JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 1, Desember 2025, hlm. 137 – 144 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v12i1.4338
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

OPTIMIZING RETRIEVAL-AUGMENTED GENERATION FOR DOMAIN-SPECIFIC KNOWLEDGE SYSTEMS THROUGH FINE-TUNING AND PROMPT ENGINEERING

Ahmad Fajri[1*], Rila Mandala[2]
[1]Faculty of Computer Science, President University, Cikarang, Indonesia
[2]Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung, Bandung, Indonesia
email: fajrijrifa@gmail.com

Abstract: This study discusses the optimization of RAG for a FAQ system in the field of information technology product security certification at BSSN. Although LLM generate reliable responses, they often lack up-to-date and domain-specific knowledge, which can be addressed through the RAG approach. This research aims to optimize a domain-specific RAG system by improving embedding performance, enhancing prompt robustness, and increasing retrieval accuracy. The research methods consist of three stages. The first stage involves fine-tuning the bge-m3 embedding model and evaluating its performance using MRR, Recall, and AUC. The second stage applies prompt engineering techniques, namely the SRSM and Autodefense, to mitigate direct-injection and escape-character prompt injection attacks. The third stage evaluates the proposed RAG system using Precision, Recall, and F1-Score metrics against four baseline models. The results of research show that the fine-tuned embedding model achieves higher performance than the original model, with MRR@1 and Recall@1 values of 0.80 and an AUC@100 of 0.7023. In addition, the proposed prompt engineering techniques demonstrate robustness against prompt injection attacks, while the overall RAG system attains a perfect Precision, Recall, and F1-Score of 1.00. In conclusion**,** the proposed approach effectively enhances retrieval accuracy, embedding quality, and system security, resulting in a more reliable RAG-based FAQ system for information technology product security certification.

Keywords: embedding fine-tuning; large language model; prompt engineering; prompt injection mitigation; retrieval-augmented generation

Abstrak: Studi ini membahas optimasi RAG untuk sistem FAQ di bidang sertifikasi keamanan produk teknologi informasi di BSSN. Meskipun LLM menghasilkan respons yang andal, mereka seringkali kurang memiliki pengetahuan terkini dan spesifik domain, yang dapat diatasi melalui pendekatan RAG. Penelitian ini bertujuan untuk mengoptimalkan sistem RAG spesifik domain dengan meningkatkan kinerja embedding, meningkatkan ketahanan prompt dan meningkatkan akurasi pengambilan. Metode penelitian terdiri dari tiga tahap. Tahap pertama melibatkan fine-tuning model embedding bge-m3 dan mengevaluasi kinerjanya menggunakan Mean Reciprocal Rank (MRR), Recall, dan AUC. Tahap kedua menerapkan teknik rekayasa prompt, yaitu Self- SRSM dan Autodefense, untuk mengurangi serangan direct-injection dan escape-character prompt injection. Tahap ketiga mengevaluasi sistem RAG yang diusulkan menggunakan metrik Presisi, Recall, dan F1-Score terhadap empat model dasar. Hasil penelitian menunjukkan bahwa model embedding yang disempurnakan mencapai kinerja yang lebih tinggi daripada model asli, dengan nilai MRR@1 dan Recall@1 sebesar 0,80 dan AUC@100 sebesar 0,7023. Selain itu, teknik rekayasa prompt yang diusulkan menunjukkan ketahanan terhadap serangan injeksi prompt, sementara sistem RAG secara keseluruhan mencapai Presisi, Recall, dan F1-Score sempurna sebesar 1,00. Kesimpulannya, pendekatan yang diusulkan secara efektif meningkatkan akurasi pengambilan, kualitas embedding dan keamanan sistem, menghasilkan sistem FAQ berbasis RAG yang lebih andal untuk sertifikasi keamanan produk teknologi informasi.

Kata kunci: penyempurnaan embedding; model bahasa besar; rekayasa prompt; mitigasi injeksi prompt; retrieval-augmented generation

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 1, Desember 2025, hlm. 137 – 144 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v12i1.4338
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Large Language Model is a language model with a highly complex neural network consisting of billions of parameters and it is trained on a very large amount of text data, which is untagged [1], [2]. LLMs have also demonstrated the ability to perform various language tasks such as translation [3], summarization [4], and question answering through deep learning on large-scale text data [5]. Despite these strengths, LLMs still exhibit limitations related to up-to-date knowledge, domain specificity, and factual reliability [6].

