**MINOR PROJECT REPORT**

On

**“PolyDoc – Multi‑lingual Document Understanding System”**

**BACHELOR OF COMPUTER APPLICATIONS**

Of

**KLE TECHNOLOGICAL UNIVERSITY**

By

|  |  |  |
| --- | --- | --- |
| **Sl. no** | **NAME** | **USN** |
| 1 | Darshan Janganure | 01FE23BCA071 |
| 2 | Pooja Madiwalar | 01FE23BCA074 |
| 3 | Sayeda Iram | 01FE23BCA111 |
| 4 | Mohemmed Yaseen | 01FE23BCA118 |

Under the Guidance of

Prof. Shivayogi V Hublikar

A red and white logo

AI-generated content may be incorrect.

**DEPARTMENT OF COMPUTER APPLICATIONS**

**KLE TECHNOLOGICAL UNIVERSITY**

**Vidyanagar, Hubballi-580031 Karnataka.**

**Academic Year 2024-2025**

DEPARTMENT OF COMPUTER APPLICATIONS

**KLE TECHNOLOGICAL UNIVERSITY**

****

**CERTIFICATE**

This is to certify that the minor project work entitled

**“PolyDoc –** **Multi‑lingual Document Understanding System”**

**Department of Computer Applications Of**

**KLE Technological University,**

**Hubballi, Karnataka**

Is a result of the bonafide work carried out by

|  |  |  |
| --- | --- | --- |
| **Sl. no** | **NAME** | **USN** |
| 1 | Darshan Janganure | 01FE23BCA071 |
| 2 | Pooja Madiwalar | 01FE23BCA074 |
| 3 | Sayeda Iram | 01FE23BCA111 |
| 4 | Mohemmed Yaseen | 01FE23BCA118 |

**During the academic year 2024-2025**

**Prof. Shivayogi V Hublikar Dr. P. R. Patil ME(CS), PhD**

**Professor and HoD**

**Project Guide**

***Viva-Voce Examination***

Name of the Examiners Signature with Date

1. :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
2. :\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ACKNOWLEDGEMENT**

Every successful completion of any undertaking would be complete only after we remember and thank the almighty, the parents, the teachers, and the personalities, who directly or indirectly helped and guided during the execution of that work. The success of this work is equally attributed to all the well-wishers who have encouraged and guided throughout the execution.

We express our deepest gratitude to **Prof. Shivayogi V Hublikar** for their guidance and assistance throughout the project with great interest.

We avail this opportunity to express our deepest gratitude to **Dr. P. R. Patil, Professor & HoD** for encouraging us with his valuable tips.

We are grateful to **Dr. B. S. Anami,** Registrar for his blessings. And we are grateful to **KLE Technological University, Hubballi** which has given us a bright future. We would like to thank all the people for their guidance and valuable suggestions throughout the project.

We also thank all teaching and non-teaching staff members of the BCA department for their invaluable cooperation.

Finally, we would like to express our sincere thanks to our **Parents and Friends** for their enormous encouragement and all others who have extended their helping hands towards the completion of our project.

**ABSTRACT**

**PolyDoc**, an innovative multi-lingual document understanding and layout preservation system designed to address the growing challenges in intelligent document processing across diverse linguistic contexts. PolyDoc represents a significant advancement in the field of Natural Language Processing (NLP) and Optical Character Recognition (OCR) technologies, specifically engineered to handle underrepresented languages such as Hindi and Kannada alongside English, while maintaining the structural integrity and formatting of original documents during the extraction and analysis process.

The system leverages cutting-edge artificial intelligence models, including HuggingFace Transformers and state-of-the-art vector-based semantic search technologies, to provide a comprehensive suite of document processing capabilities. These include automated text extraction from multiple document formats (PDF, DOCX, images), intelligent document summarization, real-time conversational interfaces for document interaction, and advanced search functionalities that understand semantic context rather than relying solely on keyword matching. The architecture follows modern microservices principles, utilizing FastAPI for robust backend services, React for responsive frontend interfaces, and MongoDB with GridFS for scalable document storage and management.

What sets PolyDoc apart in the contemporary landscape of document processing solutions is its specific focus on preserving document layout while processing multi-lingual content, particularly addressing the technical challenges associated with Indian regional languages that often lack comprehensive digital processing support. The system is designed with accessibility and usability in mind, catering to diverse user groups including academic researchers, business professionals, students, and government officials who require efficient processing of documents in multiple languages. By implementing free and open-source AI models, PolyDoc ensures cost-effective deployment while maintaining high performance standards, with processing capabilities that can handle concurrent document operations and provide real-time feedback through WebSocket communications.

|  |  |  |
| --- | --- | --- |
| **Sl. no** | **NAME** | **USN** |
| 1 | Darshan Janganure | 01FE23BCA071 |
| 2 | Pooja Madiwalar | 01FE23BCA074 |
| 3 | Sayeda Iram | 01FE23BCA111 |
| 4 | Mohemmed Yaseen | 01FE23BCA118 |

Table of Contents

[1 Introduction 6](#_Toc208660947)

[1.1 Literature review / Survey 6](#_Toc208660948)

