Sharif College of Engineering & Technology

Computer Science Department

Day Month Year

DATE | | | – | | | – | | | | |

PROJECT TITLE	Intelligent Document Retrieval and Question Answering System using On-Premise AI and Retrieval-Augmented Generation
KEY WORDS	Document AI, Question Answering, Retrieval-Augmented Generation, OCR, NLP, Information Retrieval, On-Premise AI,x` Document Management, Vector Embeddings, Secure Document Processing
DOMAIN OF THE PROJECT	Artificial Intelligence, Natural Language Processing, Information Retrieval Systems
SUPERVISOR	Mam Hirra Mustafa
CO-SUPERVISOR	Dr Mazhar Iqbal

STUDENT INFORMATION			
Sr.	**Student ID**	**Name**	**Email**
1.	UET-SHCET-LHR-CS-02	AMNA IKRAM	Ikramamna980@gmail.com
2.	UET-SHCET-LHR-CS-12	ZAIN UL HASSAN	Zainulhassan4330@gmail.com
3.	UET-SHCET-LHR-CS-13	EZZA ANSAR	Ezzaansar05@gmail.com

TABLE OF CONTENTS

1. PROBLEM STATEMENT

Organizations struggle to efficiently search, extract, and retrieve information from large collections of digital and scanned documents. Manual review is slow, error-prone, and unsuitable for workflows such as HR resume screening where timely decisions are critical. Cloud-based tools exist, but they require sending sensitive data outside the organization, creating privacy and compliance risks. There is a need for a secure, on-premise AI system that can automate document understanding, enable intelligent search and question answering, and support automated resume screening without exposing confidential information.

2. EXECUTIVE SUMMARY

This project delivers a secure, on-premise AI system designed to process, understand, and retrieve information from large volumes of documents, including scanned files and unstructured text. The solution combines OCR, natural language processing, vector embeddings, and retrieval augmented generation to provide accurate search, classification, and question answering across diverse document types.

A key feature of the system is the Automated HR Resume Screening Module. This module connects with Gmail to collect incoming resumes, extracts essential details such as skills, experience, qualifications, and contact information, and evaluates candidates based on predefined criteria. It automatically shortlists suitable applicants and reduces the manual workload traditionally associated with resume filtering. This ensures faster hiring decisions while keeping all applicant data protected within the organization.

The platform integrates advanced document understanding with an efficient workflow engine, allowing users to upload documents, run automated processing pipelines, and retrieve insights through a unified interface. All processing occurs within the organization's environment, which supports strict data privacy, regulatory compliance, and customizable access controls.

The system is built with a scalable architecture that supports modular expansion and integration with enterprise tools. Its design allows organizations to incorporate additional AI driven components, automate department specific processes, and create domain tailored search models. This flexibility ensures that the solution can evolve with changing business needs and support long term digital transformation goals.

By automating time-consuming document review tasks and improving information retrieval accuracy, the system helps organizations enhance operational efficiency, reduce human error, and accelerate decision making. The project provides a robust foundation for AI powered document intelligence and can be extended to support legal review, compliance monitoring, and enterprise knowledge management.

3. INTRODUCTION

Modern organizations handle large volumes of documents in digital and scanned formats, but traditional manual review processes are slow, inconsistent, and costly. As the scale of information grows, the ability to quickly extract knowledge, conduct accurate searches, and automate routine tasks has become essential for effective operations. The problem is especially significant for workflows such as HR screening, compliance

checks, financial auditing, and legal review, where timely access to reliable information directly affects decision making and organizational efficiency.

Advances in artificial intelligence have created new opportunities for automating document understanding. Technologies such as optical character recognition, natural language processing, vector embeddings, and deep learning-based retrieval models now enable systems to interpret both structured and unstructured documents with high accuracy. These capabilities support intelligent search, automated classification, and context aware question answering, which are far beyond the limitations of traditional keyword-based tools.

