Towards Client Engagement Using RAG System with Pattern Prediction Framework

Hanmin Jung[1,2] and Athiruj Poositaporn[1,2]
[1]*University of Science and Technology, 217, Gajeong-ro, Yuseong-gu, Daejeon, Gyeonggi-do, Republic of Korea*
[2]*Korea Institute of Science and Technology Information, 245, Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea*

Keywords: Client Engagement, Retrieval-Augmented Generation, Large Language Model, Q&A System.

Abstract: Client engagement refers to the process of companies and customers building and maintaining relationships through communication, personalized marketing, and value-added services. This often results in analysis reports, consulting services, and strategic planning documents. Tools like GPT-4o have significant potential to support these interactions in sectors such as meteorological organizations. However, standalone generative models like GPT-4o face challenges in accessing external datasets and often produce generic outputs. To overcome these limitations, this study introduces a chat-based Retrieval-Augmented Generation (RAG) system integrated with a pattern prediction framework. We demonstrate our RAG system in analyzing air pollution pattern prediction results from our prior study and compare its generated answers with a standalone GPT-4o model. Experimental results show that the RAG system delivers actionable recommendations and contextually enriched outputs grounded in domain-specific data. In future work, we aim to explore the potential of RAG in real-world applications, such as improving client engagement by generating client-focused reports.

1 INTRODUCTION

Client engagement refers to all activities performed by companies and customers to build and maintain meaningful relationships (Rana et al., 2020). This involves a wide range of interactions, including communication through various channels, personalized marketing efforts, and the provision of value-added services. The outputs of these activities often include analysis reports, consulting services, and strategic planning documents that cater to the specific needs of clients.

One area where client engagement plays a critical role is in organizations that deal with complex data analysis and reporting, such as meteorological organizations. These organizations face unique challenges in interpreting and disseminating large volumes of environmental data, which is crucial for addressing issues like air quality monitoring and pollution control. The ability to provide stakeholders with tailored, data-driven insights through real-time Q&A systems and strategy reports can significantly enhance their decision-making processes. However, existing systems such as GPT-4o may not always include the latest or most relevant information. This limitation results in generalized outputs that lack specificity and context. Moreover, their limitation to access external datasets restricts their capacity to provide precise answers and data-intensive tasks (Afzal et al., 2023; Han et al., 2024; Shahriar et al., 2024).

Retrieval-Augmented Generation (RAG) addresses these issues by integrating a retrieval mechanism with generative capabilities. RAG systems can access external databases or documents in real time, enabling more contextually accurate and evidence-based responses. This feature significantly enhances their utility for data analysis in the scenarios where insights must be derived from specific datasets (Chen et al., 2024). This highlights RAG as a transformative tool for enhancing decision-making processes across various industries (Ranade et al., 2024).

This study presents the development of a chat-based RAG system integrated with a pattern prediction framework. We design our RAG system to improve client engagement by providing data-driven insights that are customized to meet the requirements of a specific domain. The system combines the retrieval capabilities of RAG with GPT-4o for

Jung, H. and Poositaporn, A.
Towards Client Engagement Using RAG System with Pattern Prediction Framework.
DOI: 10.5220/0013474400003944

In *Proceedings of the 10th International Conference on Internet of Things, Big Data and Security (IoTBDS 2025)*, pages 436-441
ISBN: 978-989-758-750-4; ISSN: 2184-4976

