VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i2.5916

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

UTILIZING RETRIEVAL-AUGMENTED GENERATION IN LARGE LANGUAGE MODELS TO ENHANCE INDONESIAN LANGUAGE NLP

Herdian Tohir[1]; Nita Merlina[2*]; Muhammad Haris[3]

Computer Science[1]
Informatics[2,3]
Universitas Nusa Mandiri, Jakarta, Indonesia[1,2,3]
www.nusamandiri.ac.id[1,2,3]
htohir.ht@gmail.com[1], nita@nusamandiri.ac.id[2*], muhammad.uhs@nusamandiri.ac.id[3]

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *The improvement of Large Language Models (LLM) such as ChatGPT through Retrieval-Augmented Generation (RAG) techniques has urgency in the development of natural language translation technology and dialogue systems. LLMs often experience obstacles in addressing special requests that require information outside the training data. This study aims to discuss the use of Retrieval-Augmented Generation (RAG) on large-scale language models to improve the performance of Natural Language Processing (NLP) in Indonesian, which has so far been poorly supported by high-quality data and to overcome the limitations of traditional language models in understanding the context of Indonesian better. The method used is a combination of retrieval capabilities (external information search) with generation (text generation), where the model utilizes broader and more structured basic data through the retrieval process to produce more accurate and relevant text. The data used includes the Indonesian corpus of the 30 Juz Quran translation into Indonesian. The results of the trial show that the RAG approach significantly improves the performance of the model in various NLP tasks, including token usage optimization, text classification, and context understanding, by increasing the accuracy and relevance of the results.*

Keywords: *GPT, indonesian language, LLM performance, performance evaluation, RAG technique.*

Abstrak— *Peningkatan Large Language Model (LLM) seperti ChatGPT melalui teknik Retrieval-Augmented Generation (RAG) memiliki urgensi dalam pengembangan teknologi pemrosesan bahasa alami dan sistem dialog. LLM sering mengalami kendala dalam mengatasi permintaan khusus yang memerlukan informasi di luar data latihan. Penelitian ini bertujuan untuk membahas pemanfaatan Retrieval-Augmented Generation (RAG) pada model bahasa skala besar untuk meningkatkan kinerja Natural Language Processing (NLP) pada Bahasa Indonesia, yang selama ini masih kurang didukung oleh data berkualitas tinggi dan untuk mengatasi keterbatasan model bahasa tradisional dalam memahami konteks bahasa Indonesia dengan lebih baik. Metode yang digunakan adalah penggabungan kemampuan retrieval (pencarian informasi eksternal) dengan generation (pembangkitan teks), di mana model memanfaatkan basis data yang lebih luas dan terstruktur melalui proses retrieval untuk menghasilkan teks yang lebih akurat dan relevan. Data yang digunakan mencakup korpus Bahasa Indonesia dari terjemahan Quran 30 Juz dalam Bahasa indonesia. Hasil uji coba menunjukkan bahwa pendekatan RAG secara signifikan meningkatkan performa model dalam berbagai tugas NLP, termasuk optimasi penggunaan token, klasifikasi teks, dan pemahaman konteks, dengan meningkatkan akurasi dan relevansi hasil.*

Kata Kunci: *GPT, bahasa Indonesia, kinerja LLM, evaluasi kinerja, teknik RAG.*

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5916

INTRODUCTION

Large Language Models (LLMs), represent a significant leap forward in the field of natural language processing (NLP) and dialogue systems. Despite their impressive capabilities, these models often encounter difficulties when responding to specific requests that necessitate information beyond their training data, especially in the Indonesian language. In recent years, there has been an explosion of various LLMs, including the GPT series from OpenAI, such as GPT-4 [1][2]. and open-source models like Llama-3 from Meta.

LLMs are built on the transformer architecture [3]. with larger models containing hundreds of billions of parameters. They are trained on extensive training datasets, including books, crawled web pages, and social media conversations [4]. Their language capabilities make LLMs suitable for derivative applications such as question answering [5][6]. However, LLMs face limitations in handling queries that are domain-specific or highly specialized, requiring information beyond their training corpus [7][8]. LLMs can be pre-trained for specific domains such as finance [9] or geographic language for mapping applications [10], but this requires large training datasets and expensive computational resources. This is especially challenging for Indonesian, where resources are still very limited. Various approaches have been developed to build domain-specific applications with LLMs, which we review here, focusing primarily on the Indonesian language domain [11][12][13][14].