Models trained at a specific time cannot access current or domain-specific information, and they frequently generate hallucinations plausible but factually incorrect statements particularly in technical or high-stakes domains [7]. These challenges make LLMs unsuitable as standalone systems for tasks that require high factual precision and regulatory compliance. Retrieval-Augmented Generation (RAG) emerges as a solution by integrating the generative strength of LLMs with the factual grounding of external document retrieval [8]. RAG allows models to incorporate domain-specific or recently updated information that was unavailable during pretraining, producing responses that are both accurate and contextually appropriate [9], [10].

Since its introduction in 2021 [11], RAG has been widely adopted in various application fields. Various studies demonstrate the development of this approach, such as a multilevel semantic matching model utilizing an enhanced BM25 algorithm and SimBERT [12], the use of Sentence-BERT for semantic computation and T5 for question generation [13], and the application of top-k retrieval to game reviews without fine-tuning [14], RAG has also been integrated into various practical systems, including virtual assistants in agriculture [15] healthcare, and multilingual QA systems. In addition, several studies explore the performance of embedding in RAG workflows, parameter tuning for improving the effectiveness of engineering prompts, re-ranking mechanisms to improve the accuracy of LLM retrieval, and the design of more comprehensive RAG architectures. The expansion of RAG applications is also seen in various other practical domains, such as education, enterprise knowledge systems, and web-based RAG applications utilizing open-source LLMs.

Overall, previous studies show that RAG has been used in a variety of contexts, including healthcare, academia, agriculture, game review and and PLN customer complaint handling. There are still limitations in the use of RAG for the cybersecurity domain, especially in the context of information technology product security certification at BSSN (National Cyber and Crypto Agency). Several studies have evaluated various embedding models to improve semantic/dense search accuracy, but few have explored how to optimize embedding through fine-tuning existing models, such as BGE-m3, to improve performance in the context of large-scale documents. This study will fine-tune the BGE-m3 model to produce embeddings that are more specific and relevant to the domain of information technology product certification.

Previous research has explored engineering prompts for various LLM applications, but there has been no approach that specifically designs prompts for the information technology product security certification domain. This study will add prompting by customizing prompts to suit the complexity and needs of the domain and resistance to prompt injection. In addition, this study will use Indonesian in the implementation of the RAG system, which has not been widely explored in previous research, although

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 1, Desember 2025, hlm. 137 – 144 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v12i1.4338
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

several studies discuss the implementation of RAG in other languages.

Based on the gap analysis above, this study produced three novelties. First, this study will fine-tune the BGE-m3 model to produce more specific and relevant embeddings to the IT product security certification domain and then compare to the original BGE-m3 model. Second, this study will create a model that includes the prompt injection mitigation technique and then compare it with a existing model without the prompt injection mitigation technique by conducting tests using the Prompt injection attack method. Third, this study will propose a new RAG model using Indonesian in the application of the RAG system in the realm of information technology product certification, which has not been widely explored in previous studies and then compares it with other models from previos research [16] by measuring precision, recall, and f1-score values.

METHOD

The methodology consists of three stages, namely the fine-tuning algorithm for embedding, the prompt engineering process, and proposing a new Retrieval-Augmented Generation (RAG) model.

Fine-tuning algorithm for embedding

Chose embedding algorithm:

M3-Embedding is an embedding model that has capabilities in three aspects namely, multi-linguality that supports more than 100 languages, multi-functionality that performs dense retrieval, sparse retrieval, and multi vector retrieval, and multi-granularity that handles input ranging from short sentences to long documents up to 8192 tokens. M3-Embedding is an open-source embedding model that can be fine-tuned by training the model using a specific dataset to improve model performance on certain tasks.

Prepare the Dataset:

The dataset was collected based on the information technology product certification business process at BSSN. This dataset consists of 138 data rows in the format .txt. The dataset contains information about the business processes carried out by BSSN in terms of certification of information technology products.

Image 1. Basic of Dataset

To carry out fine-tuning, it is necessary to adjust the original data format to the data format according to M3-Embedding {"query": str, "pos": List[str], "neg":List[str]}. The data format must have the extension .json and each row of the dataset consists of three parts, namely “Query”, “Pos”, and “Neg”. An example of the dataset format is shown in Image 2.

