

QueryMate AI: An Offline Intelligent Multilingual Document Question-Answering System

Miss. Pranjali Tanaji Jadhav

Ashokrao Mane Group Of institution, vathar

Department of Computer Science and Engineering

pranjali_jadhav43@gmail.com

Miss. Sudharani Maruti Bhise

Ashokrao Mane Group Of institution, vathar

Department of Computer Science and Engineering

sudharanibhise2@gmail.com

Mr. Dayanand Mahadaji Patil

Ashokrao Mane Group Of institution, vathar

Department of Computer Science and Engineering

dp6222097@gmail.com

Mr. Sainath Baburao Nandiwale

Ashokrao Mane Group Of institution, vathar

Department of Computer Science and Engineering

sainathnandiwale313@gmail.com

Miss. Sakshi Anil Mahapure

Ashokrao Mane Group Of institution, vathar

Department of Computer Science and Engineering

sakshimahapure1415@gmail.com

Miss. Anupama Deepak Patil

Ashokrao Mane Group Of institution, vathar

Department of Computer Science and Engineering

adp@amgoi.edu.in

Abstract: The rapid growth of digital documentation has increased the need for intelligent systems capable of retrieving relevant information efficiently from large files. Conventional PDF readers mainly depend on keyword matching, which often fails to provide context-aware answers. This paper presents QueryMate AI, an offline AI-driven document question-answering platform developed to simplify interaction with PDF and DOCX files using Natural Language Processing (NLP) and semantic search techniques.

The proposed system enables users to upload documents and ask questions in natural language through text or voice input. QueryMate AI employs transformer-based embeddings, vector similarity search, and contextual answer generation to retrieve meaningful responses directly from

uploaded documents. The architecture integrates Streamlit for the user interface, Sentence Transformers for embedding generation, FAISS for semantic indexing, and transformer-based NLP models for response generation.

Keywords: Natural Language Processing, Semantic Search, Document Question Answering, Offline AI, FAISS, Transformers, Multilingual System, Artificial Intelligence.

1. INTRODUCTION

Digital documents have become the primary medium for storing academic, professional, legal, and organizational information. Formats such as PDF and DOCX are extensively used because they preserve formatting consistency across multiple platforms. However, extracting precise information from lengthy documents remains a difficult and time-consuming process. Users often depend on manual reading or basic keyword-based search tools, which are limited in understanding the actual intent behind a query. Traditional search approaches only identify exact word occurrences and fail to analyze semantic meaning. As a result, users may receive irrelevant matches even when the document contains the required information. This limitation creates difficulties for students reviewing academic material, researchers analyzing publications, employees handling

corporate reports, and legal professionals examining contracts or policies.

QueryMate AI is proposed as a smart document assistance platform capable of processing uploaded files and answering user questions in natural language. The system combines semantic search, transformer embeddings, and contextual answer generation to deliver accurate responses from document content. Additionally, the platform supports offline execution and multilingual communication, improving accessibility in low-connectivity environments. The primary objective of QueryMate AI is to reduce manual effort, improve retrieval accuracy, and provide a user-friendly mechanism for interacting with large documents using AI-powered technologies.

2. LITERATURE REVIEW

Research in document intelligence and semantic information retrieval has evolved rapidly with the advancement of transformer architectures and deep learning techniques.

Early document retrieval systems relied mainly on keyword indexing and rule-based matching algorithms. Although these methods improved search efficiency, they lacked contextual understanding and semantic interpretation capabilities.

The introduction of transformer architectures revolutionized NLP applications. Devlin et al. proposed BERT, which enabled bidirectional contextual understanding of language and significantly improved performance in question-answering tasks. Later, Sentence-BERT extended this capability by generating semantically meaningful sentence embeddings suitable for similarity comparison and retrieval systems.

Several researchers explored semantic document retrieval using vector databases such as FAISS. These systems convert textual information into dense vector representations and perform similarity-based matching for accurate context retrieval. Modern retrieval systems have also integrated large language models for generating natural conversational responses from retrieved contexts.

Multilingual NLP models further expanded accessibility by enabling systems to understand multiple languages. Cross-lingual embedding techniques allowed semantic matching even when queries and document content belonged to different languages.

