International Journal of Innovative Computing, Information and Control
Volume **21**, Number **2**, April **2025**

ICIC International ©2025 ISSN 1349-4198
pp. **481–490**

INNOVATIVE APPLICATIONS OF RAG-ENHANCED SMALL LLM FOR CLOSED-DOMAIN Q&A

Youngpyo Hong and Dongsoo Kim*

Department of Industrial and Information Systems Engineering
Soongsil University
369 Sangdo-Ro, Dongjak-Gu, Seoul 06978, Korea
ypyohong@soongsil.ac.kr; *Corresponding author: dskim@ssu.ac.kr

Received August 2024; revised November 2024

Abstract. *Efficiently integrating Large Language Models (LLMs) into business operations requires addressing hallucination issues, incorporating proprietary internal datasets, and ensuring economic viability. Key strategies to tackle this challenge involve implementing a Retrieval-Augmented Generation (RAG) methodology, fine-tuning open-source small LLMs using proprietary internal datasets, and establishing on-premises GPU infrastructure. In this study, we validate the performance of this methodology through practical applications. Our selected use case involves responding to queries related to operational knowledge, including regulations, guidelines, and manuals. To achieve this, we deploy a Machine Reading Comprehension (MRC)-based RAG system. Additionally, the Llama 2 model undergoes fine-tuning on both internal and external datasets to enhance Korean language understanding and acquire domain-specific knowledge. The system's performance was evaluated based on the accuracy achieved using two datasets: a set of 200 Q&A pairs prepared by task managers and a dataset comprising 150 Q&A pairs derived from training exam questions and answers for new employees. The evaluations yielded accuracy scores of 92.7% and 79.3%, respectively.*

Keywords: Retrieval-augmented generation, Machine reading comprehension, Large language models, Generative AI, Closed-domain question and answering

1. **Introduction.** Recent advancements in natural language processing have led to significant interest in Retrieval-Augmented Generation (RAG) methods, particularly for closed-domain question answering systems [1]. RAG combines retrieval-based models and generation-based models to enhance the accuracy and relevance of generated responses. Previous studies, such as those by Lewis et al. [2] and Izacard and Grave [3], have demonstrated the efficacy of RAG in open-domain settings. However, there remains a notable gap in research concerning the application of RAG methods within closed-domain environments. Existing literature, including studies by Guu et al. [4] and Karpukhin et al. [5], predominantly focuses on the retrieval and generation processes using public, open-domain datasets. These studies highlight the challenges of maintaining context and coherence in responses, a problem that is exacerbated in closed-domain settings where the dataset is proprietary and domain-specific [6]. Therefore, recent research has actively focused on RAG methodologies that leverage closed-domain environments. Siriwardhana et al. [7] focus on enhancing the domain adaptation of RAG models for Open-Domain Question Answering (ODQA). They propose RAG-end2end, an extension that adapts to domain-specific knowledge bases by updating all components during training. Their approach includes an auxiliary training signal to inject more domain-specific knowledge, demonstrating significant performance improvements in specialized domains such as healthcare

DOI: 10.24507/ijicic.21.02.481

and news. Han et al. [8] examine the use of RAG-based Large Language Models (LLMs) to automate Systematic Literature Reviews (SLRs). They detail the RAG framework's processes – retrieval, augmentation, and generation – and propose a comprehensive framework for automating SLRs. This work underscores the potential of RAG in efficiently synthesizing domain-specific literature, thereby facilitating research in closed-domain environments. Wu et al. [9] review significant techniques of RAG, focusing on the retriever and retrieval fusions. They discuss how RAG leverages external knowledge databases to augment LLMs, addressing issues like hallucinations and knowledge update challenges. The paper also explores applications of RAG in natural language processing tasks and industrial scenarios, highlighting future directions and challenges. Our research aims to bridge this gap by exploring the feasibility and effectiveness of RAG-enhanced small Large Language Models (LLMs) for closed-domain question answering. We utilize a proprietary dataset to fine-tune small LLMs, configuring an RAG architecture to assess its performance [10]. This approach addresses several challenges identified in previous studies, such as data security and data sovereignty. By establishing an on-premises GPU infrastructure, we mitigate the risk of data leakage and ensure compliance with regulatory requirements. Moreover, our study responds to the call for more detailed examinations of how closed-domain datasets can be leveraged within the RAG framework. Unlike open-domain datasets, closed-domain datasets require specialized handling to maintain data integrity and relevance. This research not only contributes to the existing body of knowledge but also provides practical insights for businesses looking to deploy RAG-enhanced models in a secure and feasible manner.

