# Leveraging Large Language Models for the Creation of an AI Lecture Chatbot Avatar.

AUTHOR CONTRIBUTIONS

**J. Obobairibhojie:** Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **B. O. Lawal:** Supervision, Writing – original draft, Writing – review & editing.

***ABSTRACT***

This project aimed to design an AI lecture chatbot by combining large language models (LLMs) and learning management systems (LMS). LLMs like ChatGPT are trained on vast data, enabling advanced natural language processing. A survey of 36 students from Edo State University Uzairue revealed that 80.6% were from the Faculty of Engineering, 40% were familiar with AI chatbots, and 85% supported AI in education. The chatbot uses a Retrieval-Augmented Generation (RAG) pipeline where users upload files, which are processed into vector embeddings for generating accurate responses. Despite the small sample size and focus on engineering students, the study highlights the positive reception of AI in education. Future research should include a larger, more diverse sample to confirm these findings. This project shows the potential for AI chatbots to enhance education.

***KEYWORDS***

Large Language Models (LLMs), Retrieval Augmented Generation (RAG), Artificial Intelligence (AI) Chatbots, Learning Management Systems (LMS).

## I. INTRODUCTION

A large language model (LLM) is an infrastructure built upon large corpora of data, numerous layers of prediction models, and deep learning algorithms for proper natural language understanding (NLU) and natural language processing (NLP) (Gan et al., 2023 Help Net Security 2023 and Lake, 2023). Using numerous amounts of data, LLMs are trained using processes such as pre-tuning and fine-tuning, which equip the model with a vast repository of context, world knowledge, language patterns, and domain-specific knowledge. LLMs such as ChatGPT and many others are popularly known and widely used in various fields.

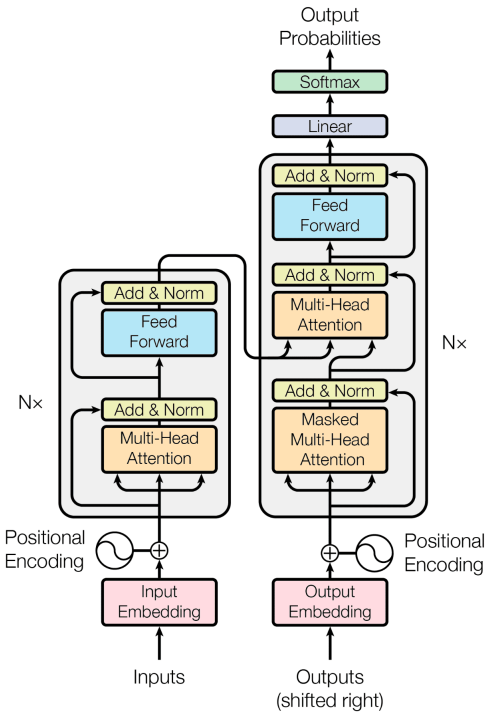
A Generative Pretrained Model (GPT) is a collection of artificial intelligence models that serve as a foundation for various applications that leverage human-like text generation output (Guinness, 2023; Shulze, 2023). According to the research (Duarte, 2023), chatGPT has over 100 million users and gets over 1.6 billion monthly visits. The LLM collects queries as inputs and generates a text output.

A lecture is a form of communication between an educated and qualified lecturer and students, conveying information, knowledge, and skills verbally or visually, thereby invoking

an understanding in the student's mind (Bruff, 2010). Traditional lectures have been the foundation of education for centuries, especially in Nigeria. However, with the advancement of technology, especially during the COVID-19 pandemic, lectures have transformed. According to the research (Li et al., 2020), educational technology experienced exponential growth before the pandemic, with investments reaching $18.66 billion in 2019 and projected to reach $350 billion by 2025. Multiple online learning platforms offer free access to their service, including Tencent Classroom, which the Chinese government has utilized to migrate 250 million students to continue their online studies. BYJU, an education technology company based in Bangalore, has seen a 200% increase in student traffic. Higher institutions such as Zhejiang University migrated more than 5000 courses online within two weeks. Among the numerous advantages of online learning, According to the research (Li et al., 2023; Online Education Trends report, 2023), 96% of online program graduates, 98% of current online students, and 94% of hybrid students would recommend online learning to others. On average, students retain 25–60% more material online compared to 8–10% in a traditional classroom. E-learning takes 40–60% less time because students feel comfortable enough to progress at their own pace. Online learning enrolments in health professions and computer and information science majors are expected to increase over the next five years. However, online learning also has numerous adverse effects. According to the research (Li et al., 2020), many students lack fast internet access or technological gadgets to participate in online lectures. In third-world countries like Indonesia, only 34% of students have computers, while 95% do in first-world countries like Switzerland, the United States of America, and Austria. Despite the end of the pandemic, online learning has continued to be integrated into traditional lectures, which leaves educators searching for the best ways to take advantage of both traditional and online lecture formats.

The most recent technological advancement that is currently gaining significant attention is artificial intelligence (AI), which is currently being integrated into traditional learning and education sectors, with LLMs like ChatGPT, Claude, and Bard being used by multiple students and educators for creating chatbots, automatic feedback and grading systems, learning platforms, intelligent tutoring systems, and more. According to the research (Ames, 2023), the generative AI economy is predicted to grow at a rate of 76.9% per year, and in education, it is expected to reach $80 billion by 2032. A Forbes Advisor survey (Hamilton, 2023) stated that 55% of educators believe AI has improved educational outcomes. Over 60% of educators use AI in the classroom. Over 55% of educators use AI-powered educational games for game-inspired learning, 43% use adaptive platforms, and 41% use automated grading and feedback systems. Despite the benefits of AI in education, some educators have concerns about academic dishonesty. 65% of educators show concerns about plagiarism in essays, 62% in self-learning, 42% in information privacy and security, and 30% in job displacement.

To the best of researchers' knowledge, a few studies have provided an understanding of artificial intelligence and education; a notable gap exists in integrating learning management systems (LMS) features with LLMs like GPT3. This research aims to leverage LLMs and LMS to produce an AI lecture chatbot.




## II. THEORETICAL ANALYSIS

### *A. artificial intelligence*

A machine or computer system's ability to replicate and accomplish tasks that typically necessitate cognitive abilities, such as logical reasoning, learning, and problem-solving, refers to Artificial intelligence (AI). Artificial intelligence is based on applying neural network algorithms and modern technologies to allow computers to use specific cognitive skills to execute activities independently or semi-autonomously (Morandín-Ahuerma, 2022).

### *B. Natural Language Understanding*

Natural language understanding is the comprehension ability of a machine or computer to utilize linguistic knowledge of dynamic complexities to perform specific functionalities (Chatterjee, 2017). NLU involves understanding the intent of a user's query. The earliest project, Daniel Bobrow's program STUDENT (1964), used NLU for algebraic word problems. Recently, NLU has made various technological advancements and improvements in understanding simple natural languages.

### *C. Large Language Models*

A large language model (LLM) is an infrastructure built upon large corpora of data, numerous layers of prediction models, and deep learning algorithms for proper natural language understanding (NLU) and natural language processing (NLP) (Gan et al., 2023; Help Net Security 2023; Lake, 2023).

*D. transformers*

A transformer is a natural language processor that uses vector representations to build a contextual representation of tokens. Figure 1 shows that Transformers comprise three key components: embedding layer, encoder, and decoder. The embeddings layer converts each token into an array of vector representations representing an input. The input is then processed using an encoder and a self-attention mechanism, creating contextual representations. These encoded inputs are then processed using another self-attention mechanism and a decoder to produce an output of tokens (Vaswani et al., 2017).

