An Autonomous Knowledge Assistant Using Hybrid Vector Search and Ollama

A Hybrid Retrieval-Augmented Generation Approach for Offline Institutional Conversational AI

Peese Vamshi [#1], Dr.D.William Albert [#2], G Sree Ramulu [#3]

[#1]*M.Tech Student, Dept. of CSE, Bheema Institute of Technology and Science, Adoni, India*

[#2]*Professor, Head of Dept. CSE, Bheema Institute of Technology and Science, Adoni, India*

[#3]*Professor, Dept. of CSE, Bheema Institute of Technology and Science, Adoni, India*

[#1] vamshipeesay@gmail.com, #2 dr.albertdwgtl@gmail.com, #3 gudla.sri698@gmail.com

***Abstract*—**
Recent advancements in Artificial Intelligence and Large Language Models (LLMs) have significantly improved conversational systems and intelligent information retrieval applications. However, most modern AI assistants rely on cloud-based infrastructures, leading to concerns regarding data privacy, API dependency, operational cost, and internet connectivity. This paper presents the development of an Autonomous Knowledge Assistant using Hybrid Vector Search integrated with Retrieval-Augmented Generation (RAG) for secure and offline conversational information retrieval.

The proposed system combines dense semantic retrieval using FAISS with sparse keyword-based retrieval using BM25 to improve response accuracy and contextual relevance. The architecture integrates Streamlit for the user interface, LangChain for retrieval orchestration, Ollama for local LLM deployment, and FAISS for vector indexing. The system also includes query rewriting, exact answer extraction, response polishing, and administrative knowledge management functionalities.

The proposed assistant enables users to interact with institutional knowledge bases using natural language queries without requiring technical expertise. Experimental evaluation demonstrates that hybrid retrieval significantly improves precision and reduces hallucinated responses compared to standalone retrieval methods. The developed system provides a cost-effective, scalable, privacy-preserving, and fully offline conversational AI solution suitable for educational institutions and enterprise knowledge systems.

Keywords— Retrieval-Augmented Generation (RAG); Hybrid Search; FAISS; BM25; Conversational AI; Offline LLM; LangChain; Ollama; Knowledge Assistant; Natural Language Processing.*Keywords— Natural Language Processing (NLP); Dialogflow; FastAPI; Conversational AI; Chatbot System; Database Automation; MySQL; Intent Recognition; Webhook Integration; Human–Computer Interaction.*

I. INTRODUCTION

Artificial Intelligence has transformed modern information systems by enabling machines to understand and process human language effectively. Conversational AI systems are now widely used in domains such as healthcare, education, customer support, banking, and enterprise management. Traditional information retrieval systems primarily rely on keyword-based search methods that lack semantic understanding and contextual awareness.

Large Language Models (LLMs) have significantly improved conversational capabilities by generating context-aware responses. However, standalone LLMs often suffer from hallucination, lack of domain-specific knowledge, and dependency on cloud-based infrastructures. Most commercial AI assistants require internet connectivity and external API services, creating challenges related to privacy, cost, and institutional data security.

Educational institutions frequently manage repetitive queries related to admissions, fee structures, examination schedules, faculty details, academic policies, and administrative procedures. Conventional FAQ systems fail to handle paraphrased queries and contextual interaction effectively. There is a growing demand for intelligent offline systems capable of accurate institutional knowledge retrieval using natural language communication.

Retrieval-Augmented Generation (RAG) has emerged as an effective solution for grounding LLM responses using external knowledge sources. Dense vector retrieval methods such as FAISS capture semantic similarity, while sparse retrieval methods such as BM25 provide precise keyword matching. Hybrid retrieval combines both approaches to improve retrieval accuracy and response relevance.

This paper presents an Autonomous Knowledge Assistant using Hybrid Vector Search integrated with Retrieval-Augmented Generation. The proposed system operates completely offline using locally hosted Large Language Models through Ollama. The architecture integrates FAISS vector indexing, BM25 sparse retrieval, LangChain orchestration, and Streamlit-based conversational interaction.

The system allows users to communicate with institutional knowledge bases using natural language queries. Query rewriting, exact answer extraction, and response polishing mechanisms improve retrieval precision

and reduce hallucinated responses. An administrative module enables dynamic knowledge management through structured CSV uploads.

The proposed approach improves accessibility, eliminates API dependency, enhances institutional data privacy, and provides an efficient conversational information retrieval framework for offline environments.

