**MINOR PROJECT**

**Synopsis Report On**

**TITLE: PolyDoc AI –Multi-lingual Document**

**Understanding and Layout Preservation System.**

**BACHELOR OF COMPUTER APPLICATION (BCA)**



**KLE TECHNOLOGICAL UNIVERSITY**



by

|  |  |
| --- | --- |
| **NAME** | **USN** |
| Darshan Janganure | 01FE23BCA071 |
| Pooja Madiwalar | 01FE23BCA074 |
| Sayeda Iram | 01FE23BCA111 |
| Mohemmed Yaseen | 01FE23BCA118 |

**Under the guidance of**

Prof.Shivayogi V Hublikar





**KLE TECHNOLOGICAL UNIVERSITY**

**Earlier Known as**

B. V. Bhoomaraddi College of Engineering

**Vidyanagar Hubballi-580031**

**2025-2026**

# Title: PolyDoc AI –Multi-lingual Document Understanding and Layout Preservation System

1. **Introduction**

In today’s digital world, documents come in a variety of formats such as PDFs, DOCX, PPTX, and scanned images. Extracting meaningful information while preserving the original layout and formatting is a challenging task, especially for multi-lingual and complex documents. PolyDoc AI is designed to address this challenge by leveraging state-of-the-art AI and machine learning techniques to automate document understanding, translation, summarization, and layout preservation.

This project was chosen due to the increasing need for automated document processing in businesses, education, and government sectors, where manual processing is time-consuming and error-prone. With globalization, handling multi-lingual documents is crucial, making PolyDoc AI relevant for real-world applications in document management systems, digital archiving, and knowledge extraction.

# Problem Statement

Digitization/ Text Extraction for multi-lingual, noisy documents while preserving document Layout including hand written documents.

**PS Description:**

In today’s global and digitized world, documents are far more than just sequences of text. They are structured, visually rich, and multilingual—ranging from legal contracts and academic papers to business reports, government forms, and presentation decks. These documents exist in diverse formats such as Word documents (DOCX), PDFs, PowerPoint slides (PPT), and scanned images including handwritten documents, often containing mixed scripts (e.g., English-Arabic or Hindi- English). Modern AI systems need to read, understand, and generate structured outputs from these documents.

# Objectives

* + To develop an AI-powered system that extracts text and preserves the layout from various document formats such as PDF, DOCX, PPTX, and images.
  + To implement multi-lingual language detection and automatic translation features.
  + To summarize extracted content for quick information retrieval.
  + To maintain the document's original visual and structural layout in the output.
  + To provide structured outputs in formats like JSON and Markdown for easy integration.
  + To design a user-friendly frontend for document upload and output visualization.

# Scope of the Project

**Included:**

* + Text extraction from multiple document types including scanned images.
  + Multi-lingual language detection and translation support.
  + Summarization of extracted content.
  + Layout preservation using AI-based layout analysis.
  + Output formatting in JSON and Markdown.

**Excluded:**

* + - Handwritten text recognition (currently focusing on printed documents).
    - Real-time collaborative editing features.
    - Offline standalone application (web-based system only).

