A Retrieval-Augmented Generation Framework for a Study Program Accreditation Question-Answering System

Rifki Afina Putri
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia
rifki.putri@ugm.ac.id

Helena I. Nurramdhani
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia
helenanurramdhaniirmanda@mail.ugm.ac.id

Sri Hartati
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia
shartati@ugm.ac.id (corresponding author)

Received: 17 November 2025 | Revised: 16 January 2026 | Accepted: 2 February 2026

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | *DOI: https://doi.org/10.48084/etasr.16326*

ABSTRACT

Accreditation agencies play a crucial role in ensuring the quality of higher-education institutions by evaluating programs based on established standards and guidelines. However, accreditation processes often involve complex documents that applicants may find difficult to interpret, leading to repetitive inquiries and administrative burden. This study proposes an Artificial Intelligence (AI)-based Question-Answering (QA) system built using a Retrieval-Augmented Generation (RAG) architecture to support accreditation-related information services. The system is developed specifically in the Indonesian language, enabling users to understand accreditation documents more easily by retrieving authoritative content and generating contextually appropriate explanations. As a case study, we apply this approach to LAM INFOKOM, Indonesia's Independent Accreditation Agency for Informatics and Computing. Experimental results show strong retrieval performance, with a Mean Reciprocal Rank (MRR) of 0.88, and high-quality response generation, with a BERTScore F1 of 0.76 and a manual accuracy of 83.67%. These results demonstrate the effectiveness of the Indonesian RAG-based system in improving accreditation information services with the potential to reduce manual workloads.

Keywords-accreditation; Question-Answering (QA); Retrieval-Augmented Generation (RAG); Indonesian

I. INTRODUCTION

Accreditation agencies play an essential role in maintaining and enhancing the quality of higher-education institutions worldwide [1]. By evaluating academic programs against established standards and guidelines, these agencies ensure that institutions meet the expectations of educational quality, governance, and continuous improvement. Although accreditation documents are designed to be comprehensive, they are often lengthy, technical, and difficult for applicants to interpret [2]. As a result, accreditation staff frequently receive repetitive inquiries that have already been addressed in official documents, creating administrative inefficiency and slowing down service delivery.

To address these challenges, automated Question-Answering (QA) systems and chatbots have increasingly been used in educational administration, customer support, and public service environments [3]. Prior research shows that Artificial Intelligence (AI)-driven assistants can reduce workload, provide consistent responses, and help users navigate complex information systems [4]. However, for regulatory domains such as accreditation, response accuracy is critical. Any automated system must rely strictly on authoritative documents while still providing natural, easy-to-understand answers [5]. Retrieval-Augmented Generation (RAG) architectures offer a promising solution, combining information retrieval with generative language models to produce grounded, non-hallucinatory responses [6].

While RAG has been widely explored in English-speaking contexts, its application to low-resource languages and regulatory domains remains limited [7]. In particular, accreditation-related documentation in Indonesia is written in formal, technical Indonesian, posing unique challenges for retrieval, semantic matching, and natural language generation. To the best of our knowledge, no prior work has developed an Indonesian QA system specifically designed to support accreditation processes.

This study addresses that gap by developing a RAG-based Indonesian QA system for accreditation documents. Although the proposed approach is general and applicable to any accreditation agency, we use LAM INFOKOM, Indonesia's Independent Accreditation Agency for Informatics and Computing, as the case study because it provides standardized accreditation manuals, guidelines, and procedural documents [8]. The system is designed to help accreditation applicants interpret key indicators, understand assessment components, and locate relevant regulatory references more efficiently.

Compared to existing RAG systems, this work targets a high-stakes regulatory domain where factual grounding, traceability, and linguistic precision are essential. Most prior RAG studies focus on open-domain or English-centric settings, whereas accreditation documents in Indonesia are written in formal legal language and require strict interpretation. Rather than proposing a new RAG architecture, the novelty of this work lies in adapting and empirically validating a RAG pipeline for Indonesian accreditation regulations, through careful retrieval design and human-centered evaluation. This adaptation demonstrates how RAG can be made reliable and practically usable for accreditation services in a low-resource language context. To sum up, the contributions of this research are:

- We present the Indonesian RAG-based QA system tailored for a university accreditation agency, a domain that has received little attention in existing literature.
- We introduce an end-to-end retrieval and generation pipeline capable of processing regulatory documents.
- We provide a comprehensive empirical evaluation, achieving promising retrieval performance with a Mean Reciprocal Rank (MRR) of 0.88, and high-quality answer generation with a BERTScore F1 of 0.76 and a manual accuracy of 83.67%.
- We demonstrate the practical value of the system through a prototype chatbot that can reduce repetitive inquiries and help accreditation stakeholders navigate formal documentation.

