

Enhancing Radiology Report Interpretation with Large-Scale Language Models: A Two-Stage Fine-Tuning Approach

Sun-Woo Pi *, Myeong-Soo Park**, Byoung-Dai Lee*,**

* Department of Computer Science, Graduate School, Kyonggi University, Suwon, Republic of Korea

** Division of AI and Computer Engineering, Kyonggi University, Suwon, Republic of Korea

ppssww1202@kyonggi.ac.kr, watermelon@kyonggi.ac.kr, blee@kyonggi.ac.kr

Abstract— Radiology reports contain complex medical terminology and specialized knowledge, making them difficult for both patients and medical professionals to interpret. This study aims to address this challenge by developing a large-scale language model specifically designed for interpreting chest radiology reports. We focus on four key natural language processing (NLP) tasks—summarization, paraphrasing, abbreviation interpretation, and question answering—using a synthetic dataset derived from the MIMIC-CXR reports and GPT-3.5 Turbo. To enhance the model's performance, we propose a two-stage supervised fine-tuning (SFT) process, incorporating real-world medical data from PubMedQA and MedQA, in addition to the synthetic dataset. The resulting models, Model-1 and Model-2, were evaluated based on accuracy, conciseness, and clarity, using test data not seen during training. Experimental results demonstrated that the proposed two-stage SFT method achieved strong performance across all four tasks, providing comparable performance to models such as GPT-3.5, Bard, Llama2, and MedAlpaca in key evaluation metrics, despite using a relatively smaller number of parameters. These findings suggest that synthetic data, when combined with domain-specific datasets, can significantly improve the interpretive capabilities of large-scale language models in the medical domain.

Keywords— Large-Scale Language Model, Radiology Report, Self-Supervised Finetuning, Synthetic Data, Prompt Engineering

I. INTRODUCTION

Radiology reports are documents in which radiologists gather and analyze patient information, including medical history, symptoms, and imaging results and evaluations based on medical images such as X-rays, CT scans, and MRIs. These reports play a crucial role in diagnosing the patient's condition and establishing treatment plans. However, numerous studies have shown that these reports contain complex medical terminology and specialized knowledge, making them difficult for the public to understand. Even medical professionals often require additional interpretation and consultation to fully comprehend the reports [1]. The difficulty is particularly evident in chest radiology reports, which include complex medical findings related to various anatomical structures such as the lungs, heart, and ribs. This further emphasizes the need for methods that can efficiently and accurately analyze these reports.

Recently, there have been various attempts to apply artificial intelligence (AI), particularly natural language processing (NLP) technologies, to the medical field. Large-scale language models (LLMs) such as GPT-3 [2] and Llama [3] have demonstrated the ability to understand and generate human language by learning from vast amounts of text data. However, because these models are trained on general domain data, they often fail to fully reflect the specificity and expertise required in the medical domain.

Therefore, this study aims to develop a large-scale language model specifically designed to perform four natural language processing tasks—summarization, paraphrasing, abbreviation interpretation, and question answering—targeting chest radiology reports. For these tasks, training data consisting of (report, question, answer) pairs are required, but challenges such as data privacy concerns, limited access to medical data, and the high costs associated with expert annotation make it difficult to construct large-scale datasets. This makes it difficult to acquire sufficient high-quality data required for deep learning models, especially for highly specialized texts like radiology reports.

To overcome these limitations, this study generates synthetic data based on radiology reports using the GPT-3.5 Turbo [2], following the methodology presented by Wang et al. [4]. The synthetic data is used to apply a Supervised Fine-Tuning (SFT) approach to large-scale language models, thereby enhancing the accuracy and reliability of the model's medical knowledge.

II. RELATED WORK

Various studies have explored methods of generating synthetic data to improve the performance of NLP models. Among them, Wang et al.'s self-instruct approach [4] leveraged large-scale language models like GPT-3 to autonomously generate questions and answers, effectively creating synthetic data across different domains. This self-instruct method helped overcome the limitations of data collection by allowing the model to generate and utilize knowledge independently. However, this approach also has limitations in terms of the reliability and diversity of the generated data. In this study, we propose a method for generating synthetic question-answer pairs using the MIMIC-CXR [5] radiology reports. Since the

data is generated based on real medical records, the resulting question-answer pairs are highly relevant and meaningful from a medical perspective. Moreover, traditional self-instruct methods risk reflecting the biases inherent in the data on which the language model was trained, potentially resulting in inaccurate answers in certain fields. Our proposed method improves reliability by utilizing real-world radiology reports based on diverse cases involving various patients.

