ArchiveChat: AI-based platform for biological knowledge mining to enhance accessibility and reusability in the exploration of NGS data

Joo-Seok Park[1] and Yang Jae Kang[1*]

[1]Research Institute of Molecular Alchemy, Gyeongsang National University, Jinju, Gyeongnam 52828,
South Korea

The exponential growth of next-generation sequencing (NGS) data and biological metadata has created a critical need for intelligent systems that enable efficient data access, interpretation, and automation. Although public databases provide access to large-scale NGS data, their interfaces often lack intuitiveness, limiting accessibility for researchers. The complexity and non-intuitive design of these databases frequently reduce both accessibility and usability, making it challenging for researchers to navigate and effectively utilize the information. Recent advancements in artificial intelligence (AI), particularly in natural language processing (NLP) and large language models (LLMs), offer promising solutions to address these limitations by enhancing retrieval processes. However, LLMs are limited in accessing real-time, domain-specific information. Retrieval-augmented generation (RAG) systems overcome these limitations by combining the generative capabilities of LLMs with external retrieval mechanisms, enabling access to up-to-date and specialized information. To improve metadata accessibility and usability, we developed ArchiveChat, an AI-based conversational platform to support efficient and user-friendly exploration of NGS metadata. ArchiveChat enables natural language queries and integrates retrieval-augmented mechanisms to provide structured and relevant results. As a result, ArchiveChat lowers the barrier to metadata navigation and promotes reusability, transparency, and productivity in bioinformatics research.

* Corresponding Author: , Tel: +82-55-772-1352, E-mail: kangyangjae@gnu.ac.kr