IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 237~246
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp237-246

Real-time intelligent virtual assistant based on retrieval augmented generation

I Ketut Resika Arthana[1], Ni Putu Novita Puspa Dewi[1], Gede Arna Jude Saskara[2], I Made Ardwi Pradnyana[2], Luh Indrayani[3]
[1]Program Study of Computer Science, Faculty of Engineering and Vocational, Universitas Pendidikan Ganesha, Singaraja, Indonesia
[2]Program Study of Information Systems, Faculty of Engineering and Vocational, Universitas Pendidikan Ganesha, Singaraja, Indonesia
[3]Program Study of English Language Education, Faculty of Language and Arts, Universitas Pendidikan Ganesha, Singaraja, Indonesia

Article Info

Article history:

Received Oct 28, 2024
Revised Dec 24, 2025
Accepted Jan 10, 2026

Keywords:

Design science research
Large language models
RIVA
Retrieval-augmented generation
RAGAS

ABSTRACT

Improving user experience in accessing information on organizational websites remains a challenge. Users often face complex navigation and multi-step searches that slow information retrieval. This study introduces the real-time intelligent virtual assistant (RIVA), which integrates large language models (LLMs) with the retrieval-augmented generation (RAG) framework to support real-time interaction with website content. The system was implemented on the Universitas Pendidikan Ganesha (Undiksha) website using a WordPress content management system (CMS) and developed following the design science research (DSR) approach, which includes six stages: problem identification, solution objectives, design and development, demonstration, evaluation, and communication. The retrieval-augmented generation assessment (RAGAS) evaluation indicated that combined model of text-embedding-ada-002 and semantic chunking yielded the best results, with context precision=0.83, context recall=0.90, response relevancy=0.91, faithfulness=0.83, and answer correctness=0.85. User experience questionnaire (UEQ) testing performed well, particularly in the novelty and stimulation dimensions. These results demonstrate that RIVA can provide users with access to relevant and engaging information. As a result, future research will focus on improving retrieval and developing adaptive semantic chunking for structured and complex data.

Corresponding Author:

I Ketut Resika Arthana
Program Study of Computer Science, Faculty of Engineering and Vocational, Universitas Pendidikan Ganesha
Udayana Street, Singaraja, Buleleng, Bali, Indonesia
Email: resika@undiksha.ac.id

1. INTRODUCTION

Organizational websites are a primary channel for information access; however, users often face difficulties due to manual searches, complex navigation, and unintuitive interfaces, which reduce user satisfaction and service effectiveness. To address the growing demand for fast and accurate information access, artificial intelligence (AI), particularly AI based virtual assistants, has emerged as a promising solution to simplify information retrieval and enhance user experience. In this context, large language models (LLM) as generative AI has changed the way people interact with information systems. LLMs research and practical application continuously advance, generating various technological innovations [1], [2]. As one of the popular LLM-based applications, ChatGPT has attracted global attention due to its ability to manage language-related tasks in conversation [3] and enhance productivity in research and academia [4], [5]. Large text datasets are used to train LLMs, enabling them to generate coherent and contextually relevant responses [6]. LLMs have

been widely applied in various fields and domains, which include education [7], healthcare [8]–[10], software development [11], [12], and industry [13]. LLMs existence also creates opportunities for organizations to implement them as virtual assistants, improving information services by receiving and following up on customer feedback to improve service quality and user experience [14], [15].

LLMs still face limitations when integrated into organizational websites although they have made rapid progress. Although trained using large datasets, their knowledge stays static and often fails to capture the specific context of an organization, necessitating extensive retraining [16]. Additionally, LLMs face challenges like hallucinations [17], falsification [18], and irrational responses in certain contexts [19]. The retrieval-augmented generation (RAG) approach has emerged as a promising solution to addressing these issues. RAG allows LLMs to access and integrate real-time information from external sources, resulting in contextually relevant and up-to-date responses [6]. Combining LLMs with RAG assists in overcoming the limitations of static knowledge and expands their role as intelligent assistants in organizational environments. RAG research is constantly evolving, highlighting its evolution through three paradigms: naive RAG, advanced RAG, and modular RAG [20]. Naive RAG implements a simple retrieve-and-read process, while advanced RAG introduces query rewriting and context optimization. Lastly, modular RAG connects retrieval, memory, and generation for domain-specific reasoning. This evolution indicates a shift in RAG from a static data retrieval model to an adaptive framework that supports complex decision-making. Its effectiveness has been validated in various fields, including autonomous vehicle systems that integrate RAG and LLMs for accurate real-time information delivery [21]. Retrieval-augmented generation assessment (RAGAS) framework generally used to evaluate performance [22], measuring the quality of information retrieval and generation through metrics such as context precision, context recall, faithfulness, and answer relevance.