The structure of this paper is organized as follows. Section 2, Related Work, introduces the foundational concepts of Retrieval-Augmented Generation (RAG), Machine Reading Comprehension (MRC), and Llama 2. Section 3, Method, details the research approach, comprising three key components. Section 3.1, Data Preprocessing, discusses the processes of dataset selection, collection, and data manipulation. Section 3.2, System Architecture and Implementation, describes the design and deployment of the on-premises GPU infrastructure, the software architecture, model fine-tuning, the configuration of a pipeline architecture integrating RAG and a small LLM, and the development of the user interface. Subsection 3.3, Performance Evaluation, presents the system evaluation results and their implications. Section 3.4, Discussion, examines considerations regarding data, models, and operational aspects for effective application in business environments. Lastly, Section 4, Conclusion, summarizes the key findings and implications of this study.

2. **Related Work.**

The RAG architectural paradigm offers a promising approach to enhance the efficacy of LLM applications by leveraging custom datasets [11]. This is achieved through retrieving relevant data or documents based on a given query or task, which subsequently serve as contextual information for the LLM [12]. The RAG methodology has demonstrated successful applications in support chatbots and question-answering systems that require real-time access to up-to-date information or domain-specific knowledge repositories [13-15]. The RAG concept, as explained thus far, can be schematically represented as depicted in Figure 1.

Machine Reading Comprehension (MRC) is the task of constructing a system that comprehends a given passage in order to answer questions related to it. The input to the reading comprehension model consists of a question and a context or passage. The model's output is the answer extracted from the passage [16]. When visualizing the concept of MRC, it corresponds to Figure 2.

