AI Agent Architecture Document – PDF Summarizer AI Agent

1. Overview

The PDF Summarizer AI Agent is designed to automate extraction, processing, and summarization of textual content from PDF documents. It leverages a modular architecture combining PDF parsing, text preprocessing, and AI-based summarization to create concise and meaningful summaries.

2. Components

1. **Frontend (Streamlit Web App):** - Provides a simple web interface where users can upload PDF files and receive summarized text output. - Built using Streamlit for ease of deployment and interactivity. 2. **Backend Modules (src/ directory):** - `pdf_extractor.py`: Handles reading and text extraction from PDF files using PyMuPDF (`fitz`). - `summarizer.py`: Applies transformer-based summarization models to compress and refine extracted content. - `main.py`: Orchestrates the command-line execution flow (input/output handling). 3. **Model Used:** - Pretrained transformer model (such as `facebook/bart-large-cnn` or `t5-base`) via Hugging Face Transformers for text summarization. - Selected for its balance between accuracy, fluency, and computational efficiency. 4. **AI Workflow:** - PDF → Text Extraction → Preprocessing → Summarization → Output Summary.

3. Interaction Flow

1. User uploads a PDF file in the web interface. 2. `pdf_extractor.py` extracts readable text using PyMuPDF. 3. Extracted text is sent to the summarization model through the `summarizer.py` module. 4. The summarized text is displayed and can be downloaded as a `.txt` file.

4. Design Justification

- **PyMuPDF (fitz):** Chosen for efficient and accurate PDF text extraction. - **Streamlit:** Enables fast web app prototyping with minimal backend configuration. - **Hugging Face Transformers:** Provides robust pre-trained summarization models. - **Modular Structure:** Facilitates scalability, easy debugging, and future integration with fine-tuning pipelines.

5. Future Extensions

- Add OCR support for scanned PDFs using Tesseract. - Implement fine-tuned models for domain-specific summaries (e.g., research papers, legal documents). - Integrate with cloud storage for persistent summary management.