*Figure 1. Transformer-model Architecture*

*E. retrieval augmented generation (RAG)*

Retrieval Augmented Generation (RAG) is a system architecture designed to provide LLMs with accurate and synthesized information by leveraging an external knowledge base. LLMs are trained on large corpora of data. Usually, in the information retrieval process, LLMs often need more domain-specific knowledge about the queried topic. Consequently, this leads to hallucinations and a need for more elucidation. The RAG architecture mitigates these hallucinations by providing relevant and accurate information from an external knowledge base. RAG is broken into three stages: The retrieval stage, the augmentation stage, and the generation stage. In the retrieval stage, the user sends a query to the LLM, and the RAG concurrently searches for relevant information about the user's query from the external knowledge base. This information is indexed into multiple chunks of data, then sent to the embedding layer and stored in a vector database.

## III. RELATED WORKS

### *A. tashi-bot*

(Carlos et al., 2021) Attempted to develop an AI chatbot to assist applicants and university students obtain information on academic and administrative processes within an educational institution. They discussed the applications of AI in education and the emergence of AI chatbot technology facilitating teaching and management tasks in education. They designed and developed Tashi-Bot, an AI chatbot platform that assisted university applicants and students in acquiring information about academic and managerial procedures encompassing several aspects such as well-being, tuition, fees, admissions, and other services conducted by an academic institution. The chatbot features included bot messages, cards for quick access, and specific intents and entities. A survey was conducted to define the specific features the users needed to develop the chatbot. The chatbot design process included needs, content analysis, design, implementation, and testing. The study sample included 90 students in total. After the development process, the AI chatbot was deployed on a telegram channel. The data obtained through this research highlighted that high user satisfaction levels were observed, with over 92% of students recommending the AI chatbot platform and 12% suggesting that the word length should be reduced.

Despite the significance of (Carlos et al., 2021) findings, it is essential to acknowledge the study's limitations. The study focused on the design and implementation of Tashi-Bot but did not extensively discuss the technical details of the Chatbot's functionality or the specific algorithms used for its operation.

### *B. aiva-bot*

According to the research study (Rabuan & Ping, 2021), They attempted to develop an AI chatbot system to save students time and energy. The paper introduced the state of administrative management at the University of Malaysia Sarawak (UNIMAS), highlighting the increasing population of the institution and the division in charge of undergraduate enrolment: the UNIMAS Undergraduate Studies Division, responsible for graduate registrations, evaluations, graduation and all undergraduate-related issues. They also highlighted the various communication devices used to convey information in the institution and how modern technologies like chatbots can aid students' ability to learn quickly and effectively. They conducted literature research on four systems: three English chatbots and a chatbot that supported multiple languages; each possessed distinct characteristics, benefits, drawbacks, and functionality during information transmission. They then developed an AI chatbot system called AiVA bot, a web-based chatbot platform for undergraduate students at UNIMAS. This chatbot utilised functionalities to differentiate students based on their educational level.

*C. Tutoring Postgraduate students with an AI-based chatbot*

According to the research study (Koivisto, 2023), The research aimed at exploring the use of chatbots for tutoring and counselling services. The paper introduced the recent adoption of AI chatbots in various economic sectors, including the medical, healthcare, and production sectors. The paper then examined the previous studies on using AI in universities, highlighting that higher institutions use AI for more technical purposes, such as adaptive systems and personalisation. They found that tutoring and other non-administrative tasks heavily depended on the staff members, which were limited by labour output and service hours.

The research was conducted on postgraduate students at the Finnish University of Applied Sciences, where the chatbot named Vivian was developed and implemented for student counselling in personal study planning. The study collected data through questionnaires that assessed user experience, chatbot-specific topics, and the strengths and weaknesses of chatbots in course selection tasks. The sample included 57 anonymous postgraduate students, of whom 53 adopted Vivian AI. Metrics such as CSAT (Customer Satisfaction Score), CES (Customer Effort Score), and NPS (Net Promoter Score) were used to evaluate user experience. The results of the survey highlighted that the students found Vivan easy to use with a (CES = 0.80) but were unsatisfied with the chatbot with a relatively low (CSAT = 0.68) and were unlikely to recommend the chatbot to other students, resulting in a low (NPS = 0.63).

Despite the insightful findings by (Koivisto, 2023), it is necessary to highlight the limitations of their approach. The paper limited the geographical scope to postgraduate engineering students at the Finnish University of Applied Sciences, neglecting other student populations. The research was limited to a questionnaire survey, which may have indicated a response bias from the postgraduate students. The research evaluated the chatbot's performance based on short-term user satisfaction metrics like CSAT, CES, and NPS, which may not fully capture the long-term impact and user loyalty towards the service. The paper did not extensively explore the technical aspects of developing the chatbot or the specific algorithms used, which could provide valuable insights into the chatbot's functionality and limitations

The significant findings from the reviewed studies suggest that AI chatbots can enhance educational experiences by providing personalized learning recommendations, answering frequently asked questions, and facilitating administrative tasks. However, the studies also revealed limitations, such as geographical constraints, response biases, and the need for a more extensive evaluation of user satisfaction and long-term impact.

## IV. METHODOLOGY

### *A. Choice of the methodology*

The RAG framework was chosen to provide LLMs with accurate and synthesized information by leveraging an external knowledge base. LLMs are trained on large corpora of data. Usually, in the information retrieval process, LLMs often need more domain-specific knowledge about the queried topic. Consequently, this leads to hallucinations and a need for more elucidation. The RAG architecture mitigates these hallucinations by providing relevant and accurate information from an external knowledge base.

### *B. Models and architecture of the system*

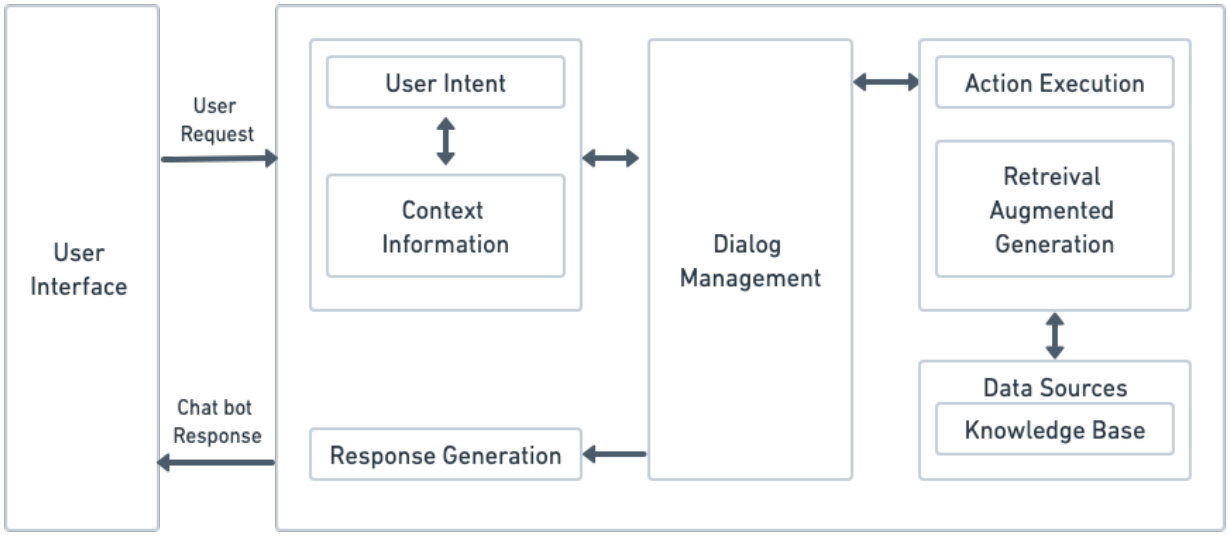



**Figure 2. Block diagram**

The block diagram depicted in Fig. 2 highlights the operation of the AI lecture chatbot. The user interface receives user input in textual format and displays the responses generated by the chatbot. This input is converted into tokens fed into the Dialog Management system. Within this system, the User Intent module analyses the tokens to determine the user's intent, while the Context Information module maintains the conversation's context. The Action Execution module decides on the appropriate action, utilizing RAG to query external data sources or knowledge bases for relevant information. The Response Generation component then crafts a contextually appropriate reply, sent back to the user interface, completing the interaction cycle. These modules are interfaced together through the VS code IDE, where the code is documented.

