

Development and Implementation of the Primakara Virtual Assistant Based on Generative Artificial Intelligence

Made Adi Paramartha Putra¹, I Putu Buda Suyasa², I Made Artana³, Nengah Widya Utami⁴

^{1,2,4} Faculty of Information Technology and Design, Primakara University, Indonesia

³ Faculty of Economy and Business, Primakara University, Indonesia

Article Info

Article history:

Received May 21, 2025

Revised Jul 19, 2025

Accepted Oct 15, 2025

Keywords:

Generative AI

Virtual Assistance

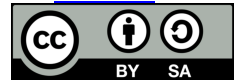
Natural Language Processing

Educational Technology

ABSTRACT

The growing need for efficient, accessible, and context-aware academic support systems has led to the exploration of Generative AI (GenAI) technologies in educational settings. However, existing virtual assistants often lack contextual relevance, adaptability, and user-friendly interaction, limiting their effectiveness in higher education environments. This study proposes a GenAI-based Virtual Assistant (VA) tailored for university-related applications, combining voice recognition, natural language understanding, and text-to-speech technologies to create an interactive and intelligent support system. The proposed work was evaluated through four key testing stages: black-box functionality testing, response similarity analysis, inference time measurement, and user acceptance testing. Black-box testing validated the system's ability to process speech input, generate accurate audio responses, and provide responsive UI/UX feedback. A TF-IDF cosine similarity analysis across 11 academic departments yielded an average similarity score of 81.86%, demonstrating semantic alignment with institutional content. The system also maintained an average response time of 3.88 seconds. User feedback from 25 participants revealed high satisfaction levels, with scores exceeding 4.0 across all indicators and large T-statistic value. These results confirm the system's potential as an effective, real-time academic assistant.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Made Adi Paramartha Putra,

Faculty of Information Technology and Design,

Primakara University,

Jalan Tukad Badung No. 135, Renon, Denpasar Selatan, Kota Denpasar, Bali 80226 Indonesia.

Email: adi@primakara.ac.id

1. INTRODUCTION

The rapid evolution of Artificial Intelligence (AI), particularly Generative Artificial Intelligence (GenAI), has catalyzed profound changes across diverse sectors, including education, healthcare, and customer service. Widely known GenAI models, such as ChatGPT, Gemini, and Claude, are capable of generating human-like responses, crafting written content, and engaging in interactive dialogue. These advancements have made AI tools increasingly attractive across various industries, including educational institutions that seek to enhance teaching, streamline administrative services, and improve student engagement. Despite its immense potential, the adoption of GenAI in education, particularly in regions like Bali remains limited and largely experimental. Most educational institutions have yet to integrate GenAI into their core academic or operational systems. Current usage is often restricted to basic AI chatbots or web application that lack contextual understanding and are not tailored to the institution's unique needs. Moreover, educational GenAI deployments are frequently limited to simulations, with few implementations being fully integrated into operational workflows.

The implementation of GenAI in the educational sector can be categorized into three main areas: teaching, evaluation, and administration. In higher education, GenAI is most frequently used for tasks such as information retrieval and text paraphrasing [1]. These functions fall under the teaching domain, where students adopt new learning method, although such usage should be guided and monitored to maintain academic integrity. In the context of evaluation, GenAI is increasingly utilized to assess student performance, ranging from task personalization to accelerating the grading process. To ensure the effective and ethical adoption of GenAI in both learning and assessment, a strong foundation in digital literacy and a holistic approach to the entire educational process are essential [2].

The adoption of GenAI in educational settings faces several significant limitations that hinder its widespread and effective implementation. One major concern is data privacy and security. Educational institutions manage sensitive information about students, staff, and internal processes, and integrating this data into GenAI systems, especially those hosted on third-party platforms raises ethical and legal challenges. Additionally, most pre-trained GenAI models are developed on general datasets and lack contextual awareness, making them incapable to handle institution-specific queries. Technical barriers also play a critical role, as many schools and universities, particularly in Bali province, lack the necessary infrastructure, such as robust servers or secure cloud environments, and do not have adequately trained personnel to maintain or manage AI systems. Furthermore, the integration of GenAI into educational environments brings concerns related to academic integrity, including the potential for students to misuse AI-generated responses, which may undermine originality and critical thinking [3]. Lastly, user acceptance and readiness are often overlooked, faculty and staff may be hesitant to adopt new technologies due to a lack of familiarity or fear of being replaced, which can further delay or obstruct successful deployment. Together, these limitations highlight the need for carefully planned, context-aware, and ethically responsible strategies when implementing GenAI in education.

Given these limitations, this research focuses on bridging the gap between theoretical potential and practical implementation by developing a customized GenAI-based Virtual Assistant (VA). The system is not only capable of interacting in natural language but is also contextually aware of internal processes, departments, and student needs. To mitigate current limitations related to data privacy and security, limited contextual understanding, and concerns over academic integrity, this research adopts a localized and controlled deployment strategy using institution-specific datasets and access restrictions. To the best of our knowledge, there has been no adoption of a VA-based approach for administrative tasks in higher education institutions. Thus, the major contributions of this work are detailed as follows:

- We present a comprehensive approach to document preprocessing and formatting for better AI ingestion, which can serve as a guideline for similar institutional implementations.
- We develop a GenAI model based on Llama3.3 that is contextually trained using curated documents from multiple departments at Primakara University, ensuring accurate and relevant responses.
- We integrate voice-to-text and text-to-speech features in the VA interface to provide a more accessible and natural user experience for users interacting via speech.
- Unlike previous works that remain at the conceptual or prototype stage, this research culminates in the deployment of a working system embedded within a website and intended to operate on microcontroller-based kiosks on campus.
- We perform intensive performance comparison of the proposed GenAI-based VA using various performance metrics, such as blackbox, cosine and semantic similarity test, along with the system delay and user feedback.

The remainder of this article is organized as follows: Section 2 provides a comprehensive overview of the state-of-the-art adoption of GenAI in the education sector. The proposed architecture and optimization techniques for the GenAI-based VA at Primakara University are presented in Section 3. Section 4 discusses the performance evaluation results. Finally, Section 5 concludes the research.