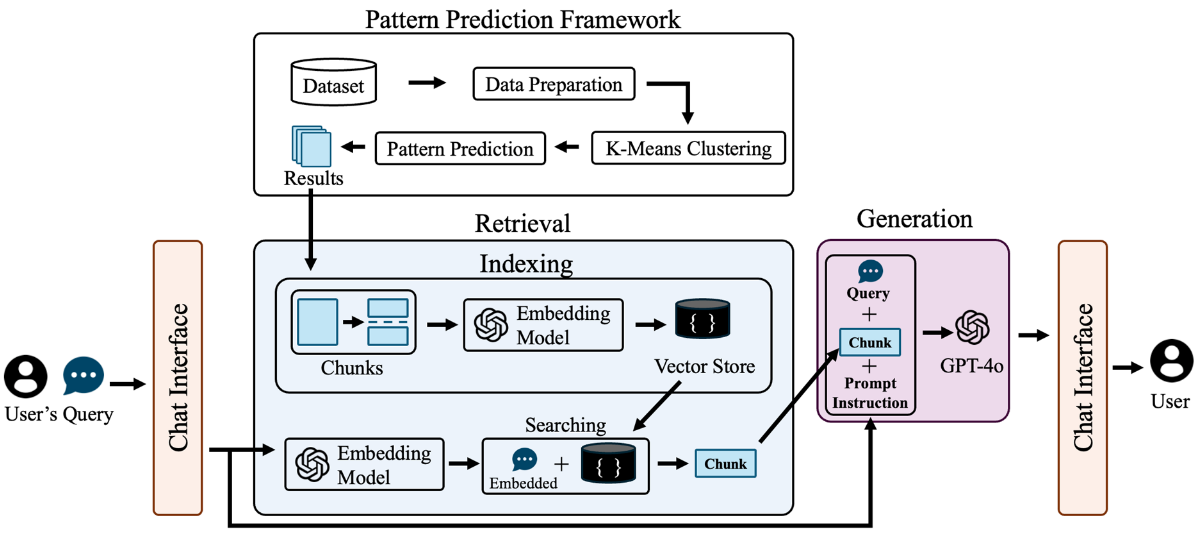

Figure 1: Process pipeline of our RAG system with pattern prediction framework.

generating more accurate and contextually relevant responses. We demonstrate its application in analyzing air pollution pattern prediction results from our prior study and generating actionable recommendations (Poositaporn & Jung, 2025). In the experiment, we compare our RAG system and a standalone GPT-4o model to evaluate their performance. We aim to highlight the advantages of the RAG system in enhancing client engagement through data analysis and explore its broader implications for real-world applications.

This study is organized as follows: Section 2 provides an overview of our RAG system with a pattern prediction framework. Section 3 details the development of the proposed RAG system with an air pollution pattern module. Section 4 presents a comparative analysis of the system's performance against a standalone GPT-4o model and discusses the implications of these findings for client engagement. Section 5 concludes with a summary of key insights and potential directions for future work.

2 RAG SYSTEM WITH PATTERN PREDICTION FRAMEWORK

In this study, we develop a chat-based interface for our RAG system that includes a pattern prediction framework (see Figure 2). The descriptions and process flows are described as follows.

The pattern prediction framework serves as an additional module that applies K-means clustering to group vector data into meaningful clusters based on key features obtained from the input data. These clusters are then used to predict future patterns by applying advanced predictive models.

The RAG system takes the results of the Pattern Prediction Framework and utilizes them in a retrieval process. As shown in figure 1, the RAG system consists of two main stages: retrieval and generation. The RAG system begins with an indexing process where the data generated by the framework are processed and stored in a vector database. Once the user's query is received, the retrieval stage retrieves relevant information from the vector database. The generation stage then uses this retrieved data and passes it to the GPT-4o[1] model to generate the final results. The final results are then presented to the user via a chat-based interface, as shown in Figure 2.

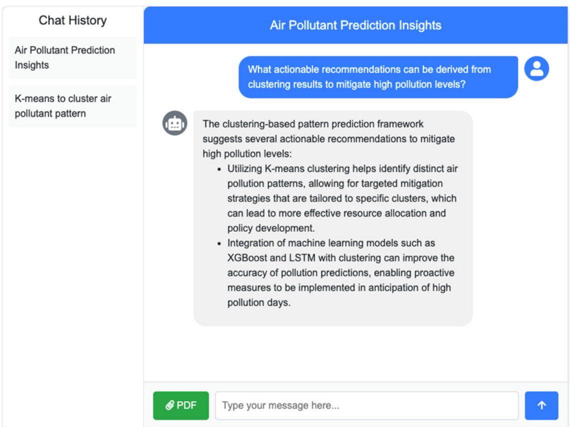

Figure 2: Chat interface of our RAG system.