Despite these advancements, many existing systems still depend heavily on cloud infrastructure, paid APIs, or continuous internet access. Additionally, several solutions focus only on text retrieval without providing offline functionality, multilingual interaction, or voice-based accessibility.

QueryMate AI addresses these gaps by integrating offline semantic retrieval, multilingual support, and intelligent contextual answering into a unified lightweight system suitable for academic and practical applications.

3. PROBLEM STATEMENT

Large digital documents often contain important information scattered across multiple pages, making manual navigation inefficient and time-consuming. Existing PDF readers

provide only basic keyword search mechanisms that cannot understand user intent or contextual meaning.

Users frequently struggle to retrieve precise answers from lengthy files because traditional systems depend solely on exact word matching. This results in reduced productivity, increased manual effort, and the possibility of overlooking critical information.

An intelligent solution is therefore required that can:

- Understand natural language queries,
- Retrieve context-aware answers,
- Support multilingual interaction,
- Operate efficiently without internet dependency,
- Reduce the complexity of document navigation.

QueryMate AI is developed to overcome these limitations using AI-powered semantic search and document understanding techniques.

4. PROPOSED SYSTEM

QueryMate AI is an intelligent document interaction system designed to provide contextual answers from uploaded documents using semantic retrieval and transformer-based NLP techniques.

The system consists of multiple integrated modules working together to process documents and answer user queries efficiently.

4.1 Document Processing Module

This module extracts textual content from uploaded PDF and DOCX files. The extracted text is cleaned, segmented into smaller chunks, and prepared for embedding generation.

4.2 Embedding Generation Module

The processed text chunks are converted into dense vector embeddings using transformer-based models. These embeddings preserve semantic meaning and contextual relationships between sentences.

4.3 Vector Search Module

FAISS indexing is used for storing embeddings and performing similarity-based nearest-neighbor searches. When a user submits a query, the system retrieves the most relevant document segments.

4.4 NLP Answer Generation Module

The retrieved contextual passages are forwarded to the NLP response generator, which creates concise and meaningful answers for the user query.

4.5 Multilingual and Voice Interaction Module

The system supports multilingual communication and voice-based query input to improve accessibility and usability for users from different linguistic backgrounds.

4.6 Offline AI Engine

Unlike cloud-dependent systems, QueryMate AI supports offline operation using locally deployed AI models and vector databases.

5. SYSTEM ARCHITECTURE

The QueryMate AI system is designed using a modular and layered architecture to ensure efficient document processing, semantic search, and intelligent response generation. The architecture separates the system into different functional layers such as the Frontend Layer, Backend Layer, AI Processing Layer, and Data Storage Layer. This structured design improves scalability, maintainability, and overall system performance. Each layer performs specific tasks and communicates with other layers through defined workflows to provide accurate and context-aware answers from uploaded documents.

1. Front End Layer

The Front End Layer acts as the interaction point between the user and the QueryMate AI system. This layer is developed using Streamlit and provides a simple, responsive, and user-friendly interface through which users can upload documents and ask questions in natural language.

The frontend provides the following functionalities:

- Document upload interface for PDF and DOCX files
- Query input box for entering questions
- Response display section for showing generated answers
- Chat-like interaction for continuous querying
- Display of contextual references and extracted content
- Upload confirmation and processing status messages

The frontend is responsible for collecting user inputs and presenting system-generated outputs in a readable and interactive format. It also improves accessibility by simplifying document interaction for non-technical users.

2. Back End Layer

The Back End Layer functions as the central processing unit of QueryMate AI. It is implemented using Python and handles all major operations including document extraction, preprocessing, semantic search, embedding generation, and answer retrieval.

The backend performs multiple operations such as:

- Extracting text from uploaded PDF and DOCX documents
- Cleaning and preprocessing extracted text
- Splitting large text into smaller chunks
- Generating semantic embeddings using transformer models

- Processing user queries and converting them into embeddings
- Communicating with the vector database for similarity search
- Retrieving relevant document segments
- Generating final responses using NLP models

The backend also manages the communication between the frontend interface, AI models, and the vector database. It ensures smooth workflow execution and accurate contextual response generation.

3. AI Processing Layer

The AI Processing Layer is responsible for intelligent semantic understanding and natural language processing. This layer integrates transformer-based AI models and semantic retrieval techniques to analyze document content and user queries.