2. RELATED WORKS

The integration of GenAI into education has introduced transformative possibilities, especially in fostering inquiry-based learning and curiosity among students [4]. Large language models like GPT-3 have been shown to simulate educational prompts that replicate a child's natural curiosity and encourage complex questioning skills. This aligns with the broader objective of GenAI to instill a spirit of inquiry and enhance engagement in learning environments where traditional methods often fall short in capturing student attention or promoting active participation [5].

Personalized learning is another domain where GenAI demonstrates considerable promise. Tools like ChatGPT have been used to tailor instructional content based on learners' profiles, pace, and preferences, overcoming limitations of one-size-fits-all teaching models. This adaptive approach is particularly beneficial for students from diverse linguistic or cultural backgrounds, as GenAI technologies can assist them in adjusting to different pedagogical styles while offering individualized feedback and progression paths [6]. Such personalization has been emphasized as a means to enhance inclusivity and promote equitable learning experiences across educational settings [7].

Moreover, GenAI has proven valuable in creating immersive and interactive learning environments. By enabling the development of simulations, virtual labs, and gamified content, GenAI increases student engagement and facilitates deeper understanding of complex subjects. The transformative potential of ChatGPT in revitalizing traditional classroom experiences has been explored, while other studies have shown that GAI-generated content can sustain learner interest through interactive storytelling, animations, and AI-driven virtual assistants [8].

In higher education and specialized fields such as medicine and engineering, GenAI serves as a powerful tool for visualization and complex concept clarification. Applications like DALL-E and Midjourney are employed to generate educational visuals, while natural language models assist in collaborative learning and virtual peer discussions [9][10]. Lastly, GenAI plays a pivotal role in enhancing the role of educators by automating routine tasks such as quiz generation, assignment grading, and feedback delivery. The LearnLM-Tutor model has been introduced as a tool that shows improved pedagogical outcomes and can be used to evaluate student understanding across multiple dimensions [11]. Furthermore, GenAI enables continuous teacher development through curated educational resources and lesson planning tools [12]. As a result, educators are empowered to focus more on high-level instruction and student mentorship, illustrating how GenAI can support and elevate teaching as a profession.

Despite the growing adoption of GenAI in educational contexts, several limitations remain, particularly concerning the deployment of GenAI-based VA. One major concern is the lack of contextual awareness in many GenAI models, which may result in vague or misleading responses when addressing domain-specific academic content. Usability challenges are also prevalent; many GenAI VA systems still require manual activation or rigid input formats, limiting natural conversational flow.

3. PROPOSED PRIMAKARA VIRTUAL ASSISTANCE

This research proposes a comprehensive framework for the adoption of a GenAI-based VA in a university setting, specifically Primakara University. The framework is structured into four key phases: (1) data collection and preprocessing, (2) model customization, (3) system integration and deployment, and (4) evaluation and feedback. This methodological design ensures not only technical feasibility but also addresses concerns related to data privacy, contextual relevance, and academic integrity. The overall research process depicted in Figure 1.

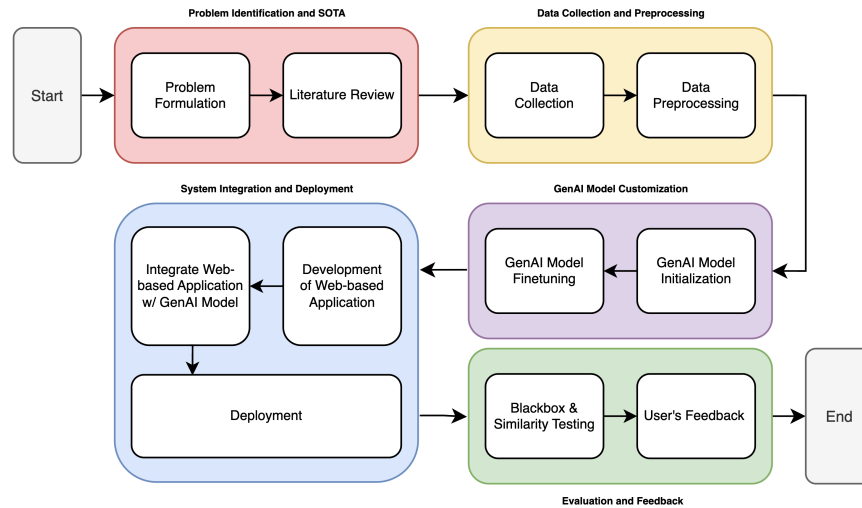


Figure 1. The proposed GenAI-based VA framework for Primakara University.

3.1. Data Collection and Preprocessing

The first phase focuses on gathering and preparing institutional data to ensure high-quality and contextually accurate AI responses. Data is sourced from internal documents, including standard operating procedures (SOPs), frequently asked questions (FAQs), academic calendars, and service workflows. Documents are prepared by each department at Primakara University, allowing for easier investigation in cases where the GenAI system generates misinformation. All relevant information is stored in .docx format to ensure compatibility and readability during the training process.

In some cases, these documents contain figures or visual representations related to specific processes or steps. Since AI models tend to be less accurate in interpreting figures, accompanying descriptions are required to provide context. Therefore, each department is also responsible for reviewing and rewriting the related documents to ensure that the information provided to the GenAI system is both clear and accurate. At the end of the day, the total collected data from 11 different department with different document count. The detailed information about department, and types of document detailed in Table 1.

