Chatbot for Student Descipline Handbook-Related Queries: A RAG-Based LLM Using Llama-3 Approach

Lysa V. Comia
School of Information Technology, Mapua University, Makati City, Philippines
lvcomia@mapua.edu.ph

***Abstract*—This study presents a Retrieval-Augmented Generation (RAG)-based chatbot designed to handle student discipline handbook-related queries using Llama-3 and ChromaDB. By employing Sentence-BERT (SBERT) for semantic similarity assessment, the chatbot ensures accurate, contextually relevant, and policy-compliant responses. Unlike conventional chatbots relying on keyword matching or rule-based systems, the proposed model integrates retrieval and generation techniques, enhancing response precision and coherence. The system is deployed using Gradio, offering a user-friendly interface for seamless interactions. Quantitative evaluation revealed a high mean similarity score of 0.9219, a median of 0.9373, and a low standard deviation of 0.0750, indicating reliable performance. A p-value of 0.0000 confirmed statistical significance, while the lowest similarity score of 0.7267 identified areas for improvement. This research demonstrates the chatbot's potential as an AI-driven educational support tool, enhancing accessibility to institutional policies and reducing administrative workload.**

Keywords— AI for Education, LLM-driven chatbot, Retrieval-augmented generation (RAG), Student Discipline Handbook, Llama-3.

I. INTRODUCTION

Educational institutions implement student discipline handbooks to establish rules, guidelines, and procedures that ensure a safe and structured learning environment [1]. These documents serve as a critical resource for students, educators, and administrators by outlining policies on academic integrity, behavioral expectations, disciplinary actions, and conflict resolution mechanisms [2]. However, despite their importance, students often find it challenging to access relevant information promptly. The sheer length and complexity of these handbooks make it difficult to locate specific policies, and institutional staff may not always be readily available to clarify doubts. To address this gap, we propose a Retrieval-Augmented Generation (RAG)-based chatbot that leverages Large Language Models (LLMs) to efficiently process and respond to student queries regarding discipline-related policies [3].

Traditional methods for accessing policy-related information, such as reading printed or digital versions of the handbook or consulting school authorities, can be time-consuming and inefficient. Many students may struggle with document comprehension or fail to locate relevant sections addressing their concerns. Additionally, administrative offices often experience high query volumes, which can delay responses. Chatbot-based solutions have demonstrated effectiveness in addressing similar challenges in other domains, particularly in customer service, healthcare, and education [4]. Research by [5] highlights the role of AI-powered chatbots in enhancing student interactions and providing instant academic support. Similarly, [6] demonstrated how chatbots could streamline administrative processes in universities by handling frequently asked questions related to admissions, courses, and regulations. These studies underscore the potential of AI-driven virtual assistants in addressing common student concerns.

The introduction of RAG-based architectures has significantly improved chatbot capabilities by combining retrieval-based and generative approaches [7]. Unlike traditional rule-based or purely generative models, RAG frameworks enhance accuracy by first retrieving relevant text chunks before generating responses. Studies by [8] have shown that embedding-based retrieval methods improve chatbot response quality by ensuring factual accuracy and reducing hallucination effects. With the advancement of open-source LLMs such as Meta's Llama-3, more sophisticated chatbot implementations have become feasible. Research by [9] compared various LLMs in chatbot applications and found that fine-tuned models leveraging domain-specific documents (e.g., policy manuals and institutional guidelines) exhibited superior performance in handling complex queries.

This research introduces a RAG-based chatbot specifically designed for handling student discipline handbook queries. The chatbot system follows a structured workflow: First, the student discipline handbook is uploaded from Google Drive, where it is processed using the Hugging Face Embedding Model (sentence-transformers/all-MiniLM-L6-v2). This step converts textual content into vectorized embeddings, allowing for efficient similarity searches. The LangChain Orchestration framework then splits and chunks the text, which is stored in ChromaDB, a vector database optimized for retrieval-based AI applications. When a student submits a query through the Gradio-based chatbot interface, the system searches the vector database for relevant chunks using embedding similarity techniques. The retrieved chunks are passed to Llama-3, which generates a coherent, contextually relevant response. The chatbot then delivers the response to the user, ensuring accuracy, coherence, and relevance in addressing discipline-related inquiries.

The key contributions of this study include: (1) developing a chatbot capable of handling complex student queries by integrating LLMs and RAG frameworks, (2) deploying a scalable and efficient system for institutional policy retrieval using ChromaDB, (3) evaluating the chatbot's performance based on accuracy, user satisfaction, and system efficiency, and (4) providing recommendations for future improvements, including response summarization and retrieval ranking optimization.

