2025 IEEE International Students' Conference on Electrical, Electronics, and Computer Science

TechAssist: A RAG-Based Chatbot for Accessing Technical Information from StackOverflow

Raj Vasaniya
Department of Information Technology
Dharmsinh Desai University
Nadiad, Gujarat, India
rajvasaniya07@gmail.com

Meet Visodiya
Department of Information Technology
Dharmsinh Desai University
Nadiad, Gujarat, India
meetvisodiya3@gmail.com

Prof. Anand K. Patel
Department of Information Technology
Dharmsinh Desai University
Nadiad, Gujarat, India
anandpatel.it@ddu.ac.in

2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) | 979-8-3315-2983-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/SCEECS64059.2025.10940184

***Abstract*—This paper presents TechAssist, an advanced artificial intelligence-driven chatbot designed to enhance the learning experience for students and professionals in tech domain. Utilizing the Mistral 7B Instruct model within a Retrieval-Augmented Generation (RAG) framework, TechAssist efficiently retrieves and generates contextually relevant answers to technical queries sourced from the StackOverflow dataset. Implemented with a user-friendly interface using Gradio, the system provides real-time responses, facilitating interactive engagement for users seeking technical support. Evaluation with learners and educators demonstrates the efficacy of TechAssist in delivering accurate and timely information, significantly improving the speed and quality of assistance. The findings indicate that the integration of cutting-edge AI technologies with an intuitive interface result in a more effective learning environment and enhanced user empowerment in addressing technical challenges.**

Keywords*— AI driven chatbot, technical question answering, Large Language Model, Mistral 7B Instruct, Retrieval-Augmented Generation, StackOverflow dataset, Gradio interface, technical support.

I. INTRODUCTION

The rapid advancement of technology has revolutionized the manner in which information is accessed and utilized across various domains. In technical education and industries, particularly in the tech domain, expeditious access to accurate and relevant information is critical for learning and problem-solving. Traditionally, technical resources such as StackOverflow have served as essential platforms, providing a wealth of community-driven insights and solutions. However, the task of manually searching for relevant information from vast resources can often be time-consuming and inefficient, especially for users seeking specific answers to technical questions.

To address this challenge, we propose TechAssist, a specialized question-answering chatbot designed to streamline the process of retrieving technical information. Unlike conventional search methodologies that rely heavily on keyword matching and manual exploration of multiple answers, TechAssist leverages advanced AI capabilities to deliver direct responses to user queries. By utilizing the Mistral 7B Instruct model within a Retrieval-Augmented Generation (RAG) framework, TechAssist efficiently synthesizes accurate answers derived from the StackOverflow dataset, enabling users to circumvent the arduous searching process. Furthermore, the chatbot provides related questions previously discussed on StackOverflow, allowing users to further explore pertinent topics and gain comprehensive insights expeditiously [1], [2].

The core functionality of the chatbot centers on its ability to comprehend and respond to user queries in natural language, emulating a conversational style that enhances user engagement. Moreover, the user interface, developed with Gradio, is constructed to provide a visually appealing and user-friendly experience. This paper explores the development and implementation of TechAssist, discussing its design principles, data sources, and the unique technical challenges addressed. Our results demonstrate the chatbot's potential as a robust solution for efficient knowledge retrieval in educational and professional settings, offering a more streamlined alternative to conventional search methods on technical forums.

Moreover, the development of TechAssist is grounded in a comprehensive examination of existing literature on Retrieval-Augmented Generation and chatbot technologies. Research has demonstrated that RAG frameworks significantly enhance the performance of language models in knowledge-intensive tasks, rendering them particularly suitable for domain-specific question answering [3]. Additionally, the integration of AI-driven chatbots into educational settings has been recognized as an effective method to facilitate learning, providing immediate support and resources to learners [5]. By synthesizing these insights, TechAssist aims to address the disparity between traditional information retrieval methods and the evolving needs of users in technical fields.

The remainder of this paper is organized as follows: Section II outlines the objectives and significance of TechAssist in enhancing technical knowledge retrieval; Section III delves into the theoretical background of the RAG framework and its role in chatbot applications; Section IV reviews related research on AI-driven question-answering systems and identifies gaps in existing approaches; Section V details the development methodology of TechAssist, including the integration of the Mistral 7B Instruct model; Section VI presents the experimental evaluation of the chatbot's performance; and Section VII concludes with insights and directions for future research in educational technology and AI support systems.