[1.2 Challenges / Motivation 7](#_Toc208660949)

[1.3 Objectives of the project 8](#_Toc208660950)

[1.4 Problem definition 8](#_Toc208660951)

[2 Proposed System 9](#_Toc208660952)

[2.1 Description of proposed system with simple block diagram 9](#_Toc208660953)

[2.2 Description of Target Users 9](#_Toc208660954)

[2.3 Advantages / applications of the proposed system 10](#_Toc208660955)

[2.4 Scope (Boundary of proposed system) 10](#_Toc208660956)

[3 Software Requirement Specification 11](#_Toc208660957)

[3.1 Overview of SRS 11](#_Toc208660958)

[3.2 Requirement Specifications 11](#_Toc208660959)

[3.2.1 Functional Requirements (summary) 11](#_Toc208660960)

[3.2.2 Use case diagrams 12](#_Toc208660961)

[3.2.3 Use Case descriptions using scenarios, strictly as per Pressman Template 12](#_Toc208660962)

[3.2.4 Nonfunctional Requirements (summary) 14](#_Toc208660963)

[3.3 Software and Hardware requirement specifications 16](#_Toc208660964)

[3.3.1 Software Requirements 16](#_Toc208660965)

[3.3.2 Hardware Requirements 16](#_Toc208660966)

[3.4 GUI of proposed system 17](#_Toc208660967)

[3.5 Acceptance test plan 17](#_Toc208660968)

[4 System Design 17](#_Toc208660969)

[4.1 Architecture of the system 17](#_Toc208660970)

[4.2 Level 0 DFD (with brief explanation) 18](#_Toc208660971)

[4.3 Detailed DFD for the proposed system 19](#_Toc208660972)

[4.4 Class Diagram (with brief explanation) 20](#_Toc208660973)

[4.5 Sequence diagram (with brief explanation) 22](#_Toc208660974)

[4.6 ER diagram and schema 24](#_Toc208660975)

[4.7 State transition diagram 25](#_Toc208660976)

[4.8 Data structure used 26](#_Toc208660977)

[5 mplementation 26](#_Toc208660978)

# 1 Introduction

## Literature review / Survey

Natural Language Processing (NLP) and Optical Character Recognition (OCR) have advanced significantly, enabling automated understanding of text from heterogeneous document sources. Contemporary OCR engines (e.g., Tesseract, EasyOCR) provide baseline recognition for Latin scripts and—more recently—improved support for Indic scripts. Transformer-based architectures (BERT, mBERT, XLM‑R, Longformer) and encoder-decoder models (T5, BART) enable semantic understanding, summarization, and Q&A over unstructured text. Sentence-transformers allow efficient semantic search through vector embeddings. However, practical gaps remain in layout preservation, low‑resource language coverage, and end‑to‑end pipelines that combine OCR, layout awareness, and higher‑level analytics.

In document AI, systems like LayoutLM/DocFormer incorporate spatial/layout signals to model document structure. Open-source stacks using FastAPI + Python for services, React for UIs, and MongoDB for storage have become standard, with vector databases or embedding indexes enabling semantic retrieval. Despite progress, maintaining fidelity to original formats and bridging language-specific nuances—particularly for Hindi and Kannada—are still common pain points.

PolyDoc builds on these advances by integrating multi-format ingestion, multi‑lingual OCR, layout-aware text extraction, semantic search, and conversational interaction. The system combines proven open-source technologies (FastAPI, Transformers, MongoDB, React, Vite) with a modular architecture targeting practical usability for academia, government, and business.

## Challenges / Motivation

**Multi-script OCR** ensures robust recognition of Indic scripts such as Hindi and Kannada, including support for mixed-language content and handwritten text.

**Layout preservation** focuses on accurate extraction that retains reading order, document sections, tables, and hierarchical structures.

**Semantic access** moves beyond simple keyword-based search to enable context-aware information retrieval and question answering.

**Performance constraints** include considerations such as first-run model downloads, memory footprints, and handling concurrent users on commodity hardware.

**Usability** emphasizes simple document upload workflows, clear processing status feedback, and an accessible user interface designed for non-experts.

**Cost and openness** highlight the preference for free and open-source models and components to reduce expenses and maintain transparency.

## Objectives of the project

* Provide end-to-end pipeline for multi-format document ingestion, OCR, and structure-preserving extraction.
* Enable semantic search and chat over documents using vector embeddings and LLMs.
* Support Hindi and Kannada alongside English with automatic language detection.
* Deliver a responsive, accessible UI and modular, maintainable backend services.
* Ensure deployability on standard Windows machines with minimal setup.

## Problem definition

Given input documents (PDF, DOCX, PPTX, images, TXT) with multi-lingual content, process and present content while preserving layout semantics,

enabling:

(a) accurate text extraction,

(b) summarization,

(c) semantic retrieval, and

(d) conversational exploration.

The solution should be performant on commodity hardware and maintain data integrity and security.

