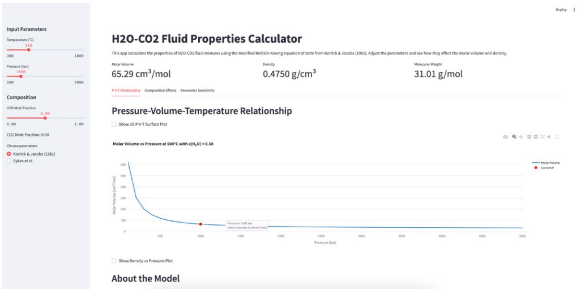
Goldschmidt 2025 Abstract
https://doi.org/10.7185/gold2025.31199



Leveraging Fine-Tuned Large Language Models (LLM) and Retrieval-Augmented Generation (RAG) for Modeling Complex Geological Fluid Systems (Geofluids) at Crustal and Mantle Conditions

ROBERT BODNAR, SCOTT MUTCHLER AND GRACE GRIFFITH

Virginia Tech

During his long and distinguished career, Dr. I-Ming Chou has been a leading innovator in the development of experimental protocols to determine the properties of geofluids. In recent decades numerous models have been developed to estimate geofluid properties at elevated PT conditions. At the same time, our knowledge and understanding of the broad range in geofluid compositions has advanced as analytical techniques to analyze individual fluid inclusions have improved. While experimental data provide the basis upon which empirical models to estimate fluid properties are built, the number of experiments required to completely characterize multi-component systems increases with each added component.

The rapidly advancing field of Artificial Intelligence offers the possibility to build models to better extract information from existing data sets, and to identify those regions of PTX space in which fluid properties are expected to vary significantly to help direct design of experiments. In this study we leverage fine-tuned Large Language Models (LLM) with Retrieval-Augmented Generation (RAG) to examine existing data and to query the models to gain more in-depth insights and understanding of fluid behavior. The LLMs used were fine-tuned on a large corpus of relevant scientific literature. That corpus was also converted into a large number of embeddings stored in a vector database for RAG. The combination of fine-tuning and RAG enabled the LLM to leverage the findings, equations and data in the corpus to generate Python code and other useful artifacts through careful prompting.

As a first step in this evolving research effort, we demonstrate the ability to extract information from published articles, and to use AI to build Python codes to implement models in those papers. This approach obviates the need to develop equations of state that are often cumbersome and difficult to use, and the models can be trained to estimate fluid properties that have not been determined experimentally. As an example, we use LLM to extract information for the binary H_2O-CO_2 fluid system based on the equation of state developed by Kerrick and Jacobs (1981). The online calculator allows the user to independently select the fluid composition and PT conditions to examine.