Karl Oskar Magnus Holm

LLMS - The Death of GIS Analysis?

An Investigation into Using Large Language Models to Make GIS Analysis Simpler, Faster, and More Accessible

Specialization Project in Computer Science and Geomatics, June 2024
Supervisor: 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

Sammendrag

Preface

Karl Oskar Magnus Holm
Trondheim, 24th October 2023

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background and Motivation

The field of Large Language Models (LLMs) is an emerging one. Fan et al. (2023, October, p. 2) found that the proportion of papers about LLMs[1] to arXiv has skyrocketed since 2020, with a six-times increase in percent points from 2022 to 2023. They write that prompt engineering has been extensively used as a way to improve code generation (Fan et al., 2023, October, p. 7)

1.2. Goals and Research Questions

The overarching goal of this specialization project is to investigate how Large Language Model (LLM) can be utilized to make GIS analysis simpler, faster, and more accessible.

Research questions:

1. What is the potential of LLM-based GIS analysis?
2. How can OGC API Features be used in an overlay analysis using ChatGPT-4?
3. How can we give ChatGPT-4 access to external tools?

Researching the potential of Location-Based Machine Learning (LLM) in GIS could uncover new methodologies for spatial analysis, predictive modeling, and decision-making. The aim with RQ 1 would be to assess the capabilities and limitations of integrating machine learning algorithms with geographic information systems for tasks like clustering, anomaly detection, or spatial prediction.

RQ 2 focuses on the feasibility of using the OGC - Features Standads in a typical overlay analysis within a conversational AI context like ChatGPT-4, having the users be able to express themselves using natural language queries. An answer to RQ 2 would describe how OGC APIs can be called and manipulated in a flexible manner during a conversation to perform spatial queries or analyses, like intersecting layers or filtering features based on certain criteria.

RQ 3 delves into the technical and ethical considerations of expanding ChatGPT-4's capabilities through integration with external tools, such as GIS software or data analytics platforms. It would explore options for secure and efficient data exchange, as well as assess the implications of such access in terms of data privacy and user consent.

I am now referencing RQ 2.

[1]Including articles whose title or abstract includes "LLM", "Large Language Model", or "GPT".

1.3. Research Method

1.4. Contributions

1.5. Thesis Structure

2. Theory

Chapter 2 of this specialization project will talk about the leading technologies in the field of Large Language Model (LLM), which in itself is a subfield of Natural Language Processing (NLP). section 2.1 will go over the leading LLM models, their strengths and weaknesses, and briefly mention differences in model architecture. Section 2.2 will name the most prominent providers of LLM services.

Figure 2.1 shows an actor map which includes stakeholders, providers, and other groups and organizations that could have some relevance to an autonomous LLM-based GIS-agent. Following subsections will go over the different groups and explain why they are included in the actor map.

2.1. LLMs

2.1.1. The GPT Family

Generative Pre-trained Transformers (GPTs) are a type of LLM first introduced by OpenAI in 2018 (Radford and Narasimhan, 2018).

2.1.2. The BERT Family

Bidirectional Encoder Representation from Transformers (BERT) is a family of language models developed at Google (Devlin et al., 2019, May), first presented in 2018.

2.1.3. Google Products

PaLM 2

Codey

2.2. Large Language Model (LLM) Providers

2.2.1. OpenAI

Being the most widely famous actor within the field of LLMs, OpenAI has gained great influence through their vast portfolio.