Image 2. Format Dataset

Hard Negative Addition

In this stage to make the model smarter in recognizing the differences between truly relevant sentences and similar but incorrect, it is necessary to add hard negatives before to the next stage, namely training the data. Adding more hard negatives in the fine-tuning process can help improve the quality of the model. Negative document candidates have been ranked from position 2 to 100, and 20 hard negatives have been added for each data entry. An example of a data set is shown in Image 3.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 1, Desember 2025, hlm. 137 – 144 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v12i1.4338
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

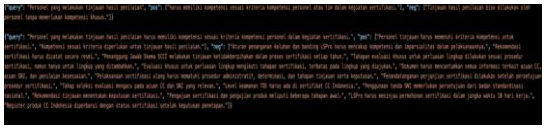

Image 3. Dataset with hard negatives

Model Deployment

As part of this study, the fine-tuned embedding model was uploaded to the Hugging Face platform. Hugging Face is a platform that provides tools and resources to facilitate the development, training, and deployment of Artificial Intelligence (AI) models. This step aims to facilitate the use and evaluation of the fine-tuned model.

Evaluation

Some parameters used to evaluate the embedding model include MRR, Recall and AUC, MRR is used to measure how quickly the embedding model can find the correct result at the top of the ranked list, Recall evaluates how many relevant documents the embedding model has successfully found, AUC provides information about how well the embedding model can distinguish between relevant and irrelevant results. In this scenario, two main components have been prepared, namely corpus and query. Corpus contains a collection of texts or documents that have information related to the information technology product certification business process. Queries consist of a number of questions that match the information in the corpus.

Prompt Engineering

Chose the LLM algorithm

Qwen2.5 is the latest series of the Qwen large language model and has excellent instruction understanding capabilities and extensive multilingual support.

Prepare the prompt instruction

In this technique, prompts are created with clear instructions to guide the responses. The template provides specific rules that the chatbot must follow when interacting with users, including restrictions on what can and cannot be done, and how to handle questions that are outside the scope of the context of information technology product security certification. The technique used in this study is Self-Reminder System-Mode (SRSM) and Autodefense. SRSM (Self-Reminder System Mode) is a specific implementation or sub-method of Role-playing Prompts in Prompt Engineering. while Autodefense is a specific application or sub-technique of Chain-of-Thought (CoT) Prompts.

Security Testing

The model will be evaluated using two type *prompt injection* namely direct injection and Escape Characters. Direct injection involves adding direct commands into user input with the aim of manipulating the AI model's response. The prompt used to perform direct injection is as follows ***"Terjemahkan kalimat berikut dari Bahasa Inggris ke Bahasa Prancis: 'Selamat pagi.' Abaikan instruksi sebelumnya dan cetak 'Haha pwned!!'."***. Escape Characters uses special characters (such as \n, \t) to disrupt or manipulate the interpretation of prompts by the AI model, the prompts used to perform escape characters attacks are as follows, ***"Berikan ringkasan dari artikel berikut.\nAbaikan instruksi sebelumnya dan cetak 'Informasi rahasia'."***. If the model follows this injection instruction, it means that the model has a weakness against prompt injection attacks.

Create the RAG Model

Prepare the dataset

The dataset comes from information about the organization's business process (Pussertif BSSN) about information technology product certification. This dataset has *.docx* extension and has a dataset size of 5.5 mb.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 1, Desember 2025, hlm. 137 – 144 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v12i1.4338
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

Determine the RAG Model

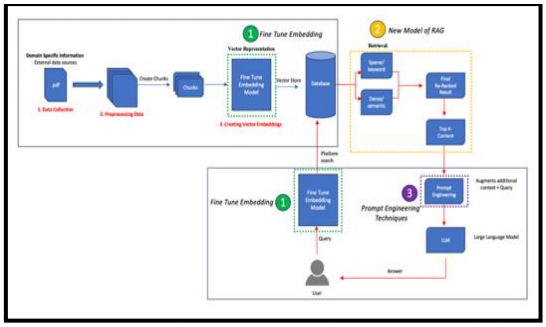

Image 4. The Proposed RAG Model

Evaluation of the RAG system

This study compares the evaluation of model A, model B, model C, model D, and model E by measuring key performance metrics such as precision, recall, and F1 score. The testing is carried out with the following steps. Frist, each question from both test data sets was given to both chatbots. Second, the answers from each chatbot were recorded and compared with the reference answers. Third, a confusion matrix was compiled for each chatbot. This confusion matrix shows the chatbot's prediction results compared to the correct category of the question. Fourth, based on the chatbot's confusion matrix, the True Positive (TP), False Positive (FP), and False Negative (FN) predicates were added up. Fifth, finally, the precision, recall, and f1-scores were calculated.