Table 1 displays a comparison between existing RAG-based systems and frameworks, by highlighting the real-time intelligent virtual assistant (RIVA) unique contributions in real-time synchronization and empirical evaluation. Most of the previous studies focused on domain-specific or general frameworks, like LangChain and ChatGPT Plugins, which have not yet addressed real-time integration with organizational websites, particularly in educational environments. This study introduces RIVA, which is built based on LLM and RAG integration to address these gaps. RIVA integrates WordPress content management system (CMS) synchronization, user experience questionnaire (UEQ)-based user experience evaluation, and local language adaptation for Indonesian users. As an example, a case study conducted at Universitas Pendidikan Ganesha (Undiksha) shows that the combination of LLM and RAG can facilitate accurate, up-to-date, and user-friendly access to organizational information.

Table 1. Benchmarking RAG-based systems and frameworks

System/paper	Focus/domain	Contribution
RAG and LLM integration [23]	General RAG-LLM concepts	Provides an overview of integrating RAG with LLMs across domains, highlighting core architectural principles
Efficient biomedical question-answering (QA) via RAG [24]	Biomedical	Proposes a reproducible and efficient RAG framework for biomedical question answering
Open-source LLM+RAG [25]	Open-source integration	Demonstrates how open-source LLMs can be integrated with RAG pipelines for flexible applications
LangChain [26]	General-purpose pipelines	A widely used open-source framework providing modular tools for RAG, memory, and agent-based applications with integration to various databases
ChatGPT plugins [27]	General-purpose assistants	A plugin system connecting ChatGPT with external APIs, allowing real-time data retrieval within the OpenAI ecosystem
RIVA (this work)	Educational website CMS	Introduces a domain-specific RAG assistant with real-time integration, evaluation using UEQ, and local language adaptation for Indonesian higher education

2. METHOD

This study utilized the design science research (DSR) approach, focusing on creating innovative artifacts such as systems, models, or methods that can be empirically verified. The DSR framework comprised six main stages: problem identification, solution objectives, design and development, demonstration, evaluation, and communication.

2.1. Problem identification

The primary issue with organizational websites is their traditional search systems and complex navigation, which often makes it difficult for users to find relevant information. As a result, LLM provides a solution through natural language-based interaction without the need to manually navigate the interface. However, integrating LLM with organizational websites poses challenges, especially in maintaining synchronization to keep information up-to-date and relevant to the organizational context. Therefore, this study

utilized RAG to integrate the LLM with the organizational database. However, RAG search results were not always relevant to the context, necessitating a comparison of various chunking and embedding techniques to determine the optimal configuration for generating accurate and contextual responses.

2.2. Objective of a solution

This study aimed to enhance the ability of LLM-based systems to provide relevant and contextually relevant answers on organizational websites. The RAG approach was applied to connect the LLM with organizational databases, ensuring that the generated information remained up to date. The study's primary focus was to determine the most effective combination of chunking and embedding for producing an accurate information retrieval process, serving as a basis for developing a responsive and contextually intelligent search system.

2.3. Design and development

The study employed a system design comprising two interrelated architectures. First, the implementation architecture (Figure 1) represented the RIVA system deployment on the organizational website as an LLM-based information retrieval solution, with the overall process integration between the organizational website and RIVA (Figure 2). Second, the experimental architecture (Figure 3) was developed to evaluate various combinations of chunking and embedding techniques, identifying the most effective configuration prior to implementation in the main system. These two architectures were iterative and complementary, where the experimental architecture's results were used to refine the implementation architecture, ensuring that the final system provided more accurate, relevant, and contextually appropriate responses.