II. OBJECTIVES

The primary objectives of the TechAssist project focus on developing an AI-powered chatbot platform specifically designed to address the unique requirements of the technology domain. Through the utilization of advanced Retrieval-Augmented Generation (RAG) techniques and the Mistral 7B Instruct model, TechAssist aims to provide expeditious and accurate retrieval of information from the extensive repository of StackOverflow, a critical resource for technical knowledge. This chatbot endeavors to optimize information access for students, educators, and professionals in computer science and information technology by delivering direct, contextually relevant responses to user queries. In contrast to conventional keyword-based search methodologies, which frequently necessitate time-intensive manual exploration, TechAssist offers an efficient, user-centered approach, enabling users to circumvent redundant searches. Furthermore, the system presents related questions from StackOverflow, empowering users to explore specific technical topics in greater depth and acquire comprehensive insights.

Beyond information retrieval, TechAssist is engineered to enhance user engagement through a conversational interface constructed with Gradio, which facilitates an interactive and intuitive experience. This interface provides a visually appealing, accessible environment that encourages natural interaction, fostering an engaging learning and problem-solving process. By incorporating educational functionalities that support users in comprehending and applying technical concepts, TechAssist aligns with the continuous learning requirements of contemporary technology professionals and students. Through these objectives, TechAssist aspires to establish a new standard for AI-driven knowledge retrieval systems in the technology domain, paving the way for a future in which intelligent tools seamlessly integrate with community resources to meet the evolving demands of the industry.

III. THEORETICAL FRAMEWORK

The development of TechAssist, a specialized question-answering chatbot, utilizes advancements in large language models (LLMs) and retrieval-based artificial intelligence frameworks to facilitate accurate and efficient technical knowledge retrieval. The project implements the Retrieval-Augmented Generation (RAG) framework with the Mistral 7B Instruct model, which enhances the chatbot's capacity to comprehend technical questions, extract relevant information, and provide synthesized responses. This theoretical framework examines the underlying mechanisms, data sources, and technical methodologies that enable TechAssist to transform the traditionally manual search process into an interactive, AI-driven solution. By integrating conversational artificial intelligence with targeted knowledge retrieval, TechAssist addresses limitations in traditional Q&A platforms, reducing the time users expend manually navigating multiple answers on platforms such as StackOverflow and facilitating expedited access to reliable technical insights.

A. Dataset and Knowledge Base:

TechAssist is built on the StackSample dataset from Kaggle, a curated subset of the StackOverflow database. This dataset comprises over 10 million technical questions and answers, spanning various programming languages, frameworks, and computer science topics. To optimize knowledge retrieval, the dataset is pre-processed to standardize question and answer formats, remove extraneous HTML tags, and filter out irrelevant entries. This preprocessing ensures that the chatbot has access to high-quality, relevant information, enabling it to address a wide range of technical queries effectively. Additionally, using a dataset grounded in real-world problem-solving makes TechAssist a valuable resource for users seeking practical solutions to their programming challenges..

B. Integration of RAG Framework:

The foundation of TechAssist is the Retrieval-Augmented Generation (RAG) framework, which integrates retrieval mechanisms with generation capabilities to enhance the chatbot's accuracy and relevance. In this configuration, the chatbot initially identifies and retrieves pertinent answers and analogous questions from the StackSample dataset based on user queries. Subsequently, it synthesizes these results utilizing the Mistral 7B Instruct model, which generates precise, concise responses tailored to the user's specific inquiry. By integrating retrieval and generation, the RAG framework equips TechAssist with the capability to deliver accurate answers while maintaining a conversational tone. This methodology minimizes user effort in analyzing multiple answers and ensures that responses are both relevant and contextually appropriate.

C. NLU and Contextual Relevance:

TechAssist's natural language understanding (NLU) capabilities are essential for interpreting user queries and providing accurate responses. The Mistral 7B Instruct model is trained on extensive textual data, enabling it to parse and comprehend complex technical language utilized in computer science and software engineering. Through its instruction-based fine-tuning, the model demonstrates the capacity to process nuanced queries, accurately identifying user intent and retrieving responses that directly address the question. Furthermore, the chatbot presents related questions from StackOverflow to provide users with a broader perspective on similar issues, thereby facilitating a more comprehensive understanding of the topic.