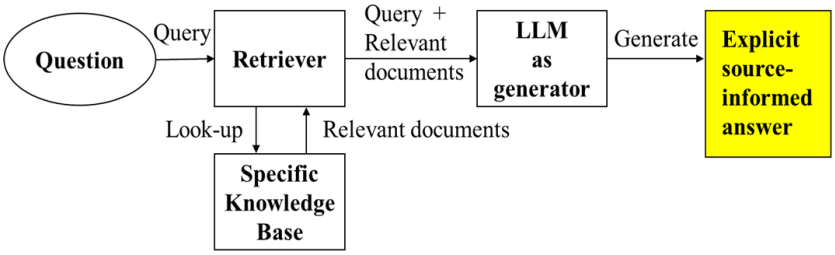

FIGURE 1. RAG concept diagram

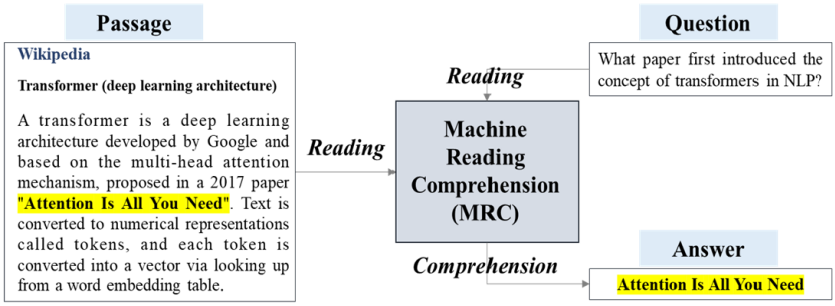

FIGURE 2. MRC concept diagram

Llama 2, announced by Meta in July 2023, is an open-source LLM that is commercially available. Therefore, even companies that cannot directly build massive GPU infrastructures can apply LLM services to their business using the Llama 2 model. Llama 2 provides three pre-trained and fine-tuned variants, each with different parameter sizes: 7 billion, 13 billion, and 70 billion. These models are specifically optimized for dialogue tasks, including the Llama 2-Chat variant. Notably, Llama 2 surpasses existing open-source chat models across various benchmarks and is considered a viable alternative to closed-source models based on human evaluations of helpfulness and safety [17].

3. **Method.**

In this study, we developed a question-answering system based on the RAG-enhanced Llama 2 model. The system was constructed in three primary phases. First, we conducted data preparation by selecting and pre-processing the necessary datasets for training. In the second phase, we established dedicated GPU computing infrastructure, fine-tuned the model, set up an RAG-enhanced small LLM pipeline, and created the user interface. Finally, the third phase involved evaluating the system's performance, where we established test datasets and conducted accuracy evaluations.

3.1. **Data preprocessing.**

The data preparation phase of this study comprises two primary procedures. First, it entails the selection and collection of datasets necessary for model training. Second, it involves preprocessing the collected data by cleaning and transforming it into a suitable format for subsequent model training.

3.1.1. *Dataset selection and collection.*

In this study, we constructed a dedicated dataset for our research by collecting and curating relevant work-related documents from within an enterprise, rather than relying on publicly available datasets. The collected materials included files from the enterprise's Knowledge Management System (KMS), attachments obtained through the enterprise's work portal, and manually managed files maintained by individual departments. These documents were available in a range of formats, including Hangul word processor files (.hwp), Word documents (.docx), Excel spreadsheets (.xlsx), presentation files (.pptx), and PDF files. While most of these documents followed

a hierarchical structure similar to that of legal documents (e.g., articles, paragraphs, subparagraphs, and items), some were presented in tabular form. Following a preprocessing stage, the data were reorganized into a hierarchical structure comprising six primary categories and eighteen subcategories defined according to task type. The classification criteria employed to form these categories are presented in Table 1.

TABLE 1. Datasets overview used in this study

Main category	Total number of subcategories	Total number of files	Total file size (MB)
1) Regulations	2	45	3.30
2) Guidelines	3	38	4.47
3) Business standards	5	570	149.37
4) Basic documents	3	691	38.67
5) Product information	1	62	39.90
6) Reference materials	4	45	270.16
Total	**18**	**1,451**	**505.87**

3.1.2. *Data manipulation.* In this phase, we ingest and extract information from the organized dataset. Subsequently, we partition and refine the dataset, followed by its storage in the database using indexing and embedding techniques. A comprehensive breakdown of the tasks associated with each step is provided in Table 2.

TABLE 2. The steps of data manipulation

Step	Task
1) File reading and data extraction	• Read files in various document formats and convert them to CSV format. • Remove unnecessary tags and extract content.
2) Data splitting and refinement	• Separate and refine passages: Parse and separate documents into passages consisting of articles, paragraphs, subparagraphs, and items (2,033 FAQ passages and 30,739 general passages). • Select data necessary for document search: select titles, content, metadata, and other relevant information.
3) Data embedding and storing	• Index and embed the dataset. • Archive each passage (stored in Vector DB). • Control the search target and scope of the search dataset.

3.2. System architecture and implementation.

3.2.1. *On-premises GPU infrastructure and software architecture.* In this study, we deployed an NVIDIA DGX A100 640GB system [18], as depicted in Figure 3, to serve as the computational resource for both training and inference. The minimum system software requirements were specified and documented in Table 3, necessitating the installation of DGX OS 6 to meet these prerequisites. However, since docker-compose was not included in the DGX OS 6 package, we had to install it separately. Our software architecture, based on this system infrastructure, comprised six layers: Landing Page, API Routing, Tools, Application Services, Engines, and Data Store, as illustrated in Table 4.