Together, these contributions highlight the potential of RAG-based systems to enhance accreditation services, particularly in low-resource languages and administrative domains.

II. RELATED WORK

Automated QA systems have gained attention in education and administrative domains as tools to improve information access and reduce repetitive inquiries [9]. Early implementations were largely rule-based or relied on direct keyword matching, which limited their ability to interpret complex or context-dependent questions [10]. Over time, QA research has shifted toward deep neural architectures, which offer better language understanding and produce more coherent answers than earlier rule-based or keyword-matching methods [11]. Recent studies have advanced chatbot design through context awareness and explainability, using dynamic knowledge graphs to deliver personalized and transparent interactions in educational and interdisciplinary domains [12]. Subsequent work in higher education applied AI-based chatbots to improve information technology governance by combining natural language processing with IT service frameworks to enhance service efficiency and decision-making [13].

More recently, the emergence of Large Language Models (LLMs) such as GPT has enabled chatbots to generate more natural and contextually rich responses, although they still face challenges in maintaining factual consistency and verifiable grounding [14]. To address these limitations, the RAG framework integrates information retrieval with generative language models, producing grounded and contextually accurate answers while reducing hallucination and improving transparency [15]. However, research on RAG-based QA systems for formal regulatory and accreditation documents, particularly in low-resource languages such as Indonesian, remains scarce and underexplored [16].

Despite its promise, existing RAG research largely concentrates on open-domain QA or English-language resources, with limited attention to domains governed by strict regulatory rules and authoritative texts. In particular, empirical studies on RAG for formal accreditation documents in low-resource languages such as Indonesian remain scarce. In such settings, precise interpretation, legal consistency, and answer traceability are mandatory rather than optional. This gap motivates the present work, which develops and systematically evaluates an Indonesian RAG-based QA system for accreditation regulations, emphasizing reliability, grounding, and practical usability in real-world administrative contexts.

III. METHODOLOGY

The development of the accreditation-focused Indonesian QA system follows the workflow illustrated in Figure 1. The process consists of three major phases: (1) data collection and preprocessing, (2) design and implementation of the QA model, and (3) evaluation and prototype development. Although the architecture is generalizable to any accreditation agency, this study applies the pipeline to LAM INFOKOM's accreditation regulations written in Indonesian as a case study.

A. Data Collection and Preprocessing

Two types of data were used in this study: reference documents and evaluation questions. The reference corpus was manually built from 42 regulatory documents issued by LAM INFOKOM published between 2022 and 2025. These documents contain official regulations, standards, indicators, and accreditation procedures. Each document was reviewed and segmented manually by two annotators to ensure that the extracted text accurately represented the authoritative content.

Irrelevant elements such as headers, footers, and duplicated formatting were removed. After segmentation, the corpus consisted of a total of 203 articles/decrees, which form the atomic textual units used in the system. The statistics are shown in Table I.

TABLE I. REGULATORY CORPUS STATISTICS

Year	Number of documents	Number of articles/decrees
2022	9	69
2023	16	99
2024	7	24
2025	10	11
Total	42	203