D. Interactive User Interface and Accessibility:

TechAssist's user interface, developed using Gradio, is designed to enhance usability and accessibility. Users interact with the chatbot through a streamlined, visually intuitive interface that allows them to ask questions naturally and receive instant responses. The Gradio interface not only provides a conversational experience but also displays related questions, enabling users to explore similar topics and gain additional insights. This user-centric design ensures that technical information is presented in an accessible format, making TechAssist a valuable tool for students, developers, and professionals in technical fields seeking fast, reliable answers.

IV. REVIEW OF LITERATURE

Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) have exerted a substantial influence on the advancement of question-answering (QA) systems, particularly in technical domains that necessitate high precision and adaptability. Nevertheless, challenges persist in providing accurate and contextually relevant responses, especially for complex fields such as computer science education. This section examines extant methodologies, identifies their limitations, and elucidates how TechAssist, through its tailored RAG-based approach, endeavors to address these challenges and offer a more specialized solution for technical education.

A. Existing Approaches And Their Limitations:

*1) Domain-Specific RAG Models:*Domain-Specific RAG Models: RAG models combine the strengths of retrieval-based systems with generative models, improving response accuracy by fetching relevant documents and generating contextually enriched responses. This hybrid approach shows promise in enhancing answer quality by combining retrieval with generation. However, the effectiveness of RAG models heavily depends on the availability of well-curated, domain-specific datasets, which are often scarce or difficult to assemble. In highly specialized fields such as computer science education, accessing relevant data that meets the quality and specificity requirements can be a significant barrier to effective model performance. This reliance on carefully curated datasets limits the scope of RAG systems, particularly in dynamic or emerging technical domains where new information constantly evolves [1].

2) General LLMs for Knowledge-Intensive QA: General-purpose Large Language Models (LLMs), which have demonstrated robust performance across various natural language processing (NLP) tasks, serve as a foundation for numerous question-answering (QA) systems. While these models exhibit proficiency in processing a wide array of topics, they often prove inadequate for technical or domain-specific applications. Such models generate responses that tend to be overly generalized, failing to address the nuanced, technical details that are critical for specialized domains such as software development. In the context of technical education, general LLMs may provide rudimentary responses but frequently lack the requisite depth for complex, detail-oriented questions that students encounter in academic environments. This limitation significantly impedes their utility in contexts where accurate, specialized knowledge is essential for addressing technical queries [2].

3) Limitations in RAG for LLMs: Despite the advancements in Retrieval-Augmented Generation (RAG) for enhancing question-answering (QA) systems, significant limitations persist in their application, particularly with regard to large language models (LLMs). The hybrid approach of combining retrieval with generation exhibits certain strengths; however, its efficacy is substantially dependent on the quality of retrieval mechanisms. Suboptimal retrieval processes may result in ineffective or incomplete responses. Moreover, real-time retrieval remains a challenge in dynamic fields characterized by rapidly evolving information. The computational resources necessary to maintain real-time, accurate retrieval systems often render them impractical for large-scale applications, especially in resource-constrained educational environments [3].

4) Hybrid Models for QA Accuracy in Educational Contexts: Hybrid models, such as RALLE, integrate multiple retrieval sources and advanced generation mechanisms to provide more nuanced and contextually accurate responses. These frameworks capitalize on the strengths of diverse data sources, offering a more reliable way of answering complex queries. However, their complexity introduces several challenges. Ensuring that the retrieved data is consistent, accurate, and relevant requires significant effort in curating data sources. Furthermore, combining multiple retrieval mechanisms increases the computational complexity of the system, which can negatively affect response times and overall system efficiency. In educational settings, especially when answering technical queries, the need for real-time, accurate answers are critical, but hybrid models may struggle to balance retrieval precision with the computational resources required for effective, timely responses [4].

5) AI-Powered Learning in Graduate Engineering Education: The research examines the use of large language models (LLMs) and chatbots in supporting graduate engineering students' learning, focusing on the adaptability and engagement potential of AI in technical education. It highlights how LLMs enhance interactive learning environments by simulating tutoring and offering dynamic feedback. Despite these benefits, the study notes challenges such as adapting responses to high-level, discipline-specific queries and maintaining accuracy without human oversight. The findings suggest that refining LLMs for technical contexts can improve their effectiveness as educational tools [5].

