Automated Extraction of Multicomponent Alloy Data Using Large Language Models for Sustainable Design

Rohit Batra[1,2], Aravindan Kamatchi Sundaram[1], Mohit Chakraborty[1], Sai Mani Kumar Devathi[1] and B. Pabitramohan Prusty[1]

[1]Department of Metallurgical and Materials Engineering, Indian Institute of Technology Madras, Chennai 600036, India
[2]Center for Atomistic Modelling and Materials Design, Indian Institute of Technology Madras, Chennai 600036, India

rbatra@smail.iitm.ac.in

The design of sustainable materials requires access to materials performance and sustainability data from literature corpus in an organized, structured and automated manner. Natural language processing approaches, particularly large language models (LLMs), have been explored for materials data extraction from the literature, yet often suffer from limited accuracy or narrow scope [1, 2]. In this work, an LLM-based pipeline is developed to accurately extract alloy-related information from both textual descriptions and tabular data across the literature on high-entropy (or multicomponent) alloys (HEA). Specifically, two databases with 37,711 and 148,069 entries respectively are retrieved; one from the literature text, consisting of alloy composition, processing conditions, characterization methods, and reported properties, and other from the literature tables, consisting of property names, values, and units. The pipeline enhances materials-domain sensitivity through prompt engineering and retrieval-augmented generation and achieves F1-scores of ~0.83 for textual extraction and ~0.88 for tabular extraction, surpassing or matching existing approaches. Application of the pipeline to over 10,000 articles yields the largest publicly available multicomponent alloy database and reveals compositional and processing-property trends. The database is further employed for sustainability-aware materials selection in three application domains, i.e., lightweighting, soft magnetic, and corrosion-resistant, identifying multicomponent alloy candidates with more sustainable production while maintaining or exceeding benchmark performance. The pipeline developed can be easily generalized to other class of materials, and assist in development of comprehensive, accurate and usable databases for sustainable materials design.

References

[1] Hira, Kausik, Mohd Zaki, and N. M. Mausam. “Large scale Extraction of Composition and Properties from Materials Tables.” NeurIPS 2024 Workshop AI4Mat.

[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.

Figures

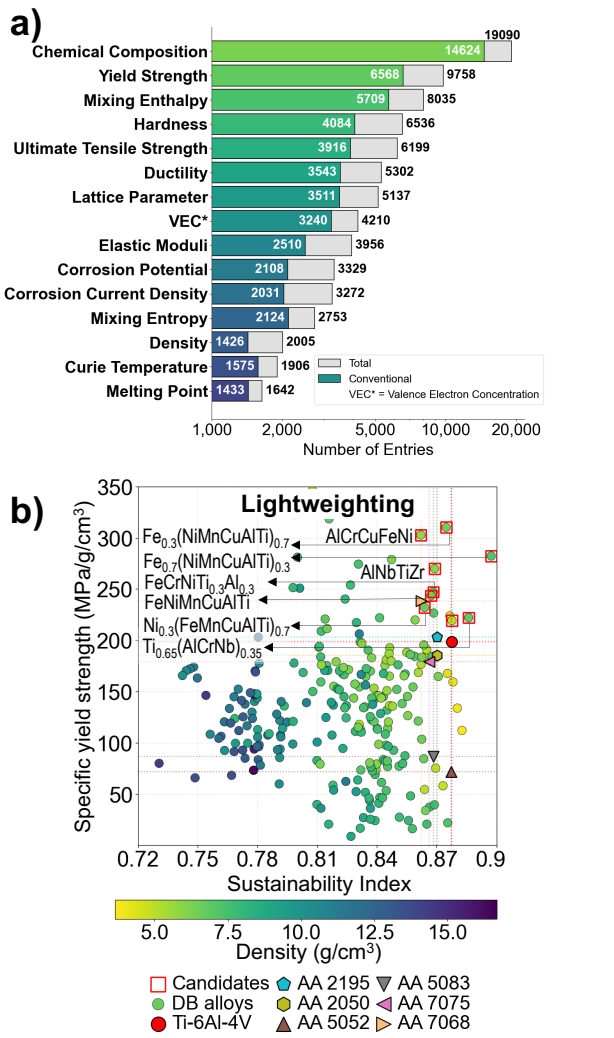

Figure 1. a) Visualization of the LLM-mined materials database showing log-scale distribution of the 20 most frequently extracted material properties. b) Sustainability maps highlighting alloy designs satisfying performance and sustainability criteria in the application domain of lightweighting.