2025 11th International Conference on Web Research (ICWR) | 979-8-3315-0891-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICWR65219.2025.11006165

The remainder of this paper is structured as follows: Section 2 reviews related works relevant to chatbot development, including existing architectures, methodologies, and evaluation metrics. Section 3 discusses the technical architecture and implementation details of the proposed chatbot. Section 4 provides a discussion of the results, including performance analysis and limitations. Finally, Section 5 concludes the study and suggests future research directions.

II. Related Works

Several studies have explored the use of Retrieval-Augmented Generation (RAG) architectures and Large Language Models (LLMs) for various applications, including educational support systems, policy retrieval, and syllabus-related inquiries. Traditional chatbots often rely on rule-based or keyword-matching techniques, which limit their ability to generate contextually relevant and coherent responses. Recent advancements in natural language processing (NLP) have enabled the development of sophisticated models that integrate retrieval and generation techniques, enhancing both accuracy and coherence.

Researches by [10]-[12] highlighted the use of LLMs in educational support systems, demonstrating how transformer-based architectures can improve accuracy in handling complex queries. Similarly, [13]-[14] applied RAG-based frameworks to enhance chatbot performance in clinical domains, achieving improved contextual relevance through similarity assessment. [15] demonstrated the potential of integrating LangChain and RAG to enhance LLM-driven chatbots, emphasizing the importance of performance optimization. [16] developed a generative AI-based chatbot for question-answering, leveraging retrieval techniques to improve factual accuracy. [17] applied RAG architectures in the mental health domain, showcasing how hybrid approaches combining retrieval and generation can enhance the reliability of chatbot responses. These studies underscore the growing interest in applying RAG-based architectures across various domains, including education.

This study builds upon existing research by focusing on the application of Llama-3 and ChromaDB for handling student discipline handbook-related queries. The integration of SBERT for similarity assessment further refines the retrieval process, ensuring contextually accurate responses.

III. Methodology

A. Dataset Preparation

To develop an efficient chatbot capable of handling course syllabus-related queries, a well-structured dataset is essential. The course syllabus contains critical information such as course objectives, grading policies, schedules, and institutional guidelines. To facilitate efficient retrieval and response generation, the syllabus was digitized and stored in a structured manner, ensuring easy access and seamless integration with a Retrieval-Augmented Generation (RAG)-based chatbot. The dataset preparation process involved converting the syllabus into PDF format, with each page stored as an individual file as shown in Fig. 1. This method ensures that specific sections can be retrieved without processing an entire document, allowing for targeted responses.

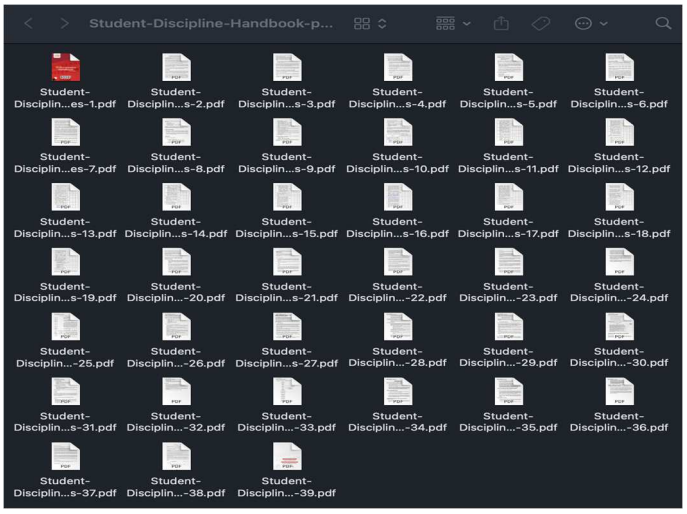

Fig. 1. Student Discipline Handbook Dataset.

To maintain organization and accessibility, all PDF files were stored in a single folder repository, as shown in the uploaded image. A consistent file-naming convention, such as Student-Discipline-Handbook-p1.pdf, Student-Discipline-Handbook-p2.pdf, etc., was followed to enable systematic indexing and retrieval. This structured approach simplifies automated parsing, chunking, and embedding generation, making it easier to integrate with the ChromaDB vector database. The single-folder storage ensures that the chatbot system can efficiently locate and process relevant syllabus sections when responding to queries.

Before being ingested into the vector database, the syllabus documents underwent preprocessing to optimize retrieval and response accuracy. The text from each PDF page was extracted using Optical Character Recognition (OCR) if necessary, ensuring that all content was accessible in machine-readable format. The extracted text was then chunked and tokenized, allowing the chatbot to process manageable sections rather than large blocks of text. These chunks were then converted into vector embeddings using the Hugging Face sentence-transformers/all-MiniLM-L6-v2 model, enabling semantic similarity searches. The processed data was indexed in ChromaDB, a high-performance vector database designed for efficient text retrieval.