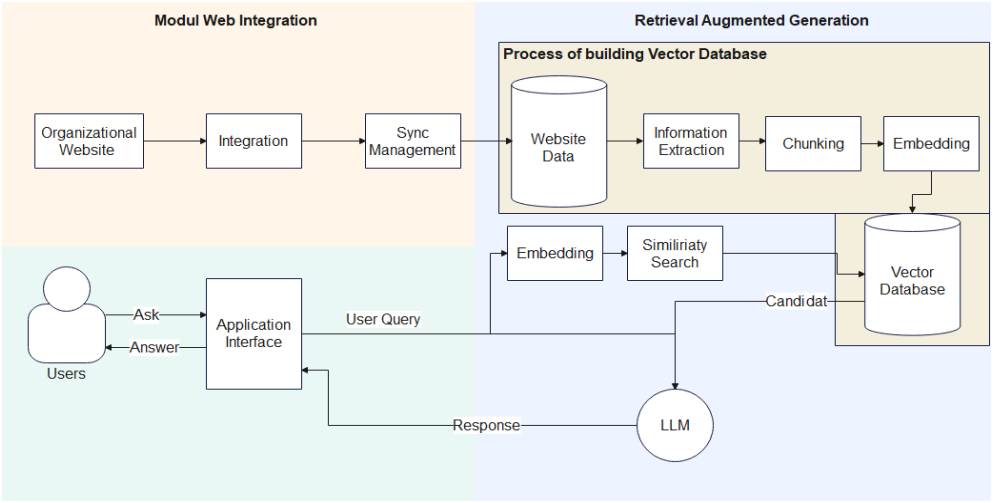

Figure 1. Architecture of RIVA

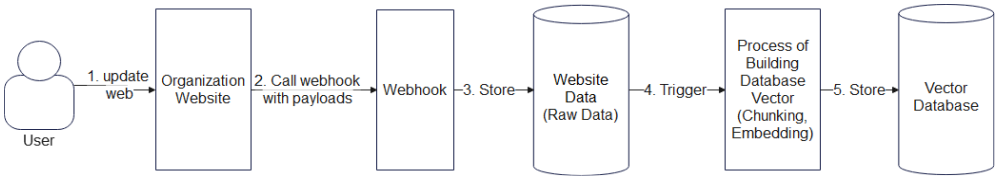

Figure 2. Process integration organization web to RIVA

2.3.1. Design architecture implementation (RIVA)

As shown in Figure 1, the RIVA architecture integrated a real-time web synchronization module and a RAG pipeline that processes user queries through the retrieval and generation stages. This figure presents a high-level system architecture that shows how user queries are processed through a combination of web integration modules and RAG components, allowing for real-time information retrieval from updated website sources via LLM virtual assistants.

i) Module web integration: various website and RAG integration mechanisms were examined to determine an efficient and secure solution. Web scraping offers implementation simplicity and data flexibility but involves legal concerns and vulnerability to changes in website structure, while built in APIs provide greater stability with limited access to content. Accordingly, this study employed a custom API to maintain full control over data flow and ensure security through token-based authentication. As shown in Figure 2, content updates activate webhooks that synchronize data to RIVA for storage, segmentation, and vector indexing, supporting accurate and reliable real time information retrieval.

ii) Module retrieval augmented generation: the RAG module served as the core component of RIVA, designed to enrich user queries and generate accurate, contextual, and reliable responses by integrating factual and real time knowledge from external sources, thereby enhancing LLM capabilities while mitigating the limitations of statically trained models and reducing hallucinations. In this study, the RAG pipeline began with document chunking and embedding, where long or unstructured texts were segmented into coherent units based on natural language boundaries and transformed into numerical representations using language models, enabling semantic similarity recognition beyond surface level word matching. These embeddings were stored in a vector database to support fast and accurate similarity search, with Facebook AI similarity search (FAISS) selected for its scalability, high performance, and suitability for real time retrieval in a university website context. Information retrieval employed a hybrid strategy that combined vector based semantic search and BM25 lexical retrieval to balance contextual relevance and keyword precision, followed by refinement using the BAAI BGE Reranker large model to select the most relevant passages. Finally, the refined context was processed by the GPT 4o model during the generation stage to synthesize fluent, factually consistent, and user aligned responses, ensuring reliable information delivery.

2.3.2. Design architecture experiment

This stage mainly focused on testing combinations of various chunking strategies and embedding models to discover the most effective settings for generating relevant context and accurate responses. The development process comprised preparing data from the organizational website, implementing RAG pipelines for retrieval and generation, and evaluating results using RAGAS metrics. As a result, this stage will form the basis for final design decisions to be applied to the RIVA system implementation architecture.

The experimental architecture flow used to determine the optimal chunking and embedding configurations in the RIVA system is shown in Figure 3. Based on the flow, the process began with data collection from the research's main source, namely the organization's official website. Based on this data, a dataset containing 200 question-answer pairs in Indonesian was developed using GPT-4o. Furthermore, each pair was verified by two experts to ensure its relevance, accuracy, and consistency with the website corpus. An example dataset is presented in Table 2.

