Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 5 (2026) 1207 - 1214 e-ISSN: 2580-0760

Reliable Indonesian Academic Service Chatbot Using Retrieval-Augmented Generation and Adaptive Refusal

Jarot Budiasto[1], Tatik Melinda Tallulembang[2], Selfina Pare[3], Muhammad Hasbi[4]
[1,2,3,4]Department of Information Systems, Faculty of Engineering, Universitas Musamus, Merauke, Indonesia

[*]*jarot@unmus.ac.id*

Abstract

Academic service information is often stored in lengthy, formal, and distributed university documents, making accurate access difficult. Generative chatbots can improve access but may answer unsupported questions unless retrieval and refusal are controlled. This study proposes a reliability-oriented Retrieval-Augmented Generation (RAG) chatbot for Indonesian academic services integrating BGE-M3, FAISS, adaptive refusal thresholding, rule-based guardrails, Qwen2.5, and retry-on-refusal. A controlled experiment used 12 official documents and 160 Indonesian questions across eight categories. The final configuration achieved Precision@5 of 0.7250, Recall@5 of 0.9125, MRR of 0.8940, nDCG@5 of 0.8888, and mean semantic similarity of 0.8516. All 16 expected-refusal questions were refused and none of the 144 answerable questions was falsely refused, yielding observed refusal accuracy of 1.0000 and a false refusal rate of 0.0000. Retry-on-refusal reduced the false refusal rate from 0.0556 to 0.0000 and increased answerable-only similarity from 0.8150 to 0.8392. These results provide initial evidence that combining retrieval confidence, guardrails, and retry supports more reliable behavior within the controlled benchmark. Because the refusal subset was small and institution-specific, larger evaluations using naturally occurring student queries are required before inferring operational reliability.

Keywords: academic service chatbot; Indonesian NLP; refusal mechanism; retrieval-augmented generation; semantic retrieval

How to Cite: J. Budiasto, T. M. Tallulembang, S. Pare, and M. Hasbi, "Reliable Indonesian Academic Service Chatbot Using Retrieval-Augmented Generation and Adaptive Refusal", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 5, pp. 1207–1214, Sep. 2026, doi: 10.29207/resti.v10i5.8018.
Permalink/DOI: https://doi.org/10.29207/resti.v10i5.8018

Received: June 4, 2026
Accepted: September 5, 2026
Available Online: September 28, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Academic service information is a critical component of university administration because it guides students through processes such as course registration, re-registration, grade reporting, academic leave, thesis procedures, and graduation. Although such information is usually available in academic guidelines, standard operating procedures, and official institutional documents, students often have difficulty locating precise answers because these documents are distributed across different sources and written in formal administrative language. In this context, an academic service chatbot can function as a natural-language access layer that helps students obtain procedural information more efficiently.

Recent advances in transformer-based large language models [1] have made chatbot responses more flexible than conventional intent-based systems. However, using a generative model without grounding is risky in an academic service domain because the model may produce unsupported or outdated information. Retrieval-Augmented Generation (RAG) addresses this issue by combining parametric generation with non-parametric retrieval from external documents [2], [3]. Dense retrieval and vector indexing further enable semantic search over document collections, making RAG suitable for domain-specific question answering over institutional documents [4-7].

Related work shows that RAG has become an important approach for educational and institutional chatbot services. Previous studies have applied RAG to academic chatbots, university student information systems, and admission services, reporting that retrieval-based grounding can improve answer relevancy, faithfulness, and context precision [8-11]. Recent educational RAG surveys further indicate that document-grounded chatbots are increasingly used for learning support and institutional information services, and that official documents and university resources can

be transformed into knowledge bases for student-facing systems [12-14]. Domain-specific RAG chatbots have also been evaluated in other professional service contexts, such as patient education, which further demonstrates the importance of grounding generated answers in authoritative documents [15]. These studies confirm the relevance of RAG in higher education, but they also show that performance depends on retrieval quality, chunking strategy, embedding model, and generator behavior.

Evaluation remains a central challenge in RAG systems. Recent surveys and frameworks emphasize that RAG evaluation should not be limited to final answer quality because errors may arise during chunking, retrieval, reranking, or generation [16], [17]. RAGAS and related evaluation frameworks further highlight the importance of measuring context precision, context recall, faithfulness, and answer relevancy [18], [19]. Trustworthiness is also important because RAG does not automatically eliminate hallucination; a retriever may still return weakly related chunks for unsupported questions, and a generator may still produce misleading answers if no refusal mechanism is applied [20-23].

Prior educational RAG studies mainly emphasize retrieval quality, answer relevance, faithfulness, or overall chatbot performance, with less attention to jointly controlling unsupported answers and unnecessary refusals in Indonesian institutional settings. This study addresses that gap by integrating semantic retrieval, adaptive thresholding, guardrails, local generation, and retry-on-refusal. Its contributions are a controlled Indonesian academic-service benchmark, separate evaluation of retrieval, generation, and refusal behavior, threshold trade-off analysis, and retrieval ablation. The objective is to determine whether the pipeline can answer supported questions and appropriately refuse unsupported or out-of-context requests.

