Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 8, No. 3, November 2025, pp. 687 – 696
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

A RAG-Based Academic Information Chatbot Using Lightweight LLaMA and Indo-Sencence-BERT

[1]Muhamad Saman, [2*]Gusti Ahmad Fanshuri Alfarisy, [3]Rizky Amelia, [4]Nisa Rizqiza Fadhliana
[1,2,3,4]Faculty of Science and Information Technology, Kalimantan Institute of Technology, Indonesia
Email: samanmuhammad077@gmail.com, [2]gusti.alfarisy@lecturer.itk.ac.id,
[3]rizky.amelia@lecturer.itk.ac.id, [4]nisafadhliana@lecturer.itk.ac.id

Article Info
Article history:
Received Aug 12th, 2025
Revised Oct 08th, 2025
Accepted Nov 02nd, 2025

Keyword:
Chatbot
Generation
Large Language Model
Natural Language Processing
Retrieval Augmented
Transformer

ABSTRACT
In the current digital era, Institut Teknologi Kalimantan (ITK) encounters challenges in delivering academic information that is fast, accurate, and easily accessible to students, lecturers, and academic staff. Access to important information, such as administrative procedures, report writing guidelines, and academic policies, remains largely reliant on manual systems and static handbooks. To address this issue, this study investigates a chatbot system that utilizes the Retrieval-Augmented Generation (RAG) framework, specifically the LLaMA model. The chatbot combines semantic retrieval and natural language generation to provide relevant and accurate answers based on existing academic documents. Evaluation was conducted on two lightweight LlaMA models: 1.5 and 3B parameters. Furthermore, different embedding vectors were also evaluated along with Indo-Sentence-BERT, as well as the chunking size. The most optimal configuration was achieved using LLaMA 3B as the generative model and Indo-Sentence-BERT as the retriever, with a chunk size of 200 tokens and an overlap of 10 tokens. This setup achieved a RAGAS score of approximately 0.9, a competitive MRR of 0.5, and response latency under 1 second. Although LLaMA 1B recorded a higher MRR (0.6), its low RAGAS score made it less favorable. Overall, the LLaMA 3B and Indo-Sentence-BERT configuration is recommended to enhance the efficiency of academic information retrieval at ITK.

Corresponding Author:
Gusti Ahmad Fanshuri Alfarisy,
Faculty of Science and Information Technology,
Department of Informatics,
Kalimantan Institute of Technology,
Soekarno-Hatta Street Km.15, Karang Joang, Balikpapan, East Kalimantan, Indonesia
Email: gusti.alfarisy@lecturer.itk.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v8i3.38150

1. INTRODUCTION

In recent years, chatbots have gained traction as a promising approach to improve responsiveness and efficiency in information systems. Early works primarily relied on rule-based methods. For example, one study developed a PHP- and MySQL-based chatbot using rule-based logic to support new student admissions at Universitas Nasional, demonstrating reliable responses tailored to user needs [1]. Another research employed Artificial Intelligence Markup Language (AIML) for a web-based academic chatbot, achieving high accuracy and user satisfaction [2]. These studies highlight the feasibility of chatbots in academic contexts, albeit with limited scalability due to their dependence on predefined rules.

With the rise of transformer networks [3] for natural language processing tasks, Large Language Models (LLMs) have significantly advanced chatbot capabilities. Unlike rule-based systems, LLM-based chatbots utilize large-scale training data to recognize language patterns and generate contextually relevant

text without explicit rules, leveraging a deep network [4]. LLM can be used to solve several task such as question answering, machine translation, and general-purpose language understanding and generation [5], [6]

Unfortunately, LLMs alone are prone to hallucination, a phenomenon where the model generates factually incorrect or nonsensical output. This arises from the core training mechanism of maximizing token likelihood, which can prioritize fluency and coherence over factual accuracy. Consequently, hallucination poses a significant threat to the trustworthiness and reliability of information generated by LLMs for real-world applications [7].

Recent developments further extend these models by combining generative components with retrieval pipelines, commonly referred to as Retrieval-Augmented Generation (RAG), to tackle hallucination. RAG works by retrieving relevant information via chunks or documents and forwarding them to the LLM to synthesize further, which provides the context [8]. This approach has been shown to reduce hallucinations and improve factual consistency, making it particularly suitable for knowledge-intensive applications [9].

Several studies have demonstrated the potential of RAG-based chatbots in higher education. One investigation integrated RAG with vector databases to improve chatbot accuracy, achieving 86.84% accuracy while significantly reducing hallucination [5]. Another introduced BARKPLUG V.2, a RAG-based LLM system tailored for university resources, which reported strong performance with an average RAGAS score of 0.96 and positive user feedback [11]. In another domain, a study on healthcare applied RAG with Indo-Sentence-BERT and LLaMA 3.1, showing strong results across metrics such as Mean Reciprocal Rank (MRR) and semantic similarity [12]. Furthermore, the survey of employing RAG for educational applications suggesting that educational policy in RAG becomes one of the important aspects [13]. These findings suggest that RAG provides both accuracy and adaptability across domains.