II. LITERATURE REVIEW

The rapid advancement of Natural Language Processing and conversational AI technologies has significantly influenced intelligent information retrieval systems. Early chatbot systems relied on rule-based architectures and keyword matching techniques for generating predefined responses. Although such systems were effective for simple interactions, they lacked contextual understanding and semantic flexibility.

Modern conversational systems utilize transformer-based Large Language Models capable of understanding contextual relationships between words and generating human-like responses. Research by Vaswani et al. introduced transformer architectures that significantly improved language understanding capabilities. Subsequent models such as BERT and GPT further advanced contextual text generation and semantic processing.

Dense vector retrieval techniques have become increasingly popular for semantic search applications. FAISS provides efficient similarity search mechanisms using vector embeddings and approximate nearest neighbor indexing. Dense retrieval systems improve semantic understanding but may fail for exact keyword matching queries.

Sparse retrieval models such as BM25 remain strong baselines in information retrieval systems. BM25 improves document ranking through term frequency and inverse document frequency weighting mechanisms. However, sparse retrieval lacks semantic understanding and struggles with paraphrased queries.

Recent studies have focused on hybrid retrieval architectures combining dense and sparse retrieval techniques. Hybrid systems improve retrieval precision and recall by balancing semantic similarity with keyword-level matching. Retrieval-Augmented Generation (RAG) further enhances conversational systems by grounding language models using retrieved external documents.

Offline AI deployment has also gained research interest due to privacy and security concerns associated with cloud-based APIs. Local LLM deployment frameworks such as Ollama enable secure conversational AI applications without external dependencies.

Despite these advancements, limited research exists on fully offline institutional knowledge assistants integrating hybrid retrieval, exact answer extraction, and administrative knowledge management. The proposed system addresses these limitations by combining FAISS, BM25, RAG, and local LLM deployment into a unified conversational framework..

III. PROPOSED METHOD

A. *System Overview*

The proposed Autonomous Knowledge Assistant is designed to provide intelligent conversational interaction with institutional knowledge bases using hybrid retrieval and Retrieval-Augmented Generation. Instead of relying on traditional search systems or cloud APIs, the proposed model operates entirely offline using local LLM deployment.

The architecture integrates:

a. Streamlit for user interaction

b. LangChain for orchestration

c. FAISS for dense vector retrieval

d. BM25 for sparse retrieval

e. Ollama for local LLM inference

f. CSV-based knowledge management

The system enables users to retrieve institutional information using natural language queries.

B. *System Overview*

The architecture consists of:

a. User Interface Layer

b. Query Rewriting Module

c. Hybrid Retrieval Engine

d. Exact Answer Extraction Module

e. RAG Generation Module

f. Response Polishing Layer

g. Administrative Knowledge Management Module

The workflow begins when the user submits a natural language query through the Streamlit interface. The query rewriting module expands ambiguous queries into structured questions. The hybrid retrieval engine performs dense semantic search using FAISS and sparse keyword search using BM25.

Retrieved documents are merged and filtered. Exact answer extraction checks whether direct responses exist within retrieved documents. If no exact response is found, the retrieved context is passed to the local LLM through a Retrieval-Augmented Generation pipeline. The generated response is polished and displayed to the user.

C. *Working Procedure*

The complete workflow of the proposed system is as follows:

Step 1: User submits query through chatbot interface.
Step 2: Query rewriting improves incomplete or ambiguous input.
Step 3: FAISS performs semantic similarity search.
Step 4: BM25 performs keyword-based retrieval.
Step 5: Results are merged and deduplicated.
Step 6: Exact answer extraction checks structured responses.

Step 7: Retrieved context is passed to local LLM.
Step 8: RAG generates grounded response.
Step 9: Final response is polished and displayed.

The conversational workflow improves accessibility and simplifies institutional information retrieval.

D. Database Design

The proposed hybrid retrieval combines:

Dense Retrieval:

- Semantic similarity search
- Embedding-based matching

Sparse Retrieval:

- Keyword ranking
- Exact phrase matching

Final retrieval set:

$H = Unique(V \cup K)$

Where:

- V = Vector retrieval results
- K = BM25 retrieval results

This approach improves both precision and recall.

E. Advantages of the Proposed System

The proposed system offers several advantages:

- Fully offline operation
- Improved retrieval accuracy
- Reduced hallucination
- No API dependency
- Enhanced data privacy
- Cost-effective deployment
- Dynamic knowledge management
- User-friendly conversational interaction

The system is suitable for educational institutions, enterprise environments, and secure organizational deployments.

