International Journal of Web of Multidisciplinary Studies
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
website: https://ijwos.com
Volume 03, Issue 10,October 2026
OPEN ACCESS
E-ISSN : 3049-2424
DOI: https://doi.org/10.71366/ijwos

Career Guidance Chatbot Using Retrieval-Augmented Generation (RAG)

ANIMIREDDI NOOKAMBIKA[1], BANIKULA SAI KIRAN[2], BHEEMISETTY DIVYA[3], K. VARAPRASAD[4]
[1,2,3] *Master of Computer Applications Final Year Students, Avanthi Institute of Engineering & Technology, Anakapalli, Andhra Pradesh, India*
[4] *Assistant Professor, Department of CSE, Avanthi Institute of Engineering & Technology, Anakapalli, Andhra Pradesh, India*

Article Info

Article History:
Published: 31 Oct 2026

Publication Issue:
Volume 03, Issue 10
October 2026

Page Number:
670-673

Corresponding Author:
ANIMIREDDI NOOKAMBIKA

Abstract:
Choosing an appropriate career path is challenging for students and job seekers because reliable, personalized guidance may be difficult to access. This paper presents a conversational career guidance chatbot based on Retrieval-Augmented Generation (RAG) and the Zephyr language model. The system retrieves relevant information from a curated career knowledge base using semantic embeddings and FAISS similarity search, then supplies the retrieved context to the language model to generate tailored guidance. A Streamlit interface supports interaction, while LangChain coordinates the retrieval and generation pipeline. The source implementation reports an MRR@5 of 0.87, a ROUGE-L score of 0.64, and a reduction in hallucination rate from 31.7% for a non-retrieval baseline to 4.2% for the RAG-based system. The proposed framework aims to provide accessible, context-aware career recommendations grounded in domain-specific information.
Keywords: Career guidance, conversational chatbot, retrieval-augmented generation, semantic search, FAISS, Sentence-Transformers, Zephyr, Streamlit.

1. Introduction

The contemporary employment landscape presents students and job seekers with a wide range of career choices, each associated with distinct educational requirements, competencies, and professional pathways. Conventional counseling remains valuable, but access to individual guidance may be limited by counselor availability, time, and the difficulty of continuously updating career information.

Static websites, aptitude-based tools, and rule-based chatbots generally provide predefined information and may not adequately support personalized, multi-turn interaction. Generative language models improve conversational interaction, but responses produced without domain evidence can contain unsupported or outdated statements.

This paper presents a conversational career guidance chatbot based on Retrieval-Augmented Generation (RAG). The system retrieves relevant passages from a curated career knowledge base and supplies them as context to the Zephyr language model. A Streamlit interface supports interaction, while semantic embeddings and FAISS provide similarity-based retrieval. The objective is to produce context-aware guidance aligned with a user's academic background, interests, and career query.

2. RELATED WORK

Early conversational systems such as ELIZA used pattern matching, whereas transformer-based language models enabled more flexible language understanding and generation. Vaswani et al. introduced the Transformer architecture, which uses attention mechanisms to model relationships between tokens [2].

Lewis et al. introduced Retrieval-Augmented Generation for knowledge-intensive NLP tasks, combining parametric generation with retrieved non-parametric knowledge [1]. This approach is relevant to career guidance because educational pathways, role requirements, and skill expectations may change over time. Reimers and Gurevych

Copyright © 2026 The Author(s) : This is an open access article under the CC BY license (http://creativecommons.org/licenses/by-nc/4.0/)

proposed Sentence-BERT for generating semantically meaningful sentence embeddings, supporting efficient semantic comparison [6].

Zephyr-7B is described as a chat-oriented model derived from the Mistral-7B architecture and aligned for instruction-following tasks [3], [4]. Gao et al. survey RAG methods and discuss their ability to mitigate limitations associated with knowledge freshness and unsupported generation [7]. These works motivate the combination of semantic retrieval and an instruction-tuned language model in the proposed system.

3. PROPOSED SYSTEM

A. System Overview

The system has two major stages: (1) offline knowledge-base preparation and indexing, and (2) online query processing and response generation. In the offline stage, career-related documents are parsed, cleaned, divided into semantically coherent chunks, converted into vector embeddings, and indexed using FAISS. In the online stage, a user query is embedded with the same embedding model, matched against the FAISS index, and combined with retrieved passages to construct a context-enriched prompt for Zephyr.

B. Main Components

Knowledge base: A curated collection of career documents, industry information, job descriptions, skill requirements, and academic programme information. Retrieval engine: FAISS performs vector similarity search and returns the most relevant chunks. Embedding model: Sentence-Transformers, such as all-MiniLM-L6-v2, converts text into dense vector representations. Language model: Zephyr generates a conversational response using the query and retrieved context. Orchestration: LangChain coordinates document loading, chunking, retrieval, and prompt construction. User interface: Streamlit provides the interactive chat experience. SQLite is used for user authentication and conversation-history management, as described in the implementation.

