performance across the dataset. I also created a **correlation matrix** to visualize the relationship between different metrics. This helped me see whether improving one metric, like fuzzy-match score, also led to improvements in others, like cosine similarity or BLEU score.

5. Visual Results and Presentation

To communicate the results clearly, I created visual representations, such as heatmaps and graphs, to show the correlations and outliers. This helped me quickly spot patterns in the data and highlighted the images that caused the most discrepancies between EasyOCR and BLIP-2.

6. Final Assessment

Finally, I wrapped up my analysis with a comprehensive assessment of the models' performance. I looked at the strengths and weaknesses across the various metrics and identified areas for improvement. I also provided suggestions for potential model improvements, such as fine-tuning specific components or diversifying the dataset to handle edge cases better.

In the end, this strategy gave me a clear picture of how well EasyOCR and BLIP-2 perform together and helped me pinpoint specific areas where I could improve the model's performance for future applications.