F. Algorithm for Hybrid Query Processing

Algorithm Steps

Step 1: Start system.
Step 2: Receive user query.
Step 3: Rewrite query if ambiguous.
Step 4: Perform vector similarity search using FAISS.
Step 5: Perform keyword retrieval using BM25.
Step 6: Merge and deduplicate retrieved documents.
Step 7: Check for exact answer extraction.
Step 8: If exact answer exists, return response.
Step 9: Else execute RAG generation.
Step 10: Polish generated response.
Step 11: Display final response.
Step 12: Stop.

TABLE III

G. Comparison Between Existing and Proposed System

Feature	Traditional FAQ System	Proposed Hybrid System
Semantic Understanding	No	Yes
Keyword Precision	Limited	High
Offline Operation	Limited	Yes
Hallucination Control	Weak	Strong
Dynamic Knowledge Update	Limited	Yes
API Dependency	High	No
Conversational Interaction	Basic	Advanced

IV. IMPLEMENTATION

The implementation of the proposed Autonomous Knowledge Assistant is carried out using Streamlit, LangChain, FAISS, BM25 retrieval, Ollama, and Python technologies. The system integrates hybrid retrieval and Retrieval-Augmented Generation to provide intelligent offline conversational interaction with institutional knowledge bases.

Streamlit is used to develop the chatbot interface for real-time conversational communication. LangChain is used for orchestrating retrieval pipelines, prompt management, document loading, and response generation. FAISS is used for dense vector similarity search, while BM25 is integrated for sparse keyword-based retrieval. Ollama is used for deploying local Large Language Models and embedding models in an offline environment.

IV. RESULTS AND DISCUSSION

The proposed Autonomous Knowledge Assistant was evaluated using multiple institutional conversational queries to measure retrieval accuracy, response relevance, hallucination

control, and overall system performance.

The evaluation focused on:

Intent understanding

Retrieval precision

Conversational response quality

Exact answer extraction accuracy

Offline operational stability

A. Experimental Evaluation

The system was tested using institutional FAQ datasets containing administrative and academic information.

Sample query categories:

Admission details

Fee structure

Examination schedules

Faculty information

Campus facilities

The system successfully generated accurate responses for most conversational queries.

B. Sample Test Cases

Test Case	User Query	Expected Output
TC01	What is admission fee?	Fee details displayed
TC02	Office timings?	Office hours displayed
TC03	Fees?	Correct rewritten answer
TC04	Random unrelated query	"Information not available"

C. Retrieval Performance Analysis

The performance of hybrid retrieval was compared with standalone dense and sparse retrieval methods.

Retrieval Method	Precision
Dense Retrieval	0.78
Sparse Retrieval	0.72
Hybrid Retrieval	0.89

VI. CONCLUSION

This paper presented the development of an Autonomous Knowledge Assistant using Hybrid Vector Search integrated with Retrieval-Augmented Generation for intelligent offline conversational information retrieval.

The proposed system combines FAISS-based dense retrieval with BM25 keyword retrieval to improve search precision and contextual relevance. The architecture integrates Streamlit, LangChain, Ollama, and local vector indexing to create a secure and cost-effective conversational AI framework.

The developed assistant enables users to interact with institutional knowledge bases using natural language queries without requiring technical expertise. Query rewriting, exact answer extraction, and response polishing mechanisms further improve retrieval accuracy and reduce hallucinated responses.

Experimental results demonstrate that hybrid retrieval significantly improves precision compared to standalone retrieval methods. The system successfully provides secure, scalable, and offline conversational knowledge assistance suitable for educational institutions and enterprise environments.

Future enhancements may include multilingual support, voice-based interaction, cloud synchronization, role-based access control, and integration with advanced re-ranking models for further improving retrieval performance and conversational intelligence.

VII. REFERENCES

[1] A. Vaswani et al., "Attention Is All You Need," Neural Information Processing Systems (NIPS), pp. 5998–6008, 2017.

[2] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020.

[3] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," IEEE Transactions on Big Data, 2017.

[4] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, 2009.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," EMNLP, 2019.

[6] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," EMNLP, 2020.

[7] T. B. Brown et al., "Language Models are Few-Shot Learners," NeurIPS, 2020.

[8] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Pearson Education, 2021.

[9] LangChain Documentation, "LangChain Framework for LLM Applications," 2023.

[10] FAISS Documentation, "Facebook AI Similarity Search Library," 2023.

[11] Ollama Documentation, "Local LLM Deployment Platform," 2024.