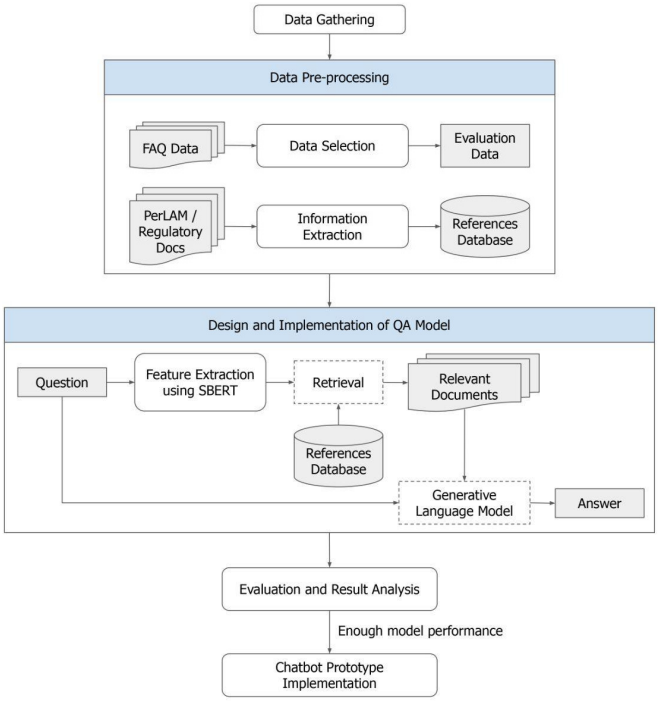

Fig. 1. Workflow of the proposed QA system.

To facilitate automated processing in the QA pipeline, each document was reorganized into a structured tabular format:

- Document number: the official unique identifier assigned to each regulation.
- Title: the name of the regulation that describes its scope and content.
- Article number and paragraph number: the position of the text within the legal hierarchy, allowing the system to preserve regulatory context.
- Decree number: the number for documents that use a decree-based structure.
- Article or decree text: the main regulatory content taken from each article, paragraph, or decree, which serves as the primary unit for sentence-level retrieval.
- Topic: a manually annotated thematic label that provides additional semantic context for retrieval.
- URL: a link to the source document in the official repository or Google Drive for verification and traceability.

To evaluate the QA system, a test dataset was created using Frequently Asked Questions (FAQ) submitted by accreditation applicants. Only questions that referred to specific clauses or topics in the Peraturan LAM (PerLAM) were selected. To increase the quantity and linguistic diversity of the evaluation data, additional questions were manually generated by varying phrasing and structure while retaining the same underlying intent. This resulted in 150 test questions, serving as the benchmark for measuring model performance.

B. Question-Answering Model Design and Implementation

The QA system follows a RAG architecture that integrates semantic retrieval and generative modeling for accreditation documents in Indonesian [17].

The retrieval module uses sentence embedding models to capture semantic similarity between user queries and regulatory text. Four SBERT-family models were evaluated to determine the most suitable option for Indonesian accreditation documents: MiniLM, E5, IndoSBERT, and BGE-M3 [18]. These models were selected because of their multilingual capabilities and their prior use in cross-lingual or low-resource retrieval tasks. Each article or decree segment in the regulatory dataset is converted into a high-dimensional embedding vector, and the user query undergoes the same embedding process. Relevance is computed using cosine similarity as shown below:

$$\text{Cosine Similarity}(q,d) = \frac{E_q \cdot E_d}{\|E_q\|\|E_d\|} \quad (1)$$

where E_q and E_d represent the embedding vectors of the query and the candidate document, respectively. Values closer to 1 indicate higher semantic similarity [19].

To study the effect of local context, two segmentation scenarios were applied: (1) a no-windowing approach in which each regulatory article or decree is treated as one text unit, and (2) a windowing approach in which longer text segments are divided into overlapping windows of up to 110 words, with an overlap of 25 words. Each window is embedded separately, and the resulting embeddings are combined using mean pooling. This technique preserves local contextual detail while allowing the embedding model to capture broader semantic meaning within longer regulatory clauses.

Once the top regulatory segments are retrieved, they are passed to the generative component. The generation stage uses the GPT-OSS model through In-Context Learning (ICL) [20]. Instead of fine-tuning, the model receives a structured prompt containing: (1) the user's question, (2) the retrieved PerLAM excerpts, and (3) explicit instructions to answer strictly based on the provided regulatory content. This structure enables the model to generate clear Indonesian explanations while maintaining fidelity to the authoritative text. By grounding the generative model in retrieved regulatory segments, the system ensures that responses remain accurate, traceable, and aligned with the formal meaning of the regulatory documents.