B. TechAssist's Proposed Solution:

TechAssist proposes a domain-specific RAG-based approach to improve technical question answering, catering specifically to the needs of computer science and engineering students. By leveraging a dedicated retrieval mechanism that sources high-quality, curated data from platforms like Stack Overflow, TechAssist addresses the limitations of general-purpose LLMs and standard RAG systems. This approach ensures that responses are timely and contextually relevant, tailored to the unique demands of students in a fast-evolving field like software development.

The combined retrieval and generation components of TechAssist enhance the accuracy, relevance, and reliability of answers provided to users. The retrieval mechanism anchors responses in updated, specialized knowledge, while the generative model synthesizes this data into clear and concise answers. This design positions TechAssist as a valuable real-time learning tool, providing students with reliable, domain-specific responses to complex questions, thus improving their learning experience and reducing time spent on information retrieval.

V. METHODOLOGY

The proposed system is a question-answering chatbot designed to handle technical queries, particularly focused on students and professionals in computer science and IT fields. This chatbot uses a Retrieval-Augmented Generation (RAG) approach, supported by embeddings and the Mistral language model, to provide precise, context-aware answers. The following sections outline the process of data acquisition, retrieval, response generation, and user interface design to create a seamless and user-friendly experience.

A. Data Acquisition and Preprocessing:

1) Dataset Collection and Structuring: The chatbot utilizes the StackSample dataset from StackOverflow, a large repository of technical questions and answers. This dataset, containing fields like question IDs, titles, bodies, and corresponding answers, is organized to ensure consistency and quality in retrieved answers. Any irrelevant content, such as HTML tags or non-technical text, is removed to focus on essential technical details.

2) Embedding Generation: Each question-answer pair in the dataset is transformed into a dense vector representation, or embedding, capturing the semantic meaning of the content. These embeddings facilitate similarity-based searches, enabling the system to match user queries to relevant stored data. FAISS (Facebook AI Similarity Search) is used to store and manage these embeddings, allowing rapid and efficient retrieval during user interactions.

B. Retrieval-Augmented Generation (RAG) Pipeline:

1) Similarity Search in FAISS Database: The first part of the system focuses on predicting stock price movements using technical indicators derived from historical stock data, for NIFTY 50 and its constituent stocks. This data is retrieved using the Yahoo Finance API and processed to calculate several important technical indicators that traders typically use to forecast future stock movement.

2) Contextual Retrieval and Re-ranking: If an exact answer is not available, the RAG pipeline provides contextually similar answers. Retrieved answers are re-ranked based on relevance to ensure the most suitable responses are prioritized. This capability allows the chatbot to handle complex or uncommon queries by suggesting related answers that could be helpful.

C. Language Model (Mistral):

The Mistral language model is used to generate responses based on the relevant content retrieved by the RAG (Retrieval-Augmented Generation) pipeline. When a user submits a query, Mistral processes the retrieved information and formulates a clear, concise answer. The model's pre-trained capabilities allow it to understand technical terminology and respond conversationally, making complex information accessible to users. This setup enables the chatbot to provide accurate, domain-specific answers without requiring additional fine-tuning of the Mistral model.

D. Answer Formatting and Post-Processing:

1) Structured Response Output: Following the language model's generation of a response, the output is formatted to enhance readability and facilitate navigation. Explanations and code snippets are clearly demarcated, with key points organized in a logical structure, such as through the utilization of headers or bullet points where appropriate. This methodology enhances clarity, enabling technical users to efficiently comprehend complex information.

2) Integration with StackOverflow: To provide users with additional insights, the system attempts to find similar questions on StackOverflow by matching keywords from the user's query. If users seek further details or are not fully satisfied with the generated response, links to relevant StackOverflow questions and answers are offered. This feature empowers users to explore deeper or alternative explanations directly on StackOverflow, ensuring they have access to a broad range of information for more comprehensive understanding.

3) Error Handling and Feedback Mechanism: In the event that the system is unable to identify an appropriate response or relevant StackOverflow matches, it provides feedback by suggesting that users reformulate their question or refine their input. This mechanism ensures that users receive guidance when their query is excessively ambiguous or complex, thereby enhancing their experience by assisting them in reformulating their question in a manner that the system can more effectively comprehend and address.

E. System Architecture and Implementation:

1) Gradio Interface: The chatbot is hosted on a user-friendly web interface built with Gradio. This interface mimics a conversational environment, allowing users to ask questions, view responses, and access a history of previous interactions. The design emphasizes usability and aesthetic appeal, making the chatbot engaging and easy to navigate.

