Domain Specific Chatbot Using RAG and Fine Tuning

1st Mudumba Sri Akshit Sri Vastav *Department of Information Technology, Vardhaman College of Engineering (Autonomous),*
Hyderabad-Telangana, 501286, India
mudumbasriakshitsrivastav22it@student.vardhaman.org

2nd Talla Ganesh Goud
Department of Information Technology, Vardhaman College of Engineering (Autonomous),
Hyderabad-Telangana, 501286, India
goudganesh790@gmail.com

3rd Vabhanagiri Ujwal
Department of Information Technology, Vardhaman College of Engineering (Autonomous),
Hyderabad-Telangana, 501286, India
samuelsam00402@gmail.com

4th Mrs. Yadla Sunanda
Department of Information Technology, Vardhaman College of Engineering (Autonomous),
Hyderabad-Telangana, 501286, India
sunanda@vardhaman.org

***Abstract*—The objective of this project is to create a domain-specific assistant chatbot by combining fine-tuning and retrieval-augmented generation (RAG), enhancing the accuracy and contextual understanding of conversational interactions. Fine-tuning a pre-trained model would allow for the exposure to domain-specific language so that the model can provide more relevant and accurate responses. With the addition of RAG, the chatbot system will be capable of real-time retrieval of external knowledge for its responses, which will make it more accurate and dynamic. The hybrid approach will impart a higher degree of complexity to the queries handled by the chatbot system, thus qualifying the chat interface to support technical and nontechnical users alike. The vector database will ensure that the chatbot remembers previous interactions, which will assist in making more consistent and informed responses over time. The system also employs natural language processing capabilities to gauge and analyze user intents, resulting in conversational and intuitive interaction styles. Scalable and adaptable, the assistant chatbot is then poised for industry-level applications, spanning data intelligence and customer support.**

***Index Terms*—Retrieval augmented generation (RAG), Fine-tuning, Vector database, Large Language Models (LLMs), Small and Medium Enterprises (SMEs).**

I. Introduction

In today's fast-changing data-science world, having rapid and precise information access is essential to survive competition. The ability to have instant access to an accurate answer can form an incredible basis for decision-making and problem-solving. This is where a domain-specific assistant chat bot shines: It has perceptive and real-time support through fine-tuning and RAG. Fine-tuning provides the ability to comprehend data science's intricate terminologies and terminologies so that the chatbot can give accurate and apt answers, whereas RAG helps the chatbot make external information available on real-time request to further boost the depth and precision of its replies. In such a scenario, users can tackle tasks such as interpreting algorithms, choosing the best preprocessing technique, or troubleshooting their models for a successful outcome. Moreover, the chat bot can greatly shorten the work process requiring excessive time for the casual user. This includes making research subjects easier to understand, rendering them adequately in layman's terms, and guiding users through analysis-related hurdles. Thus, it creates a platform for helping its user to become productive, solve workflow bottlenecks, and convert raw data into actionable insights.

Identify applicable funding agency here. If none, delete this.

A. Motivation

The impetus for this project stems from the much-needed ability of data scientists to access reliable domain-specific knowledge traditional chatbots offer inaccurate responses to technical questions simply because they lack proper context, therefore, the project will combine rag with fine-tuning to build a smart assistant retrieving real-time relevant information while also adapting to the intricacy and evolving nature of data science concepts a vectorized database will make sure that the chatbot is accumulating past interactions thus also enhancing its intelligence and the efficiency of its task

B. Contribution

this field-specific chatbot is going to be of immense use in data science education and business analytics as it shall help in algorithm selection preprocessing methods and debugging in data science thus improving efficiency in working on educational projects whereupon it shall act as a virtual tutor explaining complex concepts in extremely detailed terminology and answering student queries related to coding whereas in the case of business analytics real-time insights will be made possible by getting to relevant data and responding accordingly-this will be made sure since through the vectorized nature of a database knowledge will be retained thus ensuring accurate consistent and dynamic responses to suit time

II. Literature Review

Bhavesh Patel, B.B Chand's RAG-based Chatbot for Vikram Sarabhai Library, has been developed as a part of the IEEE paper author project. The main aim of the chatbot is to provide real-time, accurate and personalized responses to user queries in the context of library services at this Indian Institute of Management. Thus, it will revolutionize library services for users. The chatbot will try to include Large Language Models (LLMs) but also apply a custom knowledge base from libraries' materials, like service manuals, FAQs, or database PDFs, producing actual, accurate, and contextually relevant information for the user. The primary objective involves improving user satisfaction, becoming more productive in information retrieval, and maintaining an interactive library experience. In contrast, my project is specially designed for the data science domain. In particular, it instills real-time assistance to professionals and students who specialize in complex data science tasks. The library chatbot deals mostly with general information retrieval and academic support; on the other hand, mine intends to assist its users with the interpretation of algorithms, preprocessing techniques to be utilized, and troubleshooting models, hence the specialization reference within data science workflows. Furthermore, this project is also intended to pay back research time while simplifying highly complex concepts, in addition to guiding through challenges faced in data analysis-turning forth actionable insight. Both projects leverage RAG and fine-tune them to make it better in accuracy and relevance, but by aiming at different domain users-academic library patrons-and data science practitioners, it shows that it can be used in any domain [1].

