Development of a Versatile Large Language Model Platform by KAERI: Integrating Intranet and Internet Environments

Byeongha Jo[a+], Hanseung Seo[a+]
Hyuna Jeon[b], Joowon Cha[c], Jaejun Lee[c], Seungdon Yeom[c]
Yonggyun Yu[cd*]

[a]*Korea University of Technology and Education,*
[b]*jenti Co., Ltd,*
[c]*Korea National University of Science and Technology,*
[d]*Korea Atomic Energy Research Institute*
[*]*Corresponding author: ygyu@kaeri.re.kr , +contributed equally*

Keywords : LLM(large language model), sLLM(small large language model), RAG(Retrieval Augmented Generation)

1. Introduction

The recent emergence of Large Language Models (LLMs) like ChatGPT has sparked interest in systems that can enhance efficiency in document generation and information retrieval. However, the nuclear industry plays a crucial role in national security, making security concerns paramount. For this reason, Korea Atomic Energy Research Institute (KAERI) maintains a strict security system by thoroughly separating external and internal networks. By blocking internet access on the work network, it becomes challenging to utilize AI services like LLMs, leading to reduced work efficiency.

This paper presents a plan for system construction that enable the efficient use of AI language model services across internal, external, and off-network environments, while maintaining security in KAERI's network segregation policy. This approach aims to overcome the constraints on utilizing AI technology while preserving security measures.

The services are divided into three categories, considering the purpose and users of each network:

- Internal Network LLM Service: Designed for tasks involving sensitive information, accessible only to internal employees.
- External Network LLM Service: For tasks involving non-sensitive information, also limited to internal employees.
- Public Internet LLM Service: Aimed at both internal employees and the general public for institutional promotion, accessible via the public internet.

2. Service Architecture & Functionality

2.1 External LLM Service

The external network LLM service is a platform for efficiently using commercial LLM services to analyze or query non-confidential public documents or papers. This platform can utilize KAERI's own LLM model specialized for the nuclear domain and also provides commercial LLM services through APIs to improve efficiency in general tasks.

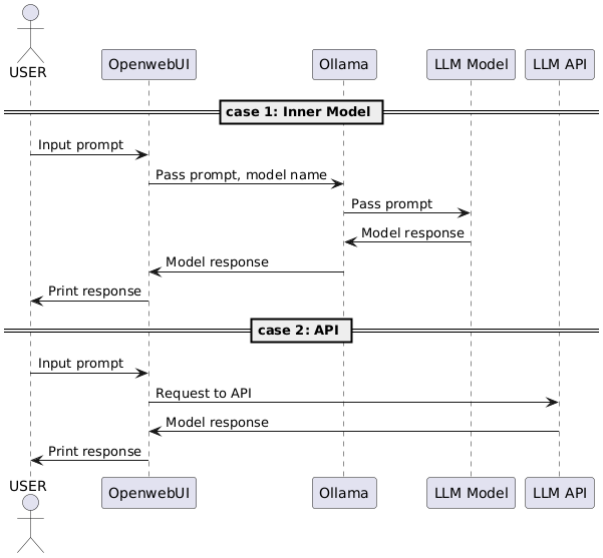

Fig 1. Sequence Diagram of External LLM Service

2.1.1 Integration external APIs by using open-source framework

The external network LLM service uses *OpenwebUI*[1] to build the UI and utilizes external APIs to access high-performance LLMs like *ChatGPT*, *Claude.AI*, and *Gemini* provided by *OpenAI*, *Anthropic*, and *Google*. These high-performance LLMs can generate answers with high accuracy and speed, allowing for efficient task processing. However, caution must be exercised when using APIs, as input information is be exposed to the API server.

In addition to external APIs, the service can also use models installed on KAERI's own servers. In this case, public models registered in Ollama[2] are installed and inference is performed on the internal server. While this

approach is better in terms of security compared to using external APIs, there are clear limitations in performance, so the appropriate method should be optimized considering the purpose and cost.

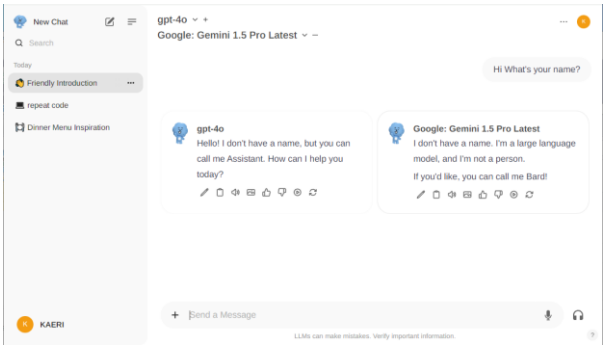

Fig 2. User Interface of External LLM Service

2.1.2 Main services offered through the API

This service allows internal employees to summarize non-confidential public documents or perform question-answering based on document content. It is also used for translation, writing assistance for paper composition, and code generation for research purposes.

2.2 Internal LLM Service

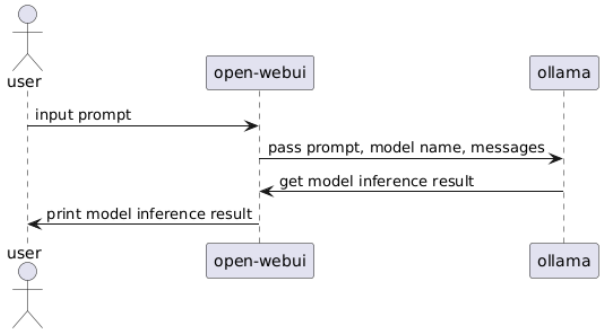

Fig 3. Sequence Diagram of Internal LLM Service

The internal network LLM service enables the use of LLM services in an environment where internet use is restricted due to network separation. It also allows for the use of internal and confidential documents that could not be handled in the external network due to security reasons.

