Automating Geospatial Analysis Workflows Using ChatGPT-4

Qianheng Zhang
Graduate Student, Geospatial Data Science Lab,
University of Wisconsin
Madison, WI, USA
qianheng.zhang@wisc.edu

Song Gao
Advisor, Geospatial Data Science Lab,
University of Wisconsin
Madison, WI, USA
song.gao@wisc.edu

Abstract

The field of Geospatial Artificial Intelligence (GeoAI) has significantly impacted domain applications such as urban analytics, environmental monitoring, and disaster management. While powerful geoprocessing tools in geographic information systems (GIS) like ArcGIS Pro are available, automating these workflows with Python scripting using AI chatbots remains a challenge, especially for non-expert users. This study investigates whether ChatGPT-4 can automate GIS workflows by generating ArcPy functions based on structured instructions. We tested prompt engineering's ability on helping large language models (LLMs) understand spatial data and GIS workflows. The overall task success rate reaches 80.5%. It is a valid and easy to implement approach for domain scientists who want to use ArcPy to automate their workflows.

CCS Concepts

• **Information systems → Geographic information systems**.

Keywords

Prompt engineering, LLM, automate workflow, GIS, GeoAI

ACM Reference Format:
Qianheng Zhang and Song Gao. 2024. Automating Geospatial Analysis Workflows Using ChatGPT-4. In *The 32nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24), October 29-November 1, 2024, Atlanta, GA, USA*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3678717.3695760

1 Introduction and Motivation

Since the introduction of large language models (LLMs), there are emerging studies on evaluating LLM agents' capabilities on automating scientific workflows in data analysis [4], workflow design and code generation [1, 7], etc., using natural languages. These approaches allow domain science experts with limited programming skills to solve scientific problems [3]. Specifically, the field of GIS, which is essential for spatial data analysis and visualization, researchers are trying to understand whether LLMs can retrieve the right dataset, understand complicated spatial relations between geographic entities, and perform executable codes [5, 6, 8]. Although some of these tasks are basic, it's promising that current LLMs can solve certain geoprocessing problems with a high accuracy.

However, while most geospatial analysis tasks being studied use Python given its powerful open-source packages, *ArcPy* in ArcGIS Pro by ESRI remains unexplored. ArcPy is a Python package designed for automating geoprocessing workflows within the software, but it might be difficult for domain scientists who are not GIS experts. Compared to *GeoPandas* and *RasterIo*, ArcPy provides access to numerous GIS tools under different domains, which often require a high-level knowledge to use the tools to connect different data layers. As the Figure 1 shows, raw data goes through several data processing functions to generate output. For example, a raw data layer has to be 'Buffered' into data layer A, which becomes layer B by 'Union', then layer B is filtered by 'Select By Attribute' before the final output. Because of that, the level of expertise that some domain scientists may lack creates a barrier to broader applications. The specialized nature of ArcPy and its critical role in automating geospatial workflows highlights the current gap. To solve this gap, the research questions are:

RQ1: How effectively can ChatGPT-4 generate accurate and reliable ArcPy functions for automating GIS workflows?

RQ2: What are the common challenges and limitations faced by ChatGPT-4 in handling complex geospatial tasks and applying correct tools?

To answer the first research question, RQ1, which involves evaluating the ability of ChatGPT-4 to generate accurate ArcPy functions, we aim to explore the automation of various GIS workflows. Each task is based on real-world geospatial questions faced by domain experts, covering core geospatial concepts such as *understanding places, determining relationships, finding locations, and detecting patterns*. These topics reflect the range of geospatial tasks that GIS professionals regularly encounter, ensuring that the evaluation captures a broad spectrum of spatial analysis challenges.

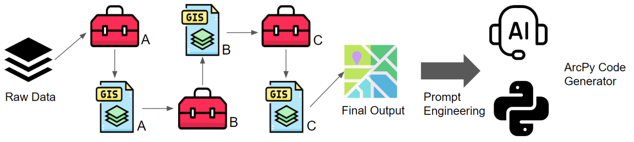

Figure 1: Raw data layers are processed through various tools to produce the output. This process can be facilitated by LLMs, making advanced spatial tools accessible.

2 Automating the Workflow

To operationalize this, we constructed a comprehensive labeled benchmark dataset consisting of 41 sub-tasks organized under 14 major tasks derived from ESRI's tutorials. The dataset, which is

SIGSPATIAL '24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1107-7/24/10
https://doi.org/10.1145/3678717.3695760

publicly available, allows for reproducibility and benchmarking. Most tasks typically performed in ArcGIS Pro can also be executed using Python code via ArcPy geoprocessing toolboxes. By comparing ChatGPT-4's performance across different workflows and task complexities, we are able to address RQ2 and assess the model's robustness in handling varying levels of difficulty in spatial analysis.