The effectiveness of such systems relies heavily on the choice of language models and embeddings. Sentence-BERT (SBERT) [14], Indo-Sentence-BERT, and LlaMA [15] have emerged as leading candidates for retrieval and generative tasks. SBERT provides semantically meaningful embeddings with efficient sentence-level similarity comparisons, while Indo-Sentence-BERT extends this capability to Indonesian corpora. On the generative side, LLaMA provides efficient autoregressive generation with fewer parameters compared to GPT-3 while maintaining competitive performance. These advancements underscore the importance of systematically evaluating different configurations of embeddings, chunking strategies, and LLMs in the context of RAG.

The rapid advancement of information technology has encouraged higher education institutions to provide information services that are fast, accurate, and easily accessible. In many higher education institutions, like Institut Teknologi Kalimantan (ITK), the delivery and accessibility of academic information remain a challenge. This is mainly due to the reliance on manual procedures and physical documents, which are increasingly inefficient in the digital era. As a result, there is a pressing need to develop innovative solutions to streamline academic information services.

Our work focuses on the challenging task of knowledge-intensive Q&A over Indonesian text in academic documents. The developed model, a combined RAG pipeline leveraging the compact LLaMA 3.2 family, was necessitated by the need for a resource-efficient and localized solution. While prior work on RAG [9] confirms the pipeline's utility and studies on model efficiency [8] highlight the potential of compact LLMs, the performance of models like the LLaMA 3.2 as both a retriever and generator paired with localized bi-encoders like Indo-Sentence-BERT remains underexplored.

In light of these developments, this study focuses on exploring the optimal configuration of RAG for academic information services at ITK. Particular attention is given to evaluating combinations of chunk sizes, overlaps, embeddings, and LLMs in terms of accuracy, contextual relevance, and system latency. The overarching aim is to identify a configuration that balances performance and efficiency, thereby offering a practical recommendation for deploying an academic information chatbot tailored to the institutional needs of ITK.

Our contribution can be summarized as follows:

1. We conducted a comprehensive comparative study to empirically demonstrate the limitations of using lightweight LLMs, specifically the LLaMA 3.2 family (1B and 3B), as general-purpose embedding/retrieval models within RAG systems. Our results validate that dedicated Indo-Sentence-BERT is critically superior for achieving effective semantic retrieval in Bahasa Indonesia.
2. We systematically investigated the impact of various document chunking and overlapping strategies to propose an optimized RAG baseline configuration. This baseline, built upon the high-performing pairing of Indo-Sentence-BERT and LLaMA 3.2, offers a validated starting point for future academic research and practical RAG implementations in the Indonesian language domain.
3. We provide a detailed multi-metric analysis, reporting the performance across RAGAS scores (Faithfulness, Answer Relevance, Context Relevance), MRR, and Latency. This comprehensive reporting provides a crucial trade-off analysis between retrieval effectiveness (MRR) and generative

efficacy (RAGAS), enabling practitioners to select the optimal configuration for efficiency-constrained environments.

2. MATERIAL AND METHOD

2.1. Retrieval-Augmented Generation (RAG)

RAG is an approach that combines retrieval mechanisms with text generation to enhance the accuracy and relevance of responses. Unlike pure LLMs, which rely solely on knowledge encoded during training, RAG enriches the model with external information before producing an output. This concept was introduced in [9] and has been shown to effectively reduce hallucinations in LLMs.

Formally, the generation process in RAG can be modeled as a conditional distribution [4] presented in Equation 1.

$$P(y|x) = \sum_{d \in C} P(y|x, d) \cdot P(d|x) \quad (1)$$

Where x represents the user input, d denotes a retrieved document from corpus C, and y is the generated output. In practice, this computation is approximated by considering only the top-k most relevant documents ($d_1,\ldots,d_k$), leading to Equation 2.

$$P(y|x) \approx \sum_{i \in 1}^{k} P(y|x, d_i) \cdot P(d_i|x) \quad (2)$$

This formulation highlights two essential components: document relevance with respect to the query ($P(d_i|| x)$), and the probability of generating an answer conditioned on the document ($P(y|| x, d_i)$). The same probabilistic framework has been revisited in recent survey work [16], confirming its consistency across later RAG-based studies.

In the context of academic chatbots, the RAG pipeline operates as follows: a user submits a query, then the retriever extracts relevant passages from academic documents such as thesis guidelines or institutional regulations. These passages are concatenated into the prompt, which is then passed to the generative model to produce more factual and context-aware responses. The overall RAG architecture is illustrated in Figure 1 [9], [16], which depicts the interaction between the retrieval and generation modules.