Rohit Tamrakar, Niraj Wan's IEEE project has been on designing and developing chatbots intended to supplement and supplement the learning process, particularly from the standpoint of Computer-Aided Design (CAD) applications. Their review will take an insight from the earliest historical origins of chatbots through their architecture and classification, design techniques, or platforms used for their actual development, focusing on how they help users in overcoming problems in procedural-based learning when using CAD software. Moreover, the paper highlights the artificial intelligence (AI) and machine learning (ML) perspectives adopted to develop a chatbot and communicate with a user through text or speech to provide solutions for certain CAD-application-specific problems. On the other hand, my project deals with developing a domain-specific chatbot for the data science field, which would be capable of assisting a user in interpreting complicated algorithms, choosing preprocessing techniques, and debugging models. If both projects involve using AI and ML, my chatbot is intended to provide real-time support for data science tasks, an exclusive product for professionals and students in this field. Not only that, but my project also focuses on fine-tuning and Retrieval-Augmented Generation (RAG) so that the chatbot can respond correctly and in a relevant-to-context manner by bringing real-time knowledge from external sources into the conversation. This will allow the chatbot to handle subtle queries and give actionable insights. In contrast, the IEEE paper's chatbot focuses on procedural help in CAD. The projects have been aimed at increasing user productivity and capability to solve problems; however, they are domain-specific: CAD and data science, respectively. The development shows the versatility for which chatbots could be employed [2].

The IEEE paper project, written by Aishwarya Bhosale, Dr. Vijay G. Gaikwad, and Sachin Deshpande, is an analytical study of the utility, processes, and significance of chatbots in business organizations, particularly commerce such as sales, customer service, marketing, and operations. The emphasis of the paper is on the ability of chatbots to automate tasks-and thereby enhance customer engagement and provide personalized experiences-primarily in areas such as e-commerce and CRM. The historical evolution of chatbots is discussed, along with their architecture and platforms for building chatbots, with particular emphasis on business application areas such as automating customer interactions, improving marketing strategies, and reducing operational costs. While the focus of my project is on developing a domain-specific chatbot for data sciences, which aims to assist users in interpreting complex algorithms, selecting preprocessing techniques, and troubleshooting models, the IEEE paper has an emphasis on the use of chatbots for business process automation and customer interaction. While the IEEE paper looks at using chatbots for business automation and customer interaction, my project is being developed as a real-time tool for providing technical assistance in data science tasks and is designed to ensure correct and contextually relevant answers through fine-tuning and Retrieval-Augmented Generation (RAG) techniques. My chatbots will also optimize workflows and shorten data-to-action research relatively, whereas the IEEE paper chatbots focus more on aiding business customer experience and operational efficiency. While both projects use AI and ML, they focus on their respective automation of business versus data science, demonstrating the versatility of the technology behind chatbots in different sectors [3].

This ie paper, written by Gooty Joshi, Naga Venkata Akhileshu Yadav, and Jaswant Hanuman, is primarily focused on the design of a chatbot system for college inquiries based on a knowledge database, which would provide information for students, parents, and staff regarding college-related activities such as admissions, course details, facilities on campus, and academic calendar. This chatbot uses natural processing algorithms like NLP and AI in interpreting user queries and generating accurate responses with better communication efficiency, thereby reducing the burden on college authorities; whereas, my project focuses on building a domain-specific chatbot in the form of the data science field, meant to assist users in understanding complex algorithms, choosing preprocessing techniques, and troubleshooting models. While both projects hinge on AI and NLP technologies, mine concentrates on

giving real-time assistance with data science tasks and building chatbots that provide responses fine-tuned and supplemented with external information through retrieval-augmented generation (RAG) activity to ensure accuracy and contextual relevance .The other aims to streamline workflow, reduce research time, and convert raw data into actionable insights, while the ie paper's chatbot focuses on basic college inquiries and administrative efficiency. Both projects demonstrate the versatility of chatbot applications, simply serving two different domains: college administration versus data science, which only showcases how adaptable chatbot technology is across domains [4].

