

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 Nossom and Arild Nomeland

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



KART&IT

fix dou-
ble page
thing if
neces-
sary

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 spesialtilpasses 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 nærmast oppgåva):

Terje Midtbø (GIS, kartografi, visualisering)

Hongchao Fan (3D modellering, fotogrammetri, laser)

Abstract

Sammendrag

Preface

Karl Oskar Magnus Holm
Trondheim, 29th April 2024

Contents

Abstract	iv
Sammendrag	v
Preface	vi
List of Figures	ix
List of Tables	x
List of Code Snippets	xi
1. Introduction	1
1.1. Background and Motivation	1
1.2. Goals and Research Questions	1
1.3. Research Method	2
1.4. Contributions	2
1.5. Thesis Structure	2
2. Background Theory	3
2.1. Approaches to Language Modelling	3
2.1.1. Early Attempts: Statistical Models and Recurrent Neural Networks	3
2.1.2. Statistical Models	3
2.1.3. Recurrent Neural Networks	3
2.1.4. Attention and the Transformer Architecture	3
2.1.5. State Space Models	4
2.2. Function Calling LLMs	4
2.3. State-of-the-Art Large Language Models	5
2.3.1. The GPT Family	5
2.3.2. The Gemini Family	6
2.3.3. The Claude Family	6
2.3.4. Open-Source Alternatives	6
3. Related Work	7
3.1. LLM-based Systems in Geospatial Technologies	7
3.2. Agent Patterns	7
3.2.1. The Multi-Agent Pattern	7
3.2.2. Patterns for Retrieval-Augmented Generation	7

Contents

3.2.3. Patterns for Self-Reflection	7
3.3. LangChain	7
3.4. Geospatial Databases and Data Catalogues	8
4. Datasets	9
4.1. Data Sources	9
4.2. Data Access	11
4.2.1. Files	11
4.2.2. SQL Database	11
4.2.3. OGC API Features	12
5. Architecture	13
5.1. High-Level Application Architecture	13
5.1.1. LangChain Server	13
5.1.2. Redis for Conversations	14
5.1.3. PostGres + PostGIS	15
5.1.4. OGC API Features	15
5.1.5. Web UI	15
5.2. Agent Architecture	15
5.2.1. LangGraph Agent Implementation	15
5.2.2. Tools	15
5.2.3. Prompt Templating	15
6. Experiments and Results	20
6.1. Experimental Setup	20
6.1.1. Quantitative Approach: Benchmarking	20
Outcome Evaluation	20
Other Metrics	22
6.1.2. Evaluating Importance of Prompt Quality	23
6.1.3. Configuration and Hardware	24
6.2. Experimental Results	24
6.2.1. Quantitative Results	24
Test Outcomes	24
Other Metrics	24
Repeatability	29
6.2.2. Prompt Quality Test Results	29
7. Evaluation and Discussion	33
7.1. Evaluation	33
7.2. Discussion	33
8. Conclusion and Future Work	34
8.1. Contributions	34

Contents

8.2. Future Work	34
8.2.1. Automated Data Access	34
Bibliography	35
Appendices	36
A. Test Results	37
B. Code	43
B.1. Python Example	43

Acronyms

AI Artificial Intelligence.

API Application Programming Interface.

AWS Amazon Web Services.

BERT Bidirectional Encoder Representation from Transformers.

BLEU BiLingual Evaluation Understudy.

CLI Command Line Interface.

DAG Directed Acyclic Graph.

EU European Union.

FAISS Facebook AI Similarity Search.

FKB Felles KartdataBase.

GAN Generative Adverserial Network.

GDAL Geospatial Data Abstraction Library.

GDPR General Data Protection Regulation.

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.

Acronyms

LATS Language Agent Tree Search.

LLaMA Large Language Model Meta AI.

LLM Large Language Model.

LSH Locality-Sensitive Hashing.

LSTM Long Short-Term Memory.

MBPP Mostly Basic Python Programming.

MLM Masked Language Modelling,

MMI Maximum Mutual Information.

MMLU Multitask Language Understanding.

NLG Natural Language Generation.

NLI Natural Language Inference.

NLIDB Natural Language Interfaces for Database.

NLP Natural Language Processing.

NLU Natural Language Understanding.

NSDI National Spatial Data Infrastructure.

NSP Next Sentence Prediction.

OGC Open Geospatial Consortium.

OSM OpenStreetMap.

PPO Proximal Policy Optimization.

RAG Retrieval Augmented Generation.

RAM Random Access Memory.

RLHF Reinforcement Learning from Human Feedback.

RNN Recurrent Neural Network.

ROUGE Recall-Oriented Understudy for Gisting Evaluation.

SDK Software Development Kit.

Acronyms

SOSI Samordnet Opplegg for Stedfestet Informasjon.

SQL Structured Query Language.

SSD Solid-State-Disk.

STAC SpatioTemporal Asset Catalog.

TC Technical committee.

UI User Interface.

WFS Web Feature Service.

WMS Web Map Service.

XML Extensible Markup Language.

List of Figures

5.1.	Architecture overview	13
5.2.	Web UI	16
5.3.	Generic tool agent graph	17
5.4.	Example of a chat trace	19
6.1.	Outcome distribution between different agent types	25
6.2.	Duration per Agent Type	26
6.3.	Cost and token usage	28
6.4.	Correlation matrix for test result metrics	28
6.5.	Outcome distribution for different levels of GIS experience	30
6.6.	Comparison between novice and expert level prompting for making Ge-oGPT's OGC API Features agent calculate the number of trees along Munkegata in Trondheim	32

List of Tables

4.1. Datasets used in experiments	9
5.1. Summary of Server Endpoints	14
5.2. Overview of each agent's tools	18
6.1. Description of Success	21
6.2. Encoding for Test Outcome	21
6.3. Questions for quantitative tests	22
6.4. Standard Deviation by Agent Type	29
A.1. Test results for quantitative tests	37

List of Code Snippets

6.1. Python code aimed at computing the difference between the Oslo outline and water features within it	26
B.1. Python example	43

List of Code Snippets

1. Introduction

The introductory chapter will explain the motivation behind the thesis, as well as its goals and the research questions it will attempt to answer. Section 8.1 will list the main contributions of the thesis, and section 1.5 will give a high-level overview over the thesis.

1.1. Background and Motivation

The release of OpenAI's ChatGPT in November, 2023 generated a hype within the general population and chat-based systems are flourishing. ChatGPT — or rather the underlying GPT-3 and GPT-4 models — is an example of how modern Large Language Models (LLMs) can provide a natural language interface between human and machine. Furthermore, significant advancements have been made within code generation, which essentially allows the LLM to execute computational tasks that can be defined within a snippet of code. Paired with the LLM's ability to often correctly interpret the user's intent regardless of the preciseness of their problem formulation, even individuals with no prior programming experience can carry out computational tasks that require the execution of code.

GIS analysis has traditionally been reserved for GIS *experts*. GIS professional are commonly required to know their way around one or more Geographic Information Systems, in addition to being proficient in programming languages suitable to data science task, such as Python or R. Extensive domain knowledge is often also necessary when tackling GIS tasks, like knowing which data to use for a particular tasks and where to get them. All of these points — and more — are barriers to entry for people that wish to make use of powerful GIS tools for their particular purposes, but lack the technical know-how required to use them correctly. This challenge serves as the overall motivation behind this master's thesis, which will mitigate these issues by utilizing the vast background knowledge and code generation abilities of modern LLMs.

1.2. Goals and Research Questions

Deriving from the motivation described in the section above, the overarching goal of this master's thesis becomes to investigate the possibilities of utilizing LLMs to have a natural language interface with a system that is capable of solving GIS-related tasks. The hypothesis is that modern LLMs are embedded with an understanding of common GIS workflows, and that their code generation abilities are of such a level that they can accomplish tasks in an autonomous manner.

1. Introduction

Based on this overarching goal, three research questions have been constructed and are listed below:

1. Can an LLM-based system perform common GIS tasks?
2. What are core challenges when developing larger systems that rely on Large Language Models to control its logic flow?
3. What are state-of-the-art methods of creating autonomous LLM-based agents?

1.3. Research Method

Prior to this master’s thesis, a specialization project on the same topic was conducted. The specialization project — as detailed in (Holm, 2023) — was of a theoretical character predominantly served as a literature study leading up to the master’s thesis. The research questions listed in the above section, however, call for a different and more practical approach. Specifically, RQ1 and RQ2 can only be addressed by drawing on insights gained from the attempt to develop such an LLM-based GIS system. Therefore, this master’s thesis will revolve around a “proof of concept” and evaluating the usefulness of the system, as well as lessons learned from the development process.

RQ3, on the other hand, will be a continuation of the theoretical work done in the specialization process, and will serve as a foundation for the “proof of concept” development process.

1.4. Contributions

1.5. Thesis Structure

2. Background Theory

Chapter 2 will lay a theoretical basis for the work done in this master thesis, providing the user with the required understanding in order to understand the contributions of the work. Section 2.1 will explain the theoretical basis of the component which most modern Large Language Models (LLMs) are based upon — namely the Transformer — and the attention mechanism within it. The section will also touch upon a new approach to language modelling called *selective state space modelling*, which has yielded very promising results for small LLMs.

Parts of the Background chapter is reused material from the specialization project (Holm, 2023) preceding this master thesis. Below are the sections in question, together with a description of the extent to which, and how, the material is reused:

- *Subsection 2.1.4: Reused without modification.*
- *Subsection 2.3.1: Reused without modification.*

2.1. Approaches to Language Modelling

Subsection 2.1.4 will explain the theoretical basis behind most modern LLMs, which are based upon the attention mechanism built into the Transformer architecture. Subsection 2.1.5 will explain modern state space-modelling approaches and why they may have a potential greater than Transformer-based models. First, however, subsection 2.1.1 will delve into earlier attempts at language modelling.

