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

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

Data and Artificial Intelligence Group
Department of Geomatics
Faculty of Engineering
Norwegian University of Science and Technology

Abstract

Sammendrag

Preface

Karl Oskar Magnus Holm
Trondheim, 23rd 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

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?

I am now referencing RQ 2.

1.3 Research Method

1.4 Contributions

1.5 Thesis Structure

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

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. Chapter 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 good.

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 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.

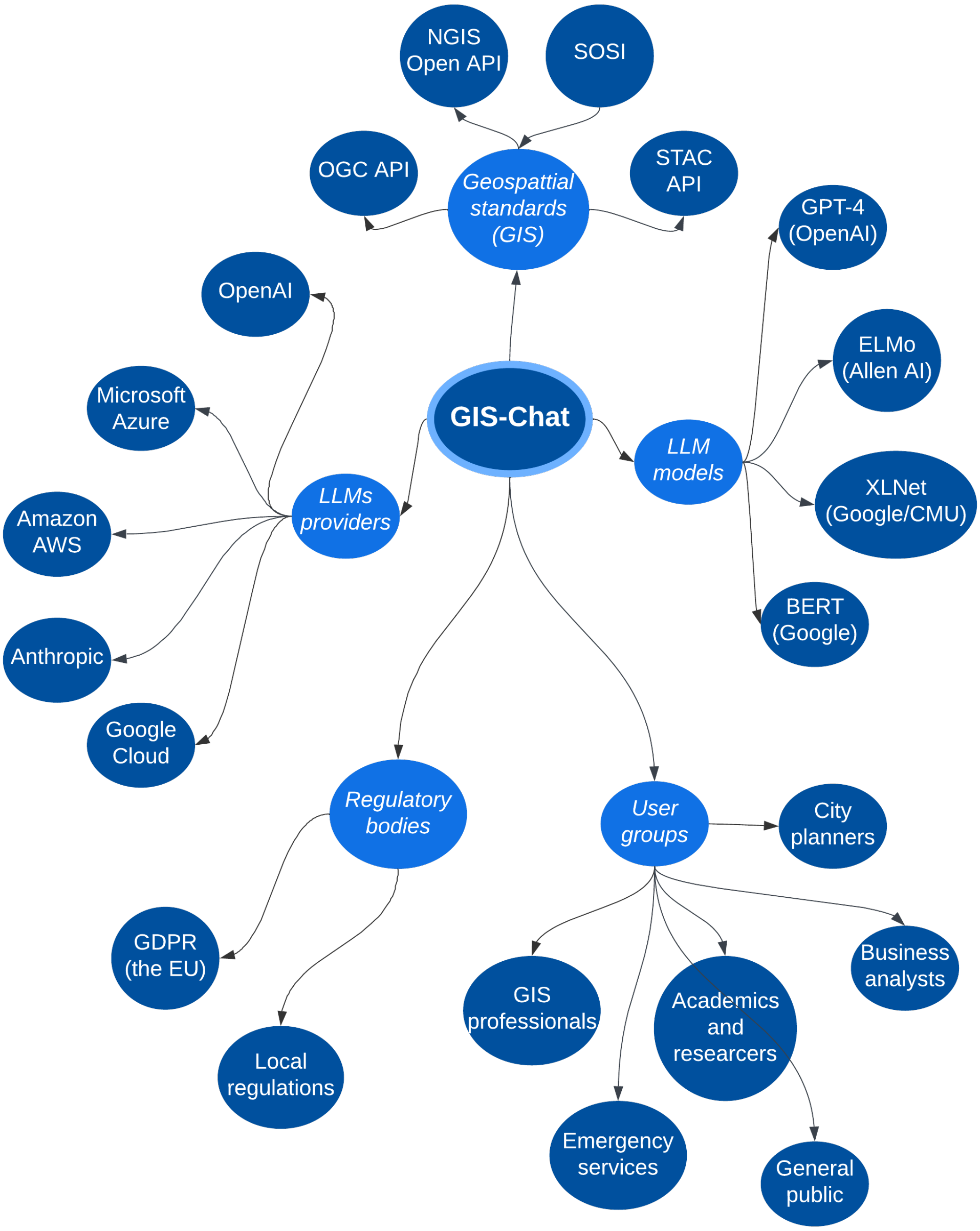

Figure 2.1: Actor map for stakeholders, providers, and other groups and organizations that could have some relevance to an autonomous LLM-based GIS-agent.

2.2.2 Microsoft Azure

2.2.3 Google Cloud

2.2.4 Amazon Web Services (AWS)

2.2.5 Anthropic

2.3 Geographic information system (GIS)

2.3.1 International Standardization Work

OGC Api Standard

The Open Geospatial Consortium (OGC) API Standards serve as the glue in the field of 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

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.

3 Experiments and Results

3.1 Experimental Plan

3.2 Experimental Setup

3.3 Experimental Results

4 Evaluation and Discussion

4.1 Evaluation

4.2 Discussion

5 Conclusion and Future Work

5.1 Contributions

Acronyms

API Application Programming Interface.

AWS Amazon Web Services.

BERT Bidirectional Encoder Representation from Transformers.

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.

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.

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/

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