[1] https://platform.openai.com/docs/models

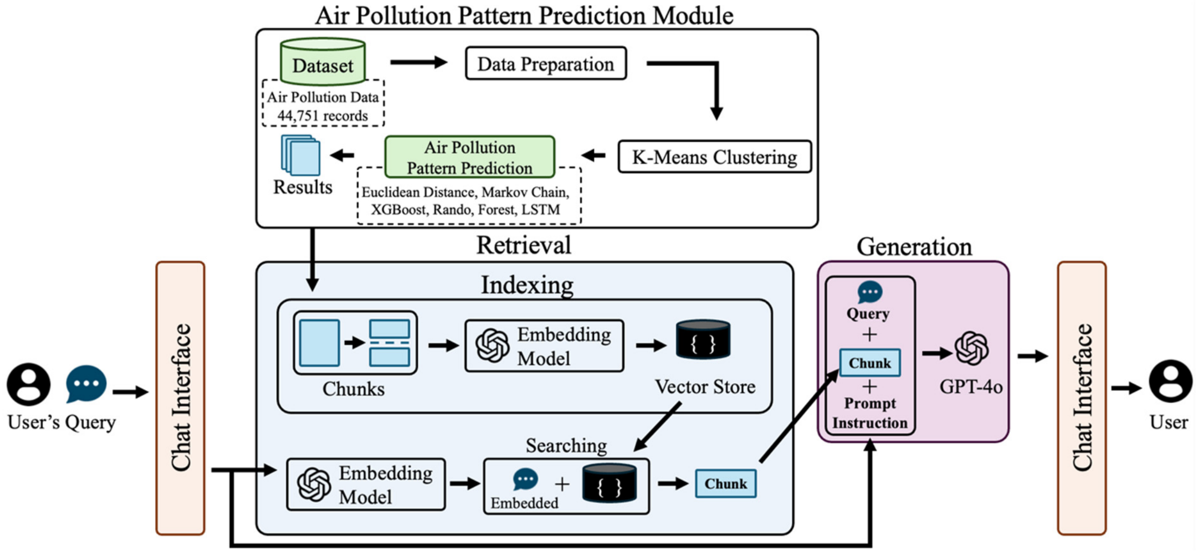

Figure 3: Process pipeline of our RAG system with air pollution pattern prediction module.

3 RAG SYSTEM WITH AIR POLLUTION PATTERN MODULE

In this section, we present how our RAG system can be applied in a real-world application. Figure 3 details the process flow of the RAG system with air pollution pattern prediction module. The system uses the results from the air pattern prediction module as input to the retrieval stage. Then, the retrieved information is passed to the generation stage for generating insight and evidence-based responses. The details of the dataset and our RAG system are presented in the following subsection.

3.1 Dataset

The dataset used in this study is derived from our previous research paper focusing on clustering-based pattern prediction for air pollution analysis (Poositaporn & Jung, 2025). It is research results from our experiment and formatted as a PDF document that includes unstructured text, tables, and figures. The dataset includes K-means clustering results, comparative analysis, performance benchmarks on different predictive models, and evaluation metrics. All of which serve as critical sources of information for the RAG system.

3.2 Our RAG System

The RAG system developed for this study utilizes the OpenAI Assistance API[2]. In this experiment, we applied the default hyperparameter settings provided by OpenAI. Figure 3 illustrates the process pipeline of our RAG system, which consists of two primary stages: retrieval and generation.

1) Retrieval
The input dataset from the air pollution pattern module is divided into smaller text chunks of 800 tokens with an overlap of 400 tokens to ensure contextual continuity between chunks. Each chunk is then embedded into a high-dimensional vector representation using the text-embedding-3-large model with a vector dimension of 256. The resulting embeddings are stored in a vector database.

When the user submits a query, it is also embedded using the same embedding model and compared against the vector database to retrieve the most relevant chunks based on semantic similarity.

2) Generation
The retrieved chunks are concatenated with the user's query and a predefined prompt instruction. A prompt instruction is an additional context or guidance to the generative model. In this study, we used the following prompt instruction in the experiment.

[2] https://platform.openai.com/docs/assistants/overview

Table 1: Example of generated responses from our RAG system and a GPT-4o model.