The research problem in this context is how to improve the performance of LLMs, particularly in the Indonesian language. One popular way to build LLM applications without requiring specialized training is through the Retrieval Augmented Generation (RAG) method [15][16][17][18][19]. When faced with domain-specific questions beyond its training data, an LLM can generate inaccurate information or even hallucinations, especially in Indonesian. RAG addresses this issue by retrieving information from external data sources in Indonesian, which is then provided as additional context to the LLM to generate a response [20]. This helps improve factual accuracy and relevance by giving the model access to additional information sources in Indonesian. Although RAG can be used during the pre-training stage [21], it is more commonly used during inference due to its practicality [22].

Retrieval-Augmented Generation (RAG) enhances the performance of LLMs on domain-specific tasks, particularly in Indonesian, by providing the model with external information.

While there are variations, we will provide an overview of RAG applications through an algorithm.

The urgency of this research lies in the need to develop language models, specifically for the Indonesian language, that can understand context more deeply and generate more appropriate content. The application of LLMs has been explored in various domains such as education for generating exam questions [23], recruitment and job recommendations [24], news recommendations, various healthcare applications [25], answering medical questions [26], searching patient health records [27], tools for mental health [28], answering legal questions [29], and IT support systems.

The utilization of RAG to enhance NLP in the Indonesian language is a highly challenging area of research to explore. Many methods build upon existing techniques. Some studies that have been conducted in the development of RAG are presented in Table 1.

Table 1. State-of-the-art

Title	Author	Year	Methods	Result
Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks	Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al	2020	RAG, fine tuning	The RAG models set new state-of-the-art results on three open-domain QA tasks [20]
Retrieval Augmented Language Model Pre-Training	Guu K, Lee K, Tung Z, Pasupat P, Chang M.	2020	REALM	Significant advancement in how language models can be trained and fine-tuned for tasks [21]
A medical question answering system using Large Language Model s and knowledge graphs	Guo Q, Cao S, Yi Z	2022	Elasticsearch, semantic matching, siamese	Robust approach to medical question answering [26]
Evaluation of AI Chatbots for Patient-Specific HER Questions	Hamidi A, Roberts K	2023	Model used, evaluation criteria, specific question	application of AI in healthcare, specifically in using LLMs for patient-specific QA [27]
Language Model Behavior: A Comprehensive Survey	Chang TA, Bergen BK	2023	Survey	understanding of transformer language models[12]

Can Large Language Models Transform Computational Social Science?	Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D	2024	Prompting and tasking	Integrating LLM into CSS pipeline, as valuable tools that can assist with specific tasks[8]
A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM	Yusuf AA, Karaarslan E, Aydin O	2024	RAG, prompt engineering	RAG-based approach to develop more accurate and respectful LLM-driven question-answering systems in the context of religious education[5]

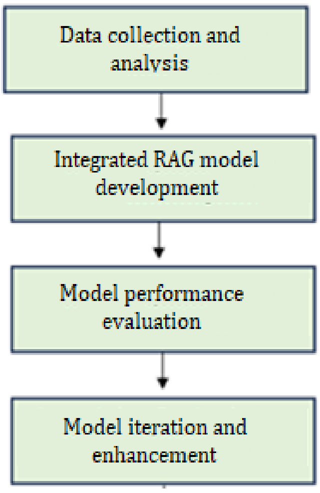
Source: (Research Results, 2024)

MATERIALS AND METHODS

Research on enhancing Large Language Models (LLM) like ChatGPT through Retrieval-Augmented Generation (RAG) techniques is crucial for the development of natural language processing (NLP) technology and dialogue systems. LLMs often face challenges in handling specific requests that require information beyond the training data.
Based on previous research, the opportunity in this study is to develop an LLM model capable of improving the contextual understanding of the Indonesian language in the translation of the Quran's 30 Juz, thereby producing more relevant and informative responses using the Retrieval Augmented Generation (RAG) technique.