Table 1. Document and Token Summary Before and After Preprocessing

No.	Description	Initial Document Count	Initial Token	Final Document Count	Final Token
1	Faculty of IT & Design	12	66113	8	31263
2	Faculty of Economy and Business	10	25738	7	15123
3	Postgraduate Department	2	15346	2	9230
4	Quality Assurance Department	1	444	1	444
5	Marketing Department	52	152438	2	23873
6	IT Department	2	3777	2	3777
7	Public Relation Department	6	15436	2	10234
8	Vice Chancellor - 1 (Academic)	14	62323	14	42323
9	Vice Chancellor - 2 (Human Resource)	4	6052	4	6052
10	Vice Chancellor - 3 (Student Affairs)	5	7234	5	7234
11	Vice Chancellor - 4 (Innovation)	4	4230	4	4230
Total		112	362,131	51	149,783

This work represents the initial development phase of a GenAI-based Virtual Assistant. We utilize the Llama 3.3 base model. To ensure the model operates efficiently as a foundation for future advancements, we limit the total number of context under 150,000 token. Table 1 outlines the initial documents and their corresponding token representations in the final processed dataset. To ensure that essential administrative information relevant to student needs is prioritized, we excluded several documents containing less frequently requested information based on feedback and observations from the university's reception desk. Additionally, we minimize repetitive

information found across various documents by reviewing all the original materials. It is worth noting that the token calculation was performed using the Llama tokenizer, which converts text input into smaller units known as tokens [13].

3.2. Model Customization

Developing a GenAI system, particularly one based on a Large Language Models (LLMs), typically requires significant computational resources, time, and large-scale datasets. To address these constraints, this research adopts the Llama model using a contextual-based learning approach rather than full model retraining [14]. This strategy enables us to utilize the power of a pre-trained model while adapting it to specific institutional needs with greater efficiency.

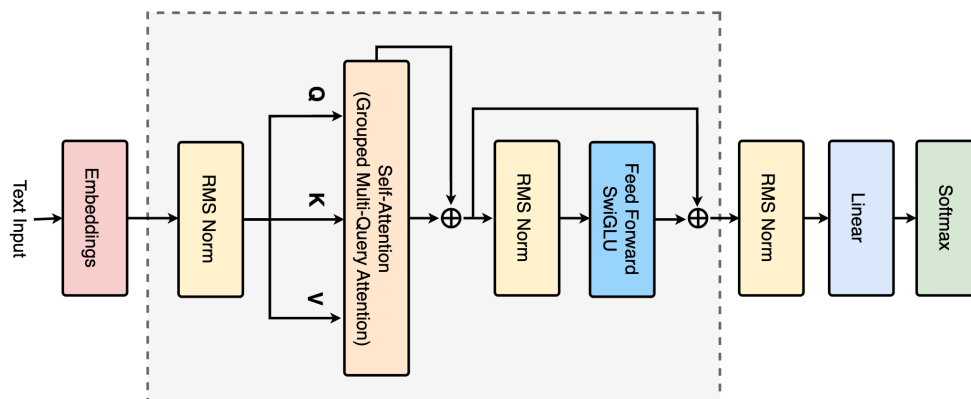


Figure 2. The vanilla architecture of the Llama model used in this work.

In this study, the Llama architecture was employed for language modeling tasks. The model begins with text inputs that are transformed into vector representations through an embedding layer. These embeddings are then normalized using RMSNorm, a variant of normalization that improves training stability. A grouped multi-query self-attention mechanism is applied to capture contextual dependencies within the sequence. This mechanism is preceded by RMS normalization and utilizes shared key and value projections for efficiency, while maintaining separate query projections. The output of the self-attention module is combined with the input through a residual connection.

Subsequently, another RMS normalization layer is applied before the output is passed into a feed-forward network, which incorporates SwiGLU activation to enhance non-linearity and learning capacity. The resulting output undergoes another residual connection and is normalized once more. Finally, a linear projection is applied to transform the hidden states into logits, which are then passed through a softmax layer to obtain the output probabilities. This architecture emphasizes efficiency through grouped attention and normalization while maintaining performance across language modeling tasks.

The finalized dataset, after extensive preprocessing, is incorporated into the model pipeline through retrieval-augmented generation (RAG) [15]. In the proposed work, the model dynamically accesses relevant documents from a structured knowledge base at query time using semantic search, which leverages vector representations of the data. These retrieved documents are embedded as part of the prompt context, allowing the model to generate responses that are both relevant and factually aligned with the university's internal information. This method improves both precision and adaptability, especially for answering localized and procedural questions that would not be available in the model's original training data. This work adopts the LangChain library, which provides extensive tools for integrating with the Ollama model. In the context of the implemented RAG pipeline, the system utilizes traditional LangChain settings, incorporating OpenAIEmbeddings and a Top-K vector similarity-based retriever. It is important to note that the retrieved documents are restricted to the latest versions listed in Table 1, with no interaction with external internet sources.

Additionally, this process can also be interpreted as a form of lightweight fine-tuning, where the base Llama model's capabilities are enhanced by injecting institution-specific knowledge without altering its core parameters. This approach strikes a balance between performance, cost-efficiency, and data governance, particularly in environments with limited infrastructure or sensitivity around internal documents [16].

Customization also involves prompt engineering to improve the model's responsiveness and control over tone and behavior. For example, specific system prompts and instructions are used to ensure that the assistant responds in a formal, helpful, and non-opinionated manner, aligned with academic communication standards. Additionally, role definition techniques are applied to specify the assistant's domain, positioning it as a university information agent rather than a general-purpose chatbot. This helps minimize hallucination and ensures responses stay within the intended scope.

3.3. System Integration and Deployment

This phase involves integrating the GenAI model into a user-accessible web application. The system utilizes a client-server architecture, with the front end built using ReactJS and the backend managed through Flask [17]. Speech recognition is enabled to allow voice input, and responses are converted back into speech [18]. In this work, we adopted Google's Indonesian-compatible engine to perform the text-to-speech process. The proposed VA platform is deployed within a controlled environment, initially accessible through a web portal, with future plans to expand to microcontroller-powered kiosks installed on campus. Access control mechanisms and usage logging are implemented to enhance both security and transparency.