Several commercial platforms have attempted to address similar needs. Enterprise content management systems offer basic indexing and search functions. Cloud based AI services provide automated document processing, but they require sending sensitive files to external servers, which is unsuitable for organizations that must maintain strict data privacy. HR tools offer applicant tracking features, but most rely on manual resume screening or third-party cloud processing. Research groups have also explored machine learning approaches for text extraction, document clustering, and resume ranking, but practical on-premise solutions that align with enterprise security needs remain limited.

This project builds upon these developments by providing a secure, on-premise AI system that automates document processing and enables intelligent retrieval across large collections of files. It also introduces an Automated HR Resume Screening Module that uses advanced extraction and evaluation mechanisms to streamline the hiring process. By combining robust document understanding with enterprise grade privacy, the system addresses a critical gap in current solutions and provides a foundation for scalable, AI driven document intelligence within the organization.

4. COMPETITORS/COMPETITIVE ANALYSIS

The market for document intelligence and automated information retrieval is growing quickly, and several commercial platforms offer related capabilities. Although no single tool provides the full combination of on-premise document processing, intelligent search, retrieval augmented generation, and automated HR resume screening in one integrated system, there are categories of competitors that overlap with parts of this solution.

4.1 Competitors

- **Cloud-based Document AI Platforms:**
  Large technology vendors such as Google Document AI, Microsoft Azure Form Recognizer, and Amazon Textract provide powerful OCR and text extraction services. These platforms support classification, structure extraction, and searchable document pipelines. Their strength lies in scalability and accuracy, but they require organizations to upload sensitive documents to external cloud environments. This makes them unsuitable for institutions that must retain complete data control.
- **Enterprise Content Management and Search Systems:**
  Solutions like SharePoint, Elastic Enterprise Search, OpenText, and IBM FileNet offer indexing, keyword search, and document organization features. They are widely used in

enterprises but rely mostly on rule-based methods and keyword matching. They do not provide advanced semantic search, embeddings, or context-aware question answering. Their automation capabilities are also limited when compared with modern AI driven workflows.

- **HR Applicant Tracking Systems:**
  Platforms such as BambooHR, Workday, Lever, Greenhouse, and Zoho Recruit include resume management and applicant tracking features. While these tools streamline hiring workflows, most of them rely on basic keyword-based resume parsing. They do not offer advanced extraction, semantic understanding, or AI powered shortlisting based on detailed job criteria. Many of these systems are also cloud hosted, which raises data privacy concerns.
- **Specialized OCR Tools:**
  Products like Adobe Acrobat OCR, ABBYY FineReader, and Tesseract based tools focus primarily on text extraction from scanned files. They perform well for OCR tasks but do not support deep document understanding, embeddings, or enterprise-level workflow automation.
- **RAG and Vector Search Platforms:**
  Tools such as Pinecone, Weaviate, Elasticsearch Vector Search, and Vespa offer vector databases for semantic retrieval. These solutions are strong in retrieval and indexing but do not include a complete document processing workflow. They require organizations to build surrounding pipelines manually, such as OCR, metadata extraction, UI layers, and domain-specific modules.

4.2 Our Competitive Edge

The system introduced in this project combines document ingestion, OCR, semantic search, structured extraction, and HR resume automation into a unified on-premise platform. This integrated design is not commonly found in existing competitors, especially in solutions that maintain full data privacy inside an organization. The Automated HR Resume Screening Module provides an additional advantage by enabling automated candidate shortlisting based on real semantic understanding rather than keyword matching.

By delivering advanced AI capabilities without depending on external cloud services, the system fills a strong market gap and offers a unique competitive position for organizations that require privacy, customization, and end-to-end document intelligence.

5. OBJECTIVES

The primary objective of this project is to develop a secure, on-premise AI system capable of processing and understanding large collections of documents in both digital and scanned formats. The system aims to automate document extraction, enable intelligent search, and support accurate question answering across diverse domains.