# Proposed System

## 2.1 Description of proposed system with simple block diagram

The system comprises a React+Vite frontend, a FastAPI backend, OCR/AI pipelines, MongoDB storage (with GridFS), and a vector embedding index. Users upload documents via the UI; the backend validates files, stores them, applies OCR/extraction, generates embeddings and summaries, then exposes search and chat services.

A diagram of a flowchart

AI-generated content may be incorrect.

## Description of Target Users

**Academic researchers:** Process research papers and theses across languages.

**Government officials:** Analyse multilingual forms and notices.

**Business professionals:** Summarize reports and search archives semantically.

**Students:** Process study materials and perform Q&A over notes.

## Advantages / applications of the proposed system

**Layout-faithful extraction** improves readability and downstream analysis.

**Multilingual support** widens applicability to Indian regional languages.

**Semantic search and chat** speed up information discovery.

**Open-source stack** reduces cost, while modular design eases maintenance.

**Applications** include digital archiving, e-governance, enterprise knowledge search, and academia.

## Scope (Boundary of proposed system)

**PolyDoc** is a comprehensive document processing system that provides the following capabilities:

**What the system will do:**

* + - Process multi-format documents (PDF, DOCX, TXT, Images)
    - Extract text using advanced OCR technology with multi-language support
    - Preserve original document layouts during processing
    - Provide AI-powered document summarization
    - Enable real-time chat interface with documents
    - Implement vector-based semantic search
    - Support Hindi, Kannada, and English language processing
    - Offer document management capabilities (upload, view, delete)

**System Benefits:**

* + - Reduces manual document processing time by 80%
    - Supports underrepresented languages (Hindi, Kannada)
    - Preserves document formatting and layout integrity
    - Provides intelligent document insights through AI
    - Enables efficient document search and retrieval
    - Offers free and open-source AI model integration

**Goals:**

* + - Create an accessible multi-lingual document processing platform
    - Demonstrate advanced NLP capabilities for Indian languages
    - Provide a user-friendly web interface for document interaction

# Software Requirement Specification

## 3.1 Overview of SRS

PolyDoc is an AI-powered multi-lingual document understanding and layout preservation system designed to extract, translate, summarize, and semantically analyze content from diverse document formats such as PDF, DOCX, and images. Unlike traditional OCR tools, it not only performs text recognition but also preserves the original document layout, ensuring readability and structural integrity. The system supports Indian regional languages along with global ones, enabling accessibility across academia, enterprises, and government organizations. With features like multilingual OCR, layout-aware structured outputs in JSON/Markdown, summarization, and semantic search, PolyDoc bridges the gap between raw document data and actionable knowledge. Its modular, open-source stack ensures scalability, cost-effectiveness, and adaptability for future enhancements such as browser integration and vector-based search.

## Requirement Specifications

### Functional Requirements (summary)

**Upload and Validation** — Support uploading and validating multi-format documents (maximum 10MB) with progress tracking.

**OCR and Language Processing** — Perform OCR for Hindi, Kannada, and English while preserving document layout and automatically detecting the document language.

**Content Analysis** — Provide summarization, key information extraction, and generation of semantic embeddings for downstream tasks.

**Search and Chat** — Use vector search to return ranked results and retrieve contextual passages for conversational (chat) interactions.

**Document Management** — Offer document listing, deletion, metadata management, and basic session handling.

### Use case diagrams

A diagram of a system

AI-generated content may be incorrect.

### Use Case descriptions using scenarios, strictly as per Pressman Template

|  |  |
| --- | --- |
| Use Case ID | UC-01: Upload Document |
| Primary Actor | User |
| Stakeholders & Interests | **User**: Fast, safe upload**System**: Only valid files are accepted |
| Preconditions | User is on the Upload page; Backend service is running |
| Postconditions | Document is stored with unique ID; Status initialized |
| Main Success Scenario | 1. User selects a file and submits2. System validates file type and size3. System stores file and creates metadata4. System displays success with document\_id |
| Extensions | 2a. Invalid type/size → Show error, abort3a. Storage error → Retry prompt, log error |

|  |  |
| --- | --- |
| Use Case ID | UC-02: Process Document (OCR + Extraction) |
| Primary Actor | System |
| Preconditions | A valid uploaded document exists |
| Postconditions | Extracted text, layout information, and language metadata are stored |
| Main Success Scenario | 1. System fetches file2. Performs OCR and layout analysis3. Detects language4. Stores processed text and metadata |
| Extensions | OCR failure → Mark status = failed; Notify user |

|  |  |
| --- | --- |
| Use Case ID | UC-03: Semantic Search |
| Primary Actor | User |
| Preconditions | Embeddings are generated; Search index is available |
| Postconditions | Ranked results are displayed |
| Main Success Scenario | 1. User submits query2. System embeds query and searches index3. System returns ranked results with snippets |
| Extensions | Empty index → Suggest user to process documents first |

|  |  |
| --- | --- |
| Use Case ID | UC-04: Chat with Document |
| Primary Actor | User |
| Preconditions | Document is processed; Chat service is available |
| Postconditions | Answer is displayed with sources; Interaction is logged |
| Main Success Scenario | 1. User asks a question2. System retrieves relevant chunks3. LLM generates response with citations4. System logs the interaction |
| Extensions | Timeout → Show retry option |