The AI layer includes the following components:

- Sentence Transformer models for embedding generation
- MiniLM and DistilBERT models for semantic understanding
- Transformer-based Question Answering models
- Semantic similarity matching algorithms
- Context-aware response generation system

The AI models convert both document text and user queries into vector embeddings that preserve contextual meaning. Using semantic similarity search, the system identifies the most relevant content from uploaded documents and generates accurate responses.

This layer significantly improves retrieval quality compared to traditional keyword-based search systems.

4. Data Storage Layer

The Data Storage Layer is responsible for storing document embeddings, processed data, and query-related information. QueryMate AI uses FAISS (Facebook AI Similarity Search) as the primary vector database for efficient semantic indexing and retrieval.

The database layer stores:

- Document embeddings
- Processed text chunks
- Query embeddings
- Retrieved contextual passages
- User interaction data
- Temporary processing information

FAISS enables fast nearest-neighbor search operations, allowing the system to retrieve semantically similar document segments within milliseconds. The data layer improves retrieval speed, scalability, and overall response efficiency.

5. System Workflow

The QueryMate AI system follows a structured workflow for intelligent document processing. Initially, the user uploads a document through the frontend interface. The backend extracts and preprocesses the text content from the uploaded file. The AI processing layer then generates vector embeddings for the document chunks and stores them in the FAISS vector database.

When the user submits a question, the query is converted into embeddings and compared with stored document vectors using semantic similarity search. The most relevant document sections are retrieved and passed to the transformer-based NLP model for contextual answer generation. Finally, the generated response is displayed on the frontend interface along with supporting references from the document.

The layered architecture ensures efficient data flow, modular development, faster processing, and improved scalability of the QueryMate AI system.



Fig 5.1 System Architecture

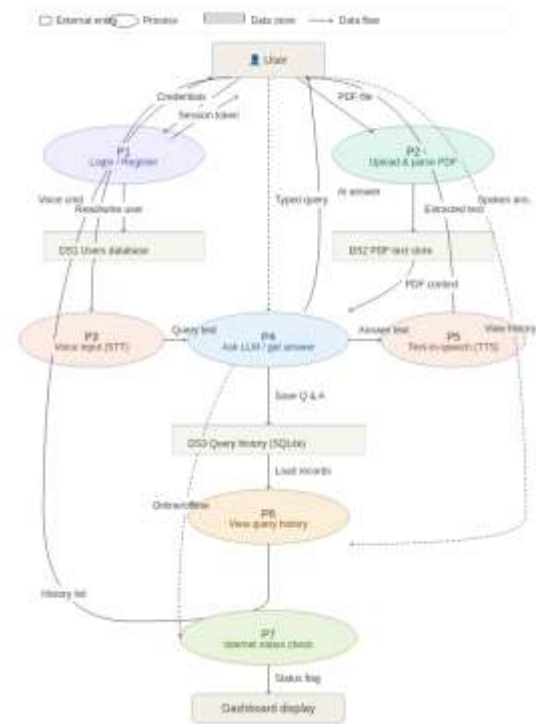


Fig 5.2 DFD Diagram

6. METHODOLOGY

The methodology of QueryMate AI is designed to provide an intelligent and efficient approach for retrieving information from digital documents using Artificial Intelligence and Natural Language Processing techniques. The system follows a modular workflow where each component performs a specific task to ensure accurate and fast response generation. Initially, the user uploads a PDF or DOCX document through the web interface. The uploaded document is then processed by the document extraction module, which reads and extracts textual content from the file. After extraction, the text is cleaned and divided into smaller meaningful segments or chunks to improve processing efficiency and semantic understanding.

Once the text preprocessing is completed, the embedding generation module converts the text chunks into dense vector representations using transformer-based Sentence Transformer models. These embeddings capture the contextual meaning of the text rather than relying only on exact keywords. The generated vectors are stored in the FAISS vector database, which enables high-speed semantic similarity search. When the user enters a question in natural language, the query is also transformed into vector embeddings using the same embedding model. The semantic search engine compares the query embedding with stored document embeddings and retrieves the most relevant text segments based on similarity scores.

