HydroGeoLLM

Develop and test a dedicated Large Language Model for hydrogeological purposes

Department: Department of Physical Geography
Research group: Earth Surface Hydrology
Supervisor(s): Dr. Gualbert Oude Essink, Dr. Daniel Zamrsky, Dr. Jude King
Email address: gualbert.oudeessink@deltares.nl / g.h.p.oudeessink@uu.nl (prefer Deltares email)

Project description

Large Language Models (LLMs) have been one of the most significant technological innovations in recent years. Most currently used LLMs are trained on vast amounts of input data, making them highly applicable across many disciplines and highly efficient at performing tedious repetitive tasks. Unfortunately, this also means that these general LLMs are not yet well-suited enough for being applied in smaller research fields such as hydrogeology. One way to overcome this is to train a new LLM with relevant input data that is directly related to hydrogeology and provide context within the training dataset to improve the LLM capabilities. Once developed, this so-called hydrogeological LLM (HydroGeoLLM) can be used in a variety of applications such as data digitizing, conducting informed literature reviews or identifying knowledge gaps.

One of the main bottlenecks in hydrogeological projects is the time-consuming task of finding, harmonizing, and digitizing local datasets, which are often only available in paper or scanned PDF format. These repetitive tasks align perfectly with the capabilities of an LLM. To address this, we aim to train the HydroGeoLLM on a dedicated dataset specifically designed to extract hydrogeological information from various file formats. We have the following six tasks in mind:

1. The assistant will first establish the generic functionalities of LLMs. Note that over the last two years we have already tested LLMs for hydrogeological purposes, in collaboration with Deltares; these LLMs could serve as a basis.
2. A hydrogeological dataset (peer-reviewed articles, grey literature reports, including digitized hydrogeological data such as geological borelogs and groundwater salinity measurements) will be compiled (note that we have already collected several hundred articles and reports).
3. A subset of this dataset will be manually digitized into a format suitable for LLM training. This dataset will include classification of typical components e.g., figures, tables, and maps.
4. Contextual relationships will be defined for HydroGeoLLM by specifying whether individual components are interlinked, e.g., if a geological cross-section provided as a separate figure is marked on a map shown earlier in the same article / report.
5. Individual components of these figures will be annotated and classified into categories, e.g., legend, axis labels, units, and coordinate systems.
6. Finally, once the training dataset is complete, the HydroGeoLLM will be trained and its output quality evaluated. If time permits, multiple open-source LLMs could be tested to identify performance differences and select the best-performing model as a foundation for further tuning and quality improvement.

Job requirements

At least intermediate (or better advanced) level of Python coding is necessary. Any experience with LLMs is beneficial.