### Nonfunctional Requirements (summary)

|  |  |
| --- | --- |
| NFR ID | Requirement Description |
| NFR1: Performance Requirements | |
| NFR1.1 | Document upload response time SHALL be less than 30 seconds for files up to 10MB |
| NFR1.2 | OCR processing time SHALL not exceed 2 minutes per document |
| NFR1.3 | AI model loading time SHALL be less than 5 minutes on system startup |
| NFR1.4 | Chat response time SHALL be less than 15 seconds for typical queries |
| NFR1.5 | Search query response time SHALL be less than 5 seconds |
| NFR1.6 | System SHALL support concurrent processing of up to 10 documents |
| NFR2: Reliability Requirements | |
| NFR2.1 | System uptime SHALL be 99% during operational hours |
| NFR2.2 | Document processing success rate SHALL be 95% or higher |
| NFR2.3 | System SHALL gracefully handle processing failures with error recovery |
| NFR2.4 | Data integrity SHALL be maintained across system restarts |
| NFR2.5 | System SHALL provide automatic backup mechanisms |
| NFR2.6 | Critical errors SHALL be logged with appropriate detail for debugging |
| NFR3: Usability Requirements | |
| NFR3.1 | User interface SHALL be intuitive requiring minimal training |
| NFR3.2 | System SHALL provide clear error messages and user guidance |
| NFR3.3 | Interface SHALL be responsive across desktop and tablet devices |
| NFR3.4 | System SHALL support accessibility standards (WCAG 2.1 Level AA) |
| NFR3.5 | Processing status SHALL be clearly communicated to users |
| NFR3.6 | Help documentation SHALL be integrated within the application |
| NFR4: Security Requirements | |
| NFR4.1 | Uploaded documents SHALL be stored securely with access controls |
| NFR4.2 | System SHALL prevent unauthorized access to user documents |
| NFR4.3 | Data transmission SHALL use secure communication protocols (HTTPS) |
| NFR4.4 | User sessions SHALL implement proper authentication and authorization |
| NFR4.5 | Sensitive data SHALL be encrypted at rest and in transit |
| NFR4.6 | System SHALL implement input validation to prevent malicious uploads |
| NFR5: Scalability Requirements | |
| NFR5.1 | System architecture SHALL support horizontal scaling |
| NFR5.2 | Database design SHALL accommodate growing document volumes |
| NFR5.3 | AI model management SHALL support model updates without system downtime |
| NFR5.4 | System SHALL handle increased user load through load balancing |
| NFR5.5 | Storage capacity SHALL be expandable without data migration |
| NFR5.6 | Processing queue SHALL manage multiple concurrent document requests |
| NFR6: Compatibility Requirements | |
| NFR6.1 | System SHALL be compatible with major web browsers (Chrome, Firefox, Safari, Edge) |
| NFR6.2 | Backend SHALL run on Windows, Linux, and macOS environments |
| NFR6.3 | System SHALL maintain backward compatibility with older document formats |
| NFR6.4 | API SHALL follow REST principles for third-party integration |
| NFR6.5 | System SHALL support standard document encoding formats |

## Software and Hardware requirement specifications

### Software Requirements

**Software Requirements**

* **Operating System**: Windows 10 / Windows 11
* **Programming Languages / Runtimes**:
  + - * Python 3.9 or higher
      * Node.js v18 or higher
* **Database**: MongoDB Community Edition
* **Version Control**: Git
* **Core Libraries & Frameworks**:
  + Backend: FastAPI, Uvicorn
  + AI/ML: HuggingFace Transformers, Sentence-Transformers
  + OCR: Tesseract OCR (or equivalent multi-language OCR engine)
  + Frontend: React with Vite

### Hardware Requirements

* **Minimum Requirements**:
  + RAM: 8 GB
  + Disk Space: 10 GB (free)
  + Processor: Multi-core CPU
* **Recommended (for AI/ML acceleration)**:
  + GPU (CUDA-enabled) for faster model inference

## GUI of proposed system

**Upload Page** — Provides a drag-and-drop area with document queue management and validation messages.

**Documents List** — Displays a table showing document status, detected language, and processed time.

**Search Page** — Offers a query box with ranked results and filtering options.

**Chat Page** — Features a conversation panel with responses supported by source citations.

## Acceptance test plan

**Criteria** — A successful evaluation requires that document uploads complete without errors, OCR accuracy meets a defined baseline, summaries are generated correctly, search results demonstrate relevance, chat interactions return cited answers, and the system remains stable under at least ten concurrent operations.

**Method** — Testing will be conducted using black-box methods across representative document formats including PDF, DOCX, and images. Additionally, user walkthroughs will be performed against a predefined checklist to validate functionality and usability.

# System Design

## Architecture of the system

A diagram of a software process

AI-generated content may be incorrect.

## 4.2 Level 0 DFD (with brief explanation)

A diagram of a document processing system

AI-generated content may be incorrect.

