IUGS-sponsored meeting on Large Language Models in the Geological Sciences – for attendees

Note the meeting followed Chatham House Rules: participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed.

Date: Tuesday 16 July 2024
Location: Geological Society of London and online.

Attendees

15 in-person and 44 virtual, spanning interested organisations across DDE, IUGS, geoscience societies and publishers, academia, industry and governmental organisations.

Agenda and notes

1. An introduction on the objectives of the meeting was given.
2. Round table introductions from all attendees (in person and virtual).
3. How do geological science entities see Large Language Models (LLMs) in the future?
    a. Presentations and statements by a number of participants were given outlining their experience and observations of the ethical development and future of LLMs.
    b. An open discussion was undertaken which touched on a variety of key themes:
        i. There was positivity that this constructive meeting had been organised between key entities and that discussion regarding development of LLMs within the geosciences was in early stages, but critical for maintaining integrity of the science.
        ii. Common themes were the importance of using authoritative, unbiased data and trusted sources to maintain scientific research credibility and ensuring attribution to these. The importance of respecting intellectual property rights was raised and concern around the erosion of critical thinking.
        iii. It was suggested that geoscience has not engaged with AI systems such as LLMs as much as it could and may risk being left behind, however this is true in other disciplines. Industry is in some cases more advanced in its use than academia/governmental organisations.
        iv. Examples were given of other geoscience data projects that did not have clear use cases and were considered to under-deliver. The importance of having a clear use case for GeoGPT was stressed.
        v. The role of societies and publishers in delivering high-quality, trusted and specialised content was highlighted. There are hidden costs to ensuring the integrity, usability and discoverability of published research, and many societies are active in curating, maintaining and encouraging FAIR (Findable, Accessible, Interoperable, Reusable) data. This is a valuable source of training material for a geoscience-specific LLM but also important for the financial sustainability of societies, allowing them to advance the science and to deliver their charitable objectives. There was

the general view that this copyright and attribution of authors, researchers and publishers must be respected.

vi. Whilst this meeting was focussed on LLMs in the geological sciences, these by their very nature focus on language which is an interpretation of geology as a physical system. It was suggested that LLMs may not be the most effective model and thought should be given to physical models.

vii. Finally, the use of AI in delivering insights, more efficient workflows and new discoveries is inevitable and valuable. It is imperative that we find collaborative routes to defining ethics and frameworks, so we can utilise these opportunities.

4. Members of the GeoGPT and DDE team presented the current state and future plans for GeoGPT. Participants were pleased that this included the most transparent view of the model to date, including the shift in technology from a generative model to a RAG (Retrieval Augmented-Generation) model. The latest version of GeoGPT was presented including the choice of use of multiple foundational models (Qwen-2, LLAMA-3 and Mistral).
    a. There was an open discussion around several themes:
        i. There was a strong recommendation that the geoscience corpus used to further train GeoGPT from its base model is transparent and be made available to the community. The ability to interrogate the model and understand data sources is necessary in research since not all sources are made equal, and transparency will build trust within the research community and ultimately improve GeoGPT's effectiveness. This includes:
            1. The Geoscience Q&A pairs, as demonstrated by the Stanford Question Answering Dataset (SQuAD, a commonly used base for Natural Language Processing). There was also the recommendation that groups involved in the Q&A process are broadened to be more globally representative to avoid bias.
            2. The make-up of data sources e.g. X% Wikipedia, X% Common Crawl, X% Published Open Access content.
            3. The geoscience-specific training data to the article level. This could be crowd-sourced 'checked' by society publishers and other parties to assess the ethics and biases of the content used as well as the legalities, to build broad trust around a number of areas including existing copyright concerns and a mechanism put in place to remove IP infringing works.

        ii. Whilst general and commercial LLMs do not make their sources transparent for competitive reasons, GeoGPT is to be not-for-profit and freely available to all researchers around the world and therefore should not rely on these practices. It was felt that GeoGPT should be setting the gold standard of transparency and ethics.

        iii. As an international project, the governance, nature of the organisation and funding and potential government control of GeoGPT (currently largely funded by the Zhejiang Lab, China) must be transparent.

        iv. iv. There was discussion around the ambition of GeoGPT, its strategy and intended use cases which were felt to be a little unclear. There has

been quick progress, but development of these models is resource intensive
and there should be focus on understanding the key audience and their needs, then defining the strategy to ensure development is aligned and affordable.

v. The mutual benefits of LLM development in the geosciences was discussed. Societies and publishers have valuable training sources but may also benefit from the development of geoscience-specific LLMs. A prerequisite for any content licensing agreement is trust, transparency and effective governance to ensure terms are adhered to and renegotiated when development advances.

5. There was then a discussion regarding governance of GeoGPT and some draft recommendations for next steps were proposed by the Chair.
    a. It was suggested that no governance model for GeoGPT is required whilst the IUGS review of DDE begins.
    b. Development of GeoGPT can continue, with community engagement and testing at IGC in August but not for public launch.
    c. There will be a reconvening of this group in Oct or Nov where a progress report will be given on GeoGPT development and delivery against the recommendations. More industry involvement will be sought.
    d. Approval on the nature of involvement of IUGS will be required before public launch of GeoGPT. IUGS will separately determine its own endorsement of DDE.