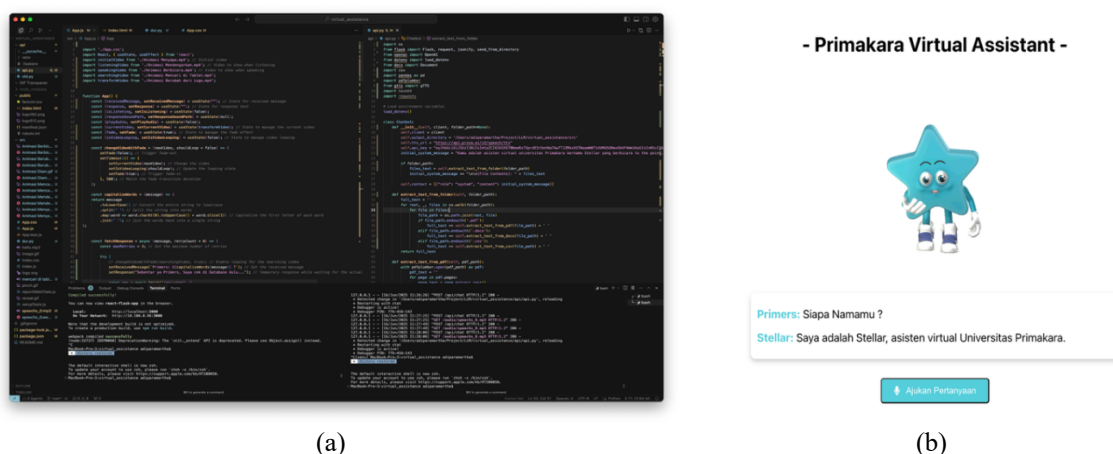


Figure 3. Codebase and User Interface of the Proposed Virtual Assistance
(a) Codebase with Flask and ReactJS, (b) Response Page UI

The codebase of the proposed system built on top of Flask and ReactJS framework along with the response UI of the proposed VA are illustrated in Figure 3. These interfaces are designed to guide users intuitively through their interactions with the system. To enhance the overall user experience and create a more engaging interaction, we introduced a virtual character named Stellar to represent the VA. Stellar serves as a visual and symbolic embodiment of the assistant, helping to humanize the interaction, especially for users unfamiliar with AI-based systems. To further improve usability and create a dynamic interface, animated visual cues were implemented. These animations indicate when the VA is listening, processing, and responding to user input, providing users with clear feedback and reinforcing a sense of real-time interaction. This feedback loop is especially important in voice-based systems, as it helps maintain user confidence that their requests are being actively handled. Such enhancements are aimed at increasing user satisfaction, reducing hesitation, and promoting sustained engagement with the system.

In terms of deployment strategy, a local development scenario was chosen over a cloud-hosted solution. This decision was made primarily to reduce operational costs and avoid the

inefficiencies of hosting a high-demand, resource-intensive system entirely online. Running a GenAI model, especially one with a relatively large parameter size requires substantial computing power, memory, and bandwidth. In high-traffic environments, such as a university with hundreds or thousands of users, relying solely on cloud resources could lead to latency issues and increased expenses. By adopting a local deployment approach, we maintain greater control over system performance, ensure faster response times, and uphold better data privacy standards, especially when handling sensitive institutional information.

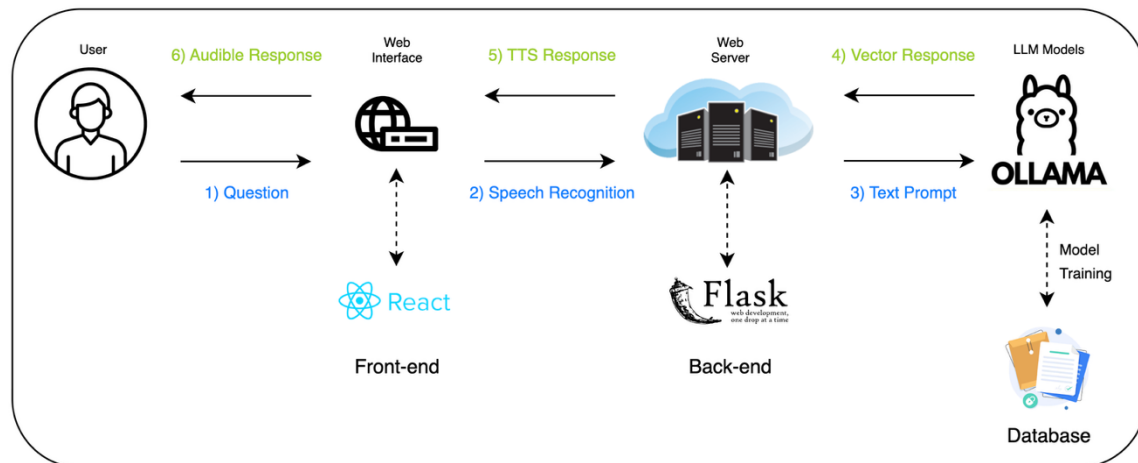


Figure 4. Workflow of the proposed GenAI-based VA system.

Furthermore, the workflow of the proposed GenAI-based VA depicted in Figure 4. The interaction process is initiated when a user submits a spoken query through a web interface. This input is captured and processed using a speech recognition module on the front end, which is built with React. The recognized speech is then converted into a text prompt and transmitted to the back-end server, which is implemented using Flask. Upon receiving the prompt, the back end forwards the query to a LLM hosted in localhost via the Llama. This LLM has been previously trained on institution-specific data retrieved from a structured database to ensure contextual accuracy. The generated response, in vector format, is returned from the LLM to the back-end server. This response is subsequently converted into speech by a TTS module. Finally, the audible answer is delivered back to the user via the web interface, completing the interaction loop. This architecture enables dynamic, speech-driven interaction while maintaining modularity across front-end, back-end, and model layers, thereby enhancing scalability and maintainability in educational and institutional contexts.

3.4. Evaluation and Feedback

To validate the effectiveness and reliability of the developed VA system, two primary evaluation strategies are employed, combining both technical performance testing and user-centered assessment.

First, black-box testing is conducted to evaluate the logical accuracy and relevance of the system's responses without accessing the internal workings of the model [19]. This approach simulates real-world usage by testing the system with various input scenarios, such as asking single or multiple questions simultaneously, submitting incomplete queries, and introducing unexpected or ambiguous inputs. The system's responses are then compared with the expected outputs based on the initial design documentation to verify the platform's performance and reliability. This qualitative evaluation also helps identify bugs or errors that may have occurred during the integration process.