2) Data Flow and Workflow Orchestration:

- User Input and Query Processing: User queries are processed through LangChain, which standardizes the input for efficient retrieval and compatibility with the RAG pipeline.
- RAG Pipeline: The RAG pipeline retrieves similar question-answer pairs from the FAISS database, based on similarity to the user query.
- LLM (Mistral): The Mistral model refines the retrieved content to generate a coherent response.
- Answer Formatting: The response is formatted for clarity and displayed in the Gradio interface.

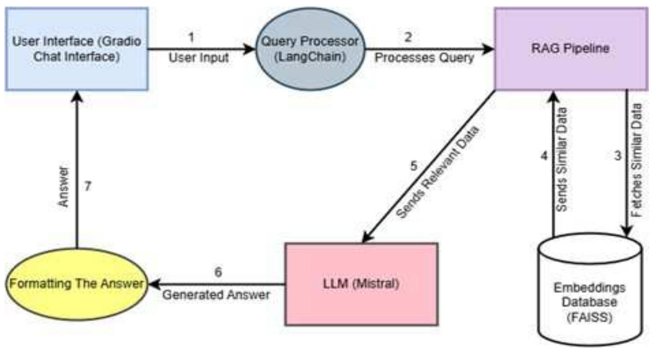

Figure 1. Architecture Diagram of the TechAssist

By integrating the RAG approach with the Mistral model, the chatbot provides users with accurate, high-quality responses tailored to technical inquiries. This methodology enables the chatbot to address a wide range of technical questions effectively, making it a valuable tool for computer science and IT users. TechAssist effectively bridges the gap between traditional methodologies and the requirement for specialized, on-demand knowledge in the contemporary, rapidly evolving academic environment.