#### *1) Mathematical models*

This project focuses on developing mathematical models to estimate both the server and cost requirements for an AI chatbot system designed for the 400 and 500-level students of

the Department of Computer Engineering at Edo State University Uzairue. The chatbot is intended for a population of around 100 students.

The first model estimates server requirements by considering factors such as concurrent users, server capacity, growth projections, redundancy needs, and load variability.

The second model calculates cost requirements based on token consumption, including concurrent users, token price per usage, and weekly costs.

*2) Mathematical model for estimating scalability: First Model*

To create a proper mathematical model we must make assumptions that apply constraints and define the boundaries in which our model operates. These assumptions help to simplify the complex real-world scenarios allowing us to focus on the key relationships that are useful to our model's accuracy.

1. The maximum 400 and 500 level computer engineering student population considered is 100.
2. No previous server infrastructure exists for this project.
3. There is a constant power supply available.
4. Server room space is already provisioned and sufficient.
5. Users interact with the chatbot with an average session length and frequency.
6. Future growth in staff and student populations is projected over a certain timeframe.
7. Load varies throughout the day, with peak times requiring additional server capacity.
8. Average response time should meet or exceed a threshold (T) for user satisfaction.
9. We assume we are purchasing an entry-level physical server (PowerEdge T30 Mini Tower Server) priced at $1,335.00 or ₦ 2,002,500.
10. Specifications of this (PowerEdge T30 Mini Tower Server) include a basic setup such as 4-8 core CPU, 8-16GB RAM, SSD with at least 256GB, and a 1Gbps Bandwidth which could easily handle 50-200 concurrent users

*3) Variables*

1. $a =$ Number of Staff members
2. $b =$ Number of students
3. $s =$ Number of servers required

4. $P =$ Percentage of staff and students likely to use the chatbot concurrently $(0 \leq P \leq 1)$

5. $U =$ The number of users each server can handle simultaneously

6. $G =$ Growth Factor for future staff and student populations

7. $L =$ Load factor accounting for peak vs. average usage

*Estimate the number of concurrent users*

$$\text{Total concurrent users} = P \cdot (a + b) \tag{1}$$

Therefore, to calculate the number of concurrent users, we will add the number of staff and students and then multiply the result by the percentage of staff and students likely to use the chatbot concurrently.

*Growth and load adjustments*

$$\text{Adjusted concurrent users} = \text{Total concurrent users} \cdot G \cdot L \tag{2}$$

Equation 2 explains the case where there is notable growth in the staff and student populations. In such cases, we multiply the total concurrent users from eqn. 1 by the growth and load factors

*Server Capacity*

$$\text{c} = \frac{\text{Adjusted concurrent Users}}{U} \tag{3}$$

To calculate the server capacity, we divide the result obtained from eqn. 2 by the number of servers each server can handle.

*Redundancy*

$$\text{Total servers with Redundancy} = c \times R \tag{4}$$

This gives the final number of servers needed to ensure the system is robust against failures and can handle maintenance without downtime.

*4) Model for estimating the cost of token consumption by users*

1. The number of tokens concurrent users use weekly is 5 million for the Mistral-7B LLM.

2. The number of concurrent users is based on the 36 respondents from the questionnaire, considering only the 80.3% percent of Faculty of Engineering students.

3. The price per token is $0.25 per million for the Mistral-7B LLM.

4. The questionnaire had a higher percentage of computer engineering respondents, so we can equate the faculty of engineering respondents to the Department of Computer Engineering 400 and 500-level students.

5. Total tokens used by concurrent users in a week: 5,000,000 tokens.

*5) Variables*

1. $U$: Number of concurrent users

2. $T$: Total tokens used in a week

3. $C$ : Cost per million tokens

4. $W$: Weekly cost for token consumption

5. $P$: Percentage of respondents that are in the faculty of engineering.

6. $E$: Concurrent users in the Faculty of Engineering

7. $X$: Number of tokens consumed by a user in a week

*6) Equations*

1. Number of Concurrent Users in the Faculty of Engineering

$$\mathsf{E} = P \cdot U \tag{1}$$

To calculate the number of concurrent users in the engineering faculty, we use Eqn. 1. We multiply the percentage of respondents $(P)$ by the total number of concurrent users $(U)$.

2. Number of tokens consumed by a user in a week

$$\mathsf{X} = \frac{T}{E} \tag{2}$$

To calculate the total number of tokens consumed by a user in a week, the total number of tokens ($T$) would be divided by the number of concurrent users in the engineering faculty ($E$).

We then scale the model to accommodate potential increases in the number of users and token usage. We will consider three scenarios:

1. Scenario 1: 100 Users

**2.** Scenario 2: 1,000 Users

**3.** Scenario 3: 10,000 Users

*C. Research Survey*

The researcher conducted a survey where 80.6% of students were from the Faculty of Engineering, 13.9% from the Faculty of Basic Medical Sciences, and 5% from the Faculty of Sciences. The sample was limited to 400 and 500-level students ages 17-23, where 40% of students were already conversant with AI chatbots like ChatGPT, Claude, and Google Bard for educational purposes, 42.9% rated their experience as excellent, and 85% endorsed the adoption of AI in education.

The total number of responses accumulated to 38 responses, which served as a basis for the development of the chatbot's primary features and model calculations

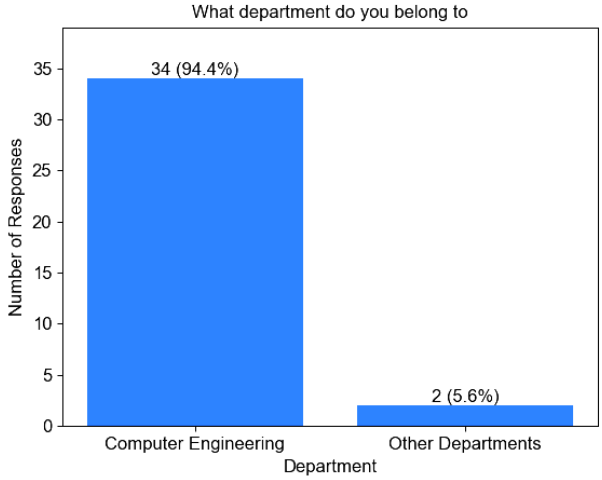



**Figure 3. Statistics displaying the number of responses about the various departments.**

94.4% of students who participated in the survey were computer engineering students, whereas 5.6 % of students were from other departments, further narrowing the scope of the study to focus more on the computer engineering department to conserve more computing resources.

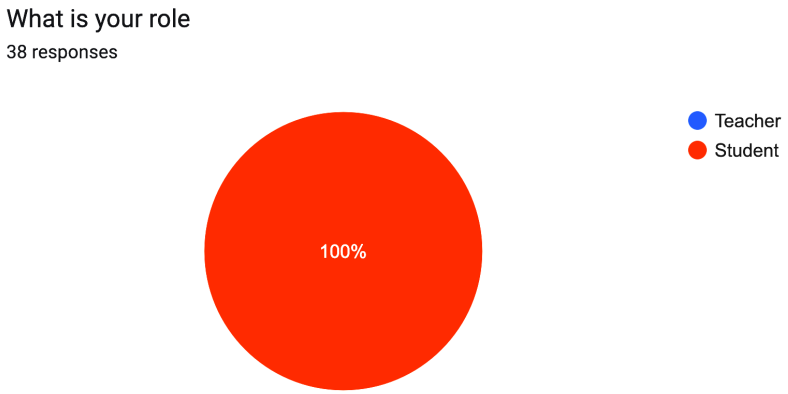



**Figure 4. Chart displaying the statistics of teacher-to-student respondents**

The pie chart highlights the total responses from only university students.