Second, a similarity-based evaluation is performed using semantic vector comparison techniques. Each user query and its corresponding response are converted into vector embeddings using natural language processing tools compatible with the underlying Llama model. The cosine similarity score is calculated to measure the semantic alignment between the intended meaning of

the query and the generated response [20]. A minimum similarity threshold of 90% is established as the benchmark for acceptable performance. Responses falling below this threshold are flagged for manual review and further refinement, ensuring that the VA delivers contextually appropriate and accurate information.

In addition to these technical assessments, user acceptance test is actively collected during the pilot deployment phase through structured questionnaires and informal interviews. The students, who were the main target of this work, were asked to evaluate their experience based on clarity, accuracy, response speed, and overall satisfaction. Suggestions and insights gathered from users are considered for future system updates, helping to fine-tune both the VA's behavior and user interface. This ongoing feedback loop contributes to the development of a more personalized, accessible, and user-friendly virtual assistant.

4. RESULTS AND DISCUSSION

This section presents the results of the proposed work, which were obtained through four different testing procedures. The evaluation began with black-box testing, which was conducted to examine the overall functionality of the system. This was followed by a similarity evaluation, aimed at assessing the system's capability to generate accurate responses. Subsequently, inference time testing was performed to measure the time required by the system to generate and deliver responses. Finally, a User Acceptance Test (UAT) was carried out to collect user feedback and assess the overall usability and performance of the proposed GenAI-based virtual assistant.

4.1. Blackbox Testing

To ensure that the proposed GenAI-based VA system works properly, not only in the training process, but also in actual implementation. This study covers comprehensive blackbox testing which divided into three categories, such as input voice, output voice and UI/UX testing to guarantee the system delivery. The results of the blackbox testing method detailed in Table 2.

Table 2. Summary of Blackbox Testing for Voice Interaction, UI Elements, and Audio Response

Test Category	Test Scenario	Test Steps	Expected Results	Actual Results
Voice Input	Basic Voice Recognition	Ask a simple question and wait for recognition text showed	System should able to translate the question correctly	Success
	Voice Input with Noise	In noisy environment, ask a simple question and wait for recognition text showed	System should able to translate the question correctly	Success
	Long Voice Input	Ask a long question and wait for recognition text showed	System should able to translate the question correctly	Success
UI/UX	Button Interaction	Press the <i>Ajukan Pertanyaan</i> button and observe the UI changes	System should changes the avatar gestures and start listening	Success
	Video Transitions	Press <i>Ajukan Pertanyaan</i> button and observe any video transitions during listening, talking and greeting.	System should provides different animation	Success
Voice Output	Basic Audio Generation	Ask simple question and check audio playback	System should generated audio file	Success
	Format Consistency	Ask simple question and check generated audio format	System should generate .mp3 file	Success
	Audio Quality	Ask simple question and listen to the response in terms of quality.	System should generate file without any noise	Success

Initially, we investigate the voice input provided by user which is captured by the proposed system. In total, three test scenarios were conducted: basic voice recognition, voice input with noise, and long voice input. In each case, users were instructed to ask questions, and the system was expected to correctly recognize and transcribe the speech input into text. All tests were successfully completed, demonstrating the effectiveness of the speech recognition module even under noisy and extended input conditions.

Additionally, the UI/UX category focused on the responsiveness and visual feedback of the user interface. Button interaction was tested by pressing the Ajukan Pertanyaan button and observing whether the system altered the avatar's gestures and initiated the listening process. In the video transitions scenario, animations during different assistant states (listening, speaking, greeting) were evaluated. Both scenarios resulted in successful outcomes, indicating that the UI was responsive and visually consistent. Furthermore, in terms of Voice Output provided by the proposed system, three aspects were evaluated: audio generation, format consistency, and audio quality. The system was tested for its ability to generate and play back audio responses, produce output in the .mp3 format, and maintain clarity without background noise. Each of these tests was successfully passed, confirming the reliability of the text-to-speech component. The results indicate the system's ability to accurately process spoken queries, provide appropriate user feedback through UI interactions, and generate high-quality audio responses.

4.2. Response Similarity Testing

To evaluate the relevance and factual accuracy of the system's responses, a similarity test was conducted between the generated outputs and a reference corpus consisting of ground-truth answers or official university documentation. The similarity testing in this study was carried out in collaboration with all departments at Primakara University. This approach was selected to ensure that the questions and answers provided by the GenAI-based virtual assistant were contextually relevant and institution-specific. It was assumed that each department possesses insight into the most frequently asked questions by students.

As shown in Figure 5, a total of 11 departments were considered in this evaluation. Accordingly, one representative question was selected from each department, resulting in 11 distinct test cases. For each question, the evaluation process was repeated across 10 trials to calculate the average similarity score and the corresponding standard deviation, thereby measuring the consistency and semantic alignment of the system's responses. It is worth noting that the similarity of the actual and generated answer was calculated using Term Frequency-Inverse Document Frequency (TF-IDF) that transform text into numerical vector form, which were then calculated using cosine similarity function. Additionally, semantic similarity was evaluated using the SentenceTransformer library with the pre-trained model all-MiniLM-L6-v2, by comparing the generated outputs against the actual information from the database.