C. Prototype Development

A lightweight prototype was implemented using the Gradio framework to demonstrate the integration of retrieval and generation within the proposed QA system [21]. The prototype enables users to submit questions related to accreditation. The retrieval module identifies relevant text segments from PerLAM, which are the official regulations issued by the Indonesian Independent Accreditation Agency, whereas the generation module produces concise and contextually grounded answers. Retrieved text excerpts and generated responses are displayed together in the interface to ensure transparency and verifiability, allowing users to trace the origin of each answer to its regulatory source. The prototype interface is illustrated in Figure 2, which shows an example query asking, "When can a new study program apply for accreditation?" and the system's generated answer retrieved from PerLAM No. 15/PERLAM/MA/LAM-INFOKOM/XII/2023. The chatbot accurately identifies the relevant clause and explains that a new study program must apply for accreditation within two years after first admitting students. This demonstration highlights how the RAG-based QA system can support accreditation information services by improving accessibility, factual reliability, and user understanding of formal regulations.

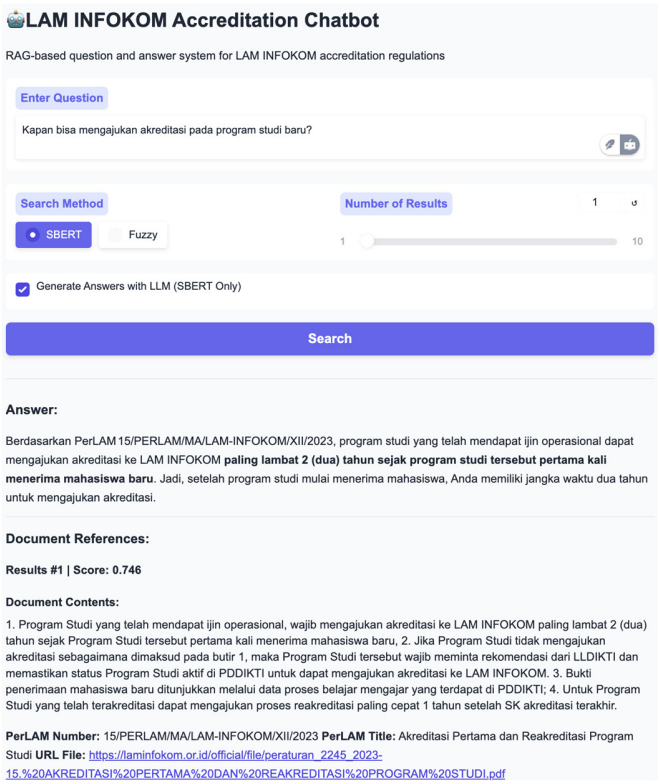

Fig. 2. User interface of the chatbot prototype.

IV. EXPERIMENTAL SETUP

In the experiment, we evaluate both the retrieval and generation components of the proposed Indonesian RAG-based QA system.

A. Baseline Retrieval Model

To provide a point of comparison for the semantic retrieval approach, a non-semantic fuzzy matching method was used as the baseline [22]. This baseline employs three RapidFuzz functions: token_set_ratio, token_sort_ratio, and partial_ratio. The token_set_ratio function measures the overlap of unique tokens between a query and a candidate text, making it robust to variations in word order. The token_sort_ratio function alphabetically sorts tokens before comparison, reducing sensitivity to differences in text arrangement. The partial_ratio function is used when a query represents only a portion of a longer regulatory statement. All three functions compute similarity based on normalized Levenshtein distance [23]. Each score is normalized to the 0–1 range, and the highest value among the three is selected as the baseline similarity score. This baseline illustrates how traditional lexical matching performs relative to embedding-based semantic retrieval.

B. Evaluation Metrics

System performance was assessed using metrics for both retrieval and answer generation. Retrieval quality was measured using MRR, which evaluates how early the correct regulatory clause appears within the ranked retrieval results [24]. For the generation component, answer quality was evaluated using BERTScore F1, which measures semantic similarity between generated responses and reference answers [25]. In addition, human evaluators assessed manual correctness, determining whether each answer was factually aligned with the retrieved PerLAM text, and assigned a naturalness score based on clarity and readability using a Likert scale. Together, these metrics provide a comprehensive view of the system's semantic accuracy, factual grounding, and linguistic quality.