In previous research, the development of domain-specific medical language models has been widely researched, yet limitations remain in fully capturing the complexity of certain medical domains. For example, Zhang et al. [6] fine-tuned the LLaMA model to develop MedAlpaca, a medical-specific language model, but encountered limitations in fully capturing the complexity of specialized medical texts, such as chest radiology reports. This is due to the model's insufficient understanding of the specialized terminology and complex context within radiology reports, preventing it from delivering the precise interpretations necessary for clinical judgement. In this study, we address these limitations by utilizing domain-specific medical datasets and synthetic datasets based on radiology reports, thereby improving the model's performance in interpreting radiology reports.

III. PROPOSED METHOD

A. Datasets

The MIMIC-CXR dataset is a large-scale, publicly available medical dataset containing chest X-ray images and corresponding radiology reports, consisting of over 370,000 chest radiographs and more than 200,000 associated reports. This dataset is provided to researchers for use in medical image analysis and natural language processing (NLP) studies. PubMedQA [7] is a question-answering dataset based on biomedical literature, while MedQA [8] is a question-answering dataset derived from United States Medical Licensing Examination (USMLE) questions. The PubMedQA and MedQA datasets were employed to enhance the coverage and accuracy of the medical knowledge in the pre-trained large language model, and the MIMIC-CXR dataset was used for generating synthetic data for individual NLP tasks.

B. Development of a Large-Scale Language Model for Radiology Report Interpretation

Figure 1 illustrates the development process of the proposed large-scale language model based on synthetic data for radiology report interpretation. In the data preprocessing stage, key sections of the MIMIC-CXR reports—such as History, Comparison, Findings, and Impression—are used to generate comprehensive report data by integrating critical information regarding patient history, medical symptoms, imaging findings, and diagnoses. Subsequently, the process of generating questions for the synthetic dataset comprising (report, question, answer) pairs used in SFT was divided into two stages: the preparation stage and the actual generation stage. In the preparation stage, 200 radiology reports containing integrated key information were selected, and the GPT-3.5 Turbo model was used to generate 50 sample questions for each of the four

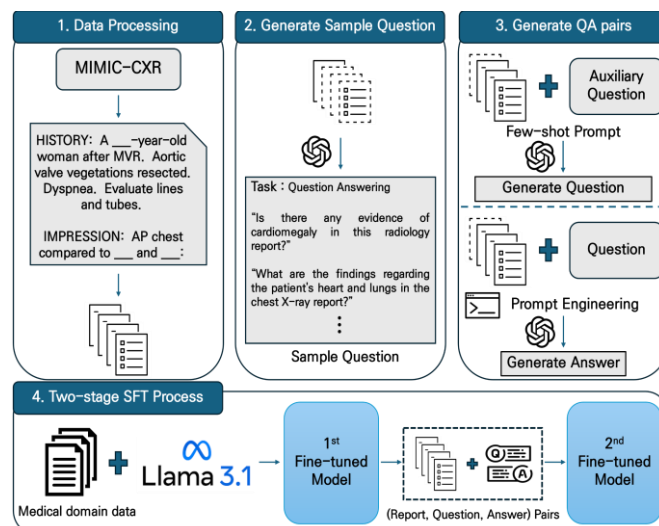


Figure 1. Development process of a large-scale language model using synthetic data

tasks—summarization, paraphrasing, abbreviation interpretation, and question answering. The generated sample questions were evaluated for task relevance and to ensure there was no redundancy among the questions. From these, the top five sample questions for each task were selected. In the actual generation stage, for the remaining radiology reports, a few-shot prompt method was employed. This method provided five auxiliary questions alongside each radiology report to guide the model in generating appropriate questions. These auxiliary questions consisted of three randomly selected from the top five sample questions in the preparation stage, and two randomly selected from the questions generated during the actual generation stage. Additionally, during this phase, one of the four tasks was randomly assigned to each radiology report to ensure the generated questions were task-specific, while maintaining an even distribution of tasks across the dataset. This process resulted in the generation of approximately 160,000 questions in total.