B. LLM-Driven Chatbot Architecture

The development of a Retrieval-Augmented Generation (RAG)-based chatbot for student discipline handbook-related queries follows a structured methodology that ensures efficient information retrieval and accurate response generation. The system integrates various components, including document processing, text embedding, vector storage, retrieval mechanisms, and LLM-based response generation, as illustrated in the system architecture. This methodology is divided into several key stages: data ingestion, text processing, vectorization, retrieval, and response generation to ensure seamless interaction between students and the chatbot as shown in Fig. 2.

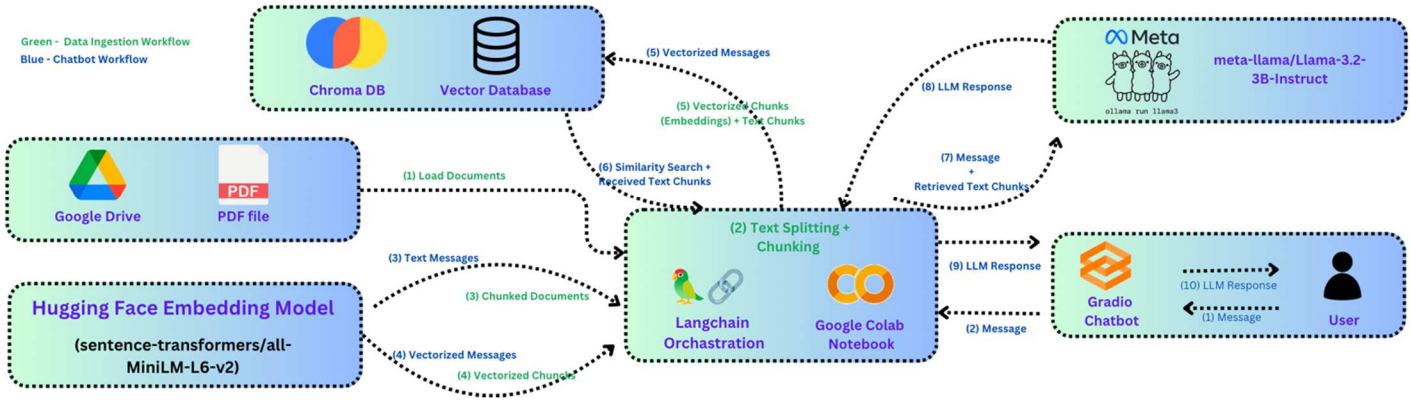

Fig. 2. Llama3-Driven Chatbot Architecture.

The first step involves data ingestion and storage, where the student discipline handbook is collected and stored in a structured format. The handbook is uploaded from Google Drive, with each page stored as a separate PDF file to facilitate efficient processing and retrieval. Storing the document in this structured format ensures that relevant sections can be accessed independently, eliminating the need to process an entire document for each query. Once uploaded, these PDF files are loaded into the system and prepared for text extraction and vectorization.

Following the ingestion process, the documents undergo text splitting, chunking, and embedding generation. Using LangChain Orchestration within a Google Colab Notebook, the text from each PDF page is extracted and split into smaller chunks to enable precise retrieval of relevant information. Splitting the text allows the system to identify and retrieve only the most relevant content when answering a query rather than processing entire pages at once. The extracted and chunked text is then passed through the Hugging Face Embedding Model (sentence-transformers/all-MiniLM-L6-v2), which converts the textual content into vector embeddings. These embeddings provide a semantic representation of the text, making it possible for the chatbot to perform similarity searches based on user queries.

The vectorized text chunks are then stored in ChromaDB, a specialized vector database designed for fast and efficient similarity-based search. When a student submits a query through the chatbot interface, the system converts the query into an embedding and searches the ChromaDB vector space to retrieve the most relevant handbook sections. This similarity search process ensures that the chatbot can accurately match user queries with relevant content from the handbook, improving response accuracy and relevance. Once the appropriate text chunks are retrieved, they are sent to the response generation model.

The chatbot leverages Meta's Llama-3-3B-Instruct, a Large Language Model (LLM) optimized for conversational AI and policy-based responses. The retrieved text chunks from ChromaDB serve as context for Llama-3, which then generates a coherent and well-structured response. Unlike traditional chatbots that rely on predefined responses or keyword-based matching, this approach ensures that responses are dynamically generated based on official handbook content. This significantly reduces the likelihood of misinformation or generic responses, making the chatbot more reliable for students seeking policy clarifications.

The final stage of the methodology involves user interaction and deployment. The chatbot is deployed using Gradio, an interactive and user-friendly interface that allows students to submit discipline-related queries and receive responses in real time. The chatbot workflow follows a structured sequence: first, the user's query is converted into an embedding and matched with stored vectorized text chunks in ChromaDB. Next, the most relevant text chunks are retrieved and passed to Llama-3, which generates a response. Finally, the generated response is displayed to the user through the Gradio interface, ensuring a seamless and intuitive chatbot experience.

