LitCurate: LLM-Assisted Extraction of Lower Mantle Mineral Equation of State Parameters for Database Construction

Dominica Wilson[1,2], Abin Shakya[2], Renata Wentzcovitch[2]

[1]Tuskegee University, Tuskegee Alabama, [2]Lamont-Doherty Earth Observatory of Columbia University, Palisades New York

Analyzing lower mantle minerals requires equation-of-state (EOS) parameters. We focus on 3 parameters in particular: molar volume (V_0), bulk modulus (K_0), and its pressure derivative (K_0'). These parameters are scattered across decades of scientific literature about mineral physics, making it difficult for geologists and geophysicists to search for data and interpret data by hand. Manual literature curation is labor intensive and difficult to scale. To aid in this process, we propose a multi-stage configurable pipeline called LitCurate that automates the process of finding papers, extracting data from these papers, and placing the data in a unified dataset. In our use case, we use LitCurate to mine and extract EOS parameters from scientific literature. The pipeline generates search queries using a large language model (LLM), gathers the papers using OpenAlex, and uses a cascading filter process which involves abstract filtering, markdown conversion, full-text filtering, and more before the final extraction of EOS parameters. Starting from 1855 papers, successive filtering reduced the corpus to 205 with valid EOS extractions. The final 205 papers resulted in 1350 rows to be placed in the final database. The full-text filter stage presented the highest proportional loss of papers with a retention rate of 37% (314 of 850 papers), suggesting that full-text review is more reliable than abstract and keyword level filtering alone. These results demonstrate that LLM-assistance in literature mining pipelines can help to obtain a structured dataset of EOS parameters from a literature corpus, setting the foundation for building a searchable lower mantle mineral database while illustrating an efficient and scalable workflow that can be adapted across disciplines.