This research aims to enhance LLM (Large Language Models) using RAG (Retrieval-Augmented Generation) techniques to develop a better language model in understanding context and generating relevant content. The research methods include data collection, the creation of a RAG model that integrates retrieval and generation elements, and performance evaluation using metrics such as fluency, factual accuracy, and response diversity.

The approach to be used in this study is experimental, with the research flowchart as shown in Figure 1.



Source: (Research Results, 2024)
Figure 1. Research flow diagram

The stages in Figure 1 represent the research procedure, which includes the following steps that can be explained as follows:

1. Data Collection and Analysis:

Relevant data sources will be identified, including large text corpora and structured datasets that align with the context of the research, in this case, in the Indonesian language. Data will be collected from various sources, followed by an initial processing step to clean and format the data. This step is an integral part of data collection and analysis, ensuring that the data used is of high quality and meets the research requirements. This will enable accurate and relevant insights to be obtained from the subsequent analysis.

2. Development of the Integrated RAG Model:

The first step is the development of a foundational Large Language Model (LLM) that will serve as the basis for integration with the Retrieval-Augmented Generation (RAG) technique. An RAG model will be designed and trained, considering the appropriate architecture to combine information retrieval elements with text generation. This process will involve experimenting with various model configurations to identify the most effective one.

In this process, emphasis is placed on seamless integration between information retrieval and text generation capabilities to produce a model that can generate high-quality text that is relevant to the context. Therefore, the development of the integrated RAG model will provide an optimal solution for addressing complex problems in its field. The stages of developing the integrated RAG model can be seen in Figure 2. The following are the steps in the RAG process:

A. Dataset Preparation (Data Sources):

Start with a dataset containing the necessary text data.

B. Chunking the Text:

The dataset is divided into manageable chunks of text. This step ensures that the text is broken down into coherent, small parts suitable for processing.

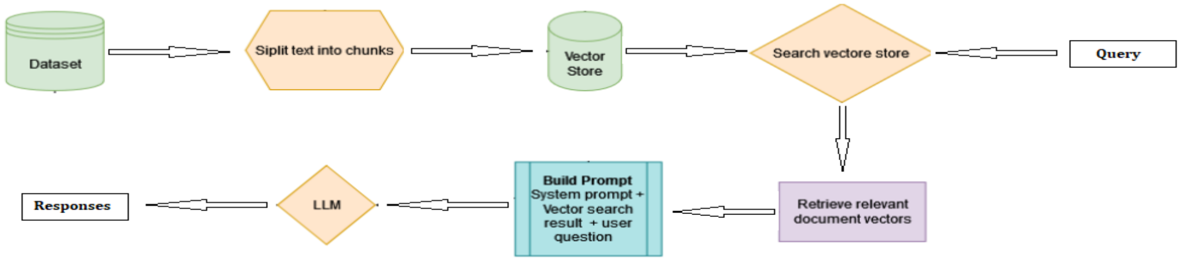
C. Store Text Chunks in a Vector Store:

The text chunks are then stored in a vector store, a specialized database that indexes and stores vectors (numerical representations) of these text chunks.

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5916

D. Search in the Vector Store:

When a request (query) is made, the vector store is searched to find relevant text chunks. This involves comparing the query vector with the vectors stored in the database.



Source: (Research Results, 2024)

Figure 2. RAG flow process diagram

E. Retrieve Relevant Document Vectors:

The relevant document vectors are retrieved based on the search results from the vector store. These vectors represent the text chunks most relevant to the query.

F. Build the Prompt:

A prompt is constructed by combining the system prompt, search results from the vector store, and the user's question. This combined prompt is designed to provide context and guidance for the model to generate a response.

G. LLM (Large Language Model):

The constructed prompt is then passed to the Large Language Model (LLM), which processes the input to generate a coherent and contextually relevant response.

H. Generate Response:

Finally, the LLM generates a response based on the prompt, which is then returned or presented to the user. Each of these steps is interconnected, forming a pipeline that processes the user's query and returns a relevant and accurate response based on the underlying dataset or data sources.

