GIST develops AI legal service with 23% higher accuracy... RAG technology that increases reliability of AI's ability to ask and find answers on its own when it doesn't know

- Professor Heung-No Lee's team from the School of Electrical Engineering and Computer Science develops RAG framework technology optimized for the legal field... Search and response accuracy improved by 23% compared to existing RAG, performance increased by 14% compared to fine-tuned LLM

- Expected to be helpful not only for legal practice but also for vulnerable groups who have difficulty receiving legal consultation services

- Published in the international academic journal 《IEEE Access》

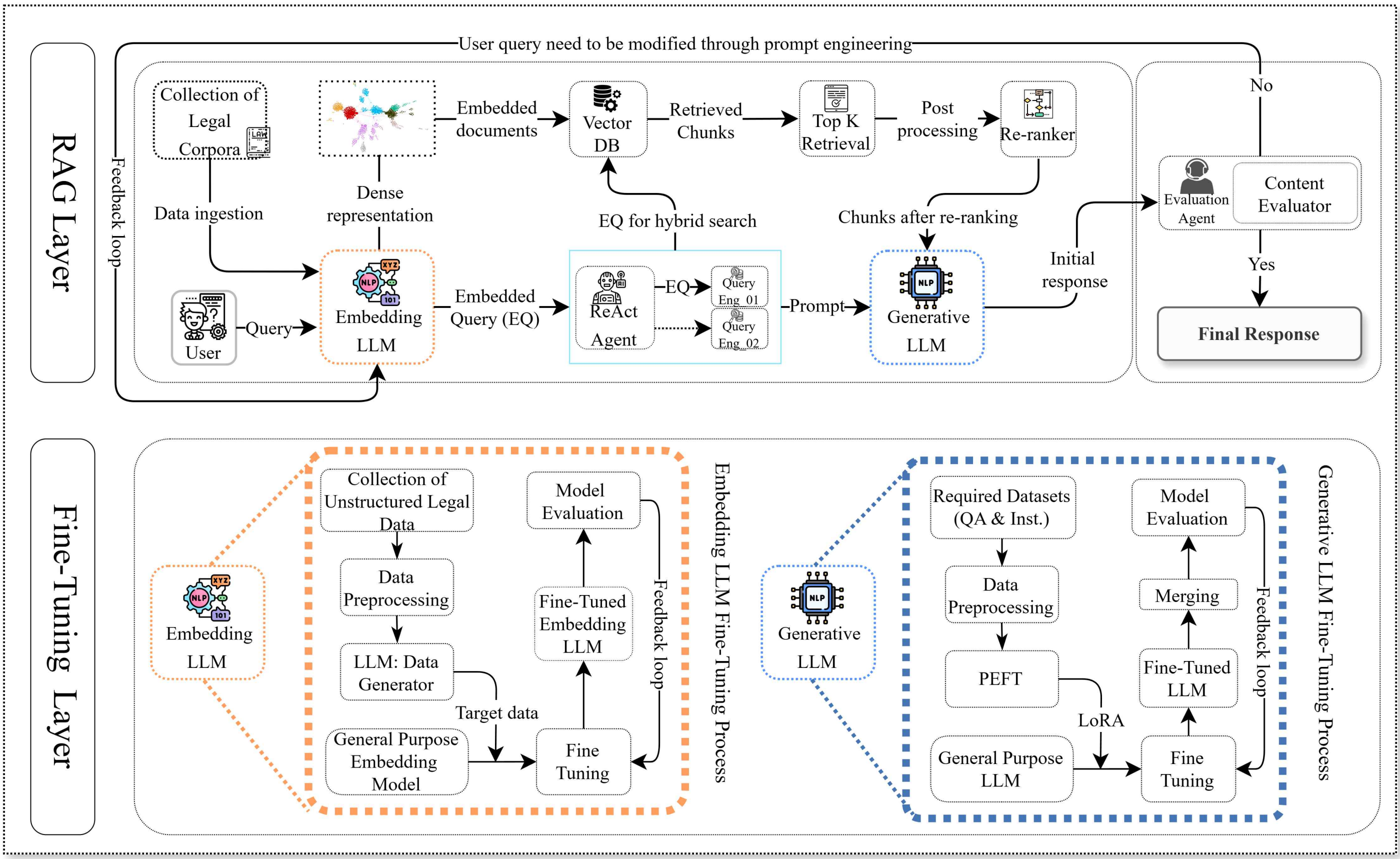
▲ (From left) Professor Heung-No Lee of the School of Electrical Engineering and Computer Science and student RAHMAN S M WAHIDUR

The recent rapid development of artificial intelligence (AI) technology is causing innovative changes across society and economy, and is also rapidly spreading in the legal field.

However, large-scale language models (LLMs) that generate answers based on learned data have limitations in accurately processing complex legal questions. In particular, serious problems can arise when AI provides incorrect information in case analysis or contract drafting where interpretation is important.