# Literature Review

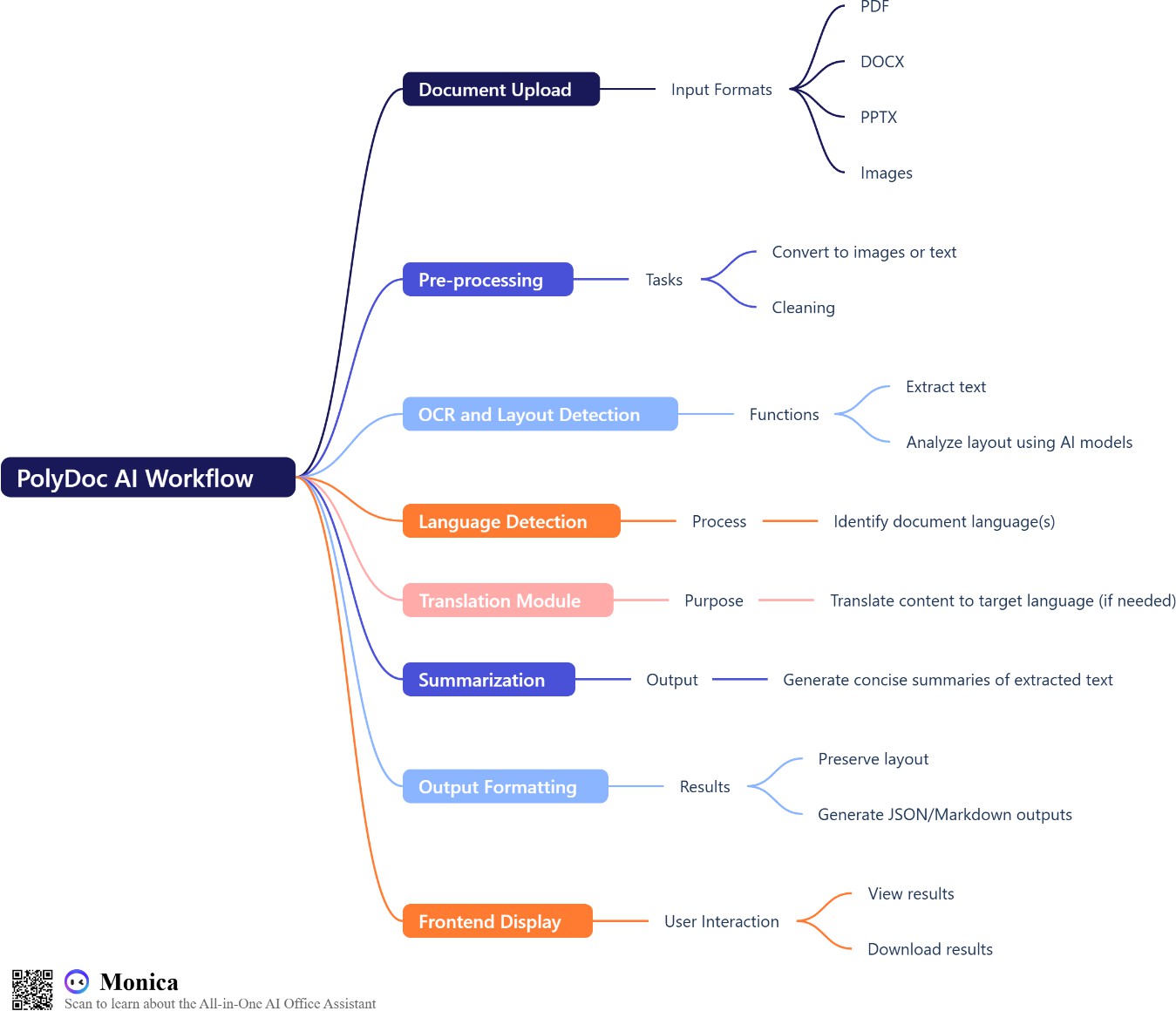
Existing document processing tools such as Tesseract OCR provide basic text extraction but lack sophisticated layout preservation and multi-lingual support. Tools like LayoutParser and Transformers-based models improve layout and semantic understanding but are often limited by language or format support. Machine translation models like mBART and NLLB offer strong translation capabilities but are not commonly integrated with layout-aware document systems.

PolyDoc AI bridges these gaps by combining OCR, layout analysis, advanced translation, and summarization models into a single, cohesive platform, addressing the challenges of multi-format and multi-lingual document understanding.