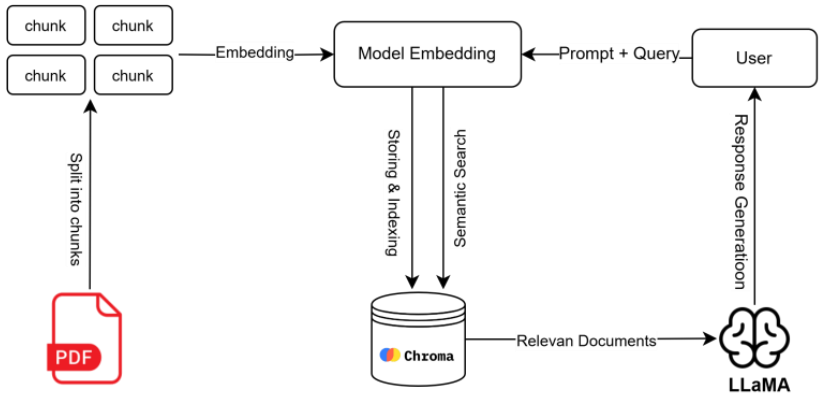

Figure 1. RAG Architecture

RAG provides several advantages over pure LLMs. First, it enables dynamic access to external knowledge bases without requiring model fine-tuning. Second, response quality is improved through semantic retrieval. Third, the system gains transparency, since retrieved passages can be traced back as evidence. These benefits are further supported by dense passage retrieval techniques [17], the application of contrastive learning for improved retrieval representations [18], and large-scale retrieval-enhanced architectures such as Retro [19], which demonstrate that retrieval integration significantly boosts generative quality in massive LLMs.

2.2. LLaMA

Large Language Model Meta AI (LLaMA) is a large-scale language model released by Meta AI with parameter sizes ranging from 7 billion to 65 billion. One of its key advantages lies in its architectural efficiency, enabling relatively smaller models to surpass the performance of much larger LLMs. For instance,

LLaMA-13B outperformed GPT-3, which has 175 billion parameters, on most standard benchmarks, despite having only one-tenth the parameter size [15].

The development of LLaMA was carried out by training on large-scale text datasets drawn entirely from public domains, ensuring compatibility with the open-source ecosystem. The data included CommonCrawl, Wikipedia, GitHub, and other publicly available sources. Text tokenization was performed using the byte-pair encoding (BPE) algorithm implemented via SentencePiece, allowing for more flexible token representation across languages and domains [15].

From an architectural perspective, LLaMA adopts the decoder-only Transformer framework originally introduced in [3], incorporating several crucial modifications to enhance training stability and computational efficiency. These innovations include pre-normalization within Transformer blocks, the SwiGLU activation function, and Rotary Positional Embeddings (RoPE) as an alternative to absolute positional embeddings. Optimization was performed using the AdamW optimizer, combined with strategies such as warm-up steps and controlled weight decay, which together accelerated convergence and reduced computational costs [15].

2.3. Indo-Sentence-BERT

Sentence-BERT (SBERT) is an extension of the BERT model that was modified with a siamese or triplet network architecture to generate semantically meaningful sentence representations. This modification was introduced to address the limitations of BERT and RoBERTa in sentence pair regression tasks, which require the joint processing of both sentences and result in high computational costs. SBERT enables efficient comparison between sentences through cosine similarity [14].

Subsequent studies have focused on improving the quality and efficiency of SBERT-based embeddings. Multiple Negative Ranking (MNR) loss was proposed as an effective method to distinguish between semantically similar and dissimilar sentence pairs [20]. SimCSE introduced a simple contrastive learning method that produces high-quality sentence representations without large-scale annotation [18], while ConSERT reinforced this approach with a self-supervised contrastive learning framework [21]. From a multilingual perspective, L3Cube-IndicSBERT was developed for Indic languages [22]. More recent work has applied layer pruning techniques to improve efficiency while maintaining performance [23].

In the Indonesian context, Indo-Sentence-BERT was developed as an adaptation of SBERT trained on Indonesian sentence pairs annotated with semantic similarity labels. The model employs MNR loss to enhance the effectiveness of embedding for semantic search and retrieval tasks [14]. Indo-Sentence-BERT has been publicly released on the HuggingFace platform under the Apache 2.0 license, making it widely accessible for both academic research and industrial applications.

2.4. Dataset

The dataset used in this study consists of official academic regulation documents issued by the ITK. These documents include guidelines for conducting academic activities such as practical work, internship, final projects, and the Merdeka Belajar Kampus Merdeka (MBKM) program, all provided in PDF format. The documents serving as data sources include several rector regulations issued by ITK, namely:

1. Regulation No. 11 of 2020, containing provisions for internship implementation
2. Regulation No. 12 of 2020, providing comprehensive guidelines for final projects
3. Regulation No. 13 of 2020, concerning the implementation of industrial training
4. Regulation No. 10 of 2021, regarding the implementation of the MBKM program

2.5. Experimental Settings

2.5.1. Experimental Environment

All experiments were conducted on a computer equipped with an AMD Ryzen Threadripper processor, an NVIDIA RTX 3090 GPU (24 GB VRAM), and 32 GB of RAM. The operating system used was Ubuntu 24.04.2 LTS with Python version 3.10. Model implementation was carried out using several libraries, including LangChain to facilitate component integration, SentenceTransformers for access to embedding models, and RAGAS for automatic evaluation of model performance.