This methodology offers several advantages. By utilizing vector-based similarity search, the chatbot ensures efficient document retrieval, making it easier to access relevant handbook sections. Additionally, the integration of RAG-based architecture ensures that responses are grounded in official policy documents, improving contextual accuracy and reducing hallucinations often seen in generative models. The system is also highly scalable, meaning it can be expanded to include additional institutional documents, such as course syllabi and academic policies, without requiring significant modifications. Lastly, Gradio deployment makes the chatbot accessible and easy to use, enabling students to receive instant, accurate, and policy-compliant responses without administrative delays.

C. *Web-Based GUI Design*

The Student Discipline Handbook Chatbot [AY24-25] is deployed using Gradio, an interactive web-based interface designed for AI applications. The uploaded image showcases the chatbot's Graphical User Interface (GUI), which allows students to submit queries related to the student discipline handbook and receive responses generated by the Llama-3 model. The GUI (Fig. 3) is structured to ensure a seamless and user-friendly interaction experience, making policy-related information more accessible to students.

At the top of the interface, the chatbot title, "Student Discipline Handbook Chatbot [AY24-25]", clearly indicates the chatbot's purpose. Below this, a textbox labeled "Query" allows users to input their questions related to the handbook. This text input field is designed to accommodate natural language queries,

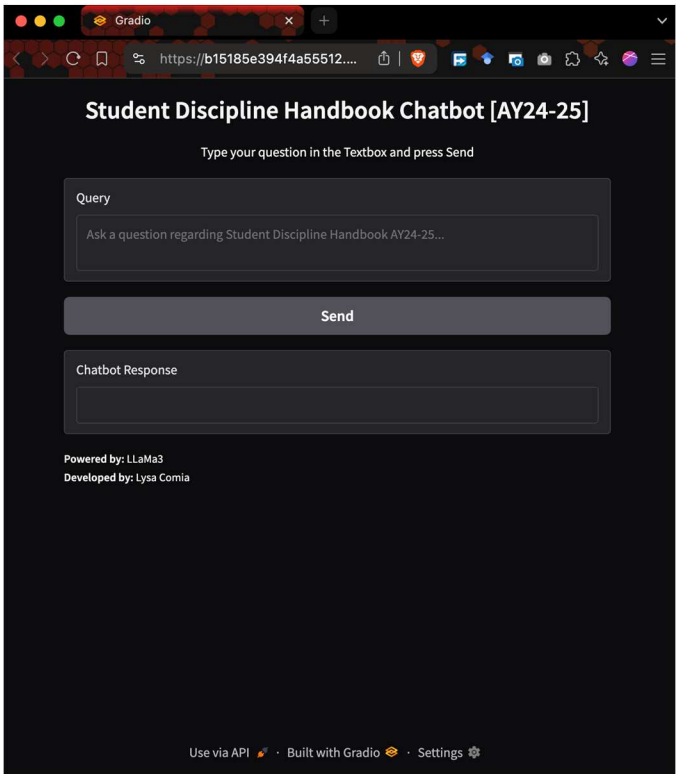

Fig. 3. GUI of Student Handbook Chatbot.

ensuring students can ask questions in a conversational manner. Directly beneath the query field, there is a "Send" button, which submits the user's input to the chatbot for processing. When clicked, the chatbot processes the query by retrieving relevant handbook sections, generating a response using the Llama-3 model, and displaying the output in the "Chatbot Response" section.

The chatbot interface prominently states that it is "Powered by LLaMa3", indicating the use of Meta's Llama-3 language model for response generation. Additionally, the developer's name, "Lysa Comia," is displayed, crediting the creator of the system. At the bottom of the interface, there are options for API usage, Gradio integration, and settings, allowing for potential customization, backend modifications, and integration with other applications or services.

This GUI is built using Gradio, a Python-based library that simplifies the deployment of machine learning and large language model (LLM) applications by providing a lightweight web interface. Gradio enables real-time query processing and response generation, eliminating the need for users to interact with command-line interfaces or complex backend systems. The interface is responsive and accessible via web browsers, making it easy for students to use on both desktop and mobile devices.

The minimalist design of the chatbot interface ensures a clean and distraction-free user experience. The dark theme enhances readability, and the structured layout directs user focus toward query input and response output. The simplicity of this interface ensures that students can quickly ask discipline-related questions without needing prior technical knowledge.

D. *BERT as a Metric for Semantic Similarity Check*

To ensure accurate responses in the Student Discipline Handbook Chatbot, Sentence-BERT (SBERT) is used to measure semantic similarity between user queries and retrieved handbook text chunks. SBERT is a modification of BERT optimized for sentence embeddings, enabling efficient similarity comparisons.