3. Model Performance Evaluation:

The performance of the developed model will be evaluated using various metrics, such as contextual relevance, factual accuracy, and response diversity. Testing will be conducted using a validation dataset and standard benchmarks to validate the model's effectiveness. This evaluation process is crucial to ensure that the model can generate text that is not only fluent and factually accurate but also diverse and contextually appropriate. By using a validation dataset and standard benchmarks, an objective comparison of the model's performance can be made. This comprehensive performance evaluation is an essential step in determining whether the model can meet the research objectives and its intended use in practice.

4. Model Iteration and Improvement:

Based on the evaluation results, the model will be refined and enhanced through repeated iterations. Optimization will be carried out to improve the quality of the model's responses and enhance its overall performance. This iterative process involves optimization to improve the quality of the model's responses as well as its overall performance. Each iteration allows for adjustments and improvements to the model based on the findings from previous evaluations, focusing on enhancing the model's ability to generate high-quality and relevant responses. The goal of this optimization is to ensure that the model meets or even exceeds the established standards in terms of quality and performance. Therefore, the process of iterating and improving the model is key to ensuring that the solution delivered can provide optimal added value.

RESULTS AND DISCUSSION

The results of this research are as follows:

1. Source document tokens and RAG document tokens

This section illustrates the size and complexity of the documents being processed. The source documents from the Quran can be seen in Table 2.

Table 2. Source documents from the Quran, Juz 29

No	Documents source	Tokens
141...145	Surah-Al-Mulk/1: Glory be to Allah who has dominion over all kingdoms, and He is almighty over all things. over all things.	16609

No	Documents source	Tokens
	Surah-Al-Mulk/2: Who created death and life, to test you as to which of you is better in deeds. who among you is better in deeds. And He is the Mighty, the Forgiving. Surah-Al-Mulk/3: Who created the seven heavens in layers. You will not see anything that is unequal in the creation of the Most Merciful. So look once more, do you see any see anything that is defective? Surah-Al-Mulk/4: Then repeat (your gaze) once more (and) once more, surely your gaze will return to you without finding any defect, and it will be in a tired state. a state of weariness. Surah-Al-Mulk/5: And indeed, We have adorned the near sky with stars and We have made them the instruments of the devils, and We have prepared for them the punishment of a fiery hell. them the torment of a blazing hell. ... Surah-Al-Mursalat/46: (Say to the disbelievers), 'Eat and be merry for a little while. (in the world) for a little while, surely you are the wrongdoers!' Surah-Al-Mursalat/47: Woe on that Day to those who deny the truth. Surah-Al-Mursalat/48: And when it is said to them, 'Bow,' they will not bow. bow. Surah-Al-Mursalat/49: Woe on that Day to those who deny the truth! Surah-Al-Mursalat/50: So in which of these teachings (other than the Qur'an) will they believe?	

Source : (Research Result, 2024)

Source Document Tokens: Ranging from 7,749 to 16,205 tokens, as shown in Table 3.

Table 3. RAG Documents with relevant context

No	Query	RAG Documents	Tokens
141	What is the essence of Surah-Al-Mulk	are they burdened with debt? Surah-Al-Qalam/47: Or do they know the unseen and write it down? Surah-Al-Mulk/1: Glory be to Allah, who has dominion over all kingdoms, and He is almighty over all things. Surah-Al-Mulk/2: Who created death and life, to test you as to which of you is better in deeds. And He is the Mighty, the Forgiving. Surah-Al-Mulk/3: Who created the seven heavens in layers. You will not see anything unequal in the creation of the Most Merciful. So look again, do you see any defect? Surah-Al-Mulk/4: Then repeat (your gaze) once more (and) once more, surely your gaze will return to you without finding any defect and it (your gaze) will be weary. Surah-Al-Mulk/5: And indeed, We have adorned the near sky with stars, and We have made it (the Qur'ān) no other than the Qur'ān. (The Qur'an) is nothing but a warning for all the worlds. ... Surah-Al-Muzzammil/1: O one who is covered (Muhammad)! Surah-Al-Muzzammil/2: Rise (for prayer) in the night, except for a small part, Surah-Al-Muzzammil/3: (i.e.) half of it or a little less than that, Surah-Al-Muzzammil/4: or more than that, and recite the Qur'an slowly. Surah-Al-Muzzammil/5: Surely We will send down a heavy word upon you. Surah-Al-Muzzammil/6: Indeed, the rising of the night is more powerful (filling the soul); and (the recitation at that time) is more memorable. Surah-Al-Muzzammil/7:	1010

Source : (Research Result, 2024)

In some queries, not all contexts can be effectively responded to by the RAG documents, as shown in Table 4.