Specific objectives include the following:

1. Develop an automated pipeline that performs OCR, text extraction, metadata generation, and document structuring with high accuracy.

2. Implement a vector based semantic search engine that allows users to retrieve relevant information using natural-language queries.
3. Integrate retrieval augmented generation to deliver context aware answers and document summaries.
4. Create an Automated HR Resume Screening Module that collects resumes from Gmail, extracts candidate information, evaluates applicants using predefined criteria, and identifies suitable candidates automatically.
5. Provide a unified, user friendly interface for uploading documents, searching content, reviewing results, and managing workflows.
6. Ensure that all processing is performed within the organizational environment to maintain confidentiality and compliance with data protection requirements.
7. Build a scalable and modular architecture that can be extended to additional domains such as legal review, finance, compliance, and enterprise knowledge management.

These objectives collectively ensure that the system delivers measurable improvements in efficiency, accuracy, and decision support while maintaining strict data privacy.

6. MOTIVATION

The growing volume of unstructured and scanned documents in modern organizations has created a strong need for efficient and intelligent document processing. Manual document review is time consuming, inconsistent, and costly, which affects the quality and speed of decision making across departments such as HR, finance, compliance, and administration. Traditional search tools are unable to interpret context or provide meaningful insights, which limits their usefulness in large document repositories.

Advances in artificial intelligence now make it possible to extract knowledge, analyze content, and automate routine tasks at a scale that was previously unattainable. However, most advanced AI tools operate in the cloud, which introduces data privacy concerns for organizations that handle confidential information. This gap highlights the importance of creating a secure, on-premise system that offers powerful AI capabilities without compromising data control.

The need is especially strong in HR workflows where resume screening consumes significant time and is prone to human bias and error. An automated resume screening module can accelerate hiring decisions, improve consistency, and reduce workload while keeping applicant information protected. Addressing this problem has the potential to deliver organizational efficiency, cost savings, and improved accuracy, which makes the project both relevant and impactful.

7. FEATURES OF PROJECT

The proposed Document Retrieval and Question Answering System offers multiple features designed to automate document processing, enable intelligent information retrieval, and maintain complete data privacy. Each feature is described in detail below:

1. **Document Upload and Processing**
   Users can upload documents in various formats such as PDF, Word, or scanned images. Each document is automatically processed through classification, metadata extraction, and Optical Character Recognition (OCR) to convert images into searchable text. The system ensures that all documents are digitized, structured, and ready for further analysis.
2. **Document Classification and Metadata Generation**
   The system categorizes documents into relevant types, such as resumes, reports, contracts, or letters. Metadata including document type, author, date, and other relevant attributes is generated automatically. This structured information allows faster indexing and retrieval, improving workflow efficiency.
3. **OCR and Content Extraction**
   Using OCR technology, the system extracts textual content from scanned or image-based documents. Extracted content is cleaned, structured, and stored in a searchable format, enabling semantic search and question answering.
4. **Vector Embedding and Semantic Search**
   All processed text is converted into vector embeddings and stored in a vector database. Semantic search allows retrieval of documents and text segments based on meaning rather than exact keywords. Users can perform searches using natural language queries for more accurate results.
5. **Question Answering using Retrieval-Augmented Generation (RAG)**
   The system leverages a RAG framework to answer user queries contextually. It retrieves relevant content from the document collection and generates concise, accurate, and context-aware responses. Users can ask questions like “Extract skills from a resume” or “Summarize the results of student XYZ” and receive immediate answers.
6. **Document Retrieval**
   Users can directly request specific documents or related files. The system identifies the most relevant matches based on semantic similarity and metadata, supporting efficient document discovery and decision making.
7. **Automated HR Resume Screening Module**
   The system includes a dedicated HR module that integrates with Gmail to fetch incoming resumes automatically. It extracts key information such as skills, experience, qualifications, and contact details. Based on predefined job criteria, the module shortlists suitable candidates, significantly reducing manual HR effort, speeding up the recruitment process, and maintaining full confidentiality of applicant data.
8. **API Integrations and Automation**
   A well-defined API layer enables integration with external applications and workflows. For example, resumes received via email can be automatically uploaded and processed, and results can be shared with HR management systems or reporting tools.
9. **Privacy and On-Premise Deployment**
   All processing, storage, and AI operations are performed on local infrastructure, ensuring that sensitive information never leaves the organization. This approach satisfies privacy, security, and compliance requirements.
10. **User-Friendly Interface**
    The system provides an intuitive web interface that allows users to upload documents, perform