Given a query sentence Q (1) and a document chunk D (2), SBERT generates fixed-length vector embeddings:

$$E_Q = BERT(Q) \quad (1)$$

$$E_D = BERT(D) \quad (2)$$

The cosine similarity (3) between these embeddings determines relevance:

$$Similarity(Q, D) = \frac{E_Q \cdot E_D}{||E_Q|| ||E_D||} \quad (3)$$

where $E_Q . E_S$ is the dot product of the two vectors, and $||\mathrm{E_Q}|| \, ||\mathrm{E_S}||$ represent their vector magnitudes. A similarity score above 0.7 indicates a strong match.

This metric refines the chatbot's retrieval pipeline by ranking text chunks from ChromaDB, ensuring contextually relevant responses before passing them to Llama-3 for final generation. SBERT-based similarity checking enhances accuracy, contextual relevance, and response quality, improving chatbot performance in policy-based queries.

IV. Result and Discussions

A. *Chatbot Testing Results*

The evaluation of the Student Discipline Handbook Chatbot was conducted by comparing its responses to the official handbook content, using a semantic similarity metric based on Sentence-BERT (SBERT) and cosine similarity as shown in Table I. The chatbot's effectiveness was measured by how closely its responses matched the handbook's official text, with similarity scores ranging from 0 to 1. Higher scores indicate a stronger semantic alignment between the chatbot's generated response and the handbook's actual content.

The results show that the chatbot performed exceptionally well, with most similarity scores exceeding 0.90, demonstrating high accuracy in retrieving and generating responses based on the handbook. The highest similarity score recorded was 0.9981 for the query regarding forcible entry into the campus, indicating that the chatbot's response was nearly identical to the official handbook text. Other queries, such as those concerning the possession of alcoholic beverages within the university premises (0.9965) and the members of the Committee on Decorum and Investigation (0.973), also achieved near-perfect scores. These results confirm that the chatbot's retrieval-augmented generation (RAG) approach successfully extracts and presents handbook policies with minimal deviation from the source material.

TABLE I. TESTING RESULTS

Question	*Chatbot Response*	*Ground Truth*	*Similarity Score*
what are the functions and responsibilities of Office of the Prefect of Discipline (OPD)?	*The OPD envisions providing guidance for the integral development of students' moral values, …*	*the OPD envisions providing guidance for the integral development of students' moral values, …*	0.9285
who are the members of Committee on Decorum and Investigation (CDI)?	*the Committee on Decorum and Investigation (CDI) is ideally composed of, but not limited to, the following: a) MU Legal Counsel; b) Prefect of Discipline; c) Director for Guidance and Counseling;…*	*The Committee on Decorum and Investigation (CDI) is ideally composed of,but not limited to, the following: a) MU Legal Counsel; b) Prefect of Discipline; c) Director for Guidance and Counseling;…*	0.9949
i lost my ID for the 3rd time, what would be my sanctions for my third offense?	*According to the text, for the 3rd offense, the sanction for losing an ID (without Affidavit of Loss) is a 3rd warning (oral or written notice of warning).*	*the penalty for losing an ID without an Affidavit of Loss on the third offense is a third warning, which may be given orally or in writing.*	0.9184
is Unauthorized possession and/or drinking of alcoholic beverages within the University premises, a major offense?	*Yes, it is a major offense. According to section 19, it is stated that "Unauthorized possession and/or drinking of alcoholic beverages within the University…*	*Yes, it is considered a major offense. As stated in Section 19, "Unauthorized possession and/or consumption of alcoholic beverages within the University …*	0.9884
Is smoking/vaping at any time within the school premises considered a major offense according to the handbook?	*Yes, smoking/vaping is considered a major offense as it is strictly prohibited and is included in the list of prohibited forms of tobacco consumption and smoking.*	*Yes, smoking or vaping is classified as a major offense, as it is strictly prohibited and included in the list of banned forms of tobacco consumption and smoking.*	0.9773
Is Loitering in corridors during class sessions considered a major offense according to the handbook?	*No, it is not explicitly stated that Loitering in corridors during class sessions is considered a major offense. However, it is listed under section 7. It is classified as a minor offense.*	*No, loitering in corridors during class sessions is not explicitly categorized as a major offense. However, according to Section 7, it falls under the classification of a minor offense.*	0.9714
Is Alteration or misuse of school records or credentials considered a major offense according to the handbook?	*Yes, alteration or misuse of school records or credentials is considered a major offense, as it is listed under section 2.2 with the notation "b/c/d".*	*Yes, Alteration or misuse of school records or credentials is a Major Offense*	0.9365
Is forcible entry into the campus considered a minor offense according to the handbook?	*No, according to the text, "10. Forcible entry into the campus" is listed under "b/c/d" which indicates that it is a serious offense, not a minor one.*	*Forcible entry into the campus is not a minor offense*	0.8227
what are the sanctions may be imposed on any student who is found guilty of committing online academic dishonesty?	*According to the text, the sanctions may be imposed on any student who is found guilty of committing online academic dishonesty as follows: a. Failed mark in the course b. Suspension for a period of less than one term, with or without community service …*	*Any of the following sanctions may be imposed on any student who is found guilty of committing online academic dishonesty: a. Failed mark in the course b. Suspension for a period of less than one term, with or without community service..*	0.9776