2.1 Experiment

Our approach involves a systematic framework to evaluate ChatGPT-4's performance on automating geospatial analysis workflows:

1. Data Preparation: We created a CSV file containing sub-tasks derived from the ESRI tutorials [2]. Each entry includes columns for sub-task ID, instruction, expected tool/function, and its parameters. For example, the task "Find liquor stores within 1,000 feet of a school, park, or library" involves using Buffer tool to create zones of influence around each feature.
2. Prompt Generation: For each sub-task, we crafted a structured prompt template incorporating the previous, current, and next instructions. This context is crucial for guiding ChatGPT in generating the precise ArcPy functions required. This ensures that the agent can successfully output the necessary instruction for the current task. Our method improves reliability and stability by focusing on transforming the data layer from the previous step to the required data layer for the next step. All prompts use one shot prompting, which means one example is given in prompt.
3. Model Invocation: Utilizing LangChain and OpenAI APIs, we invoked the ChatGPT-4 to generate ten different outputs for each sub-task with the generated prompts from the last step. We ran each task with temperature setting of 0.0 for 10 times. This multi-sample approach enhances the robustness of our evaluation [1, 4].
4. Evaluation and Error Analysis: The outputs were assessed using the Pass@3 metric, which measures the model's ability to produce at least one correct solution out of three attempts. This approach is more practical than traditional text similarity measurements for evaluating LLMs [9].

2.2 Results

Main Task	Sub-task Count	Average Success Rate
Understand Places	11	72.7%
Determine Relationships	14	92.9%
Find Locations	5	80%
Detect Patterns	11	72.7%
All	41	80.5%

Table 1: Task Performance Summary

As shown in Table 1, the experiment achieved an overall success rate of 80.5% for generating correct Python functions. The most common errors stemmed from using incorrect function names or selecting the right functions from the wrong toolboxes, suggesting that ChatGPT-4's internal knowledge of ArcPy may be based on outdated versions. Additionally, generating the correct parameter settings from abstract instructions without context is challenging with a 64.1% success rate. A notable pattern observed was the model's consistency in generating similar content across multiple rounds, both for correct and incorrect answers.

The model demonstrated high accuracy for simpler geoprocessing tools such as buffer, dissolve, and clip, which involve fewer operations and parameters. However, it struggled with more complex spatial analysis tasks, such as hot spot analysis and statistical summaries, which require multiple steps and precise parameters. These complex tasks also appeared less frequently in the dataset, and the LLM was less likely to be familiar with the latest versions of the required tools. This suggests that additional training or fine-tuning using up-to-date GIS toolbox descriptions could significantly enhance the model's performance.

3 Discussion and Future Works

The experimental results highlight both the potential and limitations of using ChatGPT-4 for automating GIS workflows. The 80.5% success rate is promising but also reveals several areas for improvement, particularly in handling complex geospatial tasks and applying correct function parameters. To address these issues, future GeoAI research should focus on enhancing the model's contextual understanding and parameter handling.

A key future direction is the integration of Retrieval-Augmented Generation (RAG) methodology to reference external knowledge sources. The concise and up-to-date ArcPy database would allow the model to stay current with the latest API versions and reduce errors caused by outdated knowledge. Additionally, fine-tuning on specific GIS tool descriptions could improve the model's performance for complex tasks like hot spot or statistical analysis, which consistently failed during testing.

In conclusion, while ChatGPT-4 shows great promise in automating geospatial analysis workflows, future research should aim to address its limitations in handling complex tasks. By integrating external knowledge and improving model fine-tuning, LLMs-powered AI chatbots can be better equipped for automating advanced GIS analyses and supporting geospatial question-answering tasks.

References

[1] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).

[2] Esri. 2024. Case Studies Overview. https://desktop.arcgis.com/en/analytics/case-studies/case-studies-overview.htm. Accessed: 2024-07-25.

[3] Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. 2024. OpenAGI: When llm meets domain experts. *Advances in Neural Information Processing Systems* 36 (2024).

[4] Jacqueline Jansen, Artur Manukyan, and Altuna Akalin. 2023. Leveraging large language models for data analysis automation. *bioRxiv* (2023), 2023–12.

[5] Yuhan Ji and Song Gao. 2023. Evaluating the Effectiveness of Large Language Models in Representing Textual Descriptions of Geometry and Spatial Relations. In *GIScience 2023*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 1–6.

[6] Yongyao Jiang and Chaowei Yang. 2024. Is ChatGPT a Good Geospatial Data Analyst? Exploring the Integration of Natural Language into Structured Query Language within a Spatial Database. *ISPRS International Journal of Geo-Information* 13, 1 (2024), 26.

[7] Zhenlong Li and Huan Ning. 2023. Autonomous GIS: the next-generation AI-powered GIS. *International Journal of Digital Earth* 16, 2 (2023), 4668–4686.

[8] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. GeoLLM: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213* (2023).

[9] Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-tau Yih, Sida Wang, and Xi Victoria Lin. 2023. Lever: Learning to verify language-to-code generation with execution. In *International Conference on Machine Learning*. PMLR, 26106–26128.