Karl Oskar Magnus Holm

LLMs - The Death of GIS Analysis?

An Investigation into Using Large Language Models for GIS Data Analysis

Master Thesis in Computer Science and Geomatics, June 2024
Supervisor at NTNU: Hongchao Fan
External supervisors from Norkart: Alexander Salveson Nossum, Arild Nomeland, and Rune Aasgaard

Department of Geomatics
Faculty of Engineering
Norwegian University of Science and Technology

Abstract

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background and Motivation

1.2. Goals and Research Questions

1.3. Research Method

1.4. Contributions

1.5. Thesis Structure

2. Background Theory

3. Related Work

Mai et al. (2023) writes about the opportunities and challenges of Large Language Models (LLMs).

4. Datasets

4.1. Data Sources

4.2. Data Access

While leading Large Language Models (LLMs) are trained on increasingly large corpora, they are still only as familiar with a topic as the extent to which the training data exposes it to said topic. For instance, many LLMs are trained specifically to generate Python code, and are therefore fed with a vast number of Python code examples during training in the hopes of improving its performance on benchmarks like . As it is unlikely that the training data is evenly distributed among many different topics, it is useful to get familiarized with a model's capabilities in the areas of interest for a particular use case. In the case of an LLM-powered GIS agent that should be capable of performing geospatial analyses, it is useful to know what data formats such an agent is most comfortable to understand and work with.

Insert Python benchmark examples here

The upcoming experiments therefore seek to benchmark model performance on three different data access methods. The datasets from section 4.1 are presented to the model in three different ways, as subsection 4.2.1 through subsection 4.2.3 elaborate upon.

4.2.1. Files

The first method of presentation is to have the files from section 4.1 remain untouched.

4.2.2. STAC API

The second method is to create a STAC API from the data.

4.2.3. SQL Database

The third and final method is to load the data into a spatial SQL database and provide the model with database schemas that can be used to generate queries.

5. Architecture

6. Experiments and Results

6.1. Experimental Plan

6.2. Experimental Setup

6.3. Experimental Results

7. Evaluation and Discussion

7.1. Evaluation

7.2. Discussion

8. Conclusion and Future Work

8.1. Contributions

Bibliography

Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., Cundy, C., Li, Z., Zhu, R., & Lao, N. (2023). On the Opportunities and Challenges of Foundation Models for Geospatial Artificial Intelligence. https://doi.org/10.48550/arXiv.2304.06798

Appendices

A. Task Description from Norkart

Oppgåve med omfang som kan tilpassast både prosjekt og masteroppgåve

LLMs - GIS-analysens død

(kan justerast seinare)

BAKGRUNN

Nyere modeller for kunstig intelligens har demonstrert spesielt gode evner til å kunne lære av store mengder ustrukturert og semi-strukturert informasjon. ChatGPT fra OpenAi tok verden med storm – og chat-baserte systemer florerer. Kan chat-baserte modeller skapes for å hente ut GIS-data effektivt? Norkart har en stor dataplattform hvor brukere utvikler mot API'er som i stor grad har GIS/Geografiske data i bunn. GeoNorge er en stor datakatalog hvor brukere slår opp, eller søker kategorisert for å finne data. QGIS, Python, PostGIS, FME og andre verktøy brukes ofte til å gjennomføre GIS-analyser – hvor en GIS-analytiker/data-scientist gjennomfører dette.

«*Finn alle bygninger innenfor 100-meters-belte som er over 100 kvm og har brygger*»

Er dette mulig å få til med dagens tilgjengelige chat-modeller?

OPPGAVEBESKRIVELSE

Oppgaven har som hovedmål å undersøke hvordan nyere språkmodeller kan benyttes for å gjennomføre klassiske GIS-analyser ved å bruke standard GIS-teknologi som PostGIS/SQL og datakataloger (OGC API Records fks). Hva finnes av tilgjengelig chat-løsninger? Hvordan spesialtilpasse til GIS-anvendelser? Hvor presise kan en GIS-Chat bli?

Relevante delmål for oppgaven:

1. Kartlegge state-of-the-art
2. Utvikle proof-of-concepts
3. Analysere begrensninger og kvalitet

Oppgaven vil med fordel deles i prosjektoppgave og masteroppgave

- Prosjektoppgave
  - State-of-the-art: Ai-modeller og multi-modal maskinlæring
  - Innhente og utvikle datagrunnlag og API-tilgjengelighet
- Masteroppgave
  - Utvikle proof-of-concepts med tilgjengelige åpne modeller/teknologi
  - Gjennomføre eksperimenter for analyse av kvalitet

Detaljert oppgavebeskrivelse utvikles i samarbeid med studenten.

ADMINISTRATIVT/VEILEDNING

Ekstern veileder: (en eller flere)
Mathilde Ørstavik, Norkart
Rune Aasgaard, Norkart
Alexander Nossum, Norkart

Aktuelle vegleiarar og ansvarleg professor ve NTNU (den som har fagansvar nærast oppgåva):

Terje Midtbø (GIS, kartografi, visualisering)
Hongchao Fan (3D modellering, fotogrammetri, laser)