INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 9.47]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 12, Issue - 4, April - 2026

DOIs:10.2015/IJIRMF/202604027 --:-- Research Paper / Article / Review

IndiraGPT: College Website Chatbot System

[1] Prasad Munde, [2] Vaibhav Kulkarni, [3] Prof. Monika Patil,
[1]Student, Department of Artificial intelligence and Data science Engineering, Indira College of Engineering, Pune, India
[2] Student, Department of Artificial intelligence and Data science Engineering, Indira College of Engineering, Pune, India
[3] Professor, Department of Artificial intelligence and Data science Engineering, Indira College of Engineering, Pune, India

Email – [1] prasadmunde999@gmail.com, [2]vaiku2656@gmail.com, [3]monikapatil@indiraicem.ac.in

Abstract: *Contemporary educational institutions have difficulties in providing direct and reliable access to information over the use of conventional Internet interfaces. In review, IndiraGPT is proposed, which makes use of Retrieval Augmentation and Generative Models for effective data dissemination in academic institutions. The framework makes it possible to retrieve dependent facts and produce responses based on the context, thus ensuring accurate and well-timed communication. The proposed architecture enables nonstop information retrieval, multi-flip dialogs, and query understanding. Empirical results indicate higher accuracy, less effort-intensive tasks, and extensive user interaction compared to conventional frameworks.*

Key Words: *Chatbot System, Retrieval Augmented Generation, Natural Language Processing, Large Language Model, Semantic Retrieval.*

1. INTRODUCTION:

As a result of the fast integration of virtual technology into training, strategies of conversation among companies and college students have passed through widespread changes [1]. Although the passage of time has undoubtedly influenced education, a few departments use earlier fact systems on their websites that now do not meet the desires of contemporary customers. In this type of case, humans would possibly spend extra time to locate the required data, thus reducing the usual character delight.

A full-size disadvantage of traditional systems lies in the fact that records cannot be tailored to the needs of each person. Regardless of the motive of the visit, all users will stick to the equal path through websites, which brings additional difficulties and reduces the extent of involvement. This disadvantage was mentioned in many studies regarding common data systems in academic environments [1].

The development of artificial intelligence has led to the emergence of promising solutions. In particular, customers can now have interactions with structures the use of interviews, which makes the procedure more intuitive. The rapid improvement of herbal medicine and conversational stores allows better information about user requests [2]. At the same time, it could also appear as a formidable task to provide a correct reaction to inquiries if reliable statistics are not available.

To overcome these risks, one should introduce a system primarily based on a combination of retrieval and temporal models. For example, a brand new tool called IndiraGPT can be created with the use of large language models and retrieval-enhanced technology strategies. The proposed system can be useful in providing accurate solutions primarily based on the actual statistics available at the institution.

The purpose of this paper is to extend an architecture that would not only address the desires of query processing but could improve the complete virtual experience that the user has while accessing such platforms. The use of scalable query processing structure mixed with efficient information management practices can be considered right here [4].

2. LITERATURE REVIEW:

In popular, the primary systems that used chatbots have been built on a directive based on rule-based total common sense. The reaction immediately became associated with some unique sample, which meant that if-else selections have been used [1]. The disadvantage of such a technique was that it was not able to generalize and expect responses to diverse requests of different nature. The slightest alternative in a person's request may result in an incorrect response because there was no known policy on interaction between a bot and a user. Such systems are fairly handy while answering repetitive questions and are highly constrained.

Further development led to a situation when computer linguistics would give chatbots greater superiority. The idea of developing structures based entirely on NLP techniques became considerable within the discipline [2]. The techniques included tokenization, part-of-speech analysis, and intents recognition among others. Furthermore, gadget learning is performed, making NLP-based totally chatbots perform much better than those of the previous generation.

Nevertheless, as mentioned in [2], NLP systems are proficient on large volumes of CDs and thus will have a whole lot of time to develop and become useful. Moreover, such structures may not recognize certain distinct sentences.
FAQ solutions proved to be very useful due to their simplicity. They allowed the chatbot to answer commonly asked questions, however, they did not adapt to customer preferences. Every time a few new statistics or updates on agency coverage appeared, the developers would have been responsible for repairing the machine. Furthermore, as stated in [1], FAQ chatbots could not answer any question that becomes not previously considered and defined, due to this that they would have been very inflexible and static in nature.