2.5.2. Model Configuration

This study's configuration encompasses three main aspects. First, on the retrieval side, three different models were employed: Indo-Sentence-BERT, LLaMA 3.2 1B, and LLaMA 3.2 3B. Second, on the generative side, two LLMs were utilized, namely LLaMA 3.2 1B and LLaMA 3.2 3B. Third, at the preprocessing stage, variations were applied to the chunk size (200, 250, 300, 350, and 400 tokens) and the overlap size (0, 30, 50, 70, and 100 tokens). These configurations were designed as experimental scenarios to evaluate the performance of the RAG system.

2.5.3. Evaluation Metrics

The system evaluation in this study employs several metrics to measure the chatbot's performance based on the RAG approach. The metrics used include RAGAS, MRR, and latency. Each metric assesses different aspects of system performance.

RAGAS comprises three main indicators: faithfulness, answer relevance, and context relevance, which are defined as follows:

1. Faithfulness measures the extent to which the generated answer is based on truly relevant context, by calculating the ratio of validated statements to the total number of statements.
2. Answer relevance evaluates how well the answer aligns with the question posed, by comparing the original question with questions reconstructed from the answer using a large language model (LLM).
3. Context relevance assesses how relevant the retrieved context is, based on the proportion of key sentences successfully extracted from the entire available context [24].

Additionally, MRR is used to evaluate the system's effectiveness in ranking the most relevant documents or contexts at the top [25]. Latency is also measured as an indicator of efficiency, defined as the average time it takes the system to respond to a given query.

2.5.4. Experimental Procedures

The experimental procedure commenced after integrating all RAG components, including the retrieval module, generative model, and evaluation pipeline. The experiments were conducted based on 20 test questions that had been prepared in advance from academic documents of the ITK, covering guidelines for practical work, internships, final projects, and MBKM. These questions were applied consistently across all system configurations to ensure fairness in evaluation.

The first stage of the experiment focused on assessing the quality of responses using RAGAS, which consists of three main indicators: faithfulness, answer relevancy, and context relevancy. The evaluation was conducted for each variation of chunk size, overlap, and the combination of retrieval and generative models, as defined in the experimental design stage. Subsequently, the effectiveness of retrieval was evaluated using MRR. At this stage, the retrieved documents for each question were ranked, and the MRR score was calculated for each combination of retrieval and generative models. The MRR evaluation was conducted while retaining the best chunk size and overlap configurations obtained previously.

In addition to evaluating response quality, the experiments also measured system latency as an indicator of efficiency. Latency was defined as the time difference between receiving the input query and producing the output answer. The latency values were then averaged for each retrieval–generative configuration, thereby providing a comprehensive overview of the trade-off between accuracy and system speed.

2.5.5. Test Questions and Ground Truth

The test questions consist of a set of queries posed during the experiment, while the ground truth refers to the reference answers constructed from the relevant documents. The details of the test questions and their corresponding ground truth are presented in Table 1.

3. RESULTS AND DISCUSSIONS

3.1. Evaluation of RAGAS on Chunk Size, Overlap, Embedding, and LLM Variations

This study followed several stages in developing a chatbot model based on the RAG approach. In the implementation phase, all components of the RAG pipeline were integrated into a unified system. Subsequently, experiments were conducted to evaluate the model's performance across various configurations of chunk size, overlap, embedding models, and LLMs. The initial evaluation was performed by calculating the RAGAS score, which includes faithfulness, answer relevance, and context relevance for each configuration. The scores from each configuration were then averaged using the arithmetic mean, and the results were grouped based on the combination of embedding and generative models to facilitate analysis. To determine the optimal text segmentation for the RAG system, we tested a range of chunk sizes (200, 250, 300, 350, and 400 tokens) and corresponding overlap sizes (0, 30, 50, 70, and 100 tokens).

Tuning these sizes is imperative to obtain relevant context. The chunk size determines how much text will be included for indexing purposes, while the overlap size determines how much text is shared between adjacent chunks. If the chunk size is too small, it may not capture the full context. In contrast, a large size could potentially include noise that is not relevant to the query. On the other hand, if the overlap size is too small, retrieval may result in an information gap with potentially no meaningful connection. If it is too large, it will introduce redundancy to the retrieved context.

The complete details of the tested configurations are provided in Table 2. The specific chunk and overlap size combination that yielded the highest RAGAS score is visually presented in Figure 2.