Figure 2 illustrates an example of the few-shot prompting method, where auxiliary questions are provided alongside radiology reports to guide question generation. To generate responses for each question, the previously generated questions and radiology reports were provided as input to the model, which then produced the corresponding answers. Prompt engineering was applied to ensure that the responses met the key evaluation criteria of accuracy, conciseness, and clarity. For instance, instructions such as “The answer must be accurate.”, “The answer should be as concise as possible.”, and “The answer should be understandable to both medical professionals and non-experts.” were added to the input to generate high-quality data consistent with the research objectives. Additionally, given that providing inaccurate information in the medical field poses significant risks to patient safety and health, instructions like “If an accurate answer cannot be provided, do not generate a response and explain why it is difficult to answer.” were incorporated to prevent the model from generating incorrect information. These guidelines were designed to enhance the reliability of responses

Question Generation Prompt

As a healthcare professional, you frequently employ a clinical language model to conduct thorough analyses of radiology reports. You ask the model some questions based on the radiology report, and get answers that helps your decision-making.

Let's suppose you want to ask a question about the task of "[task]" related to the given radiology report. What specific question would you formulate?

[Radiology Report Begin] [report] [Radiology Report End]

The subsequent questions serve as examples for reference. You can freely refer to the examples, but try not to just replicate them. Your question should be answerable solely based on the information available in the radiology report. Your question must be an appropriate question for the task presented above.

[Auxiliary Questions Begin] [auxiliary question] [Auxiliary Questions End]

I encourage you to generate a question that is clear, novel, and analytically important. You must output solely a question.

Figure 2. Example of the few-shot prompting method with auxiliary questions for question generation

by guiding the model to avoid answering, with a clear explanation, in situations where hallucinations could lead to generating non-existent or false information. The synthetic dataset comprising (report, question, answer) pairs generated for each task was then used in the SFT process of the pre-trained large-scale language model.

In this study, we adopted the pre-trained Llama3.1-8B-Instruct [3] model as the base model for radiology report interpretation. During training process of the model, radiology reports and their associated questions were provided as input to generate answers as output. Cross-entropy loss was applied to quantify the difference between the probability distribution of the model's output and the actual labels, which was used to optimize the model's parameters. This model, fine-tuned using the synthetic dataset of (report, question, answer) pairs, is referred to as Model-1.

When applying SFT to a pre-trained large language model using only radiology report-based synthetic data, the model can suffer from performance degradation due to a narrow focus on the highly specialized content of radiology reports, without sufficient knowledge of the broader medical domain. To address this limitation, we propose a two-stage fine-tuning process. In the first stage, an initial SFT was performed on the Llama3.1 model using the MedQA and PubMedQA datasets, specializing the model in the medical domain and enabling it to more effectively understand basic medical knowledge and the context of question-answering tasks. In the second stage, an additional SFT was conducted using the synthetic dataset based on radiology reports to further enhance the model's performance. This two-stage fine-tuning process significantly improved the model's ability to interpret chest radiology reports. The model resulting from this process is referred to as Model-2.

IV. RESULTS

Figure 3 shows an example of (question, answer) pairs generated for each task based on a specific radiology report. The four natural language processing (NLP) tasks—summarization, paraphrasing, abbreviation interpretation, and question answering—were selected to enable the model to effectively utilize and interpret the information in radiology reports. Summarization is the task of concisely summarizing

Summarization

Question : What are the main findings in the radiology report concerning the patient's lung and cardiac conditions?

Answer : The main findings in the radiology report indicate that the patient's lung condition has deteriorated. There is worsening interstitial edema, unresponsive right upper lobe consolidation possibly due to bronchial obstruction, and moderate bilateral pleural effusions. Additionally, the report notes cardiomegaly and mediastinal vascular engorgement, which suggest worsening heart function or cardiac decompensation.

Paraphrasing

Question : How would you rephrase the statement "Previous mild interstitial edema has worsened generally" into a more understandable form for a patient?

Answer : The fluid buildup in your lungs, which was previously mild, has now become more pronounced. This means there is more swelling in the lungs than before. It's important for your treatment to know this change, as it may affect your breathing.

Abbreviation Interpretation

Question : What does the abbreviation 'AP' stand for in the context of the radiology report?

Answer : In the context of the radiology report, 'AP' stands for "anteroposterior." This refers to the direction in which the X-ray is taken, specifically from the front (anterior) of the body to the back (posterior). This positioning helps in visualizing the structures within the chest, such as the lungs and heart, more clearly.

Question Answering

Question : What does the report indicate about the status of the right upper lobe consolidation and its possible causes?

Answer : The report indicates that the right upper lobe consolidation has not improved and now shows signs of volume loss, which suggests that there may be a blockage in the bronchial tubes. This blockage could be due to either an infection or bleeding within the lung. Overall, the findings point to a worsening condition that requires further evaluation and management.