2.1.1. Early Attempts: Statistical Models and Recurrent Neural Networks

2.1.2. Statistical Models

2.1.3. Recurrent Neural Networks

2.1.4. Attention and the Transformer Architecture

Vaswani et al. (2017) managed to achieve new state-of-the-art results for machine translation tasks with their introduction of the Transformer architecture. The Transformer has later been proved effective for numerous downstream tasks, and for a variety of modalities. Titling their paper *Attention Is All You Need*, Vaswani et al. suggest that their attention-based architecture renders network architectures like Recurrent Neural

2. Background Theory

Networks (RNNs) redundant, due to its superior parallelization abilities and the shorter path between combinations of position input and output sequences, making it easier to learn long-range dependencies (Vaswani et al., 2017, p. 6).

The Transformer employs self-attention, which enables the model to draw connections between arbitrary parts of a given sequence, bypassing the long-range dependency issue commonly found with RNNs. An attention function maps a query and a set of key-value pairs to an output, calculating the compatibility between a query and a corresponding key (Vaswani et al., 2017, p. 3). Looking at Vaswani et al.'s proposed attention function (2.1), we observe that it takes the dot product between the query Q and the keys K , where Q is the token that we want to compare all the keys to. Keys similar to Q will get a higher score, i.e., be *more attended to*. These differences in attention are further emphasized by applying the softmax function. The final matrix multiplication with the values V (the initial embeddings of the input tokens) will yield a new embedding in which all individual tokens have some context from all other tokens. We improve the attention mechanism by multiplying queries, keys, and values with weight matrices that are learned through backpropagation. Self-attention is a special kind of attention in which queries, keys, and values are all the same sequence.

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (2.1)$$

Attention blocks can be found in three places in the Transformer architecture (Vaswani et al., 2017, p. 5) (I will use machine translation from Norwegian to German as an example):

1. In the encoder block to perform self-attention on the input sequence (which is in Norwegian)
2. In the decoder block to perform self-attention on the output sequence (which is in German)
3. In the decoder block to perform cross-attention (also known as encoder-decoder attention) where each position in the decoder attends to all positions in the encoder

The Transformer represented a breakthrough in the field of Natural Language Processing (NLP), and is the fundamental building block of modern LLMs, most famous of which are the GPT's.

2.1.5. State Space Models

2.2. Function Calling LLMs

Function calling—first introduced by OpenAI (Eleti et al., 2023)—allows developers to provide function definitions to an LLM and have said LLM output a JSON object containing the name of one or more of the functions provided, as well as suitable arguments

2. Background Theory

to these. Made possible through fine-tuning models to detect when functions should be calling, function calling makes it possible to give an LLM *hooks* into the real world, and provides a more reliable way for developers to integrate LLMs into applications.

Possible use cases include using functions provide correct and up-to-date information that would otherwise require extensive training and fine-tuning. Having the LLM use function calling for information retrieval also make them more transparent, making it possible to trace a claim back to its source, something that is normally a difficult feat with LLM. Another use case might be code execution. One could imagine a rather simple function `execute_python_code(code: string) -> string` that takes Python code as a string and returns the standard output that results from executing that code. This is likely the principle behind products like OpenAI's Data Analysis mode (previously Code Interpreter), in which ChatGPT functions as a code executing agent that can generate, execute, and self-correct its own code. Similar functions could be constructed for SQL, making it possible for LLMs to work against relational databases. As Eleti et al. (2023) describes, function calling can also be used to extract structured data from text.

2.3. State-of-the-Art Large Language Models

2.3.1. The GPT Family

Generative Pre-trained Transformer (GPT) is a type of LLM that was introduced by OpenAI in 2018 (Radford and Narasimhan, 2018). Specifically designed for text generation, a GPT is essentially a stack of Transformer *decoders*. It demonstrates through its vast pre-training on unlabelled data that such unsupervised training can help a language model learn good representations, providing a significant performance boost while alleviating the dependence on supervised learning. While the original Transformer architecture as described by Vaswani et al. (2017) was intended for machine translation—thus having encoders to learn the representation of the origin language representation of a given input sequence and decoders to learn the representation in the target language and perform cross-attention between the two—the GPT is designed only to *imitate* language. This is why there are no encoders to be found in the GPT architecture, only decoders. The model employs masked multi-head attention (running the input sequence through multiple attention heads in parallel), and is restricted to only see the last k tokens—with k being the size of the context window—and tasked to predict the next one.

Training consists of two stages: unsupervised pre-training and supervised fine-tuning. The former is used to find a good initialization point, essentially teaching the model to imitate the corpora upon which it is trained. This results in a model that will ramble on uncontrollably, just trying to elaborate upon the input sequence it's given to the best of its knowledge. This will naturally produce undefined behaviour, and it is therefore necessary to fine-tune the model on target tasks in a supervised manner. Radford and Narasimhan (2018, p. 4) explain how the model can be fine-tuned directly on tasks like text classification, but how one for other tasks needs to convert structured inputs into ordered sequences because the pre-trained model was trained on contiguous sequences

2. Background Theory

of text. In the case of ChatGPT, OpenAI used Reinforcement Learning from Human Feedback (RLHF) by employing a three-step strategy: first training using a supervised policy, then using trained reward models to rank alternative completions produced by ChatGPT models, before fine-tuning the model using Proximal Policy Optimization (PPO), which is a way of training AI policies. This pipeline is then performed for several iterations until the model produces the desired behaviour (OpenAI, 2022).

2.3.2. The Gemini Family

2.3.3. The Claude Family

2.3.4. Open-Source Alternatives

3. Related Work

3.1. LLM-based Systems in Geospatial Technologies

3.2. Agent Patterns

LLM-based agents can be implemented in many ways, and researchers have developed a plethora of *agent patterns* that seek to improve upon areas where LLMs tend to be less effective. This includes patterns for retrieval of external data, as well as multi-agent patterns. Such patterns can improve the performance of LLM-based agents within any domain, and should also be considered when developing one for geospatial purposes. The following sections will shortly explain some of the patterns that were considered in the development of GeoGPT.

3.2.1. The Multi-Agent Pattern

The multi-agent pattern that takes inspiration from human collaboration in that it is made up from multiple specialized agents that work together to achieve some objective. There have been several implementations of the pattern, with certain differences. MetaGPT (Hong et al., 2023) is a LLM-based multi-agent system consisting of agents with human-level domain expertise. Using an assembly line paradigm, where the overall goal is divided into subtasks, Hong et al. showed that MetaGPT could generate more coherent solutions compared to the previous state-of-the-art multi-agent systems. At the time of release, MetaGPT set a new state-of-the-art performance on the HumanEval and MBPP benchmarks (Hong et al., 2023, p. 7), demonstrating the potential of the multi-agent pattern.

3.2.2. Patterns for Retrieval Augmented Generation

3.2.3. Patterns for Self-Reflection

3.3. LangChain

LangChain (LangChain AI, 2022) is an open-source project that provides tooling which simplifies the way developers interface with Large Language Models (LLMs). This tooling includes composable tools and integrations that can be used to build prompts for LLMs, as well as off-the-shelf chains that perform higher level tasks. Chains are Directed Acyclic Graphs (DAGs) — or sequences of runnables — that take an input and produces an output. A runnable can be a prompt template with template literals that are substituted

3. Related Work

with values that are passed into the runnable. The output is the template with the template literals filled in. This output can then be chained into an LLM runnable calls a language model using the prompt template. The output from the LLM runnable could then be passed into an output parser, e.g. a JSON parser, that ensures that the chain outputs a JSON object. Such chains are the buildings blocks that make up LangChain.

Common use cases for LangChain are:

- Building chatbots for question answering that use semantic retrieval from document store
- Creating agents with access to external tools by leveraging function calling (see section 2.2)
- Creating code executing agents for Python, SQL, or other programming languages

In January 2024, LangChain AI rolled out a new framework called LangGraph which builds on top of the LangChain ecosystem. While the chains commonly found in LangChain are good for DAG workflow, they are not suited to creating cyclic graphs. LangGraph can be used to add cycles to LLM applications, which are important for agent-like behaviours (LangChain AI, 2024). A graph in LangGraph is a set of nodes that pass some state around, state that can be modified by each node. The nodes are connected together by edges that define what node can succeed another node. These edges can also be conditional, which routes execution to a given node based on the output from a function giving the current state. This allows for complex logic and simplifies implementation of advanced agent patterns, some of which are discussed in section 3.2.

3.4. Geospatial Databases and Data Catalogues

4. Datasets

With the interest of investigating the ability of a Large Language Model (LLM)-based system to perform geospatial analysis, relevant datasets should be accessible to said system. Section 4.1 provides a description of the datasets used in the experiments. Furthermore, it was decided to explore different access channels to this data. Section 4.2 elaborates on this.

4.1. Data Sources

A total of eighteen datasets were used in the experiments. The data was downloaded from Geofabrik’s website¹. Geofabrik — German for “geo factory” — is a company that “extract, select, and process free geodata”. They have gathered data from OpenStreetMap and published them as a collection of shapefiles, dividing them into categories such as “places of worship”, “points of interest”, and “traffic”. Data can be downloaded for different regions of the world, and for experiments conducted in this thesis, data for Norway was used. Table 4.1 lists all datasets used, along with a short description of their contents. Common for all datasets are their *fclass* attribute, which is short for *feature class*. Some datasets have additional attributes, such as the *maxspeed* attribute in the road data and the *type* attribute in the building data.

