Domain Adaptation of Local LLMs for Metal-Sulfide Photocatalyst Discovery

Wataru Takahara
Nara Institute of Science and Technology

Large language models (LLMs) have the potential to serve as collaborative assistants in scientific research. However, adapting them to specialized domains is difficult because it requires the integration of domain-specific knowledge. We present a framework called Material Dual-Source Knowledge RAG (MDSK-RAG). This framework is a type of retrieval-augmented generation (RAG) and enables the domain adaptation of a local LLM without fine-tuning. We applied MDSK-RAG to research on metal-sulfide photocatalysts. The framework uses two retrievers. One handles tabular experimental data in CSV format. The other deals with unstructured scientific literature in PDF format. We convert the tabular data into scientific language by predefined templates. The local LLM summarizes the retrieved texts based on a fixed prompt. This prompt asks for information on five key aspects: composition, crystal structure type, synthesis method, reaction conditions, and hydrogen evolution activity. We evaluated the responses using 12 expert-defined questions. The results showed improvements in factual accuracy and the quality of domain-specific reasoning. The present framework works in a fully local environment and improves data security and reduces costs. We have also developed a user-friendly application. The present framework supports practical use of the framework in materials development workflows and other scientific domains.