2. Methods

This study used a controlled experimental design to evaluate how alternative pipeline configurations affected retrieval performance, generated-answer alignment, and refusal behavior under a fixed annotated benchmark. This design permitted systematic comparison of retrieval settings, refusal thresholds, and retry-on-refusal behavior while keeping the underlying language model fixed. The experimental procedure comprised document preprocessing, knowledge-base construction, semantic retrieval, refusal control, local generation, retry-on-refusal, and performance evaluation.

Figure 1 separates offline knowledge-base construction from online question answering. During offline processing, official documents are cleaned, segmented into chunks, encoded using BGE-M3, and stored in a FAISS vector index. During online processing, the system retrieves the top-*k* document chunks, evaluates the maximum retrieval score against the adaptive refusal threshold, and screens the query using rule-based guardrails. Queries that fail either reliability control are refused; otherwise, Qwen2.5 generates an answer grounded in the retrieved evidence. If the initial generation produces a refusal for a non-guardrail query with a maximum retrieval score of at least 0.58, one retry is performed, followed by top-chunk fallback when necessary. This architecture allows retrieval, refusal, and generation behavior to be evaluated separately.

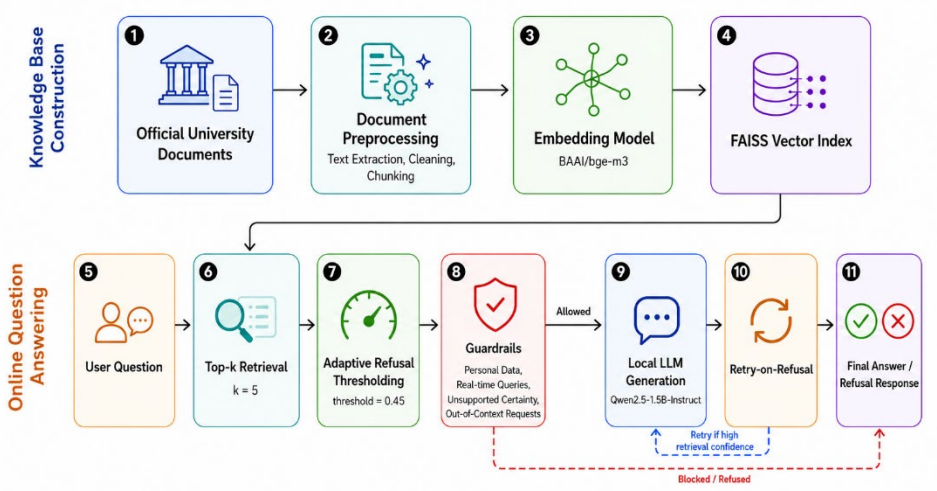


Figure 1. Overall architecture of the proposed RAG-based academic service chatbot

2.1 Document Collection and Preprocessing

The knowledge base consisted of 12 official academic documents, including academic guidelines and standard operating procedures for re-registration, course registration plans (*Kartu Rencana Studi*, KRS), course registration revision and cancellation, semester grade reports (*Kartu Hasil Studi*, KHS), academic leave, academic reactivation, thesis title submission, thesis examination scheduling and implementation, graduation approval (*yudisium*), and active student certificates.

Text was extracted from the documents, cleaned by removing excessive whitespace and formatting artifacts, and stored with metadata such as document identifier, title, and page number. The extracted collection produced 99 pages or text units, which were segmented into 99 retrievable chunks using a chunk size of 500 words and an overlap of 100 words. This overlap was intended to preserve contextual continuity for procedural academic information that may span adjacent paragraphs.

2.2 Embedding, Retrieval, and Vector Indexing

Each text chunk was encoded using BAAI/bge-m3, a multilingual embedding model [24] selected for semantic retrieval over Indonesian academic documents. The resulting vectors were stored in a FAISS index to support efficient similarity search. Query and document similarity was computed using cosine similarity as expressed in Equation 1, where q denotes the query embedding and d_i denotes the embedding of document chunk i.

$$\text{sim}(q, d_i) = \frac{q \cdot d_i}{\|q\| \, \|d_i\|} \quad (1)$$

For each input question, the retriever returned the top-k chunks with the highest similarity scores. The final system used $k = 5$ because the ablation study showed that this setting provided stronger recall while maintaining high ranking quality. The technical design is consistent with dense passage retrieval, vector similarity search, sentence embeddings, and recent RAG implementation practices [4-7], [14], [25-28].