V. RESULTS AND DISCUSSION

A. Retrieval Results

The retrieval experiments revealed substantial performance differences across the evaluated models, as shown in Table II.

TABLE II. RETRIEVAL PERFORMANCE

Model	Preprocessing	Windowing	Prec@3	Recall@3	MRR
Fuzzy	no	no	0.193	0.580	0.757
IndoSBERT	no	no	0.227	0.680	0.549
MiniLM	no	no	0.211	0.633	0.591
E5	no	no	0.264	0.793	0.694
BGE-M3	no	no	0.289	0.867	0.805
E5	yes	yes	0.287	0.860	0.759
BGE-M3	yes	yes	0.293	0.880	0.804
E5	yes	no	0.304	0.913	0.828
BGE-M3	yes	no	0.309	0.927	0.881

The best-performing configuration was BGE-M3 with preprocessing and no windowing (MRR 0.881), indicating that larger multilingual embedding models with stronger representational capacity are more effective for Indonesian accreditation documents. Preprocessing consistently improved retrieval accuracy for all models, suggesting that adding cleaned text, document titles, and topic annotations helps the

embeddings capture the formal, structured nature of the regulatory documents. Meanwhile, the small differences observed between windowed and non-windowed settings show that models can already encode long regulatory sentences adequately, so additional segmentation provides limited benefit in this domain. Across all configurations in Table II, similar performance trends are observed, with semantic retrieval models outperforming lexical fuzzy matching. This indicates that the reported gains are stable across preprocessing and windowing settings rather than tied to a single experimental configuration. This robustness across configurations supports the validity of the results within a highly domain-specific accreditation setting, where no publicly available Indonesian QA dataset currently exists.

In contrast, the fuzzy matching baseline achieved an MRR of 0.757, indicating a surprisingly reasonable performance given its purely lexical nature. This approach works effectively when user queries share a high degree of surface-form similarity with the regulatory text. However, its performance degrades on paraphrased or semantically implied questions, revealing the inherent limitations of lexical similarity methods in accreditation contexts, where users rarely reproduce regulatory wording verbatim. While fuzzy matching can capture approximate string overlaps, it fails to generalize beyond exact or near-exact term matches, underscoring the necessity of retrieval mechanisms that can handle semantically equivalent queries expressed in different linguistic forms. In contrast, BGE-M3 consistently retrieves the correct clauses, demonstrating stronger robustness to paraphrasing and varied user expressions, and reinforcing the importance of semantic retrieval for reliable QA performance. Some examples are shown in Table III.

B. Generation Results

To evaluate the performance of the generator component, we conduct a series of comparative experiments. First, we compare several LLMs as candidate generators under the same retrieval setting to analyze their impact on generation quality. After selecting the best-performing generator, we further evaluate the overall system performance by comparing different retrieval methods while keeping the generator fixed. Table IV compares different LLMs used as generators under the same retrieval setting (BGE-M3).

The results show that GPT-OSS (20 B) achieves the highest BERTScore F1 (0.763), outperforming larger models such as Gemma 3 and Gemma-SEA-LION-v4, both with 27 B parameters. Although the Gemma-based models obtain slightly higher recall, GPT-OSS provides a better balance between precision and recall, leading to the best overall F1 score. This result is notable given that Gemma-SEA-LION-v4 is a multilingual LLM specifically trained on Southeast Asian languages, including Indonesian. Despite this targeted multilingual training, GPT-OSS still delivers superior overall performance, suggesting that generator effectiveness in this task is influenced not only by language coverage but also by factors such as instruction tuning, response calibration, and alignment with the downstream RAG objective.