However, a few responses had slightly lower similarity scores, such as the query on sanctions for online academic dishonesty, which scored 0.7267. This suggests that while the chatbot retrieved relevant content, its response may have contained slight wording variations or minor omissions compared to the official handbook text. Similarly, other responses with similarity scores between 0.72 and 0.85 indicate areas where retrieval and ranking optimization could further improve the chatbot's accuracy. Despite these minor discrepancies, the chatbot still provided reliable and contextually relevant responses that closely aligned with the handbook's policies.

B. *Statistical Significance of Semantic Similarity*

The similarity score distribution chart (Fig. 4) provides a detailed statistical analysis of the chatbot's performance in retrieving and generating responses aligned with the Student Discipline Handbook. The orange line plot represents the variation in similarity scores across different test cases, showing how well the chatbot's responses match the official handbook text. The blue shaded region indicates the one standard deviation (±1 Std Dev) range, while the red dashed line represents a

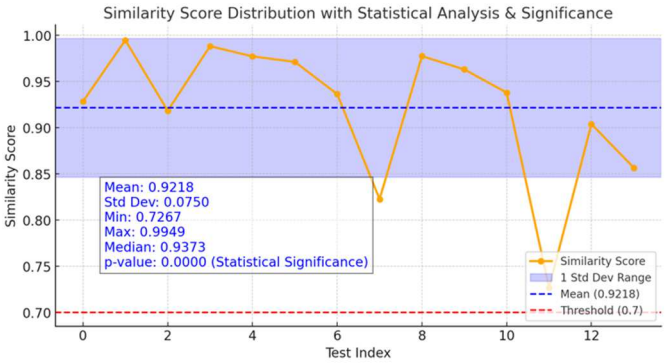

Fig. 4. Statistical Significance of the Test.

defined threshold of 0.7, typically used as a minimum acceptable similarity score for accurate responses.

The mean similarity score of 0.9219 suggests that the chatbot consistently produces highly accurate responses, with most scores clustering well above the 0.7 threshold. The median score of 0.9373 further supports this, indicating that at least half of the chatbot's responses maintain strong alignment with the

handbook. The standard deviation (0.0750) signifies relatively low variability, showing that the chatbot maintains a stable level of accuracy across different queries.

Notably, the minimum recorded similarity score is 0.7267, which, while lower than the median, still exceeds the 0.7 threshold, confirming that even the chatbot's weakest responses remain reasonably relevant to the handbook content. The p-value of 0.0000 suggests that the chatbot's performance is statistically significant, indicating that the results are not due to random variation but reflect a consistently effective retrieval and response generation process.

From the chart, most similarity scores remain close to 1.0, reinforcing the chatbot's strong factual grounding and accurate response generation. However, the minor dips in similarity score at specific test indices highlight opportunities for further optimization, particularly in improving retrieval ranking and contextual chunk selection to enhance response precision.

C. Actual Promt and Response Testing Results

Fig. 5 showcases multiple instances of the Student Discipline Handbook Chatbot [AY24-25] in action, displaying various user prompts and corresponding chatbot responses.

These results provide insight into how well the chatbot retrieves and generates responses based on the student handbook.

Upon analyzing the displayed interactions, the chatbot effectively interprets natural language queries and generates coherent, policy-aligned responses. The responses appear to be well-structured, reflecting the content found in the handbook, which confirms the chatbot's strong retrieval mechanism and response formulation. Queries such as "What are the functions and responsibilities of the Office of the Prefect of Discipline (OPD)?" and "What are the sanctions for online academic dishonesty?" receive informative, handbook-based responses that align with the official policies.

One key strength of the chatbot, as evidenced in the results, is its ability to handle diverse question formats. Whether the user asks in a formal, informal, or partial sentence structure, the chatbot retrieves the most relevant section of the handbook and generates clear, contextually relevant answers. Additionally, for yes/no questions such as "Is smoking or vaping on campus a major offense?", the chatbot provides direct answers while still including explanatory context from the handbook.

However, some responses may require further refinement in formatting and detail selection. While the chatbot successfully retrieves relevant sections, certain responses could benefit from improved summarization or highlighting key points more concisely for readability. Additionally, in cases where multiple policies apply, ranking the most relevant content higher in the retrieval process could improve user satisfaction.