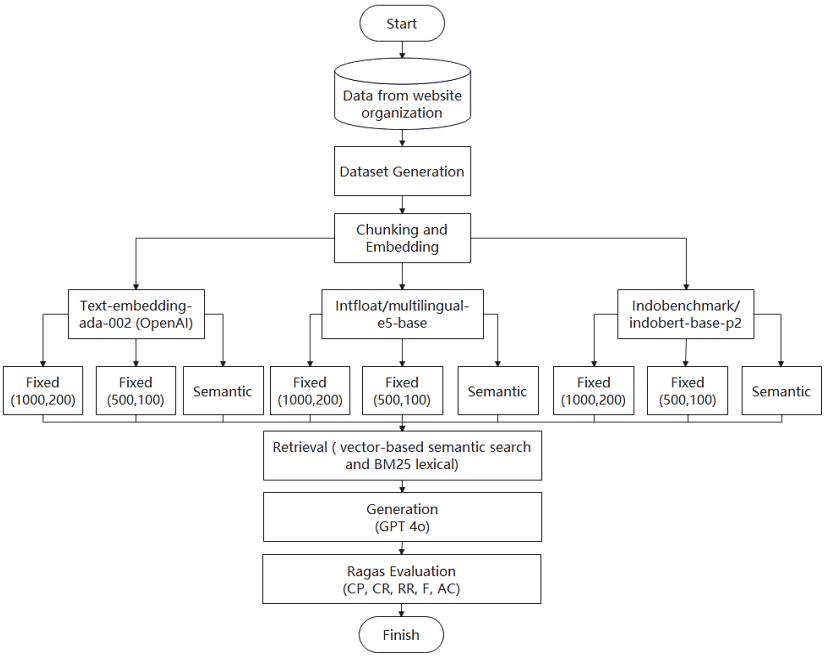

Figure 3. Design architecture experiment

Table 2. Example datasets

Input	Expected output
When to fill in tuition fee data for SNBP at Undiksha?	Tuition data submission for SNBP at Undiksha is open from March 28 to April 5, 2024.
Can students who pass SNBP follow the SNBT program?	Students who pass the SNBP are not permitted to participate in the SNBT or the independent program selection (SMBJM).

The next stage involves the chunking and embedding process. The text is divided using three strategies: fixed chunks of 1,000 characters with 200-character overlap, fixed chunks of 500 characters with 100-character overlap, and semantic chunking based on natural language boundaries. Each strategy is tested with three embedding models' text-embedding-ada-002 (OpenAI), intfloat/multilingual-e5-base, and indobenchmark/indobert-base-p2 to generate semantic vector representations. These embeddings are then used in a hybrid retrieval process combining vector-based semantic search and BM25 lexical search to obtain the most relevant context. Furthermore, this study utilized GPT-4o to process the retrieved context and generate responses, which were evaluated using the RAGAS framework with five metrics: context accuracy, context recall, response relevance, fidelity, and answer accuracy. The process enabled the comparison of different chunking and embedding configurations to identify the most effective settings for RIVA implementation.

2.4. Demonstration

The next step in this study was a demonstration stage. This stage was performed to prove that the RIVA system design was sufficient to operate as intended for the research objectives. The optimal configuration of the chunking and embedding experiments at this stage was implemented into the RIVA system. Furthermore, real-world usage would test the system, where users search for information through natural language conversations. As a result, each request was processed through RAG, which involved vector-based retrieval and answer generation using GPT-4o.

2.5. Evaluation

The evaluation stage comprised two components: system performance evaluation using the RAGAS framework and user experience evaluation using the user UEQ. RAGAS was used to assess retrieval and generation quality by measuring the relevance and factual consistency of retrieved contexts and generated responses. UEQ evaluated users' perceptions of usability, efficiency, and overall satisfaction, providing insights into the system's practical applicability.

2.5.1. Performance evaluation using RAGAS

The RAGAS evaluation stage comprised dataset preparation, context retrieval, answer generation, and metric computation to assess retrieval and generation quality. Five metrics were employed: context precision, context recall, answer relevancy, faithfulness, and answer correctness. Context precision measured the proportion of relevant chunks within the retrieved context using precision@k, indicating the accuracy of the retrieval process. Context recall evaluated the system's ability to retrieve all relevant reference contexts, ensuring completeness of the retrieved information. Answer relevancy assessed the semantic relevance of generated responses to user queries by computing the average cosine similarity between embeddings of the original question and generated questions derived from the response. Faithfulness measured factual consistency by calculating the proportion of answer claims that could be inferred from the retrieved context. Answer correctness evaluated overall response quality using a weighted combination of factual correctness and semantic similarity between generated answers and ground truth, with default weights of 0.75 for factual consistency and 0.25 for semantic similarity.

2.5.2. User experience evaluation

User experience was evaluated using the standard UEQ, which consists of 26 items designed to assess both pragmatic and hedonic qualities of system interaction. The evaluation covered six dimensions: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty, representing overall impression, ease of use, task effectiveness, system reliability, user engagement, and perceived innovation. This instrument provided a structured and quantitative assessment of users' perceptions of RIVA during system use.

2.6. Communication

This stage documented and disseminated the study results through this publication, highlighting the RIVA's conceptual and practical contributions. The study presented an innovative integration between RAG and university websites. It also provided an empirical evaluation framework that could be replicated by other educational institutions.

3. RESULTS AND DISCUSSION

In this section, the study describes the results from implementing and evaluating the RIVA system. The discussion focuses on analyzing the experimental results, interpreting the system's performance in various configurations, and highlighting important patterns observed during testing. Thus, these results are organized to provide quantitative evidence and interpretive insights that support the study's objectives.