Recently, there was an awful lot of communicating around Large Language Models as a new trend in chatbot improvement. Such modes allow one to create an interactive response capable of informing a user's request in its context, finding a solution, and producing it. The niceness of the generated text is particularly elevated, making such bots much friendlier and less difficult to interact with. Despite the benefits, there is a crucial drawback: An LLM would possibly produce false data or hallucinate something. As stated in [2], this problem is particularly pressing for academic structures.

In response to these challenges, Retrieval Augmented Generation has been mounted as a progressive technology that leverages the benefits of information retrieval and generative modes. In this approach, record retrieval methods are implemented first to retrieve contextual records from either substantiated or unstructured expertise bases. The retrieved information is then fed into the generative model, with the latter producing correct and contextually aware responses to queries. As mentioned in [3], combining those techniques provides a mechanism to ensure the accuracy and consistency of the generated answers, which are primarily based on established facts.

The proposed response uses this concept so one can use the detailed institutional expertise in the form of statistics to be included on college websites and other applicable documents. With such a method, the information retrieval procedure can offer the chatbot contextual facts related to the persona question, thus providing fairly accurate answers. This concept corresponds to modern trends in the approximation of large returns [4] and allows us to create an effective strategy for the hassle.

3. METHODOLOGY :

Input in this system comes from users as requests made using a chatbot interface. The input is analysed using Natural Language Processing techniques, including tokenization and intent detection [2]. In this way, it becomes possible to detect intent and identify key elements of a query, even if phrased differently.

Next, the question was vectorized based totally on an embedding model. This technique allows a semantic specialization of a request that can be in terms of facts available in the vector database [3]. Thanks to this solution, the system can perform semantic searches, instead of specific keyword searches, and retrieve applicable statistics.

A similarity search set of rules is used to perceive valid documents within the vector database primarily based on their proximity to the vectorized query [4]. Cosine similarity can be used to evaluate similarity of vectors as a way of retrieving valid information. Dense retrieval proposed in [3] helps to reap proper contextual relevance in the course of fact retrieval.

Ranking is another part of this process that helps filter out unnecessary information which could otherwise end up in the final result. Ranking makes sure only helpful information moves forward in the process to help create answers from the retrieved documents [2].

Then, the chosen information is sent to the generator, often a Large Language Model (LLM), to create a clear and meaningful response. It's important to note that LLMs by themselves don't have the background information needed to give accurate answers, so embedding information helps ground the responses in real-world facts [2].

Additionally, the system keeps track of the conversation, allowing it to handle several questions from the same user in a way that feels natural and connected. The ability to carry on a conversation is one of the main strengths of modern chat systems [1].

Lastly, there's a separate part of the system that handles data collection and updating the knowledge base using information from the organization's sources [4].A good way of organizing and finding data supports the system's ability to grow and work quickly.

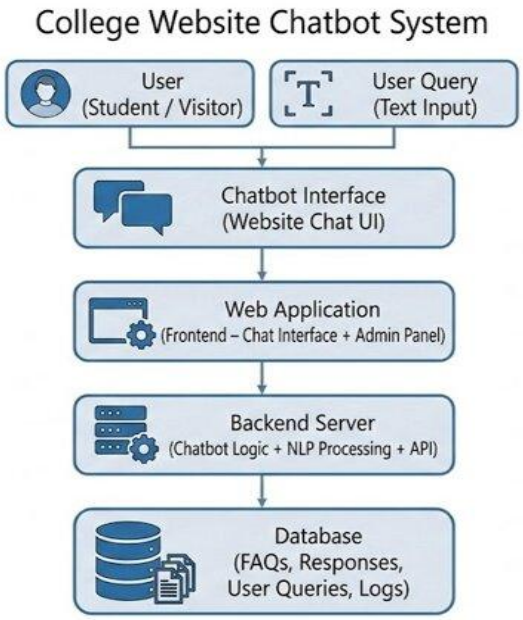
4. DATA FLOW DIAGRAMS :

The diagram shows a data flow chart that explains the steps the system follows, starting from when it gets input from a user until it sends back a response. It clearly shows how different parts of the system work together, like handling the user's query, getting information, ranking the results, and creating the response. This helps show the order in which information changes as it moves through the system. This kind of setup is important in modern chat systems to make them work well and efficiently [3].

Each part of the system plays a key role in making sure everything runs smoothly. The user interaction begins when the system gets data from the user, and then a series of steps happen to understand the query and find the right information. As you can see from the structure, the way it's designed is similar to how conversational agents are built [2]. In the end, the system creates and sends the response back to the user.