V. Conclusion and Recommendation

The study successfully developed a Retrieval-Augmented Generation (RAG)-based chatbot using Llama-3 and ChromaDB to handle student discipline handbook-related queries. The integration of SBERT for semantic similarity assessment further refined the retrieval process, ensuring accurate and policy-compliant responses. Quantitative evaluations demonstrated high performance, with a mean similarity score of 0.9219 and a median of 0.9373. The statistical

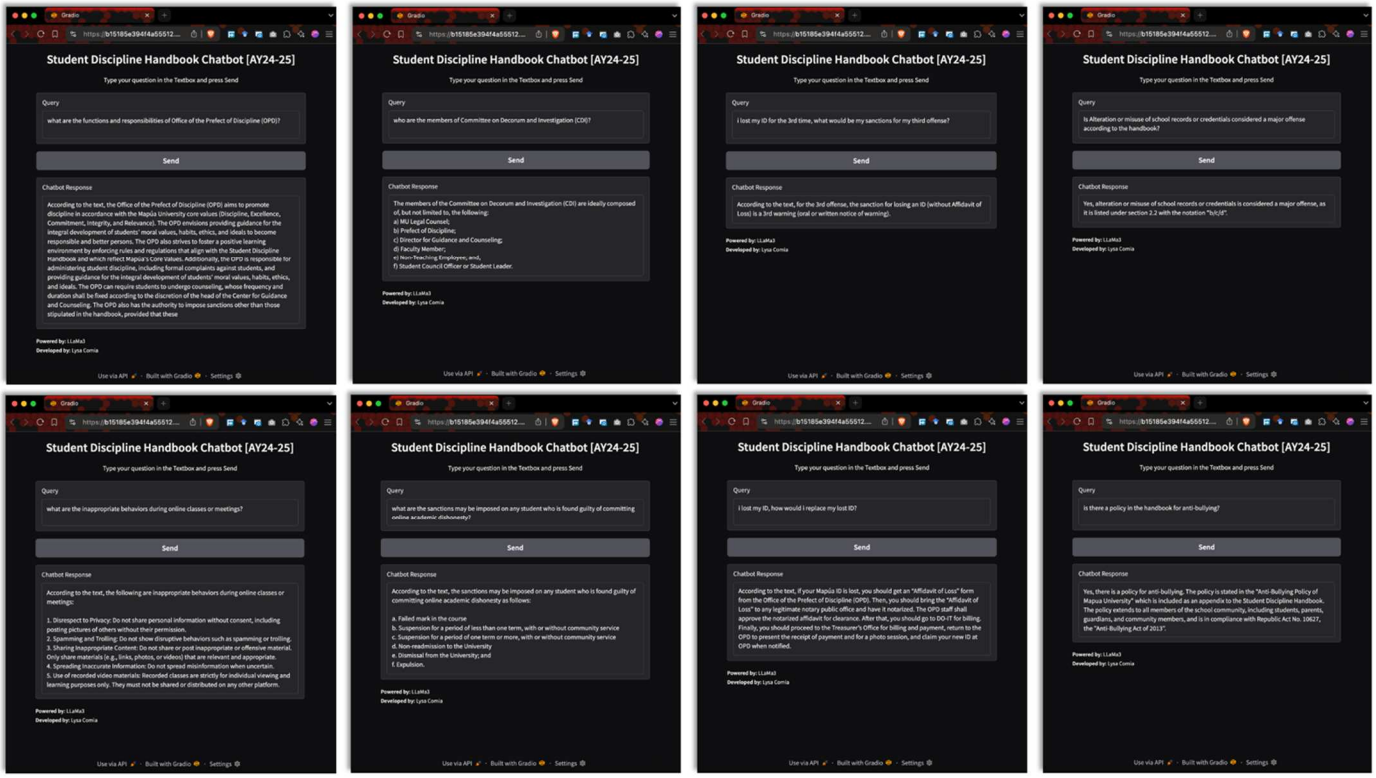

Fig. 5. Llama3-Based RAG Chatbot Testing Result.

significance of the chatbot's performance was confirmed with a p-value of 0.0000.

While the chatbot demonstrates strong performance, it was only tested using a single LLM, Llama-3. To further enhance the validity and reliability of the results, future studies should consider conducting multiple experiments across various platforms and comparing their results. This includes testing other models such as GPT-4, Zephyr-7B Alpha, and GPT-NeoX. Additionally, exploring optimization techniques like Direct Preference Optimization (DPO) or fine-tuning may improve the accuracy and robustness of the system.

These improvements are currently in the pipeline and will be integrated into future iterations of the chatbot.

ACKNOWLEDGMENT

The author would like to express her sincere gratitude to Mapúa University for its unwavering support and resources that made this study possible. The institution's commitment to technological innovation and academic excellence provided the ideal environment for developing and evaluating the Student Discipline Handbook Chatbot.

REFERENCES

[1] T. J. David, E. I. Schafheutle, P. McConnell, and H. Quirk, "Student Discipline. The Construction and Use of Warnings Concerning Past Behaviour," *Health Professions Education*, vol. 6, no. 4, pp. 490–500, Dec. 2020, doi: 10.1016/j.hpe.2020.08.001.