4. METHODOLOGY

A. Knowledge-Base Ingestion

Career resources are collected and parsed into machine-readable text. The text is cleaned and divided into manageable, semantically coherent chunks. Chunk overlap is used to preserve continuity when a concept spans adjacent passages. Each chunk is encoded using the selected Sentence-Transformers model. The resulting vectors, together with references to their source chunks, are added to a FAISS index. Fig. 3 illustrates this offline process.

B. Query Processing and Retrieval

When a student or job seeker submits a question through Streamlit, the application validates the input and converts it into a query embedding using the same embedding model used during indexing. FAISS compares the query vector with indexed vectors and returns the top-ranked relevant chunks according to vector similarity. Using a common embedding space is essential for meaningful query-to-document comparison.

C. Context Augmentation and Response Generation

The retrieved passages are combined with the original question to form an augmented prompt. The prompt instructs Zephyr to answer using the supplied career context and to relate recommendations to the user's stated academic background and interests. The generated response may describe suitable career roles, required skills, learning paths, expected salary ranges, and market outlook where such information is available in the retrieved material. The response is returned to the interface, where the user can continue the conversation. Fig. 2 presents the online workflow.

D. Personalization and Session Management

User-provided academic details, interests, and conversation context can be used to tailor the response. Session state supports continuity across turns. SQLite is identified in the source implementation for authentication and history tracking. Sensitive profile information should be handled with appropriate access controls and data-minimization practices.

5. SYSTEM ARCHITECTURE

Fig. 1 shows the end-to-end architecture. The knowledge corpus is transformed into vectors before user interaction. At runtime, the query follows a separate retrieval-and-generation path. The retrieval component supplies evidence to

the language model; it does not retrain the language model. This separation allows the knowledge base to be updated independently of the generative model.

Fig. 1. Proposed RAG-based career guidance chatbot architecture

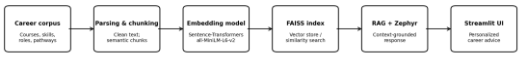
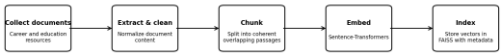


Fig. 2. Online query-processing workflow



Fig. 3. Offline knowledge-base ingestion workflow



6. IMPLEMENTATION

The implementation described in the source paper uses Python 3.9 or later, Hugging Face Transformers for language-model integration, LangChain for RAG orchestration, FAISS for vector indexing, Sentence-Transformers for embeddings, Streamlit for the web interface, and SQLite for authentication and conversation history. The principal model configuration retained in this IEEE-formatted version is Zephyr-7B, with all-MiniLM-L6-v2 as the example embedding model.

The source draft also mentions Flask and LLaMA 3.2 in one implementation paragraph. These alternatives conflict with the Streamlit and Zephyr configuration described elsewhere. To make the paper internally consistent, this version uses Streamlit and Zephyr as the stated implementation path. The authors should verify the actual deployed configuration before submission.

7. EXPERIMENTAL RESULTS

The reported evaluation in the source paper gives an MRR@5 of 0.87 for retrieval, a ROUGE-L score of 0.64 for generated responses, and a reduction in hallucination rate from 31.7% for a non-retrieval baseline to 4.2% for the RAG-based system. These values are retained as reported results; the source draft does not provide the dataset composition, sample count, annotation protocol, confidence intervals, or detailed baseline configuration.

TABLE I. REPORTED PERFORMANCE

Metric	**Reported value**
Retrieval Mean Reciprocal Rank at 5 (MRR@5)	0.87
Response ROUGE-L	0.64
Hallucination rate — non-retrieval baseline	31.7%
Hallucination rate — proposed RAG system	4.2%

These results indicate improved retrieval ranking and a lower reported hallucination rate when retrieval is used. For a reproducible evaluation, the final submission should specify the test queries, relevance judgments, number of runs, definition of hallucination, and the exact baseline and RAG configurations.

8. CONCLUSION AND FUTURE WORK

This paper presents a career guidance chatbot that combines semantic retrieval with a generative language model. The proposed pipeline uses a curated career knowledge base, Sentence-Transformers embeddings, FAISS similarity search, LangChain orchestration, Zephyr generation, and a Streamlit interface. The reported evaluation values suggest

that grounding generation in retrieved career information can improve response relevance and reduce unsupported output compared with a non-retrieval baseline.

Future work identified in the source includes multilingual support, resume parsing for automated skill-gap analysis, and integration with live job-board APIs to provide dynamic salary trends and current market-demand information. Further evaluation should also examine recommendation usefulness, retrieval coverage, factual correctness, latency, and user satisfaction.

References

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv:2005.11401, 2020.

[2] A. Vaswani et al., "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems, vol. 30, 2017.

[3] Hugging Face, "Zephyr-7B," model documentation and model card.

[4] L. Tunstall et al., "Zephyr: Direct Distillation of LM Alignment," arXiv:2310.16944, 2023.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, pp. 3982–3992, 2019.

[6] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv:2312.10997, 2023.