Question: What actionable recommendations can be derived from clustering results to mitigate high pollution levels?	
RAG	**GPT-4o**
The clustering-based pattern prediction framework suggests several actionable recommendations to mitigate high pollution levels: - **Utilizing K-means clustering helps identify distinct air pollution patterns, allowing for targeted mitigation strategies that are tailored to specific clusters, which can lead to more effective resource allocation and policy development.** - **Integration of machine learning models such as XGBoost and LSTM with clustering can improve the accuracy of pollution predictions, enabling proactive measures to be implemented in anticipation of high pollution days.**	Clustering results can identify specific areas or sources contributing most to high pollution levels. **Based on these insights, targeted interventions such as implementing stricter emissions regulations for industries in high-pollution clusters, enhancing public transportation options to reduce vehicle emissions, and increasing green spaces in urban areas can be recommended. Additionally, community awareness programs can be tailored to educate residents in affected clusters about pollution reduction practices.**

- You are a researcher tasked with answering user questions using information provided in the user's files.
- Use the content of the user's files to provide well-supported answers.
- Use bullet points for clarity when appropriate.
- If the user's files do not contain sufficient information to answer the question, clearly state this limitation.

Ensure responses maintain a formal, academic tone and stay directly relevant to the user's question. The combined input is sent to GPT-4o for response generation. Once the response is generated, it is presented to the user through the chat interface.

4 EXPERIMENTAL RESULTS AND DISCUSSION

This section demonstrates the performance of our RAG system and a GPT-4o model. We compare their ability to interpret and generate insights using five predefined prompts. These prompts were designed to test their capacity for understanding air pollution pattern results and generating actionable recommendations. All questions and generated answers are shown in Appendix A.

Table 1 revealed that the GPT-4o provided a coherent and contextually relevant response. It highlighted potential applications in urban planning and environmental policy making. However, the explanation was generic with no reference to specific data points, methods, or case studies.

On the other hand, the RAG system outperformed the GPT-4o by grounding its response in specific data retrieved from the database. For the same prompt, the RAG system provided detailed examples, such as using K-means clustering to target specific clusters, which could help in efficient resource allocation.

These results demonstrate that our RAG system is uniquely designed to enhance client engagement by delivering customized, data-driven insights adapted to domain-specific needs and supporting decision-making.

5 CONCLUSION

This study focused on enhancing client engagement by developing a chat-based RAG system designed for real-time Q&A interactions. This system enables organizations to deliver precise, context-aware responses and actionable insights. We demonstrated the system in a real-world application by integrating the RAG system with a pattern prediction framework that allows it to retrieve and generate responses grounded in clustering-based air pollution analysis and predictive modeling results.

In the experiment, we compared the performance of the RAG system with a standalone GPT-4o model. The GPT-4o demonstrated its ability to generate coherent and contextually relevant answers. However, its outputs often lacked the depth and specificity required for tasks involving complex datasets or domain-specific insights. This limitation highlights the challenges of standalone generative models when addressing data-intensive tasks.

In contrast, our RAG system presented outstanding performance by retrieving relevant information from external datasets. Its ability to ground responses in specific data resulted in outputs that were both precise and actionable. Additionally, this also highlights that our RAG system can be applied to a wide range of use cases and shows its ability to be a practical tool for a Q&A system in various fields.

In future work, we aim to further explore the potential of RAG and enhance the overall performance in handling complex and data-intensive tasks. A particularly promising application of this integration lies in client engagement, specifically in generating detailed, data-driven reports tailored to the needs of various stakeholders.

ACKNOWLEDGEMENTS

This work was supported by UST Young Scientist+ Research Program 2024 through the University of Science and Technology. (No. 2024YS12)

REFERENCES

Rana, N. P., Slade, E. L., Sahu, G. P., Kizgin, H., Singh, N., Dey, B., Gutierrez, A., & Dwivedi, Y. K. (2019). Digital and social media marketing. In *Advances in theory and practice of emerging markets*.

Afzal, A., Vladika, J., Braun, D., & Matthes, F. (2023). Challenges in Domain-Specific Abstractive Summarization and How to Overcome Them. *arXiv preprint arXiv:2307.00963*.

Han, S., Wang, M., Zhang, J., Li, D., & Duan, J. (2024). A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges. *Electronics*, *13*(24), 5040.

Shahriar, S., Lund, B. D., Mannuru, N. R., Arshad, M. A., Hayawi, K., Bevara, R. V. K., Mannuru, A., & Batool, L. (2024). Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. Applied Sciences, 14(17), 7782.