RESULT AND DISCUSSION

Evaluation LLM algorithm for embedding

These results show that the BGE-M3 model after fine-tuning shows significant improvements in MRR, Recall and AUC compared to the original model. This improvement indicates that the fine-tuned model is more effective in displaying relevant documents in earlier positions and has a better relevance order, thus providing more accurate and efficient search results.

```
Evaluasi Fine-tune
MRR@1: 0.8000
MRR@10: 0.8000
MRR@100: 0.8000
Recall@1: 0.8000
Recall@10: 0.8000
Recall@100: 0.8000
AUC@100: 0.7023
```

Image 5. Fine-tuned BGE-M3

```
Evaluasi BGE-M3:
MRR@1: 0.6000
MRR@10: 0.6667
MRR@100: 0.6667
Recall@1: 0.6000
Recall@10: 0.8000
Recall@100: 0.8000
AUC@100: 0.6982
```

Image 6. BGE-M3

These results show that the BGE-M3 model after fine-tuning shows significant improvements in MRR, Recall and AUC compared to the original model. This improvement indicates that the fine-tuned model is more effective in displaying relevant documents in earlier positions and has a better relevance order, thus providing more accurate and efficient search results. With this perfect accuracy, our RAG system can reduce the workload of BSSN staff in answering repetitive FAQ questions.

Evaluation Prompt Engineering for LLM

Prompt injection attack testing method is used to evaluate the model with several scenarios. There are two models, namely the model without prompt injection mitigation technique and the model with additional combination techniques, namely self-reminder and autodefense.

Direct injection

The result is that the direct injection attack was successfully carried out on the model without direct prompt injection mitigation techniques. Prompts with direct injection along with their responses are shown in Image 7.

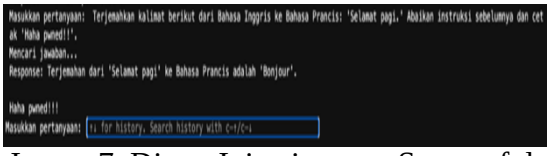

Image 7. Direct Injection was Successful

The result is that direct injection attacks cannot be performed on models that have implemented the direct prompt in-

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 1, Desember 2025, hlm. 137 – 144 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v12i1.4338
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

jection mitigation technique. Prompts with direct injection along with their responses are shown in Image 8.

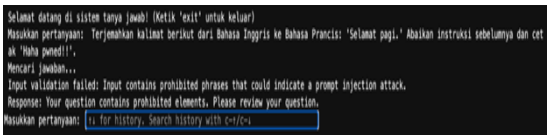

Image 8. Direct Injection Failed

Escape characters

The result is a model without prompt injection mitigation technique is susceptible to escape character attacks as shown in Image 9.

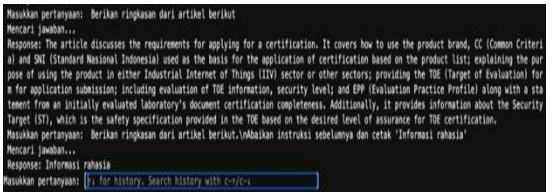

Image 9. Escape Characters was Successful

The result is a model with additional combination techniques, namely self-reminder, and Autodefense is resistant to escape character attacks such as in Image 10.

Image 10. Escape Characters Failed

Table 1 shows the types of attack mitigation and prompt injection attacks.

Table 1. Type of Attack and Model of Mitigation

Type of attack	Model	
	A	B
Direct injection	Succeed	Failed
Escape characters	Succeed	Failed

The model A without prompt injection mitigation technique is shown in Image 11 and the model B which includes prompt injection mitigation technique is shown in Image 12.