TABLE III. RETRIEVAL SAMPLES

Query	Retrieval method	Top retrieved clause	Result
Berapa biaya banding tanpa asesment lapangan? (How much is the appeal fee for appeals resolved without a site visit assessment?)	SBERT (BGE-M3)	Jika hasil pemeriksaan memutuskan untuk menerima keberatan yang diajukan program studi tanpa melalui asesmen lapangan, maka program studi wajib membayar biaya banding sebesar Rp. 10,000,000 (If the review concludes that the objection submitted by the study program is accepted without conducting a site visit assessment, the study program is required to pay an appeal fee of Rp 10,000,000)	Correct
	Fuzzy matching	Jika hasil pemeriksaan memutuskan untuk menerima keberatan yang diajukan program studi tanpa melalui asesmen lapangan, maka program studi wajib membayar biaya banding sebesar Rp. 10,000,000 (If the review concludes that the objection submitted by the study program is accepted without conducting a site visit assessment, the study program is required to pay an appeal fee of Rp 10,000,000)	Correct
Kapan program studi bisa mengajukan usulan akreditasi ulang? (When is a study program eligible to apply for re-accreditation?)	SBERT (BGE-M3)	Untuk Program Studi yang telah terakreditasi dapat mengajukan proses reakreditasi paling cepat 1 tahun setelah SK akreditasi terakhir (An accredited study program may apply for re-accreditation no earlier than one year after the issuance of its most recent accreditation decree)	Correct
	Fuzzy matching	Pemimpin PTNBH mengajukan permohonan rekomendasi program studi yang akan dibuka kepada LAM INFOKOM dengan mengirimkan surat permohonan… (The head of the PTNBH submits a request for a recommendation for the new study program to be established by sending a formal request letter…)	Incorrect

TABLE IV. COMPARISON OF MODELS USED AS GENERATORS

Model	Parameter size	BERTScore		
		Prec	Recall	F1
LLaMA 3.1	8 B	0.717	0.797	0.753
Gemma 3	27 B	0.718	0.803	0.756
Gemma-SEA-LION-v4	27 B	0.711	0.804	0.752
GPT-OSS	20 B	0.729	0.805	0.763

After selecting GPT-OSS as the generator, we further evaluated different retrieval methods while keeping the generator fixed, as reported in Table V. Among all approaches, BGE-M3 achieved the best performance, obtaining the highest BERTScore F1 of 0.763, followed closely by E5, which is consistent with the retrieval performance observed in the previous section. While fuzzy retrieval shows relatively competitive precision, its lower recall suggests limitations in handling semantic variation and paraphrased queries. IndoSBERT yields the weakest overall performance, indicating

that older or less instruction-aligned embedding models are less effective for retrieval in this setting.

TABLE V. COMPARISON OF RETRIEVAL METHODS ON FINAL ANSWER QUALITY

Retrieval method	BERTScore		
	Prec	Recall	F1
Fuzzy	0.718	0.786	0.748
IndoSBERT	0.680	0.684	0.681
MiniLM	0.717	0.779	0.745
BGE-M3	0.729	0.805	0.763

Since BERTScore primarily measures surface-level semantic similarity, it does not fully capture factual correctness, reasoning quality, or readability of generated responses. This discrepancy arises because regulatory explanations can be semantically correct while differing stylistically from reference answers. To address this, a detailed manual evaluation was conducted on 50 randomly selected test samples. The result confirms that the model was able to translate formal regulatory language into clear and accurate Indonesian explanations. In total, 83.67% of generated answers were judged to be factually correct and aligned with the authoritative PerLAM text, and the responses were highly readable, reflected in an average naturalness rating of 4.3 out of 5. These human-evaluated results provide a more realistic reflection of the system's effectiveness and clarify that the model's lower BERTScore does not indicate weak reasoning or factual grounding, but rather stylistic variation between generated and reference responses.

Overall, the findings reinforce that generation quality is strongly dependent on retrieval performance. When the correct clause is supplied as context, the model consistently produces precise, coherent, and faithful answers. Retrieval errors, however, propagate directly into the generated output, a characteristic typical of RAG architectures. The results demonstrate that ICL is sufficient for producing grounded Indonesian-language answers in a low-resource regulatory domain, provided that retrieval supplies accurate and relevant evidence.