FIGURE 3. Installation of DGX A100

TABLE 3. The minimum system software requirements

Red Hat Enterprise Linux	8.8
CUDA	12.0
nvidia-driver	525.60.13
nvidia-docker2	2.13.0
docker-ce	24.0.0
docker-compose	1.28.0

TABLE 4. Layered software architecture

Landing Page	User Interface			
API Routing	API Interface			
Tools	Messaging	Logging	Monitoring	Statistics
Application Services	Engine and Services		Management Tools	
Engines	RAG Passage Search Engine	RAG FAQ Engine	RAG MRC Engine	Small LLM
Data Store	Elasticsearch, Milvus/Kafka, MariaDB			

3.2.2. *Model fine-tuning.* In this phase, we conducted fine-tuning for two distinct models: an MRC model and a small LLM. We selected MRC and LLM models by investigating the foundation models of top-ranking state-of-the-art models on Hugging Face and KorQuAD leaderboards. Additionally, we designed the architecture and set the parameters by referencing the guidelines of AI-based search companies [4]. First, for the MRC model, we employed RoBERTa as the foundation model [19]. Our fine-tuning process involved using a dataset of 7,321 samples formatted as Passage, Query, and MRC Output, sourced from our internal datasets. Second, for the small LLM, we selected Llama 2 (13B) as the foundation model and followed a three-step fine-tuning procedure: 1) pre-training the model using a customized corpus (7,954 KB) derived from our internal datasets, 2) continual training using an external Korean dataset (87,888 samples) publicly available on HuggingFace (as outlined in Table 5), and 3) instruction fine-tuning using our internal dataset (3,595 samples) in the format of Passage, Query, and LLM Output. The fine-tuning processes for both the MRC model and the LLM are visually depicted in Figure 4.

TABLE 5. The sources of external datasets employed for fine-tuning small LLM

Datasets	Sources
ko_alpaca.json	http://huggingface.co/datasets/nlpai-lab/kullm-v2
ko_lima_train/test.json	http://huggingface.co/datasets/taeshahn/ko-lima
ko_openorca.json	http://huggingface.co/datasets/kyujinpy/OpenOrca-KO
ko_openorca_gugugo.json	http://huggingface.co/datasets/squarelike/OpenOrca-gugugo-ko
ko_platy.json	http://huggingface.co/datasets/kyujinpy/KOR-OpenOrca-Platypus