searches, ask questions, review results, and manage workflows without requiring technical expertise.

11. **Scalability and Extensibility**
    The modular architecture allows the system to scale easily with additional storage, computing resources, or AI models. Future enhancements such as document redaction, advanced summarization, or integration with enterprise knowledge management systems can be added seamlessly.

8. ARCHITECTURAL DESIGN

The proposed system follows a modular and secure architecture that enables efficient document processing, intelligent search, and privacy-preserving question answering. It integrates several software, hardware, and network components to ensure smooth operation and scalability.

8.1 Hardware Components

- **Local Server or Cloud Instance:** Hosts the main system, including backend services, databases, and AI models.
- **GPU/TPU (Optional):** Accelerates model inference and embedding generation.
- **Client Devices:** Allow users to interact with the system through a web application or integrated APIs.

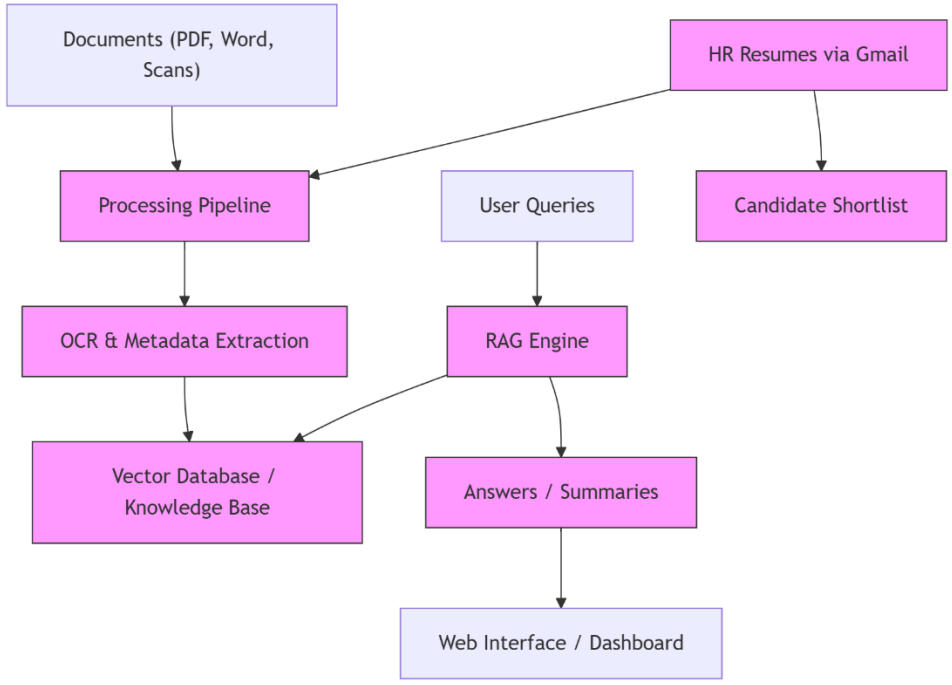
8.2 Software Components

- **Frontend Application:** Provides a user-friendly interface for document upload, searching, and query interactions.
- **Backend Services:** Handle request routing, document workflows, model execution, and API communication.
- **Databases:** Store metadata, extracted text, and vector embeddings to enable efficient retrieval and search.
- **AI and NLP Modules:** Perform document classification, OCR or text extraction, embedding generation, and Retrieval-Augmented Generation (RAG) for question answering.
- **Integration Layer:** Allows seamless connection with other systems, such as email or document management tools, for automated document ingestion.

