Symposium: MT-02

Title: Development and analysis of largest, accurate and comprehensive Multicomponent (high entropy) alloy database using large language models

Authors: Aravindan Kamatchi Sundaram[1], Sai Mani Kumar Devathi[1], B. Pabitramohan Prusty[1], Mohit Chakraborty[1], Rohit Batra[1,2,*]
[1] Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai
600036, India
[2] Center for Atomistic Modelling and Materials Design, IIT Madras, Chennai 600036, India
[*] Corresponding author; ORCID: 0000-0002-1098-7035

Abstract:
Machine learning (ML) methods have shown success in material design and optimization, but they rely on pre-existing, expensive, or manually curated datasets, limiting their potential for discovery. Natural language processing (NLP) methods, particularly large language models (LLMs), have been explored to extract materials data from literature corpus [1, 2], but they often lack desired accuracy or are too narrow in their scope. In this work, we develop an LLM-based pipeline to accurately extract alloy-related information from both textual descriptions and tabular data in the literature. From textual sources, we extract various alloy details such as composition, processing conditions, characterization methods, and reported physical and mechanical properties. From tabular data, we extract alloy properties along with their corresponding units and values, supported by a comprehensive set of alloy properties commonly reported in high-entropy alloy (HEA) studies that we compiled. Our pipeline aims at increasing the sensitivity of LLMs to material domain knowledge which is vital to improve data extraction accuracy. We employ techniques such as prompt engineering (few-shot examples, context, roles, confirmations) and retrieval-augmented generation with various LLMs (GPT-4o, GPT-4o-mini, Llama-3, GPT-3.5 turbo) for this and achieved an F1-score of ~0.9 on textual data extraction and ~0.95 on tabular data extraction, significantly surpassing current models. We also developed evaluation methods to comprehensively assess the LLMs' performance, testing the pipeline against existing alloy datasets. The developed pipeline was applied to over 10,000 articles, producing the largest publicly available alloy dataset, and uncovering chemical trends such as correlations between alloy composition, mechanical properties, and processing conditions. The developed framework is quite versatile and can be used to develop LLMs for other material domains like polymers, metal-organic frameworks (MOFs), and ceramics.

[1] Dagdelen, John, et al. "Structured information extraction from scientific text with large language models." *Nature Communications* 15.1 (2024): 1418.
[2] Polak, Maciej P., and Dane Morgan. "Extracting accurate materials data from research papers with conversational language models and prompt engineering." *Nature Communications* 15.1 (2024): 1569.