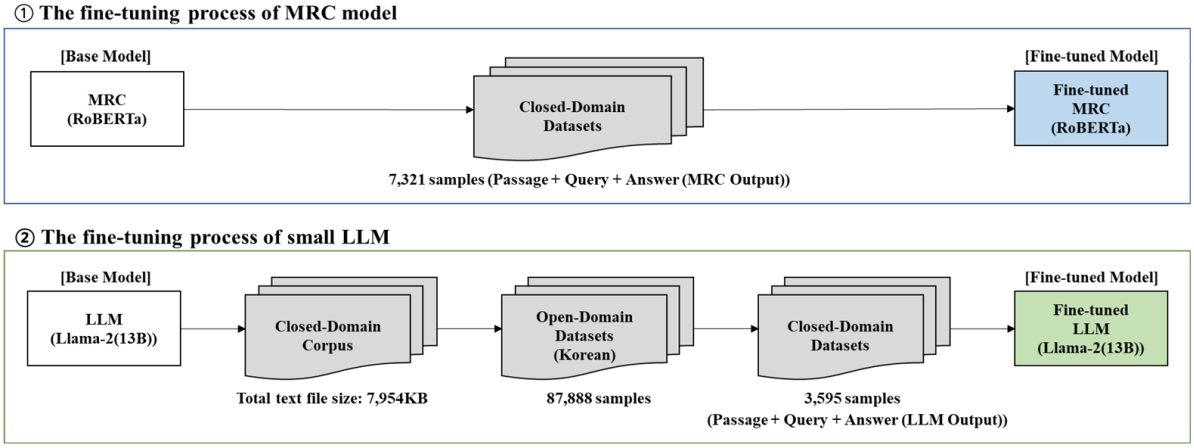

FIGURE 4. The fine-tuning process of MRC models and small LLM

3.2.3. *Configuration of a pipeline architecture integrating RAG and small LLM.* In this study, we established an RAG pipeline comprising three major engines. First, the Passage Retrieval Engine performs information retrieval and semantic retrieval for user queries, exploring the most relevant passages aligned with query intent. Subsequently, the user's query and the retrieved passages are forwarded to the FAQ Engine. Second, the FAQ Engine assesses whether the query-passage pair exhibits similarity to the query-answer pairs stored in the FAQ database. If the FAQ similarity score exceeds 0.7, the corresponding FAQ answer is promptly provided to the user. Third, if the FAQ similarity score falls below 0.7, the MRC Engine evaluates accuracy using a fine-tuned MRC model. If the F1 score-based MRC score surpasses 0.8, a prompt is constructed, incorporating the query, top two passages, and the MRC-generated answer. This prompt is then fed into the fine-tuned small LLM, which generates an appropriate response based on the provided information. However, if the MRC score remains below 0.8, we directly pass a prompt containing the query and the top three passages to the small LLM. This small LLM generates an answer referring to the given context. The overall process is visually depicted in Figure 5.

3.2.4. *User interface.* In accordance with Section 3.2.3, user queries receive responses through three distinct pathways: First, if the query aligns with an entry in the FAQ database; Second, if the answer is retrieved via the MRC model; Third, if the answer is directly generated by a small LLM. To indicate which of these three cases applies to a given query response, the 'Source' section title at the bottom of the user interface displays the corresponding answer generation pathway. Notably, for answers stemming from the MRC pathway, the system highlights supporting evidence sentences that were deemed relevant for satisfying the query. The user interface concepts corresponding to each of these three cases are visually depicted in Figure 6.