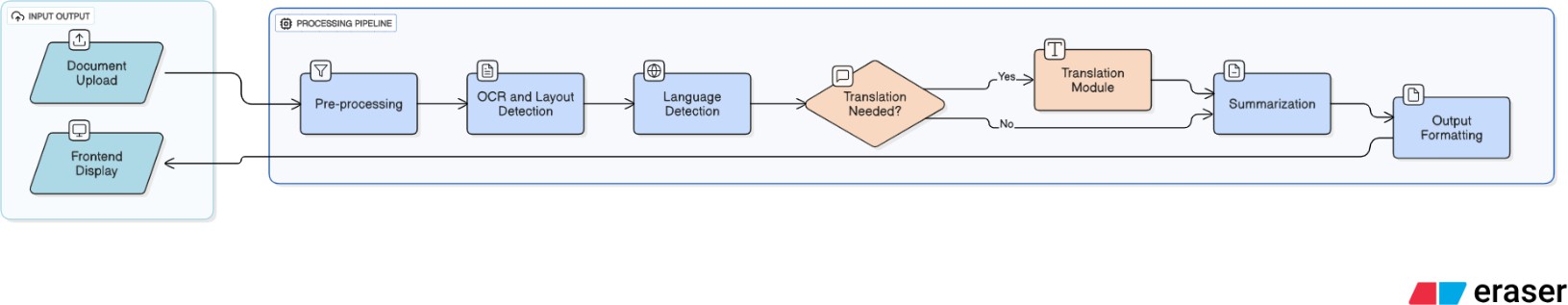
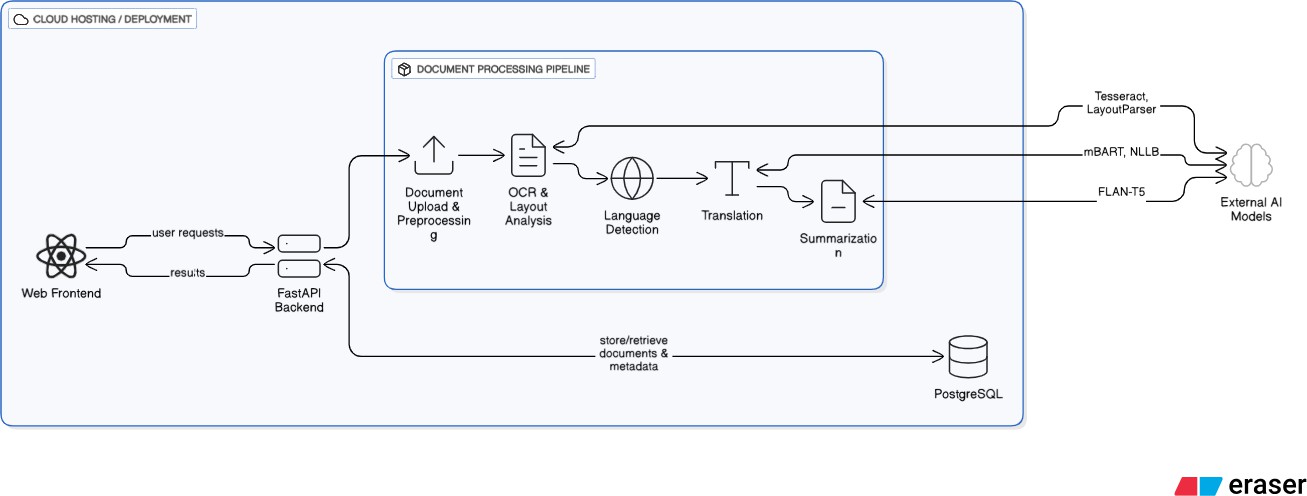
# Methodology

* + **Phase 1:** Document input and pre-processing — upload documents and convert them to image or text form.
  + **Phase 2:** OCR and layout detection — extract text and analyze document layout using LayoutParser and Detectron2.
  + **Phase 3:** Language detection and translation — detect the language using NLP techniques and translate using models like mBART or NLLB.
  + **Phase 4:** Content summarization — generate summaries with transformer-based summarizers such as FLAN-T5.
  + **Phase 5:** Output formatting — structure data into JSON and Markdown while preserving layout.
  + **Phase 6:** Frontend development — create React-based UI for document upload and viewing results.
  + **Phase 7:** Deployment — host the system on cloud platforms for public accessibility.
  + **Technologies:** Python, PyTorch, Transformers, LayoutParser, Tesseract OCR, React.

**Work Flow Diagram**

****

**Architecture Diagram**



**Flow Diagram**

# Hardware and Software Requirements

**Hardware:**

* + Processor: Intel i3 or above
  + RAM: Minimum 8GB (16GB recommended)
  + Storage: SSD with at least 5GB free space

**Software:**

* + OS: Windows
  + Python 3.8+
  + Libraries: PyTorch, Transformers, LayoutParser, Tesseract OCR, FastAPI, React.js
  + Development Tools: VS Code, Git

# Expected Outcome

* + A fully functional AI-based system capable of extracting, translating, summarizing, and preserving the layout of multi-lingual documents.
  + Structured outputs that facilitate integration with other systems.
  + A user-friendly web interface for easy document upload and result visualization.
  + Improved accuracy and usability over existing solutions, saving manual processing time.

# Applications

* + Automated document management in enterprises.
  + Digital archiving and retrieval systems for libraries and governments.
  + Multi-lingual document translation and summarization services.
  + Educational tools for analyzing research papers and study materials.
  + Enhancing accessibility by converting complex documents into simpler formats.

# Conclusion

PolyDoc AI addresses the complex problem of understanding and processing multi-format, multi- lingual documents while preserving their structure and readability. By integrating cutting-edge AI models, this project aims to automate and streamline document workflows, making information extraction efficient and accessible. The system’s adaptability to various languages and formats makes it highly relevant in today’s globalized digital ecosystem.

# References

* + LayoutParser Documentation: https://layoutparser.org/
  + Vaswani, A., et al. "Attention is All You Need." (2017).
  + Liu, Y., et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." (2019).
  + Tesseract OCR: <https://github.com/tesseract-ocr/tesseract>
  + Facebook AI Research. "mBART: Multilingual Sequence-to-Sequence Pre-training for Neural Machine Translation." (2020).
  + Lewis, M., et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation." (2020).
  + FastAPI Documentation: https://fastapi.tiangolo.com/
  + ReactJS Documentation: https://reactjs.org/docs/getting-started.html