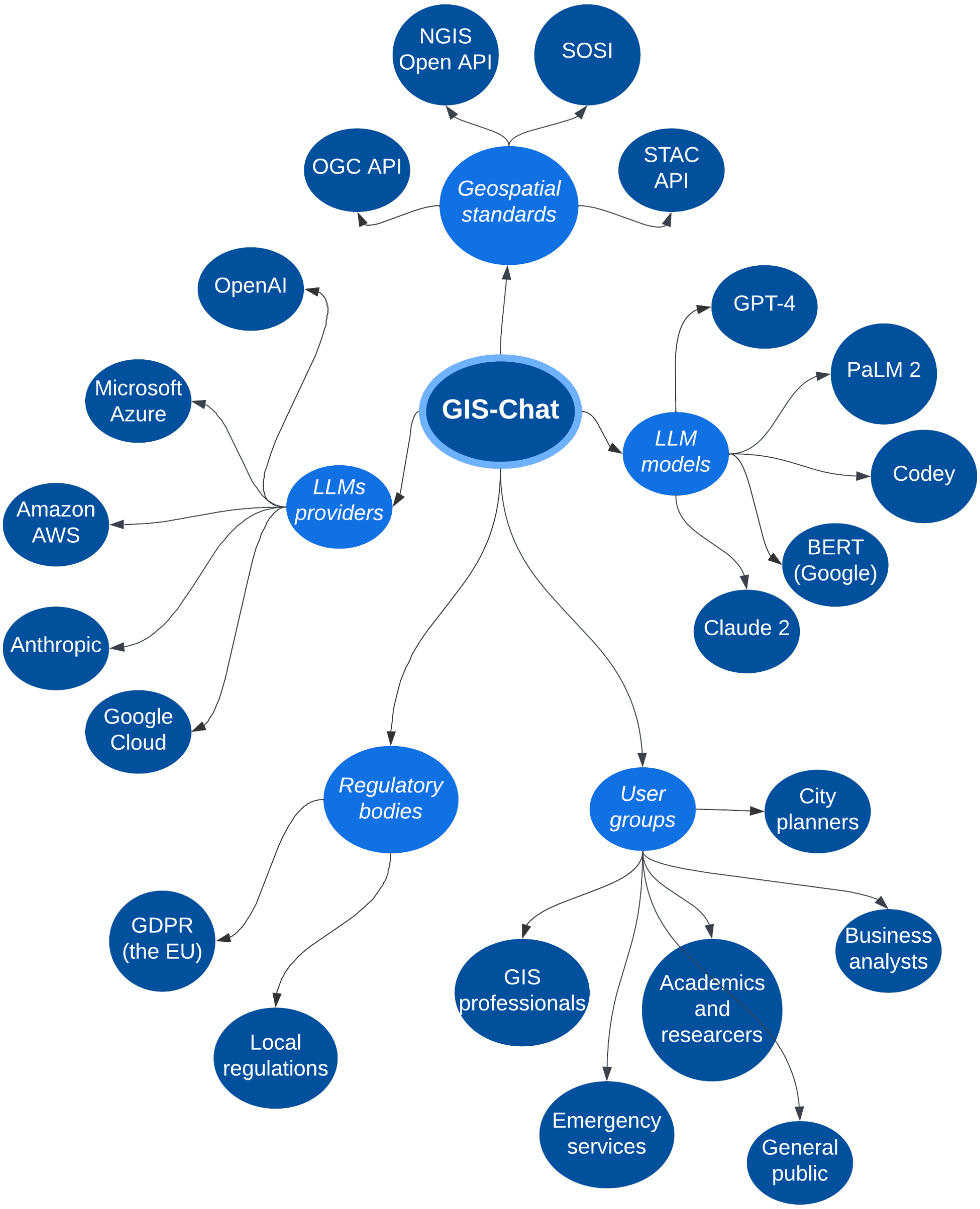

Figure 2.1.: Actor map for stakeholders, providers, and other groups and organizations that could have some relevance to an autonomous LLM-based GIS-agent.

2.2.2. Microsoft Azure

2.2.3. Google Cloud

2.2.4. Amazon Web Services (AWS)

2.2.5. Anthropic

2.3. Geospatial Standards

2.3.1. International Standardization Work

OGC Api Standard

The Open Geospatial Consortium (OGC) API Standards serve as the glue in the field of Geographic information system (GIS), paving the way for interoperability and data exchange between diverse systems. Leveraging common web protocols like HTML and supporting multiple data formats including JSON, GML, and HTML. The OGC API standard provides a modular architecture consisting of a core specification and various extensions. This modularity allows for flexibility, enabling users to customize their services according to specific needs. According to their webpages, they provide 80 different standards, each for a specific geospatial purpose. Notable examples are 3D Tiles, CityGML, GeoTiff, and OGC API - Features (OGC, 2023, October).

OGC API Standards function as modern replacements to older standards like WMS and WFS, and presents an evolved and more adaptable framework for spatial data operations, setting the stage for future innovations in the GIS domain.

STAC Api Standard

The SpatioTemporal Asset Catalog (STAC) API is a standardized way to expose collections of spatial temporal data for online search and discovery. Built upon a JSON core, it aims to be a uniform and flexible environment from which developers can customize the API infrastructure to their domain. STAC API provides a powerful query language that allows users to search by various parameters like time, location, and keywords, making widely applicable. The STAC community has also defined specification in order remove the complexity associated with having to create unique pipelines when consuming different spatial-temporary collection. The significance of the STAC API lies in its ability to democratize access to large volumes of geospatial data. By offering a common standard for data cataloguing and discovery, it reduces the barriers that often exist due to incompatible data formats. Developers or GIS professionals can take advantage of this through built-in tooling in QGIS, a desktop GIS for viewing, editing, and analysing spatial data, or through third-party packages in the Python and R programming languages. The API is also accessible through the command line interface when using GDAL (*STAC Tutorials* n.d.).

As OGC board member Chris Holmes puts it: "The STAC API implements and extends the OGC API — Features standard, and our shared goal is for STAC API to become a full OGC standard." (Holmes, 2021, January).