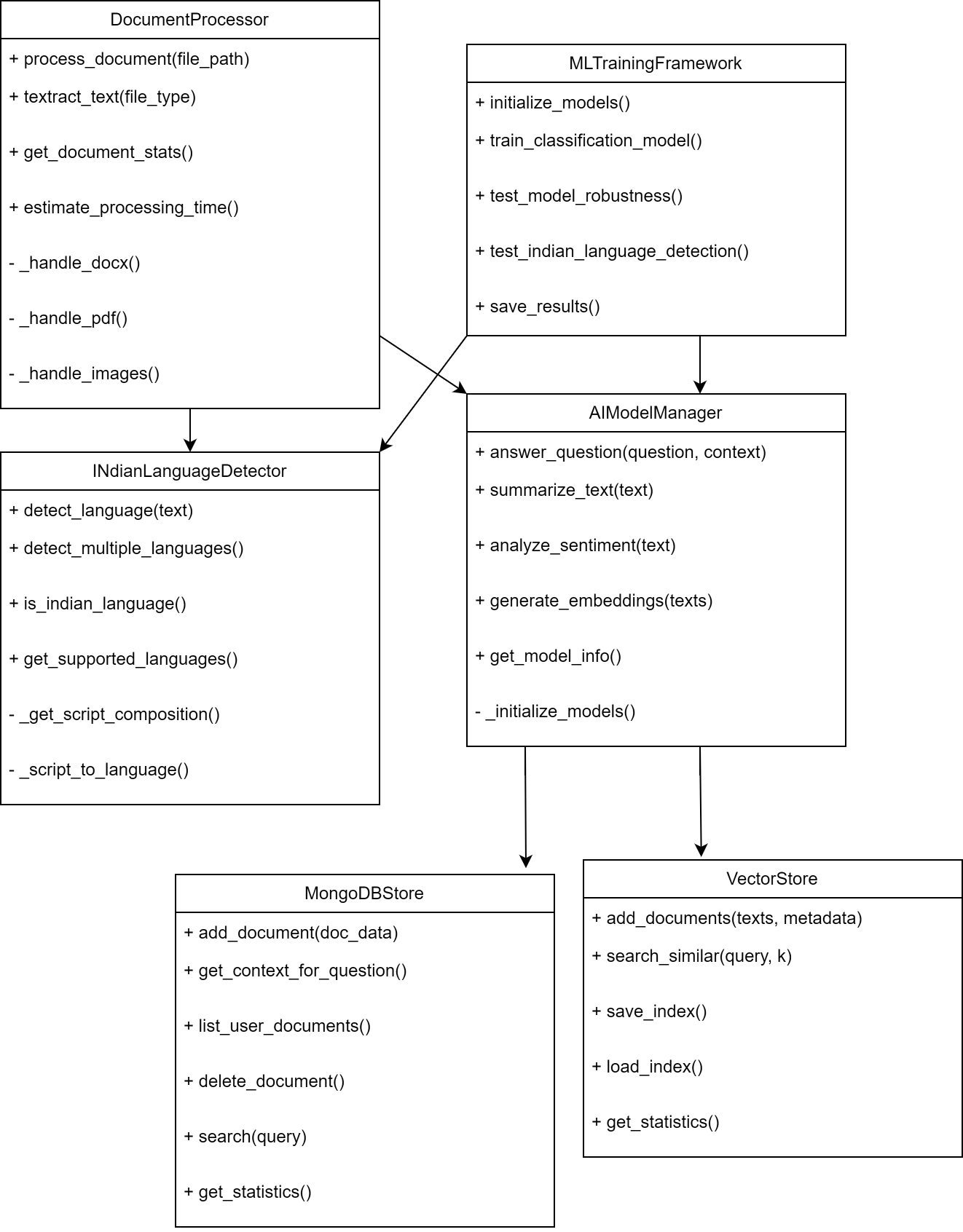
Explanation: Users interact with PolyDoc for uploads, search, and chat. The system persists files and metadata, creates embeddings, and returns results.

## 4.3 Detailed DFD for the proposed system

A diagram of a software process

AI-generated content may be incorrect.