The IEEE paper by bayan abu shawar and eric atwell presents a chat-generating corpus system that turns the bnc text dataset into conversational dialogues their chatbot employs aiml artificial intelligence markup language within a pattern-matching paradigm that defines it as a static system whose responses are generated based entirely on predefined corpus data this constrains the chatbot to merely expressing the application of language and knowledge exhibited by the corpus hence it is better defined as a tool for visualizing text rather than a dynamic interactive assistant conversely your project uses rag and fine-tuning to develop an advanced and domain-specific chatbot that can provide real-time contextually relevant responses to queries what fine-tuning does is to enable the chatbot to learn the lingo of data science while rag enhances the correctness and depth of its responses by retrieving knowledge from external sources vectorizing the database allows your chatbot to recall previous chats to give consistent contextually flexible answers that are related to your current inquiries your chatbot would then give consistent contextually aware answers your system is much practical for real-life implementation unlike the ieee chatbot that uses static dialogue rules pre-trained for single-time use since it can be scaled adapted and deal with changing and complex queries in the realm of data science [5].

As an IEEE paper and research work of Bii Patrick Kiptonui, the use of chatbot technology was explored by the author to improve the learning of students in constructivism. The chatbot, known as knowie, is to help students learn and develop 21st century skills such as problem-solving, collaboration, and critical thinking. With the application of AIML based pattern matching, responses are generated based on predefined templates and hence it is a fully static and parametrically limited system without much adaptability. Knowie focuses purely on enhancing education and operates within self-contained systems, thus reaching schools that do not have access to even the most rudimentary internet facilities in developing world regions. Unlike the chatbots built for providing interactive support through specific subject areas for improving learning outcomes, it introduces its domain-specific chatbot using RAG and employing fine-tuning which provides real-time context-aware on-data modeling tasks assistance for the enhancement of learning. Fine-tuning which allows your chatbot to specialize in domain-specific terminology, builds up a better accuracy and relevance of response. RAG dynamically retrieves information externally enabling up-to-date and precise answers that are provided by the chatbot. Your architecture makes use of a vectorized database to record previous interactions for more consistent and contextually rich responses over time. Unlike the static pre-trained AIML-based chatbot in the IEEE paper, your system is highly adaptable and able to handle extremely complicated queries that evolve, making it fit for real-world data science applications while allowing greater scalability and efficiency [6].

Pharmabot is an artificial intelligence initiative conceptualized by Benilda Eleonor V. Comendador and team members, which is meant to serve as a consultant on pediatric generic medicine. It is meant to provide the parents and patients with relevant medication information in prescribing it to their children. The chatbot receives user queries through Left-Right Parsing algorithm to a static database with predefined information about different medicines. As a result, Pharmabot is efficient in a specific, rule-based undertaking but, certainly, restricts itself from current adaptation to new or changing medical knowledge. Unlike this project, my focus is domain-specific chatbot which uses RAG and fine-tuning for data science. While Pharmabot is static and rule-based, my chatbot retrieves the augment generation dynamically to fetch real-time, external information for more accuracy and context-rich responses. Further fine-tuning on domain-specific data will better equip the model to understand and respond effectively to technical queries. Besides, my human-like chatbot has a vectorized database tool integrated into it, which will enable it to remember previous interactions and give consistent informed responses. Hence, it would likely be a more adaptive, scalable, and versatile solution than the predefined static consultation system of Pharmabot [7].

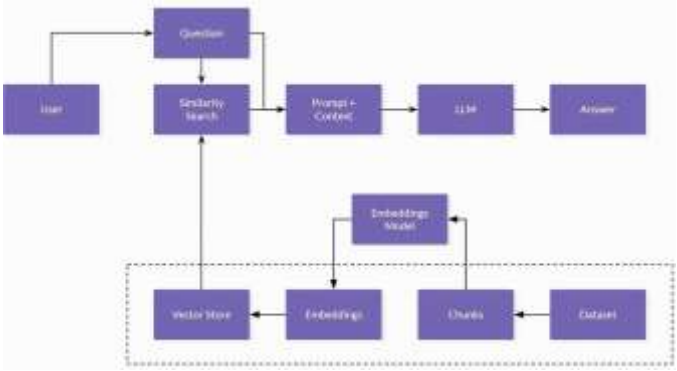
Nabeel Zaim and Vinh Duong's work involves creating a conversational chatbot trained on raw Twitter data with seq2seq and RNN models. It is an open-domain conversation generator, producing relaxed, human-like responses from large-scale, unfiltered tweet datasets. Although suitable for general dialogue, it does not have contextual accuracy or domain knowledge. My work, however, is a domain-specific data science chatbot using retrieval-augmented generation (RAG) and fine-tuning. In contrast to the Twitter chatbot, which gives generic responses, my system retrieves dynamic, external, real-time information, hence providing more accurate and relevant responses. Fine-tuning also makes it better equipped to interpret complicated technical jargon, thus making it very efficient in handling data science questions. My chatbot also incorporates a vectorized database, allowing it to store previous interactions, thereby ensuring consistency and continuity in responses. This makes my chatbot more stable, responsive, and accurate for specific tasks, while the Twitter chatbot is more ideal for general, unspecified conversations [8].