The retrieved contextual information is then forwarded to the NLP-based response generation module, where transformer models analyze the context and generate meaningful answers for the user. The final response is displayed on the frontend interface along with related contextual references from the document. The system also supports multilingual interaction

and offline processing, making it accessible in low-connectivity environments. By integrating semantic search, transformer embeddings, vector databases, and intelligent NLP models, QueryMate AI provides a faster, smarter, and more user-friendly solution for document understanding and information retrieval.

7. IMPLEMENTATION

The implementation of QueryMate AI is carried out as an intelligent AI-powered document question-answering system that combines semantic search, Natural Language Processing, and transformer-based machine learning models. The complete system is developed as a web-based application capable of processing uploaded PDF and DOCX documents and generating context-aware answers for user queries. The implementation focuses on improving document interaction, reducing manual effort, and enabling efficient information retrieval through semantic understanding.

The implementation process begins with the document upload functionality provided through the Streamlit-based web interface. Users can upload PDF and DOCX files directly into the system using a simple and interactive frontend interface. After successful upload, the backend processing module extracts textual information from the uploaded files using document processing libraries such as PyPDF2 and python-docx. The extracted content is then cleaned and preprocessed to remove unnecessary spaces, symbols, and formatting inconsistencies. The text is further divided into smaller chunks to improve semantic retrieval accuracy and processing performance.

Once preprocessing is completed, the system generates vector embeddings for each text chunk using Sentence Transformer models. Transformer-based models such as MiniLM and DistilBERT are used because they provide contextual understanding while maintaining efficient processing speed. These embeddings capture semantic meaning from the document rather than relying only on exact keyword matches. The generated embeddings are stored inside the FAISS vector database, which enables high-speed similarity search and efficient contextual retrieval.

When a user enters a question in natural language, the system converts the query into vector embeddings using the same transformer model. The semantic search engine then compares the query embeddings with stored document embeddings to identify the most relevant text segments based on similarity scores. The retrieved contextual passages are forwarded to transformer-based Question Answering models from Hugging Face Transformers for intelligent answer generation.

8. RESULTS AND DISCUSSION

The QueryMate AI system was successfully implemented and evaluated using multiple types of documents including academic PDFs, reports, resumes, and technical files. The system demonstrated effective performance in retrieving context-aware answers from uploaded documents using

semantic search and transformer-based Natural Language Processing techniques. During testing, users were able to upload documents, ask questions in natural language, and receive meaningful responses within a short processing time. The system efficiently extracted relevant information from large documents, reducing the need for manual searching and improving overall productivity.

The semantic retrieval mechanism showed significantly better performance compared to traditional keyword-based search methods. Instead of relying only on exact word matching, QueryMate AI analyzed the contextual meaning of both the document content and user queries. This enabled the system to provide more accurate and relevant answers even when the query wording differed from the actual document text. The integration of Sentence Transformer embeddings and FAISS vector similarity search improved retrieval speed and contextual matching accuracy.

The system also performed effectively in offline environments because all AI models and vector databases were deployed locally. This offline capability increased accessibility and reduced dependency on internet connectivity. Experimental observations showed that the system generated faster responses for small and medium-sized documents, while processing time slightly increased for very large documents containing hundreds of pages. However, the overall response quality remained acceptable for academic and professional applications.

9. FUTURE SCOPE

QueryMate AI has significant potential for future enhancement and expansion in the field of intelligent document processing and AI-based information retrieval. One major improvement can be the integration of Optical Character Recognition (OCR) technology to support scanned documents, handwritten notes, and image-based PDFs. This would allow the system to extract and process text from a wider variety of document formats. Future versions of the system can also support additional file types such as PPT, XLSX, TXT, HTML, and image documents to increase flexibility and usability.

The performance and accuracy of the system can be further improved through the integration of advanced Large Language Models and Retrieval-Augmented Generation (RAG) techniques. These technologies can enhance contextual understanding, reduce irrelevant responses, and provide more human-like conversational interaction. Another important enhancement is the implementation of multilingual support for regional and international languages, enabling users from different linguistic backgrounds to interact with documents more comfortably. Cloud deployment and multi-user accessibility can also be introduced in future implementations. Hosting the system on cloud platforms such as AWS, Azure, or Google Cloud would allow real-time collaboration, remote access, and large-scale document processing. Additional features such as chat history management, personalized recommendation systems, voice-based interaction, and AI-powered document summarization can further improve user experience and system intelligence.