Table 1. Test Question List and Ground Truth

No.	Question	Ground Truth
1	How long is the internship period?	The internship activities are carried out at various partners for a minimum of 6 (six) months and a maximum of 12 (twelve) months.
2	What is meant by an internship partner?	An internship partner is an industry, government or private institution, or a legally recognized organization that accepts students to carry out internships
3	What is an internship academic supervisor?	An internship academic supervisor is a permanent lecturer at ITK who is responsible for providing comprehensive guidance to internship participants
4	What are the criteria for a field supervisor?	a. A supervisor must have at least a Bachelor's degree (Strata 1) and/or a minimum of five (5) years of work experience. b. Has prior experience in supervision. c. Has the ability to provide technical guidance to internship participants according to their area of expertise.
5	What are the regulations that must be observed by students who will undertake an internship?	1. Must have completed at least the 5th semester with a minimum of 100 credits (SKS) earned. 2. The internship is recognized as semester credit units (SKS). 3. The number of internship credits can be equated with credits from compulsory courses, practical work, final projects, and/or elective courses. 4. The technical implementation of the internship will be regulated in an Internship Agreement between ITK, represented by the student's study program, and the Internship Partner.

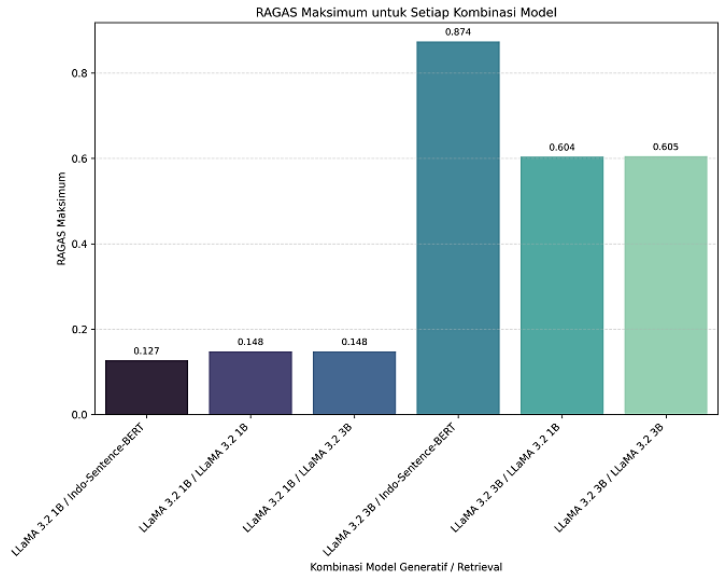

Figure 2. Arithmetic Mean of Optimal RAGAS Score for Each Model Combination

The experimental findings, as summarized in Figure 2, clearly demonstrate that the selection of the Large Language Model (LLM) and the embedder model critically influences the optimal configuration of chunk size and overlap in a RAG system. The highest performance, quantified by a RAGAS score of 0.874, was achieved using the combination of LLaMa 3.2 3B as the generative model and Indo-Sentence-BERT as the embedding model. While other configurations utilizing the LLaMa 3.2 3B LLM maintained a competitive performance level, consistently scoring around 0.6, the smaller LLaMA 3.2 1B LLM exhibited significantly lower and more stable performance, clustering around 0.1. This substantial disparity highlights the superior capability of the 3-billion parameter model in context integration for this RAG task. Consequently, based on the RAGAS metrics, the LLaMa 3.2 3B LLM paired with Indo-Sentence-BERT is recommended as the most effective configuration.

3.2. Evaluation of MRR on the Optimal Configuration of Each Model

A follow-up experiment was conducted to measure the MRR. This metric is the preferred evaluation tool when the primary objective of the RAG system is the high-precision retrieval of the single most relevant context. MRR calculates the average of the reciprocal ranks of the first relevant document across all queries, effectively assigning the highest weight to documents retrieved at the top rank (rank 1 scores 1.0), and rapidly penalizing subsequent positions. Furthermore, MRR offers superior simplicity and clarity, providing an easily interpretable measure of the system's ability to accurately rank crucial information.

For this experimentation, the configurations that previously achieved the highest RAGAS scores were utilized. The best chunk and overlap size were used in this experiment for each combination of retrieval and generative model, and the MRR results using these settings are shown in Figure 3. It clearly shows that the combination of Indo-Sentence-BERT and LLaMA 3.2 1B demonstrated the best performance with a score of approximately 0.6. This indicates that relevant documents were typically found in the first or second rank. The next best configuration was Indo-Sentence-BERT combined with LLaMA 3.2 3B, with an MRR of 0.5, suggesting that relevant documents were generally ranked second. Meanwhile, other combinations scored around 0.1, indicating a low likelihood of retrieving relevant documents using those configurations. This empirical evidence proves that the LLM model is not suitable for retrieval mechanism.