Table 4. RAG Documents with context not yet aligned.

No	Query	RAG Documents	Tokens
144	How long do angels and Gabriel ascend to God	are they burdened with debt? Surah-Al-Qalam/47: Or do they know the unseen, so they write it down? Reply (guardian angels). Surah-Al-Muddassir/31: And We made the guardians of Hell only of angels; and We appointed their number only as a trial for the disbelievers, that those who were given the Book might be convinced, that the believers might increase in faith, that those who were given the Book and the believers might not doubt; and that those in whose hearts there is a sickness and the disbelievers might say, 'What does Allah intend by this (number) as a parable?' Thus Allah misguides whom He wills and guides whom He wills. And no one knows the host of your Lord except Himself. And the Saqar is none other than ...	832

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5916

No	Query	RAG Documents	Tokens
		Surah-Al-Mursalat/41: Indeed, the pious are in the shade (of the shady trees of Paradise) and (around) springs of water, Surah-Al-Mursalat/42: and the fruits they love. Surah-Al-Mursalat/43: (Say to them), 'Eat and drink. Have We destroyed the former people? Surah-Al-Mursalat/17: Then We hasten (Our punishment on) those who come later. Surah-Al-Mursalat/18: Thus do We treat those who sin. Surah-Al-Mursalat/19: Woe on that Day, for those who deny (the truth).	

Source : (Research Result, 2024)

RAG Document Tokens: Ranging from 748 to 1,355 tokens.

2. Computation Time for Storing to Database(s)

The time required to store documents in the database is consistent across various queries, indicating stable storage performance. Average Storage Time: 96.72 ms (calculated from all entries).

3. **Computation Time for Retrieval from Database(s)**

The retrieval time from the database is generally low, ensuring fast access to stored information. Average Retrieval Time: 0.24 ms (calculated from all entries).

4. Contextual Relevance

Contextual relevance is marked with a score of 1 (relevant) or 0 (not relevant) for each query. Relevance Score: 141 out of 150 queries have a relevance score of 1, indicating high contextual relevance.

5. Prompt Template for Requesting RAG Results from LLM (GPT 3.5/GPT 4)

DOCUMENT: [RAG Document]
Based on the DOCUMENT above, answer the QUESTION: [Query]

6. Responses from LLM (GPT 3.5/4o)

For queries where the context can be effectively responded to by the RAG documents, as shown in Table 5.

Table 5. LLM Responses with aligned context

No	Query	RAG Documents	Response
141	What is the essence of Surah -Al-Mulk	are they burdened with debt? Surah-Al-Qalam/47: Or do they know the unseen and write it down? Surah-Al-Mulk/1: Glory be to Allah, who has dominion over all kingdoms, and He is almighty over all things. Surah-Al-Mulk/2: Who created death and life, to test you as to which of you is better in deeds. And He is the Mighty, the Forgiving. Surah-Al-Mulk/3: Who created the seven heavens in layers. You will not see anything unequal in the creation of the Most Merciful. So look again, do you see any defect? Surah-Al-Mulk/4: Then repeat (your gaze) once more (and) once more, surely your gaze will return to you without finding any defect and it (your gaze) will be weary. Surah-Al-Mulk/5: And indeed, We have adorned the near sky with stars, and We have made it (the Qur'ān) no other than the Qur'ān. (The Qur'an) is nothing but a warning for all the worlds. ... Surah-Al-Muzzammil/1: O one who is covered (Muhammad)! Surah-Al-Muzzammil/2: Rise (for prayer) in the night, except for a small part, Surah-Al-Muzzammil/3: (i.e.) half of it or a little less than that, Surah-Al-Muzzammil/4: or more than that, and recite the Qur'an slowly.	The essence of Surah Al-Mulk is to emphasise the power and majesty of Allah and remind people of the signs of His power in the universe and the importance of faith and devotion to Him. Here are some of the main points of Surah Al-Mulk: 1. **The Power of Allah (Verses 1-2)**: The surah begins by praising Allah, who is in control of everything, who created death and life to test which of mankind has the best deeds. 2. **The Perfection of Allah's Creation (Verses 3-4)**: Allah created the seven heavens in layers and there is no imbalance in His creation. Humans are invited to look and observe the creation, and they will find no flaw. 3. **Decoration of the Heavens (Verse 5)**: Allah adorns the near sky with stars. 4. **Threats and Warnings (Verses 17-19)**: Allah warns people of the punishments He can inflict, such as rocky storms. Also, people are reminded to pay attention to the birds that fly, which confirms Allah's power in taking care