Figure 3. (question, answer) pairs generated for each task from a specific radiology report

and analyzing the key content of a radiology report. For example, when the model receives the question "What are the main findings in the radiology report concerning the patient's lung and cardiac conditions?", it generates a concise summary of the key findings related to the lungs and heart, as shown in Figure 3. Paraphrasing is the task of converting the specialized medical terminology in the radiology report into simpler language that patients can understand. For example, when given a statement like "How would you rephrase the statement 'Previous mild interstitial edema has worsened generally' into a more understandable form for a patient?", the model rephrases it into a more understandable form for patients. Abbreviation interpretation is the task of explaining the meaning of commonly used abbreviations in radiology reports. Figure 3 illustrates this by showing a question such as "What does the abbreviation 'AP' stand for in the context of the radiology report?" with the model generating a response that explains the term behind the abbreviation and its meaning. Lastly, question answering involves responding to questions based on the status and causes of key lesions described in the radiology report. For instance, when asked "What does the report indicate about the status of the right upper lobe consolidation and its possible causes?", the model generates a

response that explains the status of the lesion and possible causes based on the report, as shown in Figure 3.

LLMs are used across various domains, and the evaluation criteria may vary depending on the specific characteristics and requirements of each field, making it challenging to accurately measure model performance with general evaluation metrics. In particular, using standard NLP metrics such as text similarity or BLEU-Score [9] to assess the validity of radiology report interpretation has limitations in evaluating clinical accuracy and relevance. Therefore, qualitative evaluation by experts is crucial to ensure clinical accuracy and appropriateness to patient cases. However, it is generally very difficult for radiologists to directly participate in evaluation tasks. Accordingly, in this study, we relied on the expertise of GPT-4 [2] to evaluate the answers generated by the model based on the given radiology report and questions.

The evaluation was conducted based on three criteria: accuracy, conciseness, and clarity, with each criterion being defined in prompts and scored on a scale of 1 to 4 [10]. The scores were assigned as follows: Accuracy was evaluated based on whether the response contained inaccurate or unnecessary information; conciseness, based on whether any parts could be omitted or summarized more briefly; and clarity, based on whether explanations were provided for complex medical terms. The test data used for the evaluation consisted of 50 (radiology report, question, answer) pairs that were not part of the training set, and these pairs were selected for each task. The results are presented in Tables I to V.

Table I shows that Model-1 and Model-2 demonstrated strong performance in the key evaluation metrics of accuracy

TABLE I. PERFORMANCE COMPARISON BY TASK: SUMMARIZATION

	Accuracy	Conciseness	Clarity
Llama2 [3]	3.88	3.04	3.49
Bard [11]	3.98	2.51	4.0
MedAlpaca [6]	3.47	3.53	2.57
GPT-3.5 [2]	3.9	3.73	3.24
Model-1	3.76	3.27	3.08
Model-2	3.94	3.53	3.37

TABLE II. PERFORMANCE COMPARISON BY TASK: PARAPHRASING

	Accuracy	Conciseness	Clarity
Llama2 [3]	3.78	3.09	3.65
Bard [11]	3.98	2.83	3.98
MedAlpaca [6]	2.5	2.44	2.13
GPT-3.5 [2]	3.83	3.5	3.54
Model-1	3.7	3.04	3.5
Model-2	3.61	2.89	3.37

TABLE III. PERFORMANCE COMPARISON BY TASK: ABBREVIATION INTERPRETATION

	Accuracy	Conciseness	Clarity
Llama2 [3]	3.02	2.96	3.15
Bard [11]	3.81	2.75	3.92
MedAlpaca [6]	2.96	3.13	2.54
GPT-3.5 [2]	3.92	3.69	3.88
Model-1	3.13	2.83	3.08
Model-2	3.52	3.0	3.56

TABLE IV. PERFORMANCE COMPARISON BY TASK: QUESTION ANSWERING

	Accuracy	Conciseness	Clarity
Llama2 [3]	3.75	2.98	3.53
Bard [11]	3.91	2.46	3.95
MedAlpaca [6]	3.42	3.54	2.77
GPT-3.5 [2]	4.0	3.67	3.58
Model-1	3.88	3.3	3.51
Model-2	3.93	3.32	3.54

and conciseness for the summarization task. Notably, Model-2 outperformed GPT-3.5 [2] in terms of accuracy and clarity, and also showed superior performance in conciseness compared to Bard, which tended to generate longer responses.

