

AI-Powered Data Extraction, Integration and Analysis on the Geochemistry of Returned Lunar Samples

Ping Wang¹, Dominick Pelaia², Zhitong Zou³, Stephen Xiao⁴, Amanda He³, Nathan He⁵, Flora Zhu⁶, Nathan Mao⁷, Jinqi Wu⁸, Yanran Chen⁹, Jian Wu¹⁰, Clive Neal¹¹, and Shichun Huang¹²

¹University of Tennessee, Knoxville

²L&N STEM Academy

³Valley Christian High School

⁴Farragut High School

⁵Ocean Lakes High School

⁶Great Neck South High School

⁷The Harker School

⁸Harvard University

⁹University of Notre Dame

¹⁰Old Dominion University

¹¹Univ Notre Dame

¹²University of Tennessee

December 24, 2025

Abstract

Our hands-on *planet+AI* curriculum development team, which includes high school students from across the nation, is developing a research-driven curriculum by taking advantage of emerging AI technologies for lunar research. With the increased interest in the Moon, especially with new sample returns, our ongoing work uses large language models (LLMs) and vision language models (VLMs) to accurately extract data from PDFs of published research on the chemical and isotopic compositions of lunar samples returned by the U.S. Apollo, the Soviet Union Luna, and the Chinese Chang'E missions.

We collect PDFs- scientific documents, including compendiums, catalogues, conference proceedings, journal articles, and supplementary documents by searching with targeted keywords. Using a range of traditional table extraction methods, such as PyMuPDF, Camelot, pdfplumber, pytesseract with pdf2image, we extract tabular data. Then, we explore the application of advanced AI models, including OpenAI's GPT-4 and GPT-4o, Google's Gemini 2.5 Flash, and Anthropic's Claude 4 Sonnet. These models are used for their Optical Character Recognition (OCR) capacities and VLMs strength in a four-step process: (1) OCR, (2) calling LLM APIs to extract tables, (3) parsing the APIs output and (4) reviewing the result. We use the *Uncertainty-Aware Complex Scientific Table Data Extraction* to detect extraction errors. We manually verify all extracted data and evaluate the quality and accuracy of outputs from each traditional and AI approach to establish a performance baseline and report our evaluations of the performance of each method. The compiled dataset will be cleaned and then analyzed by using a variety of multidimensional statistical approaches, such as, PCA, SparsePCA, t-SNE, and UMAP, dendrogram clustering. These exercises will reveal the distribution patterns of major and trace elements and mineral modes that define multiple distinct basalt groupings. We will develop AI agents that imitate the scientists' behaviors. The code, prompts, and compiled data will be made publicly available via our GitHub repository. This work is supported by the NSF DRL-2314155.