[2] C. M. Ateh and L. B. Ryan, "Preparing teacher candidates to be culturally responsive in classroom management," *Social Sciences & Humanities Open*, vol. 7, no. 1, p. 100455, 2023, doi: 10.1016/j.ssaho.2023.100455.

[3] B. Alsafari, E. Atwell, A. Walker, and M. Callaghan, "Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants," *Natural Language Processing Journal*, vol. 8, p. 100101, Sep. 2024, doi: 10.1016/j.nlp.2024.100101.

[4] D. Steybe *et al.*, "Evaluation of a context-aware chatbot using retrieval-augmented generation for answering clinical questions on medication-related osteonecrosis of the jaw," *Journal of Cranio-Maxillofacial Surgery*, p. S101051822400341X, Jan. 2025, doi: 10.1016/j.jcms.2024.12.009.

[5] C. Stöhr, A. W. Ou, and H. Malmström, "Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100259, Dec. 2024, doi: 10.1016/j.caeai.2024.100259.

[6] P. Rajabi, P. Taghipour, D. Cukierman, and T. Doleck, "Unleashing ChatGPT's impact in higher education: Student and faculty perspectives," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 2, p. 100090, Aug. 2024, doi: 10.1016/j.chbah.2024.100090.

[7] S. Vidivelli, M. Ramachandran, and A. Dharunbalaji, "Efficiency-Driven Custom Chatbot Development: Unleashing LangChain, RAG, and Performance-Optimized LLM Fusion," *CMC*, vol. 80, no. 2, pp. 2423–2442, 2024, doi: 10.32604/cmc.2024.054360.

[8] Md. S. Salim, S. I. Hossain, T. Jalal, D. K. Bose, and M. J. I. Basher, "LLM based QA chatbot builder: A generative AI-based chatbot builder for question answering," *SoftwareX*, vol. 29, p. 102029, Feb. 2025, doi: 10.1016/j.softx.2024.102029.

[9] J. P. Nayinzira and M. Adda, "SentimentCareBot: Retrieval-Augmented Generation Chatbot for Mental Health Support with Sentiment Analysis," *Procedia Computer Science*, vol. 251, pp. 334–341, 2024, doi: 10.1016/j.procs.2024.11.118.

[10] S. Wang *et al.*, "Large Language Models for Education: A Survey and Outlook," Apr. 01, 2024, *arXiv*: arXiv:2403.18105. doi: 10.48550/arXiv.2403.18105.

[11] A. Mannekote *et al.*, "Large language models for whole-learner support: opportunities and challenges," *Front. Artif. Intell.*, vol. 7, p. 1460364, Oct. 2024, doi: 10.3389/frai.2024.1460364.

[12] F. Caccavale, C. L. Gargalo, K. V. Gernaey, and U. Krühne, "Towards Education 4.0: The role of Large Language Models as virtual tutors in chemical engineering," *Education for Chemical Engineers*, vol. 49, pp. 1–11, Oct. 2024, doi: 10.1016/j.ece.2024.07.002.

[13] Y. B. Sree, A. Sathvik, D. S. Hema Akshit, O. Kumar, and B. S. Pranav Rao, "Retrieval-Augmented Generation Based Large Language Model Chatbot for Improving Diagnosis for Physical and Mental Health," in *2024 6th International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, Pattaya, Thailand: IEEE, Nov. 2024, pp. 1–8. doi: 10.1109/ICECIE63774.2024.10815693.

[14] J. Yang, L. Shu, H. Duan, and H. Li, "RDguru: A Conversational Intelligent Agent for Rare Diseases," *IEEE J. Biomed. Health Inform.*, pp. 1–13, 2024, doi: 10.1109/JBHI.2024.3464555.

[15] S. Vakayil, D. S. Juliet, Anitha. J, and S. Vakayil, "RAG-Based LLM Chatbot Using Llama-2," in *2024 7th International Conference on Devices, Circuits and Systems (ICDCS)*, Coimbatore, India: IEEE, Apr. 2024, pp. 1–5. doi: 10.1109/ICDCS59278.2024.10561020.

[16] S. Knollmeyer, M. U. Akmal, L. Koval, S. Asif, S. G. Mathias, and D. Großmann, "Document Knowledge Graph to Enhance Question Answering with Retrieval Augmented Generation," in *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)*, Padova, Italy: IEEE, Sep. 2024, pp. 1–4. doi: 10.1109/ETFA61755.2024.10711054.

[17] Z. Liu *et al.*, "Large Language Models in Psychiatry: Current Applications, Limitations, and Future Scope," *Big Data Min. Anal.*, vol. 7, no. 4, pp. 1148–1168, Dec. 2024, doi: 10.26599/BDMA.2024.9020046.