3.1. Results

3.1.1. Performance results

Table 3 presents the final experimental results. The performance of various chunking and embedding configurations implemented in the RIVA system was evaluated in this experiment. The evaluation utilized the RAGAS metric, consisting of context precision, context recall, response relevancy, faithfulness, and answer correctness. These metrics were used to measure the system's accuracy and consistency in retrieving relevant context and generating factually correct responses across various configurations.

Table 3. Evaluation results

Model embedding	Chunking	Context precision	Context recall	Response relevancy	Faithfulness	Answer correctness
Text-embedding-ada-002 (OpenAI)	Fixed (1,000; 200)	0.82	0.91	0.88	0.85	0.82
Text-embedding-ada-002 (OpenAI)	Fixed (500, 100)	0.80	0.83	0.82	0.85	0.77
Text-embedding-ada-002 (OpenAI)	Semantic	0.83	0.90	0.91	0.83	0.85
Intfloat/multilingual-e5-base	Fixed (1,000, 200)	0.83	0.89	0.89	0.85	0.79
Intfloat/multilingual-e5-base	Fixed (500, 100)	0.80	0.84	0.84	0.85	0.77
Intfloat/multilingual-e5-base	Semantic	0.81	0.88	0.88	0.88	0.81
Indobenchmark/indobert-base-p2	Fixed (1,000, 200)	0.76	0.74	0.79	0.84	0.75
Indobenchmark/indobert-base-p2	Fixed (500, 100)	0.80	0.72	0.74	0.80	0.73
Indobenchmark/indobert-base-p2	Semantic	0.75	0.75	0.78	0.84	0.74

As shown in Table 3, there are distinct differences in performance between embedding models and chunking strategies. As a whole, the highest and most consistent scores across all RAGAS metrics were achieved by the text-embedding-ada-002 (OpenAI) model. Either fixed chunking or semantic chunking produced stable results, with all values exceeding 0.80. Additionally, the best performance was observed in the semantic chunking configuration, with context precision=0.83, context recall=0.90, response relevancy=0.91, faithfulness=0.83, and answer correctness=0.85. These findings indicated that the model was extremely effective at capturing semantic relationships and maintaining accuracy in generating contextually relevant answers.

Subsequently, the Intfloat/multilingual-e5-base model also exhibited competitive performance, particularly in semantic cut-off settings (context precision=0.81, context recall=0.88, response relevancy=0.88, faithfulness=0.88, answer correctness=0.81). Despite being designed for multilingual use, this model performed favorably on Indonesian data due to its robust semantic representation. Conversely, the lowest results were recorded by the Indobenchmark/indobert-base-p2 model, particularly in context recall and answer correctness. It indicates that its embedding representation was less proficient at capturing deep contextual similarities compared to larger and general-purpose models, despite indobert being specifically trained on Indonesian. Summarizing, the Text-embedding-ada-002 (OpenAI) model with semantic chunking achieved the best overall performance, showcasing that meaning-driven semantic chunking, combined with a robust embedding model, can significantly improve retrieval precision and factual consistency in RAG pipelines.

3.1.2. UEQ results

The UEQ evaluated RIVA user experiences, measuring six dimensions: attractiveness, clarity, efficiency, dependability, stimulation, and novelty. Furthermore, the assessment aimed to explore users' perceptions of the system's usability, reliability, and innovation. The evaluation engaged users in direct interaction with RIVA to measure its ease of use and user satisfaction.

Table 4 demonstrates that RIVA scored highly across all dimensions, indicating a strong positive response from users. Novelty (2.548) and stimulation (2.452) recorded the highest average scores, indicating that users find the system innovative, engaging, and pleasant to use. Additionally, both attractiveness and dependability scored highly, indicating that the interface was visually appealing and consistently delivered reliable results. A slightly lower but remaining positive rating in the perspicuity and efficiency categories indicated that RIVA was easy to understand, intuitive, and effective in assisting users in completing tasks. Comparing with the UEQ benchmark dataset [28], across all dimensions, RIVA achieved scores in the

'excellent' range, indicating high user acceptance and recognition of its innovation. Accordingly, the results demonstrated that RIVA delivered a pleasant, efficient, and modern user experience, and it is suitable for real-time information access.

Table 4. UEQ results

UEQ scales (mean and variance)		
Attractiveness	2.317	0.06
Perspicuity	2.226	0.07
Efficiency	2.167	0.11
Dependability	2.286	0.17
Stimulation	2.452	0.06
Novelty	2.548	0.07

3.2. Discussion

The experiments demonstrated that combining hybrid retrieval and reranking significantly enhanced system performance by improving the relevance and accuracy of facts. It was achieved by retrieving candidate contexts using both semantic and lexical similarity and refining them through reranking. The system improved accuracy but incurred higher computational costs and longer response times. Those considerations can lead to scalability challenges in real-time environments, such as RIVA. The details of failed cases (RAGAS score <0.7) are presented in five categories: misretrieval, partial or redundant answers, context loss, semantic mismatch, and table parsing errors, as shown in Table 5. Misretrieval (30 cases) and partial or redundant answers (28 cases) occurred most frequently, followed by context loss (21 cases), semantic mismatch (12 cases), and table parsing errors (10 cases). These frequencies indicated that retrieval failures were the primary source of error in our implementation.