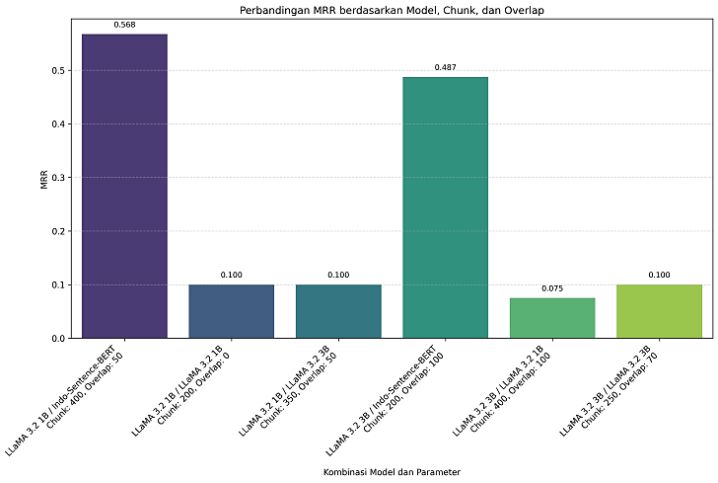

Figure 3. MRR Scores of Combinations with the Highest RAGAS

3.3. Evaluation of System Latency

The final experiment measured the latency for each RAG system configuration. This latency is quantified in seconds and represents the total time required from the submission of a query to the generation of the final answer. This latency measurement provides critical information regarding the system's readiness and operational capability, especially in resource-constrained environments.

As shown in Figure 4, all configurations demonstrated good performance, with average response times of less than one second. The highest latency was observed in the configuration that combined LLaMA 3.2 3B and Indo-Sentence-BERT, with a response time of approximately one second. In general, the LLaMA 3.2 3B model had a response time above 0.7 seconds, while LLaMA 3.2 1B was more stable with a latency of around 0.6 seconds. These findings suggest that the choice of generative model has a significant impact on the latency of the RAG system.

3.4. Recommended Optimal Model Configuration

In this analysis, we consolidate all performance metrics into a single table to recommend an optimal model combination for RAG in academic settings. These metrics comprehensively include the optimal chunk and overlap sizes, the composite RAGAS score (broken down into Faithfulness, Answer Relevancy (Answer Rel.), and Context Relevancy (Context Rel.)), MRR, and Latency.

Based on the RAGAS score evaluation across various chunk sizes and overlap levels as shown in Table 2, a range of scores was observed, reflecting differences in model performance when generating answers. The configuration with the highest RAGAS score was achieved using the LLaMA 3.2 3B model combined with Indo-Sentence-BERT, with a chunk size of 200 and an overlap of 10.

Additionally, MRR was used as a reference to assess the retrieval effectiveness. Although the combination of LLaMA 3.2 1B and Indo-Sentence-BERT achieved the highest MRR score of 0.6, this configuration was considered suboptimal due to its relatively low RAGAS score (around 0.5). Therefore, the

most balanced configuration was LLaMA 3.2 3B and Indo-Sentence-BERT, with an MRR score close to 0.5 and an average RAGAS of approximately 0.9.

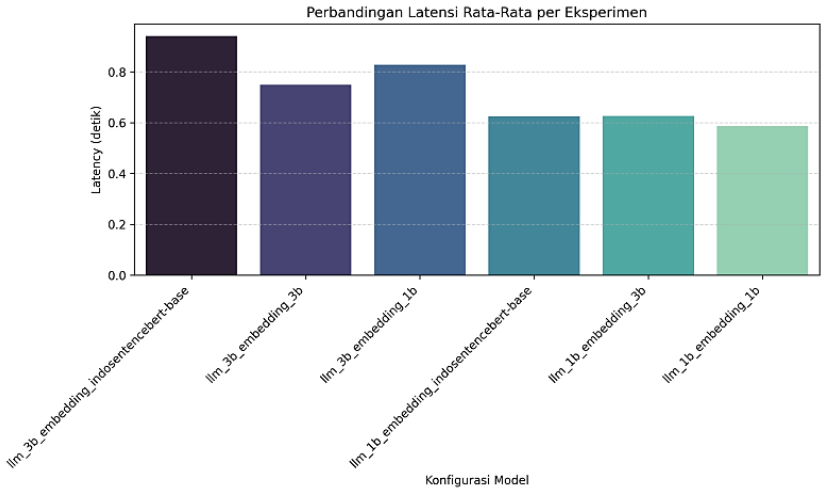

Figure 4. Latency Evaluation of Each RAG Configuration

Table 2. Overall Evaluation of Chatbot Configurations

Retrieval Model	Generative Model	Chunk Size	Overlap	Faithfulness	Answer Rel.	Context Rel.	RAGAS	MRR	Latency
Indo-Sentence-BERT	LLaMA 3.2 1B	400	50	0.21	0.17	0	0.127	0.568	0.625
LLaMA 3.2 1B	LLaMA 3.2 1B	200	0	0.13	0.32	0	0.148	0.100	0.587
LLaMA 3.2 3B	LLaMA 3.2 1B	350	50	0.20	0.24	0	0.148	0.100	0.626
Indo-Sentence-BERT	LLaMA 3.2 3B	200	100	0.81	0.90	0.91	0.874	0.487	0.828
LLaMA 3.2 1B	LLaMA 3.2 3B	400	100	0.67	0.57	0.58	0.604	0.075	0.750
LLaMA 3.2 3B	LLaMA 3.2 3B	250	70	0.67	0.58	0.56	0.605	0.100	0.940