2.3.2. Norwegian Standardization Work

Geospatial standardization work has been on the agenda of Norwegian governing powers for decades and have materialized in frameworks/collaborations like Geovekst and Norge digitalt, as wells as the SOSI file format. Subsection 2.3.2 will delve into the work that has been done and what is expected for the future. The reasoning for the conclusion of this section is that the constraints of this specialization project is set by the Norwegian borders, and thus it is important to be aware of the standards that apply.

SOSI

Samordnet Opplegg for Stedfestet Informasjon (SOSI) is a Norwegian file format for storing and exchanging geospatial data. It was first introduced in 1987 and has since approached international standards, the most important arenas currently being ISO/TC 211 and OGC (Mardal et al., 2015). SOSI is the adopted Norwegian standard for creating and delivering digital geographic data, administered by the Norwegian Mapping Authority (Statens kartverk) (Mæhlum and Rød, 2023).

In a SOSI dataset, terrain points, lines, and polygons are represented by their coordinates and classified into various object types according to the SOSI object catalog standard. However, there are few GIS systems that can read SOSI data directly, so data in SOSI format usually needs to be converted to another GIS-readable data format (Mæhlum and Rød, 2023).

Geovekst

Geovekst is a collaborative initiative in Norway aimed at collecting, managing, and distributing geospatial information. Established in 1992, it is a partnership between national, regional, and local government bodies, as well as several private companies. Geovekst's primary focus is on creating a comprehensive, standardized geographical database for Norway that is easily accessible and updated regularly. It has played a vital role in various planning and development projects across the country, from urban planning to environmental conservation.

Unlike other geospatial initiatives, Geovekst emphasizes shared responsibilities and costs among its partners. This cooperative model ensures consistent data quality and efficient use of resources. It utilizes a variety of data sources, including aerial photographs, laser scans, and mapping, making it a rich resource for both public and private sectors. Moreover, its open-access policy allows for wider dissemination of geospatial information, thus encouraging innovation and informed decision-making across multiple disciplines.

Norge digitalt

Established in 2005, Norge Digitalt is a more recent framework compared to Geovekst and is the name of Norway's national spatial data infrastructure. Norge Digitalt primarily involves governmental bodies (national, regional, and municipal), but also educational and research institutions and companies with responsibilities on a nation-wide scale; examples include Telenor and local and regional energy companies (Norge Digitalt, 2023, p. 6). Norge Digitalt aims to coordinate and streamline all geospatial activities in Norway, making it easier for users to discover, access, and use spatial data.

One key feature of Norge Digitalt is its focus on international standards and interoperability. While Geovekst is primarily a national initiative, Norge Digitalt aims to integrate Norway's geospatial data with that of other European countries. It supports a wide range of data formats and follows international standards, including those set by the Open Geospatial Consortium (OGC). The framework also provides various tools and services, like metadata catalogues and web services, making it a comprehensive solution for geospatial data management and distribution in Norway.

2.4. User Groups

2.5. Regulatory Bodies

2.6. Related Work

2.6.1. Autonomous GIS

Li and Ning (2023, May) states that "autonomous GIS will need to achieve five autonomous goals: self-generating, self-organizing, self-verifying, self-executing, and self-growing.", and provide a "divide-and-conquer"-based method to address some of these goals. Furthermore, they propose a simple trial-and-error approach to addressing the self-verifying goal. They also highlight need of a memory system in a mature LLM-based GIS system, referring to the use of vector databases in autonomous agents like AutoGPT (Richard, 2023, October). Even with its shortages, the solution that (Li and Ning, 2023, May) provide, called LLM-Geo, is able to solve provide good solutions in various case studies by providing executable assemblies in a Python environment when provided with URLs to relevant data sets, along with a user-specified query.

Zhang et al. (2023, July) uses the Langchain framework (Chase, 2022, October) in order to combine different GIS tools in a sequence in order to solve different sub-goals, and focuses on using the semantic understanding and reasoning abilities of LLMs like (e.g., ChatGPT) to call externally defined tools, employing the LLM as an agent or controller. The authors take great inspiration from the AutoGPT framework (Richard, 2023, October). The externally defined tools are described (manually) by its name and description. Said description contains information about the input parameters and output types of the tools/functions. Tools are defined for geospatial data collection, data processing and analysis, and data visualization. The effectiveness of the system is showcased in four case studies.

2.6.2. Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) is tightly interwoven with explainable AI, being a framework for retrieving facts from an external knowledge base to allow a LLM-based agent access to accurate up-to-date information (Martineau, 2023, August). A common problem when working with language models, especially those desgined to be general-purpose, is hallucination; that is, when the model provides an answer that is completely wrong but in a very convincing manner. While progress is being made with newer models even the better ones, like GPT-4, gives an incorrect answer about 1 out 5 times, and even worse for certain categories of queries (for instance 'code' and 'business') (OpenAI, 2023, March, p. 10).