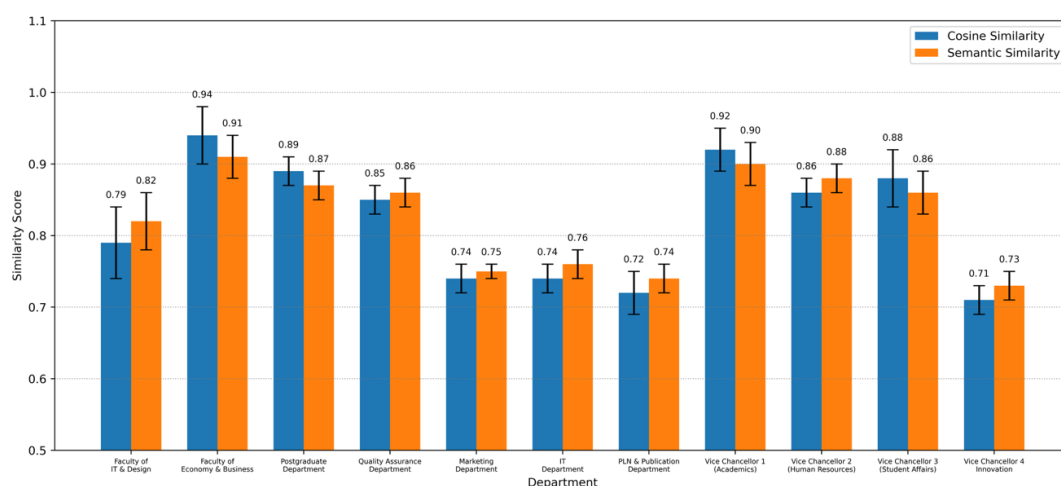


Figure 5. Cosine and Semantic Similarity test results across 11 departments.

Figure 5 illustrates the similarity scores obtained across various university departments using both Cosine Similarity and Semantic Similarity metrics, evaluated on ten different responses generated by the proposed GenAI-based VA system, along with their corresponding standard deviation values. The overall average cosine similarity across all departments is 83.7%, with the highest similarity achieved in the Faculty of Economy and Business (94%) and the lowest in the Vice Chancellor 4 department, which oversees research and innovation (71%). In contrast,

semantic similarity averages slightly lower at 82.9%, but outperforms cosine similarity in several departments such as Faculty of IT & Design and the IT Department.

To further analyze the model's consistency, standard deviation values were calculated for each department. The average standard deviation across all departments is approximately 2.6%, indicating the model's stable performance and generalization across different contexts. The lowest deviation is observed in the Vice Chancellor 3 department with a value of 1.5%, suggesting highly consistent results. Meanwhile, the highest standard deviation appears in the Faculty of IT & Design with a deviation of 5%, reflecting more variability in the system's responses for that particular department. Overall, the similarity test indicates good model performance in generating responses to user input, as the average similarity score exceeds 80%.

4.3. System Delay Testing

To better demonstrate the performance of the proposed GenAI-based VA system for university related applications, an evaluation of system delay, representing the total duration from the user's request to the playback of the response audio, was conducted. It is worth noting that the system delay covers all process proposed in this work. The test was performed by inserting a time identifier immediately after the user input was given, followed by calculating the time elapsed until the response audio was played. To ensure more reliable results, the evaluation was carried out 50 times using a randomly selected set of questions, which were chosen based on a similarity test.

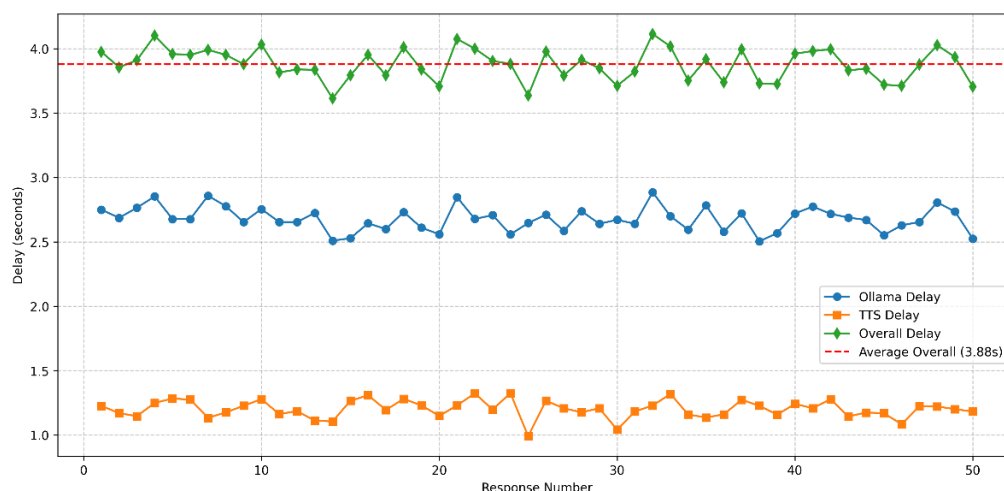


Figure 6. The overall response delay of the proposed GenAI-based virtual assistant to a user's input question.

Figure 6 presents a comparison of delay measurements over 50 response instances to evaluate the performance consistency of the proposed GenAI-based VA system. The overall delay, representing the total time from user input to the playback of the generated audio response, is plotted alongside the delay components: the time taken by the Llama language model and the TTS engine. As shown, the overall delay consistently remains close to 4 seconds, with the average overall delay marked at 3.88 seconds by the red dashed line. The Llama model contributes the largest portion of delays, averaging between 2.5 to 2.9 seconds across responses, while the TTS delay is relatively lower, fluctuating around 1.1 to 1.3 seconds. These results indicate that the system performs with stable latency, and the response generation pipeline is efficient enough for real-time applications in university settings.

4.3. User Acceptance Test

To ensure the system performs effectively during actual implementation, a user acceptance test was conducted and user feedback was collected. Figure 7 illustrates the implementation results and the testing process.



Figure 7. Implementation results and testing process of the proposed GenAI-based VA system.

A total of 25 users were randomly selected from students at Primakara University. Each student was allowed to submit up to five questions within a three-minute testing period. After the interaction, user feedback was collected through a questionnaire containing six statements rated on a scale from 1 (indicating strong disagreement) to 5 (indicating strong agreement). The evaluation indicators included aspects such as system interface, response quality, accessibility, and user recommendation. To enhance the generalization of the output from the proposed GenAI-based VA, we incorporated users with various academic backgrounds in the feedback tests, including differences in semester level and major. However, since the primary target of this system is students, the evaluation focuses exclusively on student perspectives from diverse backgrounds. To provide a clearer understanding of the proposed system's performance, evaluation results from a total of 25 students were collected. Table 3 presents 10 sample responses that were randomly selected from the full set.

