Abstract: Multimodal Document Intelligence System

With the exponential growth of unstructured digital documents—ranging from legal contracts and research papers to handwritten notes and resumes—there is a pressing need for intelligent systems that can understand, summarize, and visually interpret document content. This project presents a comprehensive Document Intelligence System that leverages transformer-based vision-language models and Retrieval-Augmented Generation (RAG) to perform end-to-end analysis of heterogeneous documents.

The system supports auto-summarization, semantic Q&A, document classification, and context-aware visual explanation, even on complex formats like PDFs and handwritten documents. A key innovation is the generation of flowcharts and structured visual representations derived from document logic—especially useful in legal or technical domains.

By combining OCR pipelines (for handwritten and scanned data) with multimodal transformers, the system extracts, understands, and converts raw documents into natural language answers, structured metadata, and interactive visual overlays.

Designed for non-expert users, the system simplifies complex documents into intuitive formats, with applications in legal, academic, and enterprise domains.