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17754-17762.

Ranade, P., & Joshi, A. (2023). Fabula: Intelligence report generation using retrieval-augmented narrative construction. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 603-610.

Poositaporn, A., & Jung, H. (2025). Towards a clustering-based pattern prediction framework. [Manuscript submitted for publication].

APPENDIX

A. Generated responses of the RAG system and a GPT-4o model

Q1: What actionable recommendations can be derived from clustering results to mitigate high pollution levels?

RAG: The clustering-based pattern prediction framework suggests several actionable recommendations to mitigate high pollution levels:

- **Utilizing K-means clustering helps identify distinct air pollution patterns, allowing for targeted mitigation strategies that are tailored to specific clusters, which can lead to more effective resource allocation and policy development.**
- **Integration of machine learning models such as XGBoost and LSTM with clustering can improve the accuracy of pollution predictions, enabling proactive measures to be implemented in anticipation of high pollution days.**

GPT-4o: Clustering results can identify specific areas or sources contributing most to high pollution levels. **Based on these insights, targeted interventions such as implementing stricter emissions regulations for industries in high-pollution clusters, enhancing public transportation options to reduce vehicle emissions, and increasing green spaces in urban areas can be recommended. Additionally, community awareness programs can be tailored to educate residents in affected clusters about pollution reduction practices.**

Q2: Explain the significance of using K-means to cluster air pollutant pattern for next day's prediction.

RAG: The significance of employing K-means clustering to analyze air pollutant patterns lies in its ability to efficiently group complex data into meaningful clusters, which represent distinct patterns of air pollution. **By calculating the centroids that represent these patterns, K-means allows for the identification of key characteristics of similar data**

points, thereby facilitating the prediction of future air pollution patterns. This approach is particularly advantageous for real-time systems where quick pattern prediction is necessary, enhancing the responsiveness and accuracy of predictive models.

GPT-4o: Using K-means to cluster air pollutant patterns is significant for next day's prediction as it helps identify distinct patterns or trends in historical air quality data. **By grouping similar data points, K-means can reveal underlying structures and relationships, enabling more accurate forecasting models to predict future pollution levels based on these identified patterns.** This approach enhances the ability to anticipate and mitigate adverse air quality conditions.

Q3: What observations can you make based on the provided data or general knowledge about air pollutant pattern prediction results?

RAG: The data explores a variety of machine learning models to predict air pollutant patterns with a focus on understanding and improving the accuracy of these predictions. **Particularly, models like XGBoost and LSTM showed better prediction accuracy for small cluster sizes, such as K = 10, indicating that fewer clusters were more effective in capturing air pollution patterns than larger ones. Additionally, the integration of clustering techniques like K-means with the Overall Difference with Crossover Penalty (OD with CP) metric has been emphasized to enhance the prediction framework by accurately capturing short-term and long-term patterns within the air pollution data.**

GPT-4o: Without specific data provided, I can offer general observations about air pollutant pattern predictions. Typically, these predictions show that air pollution levels are influenced by factors such as industrial activity, traffic density, weather conditions, and seasonal changes. Patterns often reveal higher pollution levels in urban areas and during colder months due to increased heating and stagnant air conditions, while improvements are usually observed with the implementation of stricter environmental regulations and cleaner technologies.

Q4: How can the distribution of clusters in the results help interpret underlying patterns in air pollutant levels?

RAG: The distribution of clusters in K-means clustering results is crucial for interpreting underlying patterns in air pollutant levels because it helps identify distinct groups of data points with similar characteristics. This classification allows researchers to discern prevalent patterns, seasonal trends, and potential sources of pollution. Analyzing clusters enables the examination of spatial and temporal variations in air quality, thereby assisting in the prediction and management of pollution levels.

GPT-4o: The distribution of clusters in air pollutant data can reveal spatial or temporal patterns, indicating areas or times with similar pollution characteristics. By analyzing these clusters, one can identify potential sources of pollution, such as industrial zones or traffic-heavy areas, and assess the effectiveness of regulatory measures. Additionally, clustering can highlight anomalies or trends, such as seasonal variations or the impact of specific events on air quality.