Table 5. Error analysis

Error type	Frequency	Example case	Error cause
Misretrieval	30	Which study programs requiring a TOEFL test result?	The retriever failed to find semantically relevant context, often due to weak embedding mismatches or lack of domain-specific keywords.
Partial/redundant answer	28	What documents must be scanned and uploaded for re-registration?	The model generated partially correct or repetitive responses due to incomplete context aggregation during augmentation.
Context loss	21	When should SNBT prospective students at Undiksha fill out the tuition payment form?	Relevant information was scattered across various chunks, leading to fragmented context retrieval.
Semantic mismatch	12	What documents must be prepared and uploaded during re-registration?	The system misunderstood the user's intent, resulting in retrieval from unrelated contexts.
Table parsing error	10	How much does it cost to study computer science?	The model cannot not interpret relationships between table headers and cell values extracted from structured data.

Misretrieval errors arose when the retrieval system failed to choose semantically relevant paragraphs. This was most probably caused by mismatches between embeddings and queries, or by a lack of domain-specific cues. Partial or repetitive answer errors tended to occur when the retrieved context was incomplete or insufficiently overlapping, leading the generation model to repeat content or omit crucial information. Context loss occurred when relevant information was scattered across multiple chunks, resulting in fragmented retrieval during segmentation. While semantic mismatches occurred when the system misunderstood user intent, causing retrieval from unrelated content. Additionally, table parsing errors occurred when the model was unable to correctly infer the associations between table headers and cell values in structured data, particularly in cases where chunking separated rows from their corresponding headers.

These observations were consistent with the results reported in evaluating retrieval quality in RAG [29], which highlighted that document retrieval quality was often a major factor limiting the general performance of RAG. Likewise, the survey evaluation of RAG: a survey [30] highlighted that retrieval relevance, accuracy, and faithfulness remain persistent challenges in hybrid systems. Upon successful context retrieval in our experiment, the generation module consistently generated coherent and factually accurate responses, affirming that context retrieval robustness is a critical factor affecting system accuracy and reliability. Moving forward, future improvements should focus on refining adaptive semantic chunking to

preserve contextual relationships, enhancing embedding alignment to better capture domain-specific semantics, and developing retrieval models that are aware of structural relationships in tabular and hierarchical documents, while maintaining a balance between accuracy and computational efficiency in the reordering process.

4. CONCLUSION

Based on the LLM–RAG framework, the development of RIVA demonstrated its capability to enhance real-time access to organizational information while improving user experience. By integrating LLMs with RAG, the system generated accurate, relevant, and contextually coherent responses directly grounded in active website data, and its deployment on the Undiksha website confirmed the feasibility of RAG-based intelligent assistants in educational environments. Evaluation using the RAGAS framework showed consistently high scores across all metrics, indicating reliable retrieval and generation performance, with the combination of semantic chunking and the text-embedding-ada-002 model achieving the best results. Error analysis revealed that data retrieval remains the most challenging component, particularly for tabular and structurally complex data, while the generation stage produced coherent and factually consistent outputs once relevant context was retrieved. User experience evaluation further indicated strong performance, especially in novelty and stimulation, suggesting positive user acceptance, and future work should focus on improving structured data retrieval, adaptive semantic processing, and scalability to support sustained real-time information access in higher education contexts.

FUNDING INFORMATION

This research was funded by the Directorate of Research, Technology, and Community Service (DRTPM), Ministry of Education and Culture, Republic of Indonesia, under contract numbers 081/E5/PG.02.00.PL/2024 and 371/UN48.16/LT/2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
I Ketut Resika Arthana	✓	✓	✓	✓		✓			✓	✓		✓	✓	✓
Ni Putu Novita Puspa Dewi	✓	✓				✓	✓	✓	✓	✓	✓	✓		✓
Gede Arna Jude Saskara	✓		✓	✓	✓		✓			✓			✓	✓
I Made Ardwi Pradnyana	✓	✓				✓	✓		✓					✓
Luh Indrayani						✓		✓	✓	✓			✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [IKRA], upon reasonable request. The data that are not publicly available due to institutional data management policies but can be shared for academic and research purposes.

REFERENCES

[1] M. Klenk, "Ethics of generative AI and manipulation: a design-oriented research agenda," *Ethics and Information Technology*, vol. 26, no. 1, Mar. 2024, doi: 10.1007/s10676-024-09745-x.

