

Classification of geological borehole descriptions using a domain adapted large language model

Katrijn Dirix¹, Hossein Ghorbanfekr², Pieter Jan Kerstens³

1. VITO [Environmental Intelligence, katrijn.dirix@vito.be]; 2. IMEC; 3. VITO [Data & Analytics]

Geological borehole descriptions contain rich, expert generated textual information that is essential for subsurface modelling, geotechnical applications, mineral exploration, and groundwater research. Yet, their unstructured natural language format poses a major challenge for automated interpretation. Traditional rule-based extraction methods, which rely on dictionaries and handcrafted regular expressions, struggle to capture the linguistic complexity and variability inherent in these descriptions.

In early 2024, we initiated research to explore whether large language models (LLMs) could more effectively extract structured lithological information from free text borehole descriptions. We developed GEOBERTje, a domain adapted Dutch language model derived from BERTje and further trained via masked language modelling on > 280,000 unlabeled geological descriptions. A smaller manually curated dataset of ~2,500 labeled intervals was then used to fine tune a multi label classifier capable of assigning a primary and two secondary lithology classes.

The resulting model demonstrates consistent improvements over three alternatives: a generic BERTje based classifier, traditional rule-based scripts, and a prompt engineering approach using GPT 4. GEOBERTje especially excels in capturing secondary and tertiary lithologies, where linguistic subtleties (e.g. nested descriptions, indirect grain size qualifiers, or multiple co dominant lithologies) often cause existing systems to fail or hallucinate. Unlike GPT 4, the domain adapted model rarely invents lithologies that are not present in the original descriptions, which is crucial for downstream geological modelling workflows.

This work is fully described in Ghorbanfekr et al. (2025), in which the methodology, accuracy benchmarks, and error analysis are presented in detail. For the conference contribution, we will include an updated performance benchmark incorporating GPT 5, allowing us to assess whether domain adaptation remains beneficial for this problem, or general frontier models have now caught up.

Ghorbanfekr, H., Kerstens, P. J., & Dirix, K. (2025). Classification of geological borehole descriptions using a domain adapted large language model. *Applied Computing and Geosciences*, 25, 100229.