Table 1. Main Experimental Configuration

Component	Configuration
Embedding model	BAAI/bge-m3
Vector index	FAISS
Chunk size	500 words
Chunk overlap	100 words
Retrieval setting	top_k = 5
Retrieval threshold	0.45
Local generator	Qwen/Qwen2.5-1.5B-Instruct
Reliability strategy	Adaptive threshold, guardrail, retry-on-refusal
Language	Indonesian

The main configuration is summarized in Table 1. The configuration combines a multilingual embedding model, vector indexing, threshold-based refusal, and local generation to balance retrieval coverage, response quality, and refusal reliability.

2.3 Adaptive Refusal Thresholding and Guardrails

A refusal threshold was added because a vector retriever always returns top-k chunks, even when a question is outside the supported document domain. The response decision follows Equation 2, where τ is the retrieval threshold, D_k is the set of retrieved chunks, and $g(q)$ denotes the guardrail predicate for personal, real-time, unsupported-certainty, or out-of-context requests. The threshold was tuned using retrieval confidence and false-refusal constraints, and the final value was τ = 0.45.

$$s_{\max} = \max_{d_i \in D_k} \text{sim}(q, d_i)$$

$$R(q) = \begin{cases} \text{Refusal}, & s_{\max} < \tau \text{ or } g(q) = 1 \\ \text{Answer}, & \text{otherwise} \end{cases} \quad (2)$$

Rule-based guardrails were also used for questions that should not be answered from general academic documents, including requests involving passwords, personal GPA or grades, private contact numbers, real-time information, file-processing status, or unsupported certainty such as guaranteed thesis acceptance. These guardrails were designed to reduce hallucination and unsafe responses, which remain important concerns in LLM-based systems [29-34].

2.4 Local Generation and Retry-on-Refusal

After the retrieval and refusal mechanism was stable, the local generator Qwen/Qwen2.5-1.5B-Instruct [35] was activated to produce concise Indonesian answers based on the retrieved context. The prompt required the model to answer only from official document context, avoid unsupported claims, and return a standardized refusal when the answer was unavailable.

During testing, the local generator sometimes refused answerable questions despite high retrieval scores. Therefore, retry-on-refusal was introduced. The retry was triggered when the first answer began with a refusal, the question was not blocked by guardrails, and the maximum retrieval score was at least 0.58. If the retry still produced a refusal, the system used the top retrieved chunk as fallback evidence.

2.5 Dataset and Experimental Workflow

The evaluation dataset contained 160 Indonesian academic service questions. Each question was annotated with category, reference answer, gold document identifier, gold keywords, answerability status, expected behavior, and retrieval expectation. The experimental workflow is shown in Figure 2. It begins with document preparation and knowledge base construction, continues with retrieval and refusal evaluation, and ends with generator evaluation and ablation analysis.

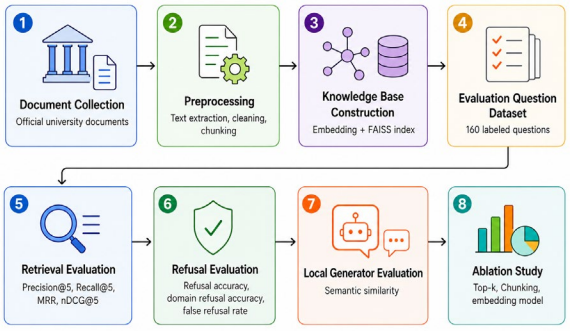


Figure 2. Experimental workflow for evaluating retrieval, refusal, generation, and ablation settings

Figure 2 summarizes eight stages: document collection, preprocessing, knowledge-base construction, dataset construction, retrieval evaluation, refusal evaluation,

local generation evaluation, and ablation. The sequence separates evidence-retrieval errors from refusal and generation errors and isolates the effects of top-k, chunk overlap, and embedding model.

Table 2. Distribution of evaluation questions

Category	Description	Number
Factual	Direct factual information	48
Procedural	Academic service procedures	40
Requirement-based	Required documents or conditions	32
Multi-hop	Information from multiple parts	16
Ambiguous	Broad or unclear questions	8
Contextual unanswerable	Related but unsupported certainty	2
Unanswerable	Not answerable from documents	6
Out-of-context	Outside academic service domain	8
Total		160

Table 2 summarizes the distribution of the 160 evaluation questions across answerable, ambiguous, contextual-unanswerable, unanswerable, and out-of-context categories. The dataset was designed as a controlled benchmark rather than as a statistically representative sample of all student-chatbot interactions. It was stratified across eight categories and used only to evaluate a fixed pipeline. The refusal analysis included 144 answerable and 16 expected-refusal questions; therefore, the resulting proportion-based estimates remain statistically fragile. Perfect observed performance should not be interpreted as error-free operational behavior, and larger datasets derived from natural helpdesk or chatbot logs are required for more precise and generalizable estimates.

After the evaluation dataset was constructed, the refusal threshold was selected based on the trade-off shown in Figure 3. Values below 0.45 did not correctly refuse all expected-refusal cases, whereas values above 0.50 increasingly produced false refusals. A threshold of 0.45 was selected as the lowest evaluated value that achieved domain refusal accuracy of 1.0000 while retaining all 144 answerable questions and producing an observed false refusal rate of 0.0000.