[2] S. Qi, Z. Cao, J. Rao, L. Wang, J. Xiao, and X. Wang, "What is the limitation of multimodal LLMs? A deeper look into multimodal LLMs through prompt probing," *Information Processing and Management*, vol. 60, no. 6, Nov. 2023, doi: 10.1016/j.ipm.2023.103510.

[3] T. Wu *et al.*, "A brief overview of ChatGPT: the history, status quo and potential future development," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1122–1136, May 2023, doi: 10.1109/JAS.2023.123618.

[4] E. Latif and X. Zhai, "Fine-tuning ChatGPT for automatic scoring," *Computers and Education: Artificial Intelligence*, vol. 6, Jun. 2024, doi: 10.1016/j.caeai.2024.100210.

[5] F. Sufi, "Generative pre-trained transformer (GPT) in research: a systematic review on data augmentation," *Information*, vol. 15, no. 2, Feb. 2024, doi: 10.3390/info15020099.

[6] P. Omrani, A. Hosseini, K. Hooshanfar, Z. Ebrahimian, R. Toosi, and M. Ali Akhaee, "Hybrid retrieval-augmented generation approach for LLMs query response enhancement," in *2024 10th International Conference on Web Research, ICWR 2024*, Apr. 2024, pp. 22–26, doi: 10.1109/ICWR61162.2024.10533345.

[7] J. Meyer *et al.*, "Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions," *Computers and Education: Artificial Intelligence*, vol. 6, Jun. 2024, doi: 10.1016/j.caeai.2023.100199.

[8] A. Suárez *et al.*, "Beyond the Scalpel: assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery," *Computational and Structural Biotechnology Journal*, vol. 24, pp. 46–52, 2024, doi: 10.1016/j.csbj.2023.11.058.

[9] H. Wang, C. Gao, C. Dantona, B. Hull, and J. Sun, "DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients," *NPJ Digital Medicine*, vol. 7, no. 1, Jan. 2024, doi: 10.1038/s41746-023-00989-3.

[10] T. Li *et al.*, "CancerGPT for few shot drug pair synergy prediction using large pretrained language models," *NPJ Digital Medicine*, vol. 7, no. 1, Feb. 2024, doi: 10.1038/s41746-024-01024-9.

[11] J. Sauvola, S. Tarkoma, M. Klemettinen, J. Riekki, and D. Doermann, "Future of software development with generative AI," *Automated Software Engineering*, vol. 31, no. 1, May 2024, doi: 10.1007/s10515-024-00426-z.

[12] G. Lu, X. Ju, X. Chen, W. Pei, and Z. Cai, "GRACE: empowering LLM-based software vulnerability detection with graph structure and in-context learning," *Journal of Systems and Software*, vol. 212, Jun. 2024, doi: 10.1016/j.jss.2024.112031.

[13] A. Saka *et al.*, "GPT models in construction industry: opportunities, limitations, and a use case validation," *Developments in the Built Environment*, vol. 17, Mar. 2024, doi: 10.1016/j.dibe.2023.100300.

[14] A. Alsumayt *et al.*, "Boundaries and future trends of ChatGPT based on AI and security perspectives," *HighTech and Innovation Journal*, vol. 5, no. 1, pp. 129–142, Mar. 2024, doi: 10.28991/HIJ-2024-05-01-010.

[15] B. J. McCloskey, P. M. LaCasse, and B. A. Cox, "Natural language processing analysis of online reviews for small business: extracting insight from small corpora," *Annals of Operations Research*, vol. 341, no. 1, pp. 295–312, Oct. 2024, doi: 10.1007/s10479-023-05816-2.

[16] J. Grudin and R. Jacques, "Chatbots, humbots, and the quest for artificial general intelligence," in *Conference on Human Factors in Computing Systems-Proceedings*, May 2019, pp. 1–11, doi: 10.1145/3290605.3300439.

[17] W. Chen *et al.*, "Systems engineering issues for industry applications of large language model," *Applied Soft Computing*, vol. 151, Jan. 2024, doi: 10.1016/j.asoc.2023.111165.

[18] R. Emsley, "ChatGPT: these are not hallucinations–they're fabrications and falsifications," *Schizophrenia*, vol. 9, no. 1, Aug. 2023, doi: 10.1038/s41537-023-00379-4.

[19] A. Barredo Arrieta *et al.*, "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[20] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: a survey," *arXiv:2312.10997*, Mar. 2024.

[21] X. Dai *et al.*, "VistaRAG: toward safe and trustworthy autonomous driving through retrieval-augmented generation," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 4, pp. 4579–4582, Apr. 2024, doi: 10.1109/TIV.2024.3396450.

[22] S. Es, J. James, L. E. -Anke, and S. Schockaert, "RAGAS: automated evaluation of retrieval augmented generation," *EACL 2024-18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*, 2024, pp. 150–158, doi: 10.18653/v1/2024.eacl-demo.16.