Table 4.1.: Datasets used in experiments

Dataset	Data Type	Description
Buildings	Polygon	Contains building outlines. Its <i>type</i> attribute can have values like <i>house</i> , <i>university</i> , and <i>restaurant</i> .
Land Use	Polygon	Represents areas designated to different purposes and activities. Its <i>fclass</i> attribute can have values like <i>forest</i> , <i>farmland</i> , and <i>residential</i> .

Continued on next page

¹<https://download.geofabrik.de/europe/norway.html>

4. Datasets

Table 4.1 continued from previous page

Dataset	Data Type	Description
Natural	Point	Contains outlines of various objects found in nature. Its <i>fclass</i> attribute can have values like <i>beach</i> , <i>glacier</i> , and <i>cave_entrance</i> .
Natural	Polygon	Similar to the point data equivalent.
Places of Worship	Point	Common values for <i>fclass</i> attribute: <i>christian</i> , <i>buddhist</i> , and <i>muslim</i> .
Places of Worship	Polygon	Similar to the point data equivalent.
Places	Point	Common values for <i>fclass</i> attribute: <i>farm</i> , <i>village</i> , and <i>island</i> . Repeated entries trimmed for brevity.
Places	Polygon	Similar to the point data equivalent.
Points of Interest	Point	Common values for <i>fclass</i> attribute: <i>tourist_info</i> , <i>bench</i> , and <i>kindergarten</i> .
Points of Interest	Polygon	Similar to the point data equivalent.
Railways	Lines	Common values for <i>fclass</i> attribute: <i>rail</i> , <i>subway</i> , and <i>tram</i> . Also has True/False attributes saying if a given line segment is a bridge or a tunnel.
Roads	Lines	Common values for <i>fclass</i> attribute: <i>rail</i> , <i>subway</i> , and <i>tram</i> . Has additional attributes <i>oneway</i> , <i>maxspeed</i> , <i>bridge</i> , and <i>tunnel</i> .
Traffic	Point	Common values for <i>fclass</i> attribute: <i>crossing</i> , <i>street_lamp</i> , and <i>parking</i> .
Traffic	Polygon	Common values for <i>fclass</i> attribute: <i>parking</i> , <i>pier</i> , and <i>dam</i> .
Transport	Point	Common values for <i>fclass</i> attribute: <i>bus_stop</i> , <i>ferry_terminal</i> , and <i>railway_station</i> .
Transport	Polygon	Similar to the point data equivalent.
Water	Polygon	Common values for <i>fclass</i> attribute: <i>water</i> , <i>wetland</i> , and <i>river_bank</i> .

Continued on next page

4. Datasets

Table 4.1 continued from previous page

Dataset	Data Type	Description
Waterways	Lines	Common values for <i>fclass</i> attribute: <i>stream</i> , <i>river</i> , and <i>canal</i> .

4.2. Data Access

While leading 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
bench-
mark
exam-
ples
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 4.2.2 elaborate upon.

4.2.1. Files

The first method of presentation is to have the files from section 4.1 remain untouched. The datasets were stored such that each dataset has its own folder. This is because some of the file types used require multiple files in order to correctly store the data—for instance, the shapefile format, which has three mandatory files: .shp, which contains the actual feature geometry; .shx, which provides a positional index of the feature geometry; and .dbf, which holds attributes for each shape.

reference

4.2.2. SQL Database

The second method used is to load the data into a spatial SQL database and provide the model with database schemas that can be used to generate queries. The datasets were uploaded to a dockerized PostGIS database using QGIS's DB Manager plugin.

Some of the datasets come in the GML data format, which can include multiple layers with potentially different geometries. For this reason, they cannot be loaded directly into a PostGIS database such that they are stored in the same database table. Furthermore, several of the layers in the multi-layer GML files are irrelevant for most analysis situations. For instance, the flood zone data were downloaded as a multi-layer GML file and includes a total of eight layers: polygon and multi-line border for the analysis area, polygon layers

4. Datasets

for rivers, ocean surfaces, and lakes, polygon and multi-line border for the flood zones, and a layer containing cross-sectional profile lines for the rivers. The quick clay dataset was similar. For brevity, a decision was made not to load all these layers into the PostGIS database. Only the polygon for the flood zones and two polygon layers from the quick clay dataset were loaded into the database.

4.2.3. OGC API Features

The third method for data access is to use the OGC API Features standard.

5. Architecture

5.1. High-Level Application Architecture

A microservice architecture was employed in order to simplify development and separate concerns between the different microservices. The services are deployed as Docker Containers, and they are orchestrated using Docker Compose. Figure 5.1 shows how the application is divided into five distinct services, and the direction of information flow between these.

quickly explain docker and why it is useful here

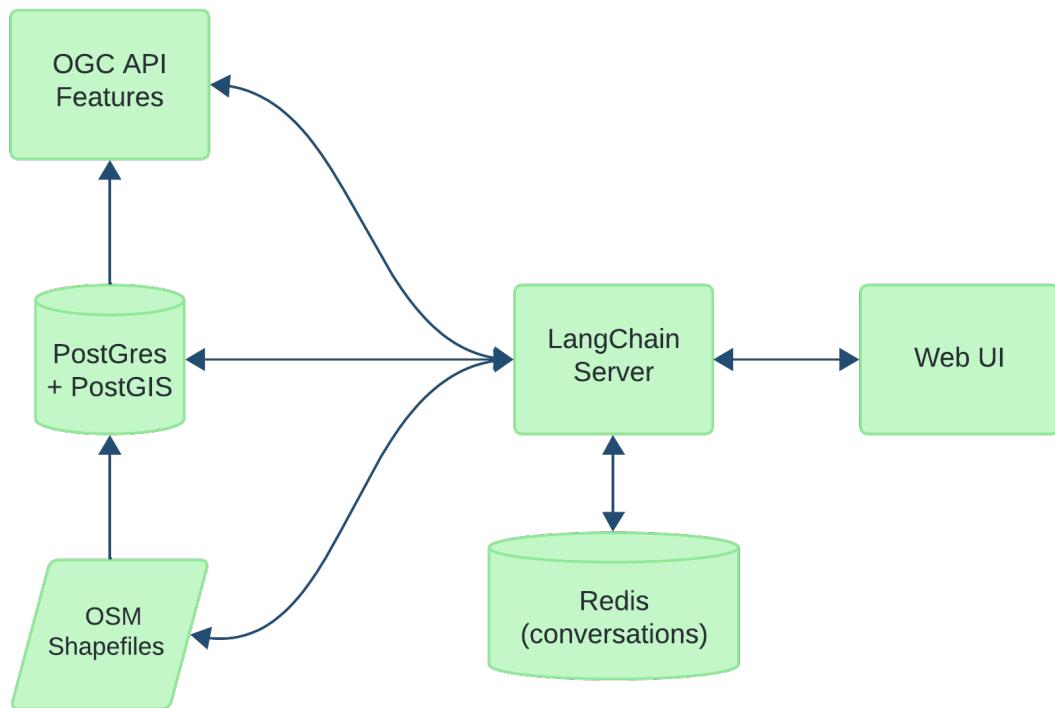


Figure 5.1.: Architecture overview

5.1.1. LangChain Server

The *LangChain Server* service is the heart of the application, and is where the Large Language Model (LLM)-related logic is situated. It is responsible for taking requests from

5. Architecture

the *Web UI* and returning suitable responses in what becomes a client-server architecture between the two services. Table 5.1 show the endpoints exposed by the server and how they can be used by a client.

Table 5.1.: Summary of Server Endpoints

Endpoint	Method	Description
/session	GET	Takes a <code>session_id</code> as a query parameter, allowing the client to continue on a pre-existing session.
/session	POST	Creates a new session with an empty conversation.
/streaming-chat	GET	Endpoint for chatting the LLM. Takes a <code>message</code> as a query parameter and returns an event stream, allowing for token streaming from server to client.
/update-map-state	POST	Send the state of the client map to the server. Keeps the server updated on what layers are present in the map, their color, etc.
/geojson	GET	Takes a <code>geojson_path</code> as a query parameter. Allows the client to retrieve a given GeoJSON file that is stored in the working directory on the server.
/history	GET	Used to retrieve the chat history of the current session.
/upload	POST	Allows the client to upload one or more files to the working directory on the server.

5.1.2. Redis for Conversations

Redis (Sanfilippo, 2009) is a fast in-memory database that is often applied as a caching database that sits on top of some persistent database. It can also be used for vector-based storage and as a simple NoSQL database. The latter option is the way it is used in GeoGPT’s architecture, and its only purpose is to store conversations. Whenever a user starts a conversation with GeoGPT an object with a unique session ID is stored to the Redis database. This object holds an array that represents the conversation. This array is written to every time either the human or GeoGPT produces a message.

Storing messages — either in memory as a simple array or in a database like Redis — is crucial to enable multi-message conversations. In order for a LLM to act as a

5. Architecture

conversational agent, some sort of chat history needs to be prepended to the prompt. In the case of GeoGPT the entire chat history is prepended. This has the advantage of providing the LLM with the complete context of the chat history, but the disadvantage of potentially bloating the context window from which it is supposed to generate tokens. Therefore, as the chat becomes longer each new token will be both more expensive and take more time to get generated. A long chat history could also make the resulting prompt exceed the token limit of the LLM, or it could confuse model by providing it with messages no longer relevant to the conversation. These issue was not considered in great detail for this project, and are left for future work.

5.1.3. Postgres + PostGIS

5.1.4. OGC API Features

On top of the PostGIS database containing OpenStreetMap (OSM) data is a RESTful geospatial feature server called *pg_featureserv* (CrunchyData, 2024).