VI. CONCLUSION

This study presents a Retrieval-Augmented Generation (RAG) system designed to support accreditation-related Question-Answering (QA) using Indonesian regulatory documents from LAM INFOKOM, the Independent Accreditation Agency for Informatics and Computing. Importantly, the results demonstrate that a carefully adapted RAG pipeline can be applied effectively to a high-stakes accreditation setting in a low-resource language, provided that retrieval is strong and responses are strictly grounded in authoritative accreditation documents. Our experiments show that high-quality semantic retrieval is crucial for handling the linguistic mismatch between user queries and formal accreditation guidelines. Models such as BGE-M3 and E5, combined with targeted preprocessing, consistently achieved strong performance and enabled the generation component to produce accurate and readable responses using In-Context Learning (ICL). The system shows clear potential to reduce repetitive inquiries, improve access to accreditation rules, and support higher-education institutions in better understanding compliance requirements.

Despite these promising results, several limitations remain. First, the system relies on in-context grounding rather than domain-specific fine-tuning, which is a deliberate design choice due to the available computing resources. This approach guarantees factual grounding but also means the model does not attempt to infer answers beyond what is explicitly present in the retrieved text, which is an appropriate constraint for high-stakes regulatory domains. Second, the system is optimized for the Indonesian language and the LAM INFOKOM accreditation framework; this specialization strengthens its reliability for the target domain, though extending it to other accreditation agencies would require additional domain adaptation.

Future work may explore several directions to address these limitations. Expanding the dataset to include more complete and fine-grained regulatory documents could further improve retrieval precision. Domain-adaptive fine-tuning or instruction tuning of open-source language models may also enhance generation robustness, particularly in edge cases where regulatory text is ambiguous or multi-layered. Integrating conversational memory or follow-up question handling could improve user experience by supporting more interactive information-seeking behavior. Finally, applying the methodology to other accreditation bodies or higher-education regulatory systems would help validate the scalability and adaptability of the proposed approach.

ACKNOWLEDGMENT

The authors would like to express their deepest gratitude to the Faculty of Mathematics and Natural Sciences, Gadjah Mada University, for providing the necessary resources and publication support under the Collaborative Research Funding Year 2025.

REFERENCES

[1] P. Kumar, B. Shukla, and D. Passey, "Impact of Accreditation on Quality and Excellence of Higher Education Institutions," *Revista Investigacion Operacional*, vol. 41, no. 2, pp. 151–167, Feb. 2020.

[2] N. Sousa *et al.*, "A hitchhikers' guide to the terminology of accreditation processes for health professionals and institutions," *MedEdPublish*, vol. 13, Feb. 2023, Art. no. 11, https://doi.org/10.12688/mep.19566.1.

[3] P. Jadhav and S. M. Jadhav, "Survey on Question Answering System with Emphasis on Method and Evaluation Metrics," in *2025 5th International Conference on Soft Computing for Security Applications*, Salem, India, 2025, pp. 2135–2140, https://doi.org/10.1109/ICSCSA66339.2025.11170976.

[4] R. Khan, S. P. Khan, and S. A. Ali, "Conversational AI," in *Conversational Artificial Intelligence*, R. Rawat, R. K. Chakrawarti, S. K. Sarangi, P. Vyas, M. S. Alamanda, K. Srividya, and K. S. Sankaran, Eds. Hoboken, NJ, USA: John Wiley & Sons, Ltd, 2024, pp. 249–268, https://doi.org/10.1002/9781394200801.ch16.

[5] D. Collarana *et al.*, "A Question Answering System on Regulatory Documents," in *31st International Conference on Legal Knowledge and Information Systems*, Groningen, The Netherlands, 2018, pp. 41–50, https://doi.org/10.3233/978-1-61499-935-5-41.

[6] Z. Jiang *et al.*, "Active Retrieval Augmented Generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2023, pp. 7969–7992, https://doi.org/10.18653/v1/2023.emnlp-main.495.

[7] N. Chirkova, D. Rau, H. Déjean, T. Formal, S. Clinchant, and V. Nikoulina, "Retrieval-augmented generation in multilingual settings," in *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, Bangkok, Thailand, 2024, pp. 177–188, https://doi.org/10.18653/v1/2024.knowllm-1.15.

[8] "Lembaga Akreditasi Mandiri Informatika dan Komputer." LAM INFOKOM. https://laminfokom.or.id/official/.

[9] R. Kumar, G. B. Singh, P. Shailendra, and N. Sharma, "A Brief Review on Question Answering System: Techniques, Challenges and Future Directions," in *2024 10th International Conference on Advanced Computing and Communication Systems*, Coimbatore, India, 2024, pp. 407–414, https://doi.org/10.1109/ICACCS60874.2024.10717204.