Langchain

Microsoft Semantic Kernel

AutoGPT

3. Experiments and Results

3.1. Experimental Plan

3.2. Experimental Setup

3.3. Experimental Results

4. Discussion

4.1. Evaluation

4.2. Discussion

5. Conclusion and Future Work

5.1. Contributions

5.2. Future Work

5.2.1. Test regime

In order to test the feasibility of different language models to serve as the brain of an autonomous GIS agent, a testing regime should be developed. In the examples of autonomous GIS agents described in the literature study of this report (see subsection 2.6.1), results have generally been presented in the form of case studies (Li and Ning, 2023, May; Zhang et al., 2023, July). This type of qualitative testing is entirely appropriate to showcase the possibilities of the technologies but may be insufficient when comparing performance of different systems. In the latter case a quantitative approach would probably be preferable.

One idea is to create a test dataset which consists of inputs and corresponding desired outputs of typical GIS tasks. Inputs would in this case be natural language queries inputted by a mock user, and the output would be what you would expect a GIS professional to return when given the same tasks/queries. Inputs should reflect the varying level of GIS knowledge in the different user groups (see section 2.4). Outputs could be files with typical geospatial extensions (.shp, .geojson, .sosi, etc.), or they could adhere to API schemas specified by geospatial standards (see section 2.3).

While the inputs should be fairly simple to construct there are several questions to be answered in regard to the outputs:

- How does one evaluate the accuracy of the output?
- How should the AI agent respond when the user does not specify an output file format?
- How does one evaluate the usefulness of outputs to questions that should not return geospatial files, e.g. answers to general questions about geo-related subjects?

These are questions outside the scope of this specialization project. They will, however, be pursued in my master thesis.

Acronyms

API Application Programming Interface.

AWS Amazon Web Services.

BERT Bidirectional Encoder Representation from Transformers.

GDAL Geospatial Data Abstraction Library.

GIS Geographic information system.

GML Geography Markup Language.

GPT Generative Pre-trained Transformer.

HTML HyperText Markup Language.

HTTP Hypertext Transfer Protocol.

ISO International Organization for Standardization.

JSON JavaScript Object Notation.

LLM Large Language Model.

NLP Natural Language Processing.

OGC Open Geospatial Consortium.

RAG Retrieval Augmented Generation.

SOSI Samordnet Opplegg for Stedfestet Informasjon.

STAC SpatioTemporal Asset Catalog.

TC Technical committee.

WFS Web Feature Service.

WMS Web Map Service.

XML Extensible Markup Language.

Bibliography

Chase, H. (2022, October). LangChain. Retrieved October 5, 2023, from https://github.com/langchain-ai/langchain

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Fan, A., Gokkaya, B., Harman, M., Lyubarskiy, M., Sengupta, S., Yoo, S., & Zhang, J. M. (2023, October). Large Language Models for Software Engineering: Survey and Open Problems.

Holmes, C. (2021, January). SpatioTemporal Asset Catalogs and the Open Geospatial Consortium. Retrieved October 23, 2023, from https://medium.com/radiant-earth-insights/spatiotemporal-asset-catalogs-and-the-open-geospatial-consortium-659538dce5c7

Li, Z., & Ning, H. (2023, May). Autonomous GIS: The next-generation AI-powered GIS.

Mæhlum, L., & Rød, J. K. (2023). SOSI. *Store norske leksikon*. Retrieved October 23, 2023, from https://snl.no/SOSI

Mardal, G., Borreb, M., Christensen, L., Jetlund, K., Ryghaug, P., & Hokstad, I. (2015). Nasjonal strategi for videreutvikling av SOSI.

Martineau, K. (2023, August). What is retrieval-augmented generation? Retrieved October 6, 2023, from https://research.ibm.com/blog/retrieval-augmented-generation-RAG

Norge Digitalt. (2023). *Generelle vilkår for Norge Digitalt-samarbeidet* (tech. rep.). Retrieved October 23, 2023, from https://www.geonorge.no/globalassets/geonorge2/avtaler-og-bilag-norge-digitalt/generelle-vilkar.pdf

OGC. (2023, October). OGC Standards. Retrieved October 23, 2023, from https://www.ogc.org/standards/

OpenAI. (2023, March). GPT-4 Technical Report.

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. Retrieved October 9, 2023, from https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035

Richard, T. B. (2023, October). AutoGPT: The heart of the open-source agent ecosystem. Retrieved October 5, 2023, from https://github.com/Significant-Gravitas/AutoGPT

STAC Tutorials. (n.d.). Retrieved October 23, 2023, from https://stacspec.org/en/tutorials/

Zhang, Y., Wei, C., Wu, S., He, Z., & Yu, W. (2023, July). GeoGPT: Understanding and Processing Geospatial Tasks through An Autonomous GPT.

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)