5.1.5. Web UI

The user interface is made with SolidJS. By design, it is very minimal. One of the goals of the project is to simplify the way we do GIS analysis. One of the key design goals was therefore to make the interface as familiar to the user as possible and lowering the chance of the user doing something wrong. The chat interface was designed to imitate the interface of OpenAI's

5.2. Agent Architecture

Three different agent were implemented for GeoGPT.

Subsection 5.2.1 will explain how the base agent was implemented using LangGraph, while subsection 5.2.3 will present a detailed description of how GeoGPT builds suitable prompts that provide the LLM with the context required to solve the problem. Finally, subsection 5.2.2 will present the different tools that were used for the different agents.

5.2.1. LangGraph Agent Implementation

5.2.2. Tools

5.2.3. Prompt Templating

5. Architecture

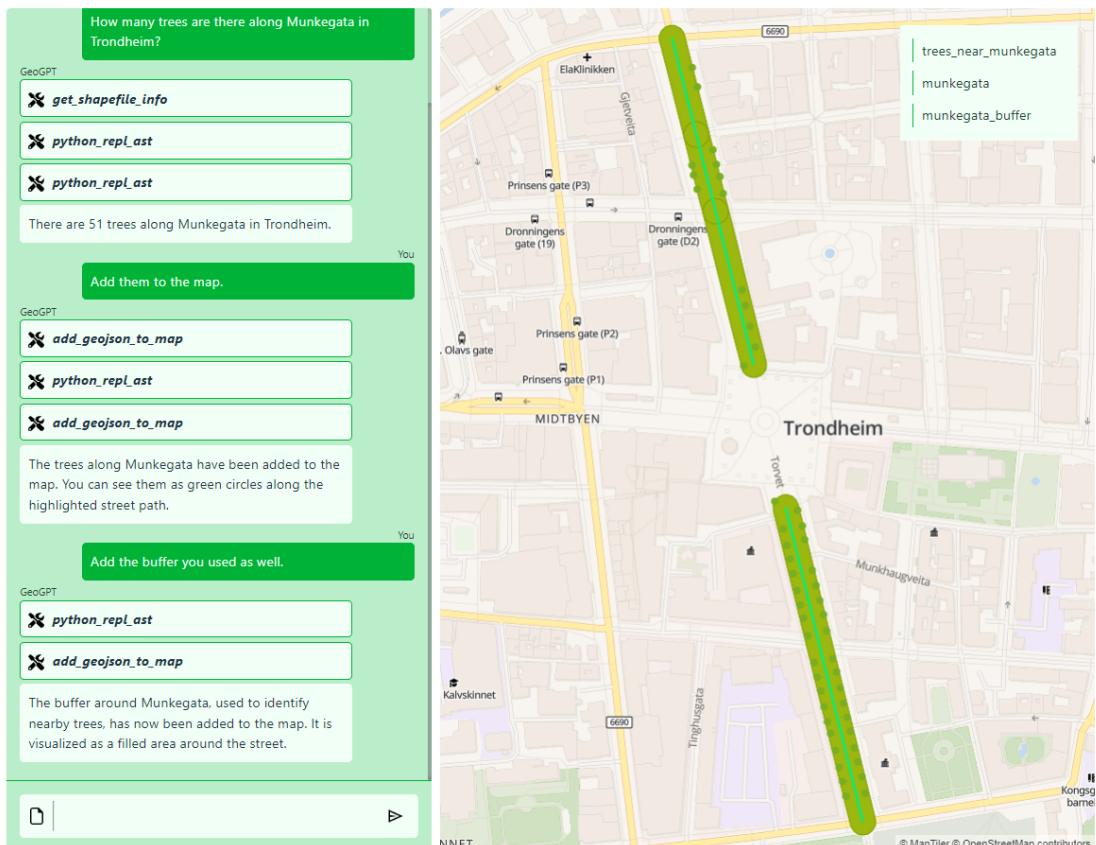


Figure 5.2.: Web UI

5. Architecture

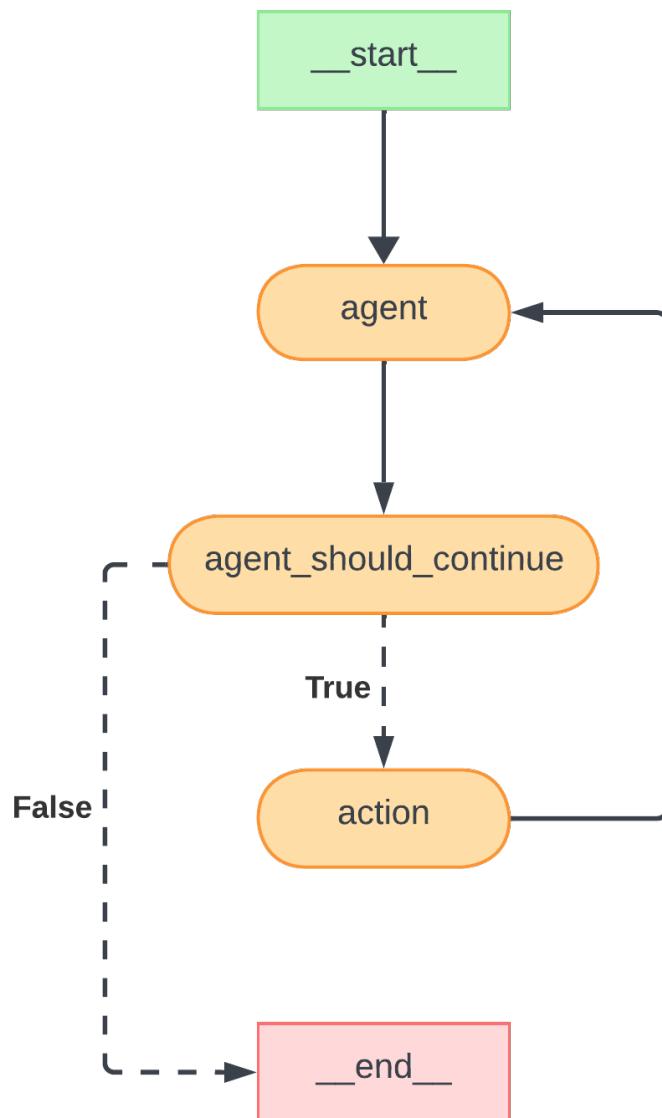


Figure 5.3.: Generic tool agent graph

5. Architecture

Table 5.2.: Overview of each agent's tools

Agent Type	Tools
OGC API Features	<code>get_collections_info</code> <code>list_collections</code> <code>query_collection</code> <code>python_repl_ast</code> <code>add_geojson_to_map</code>
Python	<code>python_repl_ast</code> <code>add_geojson_to_map</code>
SQL	<code>sql_db_query</code> <code>sql_db_list_tables</code> <code>sql_db_schema</code> <code>add_geojson_to_map</code>

5. Architecture

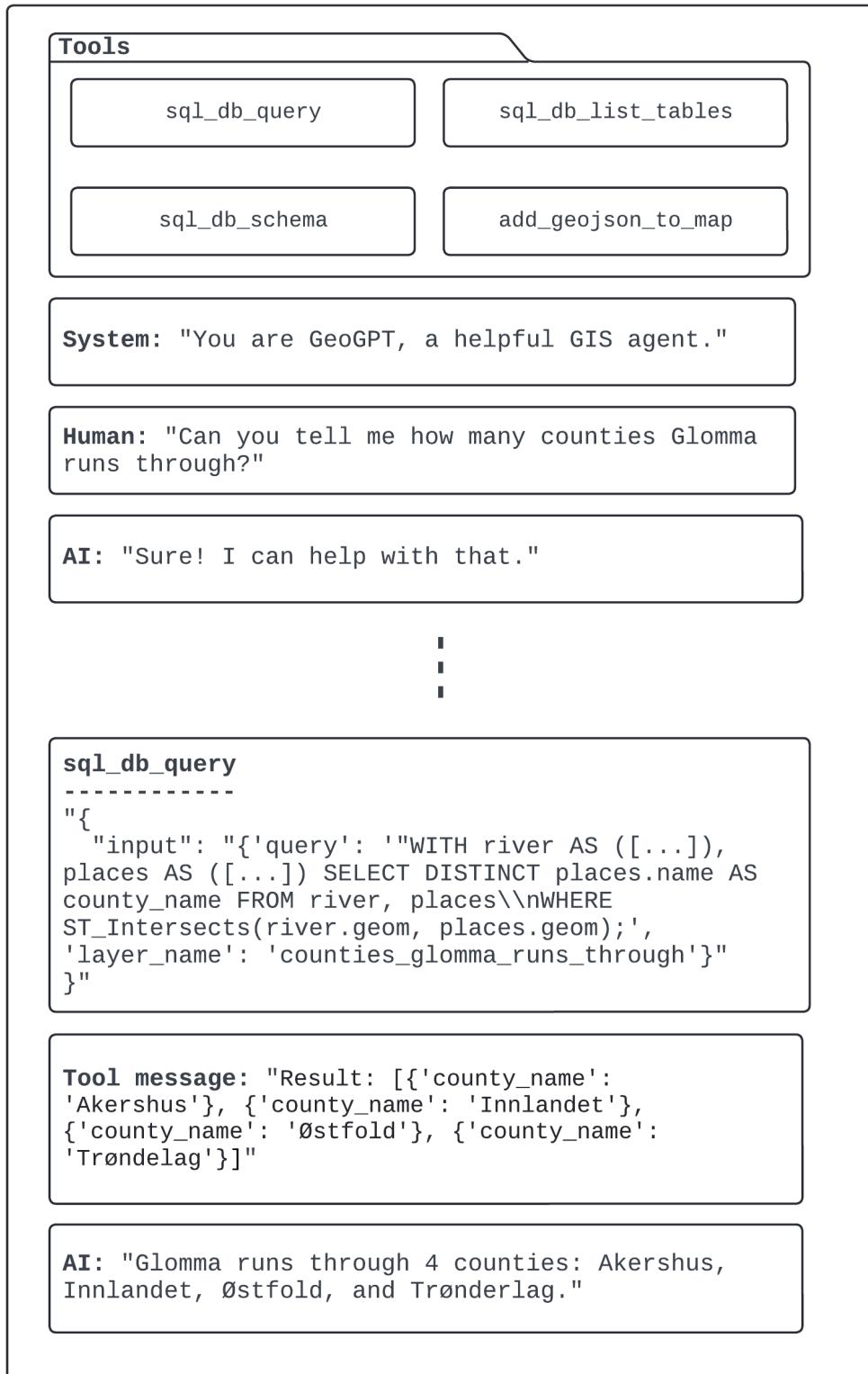


Figure 5.4.: Example of a chat trace

6. Experiments and Results

6.1. Experimental Setup

The experiments conducted to evaluate the performance of GeoGPT on geospatial tasks are divided into two approaches. These are presented in subsection 6.1.1 and subsection 6.1.2.