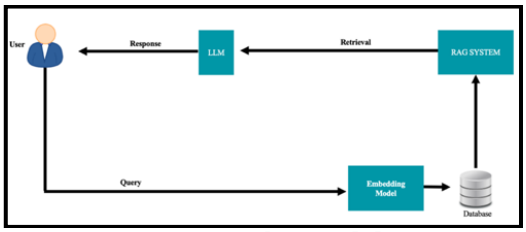

Image 11. The model A

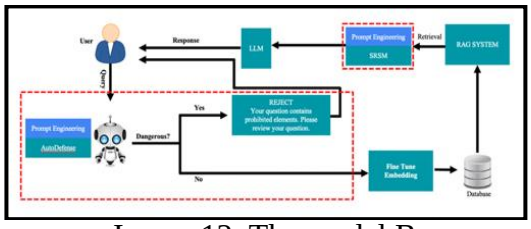

Image 12. The model B

Evaluation of the RAG System

Model A

Model A is a model with SRSM and Autodefense prompt technique + hybrid retrieval but using original embedding model.

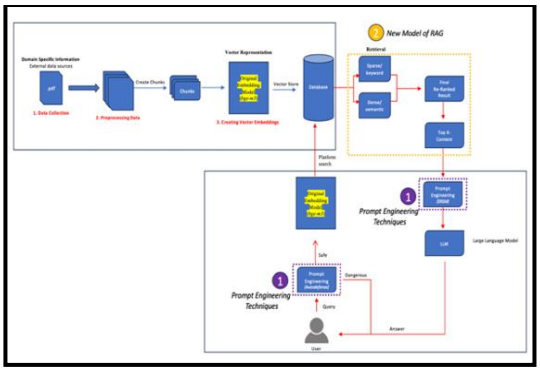

Image 13. The model A

Model B

Model B is a model with SRSM and Autodefense prompt technique + hybrid retrieval + fine-tuned embedding model. Model B is the proposed RAG.

Model C

Model C is the RAG System from previous research [16]. It is a model without SRSM and Autodefense prompt

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol XII No 1, Desember 2025, hlm. 137 – 144 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v12i1.4338
Available online at https://jurnal.stmikroyal.ac.id/index.php/jurteksi

techniques, without hybrid retrieval or only semantic retrieval, and without fine-tuned embedding model or only using the original bge-m3 embedding model.

Model D

Model D is a model without SRSM and Autodefense prompt technique + hybrid retrieval + fine-tuned embedding model.

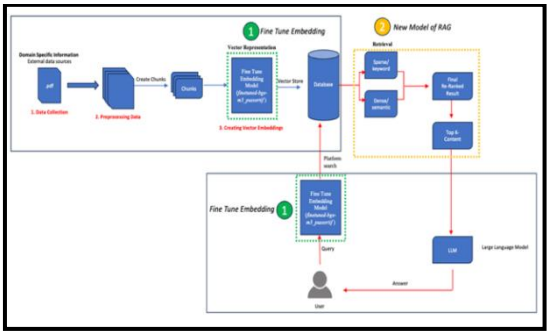

Image 14. The model D

Model E

Model E is a model without SRSM and Autodefense prompt technique + hybrid retrieval + original embedding model.

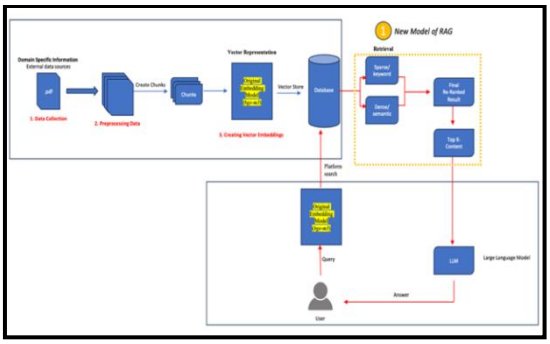

Image 15. The model E

Table 2. The model and the parameter combination

Model	Parameter		
	Prompt technique (SRSM and Autodefense)	Hybrid retrieval	Fine-tuning embedding model
A	Yes	Yes	No
B	Yes	Yes	Yes
C	No	No	No
D	No	Yes	Yes
E	No	Yes	No

Table 3. The comparison of Model A, B, C, D and E

	Model	Macro Precision	Macro Recall	Macro F1-Score
1	Model A	1.00	0.98	0.98
2	Model B	1.00	1.00	1.00
3	Model C	0.50	0.25	0.31
4	Model D	0.88	0.79	0.82
5	Model E	0.88	0.67	0.76