8.3 Network Components

All system components communicate within a secure, controlled network. Role-based access control and authentication mechanisms ensure data protection and restricted access. The entire setup can be deployed on-premises or in a private cloud environment, depending on organizational requirements.

8.4 System Architecture



9. IMPLEMENTATION TOOLS AND TECHNIQUES

9.1 Implementation Methodology

The development will follow an **incremental and modular approach**, ensuring that each system component is independently developed, tested, and integrated. The key stages include:

- **Document Processing Pipeline:**
  Building modules for document upload, classification, text extraction, and metadata generation using OCR and NLP techniques.
- **Knowledge Base Construction:**
  Generating vector embeddings from extracted text and storing them in a local vector storage system for semantic search and retrieval.
- **Question Answering Engine:**
  Implementing a Retrieval-Augmented Generation (RAG) pipeline that retrieves relevant context from stored documents and generates natural language responses.
- **API and Integration Layer:**
  Developing secure and well-defined APIs to connect the system with external tools, such as email workflows or document management systems.
- **Frontend Interface:**
  Designing an intuitive web-based interface for document uploads, searches, and interactive question answering.
- **Testing and Optimization:**
  Conducting functional and performance testing to ensure accuracy, efficiency, and scalability across document types.

9.2 Implementation Tools

To maintain flexibility, the project will use open-source or self-hosted technologies that can be replaced or upgraded easily. The tentative tools and frameworks include:

- **Programming Language:** Python (for backend logic, AI workflows, and API development)
- **Web Framework:** A lightweight backend framework such as FastAPI or Flask for API management
- **Frontend:** A modern JavaScript framework (React or Next.js) for the user interface
- **Databases:** Relational database for metadata and a vector database for embeddings and semantic retrieval
- **AI & NLP Components:** OCR library for text extraction, embedding models for vectorization, and a language model for the RAG pipeline
- **Integration & Automation:** Workflow tools like n8n or direct API endpoints for email or third-party system integration
- **Deployment:** On-premise or private cloud setup using containerization (e.g., Docker) to ensure scalability and security

9.3 Development Approach

- **Version Control:** The entire project will be managed using Git for version tracking and collaboration.
- **Testing Framework:** Automated testing will ensure each module performs as expected before integration.
- **Documentation:** All APIs, workflows, and modules will be documented to facilitate maintenance and future upgrades.

10.PROJECT PLAN

10.1. Division of Responsibilities

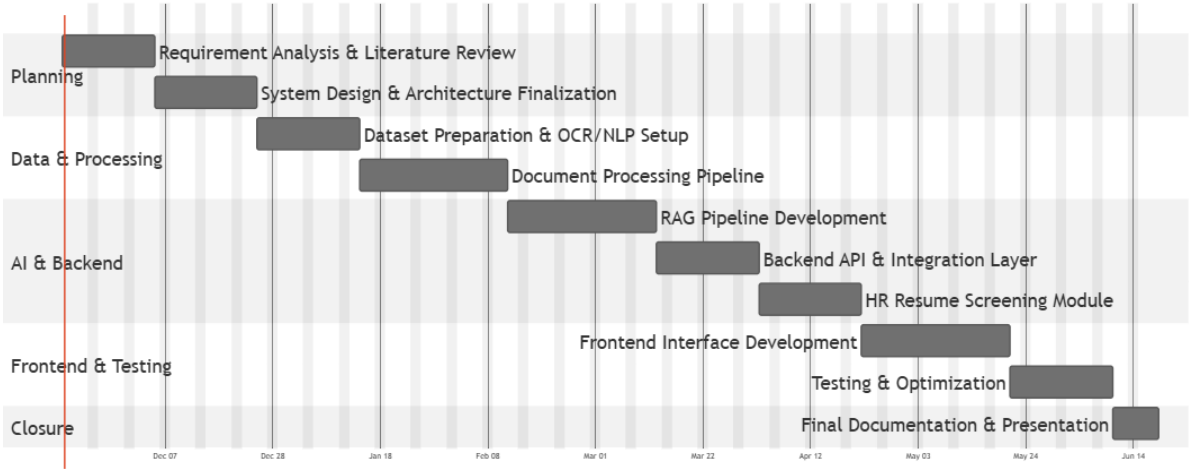
Team Member	Responsibilities
Zain Ul Hassan	**Overall project lead** responsible for system architecture, RAG pipeline development, model integration, and backend coordination. Manages codebase, ensures feature alignment, and handles final deployment and optimization.
Ezza Ansar	Focuses on document processing modules including OCR, document classification, and metadata extraction. Assists in generating and managing vector embeddings, contributes to retrieval optimization, and handles pre-processing workflows.
Amna Ikram	Designs and implements the frontend interface, integrates it with backend APIs, and ensures smooth user experience. Works on external API integration (e.g., email-to-upload workflows), handles testing.