6.1.1. Quantitative Approach: Benchmarking

The first approach seeks to evaluate its ability to successfully answer questions that have a concrete answer. To do this, a Q&A dataset was constructed. This dataset consists of set of 12 GIS-related questions with corresponding correct answers. For each record in the dataset, a description of how a human would find it natural to approach the problem. This description is provided as a step-by-step path towards the solution, and is only included to guide any reader as to how the system would be expected to solve the system. The full Q&A dataset can be found in Table 6.3. This set of experiments will allow for quantitative assessment of GeoGPT’s GIS-abilities, and is a feasible way of benchmarking the system.

Another aspect that the benchmarking approach will try to evaluate is the consistency of the system, its ability to repeatedly provide an acceptable answer to the same user question. Each of the 12 questions are therefore asked three times per agent type.

With the implementation of three different agent types, the total number of test runs becomes the following:

$$12 \text{ questions} \cdot 3 \text{ agent types} \cdot 3 \text{ repetitions} = 108 \text{ tests}$$

Outcome Evaluation

Each test run’s answer will be manually evaluated, and the outcome will be annotated as one of *success*, *partial success*, and *failure*. Table 6.1 shows the guidelines used when assigning test results.

The annotated outcomes are then encoded using the ordinal encoding presented Table 6.2. A higher value indicates a better outcome. These encoded outcome values enable standard deviation calculations, which serve as a suitable measure for assessing repeatability. This approach also allows for comparisons across different agent types and configurations.

6. Experiments and Results

Table 6.1.: Description of Success

Outcome	Guideline
Success	The question was answered correctly and little to no follow-up from the user was required to produce the desired outcome. No false assumptions were made by the system when answering the question.
Partial Success	Portions of the question were answered correctly or semi-correctly, and/or some follow-up from the user was required to guide the system toward the solution.
Failure	The question was answered incorrectly and/or false assumptions were made by the system while attempting to answer the question.

Table 6.2.: Encoding for Test Outcome

Outcome	Encoded Value
Success	2
Partial Success	1
Failure	0

6. Experiments and Results

Other Metrics

The application is hooked up to LangChain AI's tracing system, *LangSmith*. Apart from being a useful tool for debugging purposes, it provides a simple way of obtaining detailed data for token and time usage for a particular run, as well as the total cost of the run. These are metrics that will be recorded and used in the evaluation of GeoGPT.

To summarize, the following metrics are recorded for a given test run:

- The outcome of the test (*success*, *partial success*, or *failure*)
- The total duration in seconds
- The total number of tokens used
- The total cost for the run in American dollar

Table 6.3.: Questions for quantitative tests

Query ID	Query	Correct Response
aker_brygge_national	Which is the closest railway station to Aker brygge?	Nationalteateret
cliff_clusters	Locate clusters of cliffs, each containing more than 10 cliffs, with cliffs within each cluster no more than 0.1 degrees apart.	Should be about 8-9 clusters.
county_names	What are the names of the counties found in the data?	Nordland, Telemark, Troms, Rogaland, Vestland, Trøndelag, Vestfold, Buskerud, Akershus, Østfold, Innlandet, Møre og Romsdal, Agder Finnmark
glomma_counties	How many counties does Glomma run through?	4. Trøndelag, Innlandet, Akershus, and Østfold.
largest_county	Which is the largest county by size?	Nordland
nidarosdomen_polygon	Retrieve a polygon of Nidarosdomen.	Adding a polygon of Nidarosdomen to the map.

Continued on next page

6. Experiments and Results

Table 6.3.: Questions for quantitative tests

Query ID	Query	Correct Response
num_trees_munkegata	How many trees are there along Munkegata in Trondheim?	Giving the correct number of trees (about 70-80).
oslo_bergen_geodesic	Create a geodesic curve between the airports of Oslo and Bergen.	A geodesic, slightly curved line between Gardermoen and Flesland.
oslo_residential_diff	Provide an outline of Oslo but exclude residential areas by computing their difference.	The polygonal outline of Oslo with cutouts where there area areas classified as residential.
oslo_roads_gte_70_kmh	Retrieve roads in Oslo that have speed limit higher than or equal to 70 km/h.	Adding corresponding line segments to the map.
vestfold_bbox	Find the coordinates that define a the extent of a rectangular bounding box that encompasses the county of Vestfold.	(9.7553357, 58.720455, 10.6750198, 59.674011)
viken_dissolve	Combine the countries that correspond to the historical country of Viken into a single feature.	Display the outline of Viken, based on the outlines of Buskerud, Akershus, and Østfold, that are found in the data. Should be dissolved, i.e., no borders.

6.1.2. Evaluating Importance of Prompt Quality

The second set of experiments are constructed to evaluate the importance of the initial question/prompt from the human user. As stated in Background and Motivation, part of the motivation for developing an LLM-driven GIS like GeoGPT is to make GIS more accessible to non-experts. At the same time, it may be valuable to assess the extent to which a carefully constructed prompt by a GIS expert can enhance the system's output.

6. Experiments and Results

6.1.3. Configuration and Hardware

All experiments were executed locally on a Lenovo ThinkPad E490, which has an Intel® Core™ i7-8565U CPU @ 1.80GHz processor, 15.8 GB usable RAM, and 256 GB SSD storage. Everything but the LLM inference was executed locally. Text generation was done using OpenAI's API.

It is worth noting that two slightly different models were used during testing. The explanation of this is the release of the `gpt-4-turbo-2024-04-09` in mid-April. According to OpenAI, “this new model is better at math, logical reasoning, and coding” compared to `gpt-4-0125-preview`¹, which is the model that was used at the start of the experimentation phase of the master’s thesis. At the new model’s release, a decision was made to use this for the remaining experiments. The experiments that had already been conducted were not re-run due to time constraints and a belief that these slight model upgrades would not significantly change the outcome of the experiments.

6.2. Experimental Results

Subsection 6.2.1 and subsection 6.2.2 will present the outcome of the experiments presented in subsection 6.1.2 and subsection 6.1.2, respectively.

6.2.1. Quantitative Results

Graphs created for this chapter are created using Matplotlib², a Python library suitable for creating visualizations like bar charts, box plots, etc.

Test Outcomes

As described in subsection 6.1.1, a total 108 tests were run using the 12 available Q&A samples, with the same question being repeated three times for each of the three agent types , resulting in 36 test runs per agent. Figure 6.1 displays a bar chart for the outcome distribution per agent.

From Figure 6.1, we can read that the OGC API Features and Python agent have comparable results, and that the SQL-based agent performs significantly better compared to the other two in terms of producing the desired outcome.

ref some
table
with
agent
types

Other Metrics

This section features three different box plots: Figure 6.2, Figure 6.3a, and Figure 6.3b. The Matplotlib implementation of the box plot follows the description found on

¹OpenAI has a GitHub repository containing the code they use to evaluate their Large Language Models (LLMs) and benchmark results for OpenAI models and reference models from other companies: <https://github.com/openai/simple-evals>.

²<https://matplotlib.org/>

6. Experiments and Results



Figure 6.1.: Outcome distribution between different agent types

6. Experiments and Results

Wikipedia^{3,4}. Box plots allow us to easily visualize where the 0th (Q_0), 25th (Q_1), 50th (Q_2), 75th (Q_3), and 100th (Q_4) percentiles of the datasets lie, as well as the dataset's outliers. Outliers are those data points fall outside 1.5 times interquartile range, that is, the distance between Q_3 and Q_1 in each direction.

Figure 6.2 displays a box plot with a logarithmic y-axis showing the relative durations for task completion across the different agent types. Here we can see that the SQL agent spends the least amount of time per task, and that the OGC API Features agent has a slightly higher median but with a few time-consuming outliers. The Python agent is the odd one out with a median of ~ 82 seconds, a $Q_3 \sim 293$ seconds, and a $Q_4 \sim 984$. The large gap to the other two agents it largely due to the Python agent's tendency to load large datasets into memory without filtering the data on load using a bounding box. For instance, when attempting the task of calculating the difference between the polygon outlining Oslo and water polygons, the Python agent used nearly 40 minutes on the entire task. 94% of the time was spent executing the code presented in Code 6.1.