Table II indicates that for the paraphrasing task, both Model-1 and Model-2 demonstrated strong performance in accuracy and clarity, while also achieving above-average scores in conciseness. Bard [11] scored well in both accuracy and clarity but received a lower score in conciseness due to generating relatively longer responses. MedAlpaca [6], on the other hand, showed generally lower performance across all metrics.

For the abbreviation interpretation task, accurate interpretation and explanation of abbreviations is essential, making accuracy and clarity critical evaluation metrics. This task is particularly important in radiology reports, which often contain numerous specialized terms. According to Table III, both Model-1 and Model-2 outperformed MedAlpaca across all metrics, and Model-2, in particular, achieved comparable scores to Bard and GPT-3.5, despite having fewer parameters.

For the question answering task, providing accurate explanations of the information requested by the question is important. Table IV shows that both Model-1 and Model-2 demonstrated performance similar to GPT-3.5 and Bard in terms of accuracy. Notably, Model-2 outperformed Bard by a small margin in accuracy and also ranked highly in conciseness and clarity. This indicates that Model-2 not only generates accurate answers but also produces responses that are concise and easy to understand.

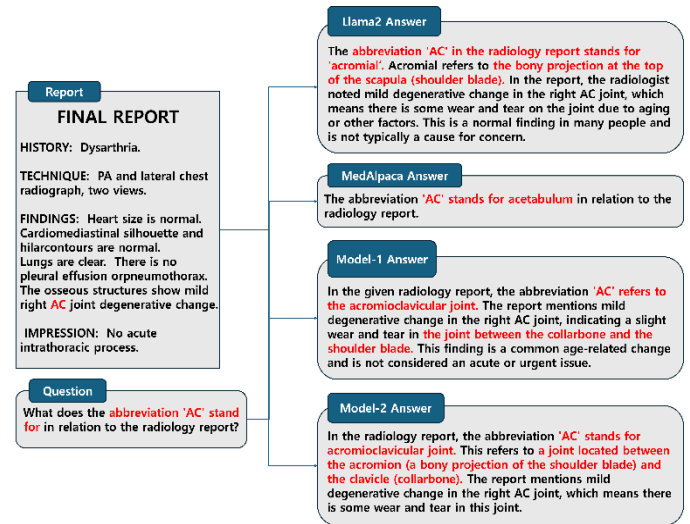
TABLE V. AVERAGE PERFORMANCE SCORES ACROSS THE EVALUATION TASKS

	Accuracy	Conciseness	Clarity
Llama2 [3]	3.6	3.02	3.31
Bard [11]	3.92	2.63	3.96
MedAlpaca [6]	3.11	3.19	2.52
GPT-3.5 [2]	3.92	3.65	3.56
Model-1	3.63	3.12	3.32
Model-2	3.76	3.2	3.46

Table V shows the average performance scores across all test data used for evaluation. In terms of accuracy, Model-1 outperformed both Llama2 and MedAlpaca, while Model-2 demonstrated exceptional performance, despite the significantly fewer parameters of the Llama 3.1 model compared to GPT-3.5 and Bard. Regarding conciseness, both Model-1 and Model-2 were designed through prompt engineering to generate responses as concise as possible, showing overall superior performance compared to other models. Clarity, a crucial factor for both experts and patients, and Model-2 demonstrated performance close to that of GPT-3.5, ranking among the top. When examining the four tasks in detail, the SFT-applied models demonstrated strong performance across all tasks. The models scored highly on key evaluation metrics specific to each task, indicating that the models effectively understood the radiology reports in a task-appropriate manner.

Figure 4 shows the question and the corresponding responses from Llama2, MedAlpaca, Model-1, and Model-2 for an abbreviation interpretation task based on a specific radiology report. In this example, we compared the responses from Llama2 and MedAlpaca, both with 7B parameters, and Model-1 and Model-2, both with 8B parameters. Despite the minimal difference in the number of parameters, the performance gap between the models is evident, as shown in Figure 4. Both Model-1 and Model-2 correctly defined 'AC' as the 'acromioclavicular joint' and provided accurate descriptions of its location and condition. In contrast, MedAlpaca incorrectly interpreted 'AC' as acetabulum, which is related to the hip joint and irrelevant to chest X-rays, without providing any further explanation. Llama2 defined 'AC' as a specific part of the shoulder bone adjacent to the joint. While this is not entirely incorrect, it is an inaccurate answer. These results highlight the superior answering ability of both Model-1 and Model-2, with Model-2 offering more detailed explanations about the 'acromion' and 'clavicle', as well as the joint's location, compared to Model-1.