VI. EXPERIMENTS AND RESULTS ON TECHASSIST

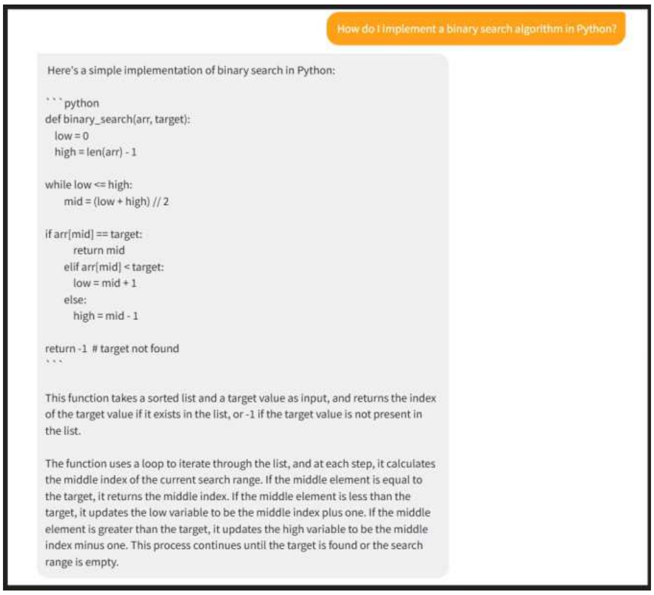

Figure 2. Automated Guidance for Binary Search in Python

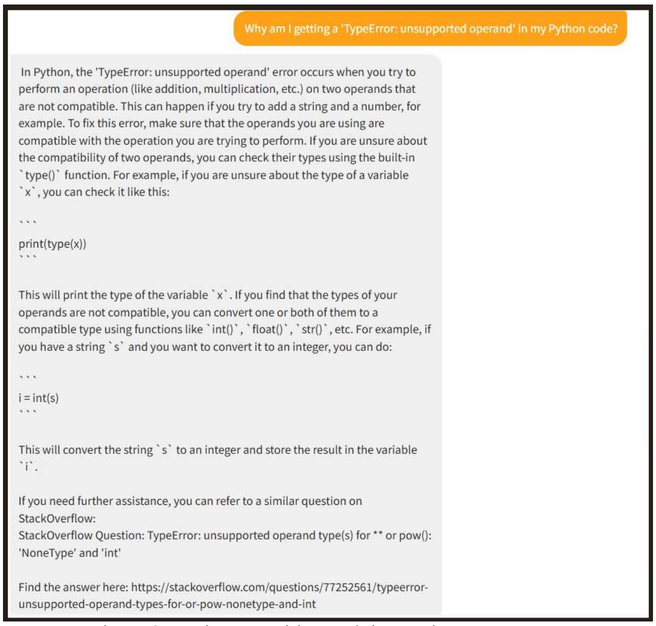

Figure 3. Assistance with Resolving Python Type Error

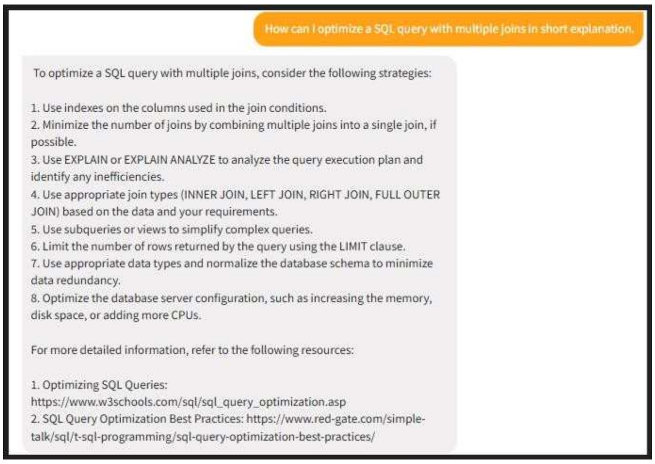

Figure 4. SQL Query Optimization Techniques

VII. COMPARATIVE ANALYSIS

Feature	StackOverflow Manual Search	OverflowAI (StackPlusOne)	TechAssist
Search Method	Keyword-based	AI-enhanced search	RAG + LLM based retrieval
User Input	Specific keywords required	Natural language with keywords	Natural language questions
Response Format	Static, multiple answers per query	Dynamic summaries, AI-generated answers	Dynamic, synthesized, conversational answers
Search Time	5–10 minutes	0–2 minutes	0–2 minutes
Answer Accuracy	Moderate, depends on user query formulation	High accuracy due to AI contextual understanding	Very high accuracy due to domain specialization
User Effort	High (manual filtering)	Low	Low
Accessibility	Open and free for all	Paid subscription required	Open and free for all

Figure 5. Comparative Analysis of Methods for Accessing Technical Information from StackOverflow

Figure 5 presents a comparative analysis of three systems designed to access technical information from StackOverflow: StackOverflow Manual Search, OverflowAI (StackPlusOne), and TechAssist. This analysis evaluates key performance metrics, including search methods, response formats, answer accuracy, search time, user effort, and accessibility, to determine the efficacy of each approach.

StackOverflow Manual Search utilizes a keyword-based approach, necessitating users to input specific terms and manually navigate through threads to locate relevant answers. While it is open and free for all users, the reliance on static, user-contributed content and the absence of dynamic aggregation of information render it time-consuming and susceptible to inaccuracies, particularly for complex queries. This traditional approach, while foundational, often requires users to sift through irrelevant or outdated answers.

OverflowAI (StackPlusOne) enhances the traditional StackOverflow experience by leveraging artificial intelligence for context-aware, dynamic summaries of content. Utilizing natural language processing (NLP) techniques, OverflowAI facilitates more efficient retrieval of relevant answers, reducing search time and improving the accuracy of responses. However, its utility is constrained by the requirement for a paid subscription, limiting accessibility for users who may not have access to premium services. OverflowAI's strength lies in its capability to synthesize AI-driven answers and provide an enhanced experience for technical users.

TechAssist, as demonstrated in the research paper, surpasses both approaches by integrating the Retrieval-Augmented Generation (RAG) framework with the Mistral 7B language model. This enables TechAssist to process natural language queries with precision, delivering real-time, context-aware, and synthesized conversational answers tailored to technical domains. TechAssist's design further enhances its relevance for computer science and engineering by leveraging curated StackOverflow datasets, ensuring high accuracy and precision in responses tailored to these fields. Additionally, TechAssist is freely accessible for users in the technical domain, democratizing access to advanced AI capabilities.

Key Advantages of TechAssist:

1. Real-Time Accuracy: With domain-specific optimization, TechAssist achieves very high accuracy in technical answers, delivering results that meet the specific requirements of users.

2. Enhanced User Experience: TechAssist's interactive Gradio-based interface ensures intuitive use, reducing user effort while maintaining seamless query resolution.