11. CONCLUSION

QueryMate AI is an intelligent AI-powered document question-answering system developed to simplify information retrieval from digital documents. The system successfully combines semantic search, transformer-based Natural Language Processing, vector databases, and contextual response generation to provide accurate answers directly from uploaded PDF and DOCX files. Unlike traditional keyword-based search systems, QueryMate AI understands the semantic meaning of user queries and retrieves contextually relevant information, significantly reducing manual effort and document navigation time.

The implementation of transformer embeddings, FAISS vector similarity search, and Hugging Face NLP models enabled the system to generate fast and meaningful responses with improved contextual accuracy. The integration of offline AI processing further increased accessibility and usability in low-connectivity environments. The user-friendly Streamlit interface simplified interaction with the system, making it suitable for students, researchers, professionals, and general users.

Experimental evaluation demonstrated that the system performs efficiently for academic and professional document retrieval tasks while providing better semantic understanding compared to conventional search approaches. Although certain limitations exist, such as reduced accuracy for ambiguous queries and increased processing time for very large documents, the overall performance of the system remains reliable and practical for real-world applications.

In conclusion, QueryMate AI serves as an effective bridge between Artificial Intelligence and document interaction technologies. The project demonstrates how semantic search and NLP-based retrieval systems can improve productivity, accessibility, and intelligent information extraction from digital documents. With future enhancements and scalability improvements, QueryMate AI has strong potential to evolve into a powerful industry-level intelligent document assistant.

10. REFERENCES

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, USA, 2019, pp. 4171-4186.
2. Reimers, N., & Gurevych, I., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 3982-3992.
3. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P., "SQuAD: 100,000+ Questions for Machine Comprehension of Text," Proceedings of

the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, USA, 2016, pp. 2383-2392.

4. Johnson, J., Douze, M., & Jégou, H., "Billion-scale Similarity Search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535-547, 2021, doi: 10.1109/TBDATA.2019.2921572.
5. Vaswani, A., Shazeer, N., Parmar, N., et al., "Attention Is All You Need," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
6. Wolf, T., Debut, L., Sanh, V., et al., "Transformers: State-of-the-Art Natural Language Processing," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38-45.
7. Lewis, P., Perez, E., Piktus, A., et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 9459-9474.
8. Conneau, A., Khandelwal, K., Goyal, N., et al., "Unsupervised Cross-lingual Representation Learning at Scale," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440-8451.

11. AUTHOR'S:

Mr. Dayanand Mahadaji Patil is currently pursuing a Bachelor's Degree in Computer Science & Engineering at Ashokrao Mane Group Of Institution, vathar. he is publishing this paper under the esteemed guidance of Miss. Anupama Deepak Patil.



Mr. Sainath Baburao Nandiwale is currently pursuing a Bachelor's Degree in Computer Science & Engineering at Ashokrao Mane Group Of Institution, vathar. he is publishing this paper under the esteemed guidance of Miss. Anupama Deepak Patil.



Miss. Sakshi Anil Mahapure is currently pursuing a Bachelor's Degree in Computer Science & Engineering at Ashokrao Mane Group Of Institution, vathar. she is publishing this paper under the esteemed guidance of Miss. Anupama Deepak Patil.



Miss. Sudharani Maruti Bhise is currently pursuing a Bachelor's Degree in Computer Science & Engineering at Ashokrao Mane Group Of Institution, vathar. she is publishing this paper under the esteemed guidance of Miss. Anupama Deepak Patil.



Miss. Pranjali Tanaji Jadhav is currently pursuing a Bachelor's Degree in Computer Science & Engineering at Ashokrao Mane Group Of Institution, vathar. he is publishing this paper under the esteemed guidance of Miss. Anupama Deepak Patil.



Miss. Anupama Deepak Patil is currently working as an Assistant Professor in Computer Science & Engineering at Ashokrao Mane Group Of Institution, vathar. she has been guiding and mentoring students in various academic and research activities. he has provided his valuable guidance for the successful completion and publication of this paper.