10.2. Project Timeline

The total duration of the project is approximately **20 weeks**, starting from **November 1, 2025** and ending in **March 2026**.

10.2.1 Key Milestones

1. Requirement Analysis & Literature Review
2. System Design & Architecture Finalization
3. Dataset Preparation & OCR/NLP Setup
4. RAG Pipeline Development
5. Backend API & Integration Layer
6. Frontend Interface Development
7. Testing & Optimization
8. Final Documentation & Presentation



10.3. Project Management Approach

- **Coordination:** Weekly meetings to track milestones, update progress, and assign new tasks.
- **Tools:** Trello or Notion for project tracking, GitHub for version control, and shared Google Drive for documentation.
- **Testing:** Iterative testing after each module is developed.
- **Supervision Reviews:** Internal milestone presentations to ensure consistent progress.
- **Documentation:** Maintained in parallel with implementation for accuracy and clarity.

11.REFERENCES

1. Google Cloud. (2024). *Document AI Overview*. Retrieved from https://cloud.google.com/document-ai
2. Amazon Web Services. (2024). *Amazon Textract – Extract Text and Data from Documents*. Retrieved from https://aws.amazon.com/textract
3. Microsoft Azure. (2024). *Form Recognizer Documentation*. Retrieved from https://azure.microsoft.com/en-us/products/form-recognizer
4. LangChain. (2024). *Building Applications with Retrieval-Augmented Generation (RAG)*. Retrieved from https://python.langchain.com
5. Hugging Face. (2024). *Document AI: Understanding Documents with Transformers*. Retrieved from https://huggingface.co/blog/document-ai
6. Hugging Face. (2024). *Transformers and Large Language Models Documentation*. Retrieved from https://huggingface.co/docs
7. PyMuPDF Documentation. (2024). *Working with PDFs in Python*. Retrieved from https://pymupdf.readthedocs.io
8. FAISS. (2024). *Facebook AI Similarity Search Library*. Retrieved from https://faiss.ai
9. Tesseract OCR. (2024). *Optical Character Recognition Engine*. Retrieved from https://github.com/tesseract-ocr/tesseract

Supervisor's Signature: ……

FYP Proposal Evaluation

(To be filled by students)

Write down the brief project topic and should not be confusing.	
PROJECT TITLE:	

STUDENT INFORMATION *Write down the detail of all group members in BLOCK LETTERS ONLY.*				
Sr.	*Student ID*	*Name*	*Email*	*Mobile*
1.				
2.				
3.				

FOR EVALUATOR'S USE ONLY			
Remarks:	☐ Accepted	☐ Accepted with Minor changes	☐ Rejected

Suggested Improvements (*if any*):

_____

_____

_____

_____

_____

_____

_____

_____

Evaluated by:

Name: _____

Signature: _____

Day Month Year

DATE | | | – | | | – | | | | |