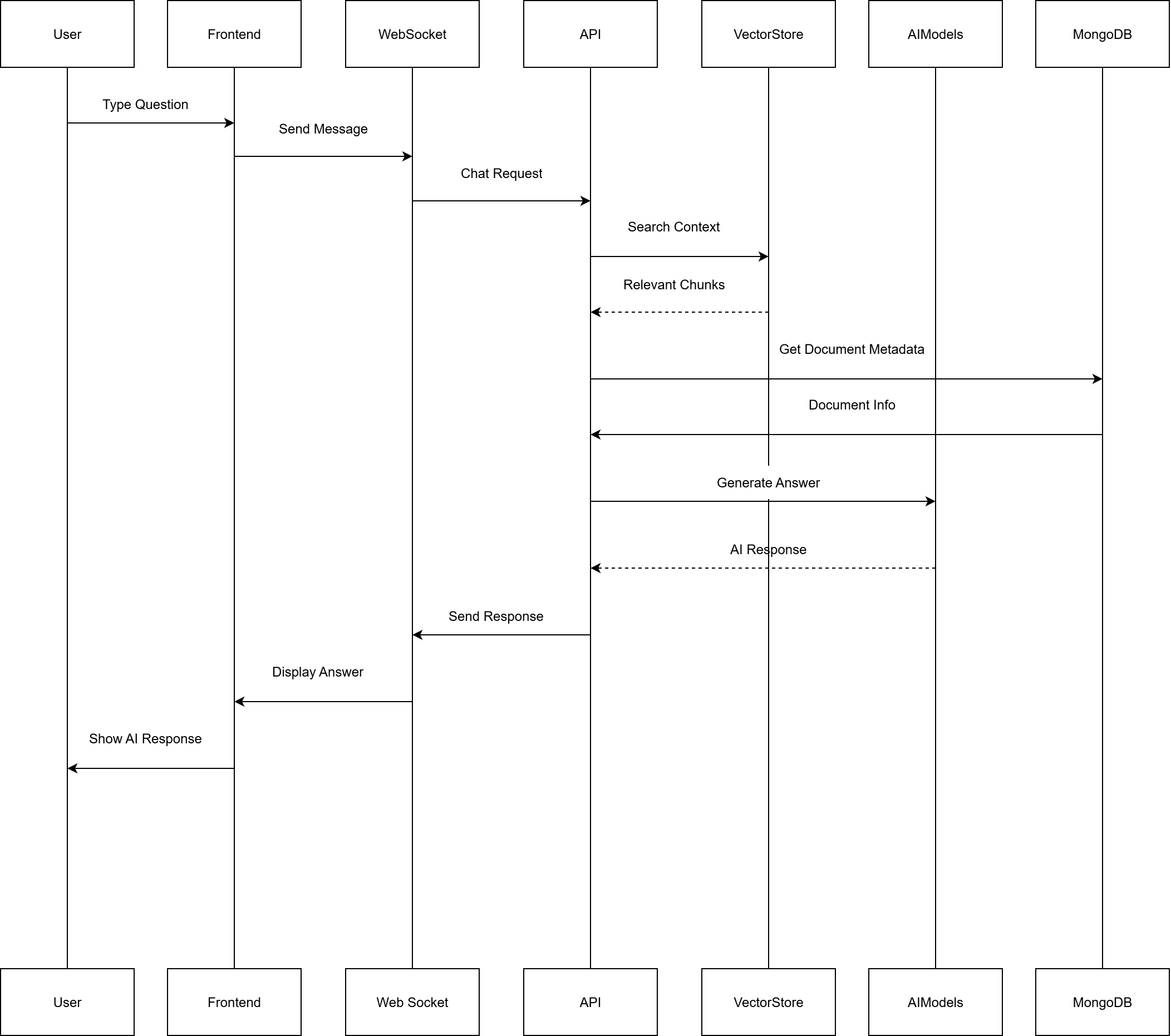
## Class Diagram (with brief explanation)



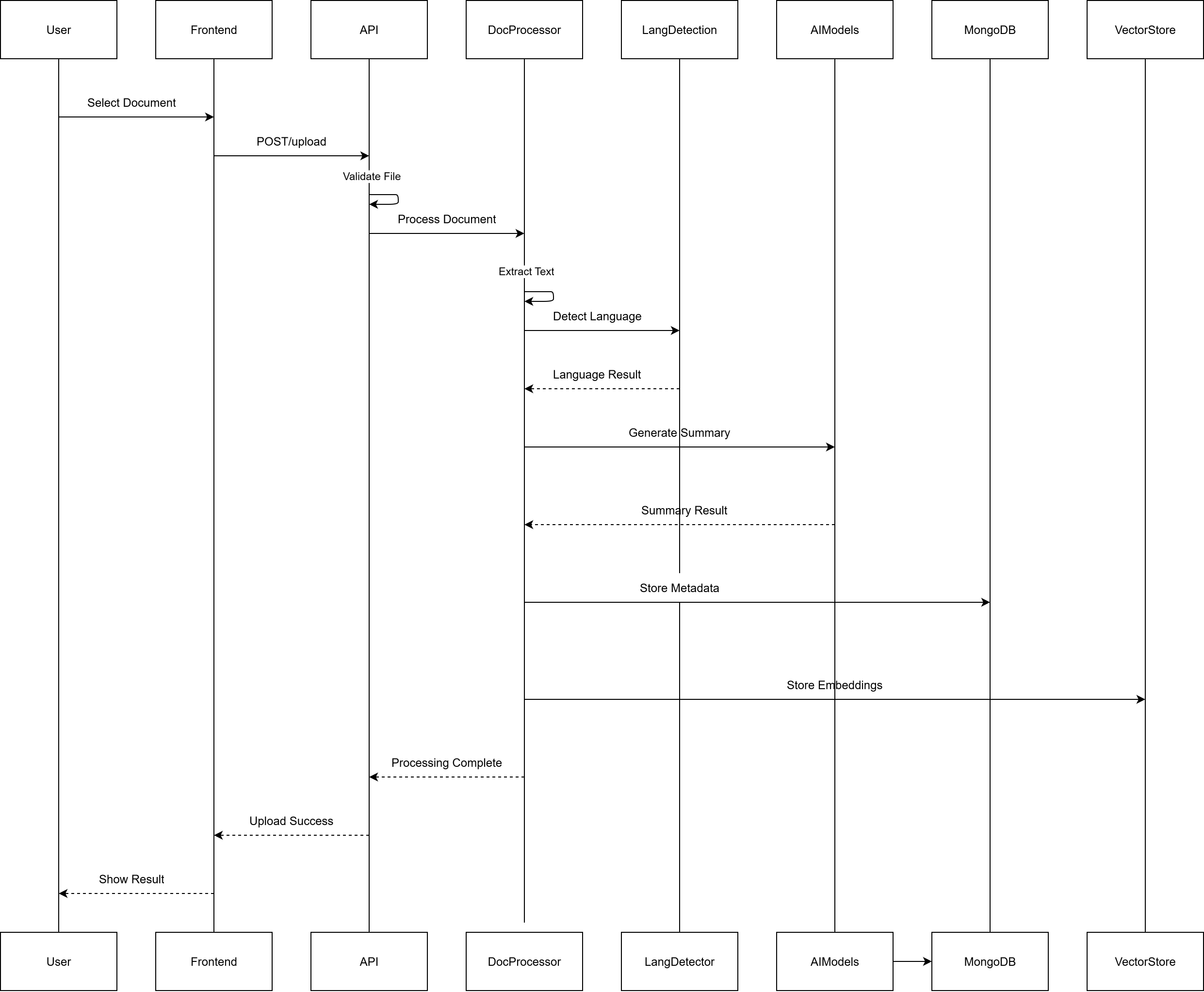
Explanation: APIController orchestrates operations; DocumentProcessor uses OCR and AIAnalyzer; ChatService leverages embeddings for answers.