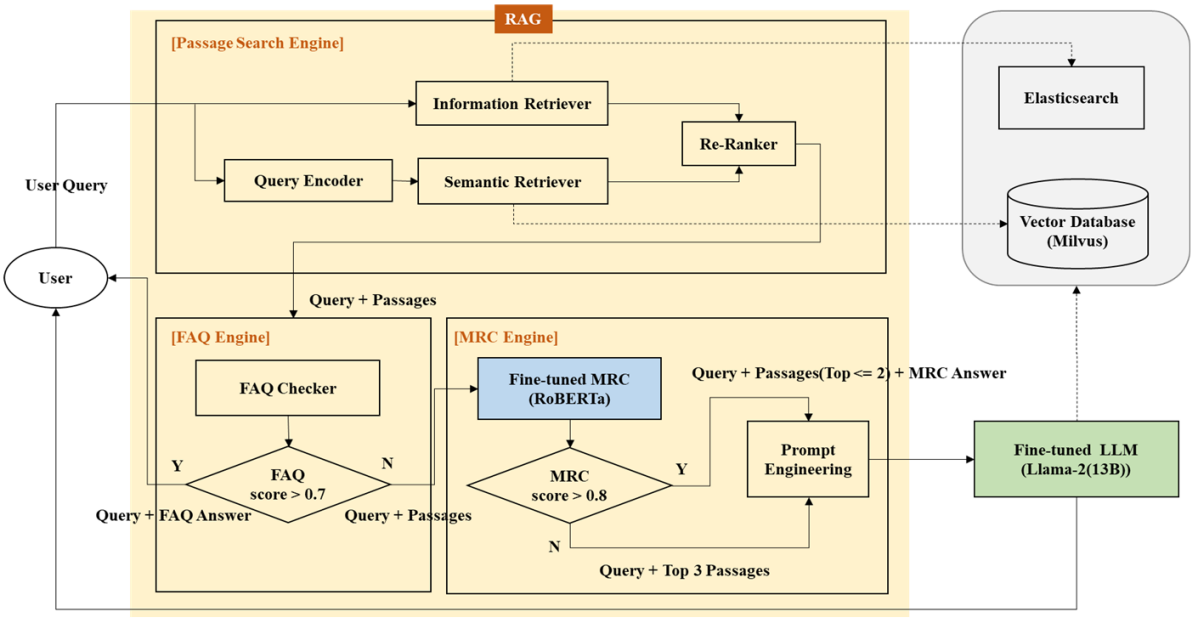

FIGURE 5. Pipeline architecture integrating RAG and small LLM

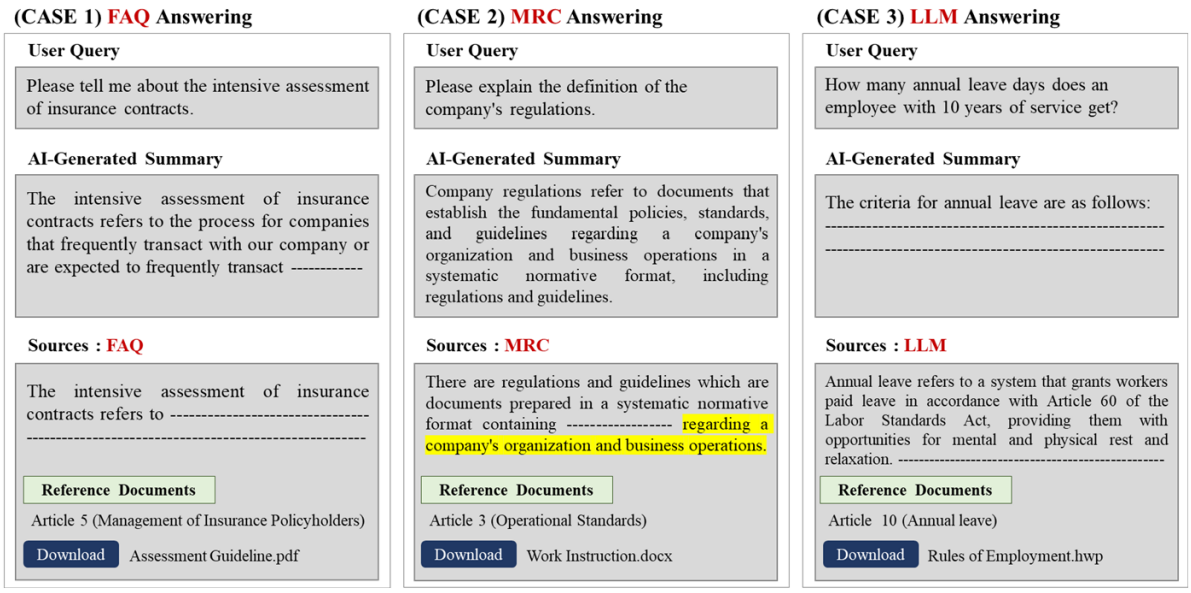

FIGURE 6. User interface representations for the three cases of query answering

3.3. **Performance evaluation.** In this study, we conducted evaluation using two distinct test datasets. The first test dataset comprised 200 question-answer pairs directly generated by task managers. The second test dataset consisted of 150 question-answer pairs extracted from 'new employee job training' exam questions spanning the past three years. Our chosen evaluation metric was accuracy, where incorrect answers received 0 points, partially correct answers received 0.5 points, and completely correct answers received 1 point. The evaluation results revealed a score of 92.7 points on the first test dataset and 79.3 points on the second test dataset. An illustrative example of the evaluation for the second dataset is provided in Figure 7.