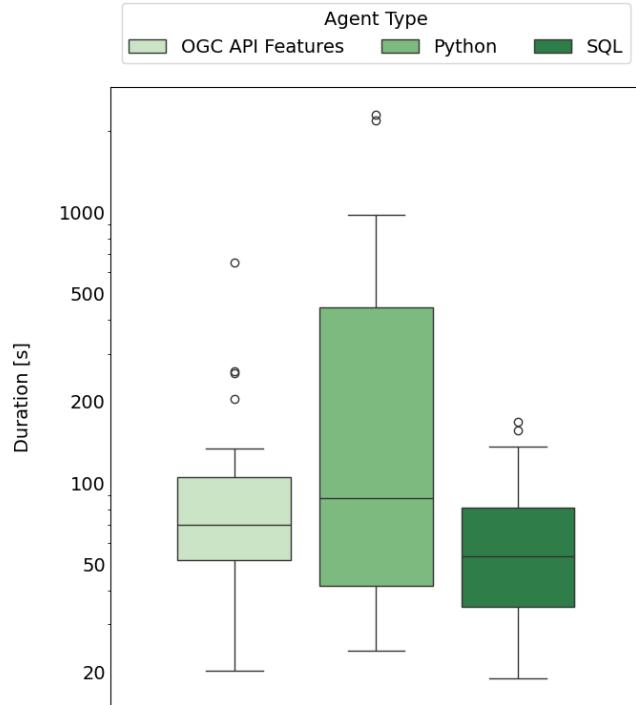


Figure 6.2.: Duration per Agent Type

```
1 import geopandas as gpd
```

³https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.boxplot.html

⁴https://en.wikipedia.org/wiki/Box_plot

6. Experiments and Results

```

2      # Paths to the shapefiles
3      landuse_path = '/tmp/tmpsutdy6it/osm_landuse_polygons.shp'
4      places_path = '/tmp/tmpsutdy6it/osm_places_polygons.shp'
5
6      # Load the data from shapefiles
7      landuse_gdf = gpd.read_file(landuse_path)
8      places_gdf = gpd.read_file(places_path)
9
10     # Filter out 'residential' areas from the landuse data
11     residential_gdf = landuse_gdf[landuse_gdf['fclass'] == 'residential']
12
13     # Compute the spatial difference to exclude residential areas from
14     # the places data
15     oslo_outline = gpd.overlay(places_gdf, residential_gdf, how='difference')
16
17     # Path for the output GeoJSON file
18     output_path = '/tmp/tmpsutdy6it/oslo_outline_no_residential.geojson'
19
20     # Save the resulting GeoDataFrame to a GeoJSON file
21     oslo_outline.to_file(output_path, driver='GeoJSON')
22
23     # Output the path to the saved file
24     print(output_path)

```

Code 6.1: Python code aimed at computing the difference between the Oslo outline and water features within it

Figure 6.3 shows box plots for the amount of tokens used per run and the total cost of these. Naturally, these figures appear very similar, as OpenAI sets fixed prices for the input tokens and generated output tokens for their GPT models. According to their websites⁵, they charge \$10 per million input tokens and \$30 per million output tokens for their GPT-4 Turbo models. From the results of the experiments, a ratio of approximately 10.7 per million tokens — either input *or* output — was calculated, which is very close to the input token price. This aligns with the observation that the number of input/prompt tokens is significantly greater than the number of output/generated tokens for the experiments conducted. The correlation matrix in Figure 6.4 confirms that this ratio is consistent.

Another observation that can be made from Figure 6.4 is that there is only a *slight* positive correlation between duration and token/usage. This supports the observation that code execution time is likely the most significant factor in determining the duration required to complete a task for the datasets and tasks used in this thesis’ experiments.

A third observation that can be made from the Figure 6.4 is the negative correlation

⁵<https://openai.com/pricing>

6. Experiments and Results

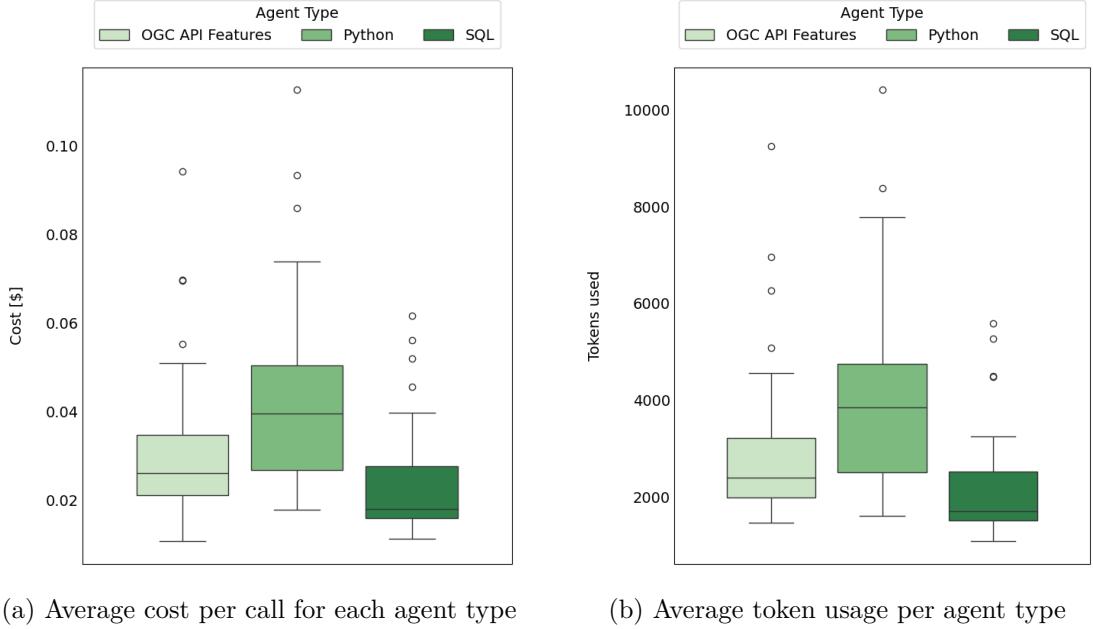


Figure 6.3.: Cost and token usage

between the encoded outcome and token usage, duration, and total cost. This suggests that a task that takes longer to complete — and thus is likely to be more expensive in terms of token usage — is more likely to produce an undesirable outcome. Possible reasons as to why this is the case will be explored in the Evaluation and Discussion.

Repeatability

Table 6.4 shows the average standard deviation for each agent type, as well as the mean of these three standard deviations. The latter serves as an overall measure of GeoGPT’s ability to repeatedly produce the desired outcome from a given query. Standard deviations were calculated for each triplet of identical test samples, in which both the question and the agent type remained the same. To produce a numerical value for the standard deviations, the encoded outcomes (see Table 6.2) were used. Taking the average of the standard deviations for all 12 triplets for each agent type produced the numbers found in Table 6.4.

6.2.2. Prompt Quality Test Results

Figure 6.6 shows a comparison between the results GeoGPT managed to produce for two different prompts that one would expect to produce identical outcomes. The *novice*-level prompt was as follows:

“Could you count how many trees there are on Munkegata street in Trondheim?”

6. Experiments and Results

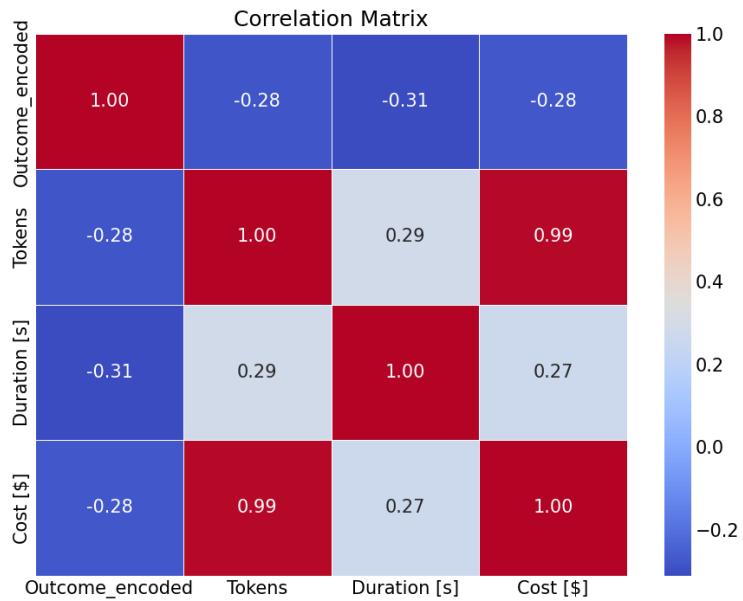


Figure 6.4.: Correlation matrix for test result metrics

Table 6.4.: Standard Deviation by Agent Type

Agent Type	Outcome Std. Deviation
OGC API Features	0.552
Python	0.337
SQL	0.337
Mean	0.408