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v10i2.5916

No	Query	RAG Documents	Response
		Surah-Al-Muzzammil/5: Surely We will send down a heavy word upon you. Surah-Al-Muzzammil/6: Indeed, the rising of the night is more powerful (filling the soul); and (the recitation at that time) is more memorable. Surah-Al-Muzzammil/7:	of His creatures. Overall, Surah Al-Mulk emphasises the greatness and power of Allah, invites people to contemplate His creation, and warns of the consequences of disbelieving Allah's warnings.

Source : (Research Result, 2024)

There are also instances of response discrepancies in the LLM, which can be seen in Table 6.

Table 6. LLM Responses with misaligned responses

No	Query	RAG Documents	Responses
144	How long do angels and Gabriel ascend to God	are they burdened with debt? Surah-Al-Qalam/47: Or do they know the unseen, so they write it down? twelve (guardian angels). Surah- ... Surah-Al-Mursalat/42: And the fruits that they love. Surah-Al-Mursalat/43: (Say to them), 'Eat and drink. have We destroyed the former people? Surah-Al-Mursalat/17: Then We hasten (Our punishment on) those who come later. Surah-Al-Mursalat/18: Thus do We treat those who sin. Surah-Al-Mursalat/19: Woe on that Day, for those who deny (the truth).	Based on the DOCUMENTS provided, there is no direct information that answers how long the angels and Jibril ascended to God. However, there is an explanation of angels in Surah-Al-Muddassir/31 that mentions angels as the guardians of hell and as a test for the disbelievers. However, this is not directly related to the duration of time that angels and Gabriel ascend to God. To answer this question properly, reference is usually made to Surah Al-Ma'arij (70:4) which states that angels and Gabriel ascended to God in fifty thousand years. Here is the verse: *"The angels and Gabriel ascend (to God) in a day of fifty thousand years. '* (Surah Al-Ma'arij, Verse 4) So, the duration of time stated in the verse is fifty thousand years.

Source : (Research Result, 2024)

This dataset includes a range of document lengths, which may affect processing time and complexity. RAG (Retrieval-Augmented Generation) tokens are significantly fewer than source tokens, which is expected since RAG models typically compress information to generate responses.

Consistent document storage times indicate reliable database performance. However, there is some variation, with the maximum storage time being 123.66 ms and the minimum being 83.55 ms. This variation could be due to the size and complexity of the documents, as well as the database load during storage.

Data retrieval times are very low, indicating that the database's indexing and retrieval mechanisms are highly efficient. The maximum retrieval time is 4.00 ms, and the minimum is 0.10 ms. This efficiency is crucial for real-time applications where rapid data access is essential.

The high relevance score (94%) indicates that the system effectively retrieves contextually relevant documents. This high relevance is important for ensuring that the retrieved information is useful and accurate.

Special Case Observations:

1. Religious Text Queries: Many queries relate to the interpretation and explanation of religious texts. The system appears to be well-configured to handle such queries, given the high contextual relevance scores.
2. Efficiency in Handling Large Documents: The system's ability to efficiently manage documents with a high number of tokens is noteworthy. This demonstrates resilience in processing and storing large texts.
3. Low Retrieval Time: The low average retrieval time indicates that the database structure and retrieval algorithms are optimized for quick access, which benefits the user experience in real-time systems.

To view the results of tokens, time, and context relevance, refer to Table 6, where every set of five queries is linked to one source document (Quran/Juz). We calculated the number of tokens in each Juz and the time required to store them in the database (in numerical vector format). In this case, we used Qdrant because it is well-suited for numerical vector databases.