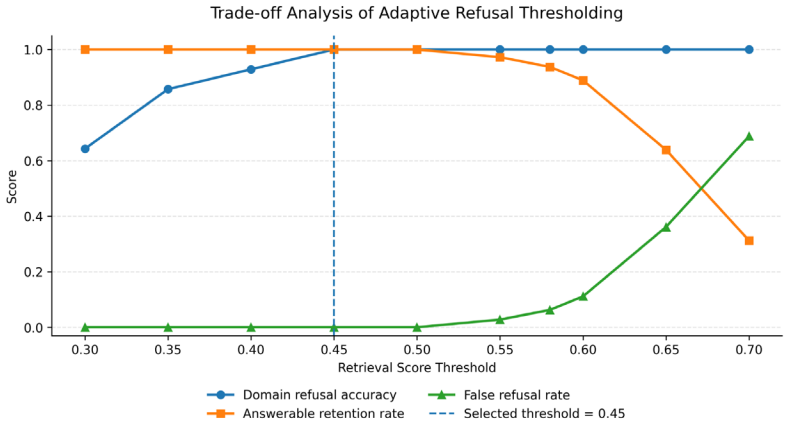


Figure 3. Trade-off analysis of adaptive refusal thresholding

2.6 Evaluation Metrics

Retrieval performance was measured using Precision@k, Recall@k, MRR, and nDCG@k. Precision@k, as shown in Equation 3, measures the proportion of retrieved chunks that are relevant. Recall@k, as shown in Equation 4, follows the implementation used in this experiment, namely a hit-based recall averaged over all questions. A question receives a recall value of one when at least one relevant gold evidence item appears in the top-k results and zero otherwise. In these equations, $R_{k,i}$ represents the top-k retrieved chunks for question i and G_i represents the relevant gold evidence for that question.

$$\text{Precision@k} = \frac{1}{N}\sum_{i=1}^{N}\frac{|R_{k,i} \cap G_i|}{k} \quad (3)$$

$$\text{Recall@k} = \frac{1}{N}\sum_{i=1}^{N} 1\left(|R_{k,i} \cap G_i| > 0\right) \quad (4)$$

Ranking quality was measured using MRR and nDCG@k. MRR, as shown in Equation 5, evaluates how early the first relevant chunk appears, where $rank_i$ is the position of the first relevant chunk for question i. If no relevant chunk is retrieved, the contribution for that question is set to zero. DCG@k and nDCG@k are given in Equations 6 and 7, where rel_i denotes the binary relevance value at rank i and IDCG@k denotes the ideal DCG value for the retrieved relevance labels.

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i} \quad (5)$$

$$\text{DCG@k} = \sum_{i=1}^{k}\frac{\text{rel}_i}{\log_2(i+1)} \quad (6)$$

$$\text{nDCG@k} = \frac{\text{DCG@k}}{\text{IDCG@k}} \quad (7)$$

Generation quality was measured using semantic similarity (SS) between the normalized embedding of the reference answer and the normalized embedding of the system-generated answer, as shown in Equation (8). Although Equations 1 and 8 use the same cosine-based formulation, they are applied to different embedding pairs: query-document embeddings for retrieval and reference-generated answer embeddings for response evaluation.

$$\text{SS}(a, \hat{a}) = \frac{a \cdot \hat{a}}{\|a\|\|\hat{a}\|} \quad (8)$$

Refusal behavior was evaluated using refusal accuracy (RA) and false refusal rate (FRR), as expressed in Equations 9 and 10. $TP_{refusal}$ denotes correctly refused questions, $FN_{refusal}$ denotes expected refusals that were incorrectly answered, $FP_{refusal}$ denotes answerable questions that were incorrectly refused, and $N_{answerable}$ denotes all answerable questions. Domain refusal accuracy was computed using the same refusal accuracy definition on the subset of expected-refusal questions that did not require academic document retrieval. Statistical uncertainty for the two proportion-based refusal metrics was summarized with two-sided 95% exact Clopper–Pearson binomial confidence intervals.

$$RA = \frac{TP_{refusal}}{TP_{refusal}+FN_{refusal}} \quad (9)$$

$$FRR = \frac{FP_{refusal}}{N_{answerable}} \quad (10)$$

2.7 Ablation Study

The ablation study compared the effects of retrieval configuration on performance. The compared settings varied top-k values, chunk overlap, and embedding model.

The experiments included BGE-M3 with fixed chunking at $k = 3$ and $k = 5$, BGE-M3 with overlap chunking at $k = 5$ and $k = 10$, and multilingual-e5-base with overlap chunking at $k = 5$. The study was intended to justify the final configuration through empirical comparison rather than assumption.