The most significant finding is the catastrophic failure of the LLaMA 3.2 family (1B and 3B) when deployed as a zero-shot retrieval model. This is empirically validated by its substantially lower MRR of approximately 0.1 compared to the dedicated Indo-Sentence-BERT model's MRR of 0.487. This demonstrates that model size is not the primary bottleneck for generation, provided the retrieved context is relevant and complete. The 3B parameter model is sufficient for synthesizing coherent answers from context.

Although the RAGAS score shows promising results, with an Answer Relevance of approximately 0.874, the MRR remains a critical bottleneck for overall system effectiveness and requires improvement in future development. To directly address the low MRR, which indicates insufficient ranking accuracy, several advanced retrieval techniques should be investigated, including Context-Aware Retrieval, optimizing the level of granularity by employing the Parent Document Retrieval method, utilizing smaller chunks for indexing and larger complete parent chunks for the Large Language Model's final generation. Query transformation can also be further investigated to improve the retrieval systems.

5. CONCLUSION

Based on the research findings, it can be concluded that the performance of the RAG system is significantly influenced by the configuration of chunk size, overlap, and the combination of retrieval and generative models used. The best results were obtained using Indo-Sentence-BERT as the retrieval model and LLaMA 3.2 3B as the generative model, with a chunk size of 200 and an overlap of 10. This configuration yielded the highest RAGAS score of 0.9, a competitive MRR of around 0.5, and stable latency under one second. Although the combination of LLaMA 1B and Indo-Sentence-BERT recorded a higher MRR (0.6), its RAGAS score was low (around 0.1), making it not recommended. Considering answer relevance, retrieval effectiveness, and response time efficiency, the combination of Indo-Sentence-BERT and LLaMA 3.2 3B is recommended as the optimal configuration for developing an academic information chatbot at ITK.

REFERENCES

[1] Lantana DA, Ningsih S, Waluyo T, Winarsih W, Rancang Bangun Chatbot Berbasis Rule-Based Sebagai Pusat Informasi Calon Mahasiswa Baru Di Universitas Nasional. J. Sist. Inf. Bisnis JUNSIBI. 2023; 4(1): 34-32. doi: 10.55122/junsibi.v4i1.695.

[2] Ajiz MF, Ramadan MFS, Mutia HD, Yanuari PD, Pengembangan Aplikasi Chatbot Informasi Akademik Berbasis Web Menggunakan Metode Artificial Intelligence Markup Language (AIML). Media J. Inform. 2023; 15(2): 143-148. doi: 10.35194/mji.v15i2.3316.

[3] Vaswani A et al., Attention Is All You Need, Dec. 05, 2017, arXiv: arXiv:1706.03762. Accessed: Sept. 19, 2022. [Online]. Available: http://arxiv.org/abs/1706.03762

[4] Al-Amin M et al., History of generative Artificial Intelligence (AI) chatbots: past, present, and future development. Feb. 09, 2024, arXiv: arXiv:2402.05122. doi: 10.48550/arXiv.2402.05122.

[5] Annepaka Y, Pakray P, Large language models: a survey of their development, capabilities, and applications. Knowl. Inf. Syst. 2025; 67(3): 2967–3022. doi: 10.1007/s10115-024-02310-4.

[6] Minaee S et al., Large Language Models: A Survey., Feb. 20, 2024, arXiv: arXiv:2402.06196. Accessed: Oct. 30, 2024. [Online]. Available: http://arxiv.org/abs/2402.06196

[7] Ji Z et al., Survey of Hallucination in Natural Language Generation. ACM Comput Surv. 2023; 55(12):248:1-248:38. doi: 10.1145/3571730.

[8] Gao Yet al., Retrieval-Augmented Generation for Large Language Models: A Survey. 2024, arXiv: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.

[9] Lewis P 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., Dec. 2020, pp. 9459–9474.

[10] Shalhan A, Implementasi Teknik Retrieval Augmented Generation Untuk Menghilangkan Halusinasi Pada Large Language Model Berbasis Vector Database. Universitas Multimedia Nusantara. [Online]. Available: https://kc.umn.ac.id/id/eprint/34785/

[11] Neupane S et al., From Questions to Insightful Answers: Building an Informed Chatbot for University Resources. 2024, arXiv: arXiv:2405.08120. doi: 10.48550/arXiv.2405.08120.

[12] Samudra G and Zy AT, Implementasi Retrieval Augmented Generation (RAG) Dalam Perancangan Chatbot Kesehatan Pencernaan. 2025; 8(1).