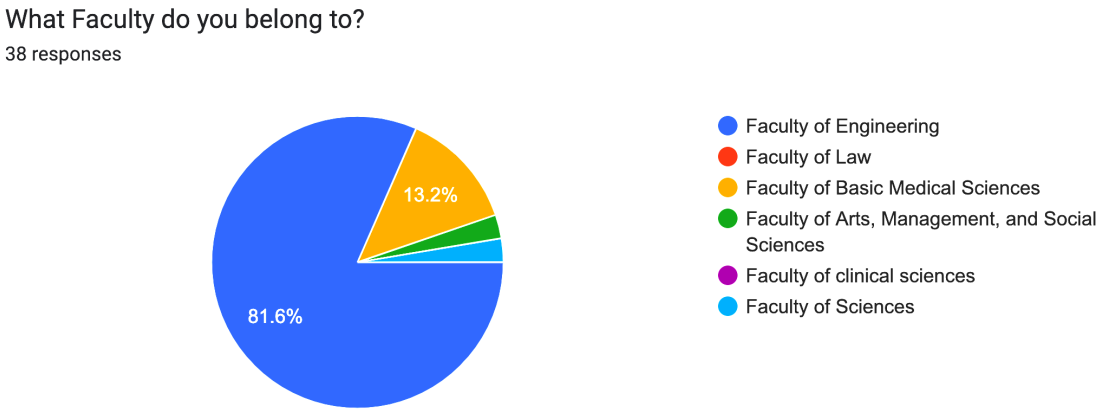



**Figure 5. Chart displaying the statistics of the various faculties that responded to the questionnaire.**

The pie chart highlights that 81.6% of respondents came from the engineering faculty, 13.2% from the faculty of basic medical sciences, and other faculties had negligible percentage values. This further solidifies the focus on the engineering faculty rather than others.

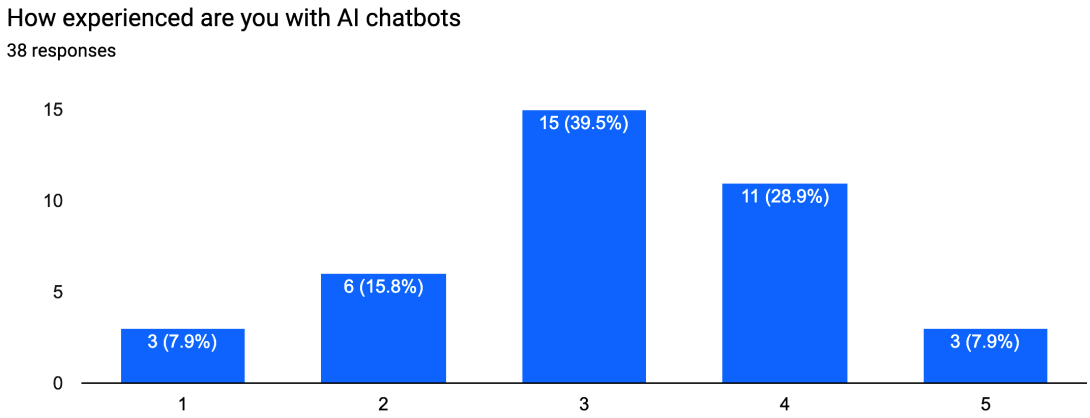



**Figure 6. Chart displaying the statistics of respondents experience with chatbots**

The bar chart highlights the various experience levels of various respondents with AI chatbots, with 28.9%, 7.9%, and 39.5% of responders having an average-to-high level of experience with AI chatbots, 15.8%, and 7.9% having a below-average level of experience with AI chatbots.

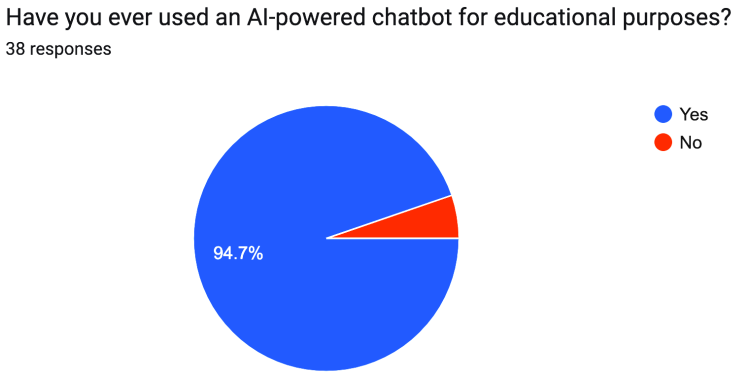



**Figure 7. Chart for the respondents' AI chatbot usage statistics for educational purposes.**

The pie chart in Figure 7 highlights that 94.7% of respondents have utilized an AI chatbot for educational purposes, hinting at an already adopted framework in students' workflows.

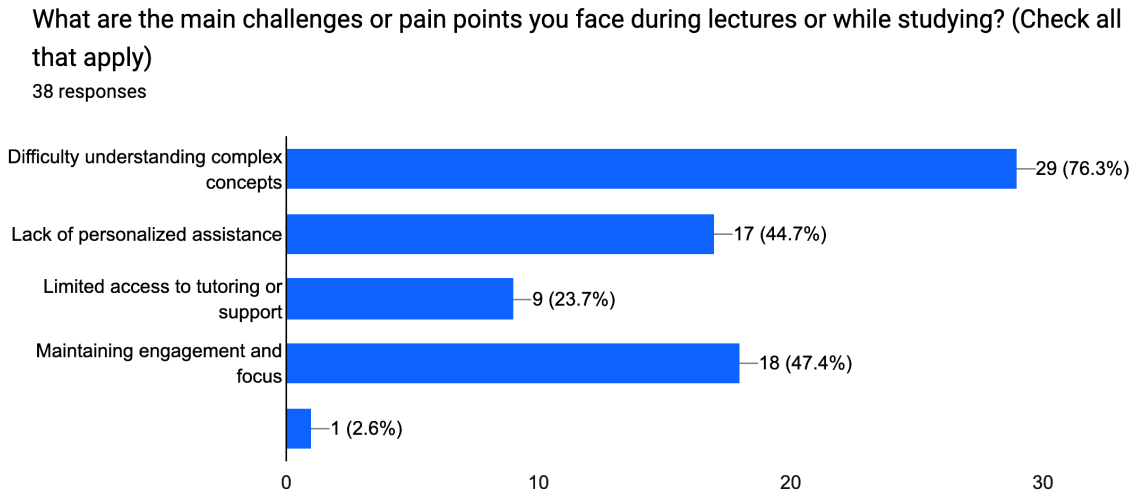



**Figure 8. Chart displaying the statistics or respondents' main challenges or pain points faced while studying**

The bar chart in Fig. 8 highlights respondents' main challenges while studying. This specific metric was evaluated to assess the student's pain points while studying. 76.3% of respondents voted for difficulty understanding complex concepts. 44.7% of students voted for lack of personalized assistance, 23.7% for limited access to tutoring or support, 47.4% for maintaining engagement and focus, and 2.6% for nothing. This evaluation tells us that most students struggle with understanding complex concepts, resulting in a need for an AI chatbot.