There are also some steps in between where changes happen, like making embeddings and applying filters to get only useful information. This ensures that only relevant data is used when making the response. Good retrieval and indexing are also important to help achieve this [4].

Figure 1: Data Flow Diagram of IndiraGPT



5. SYSTEM ARCHITECTURE:

The proposed architecture has several layers, each with a specific job, making the structure clear and organized. This setup makes it easier to scale, maintain, and connect with other systems, which are important in today's software and AI development [2]. Because of this, changes in one part of the system don't affect other parts.

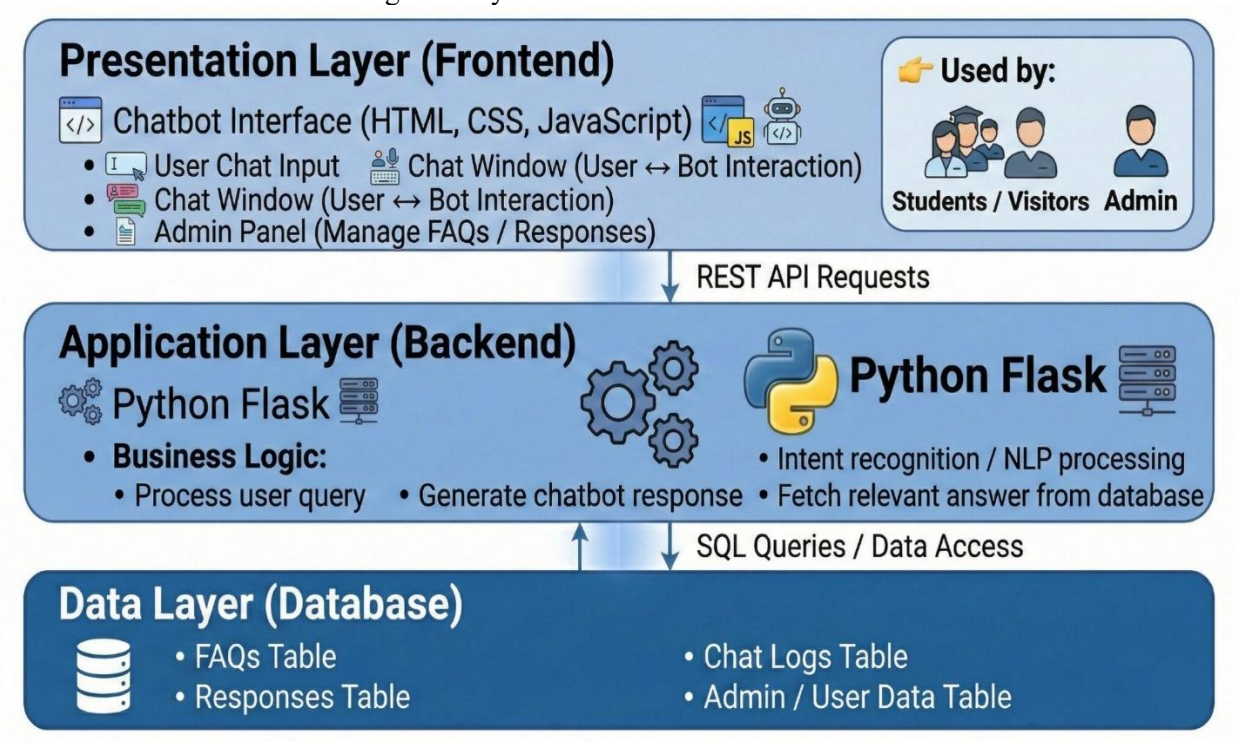
The presentation layer helps users interact with the system. It handles user questions and creates responses through conversations. This layer works with the backend services to process incoming requests smoothly. The application layer holds all the main logic, like understanding natural language, analysing meaning, getting data, and creating responses.

This layer is key in the overall process where different steps come together to provide a response. Its design is similar to how modern conversational AI systems are built [3].

The data layer stores both structured and unstructured data in a vector format. Using embedding techniques allows the system to find similar data by comparing the user's question with stored information. This method is widely used in information systems because it's accurate and adaptable [4].

External services help with extra tasks like understanding language, searching for meaning, creating embeddings, and extracting information. This allows the system to use more advanced tools for better calculations and learning new facts.

