Goldschmidt 2026 Abstract
https://doi.org/10.46427/gold2026.37649

Generative AI Applied to Geochemistry Research: A Practical Case Study on Petroleum Characterization

GABRIEL BARBERES[1], ANNA ALICE GOMES[2] AND LUCAS BASTOS[3]

[1]Universidade Federal Fluminense
[2]Universidade Federal do Rio Grande do Sul
[3]Universidade do Estado do Rio de Janeiro

The global energy sector is increasingly adopting Generative AI (GAI), particularly Large Language Models (LLMs), to manage and interpret complex geoscientific datasets. Traditionally, determining thermal maturity and depositional environments in petroleum geochemistry requires the meticulous evaluation of multi-dimensional data. This study explores applying LLMs to automate oil sample characterization, providing geoscientists with a robust and accessible analytical workflow.

To evaluate this approach, we utilized diverse geochemical datasets comprising stable carbon isotopes, biomarker concentrations, and SARA fractions. The core dataset consists of 198 oil samples from the Potiguar Basin, Brazil [1]. To validate the models' global applicability and adaptability to diverse regional signatures, the methodology was extended to three additional datasets: 31 samples from the Mesopotamian Basin, Iraq [2]; 27 samples from the Southern Persian Gulf Basin, Iran [3]; and 57 samples from the Bohai Bay Basin, China [4]. Using freely accessible LLMs like Gemini and ChatGPT, we applied Chain of Thought (CoT) prompting to enforce logical, step-by-step reasoning. Furthermore, a Retrieval-Augmented Generation (RAG) approach was implemented using NotebookLM—a free and accessible AI tool. This grounded the models in verified specialized literature (e.g., The Biomarkers Guide 1 and 2), ensuring the accurate retrieval and application of classification thresholds.

Preliminary findings demonstrate that LLMs effectively extract reference values and categorize the core dataset into distinct depositional environments and sometimes are even able to identify maturation and biodegradation degrees. Rather than applying rigid thresholds, the AI captured nuanced source rock contributions, highlighting its ability to discern subtle variations in mixed-source systems often challenging traditional cross-plotting methods.

Crucially, this work emphasizes observability and control in AI-driven research. Structured methodologies with defined criteria and controlled inputs enhance algorithmic transparency, ensure reproducibility, and mitigate AI hallucination risks. By maintaining rigorous human oversight, LLMs offer a highly efficient alternative to traditional data analysis, significantly increasing productivity while retaining strict scientific rigor.

References

[1] Barberes et al. (2023), *Geosciences* 13, 282.
[2] Abeed et al. (2012) *Organic Geochemistry* 46, 113-126
[3] Alizadeh et al. (2017) *Organic Geochemistry* 111, 67-81
[4] Zhan et al. (2019) *Organic Geochemistry* 127, 37-49