3. Results and Discussions

The proposed system was evaluated using the final configuration described in Table 1. The results are organized into retrieval performance, refusal reliability, local generation quality, ablation analysis, and research implications. Table 3 reports the overall point estimates, while Figure 4 shows the descriptive effect of retry-on-refusal on generation and refusal metrics.

Table 3. Overall performance of the final RAG configuration

Aspect	Metric	Score
Retrieval	Precision@5	0.7250
Retrieval	Recall@5	0.9125
Retrieval	MRR	0.8940
Retrieval	nDCG@5	0.8888
Generation	Mean semantic similarity	0.8516
Generation	Answerable-only similarity	0.8392
Generation	Expected-refusal similarity	0.9639
Refusal	Refusal accuracy	1.0000
Refusal	Domain refusal accuracy	1.0000
Refusal	False refusal rate	0.0000

Within the controlled dataset, the final system answered all 144 answerable questions and refused all 16 expected-refusal questions. No expected-refusal case was answered and no answerable case was refused, producing observed refusal accuracy of 1.0000 and an observed false refusal rate of 0.0000. These are sample-specific point estimates; their statistical uncertainty and interpretation are discussed in Section 3.2.

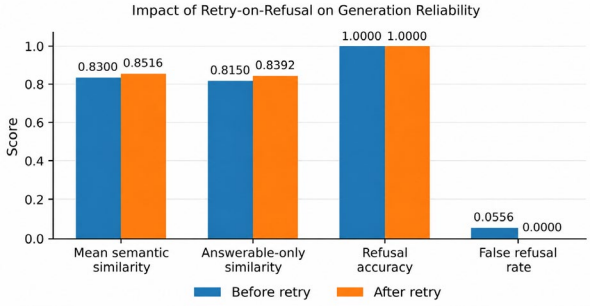


Figure 4. Impact of retry-on-refusal on generation reliability

3.1 Retrieval and Category-Level Performance

The final retriever achieved Recall@5 of 0.9125 and MRR of 0.8940, indicating that relevant academic document chunks were usually found within the top five results and often ranked near the top. Precision@5 of 0.7250 shows that some retrieved chunks were semantically related but not always directly needed. This trade-off is acceptable in academic service chatbots because missing relevant official information is more harmful than retrieving a small amount of additional context.

Table 4. Retrieval performance by question category

Category	P@5	R@5	MRR	nDCG@5
Factual	0.8917	1.0000	0.9896	0.9844
Procedural	0.7750	1.0000	0.9750	0.9758
Requirement	0.6750	1.0000	0.9714	0.9670
Multi-hop	0.7750	1.0000	0.9583	0.9456
Ambiguous	0.8750	1.0000	1.0000	0.9835
Contextual unanswerable	0.6000	1.0000	1.0000	0.9914
Unanswerable	0.0000	0.0000	0.0000	0.0000
Out-of-context	0.0000	0.0000	0.0000	0.0000

Table 4 shows that all document-related categories with annotated gold evidence reached Recall@5 of 1.0000. The retriever consistently found relevant contexts for factual, procedural, requirement-based, multi-hop, ambiguous, and contextual-unanswerable questions. The requirement category had lower precision because similar administrative terms appeared across several procedures. For unanswerable and out-of-context questions, the zero retrieval scores indicate that these questions had no annotated gold evidence in the official document collection; therefore, their success was evaluated through refusal metrics rather than retrieval recall.

3.2 Refusal Reliability and Generator Quality

On the evaluated dataset, the refusal mechanism correctly handled all 16 expected-refusal questions and did not incorrectly refuse any of the 144 answerable questions. This produced observed refusal accuracy of 1.0000, domain refusal accuracy of 1.0000, and a false refusal rate of 0.0000. Nevertheless, the 95% exact binomial confidence interval for refusal accuracy was 0.7941–1.0000, reflecting the small expected-refusal subset, while the interval for the false refusal rate was 0.0000–0.0253. The point estimates therefore demonstrate successful behavior on the controlled benchmark but do not establish error-free refusal in

deployment. Adaptive thresholding and guardrails functioned as reliability controls by preventing generation from weakly related evidence or from requests that required private, real-time, or unsupported information.

Retry-on-refusal produced descriptive improvements on the same controlled benchmark. Mean semantic similarity increased from 0.8300 to 0.8516, answerable-only similarity increased from 0.8150 to 0.8392, and the false refusal rate decreased from 0.0556 to 0.0000, while refusal accuracy remained 1.0000. Figure 4 visualizes these before-and-after changes. The expected-refusal similarity was 0.9639 because both the reference and system outputs used a standardized refusal statement; therefore, answerable-only similarity is also reported to avoid inflating the interpretation of generation quality. These changes are descriptive and are not presented as inferential evidence of population-level improvement.