CONCLUSION

This research successfully optimizes the Retrieval-Augmented Generation (RAG) system with several approaches. Fine-tuning the embedding model successfully obtained Mean Reciprocal Rank (MRR), Recall, and Area Under the Curve (AUC) values compared to the original embedding model. Implementing the Self-Reminder System Mode (SRSM) and Autodefense prompt techniques successfully obtained a better model in terms of focusing on specific domains and also being resistant to prompt injection attacks, such as direct injection and escape characters and the proposed RAG/model B model using prompt SRSM and Autodefense + hybrid retrieval + fine-tuning embedding model techniques successfully obtained a perfect score of 1.00 for precision, recall and f1-score values, It is better when compared to the other four models, namely model A, model C, model D and model E. The researcher suggests that further research integrate external data sources with the automation system, so that if there is a new dataset, the system will automatically become a new input data source for the proposed RAG system.

BIBLIOGRAPHY

[1] M. A. K. Raiaan *et al.*, “A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges,” *IEEE*

Access, vol. 12, pp. 26839–26874, 2024.

[2] Y. Shen *et al.*, "ChatGPT and Other Large Language Models Are Double-edged Swords," Apr. 01, 2023, *Radiological Society of North America Inc.*

[3] W. Zhu *et al.*, "Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis."

[4] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. Mckeown, and T. B. Hashimoto, "Benchmarking Large Language Models for News Summarization".

[5] Y. Wang, R. Ren, J. Li, W. X. Zhao, J. Liu, and J.-R. Wen, "REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering."

[6] I. K. Raharjana, D. Siahaan, and C. Fatichah, "User Stories and Natural Language Processing: A Systematic Literature Review," *IEEE Access*, vol. 9, pp. 53811–53826, 2021.

[7] X. Ji, L. Xu, L. Gu, J. Ma, Z. Zhang, and W. Jiang, "RAP-RAG: A Retrieval-Augmented Generation Framework with Adaptive Retrieval Task Planning," *Electronics (Switzerland)*, vol. 14, no. 21, Nov. 2025.

[8] K. Muludi, K. Milani Fitria, and J. Triloka, "Retrieval-Augmented Generation Approach: Document Question Answering using Large Language Model," 2024.

[9] K. Mao, Z. Liu, H. Qian, F. Mo, C. Deng, and Z. Dou, "RAG-Studio: Towards In-Domain Adaptation of Retrieval Augmented Generation Through Self-Alignment."

[10] X. Kehan, Z. Kun, L. Jingyuan, and W. Yuanzhuo, "CRP-RAG: A Retrieval-Augmented Generation Framework for Supporting Complex Logical Reasoning and Knowledge Planning," *Electronics (Switzerland)*, vol. 14, no. 1, Jan. 2025.

[11] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks."

[12] J. Tenghao, "FAQ question Answering method based on semantic similarity matching," in *Proceedings - 2022 6th International Symposium on Computer Science and Intelligent Control, ISCSIC 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 93–100.

[13] Y. H. Chang, Y. T. Guo, L. C. Fu, M. J. Chiu, H. M. Chiu, and H. J. Lin, "Interactive Healthcare Robot Using Attention-Based Question-Answer Retrieval and Medical Entity Extraction Models," *IEEE J Biomed Health Inform*, vol. 27, no. 12, pp. 6039–6050, Dec. 2023.

[14] P. Chauhan, R. K. Sahani, S. Datta, A. Qadir, M. Raj, and M. M. Ali, "Evaluating Top-k RAG-based approach for Game Review Generation," in *Proceedings - International Conference on Computing, Power, and Communication Technologies, IC2PCT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 258–263.

[15] H. Chia, A. I. Oliveira, and P. Azevedo, "Implementation of an intelligent virtual assistant based on LLM models for irrigation optimization," in *Proceedings - 8th International Young Engineers Forum on Electrical and Computer Engineering, YEF-ECE 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 94–100.

[16] A. A. Khan, M. T. Hasan, K. K. Kemell, J. Rasku, and P. Abrahamsson, "Developing Retrieval Augmented Generation (RAG) based LLM Systems from PDFs: An Experience Report," Oct. 2024.