After the storage process, we proceeded with data retrieval. For retrieval, each query produced different RAG document results (150 RAG documents). We then selected 5 chunks with the highest cosine similarity scores. We also calculated the number of tokens in each RAG document and the time required for it.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 10. NO. 2 NOVEMBER 2024
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v10i2.5916

To assess contextual relevance, we manually examined the RAG document results for each query, assigning a score of 1 if the document was contextually relevant and 0 if not. Optimization for source documents and RAG documents yielded excellent results, with an average score of 91.74%. The model will vary with each query due to the use of word embeddings. Not every word in the Indonesian language has a representation in numerical vectors. Therefore, our future task is to create word embeddings that can represent all words in numerical vector form, especially words in the Quran. In this context, to find the model, we identified queries optimized with appropriate word embedding.

Table 7. Tokens, time, and contextual relevance results.

No	Query	Sources documents token	RAG documents token	Time to store data	Time to retrieve data	Contextual relevance
1	What is the guidance for those who fear	11698	972	86.07	0.22	1
2	What is the nature of disbelievers	11698	960	86.07	0.12	1
3	What caused prophet musa to be angry with the children of israil	11698	1045	86.07	0.22	1
4	Who is the harut and marut	11698	1118	86.07	0.13	1
5	Why the devil would not bow down to adam	11698	1044	86.07	0.18	0
6	Where is the Qibla of Muslims	11439	975	83.55	0.16	1
7	What is the meaning of Surah-Al-Baqarah/153	11439	1245	83.55	0.17	1
8	What kind of people know Muhammad as they know their own children?	11439	1,189	83.55	0.22	1
...	...	...	...	...	...	...
148	How the devil whispers evil	16721	844	114.70	0.15	1
149	What is the main content of surah-al-fajr	16721	1242	114.70	0.14	1
150	How is the Godhead	16721	1126	114.70	0.10	1

Source : (Research Result, 2024)

CONCLUSION

The analysis shows that this system performs well in terms of data storage and retrieval times while maintaining high contextual relevance. These findings indicate that the system is robust and efficient, capable of handling various document lengths and complexities while ensuring quick access to relevant information.

Further optimization of storage times could enhance overall efficiency, particularly for larger documents, although the system already performs well. Conducting scalability testing to ensure consistent performance with increased data loads would also be highly beneficial. Continuously refining relevance algorithms could help maintain and even improve the high contextual relevance scores. This comprehensive analysis provides insights into the system's performance and potential areas for improvement."

REFERENCES

[1] I. L. Alberts *et al.*, “Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be?,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 50, no. 6, pp. 1549–1552, 2023, doi: 10.1007/s00259-023-06172-w.

[2] I. O. Gallegos *et al.*, “Bias and Fairness in Large Language Models: A Survey,” *Comput. Linguist.*, no. March, pp. 1–83, 2024, doi: 10.1162/coli_a_00524.

[3] P. Dufter, M. Schmitt, and H. Schütze, “Position Information in Transformers: An Overview,” *Comput. Linguist.*, vol. 48, no. 3, pp. 733–763, 2022, doi: 10.1162/coli_a_00445.

[4] M. Mandelkern and T. Linzen, “Do Language Models’ Words Refer?,” *Comput. Linguist.*, no. October 2023, pp. 1–10, 2024, doi: 10.1162/coli_a_00522.

[5] A. Y. Alan, Ö. Aydın, and E. Karaarslan, “A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM,” *SSRN Electron. J.*, pp. 1–21, 2024, doi: 10.2139/ssrn.4707470.

[6] A. Chaturvedi, S. Bhar, S. Saha, U. Garain, and N. Asher, “Analyzing Semantic Faithfulness of Language Models via Input Intervention on Question Answering,” *Comput. Linguist.*, vol. 50, no. 1, pp. 119–155, 2023, doi: 10.1162/coli_a_00493.

[7] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, “Large Language Models Struggle to Learn Long-Tail Knowledge,” *Proc. Mach. Learn. Res.*, vol. 202, pp. 15696–15707, 2023.

[8] C. Ziems, W. Held, O. Shaikh, J. Chen, Z.