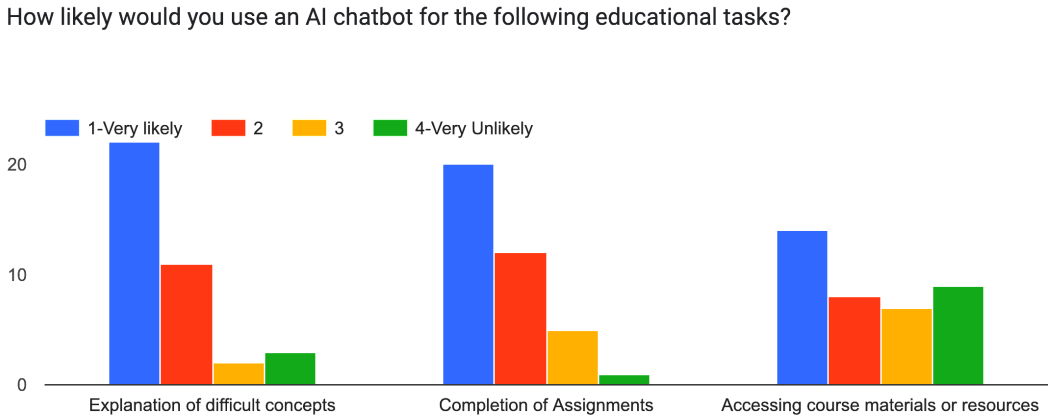



**Figure 9 Chart Displaying the likeliness of respondents to use an AI chatbot.**

The chart in Fig 9. highlights the respondents' likeliness to use an AI chatbot for a specified list of educational tasks like explaining complex concepts, completing assignments, and accessing course materials or resources. With a higher percentage of respondents voting "very

likely" for all of the tasks. This field in the survey was carried out to evaluate the number of features.

*D. Working Principle*

The principle behind the system is a RAG Pipeline. The user initializes the process by uploading a file of various formats, including (PDF, DOCX, and TXT). These files are then stored securely in a temporary location to be processed by the recursive text splitter. The recursive text splitter uses NLP algorithms such as normalization, stop word removal, and tokenization to split the file into chunks. These chunks are fed into the Mistral Large Language Model, where the chunks are converted into vector embeddings. These vector embeddings are stored in a vector database for seamless information retrieval.

The user sends a prompt via the text input, which is then processed to match the preprocessed chunks in the VectorDB. The system uses similarity search algorithms to find the most relevant chunks that match the user's query. These relevant chunks, now contextualized as vector embeddings, are passed to an input parser. The input parser prepares the data by organizing it and ensuring it is in the correct format for context.

This context is then provided to the Mistral LLM. The Mistral LLM uses the contextual information from the relevant chunks to generate a coherent and contextually accurate response. The response generation process leverages the model's deep understanding and training on vast data, ensuring the output is relevant and informative.

The response is then streamlined to ensure it meets the user's needs, possibly involving further NLP techniques such as summarisation or paraphrasing to enhance clarity and relevance. Finally, the output is delivered to the user in a user-friendly format, completing the RAG Pipeline process. This system effectively combines the strengths of information retrieval and generative modeling to provide users with accurate, contextually relevant responses based on their input queries.

## IV. RESULTS AND DISCUSSION

In this section, we discuss the findings from the survey conducted among students at Edo State University Uzairue regarding adopting AI chatbots for educational purposes and the various processes to implement the AI chatbot system. The survey aimed to gauge student interest, familiarity, and experience with AI chatbots and their willingness to integrate such technology into their learning processes. The data collected provides valuable insights into implementing the AI chatbot system.

*A. Results for Maximum Response Time*

**Table 1: Maximum response time based on MTN's network speed.**

| Test Number | Time Interval (24 hours) | Network Speed (Mbps) | Response (ms) |
|---|---|---|---|
| 1 | 00:08 | 4.3 | 60 |
| 2 | 03:00 | 2.3 | 400 |
| 3 | 10:00 | 1.8 | 433 |
| 4 | 14:00 | 3.0 | 188 |
| 5 | 20:00 | 0.3 | 3418 |

The test shows the variance of response latency at different time intervals of the day using the MTN carrier network. A 'Hello' prompt was sent to the AI chatbot. The location of the test was performed at the university faculty and the university hostel (Hall 2). From this test, we can deduce that network latency significantly affects the AI chatbots, especially at busy hours, when network traffic is paramount.