Overall, as demonstrated by these results, Model-2, which underwent the two-stage SFT process, outperformed Model-1, which used only synthetic data for SFT, in both quantitative and qualitative evaluations. This validates the effectiveness of the two-stage SFT methodology proposed in this study.

**Figure 4.** Responses from Llama2, MedAlpaca, Model-1, and Model-2 for the abbreviation interpretation task

V. CONCLUSIONS

In this study, we developed and evaluated a large-scale language model for interpreting radiology reports. By applying an SFT methodology using synthetic data to a pre-trained large-scale language model, we improved the model's performance in interpreting radiology reports. Notably, Model-2, developed through a two-stage SFT process, demonstrated superior performance across four key tasks—summarization, paraphrasing, abbreviation interpretation, and question answering—outperforming existing models in terms of accuracy, conciseness, and clarity, as verified through evaluations using GPT-4. One particularly noteworthy result is that Model-2 achieved performance comparable to larger models like GPT-3.5 and Bard, despite having significantly fewer parameters. These findings demonstrate the effectiveness of the synthetic data-based SFT methodology proposed in this study for interpreting radiology reports. Additionally, this study structured the synthetic data generation process based on MIMIC-CXR radiology reports, highlighting the potential to overcome the challenges associated with the scarcity of medical data. Future research will focus on applying this approach to a wider range of medical datasets and tasks to enhance the model's generalizability and expand its clinical applicability.

ACKNOWLEDGMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2020R1A6A1A03040583). This research was also supported by the MSIT(Ministry of Science and ICT), Korea, under the ICAN(ICT Challenge and Advanced Network of HRD) support program(IITP-2024-00436954) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

- [1] W. Kim, S. Hwang, M. Ahn, Y. Jung, and J. Jung, "Radiation Reading Results: Cancer Patients' Attitudes, Preferred Types and Opinions," *Journal of the Korean Radiological Society*, vol. 4, no. 3, 2001.
- [2] OpenAI, "ChatGPT." [Online]. Available: <https://openai.com/chatgpt/>. [Accessed: Oct. 11, 2024].
- [3] Meta, "Llama." [Online]. Available: <https://www.llama.com/>. [Accessed: Oct. 11, 2024].
- [4] Y. Wang, Y. Kordi, S. Mishra, P. Liu, N. Smith, and D. Khoshnab, "Self-Instruct: Aligning Language Models with Self-Generated Instructions," *arXiv preprint arXiv:2212.10560*, 2022.
- [5] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR: A Large Publicly Available Database of Labeled Chest Radiographs," *Nature Scientific Data*, vol. 6, 2019.
- [6] S. Zhang, Y. Gu, Q. Liu, W. Wang, M. Li, and J. Wu, "MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data," *arXiv preprint arXiv:2305.14568*, 2023.
- [7] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A Dataset for Biomedical Research Question Answering," *arXiv preprint arXiv:1909.06146*, 2019.
- [8] J. Zeng, X. Liu, S. Lai, H. Zhou, and Z. Lin, "MedQA: A Large-Scale Medical Question Answering Dataset," *arXiv preprint arXiv:2005.06411*, 2020.
- [9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311-318.
- [10] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.
- [11] Bard. "Bard." [Online]. Available: <https://bard.google.com/>. [Accessed: Oct. 11, 2024].



Sun-Woo Pi received B.S. degree in Division of AI and Computer Engineering from Kyonggi University, Suwon, Republic of Korea in 2024. He is currently pursuing the M.S. degree in the Department of Computer Science at Kyonggi University, Suwon, Republic of Korea. His research interests include deep learning, computer vision, and medical image analysis.



Myeong-Soo Park is currently a third-year undergraduate student in the Division of AI & Computer Engineering at Kyonggi University, Suwon, Republic of Korea in 2024. His current research interests include natural language processing.



Byoung-Dai Lee received the B.S. and M.S. degrees in computer science from Yonsei University, South Korea, in 1996 and 1998, respectively, and the Ph.D. degree in computer science and engineering from the University of Minnesota, Minneapolis, USA, in 2003. From 2003 to 2010, he worked with Samsung Electronics, Company Ltd., as a Senior Engineer. He is currently a Full Professor with the Division of Computer Science and Engineering, Kyonggi University, South Korea. His research interests include machine learning, deep learning, and medical image analysis.