3.3 Ablation Analysis

The ablation study results are presented in Table 5 and visualized in Figure 5. E1, which used BGE-M3 with fixed chunking and top-k = 3, achieved the highest Precision@k and nDCG@k, indicating that a smaller retrieval set produced cleaner retrieval contexts. However, its Recall@k was slightly lower than the top-five configurations, suggesting a higher risk of missing relevant procedural evidence.

E2 and E3 produced identical results across all evaluation metrics, suggesting that chunk overlap had limited impact on this dataset. This outcome is likely because the academic service documents were relatively short, structured, and procedurally explicit. E4, which increased top-k to 10, did not improve Recall@k and instead reduced Precision@k and nDCG@k. This indicates that retrieving too many chunks introduced additional noise without improving evidence coverage. E5, which used E5-base with overlap and top-k = 5, produced lower MRR and nDCG@k than the BGE-M3 configurations, indicating that BGE-M3 was more suitable for Indonesian academic service documents in this experiment.

Table 5. Ablation study of retrieval configurations

Exp.	Configuration	P@k	R@k	MRR	nDCG
E1	BGE-M3, fixed, k=3	0.8104	0.9063	0.8927	0.8921
E2	BGE-M3, fixed, k=5	0.7250	0.9125	0.8940	0.8888
E3	BGE-M3, overlap, k=5	0.7250	0.9125	0.8940	0.8888
E4	BGE-M3, overlap, k=10	0.6413	0.9125	0.8940	0.8714
E5	E5-base, overlap, k=5	0.7113	0.9125	0.8726	0.8720

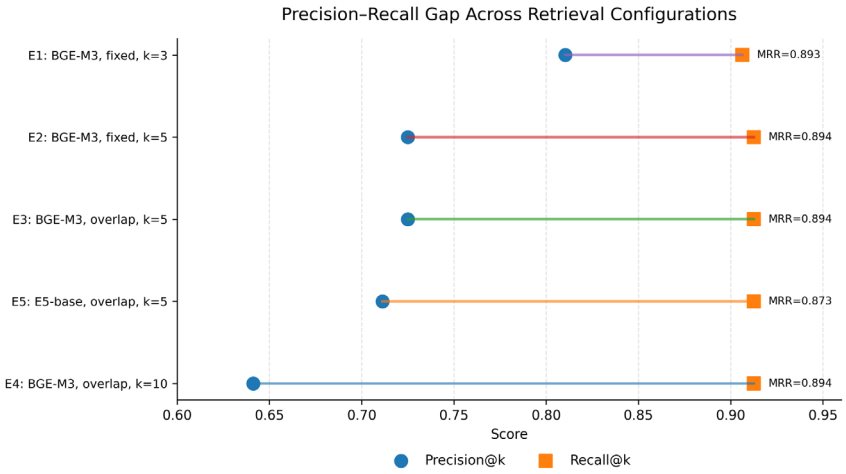


Figure 5. Precision-recall gap across retrieval configurations

The final configuration retained BGE-M3 with top-k = 5 because it provided stronger Recall@k while maintaining high MRR. Although E1 achieved higher Precision@k, E2 and E3 were more appropriate for academic service use because they reduced the risk of omitting relevant procedural evidence. In Figure 5, the horizontal distance between each precision and recall marker represents the coverage-noise trade-off, while the adjacent MRR label indicates how early the first relevant chunk was ranked. The visualization shows that top-k = 10 enlarged the precision-recall gap without improving recall or MRR, whereas the selected top-five setting preserved evidence coverage with less retrieval noise.

These results support previous findings that RAG configurations should be selected empirically, as chunking strategy, top-k selection, embedding model choice, and educational-domain retrieval design can affect both retrieval context quality and downstream generation performance [14], [16-19], [25-28].

3.4 Limitations and Future Work

This study has several limitations. First, the knowledge base was derived from a single institutional context, so performance may differ across universities with different document structures, terminology, and academic regulations. Second, the 160-question dataset was a controlled benchmark and may not fully represent

naturally occurring student messages containing typographical errors, informal language, code-mixing, incomplete context, adversarial wording, or multi-turn interactions. Third, the refusal evaluation included only 16 expected-refusal cases. Although all were handled correctly, the corresponding perfect point estimate has a wide exact confidence interval and remains statistically fragile; it should not be interpreted as a population-level guarantee. Fourth, generation quality was evaluated mainly through semantic similarity and refusal metrics, so human assessment is still needed for factual accuracy, clarity, completeness, helpfulness, and trustworthiness. Fifth, the local generator was relatively small. Larger Indonesian-capable models may improve fluency but may also increase latency and deployment cost. Future work should evaluate larger real helpdesk logs, conduct human and subgroup analyses, test hybrid retrieval and reranking, compare larger generators, support multi-turn conversations, and deploy the chatbot with appropriate privacy and access-control mechanisms.