Figure 10. A Plot of Response Latency against Time

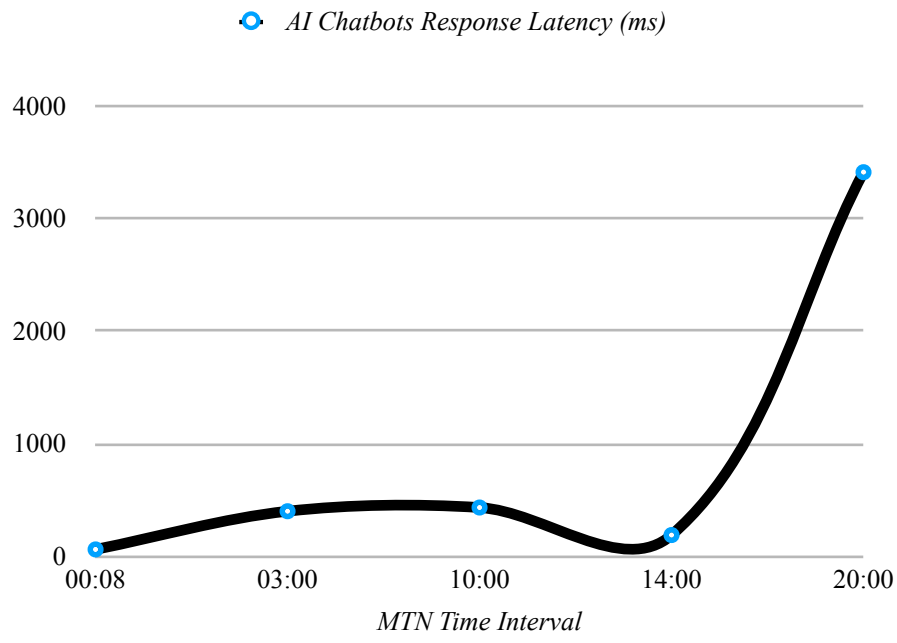


**Figure 11. A Plot of the Response Latency of the AI Chatbot against MTN Network Speed**

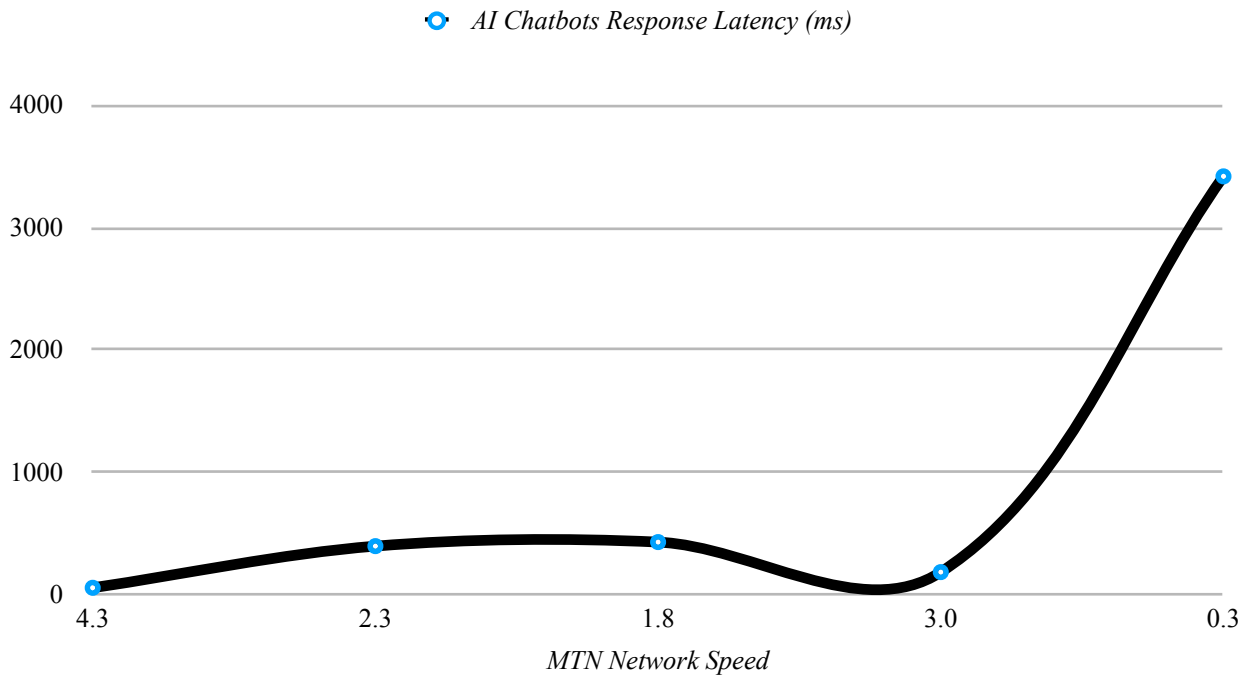



**Table 2. Maximum response time is based on Airtel's network speed.**

| Test Number | Time Interval (24 hours) | Response (ms) | Network Speed (Mbps) |
|---|---|---|---|
| 1 | 02:30 | 386 | 2.30 |
| 2 | 03:00 | 225 | 3.20 |
| 3 | 12:00 | 2994 | 0.18 |
| 4 | 14:00 | 3326 | 0.04 |
| 5 | 17:00 | 866 | 2.60 |

The test shows the variance of response latency at different time intervals of the day using the Airtel carrier network. The test maintained the previous test conditions.

**Figure 12. A Plot of Response Latency and Network Speed against Time**

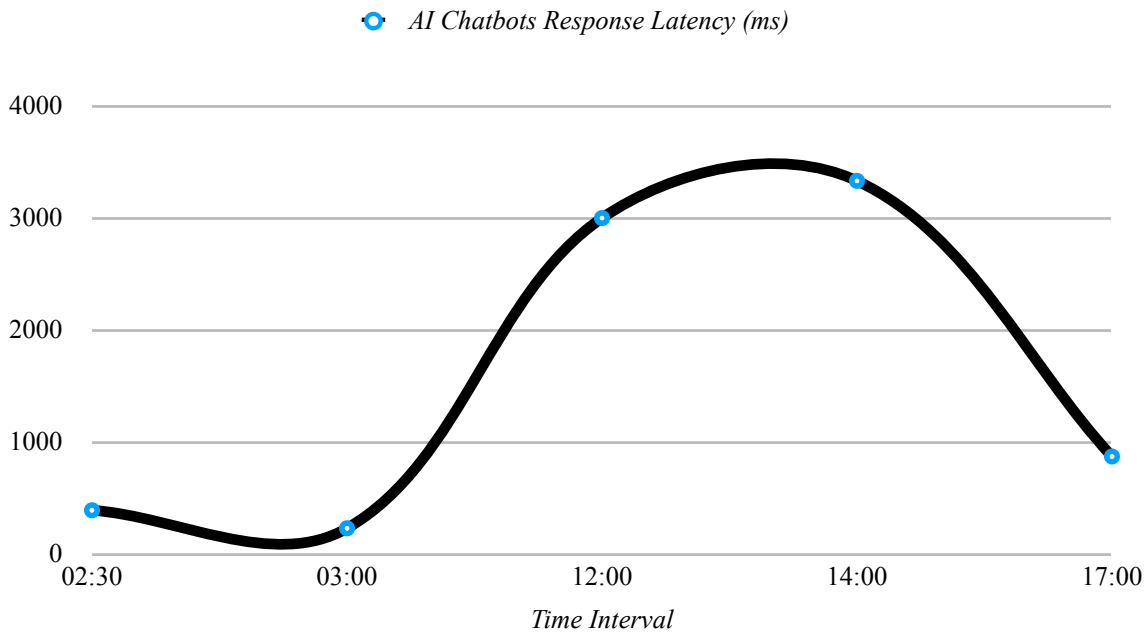
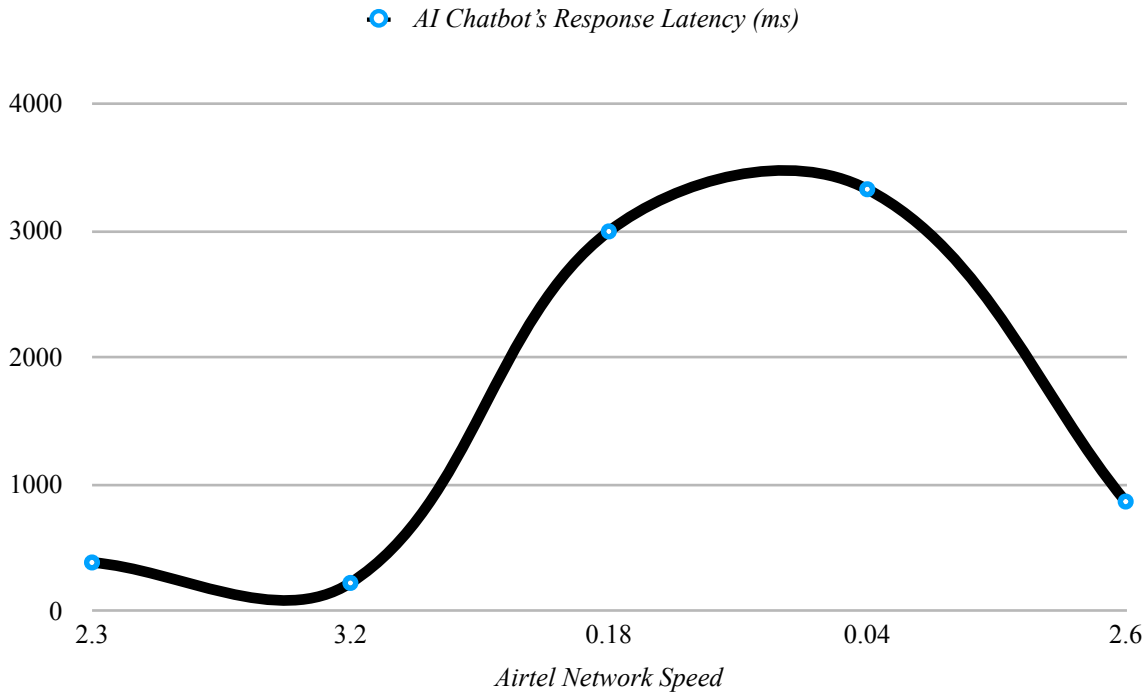



**Figure 13. A Plot of Response Latency against Airtel Network Speed**



*B. Results and Analysis Test for Mathematical Model: Scalability of Servers*

**Table 3. Simulation Results for Scalability of Servers**

| Scenarios | *a* | *b* | *c* | *P* | *U* | *G* | *R* | *L* | Estimated Server Requirements |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 5000 | 20000 | 25000 | 0.2 | 50 | 1.2 | 1.2 | 1.5 | 216 |
| 2 | 6000 | 22000 | 28000 | 0.3 | 70 | 1.3 | 1.1 | 1.4 | 241 |
| 3 | 7000 | 25000 | 32000 | 0.25 | 110 | 1.4 | 1.2 | 1.6 | 196 |
| 4 | 5500 | 21000 | 26500 | 0.18 | 150 | 1.25 | 1.15 | 1.45 | 67 |
| 5 | 4800 | 19500 | 24300 | 0.22 | 200 | 1.1 | 1.2 | 1.55 | 55 |

For Scenario 1, with 5,000 staff and 20,000 students, each server handling 50 users, the estimated number of servers required is 216. In Scenario 2, increasing the staff and students to 6,000 and 22,000, respectively, and raising the server capacity to 12,000 users requires 241 servers. Scenario 3, with the most staff, 7,000 and 25,000 students, requires 196 servers to handle 14,000 users. Scenario 4, adjusting the numbers slightly to 5,500 staff and 21,000 students, with a server capacity of 11,000 users, requires 67 servers. Lastly, Scenario 5, with 4,800 staff and 19,500 students and the highest server capacity of 15,000 users, requires only 55 servers.