6. Experiments and Results

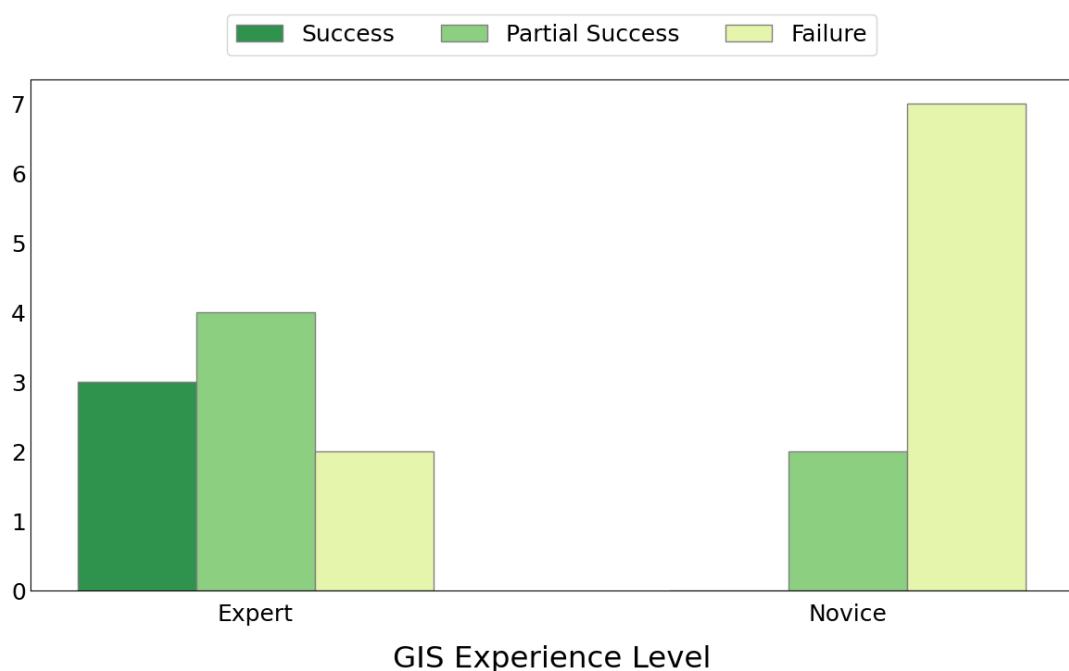


Figure 6.5.: Outcome distribution for different levels of GIS experience

6. Experiments and Results

The *expert*-level prompt, on the other hand, included a series of instructions:

- “1. List all datasets that could possibly include trees.
2. Find the correct feature class and filter the relevant dataset to access tree data for Trondheim. Use a bounding box to reduce the number of trees to analyse.
3. Fetch road data for Munkegata. Use a bounding box for Trondheim in case there are streets elsewhere named Munkegata.
4. Convert both datasets to a suitable metric CRS and add a 20-meter buffer around the road data.
5. Find all trees that lie within this buffer and count them.
6. Present the findings with a map highlighting the roads and the trees.”

Using novice-level prompt GeoGPT was unable to produce the correct outcome, and confidently answered that there are “approximately 6,915 trees on Munkegata street in Trondheim.” (see Figure 6.6a), which is far from being true. When solving the task, GeoGPT made a series of oversights that lead to this result. First, GeoGPT failed to take into account that there may be more than one street in the dataset with the name “Munkegata”, “forgetting” to use a bounding box when retrieving the road data from the API. The same mistake was made when retrieving the tree data. Due to the upper limit of 10,000 features per query in the API, it’s crucial to narrow down the query to ensure retrieval of all relevant features rather than just a subset. GeoGPT’s query for tree data lacked a bounding box, resulting in a randomly distributed subset scattered across Norway. A third mistake occurred when GeoGPT calculated a bounding box around the retrieved road data instead of creating a buffer. The latter method would have produced a more accurate result. The bounding box that was created spanned from Trondheim to Oslo, thus including far more trees than was intended.

On the other hand, the expert-level prompt provided the necessary guidance for GeoGPT for this specific task, steering it clear of the issues it encountered with the novice-level prompt. As stated by OpenAI themselves⁶, “some tasks are best specified as a sequence of steps”. Furthermore, they say that writing explicit steps required to solve a tasks “makes it easier for the model to follow them”.

move to
discus-
sion?

⁶<https://platform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions>

6. Experiments and Results

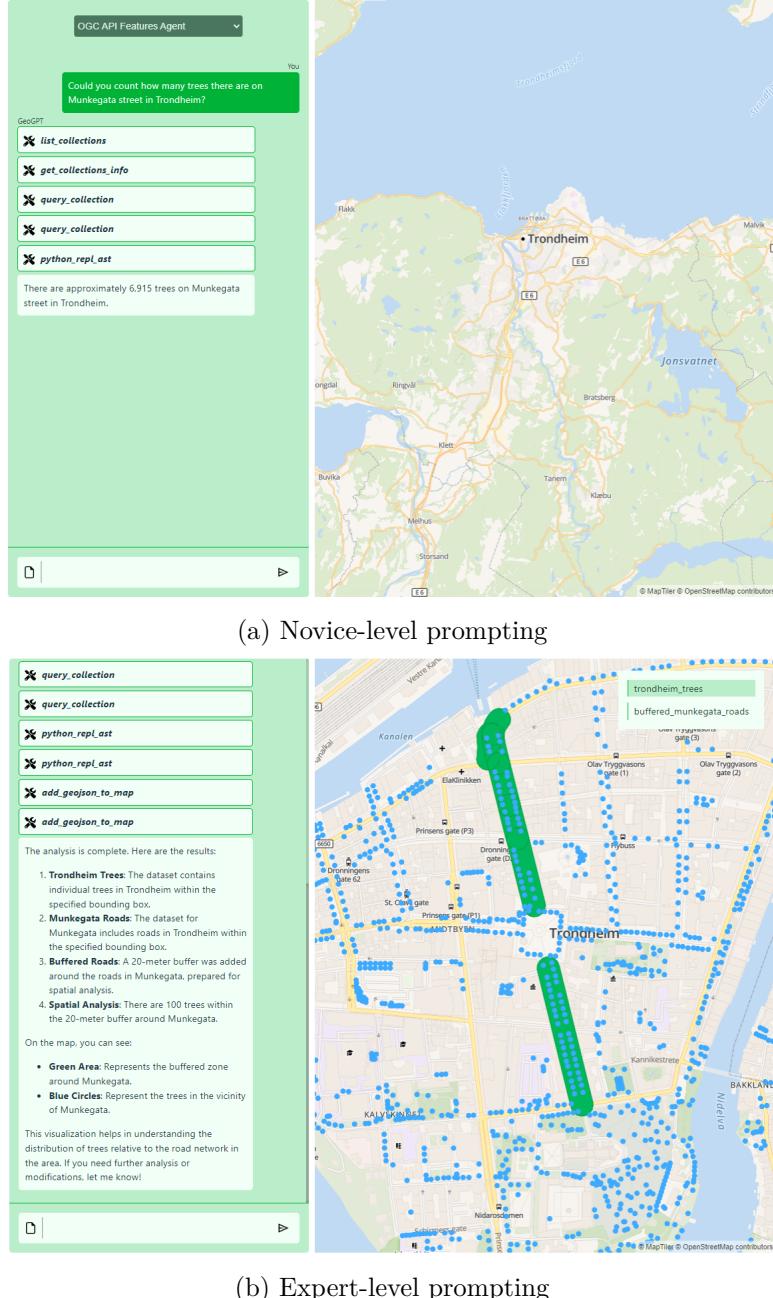


Figure 6.6.: Comparison between novice and expert level prompting for making GeoGPT’s OGC API Features agent calculate the number of trees along Munkegata in Trondheim

7. Evaluation and Discussion

7.1. Evaluation

7.2. Discussion

8. Conclusion and Future Work

8.1. Contributions

8.2. Future Work

8.2.1. Automated Data Access

The experiments in chapter 6 were based upon a pre-existing database. A fully autonomous GIS agent should, however, be able to search the web for suitable datasets, based on the user's query. In a Norwegian context, one could imagine asking for a noise analysis for a particular building. The agent would then search Geonorge for datasets related to noise (firing ranges, roads, etc.), downloading these, and then performing analysis. Simple experiments were conducted in this thesis to see if this was possible, but results were somewhat poor. Methods like semantic search based upon the documentations of datasets should be explored in future research.

Bibliography

- CrunchyData. (2024). CrunchyData/pg_featureserv. Retrieved April 19, 2024, from https://github.com/CrunchyData/pg_featureserv
- Eleti, A., Harris, J., & Kilpatrick, L. (2023). Function calling and other API updates. Retrieved March 10, 2024, from <https://openai.com/blog/function-calling-and-other-api-updates>
- Holm, O. (2023). *LLMs - The Death of GIS Analysis?* (Specialization Project). NTNU. Trondheim. <https://github.com/oskarholm/prosjektoppgave/blob/main/tex/auxiliary/main.pdf>
- Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2023). MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. <https://doi.org/10.48550/arXiv.2308.00352>
- LangChain AI. (2022). Langchain-ai/langchain. Retrieved October 5, 2023, from <https://github.com/langchain-ai/langchain>
- LangChain AI. (2024). Langchain-ai/langgraph. Retrieved March 21, 2024, from <https://github.com/langchain-ai/langgraph>
- OpenAI. (2022). Introducing ChatGPT. Retrieved October 26, 2023, from <https://openai.com/blog/chatgpt>
- 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>
- Sanfilippo, S. (2009). Redis - The Real-time Data Platform. Retrieved April 19, 2024, from <https://redis.io/>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. Retrieved October 10, 2023, from <https://arxiv.org/abs/1706.03762v7>

Appendices

A. Test Results

Table A.1.: Test results for quantitative tests

Query ID	Agent Type	Outcome	Duration [s]	Tokens
aker_brygge_national	oaf	partial success	87.16	3347
aker_brygge_national	oaf	partial success	97.85	3042
aker_brygge_national	oaf	partial success	70.36	4563
aker_brygge_national	python	partial success	68.10	4736
aker_brygge_national	python	partial success	49.48	2192
aker_brygge_national	python	partial success	76.84	4270
aker_brygge_national	sql	success	129.03	4498
aker_brygge_national	sql	partial success	114.39	5269
aker_brygge_national	sql	success	95.83	5584
cliff_clusters	oaf	partial success	93.21	3047
cliff_clusters	oaf	partial success	134.62	3713
cliff_clusters	oaf	failure	253.77	6956