[10] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning with Applications*, vol. 2, Dec. 2020, Art. no. 100006, https://doi.org/10.1016/j.mlwa.2020.100006.

[11] Z. Abbasiantaeb and S. Momtazi, "Text-based question answering from information retrieval and deep neural network perspectives: A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 6, Nov. 2021, Art. no. e1412, https://doi.org/10.1002/widm.1412.

[12] A. Raza, M. Latif, M. U. Farooq, M. A. Baig, M. A. Akhtar, and Waseemullah, "Enabling Context-based AI in Chatbots for conveying Personalized Interdisciplinary Knowledge to Users," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12231–12236, Dec. 2023, https://doi.org/10.48084/etasr.6313.

[13] S. Ahriz, H. Gharbaoui, N. Benmoussa, A. Chahid, and K. Mansouri, "Enhancing Information Technology Governance in Universities: A Smart Chatbot System based on Information Technology Infrastructure Library," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 17876–17882, Dec. 2024, https://doi.org/10.48084/etasr.8878.

[14] A. T. Neumann, Y. Yin, S. Sowe, S. Decker, and M. Jarke, "An LLM-Driven Chatbot in Higher Education for Databases and Information Systems," *IEEE Transactions on Education*, vol. 68, no. 1, pp. 103–116, Feb. 2025, https://doi.org/10.1109/TE.2024.3467912.

[15] A. J. Oche, A. G. Folashade, T. Ghosal, and A. Biswas, "A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions." arXiv, July 25, 2025, https://doi.org/10.48550/arXiv.2507.18910.

[16] J. Malhotra, S. Sharma, and Er. A. S. Rana, "Improving Retrieval-Augmented Generation for Low-Resource Languages," in *2025 International Conference on Modern Sustainable Systems*, Shah Alam, Malaysia, 2025, pp. 1164–1169, https://doi.org/10.1109/CMSS66566.2025.11182476.

[17] K. Muludi, K. M. Fitria, J. Triloka, and Sutedi, "Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, pp. 776–785, Mar. 2024, https://doi.org/10.14569/IJACSA.2024.0150379.

[18] 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*, Hong Kong, China, 2019, pp. 3982–3992, https://doi.org/10.18653/v1/D19-1410.

[19] 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*, Online, 2020, pp. 6769–6781, https://doi.org/10.18653/v1/2020.emnlp-main.550.

[20] S. Agarwal *et al.*, "gpt-oss-120b & gpt-oss-20b Model Card." arXiv, Aug. 08, 2025, https://doi.org/10.48550/arXiv.2508.10925.

[21] R. Ferreira, M. Canesche, P. Jamieson, O. P. V. Neto, and J. A. M. Nacif, "Examples and tutorials on using Google Colab and Gradio to create online interactive student-learning modules," *Computer Applications in Engineering Education*, vol. 32, no. 4, July 2024, Art. no. e22729, https://doi.org/10.1002/cae.22729.

[22] Y. Andriyani *et al.*, "Improving University Community Service Communication with Kukerti's Fuzzy String Matching Chatbot," in *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)*, Jakarta Selatan, Indonesia, 2023, pp. 398–403, https://doi.org/10.1109/ICIMCIS60089.2023.10348968.

[23] P. Janardhana Rao, K. Nageswara Rao, S. Gokuruboyina, and K. N. Neeraja, "An Efficient Methodology for Identifying the Similarity Between Languages with Levenshtein Distance," in *Proceedings of the 6th International Conference on Communications and Cyber Physical Engineering*, Hyderabad, India, 2023, pp. 161–174, https://doi.org/10.1007/978-981-99-7137-4_15.

[24] S. Reddy, "Evaluating Retrieval-Based Chatbot Systems: A Focus on Speed and Performance," in *2025 International Conference on Computing and Communication Technologies*, Chennai, India, 2025, pp. 1–6, https://doi.org/10.1109/ICCCT63501.2025.11019950.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT." arXiv, Feb. 24, 2020, https://doi.org/10.48550/arXiv.1904.09675.