4. Conclusions

This study developed and evaluated a document-grounded RAG chatbot for Indonesian academic services. The final pipeline combined BGE-M3 retrieval, FAISS indexing, adaptive refusal thresholding, rule-based guardrails, a local Qwen2.5 generator, and retry-on-refusal. It achieved Precision@5 of 0.7250, Recall@5 of 0.9125, MRR of 0.8940, nDCG@5 of 0.8888, mean semantic similarity of 0.8516, and answerable-only similarity of 0.8392. Within the controlled sample, all 16 expected-refusal questions were refused and none of the 144 answerable questions was falsely refused, corresponding to observed refusal accuracy of 1.0000 and a false refusal rate of 0.0000. Retry-on-refusal also reduced the false refusal rate from 0.0556 to 0.0000 and improved answerable-only similarity from 0.8150 to 0.8392. These results provide initial evidence that combining retrieval confidence, guardrails, and retry supports a better balance between helpfulness and unsupported-response prevention. However, the perfect refusal proportions are specific to a small, structured, single-institution benchmark and should not be treated as guarantees of operational performance. The ablation study further showed that increasing top-k did not improve evidence coverage and that BGE-M3 provided stronger ranking quality than multilingual-e5-base in this dataset. Overall, the system offers an empirically evaluated initial framework for reliable Indonesian academic-service question answering, subject to validation with larger real-world queries and human evaluation.

Declarations

AI Usage Statement

The authors used generative artificial intelligence tools, specifically ChatGPT (OpenAI) and Claude (Anthropic), to assist with language editing, manuscript organization, formatting, and limited code assistance. All scientific content, methodological decisions, data analysis, interpretation of results, and conclusions were independently reviewed and verified by the authors, who take full responsibility for the content of the manuscript.

Author Contributions

J.B.: conceptualization, methodology, software, formal analysis, visualization, writing – original draft, and writing – review and editing. T.M.T., S.P., and M.H.: validation and writing – review and editing. All authors read and approved the final manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest Statement

The authors declare no conflict of interest.

Ethical Clearance

Not applicable. This study did not involve human participants or animal subjects. The evaluation used official institutional documents and a controlled benchmark and did not involve personal or sensitive participant data.

Data and Code Availability

The anonymized evaluation dataset and implementation scripts supporting the findings of this study are available from the corresponding author upon reasonable request. The original official institutional documents are not publicly redistributed because they are institutional documents. Relevant document metadata, preprocessing settings, chunking parameters, and instructions for reconstructing the knowledge base from authorized copies can also be provided upon reasonable request.

References

[1] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, in NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010. doi: 10.5555/3295222.3295349.

[2] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, in NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020, pp. 9459–9474. doi: 10.5555/3495724.3496517.

[3] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "REALM: retrieval-augmented language model pre-training," in *Proceedings of the 37th International Conference on Machine Learning*, in ICML'20. JMLR.org, 2020. doi: 10.5555/3524938.3525306.

[4] V. Karpukhin *et al.*, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[5] J. Johnson, M. Douze, and H. Jégou, "Billion-Scale Similarity Search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021, doi: 10.1109/TBDATA.2019.2921572.

[6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[7] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. doi: 10.18653/v1/2021.emnlp-main.552.

[8] P. A. S. Artayasa, A. S. Kusuma, and P. Sugiartawan, "Enhancing Academic Chatbot Accuracy With Retrieval-Augmented Generation in Higher Education," in *2025 IEEE International Symposium on Consumer Technology (ISCT)*, 2025, pp. 427–432. doi: 10.1109/ISCT66099.2025.11297210.

[9] L. S. Hartono, E. I. Setiawan, and V. Singh, "Retrieval Augmented Generation-Based Chatbot for prospective and current university students," *Int. J. Eng. Sci. Inf. Technol*, vol. 5, no. 3, pp. 268–277, 2025, doi: 10.52088/ijesty.v5i3.951.

[10] M. Khasanova Zafar kizi and Y. Suh, "Design and Performance Evaluation of LLM-Based RAG Pipelines for Chatbot Services in International Student Admissions," *Electronics (Basel).*, vol. 14, no. 15, 2025, doi: 10.3390/electronics14153095.

[11] S. Neupane *et al.*, "From Questions to Insightful Answers: Building an Informed Chatbot for University Resources," in *2024 IEEE Frontiers in Education Conference (FIE)*, Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2024, pp. 1–9. doi: 10.1109/FIE61694.2024.10892994.

[12] M. L. Husain, Y. Wibisono, and A. Anisyah, "Development of an Academic Services Chatbot Based on Retrieval-Augmented Generation (RAG)," *Brilliance: Research of Artificial Intelligence*, vol. 5, no. 2, pp. 727–735, Aug. 2025, doi: 10.47709/brilliance.v5i2.6719.

[13] J. Swacha and M. Gracel, "Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications," *Applied Sciences*, vol. 15, no. 8, 2025, doi: 10.3390/app15084234.

[14] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100417, 2025, doi: 10.1016/j.caeai.2025.100417.