2.2.1 Difference between External LLM Service and Internal LLM Service

The main difference from the external network service is the ability to use internal confidential documents related to security. When using external APIs, there is a possibility of secret information being leaked as user input information flows into the server providing the API. However, the internal network service serves its own model, allowing the use of internal information for generating questions, translations, summaries, and other functions without the risk of external leakage.

2.2.2 Integration internal APIs by using open-source framework

Since the internal network is not connected to the internet, it is difficult to use LLM services that rely on APIs. Therefore, in the internal network, it is possible to use models specialized for specific domains by utilizing domain-specific data.

The internal network LLM service can process user queries using its own small Large Language Model (sLLM), similar to the case of using a proprietary model in the external network service. Using a proprietary sLLM facilitates the application of RAG (Retrieval Augmented Generation)[3], enabling more accurate answer generation for domain-specific questions by referencing relevant documents.

2.2.3 Implementing Retrieval Augmented Generation based Q&A function for internal regulation documents

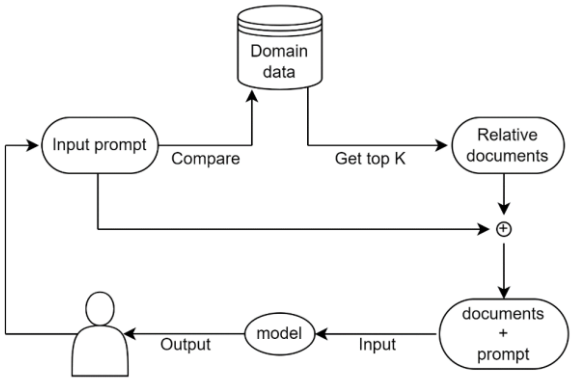

Fig 4. Retrieval Augmented Generation flow chart

By incorporating RAG into the proprietary sLLM, more accurate information about domain-specific documents can be provided. The RAG system operates by pre-storing domain-specific documents, then selecting the k most relevant documents to the user's input. It then combines these k documents with the user's prompt as input to the model, This approach enables the model to provide more accurate answers to user queries by referencing the selected documents.

2.3 Public Internet LLM Service

2.3.1 Need for Public Internet LLM Service Development

In addition to the LLM services for internal employees, an intelligent chatbot named '*Padongi-bot*' was developed to convey knowledge about nuclear energy to the general public. This system, accessible via the public internet, aims to provide objective information about nuclear energy to users based on commercial LLM APIs and RAG technology.

2.3.2 Public Internet LLM Service System Structure

The overall system architecture of '*Padongi-bot*', our public internet LLM service, is as follows:

- A public web interface that receives user queries and performs ethical content filtering.
- A classification system that categorizes verified questions into general conversations, search-required queries, and nuclear-specific inquiries.
- Specialized processing pipelines for each query type:
  - General conversations are handled directly by the *GPT* model.
  - Search-required queries utilize the Google Search API, with the top two results being processed.
  - Nuclear-specific questions engage a RAG system using the *Milvus*[4] vector database for relevant document retrieval.
- The *GPT4o-mini* model generates final responses based on the processed information.

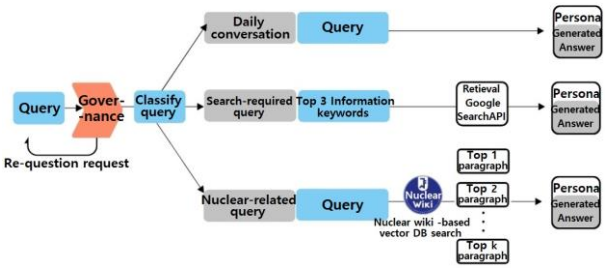

Fig 5. System Architecture

2.3.3 User Conversation History Management System

A PostgreSQL[5] relational database is used to manage user conversation history. This allows for effective storage and retrieval of each user's recent conversation content, enabling context-aware and consistent conversations. The conversation history management system stores the most recent 10 conversations for each user, which is used to understand context and generate appropriate responses in subsequent conversations.

3. Conclusions

This paper presented a method for utilizing LLM services for various purposes in environments with different security levels: internal network, external network, and public internet. Currently, internal and external network LLM services are publicly available and in service within the institute, while the public internet LLM service (*Padongi-bot*) has completed development and is being set up for public access. Looking ahead, KAERI plans to further enhance these systems to improve both internal work efficiency and public engagement. This will involve continued refinement of our proprietary models, incorporating additional training on internal regulatory documents, technical reports, and publicly available nuclear energy information.

Acknowledgement

This work was supported by KAERI R&D Program (*KAERI-524540-24*).

REFERENCES

[1] open-webui, Available: https://github.com/open-webui/open-webui
[2] ollama, Available: https://github.com/ollama/ollama
[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Proceedings of NeurIPS 2020, 2021.
[4] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, et al., Milvus: A Purpose-Built Vector Data Management System, Proceedings of the 2021 International Conference on Management of Data, pp. 2614-2627, 2021.
[5] PostgreSQL Global Development Group, PostgreSQL: The world's most advanced open source database. Available: https://www.postgresql.org