Table 3. User Acceptance Testing (UAT) Indicators and Results

No	System Interface	Response Quality	Accessibility	Audio Clarity	Overall Satisfaction	Willingness to Recommend	Comments
1	4	3	4	5	4	5	Gives wrong information about lecturer academic degree
2	3	4	4	3	4	4	Need to push button; Audio is not clear
3	5	5	5	4	5	5	-
4	4	3	3	4	3	4	Provides wrong answer on Newest Curriculum
5	3	4	4	5	4	3	Remove the requirement to push button
6	5	5	5	5	5	5	-
7	4	4	4	3	4	4	Use Indonesian services TTS
8	3	3	4	4	3	3	Provides wrong answer on tuition fee
9	5	4	5	5	5	5	-
10	4	3	4	4	4	4	Provides wrong answer on university agenda

The results show that most users provided relatively positive and constructive feedback, while some also raised concerns about the performance of the proposed GenAI-based virtual assistant in this work. To summarize the overall feedback from the primary target users, namely students, we present the average scores across all indicators. Table 4 provides a summary of the overall feedback obtained from the 25 participants in this study.

Table 4. User Acceptance Testing (UAT) Indicators and Results

Indicator	Description	Scale	Average Score (Interpretation)	T-statistic value
System Interface	Clarity, layout, and ease of navigation	1 (SD) – 5 (SA)	4.12 (Very Satisfactory)	12.74
Response Quality	Relevance and accuracy of system-generated answers	1 (SD) – 5 (SA)	4.36 (Very Satisfactory)	10.66
Accessibility	Ease of use for various users and accessibility	1 (SD) – 5 (SA)	4.20 (Very Satisfactory)	9.30

Audio Clarity	support Understandability and naturalness of the voice response	1 (SD) – 5 (SA)	4.60 (Excellent)	16.00
Overall Satisfaction	General impression and user experience	1 (SD) – 5 (SA)	4.64 (Excellent)	16.74
Willingness to Recommend	Likelihood of recommending the system to others	1 (SD) – 5 (SA)	4.68 (Excellent)	17.64

Based on the feedback collected from 25 participants, the System Interface received an average score of 4.1, indicating that the layout and navigation were perceived as very satisfactory. The Response Quality, reflecting the relevance and accuracy of the system's answers, was rated at 4.4, also falling within the very satisfactory range. Accessibility was rated slightly lower, with a score of 4.2, categorized as very satisfactory, suggesting that improvements may be needed to accommodate a wider range of users as the current interaction requires user to touch the screen to activate the listening function of the VA.

Additionally, in terms of Audio Clarity of the system, which assessed the understandability and naturalness of voice responses, received a high rating of 4.6. Similarly, the Overall Satisfaction indicator achieved a score of 4.6. The highest score of 4.7 was given to the Willingness to Recommend indicator, showing excellent user approval and confidence in the system's potential. These results demonstrate that the system was well-received by users and is considered effective and usable for its intended purpose.

Additionally, a one-sample t-test was conducted using *ttest_Isamp* library from *scipy* to assess whether user ratings across six UAT indicators were significantly greater than the neutral midpoint (3.0). The results show that all indicators yielded statistically significant results, with particularly high t-values for Audio Clarity ($t = 16.00$), Overall Satisfaction ($t = 16.74$), and Willingness to Recommend ($t = 17.64$). These findings indicate that users had an overall highly positive experience with the system and are likely to recommend it to others.

4.3. Analysis of Model Response

To further assess the performance and limitations of the proposed GenAI-based virtual assistant, a qualitative analysis was conducted based on direct user feedback. User comments, along with possible causes and mitigation strategies, were incorporated into the results as detailed in Table 5. The entries considered in this subsection were collected from comments in Table 3.

Table 5. Analysis of user comments, possible causes, and mitigation strategies

No	User Comment	Possible Cause	Mitigation Strategy
1	Gives wrong information about lecturer academic degree	Verified lecturer profile documents were not included in the RAG index; outdated or incomplete data may have been retrieved	Always check lecturer profiles is up to date to the document index and validated against the internal HR database
2	Need to push button; Audio is not clear	A manual input mechanism was used; audio was generated using a TTS engine that had not been optimized	A voice-activated trigger should be implemented, and audio quality should be enhanced with an improved TTS engine
3	Use Indonesian services TTS	An English-centric TTS engine was used, resulting in unclear or unnatural speech in Bahasa Indonesia	An Indonesian-optimized TTS engine e.g., ProsaAI should be integrated
4	Provides wrong answer on Newest Curriculum	The most recent curriculum documents were either not indexed or not retrieved correctly	Updated curriculum documents should be indexed and consistently versioned to improve retrieval accuracy
5	Remove the requirement to push button	A manual interaction design was still relied upon	A fully voice-based interaction system should be implemented, such as one with continuous listening or wake-word support
6	Provides wrong answer on tuition fee	Tuition-related documents were either outdated or excluded, leading to inaccurate retrieval	Finance and tuition documents should be added to the index
7	Provides wrong answer on university agenda	Time-sensitive agenda information was not available in the indexed documents; no real-time sync was established	A university agenda API should be integrated, or regular updates of time-based documents should be ensured

The results presented in Table 5 highlights several categories of errors observed in real-world usage scenarios, particularly those related to information accuracy and interaction experience. Notably, users reported factual inaccuracies such as incorrect academic degrees for lecturers, outdated or incorrect tuition fee details, and misleading responses regarding the most

recent curriculum and university agenda. These errors are attributed primarily to limitations in the document retrieval scope and update frequency within the RAG system, as the vector store was restricted to a fixed set of curated documents without real-time internet access or API integration.

Furthermore, interaction-related feedback revealed issues such as the need to manually trigger voice responses and lack of clarity in audio output. These shortcomings are not directly tied to the retrieval component but rather to the TTS engine and voice interaction design. Specifically, the absence of a fully voice-activated input system and reliance on English-centric TTS models contributed to reduced usability for Indonesian users. Mitigation strategies have been identified, including the integration of up-to-date internal university data, implementation of a voice-activated interface, and adoption of TTS engines optimized for Bahasa Indonesia.