[15] D. Baur, J. Ansorg, C.-E. Heyde, and A. Voelker, "Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study," *JMIR AI*, vol. 4, p. e75262, 2025, doi: 10.2196/75262.

[16] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of Retrieval-Augmented Generation: A Survey," in *Big Data*, W. Zhu, H. Xiong, X. Cheng, L. Cui, Z. Dou, J. Dong, S. Pang, L. Wang, L. Kong, and Z. Chen, Eds., Singapore: Springer Nature Singapore, 2025, pp. 102–120. doi: 10.1007/978-981-96-1024-2_8.

[17] J. Liu, R. Ding, L. Zhang, P. Xie, and F. Huang, "Cofe-rag: A comprehensive full-chain evaluation framework for retrieval-augmented generation with enhanced data diversity," *arXiv preprint arXiv:2410.12248*, 2024, doi: 10.48550/arXiv.2410.12248.

[18] S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, N. Aletras and O. De Clercq, Eds., St. Julians, Malta: Association for Computational Linguistics, Mar. 2024, pp. 150–158. doi: 10.18653/v1/2024.eacl-demo.16.

[19] Y. Lyu *et al.*, "CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, Jan. 2025, doi: 10.1145/3701228.

[20] P. Zhao *et al.*, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," *Data Sci. Eng.*, vol. 11, no. 1, pp. 1–29, 2026, doi: 10.1007/s41019-025-00335-5.

[21] Y. Hu and Y. Lu, "Rag and rau: A survey on retrieval-augmented language model in natural language processing," *arXiv preprint arXiv:2404.19543*, 2024, doi: 10.48550/arXiv.2404.19543.

[22] Y. Zhou *et al.*, "Trustworthiness in retrieval-augmented generation systems: A survey," *arXiv preprint arXiv:2409.10102*, 2024, doi: 10.48550/arXiv.2409.10102.

[23] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023, doi: 10.1145/3571730.

[24] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2318–2335. doi: 10.18653/v1/2024.findings-acl.137.

[25] C. A. Gomez-Cabello *et al.*, "Comparative Evaluation of Advanced Chunking for Retrieval-Augmented Generation in Large Language Models for Clinical Decision Support," *Bioengineering*, vol. 12, no. 11, 2025, doi: 10.3390/bioengineering12111194.

[26] R. Akkiraju *et al.*, "Facts about building retrieval augmented generation-based chatbots," *arXiv preprint arXiv:2407.07858*, 2024, doi: 10.48550/arXiv.2407.07858.

[27] A. James, M. Trovati, and S. Bolton, "Retrieval-Augmented Generation to Generate Knowledge Assets and Creation of Action Drivers," *Applied Sciences*, vol. 15, no. 11, 2025, doi: 10.3390/app15116247.

[28] A. Ammar, A. Koubaa, O. Nacar, and W. Boulila, "Optimizing retrieval-augmented generation: analysis of hyperparameter impact on performance and efficiency," *arXiv preprint arXiv:2505.08445*, 2025, doi: 10.48550/arXiv.2505.08445.

[29] B. Saha, U. Saha, and M. Z. Malik, "QuIM-RAG: Advancing Retrieval-Augmented Generation With Inverted Question Matching for Enhanced QA Performance," *IEEE Access*, vol. 12, pp. 185401–185410, 2024, doi: 10.1109/ACCESS.2024.3513155.

[30] Y. Yu *et al.*, "RankRAG: unifying context ranking with retrieval-augmented generation in LLMs," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, in NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2024. doi: 10.5555/3737916.3741766.

[31] A. Abdallah, M. Abdalla, B. Piryani, J. Mozafari, M. Ali, and A. Jatowt, "RerankArena: A Unified Platform for Evaluating Retrieval, Reranking and RAG with Human and LLM Feedback," in *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, in CIKM '25. New York, NY, USA: Association for Computing Machinery, 2025, pp. 6593–6597. doi: 10.1145/3746252.3761484.

[32] Y. Katsis *et al.*, "mtRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems," *Trans. Assoc. Comput. Linguist.*, vol. 13, pp. 784–808, 2025, doi: 10.1162/tacl.a.19.

[33] G. Mialon *et al.*, "Augmented language models: a survey," *arXiv preprint arXiv:2302.07842*, 2023, doi: 10.48550/arXiv.2302.07842.

[34] V. Raina and M. Gales, "Question-Based Retrieval using Atomic Units for Enterprise RAG," in *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, M. Schlichtkrull, Y. Chen, C. Whitehouse, Z. Deng, M. Akhtar, R. Aly, Z. Guo, C. Christodoulopoulos, O. Cocarascu, A. Mittal, J. Thorne, and A. Vlachos, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 219–233. doi: 10.18653/v1/2024.fever-1.25.

[35] A. Yang *et al.*, "Qwen2.5 Technical Report," *arXiv preprint arXiv:2412.15115*, 2025, doi: 10.48550/arXiv.2412.15115.