AIMI 플랫폼: 신소재 조성–공정–물성 추론을 위한 LLM 기반 RAG–SHAP 통합 아키텍처 설계

AIMI Platform: Design of An Integrated RAG–SHAP Architecture Based on LLM for New Materials C–P–P Inference

Seung-Jun Han*[2, 3], Won-Yong Shin†[1, 3], Jun-Chae Na[3], Sung-Il Yang[3], Young-Jin Yu[1, 3], Ju-Hye Lee[3], Yong-eun Cho[1, 3], Min-Hee Lee[1, 3], and Chan Jung[1, 3]

[1]School of Mathematics and Computing (CSE), Yonsei University. Seoul 03722, Republic of Korea
[2]Department of Mechanical Engineering (ME), Kyung Hee University, Yongin 17104, Republic of Korea
[3]KAILOS LAB Co. Ltd., Seoul 06349, Republic of Korea
Email: hansj2k@khu.ac.kr, wy.shin@yonsei.ac.kr, kailoslab@gmail.com, siyang12@gmail.com, lucy.yu0601@gmail.com, kailosmay@gmail.com, jdmeekboi@gmail.com, chloe@kailoslab.com, channthehuman@gmail.com

Abstract

신소재 개발 과정의 자동화를 위해, 조성–공정–물성 (Composition–Processing–Property, C-P-P) 관계를 LLM (Large Language Model) 기반으로 추론하고, RAG (Retrieval-Augmented Generation) 기반 문헌 검색과 SHAP (SHapley Additive exPlanations) 기반 해석 모듈을 통합한 기술 구조를 제안한다. 제안 시스템은 자연어 질의로부터 문헌 기반 조성 정보를 검색하고, 실험 데이터 기반 회귀 예측 및 해석을 통해 설명 가능한 합금 조합 추천 및 물성 예측을 수행한다. 특히 SHAP 분석 결과를 LLM 프롬프트에 삽입함으로써 정량적 예측과 자연어 해석을 연계한 설명형 실험 설계가 가능하며, AIMI (AI for Material Innovation) 플랫폼 내 실제 구현하여 기술의 실효성과 확장 가능성을 확인한다.

I. Introduction

반도체 배선 소재나 고체 전해질 등의 신소재 개발은 조성, 공정, 물성 간에서 최적 조합을 찾는 것이며, 기존의 반복적인 실험은 높은 비용과 시간이 소요된다. 기존의 기계학습 기반 예측 모델의 대부분은 정형 데이터 기반이며, 도메인 지식 통합이나 해석 가능성에서 한계를 갖는다. 최근 LLM (Large Language Model)은 자연어 기반 질의응답 및 문헌 추론 기능을 바탕으로 실험 로그, 조성 파일 등 비정형 데이터를 함께 보조하여 처리하는 등 재료 과학 분야에서 새로운 도구로 주목받고 있다. 선행 연구 중 MatSciBERT는 텍스트의 임베딩 및 유사 질의 응답을 시도하였다 [1]. 그러나 해당 연구는 문헌 요약이나 개념 수준의 탐색에 국한되며, 실험 데이터 기반 수치 예측, 실험 조건 생성을 통합하는 시스템은 전무하다. 본 연구는 이를 해결하기 위해, LLM 중심의 질의응답 구조에 RAG (Retrieval-Augmented Generation) 기반 문헌 검색과 SHAP (SHapley Additive exPlanations) [2] 기반 회귀 예측 및 해석을 통합한 통합형 추론 시스템 아키텍처 설계를 목표로 한다.

II. Methodology

LLM 기반 신소재 추론 구조는 다음 세 가지 주요 구성 요소로 이루어진다.

- 첫째, C-P-P (Composition–Processing–Property) 기반 질의 응답 구조는 사용자의 자연어 질의를 바탕으로 RAG 구조를 활용하여 논문 · 특허 등에서 구축된 벡터 데이터베이스로부터 관련 사례를 검색한다. 검색된 문헌은 LLM 프롬프트의 컨텍스트로 삽입되며, LLM은 이를 바탕으로 조성–공정–물성 관계를 반영한 응답을 생성한다.
- 둘째, 수치 예측과 해석을 위해 SHAP 기반 회귀 분석 모듈을 구성한다. 실험 데이터를 기반으로 학습된 회귀 모델은 조성 및 공정 파라미터를 입력으로 하여 물성 값을 예측하고, SHAP 분석을 통해 각 특징 (feature)의 기여도를 도출한다. 이 결과는 JSON 형태로 정리되어 LLM 프롬프트에 삽입되며, LLM은 이를 자연어 해석으로 변환해 부연설명을 생성할 수 있다.
- 셋째, 최종 프롬프트는 사용자 질의, 문헌 정보, 수치 해석 결과를 통합하여 구성되며, LLM은 이 복합 정보를 바탕으로 조성 추천, 조건 생성, 해석형 응답을 동시에 수행한다. Few-shot 예시와 함께 LangChain Tool 및 SHAP API 연계를 통하여 멀티모달 추론도 가능하도록 설계한다.

III. Conclusion

본 연구는 C–P–P 추론을 중심으로, LLM 기반 질의응답 시스템에 RAG를 기반한 지식 검색 과 설명 가능한 회귀 해석 모듈인 SHAP을 통합한 자동화 설계 프레임워크를 제안하였다. 수치 기반 예측, 해석 가능성, 조건 생성까지 포괄하는 기술 통합 구조를 제공하며, AIMI 플랫폼을 통해 실제 구현 가능성을 검토하였다. 향후에는 정량 실험 (예: 예측 정확도, SHAP-LM 해석 정합성)에 대한 검증을 기반으로, 강화학습 기반 실험 최적화, 고도화된 신경망 (예: 그래프 신경망)을 통한 멀티모달 구조 (예: SEM image) 통합, LLM 경량화 적용 등을 통해 자율 실험형 소재 개발 시스템으로 확장할 계획이다.

Acknowledgement

This research was supported by SMEs Technology Innovation Development Program through the Technology Innovation and Promotion Agency (TIPA), funded by Ministry of SMEs and Startups (RS-2024-00511332) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2021-NR059723, No. RS-2023-00220762).

References

[1] T. Gupta *et al.*, "Matscibert: A materials domain language model for text mining and information extraction," *Computational Materials*, 2022.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NeurIPS*, 2017.