Figure 2: System Architecture of IndiraGPT



6. IMPLEMENTATION:

For the system to work, we use modern web technologies and a backend framework. This ensures the system runs well and is easy for users to use. The frontend is designed to be simple and responsive, so users can interact with the bot smoothly. It's important that the frontend is easy to use and clear, so users can ask questions without needing any special skills or knowledge.

On the backend side, the system has several parts that each handle different tasks. These tasks include understanding natural language, getting data, and creating responses. All these parts talk to each other using APIs, which is a common way to build flexible and fast software [2].

The vector database holds special versions of the data, called embeddings. These embeddings help the system find the right information by understanding meaning, not just words. Semantic search, which uses this method, is explained in detail in [3] and works much better than searching by keywords alone.

The system can also use a generative AI service that connects to other tools. This lets the system use Large Language Models without needing a lot of local equipment. This setup makes it easier to update and improve the model without changing the app much.

To handle many users at once, the system needs a good setup. Using efficient methods for managing loads and data helps the system run smoothly and handle more users without slowing down. A scalable architecture is key to making the system perform well.

7. RESULTS :

Different kinds of questions were looked at to check how well the system works. These questions include ones that ask for information, ones that ask about procedures, and follow-up questions. From the analysis, it's clear that the system is very good at giving accurate answers quickly. This is because it uses a retrieval system along with language models [3].

The way the system talks helps users feel satisfied. They can ask more questions without having to repeat the same information each time, which is a big plus. Keeping a conversation going is important for chatbots, and it helps users stay interested [1]. A major benefit of this system is its ability to give correct answers.

This is because it uses a retrieval method that provides background information. In other words, this method lowers the chance of giving wrong or false information since it uses verified sources [2]. How quickly the system answers questions is a key measure of its performance.

With the new technology, this time has gone down compared to older methods like manually searching or asking an administrator. So, the system is very efficient in this area [4].

Overall, the system is good at handling different types of questions, including those that need extra thinking or more than one step.

8. CONCLUSION:

The IndiraGPT system solves issues with current college websites by using smart conversation features to help users find information. Instead of going through many pages, users can just chat with the app, making it easier for everyone to use [1].

By combining the process of finding information with the ability to create new responses, the system ensures the answers are accurate. It also checks the data to avoid giving wrong information, while still allowing the system to be flexible, which is a big advantage of the Retrieval-Augmented Generation method [3].

Having easy, back-and-forth chats improves the user experience and makes it easier for staff to handle questions. This setup makes the system more efficient and cheaper to run compared to other options [2].

In the future, the system could add features like supporting multiple languages to reach more people. Adding ways to interact, like through mobile apps or voice assistants, would also be helpful. Also, improving the system's ability to handle more users and better search for information will keep it running smoothly as it grows [4].

9. ACKNOWLEDGEMENT:

The authors wish to express their heartfelt gratitude towards Prof. Monika Patil and Prof. Vidya Dhoke for the guidance, suggestions, and help rendered during the preparation of this research work. They have played a major part in moulding the project.

The authors are equally grateful to Indira College of Engineering, Pune, for providing the required facilities and infrastructure to conduct this research work.

Also, the authors wish to thank their colleagues and well-wishers whose efforts helped complete this research.

10. CONFLICT OF INTEREST:

Authors state clearly that there is no conflict of interest concerning the publishing of this research article. The present research is conducted independently, without any influence from any outside agency or organization or individual.

11. AUTHOR'S BIOGRAPHY:

Prof. Monika Patil is one of the lecturers from the Department of Artificial intelligence and Data science Engineering, Indira College of Engineering, Pune.

Prasad Munde is one of the students pursuing studies in Artificial intelligence and Data science, Indira College of Engineering, Pune. His areas of interest are web development, artificial intelligence, and creating intelligent systems.

Vaibhav Kulkarni is one of the students pursuing studies in Artificial intelligence and Data science, Indira College of Engineering, Pune. His areas of interest are machine learning and designing intelligent systems.

REFERENCES:

1. Hussain A., et al., (2020): A Survey on Chatbots and Conversational Agents. *Journal of Artificial Intelligence Research*, 67(1), 1–35.

2. Adamopoulou E., and Moussiades L., (2020): An Overview of Chatbot Technology. *Artificial Intelligence Applications and Innovations*, 584, 373–383.

3. Karpukhin V., et al., (2020): Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781.

4. Johnson J., Douze M., and Jegou H., (2017): Billion-scale Similarity Search with FAISS. *IEEE Transactions on Big Data*, 7(3), 535–547.