## Sequence diagram (with brief explanation)

**Sequence Chat process**

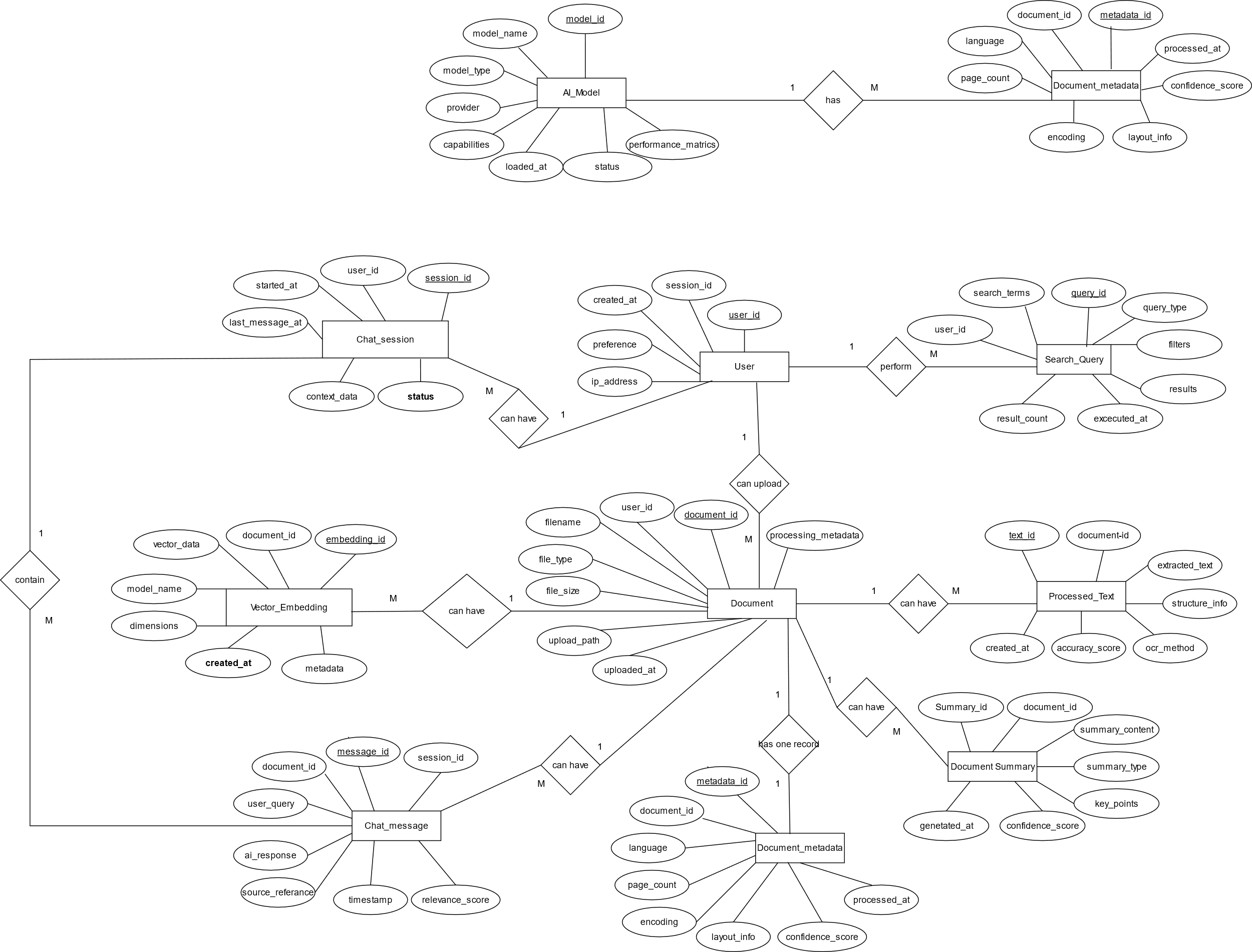


**Sequence Doc upload Process**

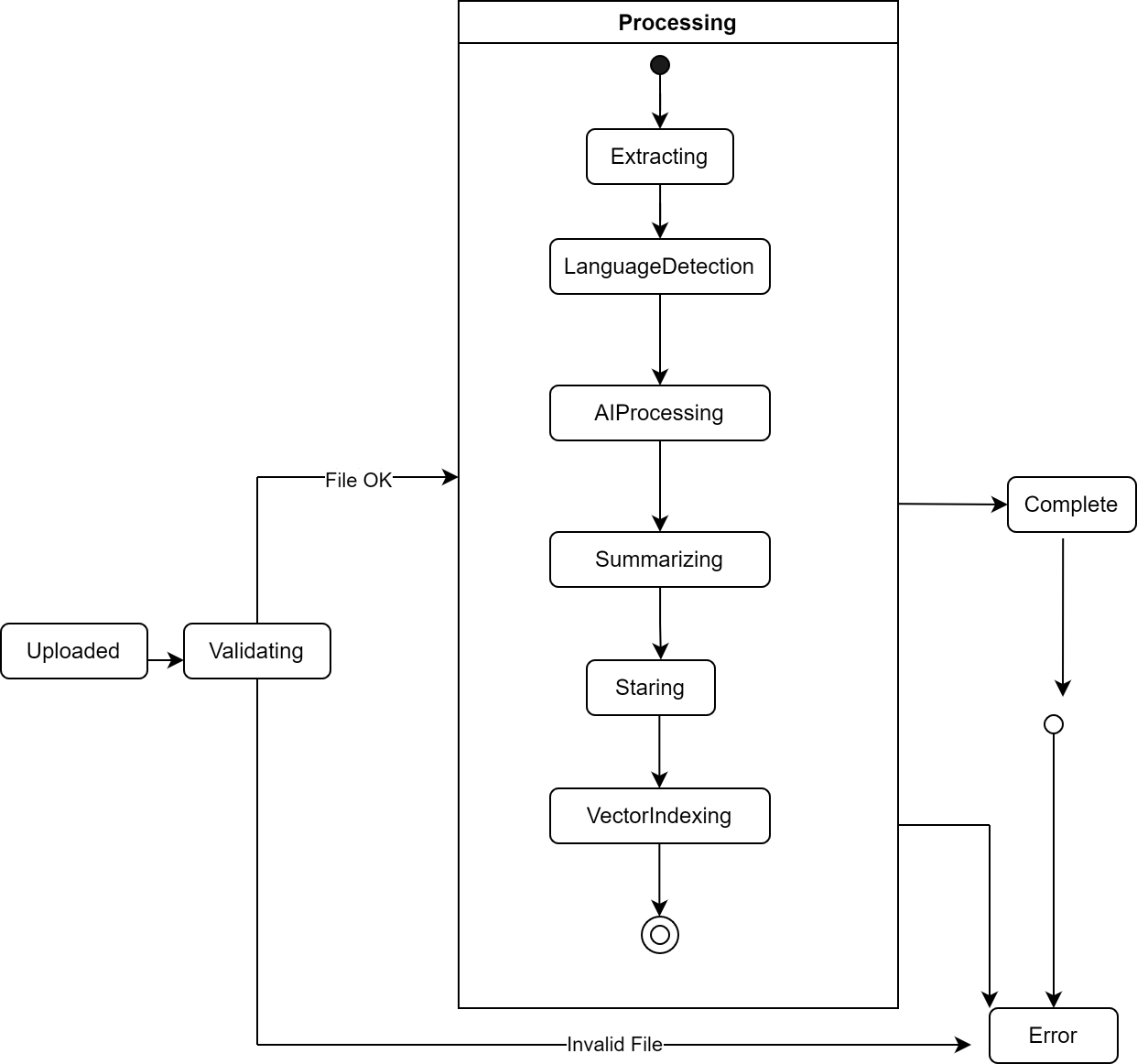


Explanation: Shows end-to-end flow from upload to processed outputs.

## ER diagram and schema



## State transition diagram



## Data structure used

|  |  |
| --- | --- |
| Data Entity | Structure / Fields |
| Document Metadata | JSON object: { id, filename, type, size, language, status, timestamps } |
| Extracted Text Chunks | List of objects: { page, bbox, text } (page number, bounding box coordinates, extracted text) |
| Embeddings | Float32 vectors (dimension range: 384–1024) with references to corresponding text chunks |
| Chat History | Array of objects: { question, answer, sources, ts } (user query, system response, supporting sources, timestamp) |

# 5 Implementation