*C. Results and Analysis Test for Mathematical Model: Estimating the Cost of Token Consumption by Users*

**Table 4. Simulation Results for Estimating the Cost of Token Consumption by Users**

| Scenarios | Number of Users | Total Tokens (Million) | Cost per Million (dollars) | Weekly cost per Million (dollars) | Concurrent users |
|---|---|---|---|---|---|
| Current | 36 | 5,000,000 | 0.25 | 1.25 | 29 |
| 1 | 100 | 13,896,552 | 0.25 | 3.47 | 80 |
| 2 | 1,000 | 138965517.24 | 0.25 | 34.7 | 806 |
| 3 | 10,000 | 1389655172.41 | 0.25 | 347.41 | 8060 |

The table above summarizes the results. In the current scenario with 29 users, the weekly cost is $0.25. For Scenario 1, with 100 users, the weekly cost increases to $3.47. In Scenario 2, with 1,000 users, the weekly cost rises to $34.7. For Scenario 3, with 10,000 users, the weekly cost escalates to $347.41. Based on the averages, the average number of users considered was 2,782.25, with 80.6% from the Faculty of Engineering, resulting in 2,738.30 adjusted users. The average weekly tokens used were approximately 386,879,310.34, leading to an average weekly cost of $19.34.

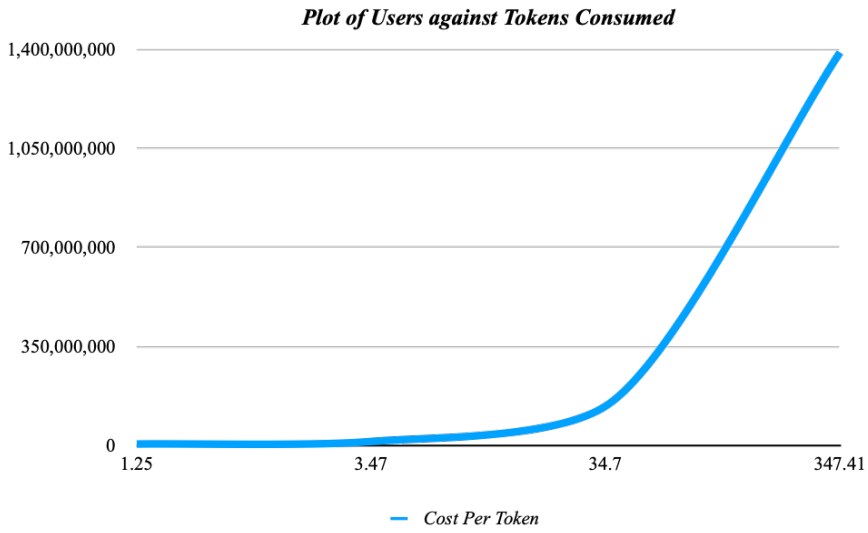



**Figure 14. Plot of Users against Tokens Consumed**

The Figure above highlights how the users' growth drastically increases the project's cost.

*D. Results and Analysis Test for Iterative Changes*

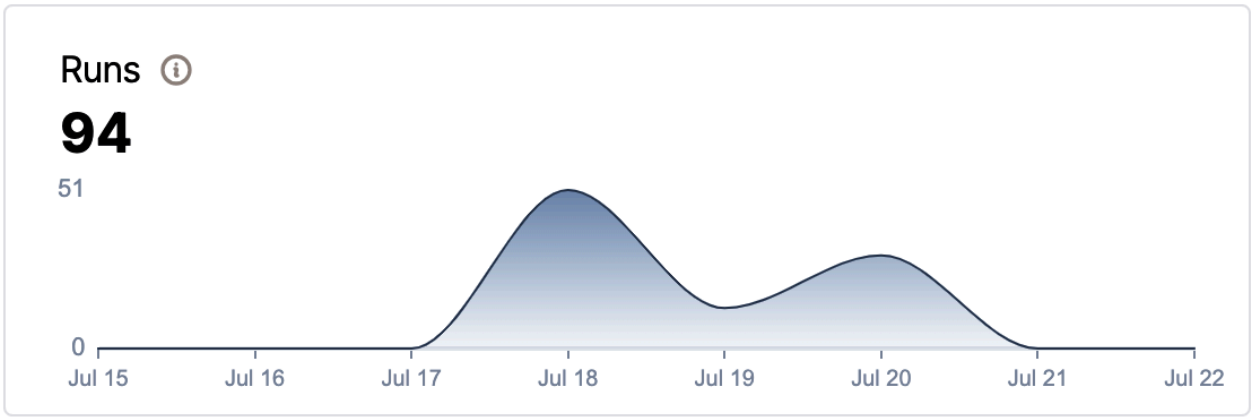



**Figure 15. Plot of Iterative Runs against Time Interval**

The figure above highlights the number of iterative changes made throughout the project.

### *E. Discussions*

#### *1) Adoption Interest amongst Students of Edo State University Uzairue*

The findings from the survey conducted among 36 students at Edo State University Uzairue indicate a high level of interest in adopting AI chatbots for educational purposes. Notably, 85% of the respondents endorsed the integration of AI into education, reflecting a positive attitude toward technological advancements in learning. Most students surveyed were from the Faculty of Engineering (80.6%), which may suggest a higher familiarity and comfort with technology-driven solutions in this faculty.

#### *2) Students Experience with other AI chatbots*

The survey also revealed that 40% of the students were already conversant with AI chatbots like ChatGPT, Claude, and Google Bard, with 42.9% rating their experience as excellent. This indicates a substantial base of students who are not only aware of AI tools but also find them beneficial in their educational activities. The positive reception can be attributed to the efficiency and accessibility that AI chatbots provide in learning environments.

#### *3) Scalability*

The scalability models highlighted the varying server requirements and costs associated with user scenarios. In Scenario 1, serving 25,000 users required 216 servers, whereas increasing server capacity in Scenario 2 allowed for handling 28,000 users with 241 servers. Scenario 3, with the highest number of users, 32,000, needed 196 servers due to a higher server capacity, while Scenario 4's 67 servers supported 26,500 users. The most efficient was Scenario 5, requiring only 55 servers for 24,300 users due to optimal server capacity. These deductions were based on the assumption that a midrange physical server with basic specifications was purchased. The analysis also examined token consumption costs, from $0.25 (₦375) per week for 29 users to $347.41 (₦521,115) for 10,000 users. On average, 2,782.25 users resulted in a weekly cost of $19.34 (₦29,010). These findings underscore the importance of strategic capacity planning and cost management to ensure the scalability and economic feasibility of the chatbot system.

#### *4) Testing and Feedback Collection*

Beta testing was conducted with a small group of 19 students to evaluate the chatbot's effectiveness. Their interactions with the chatbot were closely monitored, and feedback was collected to identify areas for improvement. This iterative testing process was crucial for refining the chatbot’s responses and functionality, ensuring it met the needs of its users.

## V. CONCLUSION

The project successfully designed and implemented an AI lecture chatbot by integrating large language models with learning management systems. The findings indicate that this chatbot

can enhance the learning experience for students, as 85% of the surveyed students endorsed adopting AI in education. The unique approach of combining LLMs with LMS tailored to Edo State University Uzairue demonstrates the potential to revolutionize the educational sector in Nigeria. However, the small sample size and limited focus on engineering students restrict generalizability. Future research should expand the sample size and include diverse faculties to validate these findings further. Recommendations for improving the chatbot include adding a dropdown component for lecturers, a lecturer recommendation system, and expanding the user base to enhance the chatbot's effectiveness.

## REFERENCES

**Ames, S. (2023)** *AI technology in education statistics 2023: Industry trends and 300+ audience survey*, *Rask.ai*. Available at: https://www.rask.ai/research/ai-in-education (Accessed: April 17, 2024).