[13] Li Z, Wang Z, Wang W, Hung K, Xie H, Wang FL, Retrieval-augmented generation for educational application: A systematic survey. Comput. Educ. Artif. Intell., 2025; 8:100417. doi: 10.1016/j.caeai.2025.100417.

[14] Reimers N and Gurevych I, 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), Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3980–3990. doi: 10.18653/v1/D19-1410.

[15] LLaMA: Open and Efficient Foundation Language Models | Research - AI at Meta. Accessed: Nov. 14, 2025. [Online]. Available: https://ai.meta.com/research/publications/llama-open-and-efficient-foundation-language-models/

[16] Sharma C, Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers. 2025, arXiv: arXiv:2506.00054. doi: 10.48550/arXiv.2506.00054.

[17] Karpukhin V 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, 2020; 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.

[18] Gao T, Yao X, Chen D, 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, 2021; 6894–6910. doi: 10.18653/v1/2021.emnlp-main.552.

[19] Borgeaud S, Mensch A, Hoffmann J, Cai T, Rutherford E, Millican K, Improving Language Models by Retrieving from Trillions of Tokens.

[20] Henderson M et al., Efficient Natural Language Response Suggestion for Smart Reply. 2017; arXiv: arXiv:1705.00652. doi: 10.48550/arXiv.1705.00652.

[21] Yan Y, Li R, Wang S, Zhang F, Wu W, Xu W, ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), C. Zong, F. Xia, W. Li, and R. Navigli, Eds, Online: Association for Computational Linguistics. 2021; 5065–5075. doi: 10.18653/v1/2021.acl-long.393.

[22] Deode S, Gadre J, Kajale A, Joshi A, Joshi R, L3Cube-IndicSBERT: A simple approach for learning cross-lingual sentence representations using multilingual BERT. in Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li, and J. Li, Eds, Hong Kong, China: Association for Computational Linguistics, Dec. 2023, pp. 154–163. Accessed: Nov. 14, 2025. [Online]. Available: https://aclanthology.org/2023.paclic-1.16/

[23] Shelke A, Savant R, Joshi R, Towards Building Efficient Sentence BERT Models using Layer Pruning. in Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, N. Oco, S. N. Dita,

A. M. Borlongan, and J.-B. Kim, Eds, Tokyo, Japan: Tokyo University of Foreign Studies. 2024; 720–725. Accessed: Nov. 14, 2025. [Online]. Available: https://aclanthology.org/2024.paclic-1.68/
[24] Es S, James J, Espinosa-Anke L, Schockaert S, RAGAS: Automated Evaluation of Retrieval Augmented Generation. 2023; arXiv: arXiv:2309.15217. doi: 10.48550/arXiv.2309.15217.
[25] Caragea C, Honavar V, Machine Learning in Computational Biology. in Encyclopedia of Database Systems, Springer, Boston, MA. 2009; 1663–1667. doi: 10.1007/978-0-387-39940-9_636.

BIBLIOGRAPHY OF AUTHORS

Muhamad Saman is a student of the Informatics Program at Institut Teknologi Kalimantan. His interests include Artificial Intelligence and Software Engineering, particularly in developing intelligent systems and efficient software solutions. Throughout his studies, he has actively enhanced his skills in programming, data processing, and applying technology to support innovative digital solutions.

Gusti Ahmad Fanshuri Alfarisy received his B.C.S. and M.C.S. degrees from Brawijaya University in 2014 and 2017, respectively. He completed his Ph.D. in Artificial Intelligence at the School of Digital Science, Universiti Brunei Darussalam in 2024. He serves as a Lecturer in the Department of Informatics at Institut Teknologi Kalimantan, Indonesia. His research interests encompass open-world lifelong machine learning, deep learning, web intelligence, and ecological and environmental informatics. He has also taught various subjects, including algorithms and programming languages, data structures, numerical methods, artificial intelligence, deep learning, machine learning, and software engineering.

Rizky Amelia earned her B.Sc. degree in Physics from Bogor Agricultural University (Institut Pertanian Bogor) in 2015, with a specialization in computational biophysics. She completed her M.Sc. degree in Remote Sensing Technology at the Indonesian Defense University (Universitas Pertahanan) in 2020. She currently serves as a Lecturer in the Department of Informatics at Institut Teknologi Kalimantan, Indonesia. Her research interests include computer vision, computational bioinformatics, machine learning, and data analysis. Throughout her academic career, she has been actively involved in teaching and research activities related to image processing, data-driven analysis, sensor technology, and intelligent systems.

Nisa Rizqiya Fadhliana is a Lecturer at the Department of Informatics, Institut Teknologi Kalimantan. Her primary research interests focus on Human-Computer Interaction (HCI), Game Technology, and Educational Games. In her academic role, she is actively involved in teaching and supervising students, particularly in the courses Human-Computer Interaction, Digital Game Development, and Graph Theory and Automata. She is committed to advancing research and education in the field of Informatics, with a particular emphasis on the design, development, and evaluation of interactive systems and educational technologies.