Zhang, and D. Yang, "Can Large Language Models Transform Computational Social Science?," *Comput. Linguist.*, vol. 50, no. 1, pp. 237–291, 2023, doi: 10.1162/coli_a_00502.

[9] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A Large Language Model for Extracting Information from Financial Text*," *Contemp. Account. Res.*, vol. 40, no. 2, pp. 806–841, 2023, doi: 10.1111/1911-3846.12832.

[10] J. Huang *et al.*, *ERNIE-GeoL: A Geography-and-Language Pre-trained Model and its Applications in Baidu Maps*, vol. 1, no. 1. Association for Computing Machinery, 2022. doi: 10.1145/3534678.3539021.

[11] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, and K. Lindén, "Automatic language identification in texts: A survey," *J. Artif. Intell. Res.*, vol. 65, pp. 675–782, 2019, doi: 10.1613/JAIR.1.11675.

[12] T. A. Chang and B. K. Bergen, "Language Model Behavior: A Comprehensive Survey," *Comput. Linguist.*, vol. 50, no. 1, pp. 293–350, 2024, doi: 10.1162/coli_a_00492.

[13] T. Sommerschield *et al.*, "Machine Learning for Ancient Languages: A Survey," *Comput. Linguist.*, vol. 49, no. 3, pp. 703–747, 2023, doi: 10.1162/coli_a_00481.

[14] T. Giallanza, D. Campbell, and J. D. Cohen, "Toward the Emergence of Intelligent Control: Episodic Generalization and Optimization," *Open Mind*, vol. 8, pp. 688–722, 2024, doi: 10.1162/opmi_a_00143.

[15] M. Fatehkia, J. K. Lucas, and S. Chawla, "T-RAG: Lessons from the LLM Trenches," pp. 1–22, 2024, [Online]. Available: http://arxiv.org/abs/2402.07483

[16] F. Cuconasu *et al.*, "The Power of Noise: Redefining Retrieval for RAG Systems," *SIGIR 2024 - Proc. 47th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, pp. 719–729, 2024, doi: 10.1145/3626772.3657834.

[17] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1–17, 2023, doi: 10.1162/tacl_a_00530.

[18] W. Fan *et al.*, "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 6491–6501, 2024, doi: 10.1145/3637528.3671470.

[19] E. Melz, "Enhancing LLM Intelligence with ARM-RAG: Auxiliary Rationale Memory for Retrieval Augmented Generation," 2023, [Online]. Available: http://arxiv.org/abs/2311.04177

[20] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, 2020.

[21] K. Guu, K. Lee, Z. Tung, and P. Pasupat, "REALM : Retrieval-Augmented Language Model Pre-Training," 2019.

[22] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," pp. 1–21, 2023, [Online]. Available: http://arxiv.org/abs/2312.10997

[23] I. Drori *et al.*, "From Human Days to Machine Seconds: Automatically Answering and Generating Machine Learning Final Exams," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 3947–3955, 2023, doi: 10.1145/3580305.3599827.

[24] C. Fang *et al.*, "RecruitPro: A Pretrained Language Model with Skill-Aware Prompt Learning for Intelligent Recruitment," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 3991–4002, 2023, doi: 10.1145/3580305.3599894.

[25] K. He *et al.*, "A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics," vol. 14, no. 8, pp. 1–32, 2023, [Online]. Available: http://arxiv.org/abs/2310.05694

[26] Q. Guo, S. Cao, and Z. Yi, "A medical question answering system using large language models and knowledge graphs," *Int. J. Intell. Syst.*, vol. 37, no. 11, pp. 8548–8564, 2022, doi: https://doi.org/10.1002/int.22955.

[27] A. Hamidi and K. Roberts, "Evaluation of AI Chatbots for Patient-Specific EHR Questions," 2023, [Online]. Available: http://arxiv.org/abs/2306.02549

[28] T. Lai *et al.*, "Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models," 2023, [Online]. Available: http://arxiv.org/abs/2307.11991

[29] A. Louis, G. van Dijck, and G. Spanakis, "Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models," *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 20, pp. 22266–22275, 2024, doi: 10.1609/aaai.v38i20.30232.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri