# Summarize and QnA PDF Using LLMs

## Title

Summarization and Question Answering from PDFs using LLM-Based NLP Pipelines

## Abstract

This project implements an LLM-based system capable of reading PDF documents, generating concise summaries, and answering user-defined questions based on the content. It leverages state-of-the-art language models from HuggingFace Transformers, combining `facebook/bart-large-cnn` for summarization and `deepset/roberta-large-squad2` for question answering. The system enhances contextual understanding using paragraph-level similarity scoring and answer post-processing for higher accuracy.

## Introduction

The goal of the project is to build an AI-powered assistant that can automatically summarize academic PDF documents and answer user queries based on the content. The tool aims to help students, researchers, and professionals quickly extract useful information from long documents without reading them end-to-end.

## Technologies Used

- HuggingFace Transformers (facebook/bart-large-cnn, roberta-large-squad2)

- PyMuPDF (fitz) for PDF text extraction

- Python (re, string, etc.)

- Kaggle GPU environment

## System Architecture

1. PDF Text Extraction: Uses PyMuPDF to extract structured text from uploaded PDF.

2. Summarization: Uses BART model to generate an abstract summary.

3. Question Answering: Uses RoBERTa model to find relevant answers from context.

4. Relevance Scoring: Calculates semantic similarity to find best paragraph context.

5. Answer Enhancement: Expands or corrects short answers using logic-based techniques.

## Evaluation & Results

The system correctly identified the dataset (MMWHS), preprocessing steps like resizing and normalization, but occasionally missed specifics like file format (.nii.gz) or model name (U-Net). Estimated answer accuracy is 70–85% on structured questions. Further improvement possible by fine-tuning or upgrading the models.

## Limitations

- May miss answers if the question is multi-part or phrased abstractly.

- Pretrained models not fine-tuned on domain-specific data.

- PDF structure variations may reduce extraction quality.

## Future Scope

- Integrate GPT-4 for deeper reasoning and generative answers.

- Add Streamlit/Gradio UI for web-based demo.

- Extend to support OCR for scanned PDFs and table summarization.

## Conclusion

This project successfully demonstrates a proof-of-concept system that combines LLMs with text processing to enable document understanding. With further tuning and UI improvements, this can become a useful tool for education and research.