**Bruff, D. (2010)** *Lecturing, Vanderbilt University*. Available at: https://cft.vanderbilt.edu/guides-sub-pages/lecturing/ (Accessed: April 16, 2024).

**Carlos, H., German, S.-T. and Dixon, S. (2021)** "Tashi-Bot: A Intelligent Personal Assistant for Users in an Educational Institution," *Preprints*. doi: 10.20944/preprints202108.0380.v1.

**Chatterjee, A. (2017)** "Automated Information Organization," in *Elements of Information Organization and Dissemination*. Elsevier, pp. 497–526.

**Duarte, F. (2023)** "Number of ChatGPT users (Apr 2024)," *Exploding Topics*, 30 March. Available at: https://explodingtopics.com/blog/chatgpt-users (Accessed: April 16, 2024).

**Gan, W. *et al.* (2023)** "Large language models in education: Vision and opportunities," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE. doi: 10.1109/bigdata59044.2023.10386291.

**Guinness, H. (2023)** *What is GPT? Everything you need to know about GPT-3 and GPT-4*, *Zapier.com*. Zapier. Available at: https://zapier.com/blog/what-is-gpt/ (Accessed: April 16, 2024).

**Help Net Security (2023)** *The security and privacy risks of large language models*, *Help Net Security*. Available at: https://www.helpnetsecurity.com/2023/05/10/security-privacy-risks-large-language-models-video/ (Accessed: April 16, 2024).

**Hamilton, I. (2023)** *Artificial intelligence in education: Teachers' opinions on AI in the classroom*, *Forbes*. Available at: https://www.forbes.com/advisor/education/it-and-tech/artificial-intelligence-in-school/ (Accessed: April 16, 2024).

**Koivisto, M. (2023)** "Tutoring postgraduate students with an AI-based chatbot," *International Journal of Advanced Corporate Learning (iJAC)*, 16(1), pp. 41–54. doi: 10.3991/ijac.v16i1.35437.

**Lake, R. (2023)** *What is a large language model (LLM)?*, *Investopedia*. Available at: https://www.investopedia.com/large-language-model-7563532 (Accessed: April 16, 2024).

**Li, C., Lalani, F. and World Economic Forum (2020)** *The COVID-19 pandemic has changed education forever. This is how*, *World Economic Forum*. Available at: https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/ (Accessed: April 16, 2024).

**Morandín-Ahuerma, F. (2022)** "What is artificial intelligence?," *International Journal of Research Publication and Reviews*, 03(12), pp. 1947–1951. doi: 10.55248/gengpi.2022.31261.

***Online Education Trends Report* (2024)** *Bestcolleges.com*. Available at: https://www.bestcolleges.com/research/annual-trends-in-online-education/ (Accessed: September 27, 2024).

**Rabuan, M. and Ping, P. (2021)** AiVA-BOT: WEB-BASED INFORMATION PROVIDER CHATBOT FOR UNDERGRADUATE STUDENT IN UNIMAS.

**Vaswani, A. et al. (2017)** "Attention is all you need," Advances in neural information processing systems, 30.

# APPENDIX

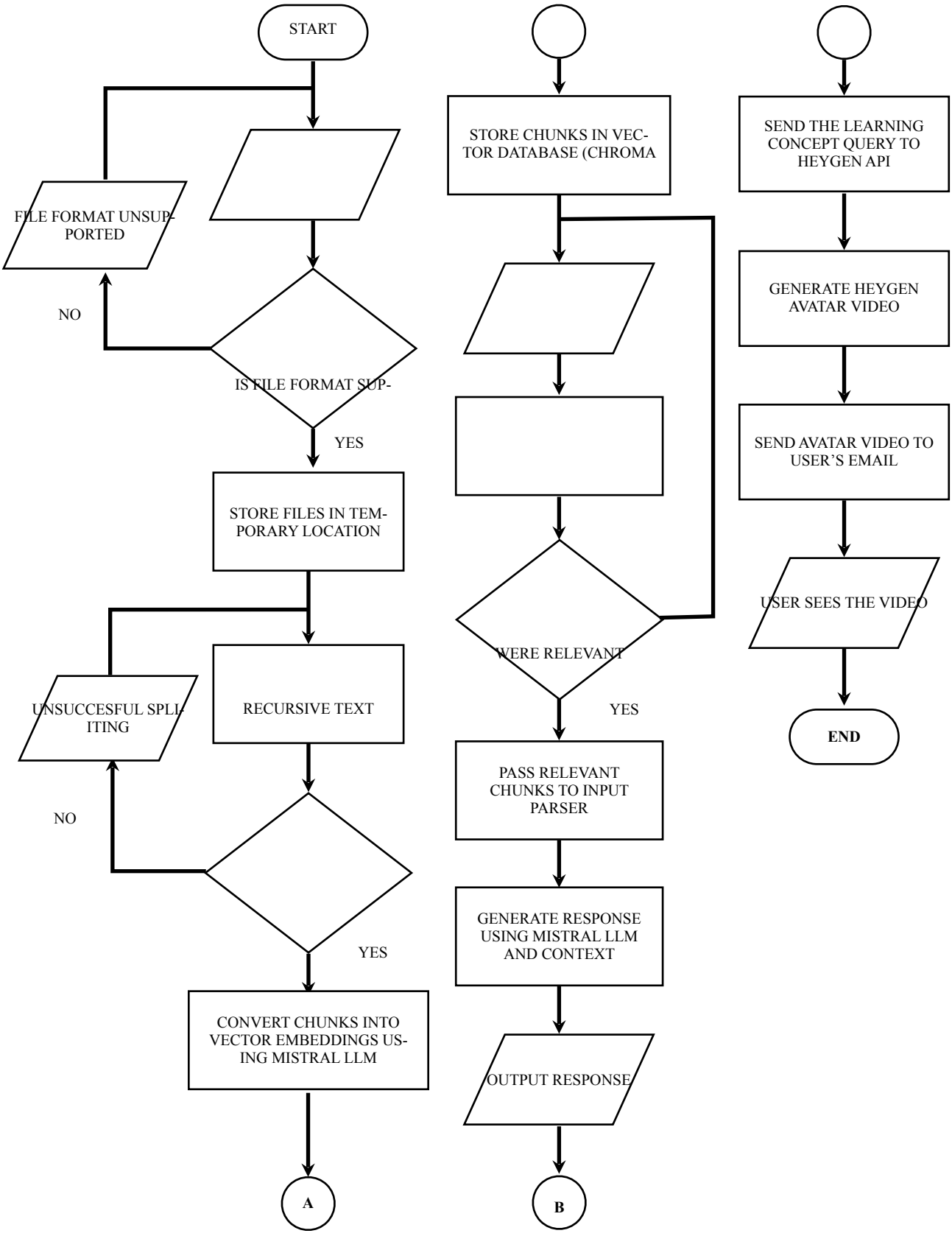

START
FILE FORMAT UNSUP-
PORTED
NO
IS FILE FORMAT SUP-
YES
STORE FILES IN TEM-
PORARY LOCATION
UNSUCCESFUL SPLI-
ITING
RECURSIVE TEXT
NO
YES
CONVERT CHUNKS INTO
VECTOR EMBEDDINGS US-
ING MISTRAL LLM
A
STORE CHUNKS IN VEC-
TOR DATABASE (CHROMA
WERE RELEVANT
YES
PASS RELEVANT
CHUNKS TO INPUT
PARSER
GENERATE RESPONSE
USING MISTRAL LLM
AND CONTEXT
OUTPUT RESPONSE
B
SEND THE LEARNING
CONCEPT QUERY TO
HEYGEN API
GENERATE HEYGEN
AVATAR VIDEO
SEND AVATAR VIDEO TO
USER'S EMAIL
USER SEES THE VIDEO
END

**Figure 16. Flowchart of the system**