3. Accessibility and Affordability: OverflowAI requires integration with platforms such as Microsoft Teams and operates under a paid subscription model, which can limit accessibility for some users. In contrast, TechAssist offers its advanced features as a standalone tool that is freely available to users in the technical domain, making it an optimal choice for professionals and students alike.

VIII. CONCLUSION

TechAssist represents a progressive step in technical question-answering for computer science and engineering education by using a specialized RAG-based model paired with the Mistral language model. This tailored system effectively combines real-time retrieval from curated sources like Stack Overflow with precise generation capabilities, enabling detailed, contextually relevant responses. By addressing the limitations of traditional QA systems in handling specialized, evolving knowledge, TechAssist provides an adaptable and dynamic support tool for students tackling complex, domain-specific queries.

In bridging the gap between general-purpose language models and the specialized demands of technical education, TechAssist offers an invaluable resource that enhances both learning efficiency and comprehension. This approach not only supports students' immediate academic needs but also aligns with the dynamic nature of technology, providing a reliable tool that adapts as the field evolves. As technical education continues to grow in complexity, solutions like TechAssist stand to redefine educational support, merging structured retrieval with generation for an optimized learning experience.

REFERENCES

[1] D. Y. Liu, S. R. Soni, and R. T. N. Cardoso, "Retrieval-Augmented Generation for Domain-Specific Question Answering," arXiv preprint arXiv:2404.14760, 2024. [Online]. Available: https://arxiv.org/abs/2404.14760

[2] R. K. P. S. Dos Santos, P. V. A. Silva, and C. L. Reis, "RAG: Retrieval-Augmented Generation for Knowledge-Intensive Tasks," arXiv preprint arXiv:2005.11401, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[3] A. J. Gupta, A. L. Toms, and M. Z. Asgarian, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

[4] J. C. Singh, S. P. Ghosh, and M. M. Gupta, "RALLE: A Framework for Developing and Evaluating Retrieval-Augmented Large Language Models," arXiv preprint arXiv:2308.10633, 2023. [Online]. Available: https://arxiv.org/abs/2308.10633

[5] J. B. Ahmed, M. F. S. Ali, and A. T. Rayhan, "Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education," ResearchGate, 2023. [Online].Available: https://www.researchgate.net/publication/385279904_Beyond_Traditional_Teaching_The_Potential_of_Large_Language_Models_and_Chatbots_in_Graduate_Engineering_Education

[6] M. Thway, J. Recatala-Gomez, F. S. Lim, K. Hippalgaonkar, and L. W. T. Ng, "Battling Botpoop using GenAI for Higher Education: A Study of a Retrieval Augmented Generation Chatbot's Impact on Learning," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2406.07796

[7] C. Dong, "How to Build an AI Tutor that Can Adapt to Any Course and Provide Accurate Answers Using Large Language Model and Retrieval-Augmented Generation," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2311.17696

[8] A. Gupta and B. E. Varghese, "Best Practices for Using Retrieval Augmented Generation (RAG) in AI Chatbots," SearchUnify Technical Blogs, 2024. [Online]. Available: https://www.searchunify.com/sudo-technical-blogs/best-practices-for-using-retrieval-augmented-generation-rag-in-ai-chatbots/

[9] R. Akkiraju, A. Xu, D. Bora, T. Yu, L. An, V. Seth, A. Shukla, P. Gundecha, H. Mehta, A. Jha, P. Raj, A. Balasubramanian, M. Maram, G. Muthusamy, S. K. Annepally, S. Knowles, M. Du, N. Burnett, S. Javiya, A. Marannan, M. Kumari, S. Jha, E. Dereszenski, A. Chakraborty, S. Ranjan, A. Terfai, A. Surya, T. Mercer, V. K. Thanigachalam, T. Bar, S. Krishnan, S. Kilaru, J. Jaksic, N. Algarici, J. Liberman, J. Conway, S. Nayyar, and J. Boitano, "FACTS About Building Retrieval Augmented Generation-based Chatbots," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2407.07858

[10] C. Antico, S. Giordano, C. Koyuturk, and D. Ognibene, "Unimib Assistant: Designing a Student-Friendly RAG-Based Chatbot for All Their Needs," arXiv, 2024. [Online]. Available: https://arxiv.org/pdf/2411.19554