In this study, we determined that human verification and accuracy assessment are more desirable for evaluating user-perceived performance in real-world applications. However,

No.	Question	Answer	AI-Generated Summary	Score
1	Please explain the definition of the company's regulations.	There are regulations and guidelines which are documents prepared in a systematic normative format containing fundamental policies, standards, etc. regarding a company's organization and business operations.	Company regulations refer to documents that establish the fundamental policies, standards, and guidelines regarding a company's organization and business operations in a systematic normative format, including regulations and guidelines.	1
2	What credit rating is applied to other companies that do not have sufficient credit evaluation materials?	abridged rating	For other companies that do not have sufficient credit evaluation materials, a simplified rating is applied.	0.5
…	…	…	…	…
150	What is the Information Protection Commission's deliberative resolution?	Information Technology Sector Plan pursuant to Article 21(4) of the Electronic Financial Transactions Act. ----------	A financial company or electronic financial institution shall establish and operate an information protection committee to deliberate and decide on matters related to the protection of important information.	0
Total Score				79.3

FIGURE 7. Example of evaluation results for the second dataset

as the dataset grows and more comprehensive performance evaluation is required, additional metrics such as F1 score, recall, and precision need to be assessed.

3.4. **Discussion.** Effectively leveraging small LLMs based on the RAG architecture with internal datasets for business applications requires addressing key considerations in data, model, and operational aspects.

From a data perspective, enhancing search accuracy demands meticulous domain-specific data curation and granularity, addressing overfitting, automating preprocessing, and managing continuous dataset updates. It is advisable to exclude frequently changing data, such as exchange rates, from the training datasets and handle it separately. Above all, the significant discrepancy in accuracy observed between the two datasets in this study highlights the complexities inherent in adapting generative AI models to specific domains and underscores the need for targeted improvements to ensure the system's suitability for mission-critical internal tasks. The observed variance in accuracy indicates that ongoing refinement is necessary to enhance the model's performance and reliability. To achieve more consistent accuracy, it is crucial to establish robust data governance suitable for the training and evaluation of generative AI models. In this study, considerable time and effort were expended on refining the data for model training. A structured data governance approach will reduce repetitive data processing tasks, allowing us to focus more on enhancing data quality and improving accuracy. This will ultimately lead to more consistent results in the training and evaluation of generative AI models. By addressing these challenges and incorporating these insights, we can significantly enhance the practical utility and reliability of our RAG-enhanced small LLM system for business applications.

Regarding the model, maintaining consistent performance requires selective fine-tuning with new data and establishing a validation process that enables swift replacement with new state-of-the-art models. Achieving natural responses necessitates parameter tuning, including reinforcement learning and token adjustments.

From an operational standpoint, establishing LLMOps (Large Language Model Operations) is crucial for efficient pipeline operations. Optimizing GPU resource allocation, implementing strict security measures, and enhancing quality through feedback loops like RLHF (Reinforcement Learning with Human Feedback) are essential. Additionally, user-friendly features such as the ability to directly navigate to the location of the source material used as the basis for answers can enhance usability.

4. **Conclusion.** In this study, we deployed an on-premises NVIDIA A100 server to create a dedicated infrastructure environment. By fine-tuning RoBERTa and Llama 2 models

with our internal dataset, we developed an MRC model and a small LLM. Our pipeline integrates the RAG architecture with the small LLM, and we evaluated the system's accuracy using our proprietary dataset. However, our evaluation revealed a significant discrepancy of approximately 13 points between the accuracy scores of the two test datasets. In this paper, we focus on the applicability of the proposed methodology to business contexts. However, to achieve a comprehensive performance evaluation, it is necessary to compare the performance with other models and assess the impact of the RAG component. This limitation highlights the need for further research.

Acknowledgment. This work is supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0017123, The Competency Development Program for Industry Specialist).

REFERENCES

[1] M. A. Arefeen, B. Debnath and S. Chakradhar, LeanContext: Cost-efficient domain-specific question answering using LLMs, *Natural Language Processing Journal*, vol.7, 100065, 2024.