III. PROPOSED METHODOLOGY

A. Data Collection and Preprocessing:

The methodology proposed for the Domain-Specific Intelligent Chatbot for SMEs Using Fine-Tuned LLMs with RAG is a multi-step process to provide precise, contextually aware answers. It starts from collecting and preprocessing data, in which the system retrieves text from PDFs uploaded by users, cleans and tokenizes text, and creates vector embeddings utilizing pre-trained transformer models. The embeddings are preserved in a vector database so that relevant information can be retrieved efficiently. The chatbot leverages a large language model (LLM) which has been fine-tuned using domain-specific data from the PDFs to improve its context understanding. By implementing Retrieval-Augmented Generation (RAG), the system both retrieves and generates content. In the retrieval part, it queries the vector database to retrieve relevant document snippets. In the generation, the fine-tuned LLM generates accurate, real-time responses using the retrieved context and the user query. The system is developed with a MERN stack having Flask responsible for the backend processing, such as PDF uploading, data retrieval, and vector database communication, whereas the frontend utilizes React.js for an interactive interface. The whole system is hosted on a cloud server so that it is accessible by SMEs. Testing and evaluation are conducted by quantifying the response accuracy, retrieval efficiency, and users' satisfaction to check whether the chatbot fulfills the domain requirements or not. This approach guarantees the chatbot provides accurate, context-aware, and real-time responses, thus qualifying as a very influential tool for automating customer interactions and quickening information acquisition for SMEs.

B. Architecture Design:



IV. CONCLUSION AND FUTURE WORK

The Domain-Specific Intelligent Chatbot for SMEs With Fine-Tuned LLMs and RAG successfully illustrates how large language models can be adapted to produce context-sensitive, real-time responses by utilizing fine-tuning and retrieval-augmented generation. The chatbot is able to properly comprehend and process domain-specific jargon through fine-tuning, while RAG makes it retrieve proper, external knowledge in order to provide accurate answers. Integration with a vectorized database also supports the efficiency of the chatbot further by making fast and contextualized retrieval possible. The solution is able to successfully automate information retrieval, increase customer interaction, and minimize manual data lookup times for SMEs. In the future, the project can be improved by adding multi-modal data support (processing images, sound, and video), fine-tuning with improved few-shot or zero-shot learning, and adding long-term memory retention to enhance contextual consistency. Furthermore, multilingual support and offering the system as a Software-as-a-Service (SaaS) will make it more scalable and accessible, positioning it as an even stronger tool for SMEs.

REFERENCES

[1] Elliott, W. S. "Computer-Aided Mechanical Engineering: 1958 to 1988." Computer-Aided Design, vol. 21, no. 5, 1989, pp. 275–88.
[2] Ranoliya, Bhavika R., Nidhi Raghuwanshi, and Sanjay Singh. "Chatbot for university related FAQs." 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). IEEE, pp. 1525–30, 2017.
[3] Rosruen, Nudtaporn, and Taweesak Samanchuen. "Chatbot utilization for medical consultant system." 2018 3rd technology innovation management and engineering science international conference (TIMES-iCON). IEEE, pp. 1-5, 2018
[4] Bhavesh Patel, Senior Library Professional - IT Applications, Vikram Sarabhai Library, IIMA, Ahmedbad, Gujarat Email: bhaveshiima@gmail.com Dr. B B Chand, Librarian Head NICMAN, Vikram Sarabhai Library, IIMA, Ahmedabad, Gujarat Email: bbchand@iima.ac.in
[5] Aishwarya Bhosale1, Dr. Vijay Gaikwad2, Sachin Deshpande3 1B.Tech student, Vishwakarma Institute of Technology, Pune, Maharashtra, IndiaInternational Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT — Volume 2 — Issue 6 — ISSN : 2456-3307
[6] GOOTY JOSHI NAGA VENKATA AKHILESH YADAV (REG NO: 39110343) YASWANTH HANUMANTHU (REG NO: 39110365)
[7] Rohit Tamrakar Sardar Vallabhbhai National Institute of Technology Surat, Niraj Wani,Sardar Vallabhbhai National Institute of Technology Surat.
[8] Benilda Eleonor V. Comendador, Bien Michael B. Francisco, Jefferson S. Medenilla, Sharleen Mae T. Nacion, and Timothy Bryle E. Serac Polytechnic University of the Philippines, Manila, Philippines
[9] Abu Shawar, BA and Atwell, ES (2005) A chatbot system as a tool to animate a corpus. ICAME Journal: International Computer Archive of Modern and Medieval English Journal, 29. 5 - 24. ISSN 0801-5775
[10] Mukherjee, Subrata (2024, February 11). Retrieval-Augmented Generation (RAG) and explore how it can address the pain points we have in the context of LLM-based applications