Continued on next page

A. Test Results

Table A.1.: Test results for quantitative tests

Query ID	Agent Type	Outcome	Duration [s]	Tokens
cliff_clusters	python	failure	66.42	5421
cliff_clusters	python	failure	125.96	4650
cliff_clusters	python	failure	109.29	4768
cliff_clusters	sql	success	65.47	1909
cliff_clusters	sql	success	61.22	2837
cliff_clusters	sql	success	156.00	4485
county_names	oaf	success	78.08	2389
county_names	oaf	success	67.45	2399
county_names	oaf	failure	20.23	1471
county_names	python	success	44.33	2532
county_names	python	success	30.50	2355
county_names	python	success	31.09	2355
county_names	sql	success	18.89	1149
county_names	sql	success	19.74	1169
county_names	sql	partial success	47.36	2489
glomma_counties	oaf	success	69.73	2253
glomma_counties	oaf	success	65.57	1945
glomma_counties	oaf	success	67.37	2397
glomma_counties	python	success	663.22	3658
glomma_counties	python	failure	295.93	4388
glomma_counties	python	success	285.23	2281
glomma_counties	sql	success	36.54	1231
glomma_counties	sql	success	63.65	1810

Continued on next page

A. Test Results

Table A.1.: Test results for quantitative tests

Query ID	Agent Type	Outcome	Duration [s]	Tokens
glomma_counties	sql	success	22.81	1079
largest_county	oaf	success	41.82	1713
largest_county	oaf	success	71.82	2223
largest_county	oaf	failure	47.00	1797
largest_county	python	failure	29.37	2415
largest_county	python	failure	41.95	2810
largest_county	python	failure	40.55	2407
largest_county	sql	success	37.52	1654
largest_county	sql	success	27.55	1483
largest_county	sql	success	41.22	1765
nidarosdomen_polygon	oaf	success	29.82	1923
nidarosdomen_polygon	oaf	success	31.73	1909
nidarosdomen_polygon	oaf	success	31.61	1922
nidarosdomen_polygon	python	success	35.10	4510
nidarosdomen_polygon	python	success	25.36	3649
nidarosdomen_polygon	python	success	33.35	3663
nidarosdomen_polygon	sql	success	34.60	1495
nidarosdomen_polygon	sql	success	63.43	1519
nidarosdomen_polygon	sql	success	34.77	1518
num_trees_munkegata	oaf	failure	75.60	2977
num_trees_munkegata	oaf	failure	80.13	2984
num_trees_munkegata	oaf	failure	57.66	2407
num_trees_munkegata	python	partial success	122.99	3930

Continued on next page

A. Test Results

Table A.1.: Test results for quantitative tests

Query ID	Agent Type	Outcome	Duration [s]	Tokens
num_trees_munkegata	python	failure	916.13	2362
num_trees_munkegata	python	failure	911.18	8385
num_trees_munkegata	sql	failure	63.93	1831
num_trees_munkegata	sql	failure	77.45	3086
num_trees_munkegata	sql	failure	64.47	1640
oslo_bergen_geodesic	oaf	failure	652.88	9253
oslo_bergen_geodesic	oaf	failure	102.38	3360
oslo_bergen_geodesic	oaf	failure	258.52	6271
oslo_bergen_geodesic	python	partial success	118.54	6683
oslo_bergen_geodesic	python	partial success	129.40	7782
oslo_bergen_geodesic	python	failure	373.37	10412
oslo_bergen_geodesic	sql	failure	80.49	3253
oslo_bergen_geodesic	sql	failure	83.78	2466
oslo_bergen_geodesic	sql	failure	88.63	2825
oslo_residential_diff	oaf	failure	119.39	2373
oslo_residential_diff	oaf	partial success	114.92	3033
oslo_residential_diff	oaf	success	121.17	2621
oslo_residential_diff	python	success	673.96	4013
oslo_residential_diff	python	failure	2190.08	3867
oslo_residential_diff	python	failure	2290.09	3650
oslo_residential_diff	sql	success	168.13	1944
oslo_residential_diff	sql	failure	94.40	2536

Continued on next page

A. Test Results

Table A.1.: Test results for quantitative tests

Query ID	Agent Type	Outcome	Duration [s]	Tokens
oslo_residential_diff	sql	failure	136.12	2511
oslo_roads_gte_70_kmh	oaf	partial success	44.44	1959
oslo_roads_gte_70_kmh	oaf	failure	58.61	1989
oslo_roads_gte_70_kmh	oaf	partial success	35.55	1961
oslo_roads_gte_70_kmh	python	failure	983.76	3912
oslo_roads_gte_70_kmh	python	failure	757.70	3819
oslo_roads_gte_70_kmh	python	failure	849.06	3827
oslo_roads_gte_70_kmh	sql	partial success	59.47	1571
oslo_roads_gte_70_kmh	sql	partial success	44.79	1565
oslo_roads_gte_70_kmh	sql	partial success	41.67	1548
vestfold_bbox	oaf	failure	56.33	2001
vestfold_bbox	oaf	partial success	50.38	3177
vestfold_bbox	oaf	success	55.82	2480
vestfold_bbox	python	success	24.08	1612
vestfold_bbox	python	success	33.63	2366
vestfold_bbox	python	success	42.36	2808
vestfold_bbox	sql	success	25.81	1170
vestfold_bbox	sql	failure	32.59	1561
vestfold_bbox	sql	success	20.00	1192
viken_dissolve	oaf	success	52.24	2336

Continued on next page

A. Test Results

Table A.1.: Test results for quantitative tests

Query ID	Agent Type	Outcome	Duration [s]	Tokens
viken_dissolve	oaf	success	204.07	5078
viken_dissolve	oaf	failure	133.40	3985
viken_dissolve	python	success	68.23	6099
viken_dissolve	python	success	86.32	6007
viken_dissolve	python	partial success	89.98	6461
viken_dissolve	sql	partial success	26.49	1447
viken_dissolve	sql	success	47.59	1963
viken_dissolve	sql	partial success	36.20	1555

B. Code

B.1. Python Example

```
1      import numpy as np
2
3      def incmatrix(genl1,genl2):
4          m = len(genl1)
5          n = len(genl2)
6          M = None #to become the incidence matrix
7          VT = np.zeros((n*m,1), int) #dummy variable
8
9          #compute the bitwise xor matrix
10         M1 = bitxormatrix(genl1)
11         M2 = np.triu(bitxormatrix(genl2),1)
12
13         for i in range(m-1):
14             for j in range(i+1, m):
15                 [r,c] = np.where(M2 == M1[i,j])
16                 for k in range(len(r)):
17                     VT[(i)*n + r[k]] = 1;
18                     VT[(i)*n + c[k]] = 1;
19                     VT[(j)*n + r[k]] = 1;
20                     VT[(j)*n + c[k]] = 1;
21
22         if M is None:
23             M = np.copy(VT)
24         else:
25             M = np.concatenate((M, VT), 1)
26
27         VT = np.zeros((n*m,1), int)
28
29     return M
30
31     import numpy as np
32
33     def incmatrix(genl1,genl2):
34         m = len(genl1)
35         n = len(genl2)
36         M = None #to become the incidence matrix
37         VT = np.zeros((n*m,1), int) #dummy variable
38
39         #compute the bitwise xor matrix
40         M1 = bitxormatrix(genl1)
```

B. Code

```

41      M2 = np.triu(bitxormatrix(genl2),1)
42
43      for i in range(m-1):
44          for j in range(i+1, m):
45              [r,c] = np.where(M2 == M1[i,j])
46              for k in range(len(r)):
47                  VT[(i)*n + r[k]] = 1;
48                  VT[(i)*n + c[k]] = 1;
49                  VT[(j)*n + r[k]] = 1;
50                  VT[(j)*n + c[k]] = 1;
51
52      if M is None:
53          M = np.copy(VT)
54      else:
55          M = np.concatenate((M, VT), 1)
56
57      VT = np.zeros((n*m,1), int)
58
59      return M
60
61
62      import numpy as np
63
64      def incmatrix(genl1,genl2):
65          m = len(genl1)
66          n = len(genl2)
67          M = None #to become the incidence matrix
68          VT = np.zeros((n*m,1), int) #dummy variable
69
70          #compute the bitwise xor matrix
71          M1 = bitxormatrix(genl1)
72          M2 = np.triu(bitxormatrix(genl2),1)
73
74          for i in range(m-1):
75              for j in range(i+1, m):
76                  [r,c] = np.where(M2 == M1[i,j])
77                  for k in range(len(r)):
78                      VT[(i)*n + r[k]] = 1;
79                      VT[(i)*n + c[k]] = 1;
80                      VT[(j)*n + r[k]] = 1;
81                      VT[(j)*n + c[k]] = 1;
82
83          if M is None:
84              M = np.copy(VT)
85          else:
86              M = np.concatenate((M, VT), 1)
87
88          VT = np.zeros((n*m,1), int)
89
90      return M
91
92      import numpy as np
93

```

B. Code

```
94     ya = "hei"
95
96     def incmatrix(genl1,genl2):
97         m = len(genl1)
98         n = len(genl2)
99         M = None #to become the incidence matrix
100        VT = np.zeros((n*m,1), int) #dummy variable
101
102        #compute the bitwise xor matrix
103        M1 = bitxormatrix(genl1)
104        M2 = np.triu(bitxormatrix(genl2),1)
105
106        for i in range(m-1):
107            for j in range(i+1, m):
108                [r,c] = np.where(M2 == M1[i,j])
109                for k in range(len(r)):
110                    VT[(i)*n + r[k]] = 1;
111                    VT[(i)*n + c[k]] = 1;
112                    VT[(j)*n + r[k]] = 1;
113                    VT[(j)*n + c[k]] = 1;
114
115        if M is None:
116            M = np.copy(VT)
117        else:
118            M = np.concatenate((M, VT), 1)
119
120        VT = np.zeros((n*m,1), int)
121
122    return M
123
```

Code B.1: Python example