[2] P. Lewis, E. Perez et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems*, vol.33, pp.9459-9474, 2020.

[3] G. Izacard and E. Grave, Leveraging passage retrieval with generative models for open domain question answering, *arXiv Preprint*, arXiv: 2007.01282, 2020.

[4] K. Guu et al., REALM: Retrieval-augmented language model pre-training, *Proc. of the 37th International Conference on Machine Learning*, pp.3929-3938, 2020.

[5] V. Karpukhin et al., Dense passage retrieval for open-domain question answering, *arXiv Preprint*, arXiv: 2004.04906, 2020.

[6] S. Sharma et al., Retrieval augmented generation for domain-specific question answering, *arXiv Preprint*, arXiv: 2404.14760, 2024.

[7] S. Siriwardhana et al., Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering, *Transactions of the Association for Computational Linguistics*, vol.11, pp.1-17, 2023.

[8] B. Han, T. Susnjak and A. Mathrani, Automating systematic literature reviews with retrieval-augmented generation: A comprehensive overview, *Applied Sciences*, vol.14, no.19, 9103, 2024.

[9] S. Wu et al., Retrieval-augmented generation for natural language processing: A survey, *arXiv Preprint*, arXiv: 2407.13193, 2024.

[10] E. J. Hu et al., LoRA: Low-rank adaptation of large language models, *arXiv Preprint*, arXiv: 2106.09685, 2021.

[11] Y. Gao et al., Retrieval-augmented generation for large language models: A survey, *arXiv Preprint*, arXiv: 2312.10997, 2023.

[12] Nurjayanti, Adiwijaya and S. Al Faraby, Learning contextual meaning for question retrieval using Siamese LSTM on Islamic question answering system, *ICIC Express Letters*, vol.16, no.9, pp.1011-1017, 2022.

[13] G. Izacard and E. Grave, Leveraging passage retrieval with generative models for open domain question answering, *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp.874-880, 2021.

[14] N. Dziri et al., Retrieval augmentation reduces hallucination in conversation, *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*, pp.1579-1595, 2022.

[15] V. Karpukhin et al., Dense passage retrieval for open-domain question answering, *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp.6769-6781, 2020.

[16] Q. Liu et al., Semantic matching in machine reading comprehension: An empirical study, *Information Processing & Management*, vol.60, no.2, 103145, 2023.

[17] H. Touvron et al., Llama 2: Open foundation and fine-tuned chat models, *arXiv Preprint*, arXiv: 2307.09288, 2023.

[18] *NVIDIA DGX A100 Datasheet*, https://images.nvidia.com/aem-dam/Solutions/Data-Center/nvidia-dgx-a100-datasheet.pdf, Accessed on February 7, 2025.

[19] Y. Liu et al., RoBERTa: A robustly optimized BERT pretraining approach, *arXiv Preprint*, arXiv: 1907.11692, 2019.

Author Biography

Youngpyo Hong joined the Department of Industrial and Information Systems Engineering at Soongsil University in Korea as a Ph.D. student in the spring of 2022. He earned his M.S. degree in Information Management from the KAIST College of Business in Korea in 2015.

His primary research interests focus on developing AI architectures for the efficient application and utilization of AI services in business, as well as optimizing the operations and processes of AI services.

Dongsoo Kim is a professor in the Department of Industrial and Information Systems Engineering, Soongsil University, Korea. He received the B.S., M.S. and Ph.D. degrees in Industrial Engineering from Seoul National University (SNU), Korea, in 1994, 1996 and 2001, respectively. He had been a team leader of e-business standard team at National Computerization Agency. After that he had worked as an assistant professor of the Graduate School of Healthcare Management and Policy at the Catholic University of Korea.

He served as the president of the Society for e-Business Studies, and served as the chair of the BPI (Business Process Innovation) working group of the Korean Institute of Industrial Engineering. His research interests include industrial intelligent system, business process analytics, healthcare process management, and information security management.