The Gwangju Institute of Science and Technology (GIST, President Kichul Lim) announced that the research team of Professor Heung-No Lee (Director of the ITRC Blockchain Intelligence Convergence Center) of the School of Electrical Engineering and Computer Science has developed a ‘Retrieval-Augmented Generation (RAG)’ framework technology specialized in the legal field. The biggest feature is that the AI itself asks and searches for information it does not know, thereby increasing the reliability and accuracy of answers.

This technology is expected to be widely used in legal practice as well as legal support for vulnerable groups by drastically reducing the 'hallucination' problem that occurs in existing AI-based legal reasoning and increasing accuracy, transparency, and reliability.



[Figure] Schematic design of the proposed LQ-RAG system. This system is divided into two main components: the Fine-Tuning Layer and the RAG Layer.
‣Fine-Tuning Layer: This is a layer that performs fine tuning to optimize the performance of the embedding LLM and the generation LLM.
‣RAG Layer: It includes the advanced RAG module, evaluation agent, prompt engineering agent, and feedback mechanism to ensure accuracy and reliability in the response generation process of the system.

Through this design, the LQ-RAG system enables more precise and reliable information retrieval and response generation in the legal domain, ultimately contributing to improving the performance of the legal AI system.

It has been reported that existing LLM-based legal AI systems have a hallucination rate of 58-82%. Accordingly, RAG* technology, which reflects legal information accurately searched by AI, is attracting attention, but the existing RAG method also has problems such as limitations in information retrieval and lack of applicability to legal context.

* RAG: It searches for necessary information from a pre-established database or document and generates answers based on the latest data that the existing LLM does not reflect. RAG also has the advantage of saving cost and time because it searches and uses external data instead of retraining a large-scale model.

To solve this, the research team developed the 'Legal Query RAG (LQ-RAG)' framework optimized for the legal field that efficiently searches and utilizes legal data while increasing the reliability and accuracy of answers.

The LQ-RAG model fine-tuned* the embedding* generation LLM and the response generation LLM by utilizing a wide range of legal texts. By learning a large amount of case law and statutory data, it was able to deeply understand specialized legal terms and document structures, and by retraining the generation model based on actual legal Q&A data, it secured more sophisticated answering capabilities.

* Embedding: A representation that converts documents into vectors to quantify meaning

* Fine tuning: It refers to the process of additionally learning a previously learned model for a specific purpose or dataset. It optimizes an already learned model to fit a specific task, which is faster and more efficient than learning a basic model from scratch.

LQ-RAG integrates four core elements: ▴ Customized Evaluation Agent ▴ Response Generator LLM ▴ Prompt Engineering Agent ▴ Embedding Generator LLM.

This structure effectively minimizes hallucination, improves domain-specific accuracy, and provides clear and high-quality answers even for complex questions. It also enables continuous performance improvement through a recursive feedback process*.

* Recursive feedback process: A mechanism that repeatedly improves the search and generation steps based on the evaluation of generated responses to induce more accurate and relevant answers.

LQ-RAG explicitly applies an agent-based iterative improvement mechanism in the inference process to derive the optimal answer. The generated answer is evaluated based on the appropriateness of the context and factual accuracy through an evaluation agent.

LQ-RAG increased reliability through a recursive feedback mechanism that continuously improves the answers generated by AI, and the research team systematically embedded legal documents into high-dimensional vectors, enabling AI to provide more accurate legal information based on this.

When comparing China's generative AI 'DeepSeek-R1' and the research team's LQ-RAG, both models use unique techniques to provide more accurate and sophisticated answers, but they differ in the way they improve the answers.

DeepSeek-R1 develops thinking ability based on reinforcement learning (RL) and improves the quality of answers through a chain reasoning process. It also performs the function of reviewing and improving the answers it generates.

While DeepSeek-R1 recursively improves answers within a single model, LQ-RAG improves answers through a multi-agent collaboration method. Both models have in common that they go through their own optimization process without direct human feedback.

As a result of applying LQ-RAG, the accuracy of legal document search and response was improved by 23% compared to the existing RAG system, and it also recorded 14% higher performance compared to the fine-tuned LLM.

This shows that domain-specific LLM combined with advanced RAG modules and feedback mechanisms can significantly increase the reliability and performance of AI in legal practice.

The research team plans to develop an agent-based legal workflow to improve the efficiency of legal work such as contract drafting and compliance monitoring, so that legal professionals can focus on their core work.

Professor Heung-No Lee said, “We plan to apply this technology to various legal work such as legal document analysis, contract drafting automation, and compliance monitoring. We will build a reliable legal AI system by combining search augmentation generation (RAG) and multi-agent collaboration technology, and provide more accurate legal analysis and reliable AI-based legal solutions.”

This research, supervised by Professor Heung-No Lee of the School of Electrical Engineering and Computer Science at GIST and conducted by integrated course students Rahman S M Wahidur, was supported by the Ministry of Science and ICT and the Institute of Information and Communications Technology Planning and Evaluation (IITP). The results of the research were published online in the international journal 《IEEE Access》 on February 14, 2025.

GIST
Since 1993