Despite a mix of positive and negative feedback from users, the T-statistic value obtained from the UAT indicates strong performance of the proposed GenAI-based virtual assistant. Also, this multi-layered error analysis validates the system's strengths in delivering grounded responses when supported by relevant data, while also revealing critical areas requiring enhancement for reliable, context-aware academic assistance.

5. CONCLUSION

This study presented the development and evaluation of a GenAI-based VA system designed to support university-related interactions. The system integrates speech recognition, natural language processing, and text-to-speech components to deliver a seamless voice-based user experience. A series of four testing phases were conducted to assess the system's functionality, relevance of responses, processing performance, and user satisfaction. Black-box testing confirmed that the system successfully handled various voice input scenarios, delivered accurate voice output, and maintained responsive UI/UX elements. Through similarity evaluation using TF-IDF and cosine similarity across multiple academic departments, the system demonstrated an average response similarity of 81.86%, with a relatively low standard deviation, indicating its consistency and contextual accuracy. Performance evaluation revealed that the system maintained an average response time of 3.88 seconds, making it suitable for real-time interactions. Moreover, user acceptance testing involving 25 randomly selected students yielded highly positive results and validated using T-statistic value. Most usability indicators, such as interface clarity, response quality, and audio clarity were rated as very satisfactory, while the willingness to recommend the system achieved an excellent rating.

Overall, the proposed GenAI-based VA system has been validated through both objective and user-centered evaluations, confirming its effectiveness and reliability in supporting information services within the university environment. Future work may explore enhancements in accessibility features and broader integration with institutional databases to further enrich the assistant's capabilities while also consider another RAG technique to provides better response for the VA.

ACKNOWLEDGEMENTS

This research was supported by Directorate of Research and Community Services, Primakara University, Bali, Indonesia.

REFERENCES

- [1] M. Belkina *et al.*, "Implementing generative AI (GenAI) in higher education: A systematic review of case studies," *Comput. Educ. Artif. Intell.*, vol. 8, p. 100407, 2025, doi: <https://doi.org/10.1016/j.caeai.2025.100407>.
- [2] M. Matsiola, G. Lappas, and A. Yannacopoulou, "Generative AI in Education: Assessing Usability, Ethical Implications, and Communication Effectiveness," 2024. doi: 10.3390/soc14120267.
- [3] H.-P. (Hank) Lee *et al.*, "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, in CHI '25. New York, NY, USA: Association for Computing Machinery, 2025. doi: 10.1145/3706598.3713778.

- [4] U. Mittal, S. Sai, V. Chamola, and D. Sangwan, "A Comprehensive Review on Generative AI for Education," *IEEE Access*, vol. 12, pp. 142733–142759, 2024, doi: 10.1109/ACCESS.2024.3468368.
- [5] R. Abdelghani *et al.*, "GPT-3-driven pedagogical agents to train children's curious question-asking skills," *Int. J. Artif. Intell. Educ.*, vol. 34, no. 2, pp. 483–518, 2024.
- [6] M. Rawal, "National Education Policy 2019: For The Transferable Job Parents Which Affect Their Children' Education," vol. 2, no. 3, pp. 1–7, 2020.
- [7] H. Yu and Y. Guo, "Generative artificial intelligence empowers educational reform: current status, issues, and prospects," *Front. Educ.*, vol. Volume 8-2023, 2023, [Online]. Available: <https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2023.1183162>
- [8] T. Adiguzel, H. Kaya, and F. Cansu, "Revolutionizing education with AI: Exploring the transformative potential of ChatGPT," *Contemp. Educ. Technol.*, vol. 15, p. ep429, Apr. 2023, doi: 10.30935/cedtech/13152.
- [9] L. Kohnke, B. L. Moorhouse, and D. Zou, "Exploring generative artificial intelligence preparedness among university language instructors: A case study," *Comput. Educ. Artif. Intell.*, vol. 5, p. 100156, 2023, doi: <https://doi.org/10.1016/j.caeai.2023.100156>.
- [10] A. R. Doshi and O. P. Hauser, "Generative artificial intelligence enhances creativity but reduces the diversity of novel content," *arXiv Prepr. arXiv2312.00506*, 2023.
- [11] I. Jurenka *et al.*, "Towards responsible development of generative AI for education: An evaluation-driven approach," *arXiv Prepr. arXiv2407.12687*, 2024.
- [12] D. Baidoo-Anu and L. Ansah, "Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning," *J. AI*, vol. 7, Mar. 2023, doi: 10.61969/jai.1337500.
- [13] Y. Ge *et al.*, "Making llama see and draw with seed tokenizer," *arXiv Prepr. arXiv2310.01218*, 2023.
- [14] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv Prepr. arXiv2302.13971*, 2023.
- [15] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020.
- [16] M. Hanisch, C. M. Goldsby, N. E. Fabian, and J. Oehmichen, "Digital governance: A conceptual framework and research agenda," *J. Bus. Res.*, vol. 162, p. 113777, 2023, doi: <https://doi.org/10.1016/j.jbusres.2023.113777>.
- [17] A. F. Anwar, R. Kusumawati, M. A. Yaqin, I. B. Santoso, and A. A. Zuhri, "Enhancing Student Collaboration in Academic Projects Through a Content-Based Filtering Recommender System," 2025. doi: 10.59395/ijadis.v6i2.1329.
- [18] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 325–351, 2023.
- [19] M. T. Abdillah, I. Kurniastuti, F. A. Susanto, and F. Yudianto, "Implementasi Black Box Testing dan Usability Testing pada Website Sekolah MI Miftahul Ulum Warugunung Surabaya," *J. Comput. Sci. Vis. Commun. Des.*, vol. 8, no. 1, pp. 234–242, 2023, doi: 10.55732/jikdiskomvis.v8i1.897.
- [20] Ylber Januzaj and Artan Luma, "Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words," *Int. J. Emerg. Technol. Learn.*, vol. 17, no. 12 SE-Short Papers, pp. 258–268, Jun. 2022, doi: 10.3991/ijet.v17i12.30375.