[23] B. Tural, Z. Örpek, and Z. Destan, "Retrieval-augmented generation (RAG) and LLM integration," in *2024 8th International Symposium on Innovative Approaches in Smart Technologies (ISAS),* İstanbul, Turkiye, 2024, pp. 1-5, doi: 10.1109/ISAS64331.2024.10845308.

[24] L. Stuhlmann, M. A. Saxer, and J. Furst, "Efficient and reproducible biomedical question answering using retrieval augmented generation," *2025 IEEE Swiss Conference on Data Science (SDS), Zürich, Switzerland,* Jul. 2025, pp. 154-157, doi: 10.1109/SDS66131.2025.00029.

[25] A. W. Bhiwgade and N. Nagrale, "Integrating open-source LLMs with retrieval-augmented generation for obstetrics and gynecology domain," in *IEEE International Conference on Next Generation Information System Engineering, NGISE 2025*, Mar. 2025, pp. 1–5, doi: 10.1109/NGISE64126.2025.11085188.

[26] Langchain, "LangChain overview," *LangChain Docs*, 2024. Accessed: Nov. 26, 2025. [Online]. Available: https://docs.langchain.com/oss/python/langchain/overview

[27] OpenAI, "ChatGPT plugins," *OpenAI Platform*, 2023. Accessed: Nov. 26, 2025. [Online]. Available: https://openai.com/index/chatgpt-plugins/

[28] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Construction of a benchmark for the user experience questionnaire (UEQ)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, pp. 40–44, Jun. 2017, doi: 10.9781/ijimai.2017.445.

[29] A. Salemi and H. Zamani, "Evaluating retrieval quality in retrieval-augmented generation," in *SIGIR 2024-Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2024, pp. 2395–2400, doi: 10.1145/3626772.3657957.

[30] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of retrieval-augmented generation: a survey," in *Communications in Computer and Information Science*, vol. 2301, 2025, pp. 102–120, doi: 10.1007/978-981-96-1024-2_8.

BIOGRAPHIES OF AUTHORS

I Ketut Resika Arthana earned a Master of Computer (M.Kom.) from Universitas Indonesia, Indonesia in 2012. Currently, he works as an associate professor at the Department of Computer Science of Universitas Pendidikan Ganesha (Undiksha), Bali, Indonesia. Over a decade of experience in academia, he has contributed to the field through his teaching, research, and administrative roles. His research interests are in computer science, including artificial intelligence (AI), especially in generative AI, software development, human-computer interaction (HCI), and the internet of things (IoT). He can be contacted at email: resika@undiksha.ac.id.

Ni Putu Novita Puspa Dewi is an assistant professor in Department of Computer Science at the Faculty of Engineering and Vocational Education, Universitas Pendidikan Ganesha, Indonesia. She holds a master's degree in Computer Science from Universitas Gadjah Mada and a Master of Information Management from National Taiwan University of Science and Technology. Her research interests include artificial intelligence, machine learning, data science, and time-series forecasting. Her current research focuses on large language models, sentiment analysis, educational data mining, and remote sensing–based machine learning applications. She can be contacted at email: novita.puspa.dewi@undiksha.ac.id.

Gede Arna Jude Saskara earned a Bachelor of Engineering (S.T.) in Telecommunication Engineering from Institut Teknologi Telkom, a Master of Engineering (M.T.) in Electrical Engineering with a concentration in Telematics and Telecommunication Networks from Institut Teknologi Bandung, and a Professional Engineer (Ir.) qualification from Universitas Udayana. Currently, he is a lecturer at Universitas Pendidikan Ganesha, teaching in the Information Systems Program, Department of Informatics Engineering, Faculty of Engineering and Vocational Studies. His research interests are focused on computer networks and network security. He can be contacted at email: jude.saskara@undiksha.ac.id.

I Made Ardwi Pradnyana is a lecturer at Department of Informatics, Faculty of Engineering and Vocational, Universitas Pendidikan Ganesha (Undiksha), Bali, Indonesia. He earned a Master of Engineering (M.T.) degree in informatics with a specialization in information systems from the Institut Teknologi Bandung. His research interests include IT governance, enterprise architecture, IT service management, business process management, and human computer interaction. He is a member of the Enterprise Information System Research Group (EIS-RG). He can be contacted at email: ardwi.pradnyana@undiksha.ac.id.

Luh Indrayani earned her bachelor and master's degrees in English education from Universitas Pendidikan Ganesha in 2015 and 2017 respectively. Currently, she is teaching in the Department of English Education at Universitas Pendidikan Ganesha. Her research interests include English as a foreign language (EFL), technology integration in english language teaching (ELT) and AI in education, innovative pedagogical practices and inclusive education. She teaches several disciplines in education and literature. She also actively participates in service as action for the community. He can be contacted at email: luh.indrayani@undiksha.ac.id.